Extended Unconstrained Features Model for Exploring Deep Neural Collapse

Tom Tirer [1] **Joan Bruna** [1 2]

Abstract

The modern strategy for training deep neural networks for classification tasks includes optimizing the network's weights even after the training error vanishes to further push the training loss toward zero. Recently, a phenomenon termed "neural collapse" (NC) has been empirically observed in this training procedure. Specifically, it has been shown that the learned features (the output of the penultimate layer) of within-class samples converge to their mean, and the means of different classes exhibit a certain tight frame structure, which is also aligned with the last layer's weights. Recent papers have shown that minimizers with this structure emerge when optimizing a simplified "unconstrained features model" (UFM) with a regularized cross-entropy loss. In this paper, we further analyze and extend the UFM. First, we study the UFM for the regularized MSE loss, and show that the minimizers' features can have a more delicate structure than in the cross-entropy case. This affects also the structure of the weights. Then, we extend the UFM by adding another layer of weights as well as ReLU nonlinearity to the model and generalize our previous results. Finally, we empirically demonstrate the usefulness of our nonlinear extended UFM in modeling the NC phenomenon that occurs with practical networks.

1. Introduction

Deep neural networks (DNNs) have led to a major improvement in classification tasks (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; He et al., 2016; Huang et al., 2017). The modern strategy for training these networks includes optimizing the network's weights even after the training error vanishes to further push the training loss toward zero (Hoffer et al., 2017; Ma et al., 2018; Belkin et al., 2019).

Recently, a phenomenon termed "neural collapse" (NC) has been empirically observed by Papyan et al. (2020) for such training with cross-entropy loss. Specifically, via experiments on popular network architectures and datasets, Papyan et al. (2020) showed four components of the NC: (NC1) The learned features (the output of the penultimate layer) of within-class samples converge to their mean (i.e., the intraclass variance vanishes); (NC2) After centering by their global mean, the limiting means of different classes exhibit a simplex equiangular tight frame (ETF) structure (see Definition 2.2); (NC3) The last layer's (classifier) weights are aligned with this simplex ETF; (NC4) As a result, after such a collapse, the classification is based on the nearest class center in feature space.

The empirical work in (Papyan et al., 2020) has been followed by papers that theoretically examined the emergence of collapse to simplex ETFs in simplified mathematical frameworks. Starting from (Mixon et al., 2020), most of these papers (e.g., (Lu & Steinerberger, 2022; Wojtowytsch et al., 2021; Fang et al., 2021; Zhu et al., 2021)) consider the "unconstrained features model" (UFM), where the features of the training data after the penultimate layer are treated as free optimization variables (disconnected from the samples). The rationale behind this model is that modern deep networks are extremely overparameterized and expressive such that their feature mapping can be adapted to any training data (e.g., even to noise (Zhang et al., 2021)).

While most existing papers consider cross-entropy loss, in this paper we focus on the mean squared error (MSE) loss, which has been recently shown to be powerful also for classification tasks (Hui & Belkin, 2020). (We note that the occurrence of neural collapse when training practical DNNs with MSE loss, and its positive effects on their performance, have been shown *empirically* in a very recent paper (Han et al., 2021)). We start with analyzing the (plain) UFM, showing that for the regularized MSE loss the collapsed features can have a more delicate structure than in the cross-entropy case (e.g., they may possess also orthogonality), which affects also the structure of the weights. Then, we extend the UFM by adding another layer of weights as well as ReLU nonlinearity to the model and generalize our

[1]Center for Data Science, New York University, New York [2]Courant Institute of Mathematical Sciences, New York University, New York. Correspondence to: Tom Tirer <tirer.tom@gmail.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

previous results.

Our contributions can be summarized as follows:

- We analyze the minima of the UFM with regularized MSE loss and show the effect of the bias term on the minimizers' structured collapse.
- We analyze the minima of a linear extended UFM and show its limitation in modeling (practical) depthwise NC behavior.
- We analyze the minima of a ReLU-based nonlinear extended UFM and show the structured collapse of the deepest features.
- We present an asymptotic analysis of the case where the features are a fixed perturbation around the structured collapse.
- We empirically demonstrate the usefulness of our nonlinear extended UFM in modeling the NC phenomenon that occurs in the training of practical networks.

2. Background and Related Work

In this section, we provide more details on the empirical NC phenomenon and its analysis via the unconstrained features model.

Consider a classification task with K classes and n training samples per class, i.e., overall $N := Kn$ samples. Let us denote by $\mathbf{y}_k \in \mathbb{R}^K$ the one-hot vector with 1 in its k-th entry and by $\mathbf{x}_{k,i} \in \mathbb{R}^D$ the i-th training sample of the k-th class. Most DNN-based classifiers can be modeled as

$$\boldsymbol{\psi}_{\boldsymbol{\Theta}}(\mathbf{x}) = \mathbf{W}\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}) + \mathbf{b},$$

where $\mathbf{h}_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^D \to \mathbb{R}^d$ is the feature mapping ($d \geq K$), and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]^\top \in \mathbb{R}^{K \times d}$ ($\mathbf{w}_k^\top$ denotes the k-th row of $\mathbf{W}$) and $\mathbf{b} \in \mathbb{R}^K$ are the last layer's classifier matrix and bias, respectively. $\boldsymbol{\Theta} = \{\mathbf{W}, \mathbf{b}, \boldsymbol{\theta}\}$ is the set of the trainable network parameters, which includes the parameters $\boldsymbol{\theta}$ of a nonlinear compositional feature mapping (e.g., $\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma(\mathbf{W}_L(\ldots\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}))\ldots)$ where $\sigma(\cdot)$ is an element-wise nonlinear function).

The network parameters are obtained by minimizing an empirical risk of the form

$$\min_{\boldsymbol{\Theta}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{W}\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_{k,i}) + \mathbf{b}, \mathbf{y}_k\right) + \mathcal{R}\left(\boldsymbol{\Theta}\right), \quad (1)$$

where $\mathcal{L}(\cdot,\cdot)$ is a loss function (e.g., cross-entropy or MSE) and $\mathcal{R}(\cdot)$ is a regularization term (e.g., squared L_2-norm). Let us denote the feature vector of the i-th training sample of the k-th class by $\mathbf{h}_{k,i}$ (i.e., $\mathbf{h}_{k,i} = \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_{k,i})$),

We now define the notions of (within-class/intraclass) feature collapse and the simplex ETF. We use $\mathbf{I}_K$ to denote the $K \times K$ identity matrix, $\mathbf{1}_K$ to denote the all-ones vector of size $K \times 1$, and $[K]$ to denote the set $\{1, 2, ..., K\}$.

Definition 2.1 (Collapse)**.** We say that the training phase exhibits a (within-class) collapse if all the feature vectors of each class are mapped to a single point, i.e.,

$$\mathbf{h}_{k,i_1} = \mathbf{h}_{k,i_2}$$

for all $k \in [K]$ and i_1, i_2 training samples of the k-th class.

Definition 2.2 (Simplex ETF)**.** The standard simplex equiangular tight frame (ETF) is a collection of points in $\mathbb{R}^K$ specified by the columns of

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

Consequently, the standard simplex ETF obeys

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

In this paper, we consider a (general) simplex ETF as a collection of points in $\mathbb{R}^d$ ($d \geq K$) specified by the columns of $\tilde{\mathbf{M}} \propto \sqrt{\frac{K}{K-1}}\mathbf{P}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$, where $\mathbf{P} \in \mathbb{R}^{d \times K}$ is an orthonormal matrix. Consequently, $\tilde{\mathbf{M}}^\top\tilde{\mathbf{M}} \propto \frac{K}{K-1}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$.

Papyan et al. (2020) empirically showed that training networks after reaching zero training error leads to collapse of the features: they converge to K inter-class means that form a simplex ETF. Moreover, the last layer's weights $\{\mathbf{w}_k\}$ are also aligned (i.e., equal up to a scalar factor) to the same simplex ETF, and as a result, the classification turns to be based on the nearest class center in feature space. This "neural collapse" (NC) behavior has led to many follow-up papers (Mixon et al., 2020; Lu & Steinerberger, 2022; Wojtowytsch et al., 2021; Fang et al., 2021; Zhu et al., 2021; Graf et al., 2021; Ergen & Pilanci, 2021; Zarka et al., 2021; Ji et al., 2021). Some of them include practical implications of the NC phenomenon, such as designing layers (multiplication by tight frames followed by soft-thresholding) that concentrate within-class features (Zarka et al., 2021) or fixing the last layer's weights to be a simplex ETF (Zhu et al., 2021).

To mathematically show the emergence of a collapse to simplex ETF, most follow-up papers have considered a simplified framework — the "unconstrained features model" (UFM), where the features $\{\mathbf{h}_{k,i}\}$ are treated as free optimization variables

$$\min_{\mathbf{W},\mathbf{b},\{\mathbf{h}_{k,i}\}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{W}\mathbf{h}_{k,i} + \mathbf{b}, \mathbf{y}_k\right) + \mathcal{R}\left(\mathbf{W}, \mathbf{b}, \{\mathbf{h}_{k,i}\}\right). \quad (2)$$

The rationale for considering this model is that modern overparameterizd deep networks can adapt their feature mapping to almost any training data. Specifically, (Mixon et al.,

2020) considered the unregularized case (no regularization $\mathcal{R}$) where $\mathcal{L}$ is the MSE loss. It is shown there that a simplex EFT is (only) *a* global minimizer. However, without penalizing the optimization variables it is easy to see that there are infinitely many global minimizers of different structures (which are not necessarily collapses). In fact, experiments with unregularized MSE loss and randomly initialized gradient descent typically convergence to non-collapse global minimizers. (See the dependency on the initialization in the experiments in (Mixon et al., 2020)). Other works considered (2) under L_2-norm regularized (or constrained) cross-entropy loss with or without the bias term (Lu & Steinerberger, 2022; Fang et al., 2021; Zhu et al., 2021). They showed that, in this case, *any* global minimizer has the simplex EFT structure.

In the following section, we first close the gap for the UFM with regularized MSE loss (this loss has been shown to be powerful for classification tasks (Hui & Belkin, 2020)). We show that in this case the collapsed features can be more structured than in the cross-entropy case. Then, we turn to mitigate a limitation of the plain UFM, namely, its inability to capture any behavior that happens across depth as it considers only one level of features. To tackle this, we extend the UFM by adding another layer of weights as well as nonlinearity to the model and generalize our previous results. We note that there is a concurrent work (Zhou et al., 2022) that also studies the UFM with regularized MSE loss. Yet, they consider only the plain UFM (with no extensions).

3. NC for Unconstrained Features Model with Regularized MSE Loss

In this section, we study the optimization of the UFM with regularized MSE loss. Let $\mathbf{H} = [\mathbf{h}_{1,1}, \ldots, \mathbf{h}_{1,n}, \mathbf{h}_{2,1}, \ldots, \mathbf{h}_{K,n}] \in \mathbb{R}^{d \times Kn}$ be the (organized) unconstrained features matrix, associated with the one-hot vectors matrix $\mathbf{Y} = \mathbf{I}_K \otimes \mathbf{1}_n^\top \in \mathbb{R}^{K \times Kn}$, where $\otimes$ denotes the Kronecker product. We consider the optimization problem

$$\min_{\mathbf{W},\mathbf{H},\mathbf{b}} \frac{1}{2Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \|\mathbf{W}\mathbf{h}_{k,i} + \mathbf{b} - \mathbf{y}_k\|_2^2 \tag{3}$$
$$+ \frac{\lambda_W}{2} \sum_{k=1}^{K} \|\mathbf{w}_k\|_2^2 + \frac{\lambda_H}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} \|\mathbf{h}_{k,i}\|_2^2 + \frac{\lambda_b}{2} \|\mathbf{b}\|_2^2$$
$$= \min_{\mathbf{W},\mathbf{H},\mathbf{b}} \frac{1}{2Kn} \|\mathbf{W}\mathbf{H} + \mathbf{b}\mathbf{1}_N^\top - \mathbf{Y}\|_F^2 \tag{4}$$
$$+ \frac{\lambda_W}{2} \|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2} \|\mathbf{H}\|_F^2 + \frac{\lambda_b}{2} \|\mathbf{b}\|_2^2,$$

where λ_W, λ_H, and λ_b are positive regularization hyperparameters and $\|\cdot\|_F$ denotes the Frobenius norm.

We provide complete characterizations of the minimizers for two settings: (i) the *bias-free case*, where $\mathbf{b} = \mathbf{0}$ is fixed (equivalently, $\lambda_b \to \infty$), and (ii) the *unregularized-bias case*, where $\lambda_b = 0$ and $\mathbf{b}$ can be optimized. From these results, several conclusions are deduced also for the case where $\lambda_b > 0$ and $\mathbf{b}$ is optimizable.

In the following subsections, we show that while in the *unregularized-bias case* the features and weights of any global minimizer are aligned in a simplex ETF structure (similarly to the results obtained for the cross-entropy loss both with and without bias), in the *bias-free case* the features and weights of any global minimizer are aligned in an orthogonal frame (OF) structure. Since any orthogonal frame can trivially be turned into a simplex ETF by reducing its global mean, in a sense, this collapse is more structured than a simplex ETF collapse. Giving a precise characterization for the minimizers of the bias-free model is important, as later, based on these results, we will study an extension of the bias-free UFM, which has another layer of weights and nonlinearity.

Remark on the optimization procedure. Despite the fact that (3) is a non-convex problem (due to the multiplication of $\mathbf{W}$ and $\mathbf{H}$), its global minimizers are easily obtained by simple optimization algorithms, such as plain gradient descent. This phenomenon follows from the fact that the optimization landscape of matrix factorization with two factors includes only global minima (no local minima) and strict saddle points (roughly speaking, such saddle points can be easily escaped from by gradient-based algorithms) (Kawaguchi, 2016; Freeman & Bruna, 2017).

3.1. The Bias-Free Case

We first consider the optimization problem

$$\min_{\mathbf{W},\mathbf{H}} \frac{1}{2Kn} \|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_W}{2} \|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2} \|\mathbf{H}\|_F^2, \tag{5}$$

which is a special case of (3) with a fixed $\mathbf{b} = \mathbf{0}$ (or equivalently, $\lambda_b \to \infty$).

The following theorem characterizes the global solutions of (5), showing that they necessarily have an orthogonal frame (OF) structure.

Theorem 3.1. *Let $d \geq K$ and define $c := K\sqrt{n\lambda_H\lambda_W}$. If $c \leq 1$, then any global minimizer $(\mathbf{W}^*, \mathbf{H}^*)$ of (5) satisfies*

$$\mathbf{h}_{k,1}^* = \ldots = \mathbf{h}_{k,n}^* =: \mathbf{h}_k^*, \quad \forall k \in [K], \tag{6}$$

$$\|\mathbf{h}_1^*\|_2^2 = \ldots = \|\mathbf{h}_K^*\|_2^2 =: \rho = (1-c)\sqrt{\frac{\lambda_W}{n\lambda_H}}, \tag{7}$$

$$[\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*]^\top [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] = \rho \mathbf{I}_K, \tag{8}$$

$$\mathbf{w}_k^* = \sqrt{n\lambda_H/\lambda_W}\,\mathbf{h}_k^*, \quad \forall k \in [K]. \tag{9}$$

If $c > 1$, then (5) is minimized by $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$.

Proof. See Appendix A. The proof is based on lower bounding the objective by a sequence of inequalities that hold with equality if and only if the stated conditions are satisfied. □

Let us dwell on the implication of this theorem. Denote $\overline{\mathbf{H}} := [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d\times K}$. In the theorem, (6) implies that the columns of $\mathbf{H}^*$ collapse to the columns of $\overline{\mathbf{H}}$ and (9) implies that the rows of $\mathbf{W}^*$ are aligned with the columns of $\overline{\mathbf{H}}$. That is,

$$\mathbf{H}^* = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top \tag{10}$$
$$\mathbf{W}^* = \sqrt{n\lambda_H/\lambda_W}\overline{\mathbf{H}}^\top.$$

The consequence of (7), (8) and (9) is that

$$\mathbf{W}^*\mathbf{W}^{*\top} = \frac{n\lambda_H}{\lambda_W}\rho\mathbf{I}_K = (1-c)\sqrt{\frac{n\lambda_H}{\lambda_W}}\mathbf{I}_K, \tag{11}$$
$$\mathbf{W}^*\mathbf{H}^* = \sqrt{\frac{n\lambda_H}{\lambda_W}}\rho\mathbf{I}_K \otimes \mathbf{1}_n^\top = (1-c)\mathbf{I}_K \otimes \mathbf{1}_n^\top. \tag{12}$$

Note that the collapse of $\mathbf{W}^*$ and $\mathbf{H}^*$ here, in the case of bias-free regularized MSE loss, is to an orthogonal frame (as $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} = \rho\mathbf{I}_K$). Yet, by defining the global feature mean $\mathbf{h}_G^* := \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^n \mathbf{h}_{k,i}^* = \frac{1}{K}\sum_{k=1}^K \mathbf{h}_k^*$, trivially, we have that $\overline{\mathbf{H}} - \mathbf{h}_G^*\mathbf{1}_K^\top = [\mathbf{h}_1^* - \mathbf{h}_G^*, \ldots, \mathbf{h}_K^* - \mathbf{h}_G^*]$ is a simplex ETF. This follows from

$$\begin{aligned}
&\left(\overline{\mathbf{H}} - \mathbf{h}_G^*\mathbf{1}_K^\top\right)^\top\left(\overline{\mathbf{H}} - \mathbf{h}_G^*\mathbf{1}_K^\top\right) \qquad (13)\\
&= \left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)^\top \overline{\mathbf{H}}^\top\overline{\mathbf{H}}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\\
&= \rho\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)^\top\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\\
&= \rho\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right),
\end{aligned}$$

where we used $\mathbf{h}_G^* = \frac{1}{K}\overline{\mathbf{H}}\mathbf{1}_K$. In that sense, $\mathbf{H}^*$ here is more structured than in the results reported by previous works that considered the UFM with regularized/constrained cross-entropy loss (Lu & Steinerberger, 2022; Fang et al., 2021; Zhu et al., 2021), where the collapse of $\mathbf{W}^*$ and $\mathbf{H}^*$ is to a simplex ETF.

3.2. The Unregularized-Bias Case

We next turn to consider the optimization problem

$$\min_{\mathbf{W},\mathbf{H},\mathbf{b}} \frac{1}{2Kn}\|\mathbf{W}\mathbf{H} + \mathbf{b}\mathbf{1}_N^\top - \mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2, \tag{14}$$

which is a special case of (3) when $\lambda_b = 0$.

The following theorem characterizes the global solutions of (14), showing that they necessarily have a simplex ETF structure.

Theorem 3.2. *Let $d \geq K$ and define $c := K\sqrt{n\lambda_H\lambda_W}$. If $c \leq 1$, then any global minimizer $(\mathbf{W}^*, \mathbf{H}^*, \mathbf{b}^*)$ of (14) satisfies*

$$\mathbf{b}^* = \frac{1}{K}\mathbf{1}_K, \tag{15}$$
$$\mathbf{h}_{k,1}^* = \ldots = \mathbf{h}_{k,n}^* =: \mathbf{h}_k^*, \quad \forall k \in [K], \tag{16}$$
$$\mathbf{h}_G^* := \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^n \mathbf{h}_{k,i}^* = \frac{1}{K}\sum_{k=1}^K \mathbf{h}_k^* = \mathbf{0}, \tag{17}$$
$$\|\mathbf{h}_1^*\|_2^2 = \ldots = \|\mathbf{h}_K^*\|_2^2 =: \rho = \frac{(1-c)(K-1)}{K}\sqrt{\frac{\lambda_W}{n\lambda_H}}, \tag{18}$$
$$[\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*]^\top[\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] = \rho\frac{K}{K-1}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right), \tag{19}$$
$$\mathbf{w}_k^* = \sqrt{n\lambda_H/\lambda_W}\mathbf{h}_k^*, \quad \forall k \in [K]. \tag{20}$$

If $c > 1$, then (14) is minimized by $(\mathbf{W}^*, \mathbf{H}^*, \mathbf{b}^*) = (\mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K)$.

Proof. See Appendix B. Similarly to the previous theorem, the proof is based on lower bounding the objective by a sequence of inequalities that hold with equality if and only if the stated conditions are satisfied. □

The consequence of (18), (19) and (20) is that

$$\mathbf{W}^*\mathbf{W}^{*\top} = \frac{n\lambda_H}{\lambda_W}\rho\frac{K}{K-1}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right), \tag{21}$$
$$\mathbf{W}^*\mathbf{H}^* = \sqrt{\frac{n\lambda_H}{\lambda_W}}\rho\frac{K}{K-1}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right) \otimes \mathbf{1}_n^\top. \tag{22}$$

Note that the results in Theorem 3.2 (contrary to those in Theorem 3.1) resemble the results that have been obtained for the cross-entropy loss (both with and without bias). However, as far as we know, no such theorem has been reported for the case of MSE loss.

Remark on the regularized-bias case. From Theorems 3.1 and 3.2, we get the following facts about the global minimizers. In the bias-free case ($\lambda_b \to \infty$), $\mathbf{H}^*$ has an OF structure, and trivially, if we subtract from it the global feature mean $\mathbf{h}_G^*$, we get that $\mathbf{H}^* - \mathbf{h}_G^*\mathbf{1}_K$ has a simplex ETF structure. In the unregularized-bias case ($\lambda_b = 0$), $\mathbf{H}^*$ has a simplex ETF structure. Trivially, this is also the structure of $\mathbf{H}^* - \mathbf{h}_G^*\mathbf{1}_K$, as the global feature mean $\mathbf{h}_G^*$ equals zero in this case. In both cases, $\mathbf{W}^{*\top}$ is aligned with $\mathbf{H}^*$, i.e., it is an OF in the bias-free case and a simplex ETF in the

unregularized-bias case. The consequence of these results[1] is that for the fully regularized MSE loss, where $0 < \lambda_b < \infty$, the global minimizers may have $\mathbf{H}^*$ and $\mathbf{W}^{*\top}$ that are neither a simplex EFT nor an OF. Yet, we empirically observed that still $\mathbf{W}^{*\top}$ is aligned with $\mathbf{H}^*$ and that $\mathbf{H}^* - \mathbf{h}_G^* \mathbf{1}_K$ is a simplex ETF (as may be expected, because these two properties hold in both extreme settings of λ_b).

4. Extended Unconstrained Features Model

The UFM, which considers only one level of features, cannot capture any behavior that happens across depth. Therefore, in this section, we extend this model, first with another layer of weights, and then with the nonlinear ReLU activation between the two layers of weights.

4.1. Unconstrained Features Model With an Additional Layer

Consider the following optimization problem that corresponds to an extended UFM with two layers of weights,

$$\min_{\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1} \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2, \tag{23}$$

where λ_{W_2}, λ_{W_1}, and λ_{H_1} are regularization hyperparameters, and $\mathbf{W}_2 \in \mathbb{R}^{K\times d}$, $\mathbf{W}_1 \in \mathbb{R}^{d\times d}$, $\mathbf{H}_1 \in \mathbb{R}^{d\times N}$. Observe the similarity between (23) and (5), where $(\mathbf{W}, \mathbf{H})$ in (5) are replaced by $(\mathbf{W}_2, \mathbf{W}_1\mathbf{H}_1)$ or by $(\mathbf{W}_2\mathbf{W}_1, \mathbf{H}_1)$. Yet, the similarity is only *partial* because, e.g., if we plug $(\mathbf{W}, \mathbf{H}) = (\mathbf{W}_2, \mathbf{W}_1\mathbf{H}_1)$ in (5) we get a regularization term $\|\mathbf{W}_1\mathbf{H}_1\|_F^2$ rather than separated $\|\mathbf{W}_1\|_F^2$ and $\|\mathbf{H}_1\|_F^2$. To the best of our knowledge, characterization of the minimizers of a multilayer extension of the unconstrained features model has not been done so far.

Remark on the optimization procedure. While both (23) and (5) are non-convex problems, obtaining the global minimizers of (23) is more challenging in practice (e.g., requires careful initializations). This follows from the fact that the optimization landscapes of matrix factorization with three of more factors (or equivalently, non-shallow linear neural networks) include also non-strict saddle points, which entangle gradient-based methods (Kawaguchi, 2016).

The following theorem characterizes the global solutions of (23). It shows that the orthogonal frame structure of the solutions is maintained despite the intermediate weight matrix that has been added. Here "$\propto$" denotes proportional, i.e., equal up to a positive scalar factor.

Theorem 4.1. *Let $d > K$ and $(\mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$ be a global minimizer of (23). Then, both $\mathbf{H}_1^*$ and $\mathbf{W}_1^*\mathbf{H}_1^*$ collapse to orthogonal $d \times K$ frames. Also, both $(\mathbf{W}_2^*\mathbf{W}_1^*)^\top$ and $\mathbf{W}_2^{*\top}$ are orthogonal $d \times K$ matrices, where $(\mathbf{W}_2^*\mathbf{W}_1^*)^\top$ is aligned with $\mathbf{H}_1^*$ and $\mathbf{W}_2^{*\top}$ is aligned with $\mathbf{W}_1^*\mathbf{H}_1^*$. Formally, we have that $\mathbf{H}_1^* = \overline{\mathbf{H}}_1 \otimes \mathbf{1}_n^\top$ for some $\overline{\mathbf{H}}_1 \in \mathbb{R}^{d\times K}$, and*

$$(\mathbf{W}_2^*\mathbf{W}_1^*)\overline{\mathbf{H}}_1 \propto \overline{\mathbf{H}}_1^\top\overline{\mathbf{H}}_1 \propto (\mathbf{W}_2^*\mathbf{W}_1^*)(\mathbf{W}_2^*\mathbf{W}_1^*)^\top \propto \mathbf{I}_K.$$

Similarly, we have that $\mathbf{H}_2^* := \mathbf{W}_1^*\mathbf{H}_1^* = \overline{\mathbf{H}}_2 \otimes \mathbf{1}_n^\top$ for some $\overline{\mathbf{H}}_2 \in \mathbb{R}^{d\times K}$, and

$$\mathbf{W}_2^*\overline{\mathbf{H}}_2 \propto \overline{\mathbf{H}}_2^\top\overline{\mathbf{H}}_2 \propto \mathbf{W}_2^*\mathbf{W}_2^{*\top} \propto \mathbf{I}_K.$$

Proof. See Appendix C. The proof is based on connecting the minimization of the three-factors objective with two sub-problems that include two-factors objectives. More specifically, the sum of the Frobenius norm regularization of two matrices is lower bounded (with attainable equality) by a suitably scaled nuclear norm of their multiplication, and the minimizers of the latter formulation, which can be expressed by the minimizers of the original problem, are analyzed. □

Remark on the choice of loss function. The proof of Theorem 4.1 mostly depends on handling the regularization terms when transforming the problem into two sub-problems, and can be potentially modified to the case where the cross-entropy loss is used instead of MSE. Thus, a similar theorem can be stated for cross-entropy loss, for which it is known that the minimizers of the plain UFM collapse as well (Zhu et al., 2021). Naturally, in such a statement the collapse will be to a simplex ETF rather than to an OF. Indeed, we empirically observed that also when using the cross-entropy loss in (23), the global minimizers $\mathbf{W}_1^*\mathbf{H}_1^*$ and $\mathbf{H}_1^*$ collapse to a simplex ETF structure.

Discussion. In practical "well-trained" DNNs (e.g., see Figure 5 in the experiments section): (1) structured collapse appears only in the deepest features; (2) decrease in within-class variability is obtained monotonically along the depth of the network[2]. However, Theorem 4.1 shows the emergence of structured (orthogonal) collapse *simultaneously* at the two levels of unconstrained features of the model in (23) — both at the deeper $\mathbf{H}_2 := \mathbf{W}_1\mathbf{H}_1$ and at the shallower $\mathbf{H}_1$ — which does not fit (1). Moreover, the linear link between $\mathbf{H}_2$ and $\mathbf{H}_1$ implies that they have the same within-class variability measured by the metric NC_1 (defined in (26) below) as long as the columns of $\mathbf{H}_1$ are not in the null space

[1] In the UFM, note that the (within-class) collapse of the global minimizers (i.e., $\mathbf{h}_{i,k}^* = \mathbf{h}_k^*$ for all $i \in [n]$) is a consequence of the symmetry of the loss and the regularization terms w.r.t. the sample index, which, in our proofs, is exploited by attaining Jensen's inequality when averaging over $i \in [n]$. Thus, it does not depend on whether we regularize the bias term.

[2] Similar depthwise progressive variability decrease is empirically observed also in (Papyan, 2020).

of $\mathbf{W}_1$. This hints that $\mathbf{H}_1$ and $\mathbf{H}_2$ may have similar values/slopes for their NC_1 metric after random initialization and along gradient-based optimization (see Appendix D for more details). Yet, this does not fit (2). Therefore, extending the model to two levels of features without the addition of a non-linearity still cannot capture the behavior of practical DNNs across layers. This encourages us to further extend the model by adding a nonexpansive nonlinear activation function (ReLU) between $\mathbf{W}_2$ and $\mathbf{W}_1$, that naturally breaks the similarity between the two levels of features.

4.2. Non-Linear Unconstrained Features Model

In this section, we turn to consider a nonlinear version of the unconstrained features model that has been stated in (23). Specifically, using the same notation as (23), we consider the optimization problem

$$\min_{\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1} \frac{1}{2Kn}\|\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{H}_1)-\mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2+\frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2+\frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2, \tag{24}$$

where $\sigma(\cdot)=\max(0,\cdot)$ is the element-wise ReLU function.

The following theorem characterizes the global solutions of (24) by exploiting the similarity of this problem to the one in (23). It shows that the orthogonal frame structure created by the optimal solution $(\mathbf{W}^*,\mathbf{H}^*)=(\mathbf{W}_2^*,\sigma(\mathbf{W}_1^*\mathbf{H}_1^*))$ is maintained despite the nonlinearity that has been added.

Theorem 4.2. *Let $d>K$ and $(\mathbf{W}_2^*,\mathbf{W}_1^*,\mathbf{H}_1^*)$ be a global minimizer of (24). Then, $\sigma(\mathbf{W}_1^*\mathbf{H}_1^*)$ collapses to an orthogonal $d\times K$ frame and $\mathbf{W}_2^{*\top}$ is an orthogonal $d\times K$ matrix that is aligned with $\sigma(\mathbf{W}_1^*\mathbf{H}_1^*)$, i.e., $\mathbf{H}_2^*:=\sigma(\mathbf{W}_1^*\mathbf{H}_1^*)=\overline{\mathbf{H}}_2\otimes\mathbf{1}_n^\top$ for some non-negative $\overline{\mathbf{H}}_2\in\mathbb{R}^{d\times K}$, and*

$$\mathbf{W}_2^*\overline{\mathbf{H}}_2\propto\overline{\mathbf{H}}_2^\top\overline{\mathbf{H}}_2\propto\mathbf{W}_2^*\mathbf{W}_2^{*\top}\propto\mathbf{I}_K.$$

Proof. See Appendix E. The proof utilizes the one of Theorem 4.1. It is based on showing equivalence between two related sub-problems w/ and w/o the ReLU operation. □

Note that the structure of $(\mathbf{W}^*,\mathbf{H}^*)=(\mathbf{W}_2^*,\sigma(\mathbf{W}_1^*\mathbf{H}_1^*))$ is the same as for the model in (23), where the non-linearity is absent (yet, here $\mathbf{H}^*$ is obviously also non-negative). This analysis benefits from the fact that the features are unconstrained, and is in contrast with the usual case, where the results obtained for linear models do not carry "as is" to their non-linear counterparts. In Section 6 we show that the nonlinearity is necessary for capturing the different behavior of features in different depths during the collapse of practical networks.

5. Toward Generalizing the UFMs Results to Other Models

Similar to the existing theoretical works that demonstrate the emergence of collapsed minimizers, in this paper we considered models where the features matrix $\mathbf{H}$ (or $\mathbf{H}_1$) is a free optimization variable. It is of high interest to make a step forward and instead of freely optimize the features connect them to some data distribution.

While we defer a comprehensive study that links the models to data for future research, in this short section we demonstrate the feasibility of this goal, even for the plain UFM, through the following theorem.

Theorem 5.1. *Consider (5) with $\lambda_H=\frac{\tilde{\lambda}_H}{n}$. Denote by $(\mathbf{W}^*,\mathbf{H}^*)$ a global minimizer of (5) for some n. Following Theorem 3.1, observe that $\mathbf{H}^*=\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top$ for some $\overline{\mathbf{H}}\in\mathbb{R}^{d\times K}$. Let $\tilde{\mathbf{H}}_n:=\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top+\mathbf{E}_n$ where $\mathbf{E}_n\in\mathbb{R}^{d\times Kn}$ whose entries are i.i.d. random variables with zero mean, variance σ_e^2, and finite fourth moment. Let*

$$\hat{\mathbf{W}}_n=\underset{\mathbf{W}}{argmin}\ \frac{1}{2Kn}\|\mathbf{W}\tilde{\mathbf{H}}_n-\mathbf{Y}\|_F^2+\frac{\lambda_W}{2}\|\mathbf{W}\|_F^2. \tag{25}$$

We have that $\hat{\mathbf{W}}_n\xrightarrow[n\to\infty]{a.s.}\frac{1}{1+\sigma_e^2K\sqrt{\tilde{\lambda}_H/\lambda_W}}\mathbf{W}^*$.

Proof. See Appendix F. The proof exploits the fact that $\hat{\mathbf{W}}_n$ has a closed-form expression (a function of the features matrix) that allows linking it to $\mathbf{W}^*$. □

Theorem 5.1 shows that as the number of samples tend to infinity we have that properties of the optimal weights such as the orthogonal structure and the alignment with $\overline{\mathbf{H}}^\top$ (stated for $\mathbf{W}^*$ in Theorem 3.1) are restored even with a fixed non-collapsed features matrix.

As discussed in Appendix F.1, the intuition that the asymptotic consequence of the deviation from "perfectly" collapsed features will only be some attenuation of $\mathbf{W}^*$ can also be seen from expending the quadratic term in (25) and eliminating the terms that are linear in the zero-mean $\mathbf{E}_n$. This intuition applies also for the extended UFMs with fixed features (where no closed-form minimizers exist).

6. Numerical Results

In this section, we corroborate our theoretical results with experiments. For each setting that is considered in the theorems of Sections 3 and 4 we tune a gradient descent scheme to reach a global minimizer. We plot the optimization's objective value curve at different iterations, as well as several metrics that measure the properties of the NC, which are computed every 5e3 iterations. The theorems are verified by demonstrating the convergence of the NC metrics to zero. We use the following metrics for measuring NC, which are

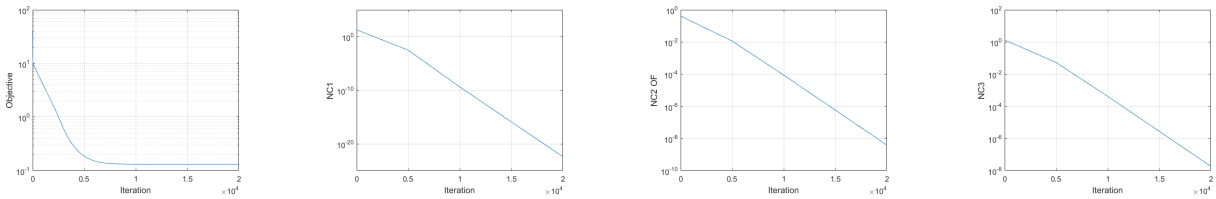

Figure 1. Verification of Theorem 3.1 (MSE loss with no bias). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to OF), and NC3 (alignment between the weights and the features).

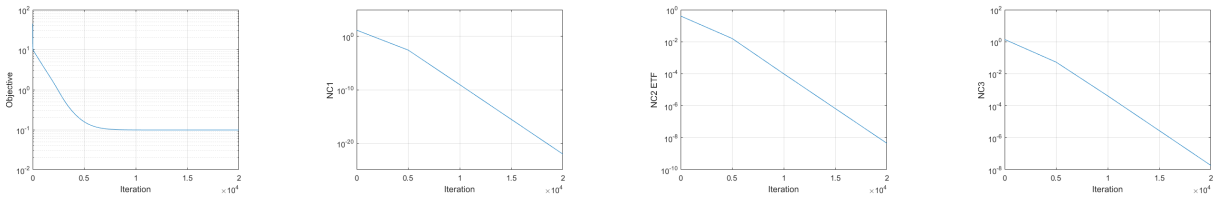

Figure 2. Verification of Theorem 3.2 (MSE loss with unregularized bias). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to simplex ETF), and NC3 (alignment between the weights and the features).

similar to those in (Papyan et al., 2020; Zhu et al., 2021) but include also a metric for collapse to orthogonal frames.

First, for a given set of n features for each of K classes, $\{\mathbf{h}_{k,i}\}$, we define the per-class and global means as $\overline{\mathbf{h}}_k := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}$ and $\overline{\mathbf{h}}_G := \frac{1}{Kn}\sum_{k=1}^{k}\sum_{i=1}^{n}\mathbf{h}_{k,i}$, respectively, as well as the mean features matrix $\overline{\mathbf{H}} := \left[\overline{\mathbf{h}}_1, \ldots, \overline{\mathbf{h}}_K\right]$. Next, we define the within-class and between-class $d \times d$ covariance matrices

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

$$\mathbf{\Sigma}_B := \frac{1}{K}\sum_{k=1}^{K}(\overline{\mathbf{h}}_k - \overline{\mathbf{h}}_G)(\overline{\mathbf{h}}_k - \overline{\mathbf{h}}_G)^\top.$$

Now, we turn to define three metrics of NC.

NC_1 for measuring within-class variability:

$$NC_1 := \frac{1}{K}\mathrm{Tr}\left(\mathbf{\Sigma}_W\mathbf{\Sigma}_B^\dagger\right), \tag{26}$$

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

NC_2 for measuring the similarity of the mean features to the structured frames:

$$NC_2^{ETF} := \left\| \frac{\overline{\mathbf{H}}^\top\overline{\mathbf{H}}}{\|\overline{\mathbf{H}}^\top\overline{\mathbf{H}}\|_F} - \frac{1}{\sqrt{K-1}}(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)\right\|_F$$

$$NC_2^{OF} := \left\| \frac{\overline{\mathbf{H}}^\top\overline{\mathbf{H}}}{\|\overline{\mathbf{H}}^\top\overline{\mathbf{H}}\|_F} - \frac{1}{\sqrt{K}}\mathbf{I}_K\right\|_F \tag{27}$$

where the simplex ETFs and the OFs are normalized to unit Frobenius norm.

NC_3 for measuring the alignment of the last weights and the mean features:

$$NC_3 := \left\|\mathbf{W}/\|\mathbf{W}\|_F - \overline{\mathbf{H}}^\top/\|\overline{\mathbf{H}}\|_F\right\|_F. \tag{28}$$

Figure 1 corroborates Theorem 3.1 for $K = 4, d = 20, n = 50$ and $\lambda_W = \lambda_H = 0.005$ (no bias is used, equivalently $\lambda_b \to \infty$). Both $\mathbf{W}$ and $\mathbf{H}$ are initialized with standard normal distribution and are optimized with plain gradient descent with step-size 0.1.

Figure 2 corroborates Theorem 3.2 for $K = 4, d = 20, n = 50$, $\lambda_W = \lambda_H = 0.005$ and $\lambda_b = 0$. All $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{b}$ are initialized with standard normal distribution and are optimized with plain gradient descent with step-size 0.1.

Figure 3 corroborates Theorem 4.1 for $K = 4, d = 20, n = 50$ and $\lambda_{W_2} = \lambda_{W_1} = \lambda_{H_1} = 0.005$ (no bias is used). All $\mathbf{W}_2$, $\mathbf{W}_1$ and $\mathbf{H}_1$ are initialized with standard normal distribution scaled by 0.1 and are optimized with plain gradient descent with step-size 0.1. The metrics are computed for $\mathbf{W} = \mathbf{W}_2$ and $\mathbf{H} = \mathbf{W}_1\mathbf{H}_1$. We also compute NC_1 and NC_2^{OF} for the first layer's features $\mathbf{H} = \mathbf{H}_1$. The collapse of both $\mathbf{W}_1\mathbf{H}_1$ and $\mathbf{H}_1$ to OF (demonstrated by NC1 and NC2 converging to zero) is in agreement with Theorem 4.1.

Figure 4 corroborates Theorem 4.2 that considers the non-linear model in (24). We use $K = 4, d = 20, n = 50$ and $\lambda_{W_2} = \lambda_{W_1} = \lambda_{H_1} = 0.005$ (no bias is used). All $\mathbf{W}_2$, $\mathbf{W}_1$ and $\mathbf{H}_1$ are initialized with standard normal distribution scaled by 0.1 and are optimized with plain gradient descent with step-size 0.1. The metrics are computed for $\mathbf{W} = \mathbf{W}_2$ and $\mathbf{H} = \sigma(\mathbf{W}_1\mathbf{H}_1)$. We also compute NC_1 and NC_2^{OF} for the first layer's features $\mathbf{H} = \mathbf{H}_1$ (as well as for the pre-ReLU features $\mathbf{H} = \mathbf{W}_1\mathbf{H}_1$).

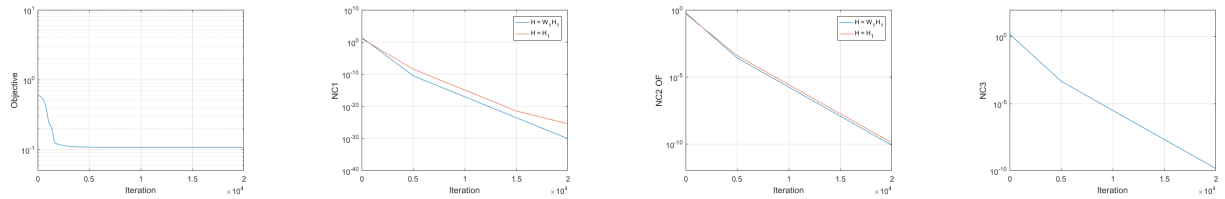

Figure 3. Verification of Theorem 4.1 (two levels of features). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to OF), and NC3 (alignment between the weights and the features).

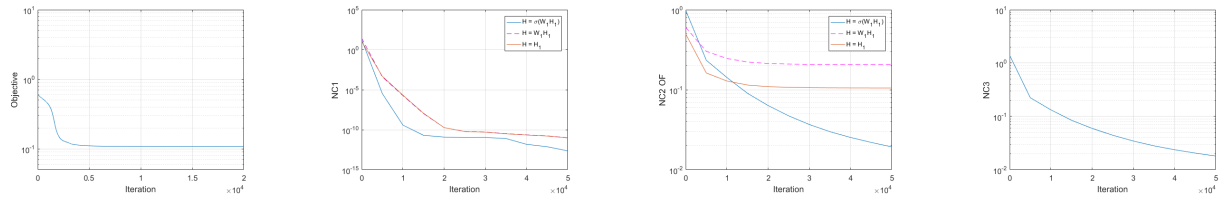

Figure 4. Verification of Theorem 4.2 (two levels of features with ReLU activation). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to OF), and NC3 (alignment between the weights and the features).

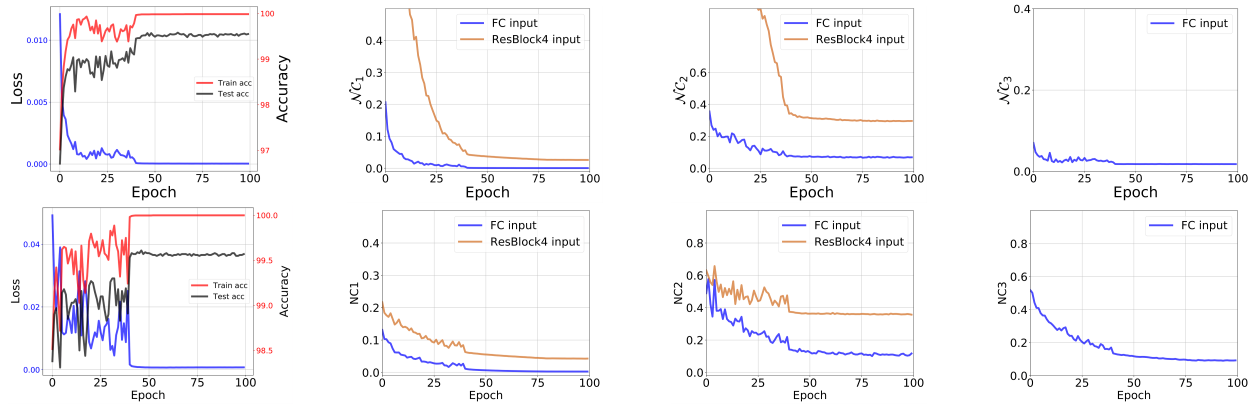

Figure 5. NC metrics for ResNet18 trained on MNIST. Top: MSE loss, weight decay, and no bias; Bottom: Cross-entropy loss and weight decay. From left to right: training's objective value and accuracy, NC1 (within-class variability), NC2 (similarity of the centered features to simplex ETF), and NC3 (alignment between the weights and the features).

Comparing Figures 3 and 4 (experiments with different hyper-parameter setting yield similar results, as shown in Appendix G), we observe that adding the ReLU nonlinearity to the model better distinguishes between the behavior of the features in the two levels, both in the rate of the collapse and in its structure.

Finally, we show the similarity of the NC metrics that are obtained for the *nonlinear* extended UFM in Figure 4 (rather than those in Figure 3) and metrics obtained by a practical well-trained DNN, namely ResNet18 (He et al., 2016) (composed of 4 ResBlocks), trained on MNIST with SGD with learning rate 0.05 (divided by 10 every 40 epochs) and weight decay (L_2 regularization) of 5e-4. Figure 5 shows the results for two cases: 1) MSE loss without bias in the FC layer; and 2) the widely-used setting, with cross-entropy loss and bias. (Additional experiments with CIFAR10 dataset appear in Appendix G). The behaviors of the metrics in both cases correlate with the one of the extended UFM in Figure 4.

7. Conclusion

In this work, we first characterized the (global) minimizers of the unconstrained features model (UFM) for regularized MSE loss, showing some distinctions from the neural collapse (NC) results that have been obtained for the UFM with cross-entropy loss in recent works (such as the effect of the bias term). Then, we mitigated the inability of the plain UFM to capture any NC behavior that happens across depth by adding another layer of weights as well as ReLU nonlinearity to the model and generalized our previous results.

Finally, we empirically verified the theorems and demonstrated the usefulness of our nonlinear extended UFM in modeling the NC phenomenon that occurs in the training of practical networks.

The aforementioned experiments further demonstrated the necessity of the nonlinearity in the model. We note, however, that adding a ReLU nonlinearity in the plain UFM, after the single level of features (with no additional layer of weights), is problematic. Optimizing such a model with simple gradient-based method after random initialization (which is the common way to train practical DNNs), is doomed to fail because the negative entries in the first layer cannot be modified. The extended model that is considered in this paper does not have this limitation.

As directions for future research, we believe that analyzing the gradient descent dynamics of the proposed extended UFM may lead to insights on gradient-based training of practical networks that cannot be obtained from the dynamics of the plain UFM. Generalizing the results that are obtained for the plain and extended UFMs to models where the features cannot be freely optimized, but are rather linked to some data distribution is also of high interest. In this front, the result in Theorem 5.1 is encouraging, though, it is only asymptotic. When the training data is limited and the question of generalization arises (as in real-world settings), it may not be possible to show positive effects of NC on the generalization without departing from the plain UFM, which has limited expressiveness when the features are fixed. On the other hand, the proposed nonlinear extended UFM seems to be more suitable for such analysis, as, in fact, it has a shallow MLP on top of the first level of features.

Acknowledgements

This work has been partially supported by the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF-1845360, and NSF CCF-1814524.

References

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619. PMLR, 2019.

Ergen, T. and Pilanci, M. Revealing the structure of deep neural networks via convex duality. In *International Conference on Machine Learning*, pp. 3004–3014. PMLR, 2021.

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), 2021.

Freeman, C. D. and Bruna, J. Topology and geometry of half-rectified network optimization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

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

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

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

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *arXiv preprint arXiv:1705.08741*, 2017.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Hui, L. and Belkin, M. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

Kawaguchi, K. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29: 586–594, 2016.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Lu, J. and Steinerberger, S. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 2022.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Mixon, D. G., Parshall, H., and Pi, J. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

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

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

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Srebro, N. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

Vershynin, R. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Watson, G. A. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170(0):33–45, 1992.

Wojtowytsch, S. et al. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. *Proceedings of Machine Learning Research*, 145:1–21, 2021.

Zarka, J., Guth, F., and Mallat, S. Separation and concentration in deep networks. In *ICLR 2021-9th International Conference on Learning Representations*, 2021.

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

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. *arXiv preprint arXiv:2203.01238*, 2022.

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

A. Proof of Theorem 3.1

Proof. The proof is based on lower bounding $f(\mathbf{W},\mathbf{H}) := \frac{1}{2N}\|\mathbf{W}\mathbf{H}-\mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2$ by a sequence of inequalities that hold with equality if and only if (6)-(9) are satisfied. First, observe that

$$\begin{aligned}
&\frac{1}{2N}\|\mathbf{W}\mathbf{H}-\mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2 \qquad (29)\\
&= \frac{1}{2Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{W}\mathbf{h}_{k,i}-\mathbf{y}_k\|_2^2 + \frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}\|_2^2\\
&\overset{(a)}{\geq} \frac{1}{2Kn}\sum_{k=1}^{K} n\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}_k^\top\mathbf{h}_{k,i}-1)^2 + \frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K} n\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}\|_2^2\\
&\overset{(b)}{\geq} \frac{1}{2Kn}\sum_{k=1}^{K} n\left(\mathbf{w}_k^\top\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}-1\right)^2 + \frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K} n\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}\right\|_2^2
\end{aligned}$$

The inequality (a) follows from ignoring all the entries except k in the $K\times 1$ vector $\mathbf{W}\mathbf{h}_{k,i}-\mathbf{y}_k$, and holds with equality iff $\mathbf{w}_{k'}^\top\mathbf{h}_{k,i}=0$ for all $k'\neq k$ and $i\in[n]$. In (b) we used Jensen's inequality, which (due to the strict convexity of $(\cdot-1)^2$ and $\|\cdot\|^2$) holds with equality iff $\mathbf{h}_{k,1}=\ldots=\mathbf{h}_{k,n}$ for all $k\in[K]$. Indeed, note that the equality condition for (b) is satisfied by (6), and the equality condition for (a) is a consequence of (6), (8) and (9) (which yield (12)).

Next, to simplify the notation, let us denote $\mathbf{h}_k := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}$. Thus, continuing from the last RHS in (29), we have

$$\begin{aligned}
&\frac{1}{2K}\sum_{k=1}^{K}\left(\mathbf{w}_k^\top\mathbf{h}_k-1\right)^2 + \frac{\lambda_W}{2}K\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2 + \frac{n\lambda_H}{2}K\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2^2 \qquad (30)\\
&\overset{(c)}{\geq} \frac{1}{2}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k^\top\mathbf{h}_k-1\right)^2 + \frac{\lambda_W}{2}K\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2\right)^2 + \frac{n\lambda_H}{2}K\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\right)^2\\
&\overset{(d)}{\geq} \frac{1}{2}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k^\top\mathbf{h}_k-1\right)^2 + K\sqrt{n\lambda_H\lambda_W}\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2\right)\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\right)
\end{aligned}$$

In (c) we used Jensen's inequality, which holds with equality iff

$$\begin{aligned}
\mathbf{w}_1^\top\mathbf{h}_1 &= \ldots = \mathbf{w}_K^\top\mathbf{h}_K,\\
\|\mathbf{w}_1\|_2 &= \ldots = \|\mathbf{w}_K\|_2,\\
\|\mathbf{h}_1\|_2 &= \ldots = \|\mathbf{h}_K\|_2,
\end{aligned}$$

which are satisfied when conditions (7) and (9) are satisfied. In (d) we used the AM–GM inequality, i.e., $\frac{a}{2}+\frac{b}{2}\geq\sqrt{ab}$, with $a=\lambda_W\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2\right)^2$ and $b=n\lambda_H\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\right)^2$. It holds with equality iff $a=b$, which is satisfied by (9) that implies $\lambda_W\|\mathbf{w}_k\|_2^2 = n\lambda_H\|\mathbf{h}_k\|_2^2$.

Note that so far all the iff conditions are satisfied by both $(\mathbf{W}^*,\mathbf{H}^*)$ that satisfy (6)-(9) and the trivial $(\mathbf{W}^*,\mathbf{H}^*)=(\mathbf{0},\mathbf{0})$. Now, it is left to show that if $K\sqrt{n\lambda_H\lambda_W}\leq 1$ then $\mathbf{w}_k$ and $\mathbf{h}_k$ must have the same direction, as implied by (9), which will also yield the orthogonality of $\{\mathbf{h}_k^*\}$ and $\{\mathbf{w}_k^*\}$. While for $K\sqrt{n\lambda_H\lambda_W}>1$, we get the zero minimizer.

As all the inequalities (a)-(d) are attainable with iff conditions, we can consider now $(\mathbf{W},\mathbf{H})$ that satisfy these conditions to further lower the bound. Specifically, using the symmetry w.r.t. k, the last RHS in (30) turns into the expression

$$\begin{aligned}
&\frac{1}{2}\left(\mathbf{w}_k^\top\mathbf{h}_k-1\right)^2 + K\sqrt{n\lambda_H\lambda_W}\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2 \qquad (31)\\
&= \frac{1}{2}\left(\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2\cos\alpha-1\right)^2 + K\sqrt{n\lambda_H\lambda_W}\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2,
\end{aligned}$$

where α is the angle between $\mathbf{w}_k$ and $\mathbf{h}_k$. Invoking Lemma A.1 with $\beta = \|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2$ and $c = K\sqrt{n\lambda_H\lambda_W}$, we get that if $K\sqrt{n\lambda_H\lambda_W} > 1$ then the minimizer is $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$ (since $\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2 = 0$), and otherwise, the minimizer must obey $\alpha = 0$. Therefore, we get the desired result that $\mathbf{w}_k^*$ and $\mathbf{h}_k^*$ must have the same direction. Together with $\lambda_W\|\mathbf{w}_k\|_2^2 = n\lambda_H\|\mathbf{h}_k\|_2^2$ (which is required to attain equality for AM-GM), we get the necessity of $\mathbf{w}_k^* = \sqrt{n\lambda_H/\lambda_W}\mathbf{h}_k^*$ in (9). Finally, the orthogonality of $\{\mathbf{h}_k^*\}$ (and similarly of $\mathbf{w}_k^*$) follows from

$$\mathbf{h}_{k'}^{*\top}\mathbf{h}_k^* = \frac{1}{\sqrt{n\lambda_H/\lambda_W}}\mathbf{w}_{k'}^{*\top}\mathbf{h}_k^* = 0 \quad \forall k' \neq k$$

where we used the previous condition $\mathbf{w}_{k'}^{*\top}\mathbf{h}_k^* = 0$ for all $k' \neq k$, which is necessary to attain equality in (29).

□

Lemma A.1. *Let*

$$\tilde{f}(\alpha, \beta) = \frac{1}{2}\left(\beta\cos\alpha - 1\right)^2 + c\beta, \tag{32}$$

where $\beta \geq 0$ and $c > 0$. Then, (i) if $c > 1$ then $\tilde{f}$ is minimized by $\beta^* = 0$ and the minimal value is $\frac{1}{2}$; (ii) if $c \leq 1$ then $\tilde{f}$ is minimized by $(\alpha^*, \beta^*) = (0, 1-c)$ and the minimal value is $c - \frac{1}{2}c^2$.

Proof. The proof is based on separately analyzing the cases $\beta = 0$, $0 < \beta < 1$ and $\beta \geq 1$.

For $\beta = 0$, we get objective value of $\frac{1}{2}$ for any α. Assuming that $0 < \beta < 1$, clearly, the minimizer of (32) w.r.t. α is only $\alpha^* = 0$ (or other integer multiplications of 2π). Thus, we have

$$\tilde{f}(0, \beta) = \frac{1}{2}\left(\beta - 1\right)^2 + c\beta = \frac{1}{2}\beta^2 - (1-c)\beta + \frac{1}{2},$$

which is a "smiling" parabola in β, with feasible minimum at $\beta^* = \max\{1-c, 0\}$. This means that if $c > 1$ we get the (feasible) minimum at $(\alpha^*, \beta^*) = (0, 0)$, for which $\tilde{f}(\alpha^*, \beta^*) = \frac{1}{2}$. If $c \leq 1$, we get minimum at $(\alpha^*, \beta^*) = (0, 1-c)$ with objective value of $\tilde{f}(\alpha^*, \beta^*) = \frac{1}{2}c^2 + c(1-c) = c - \frac{1}{2}c^2$.

Assuming that $\beta \geq 1$, the first term in (32) is minimized (eliminated) by $\alpha^* = \arccos(1/\beta)$. Thus, we get $\tilde{f}(\alpha^*, \beta) = c\beta$, which is minimized by $\beta^* = 1$, and the objective value is $\tilde{f}(\alpha^*, \beta^*) = c$. Since $c > 0$, note that this value is always larger than the minimal value obtained for $\beta < 1$.

To summarize, (i) if $c > 1$ we get the minimizers $\tilde{f}(\alpha^*, \beta^* = 0) = \frac{1}{2}$; (ii) If $c \leq 1$ we get the minimizer $\tilde{f}(\alpha^* = 0, \beta^* = 1-c) = c - \frac{1}{2}c^2$.

□

A.1. Alternative proof for Theorem 3.1

We present here an alternative proof for Theorem 3.1. While this proof requires tools that are less elementary than the preceding proof, in some sense its strategy is more similar to the one we take to handle the three layer case in Appendix C.

We start by computing the gradients of the objective $f(\mathbf{W},\mathbf{H}) := \frac{1}{2N}\|\mathbf{W}\mathbf{H}-\mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2$

$$\frac{\partial f}{\partial \mathbf{H}} = \mathbf{W}^\top \frac{1}{N}(\mathbf{W}\mathbf{H}-\mathbf{Y}) + \lambda_H \mathbf{H}, \tag{33}$$

$$\frac{\partial f}{\partial \mathbf{W}} = \frac{1}{N}(\mathbf{W}\mathbf{H}-\mathbf{Y})\mathbf{H}^\top + \lambda_W \mathbf{W}. \tag{34}$$

From these expressions we have that any stationary point $(\mathbf{W},\mathbf{H})$ of f (i.e., any point for which all the gradients equal zero) obeys

$$\lambda_W \mathbf{W}^\top \mathbf{W} = \lambda_H \mathbf{H}\mathbf{H}^\top \tag{35}$$

which follows from $\mathbf{W}^\top \frac{\partial f}{\partial \mathbf{W}} - \frac{\partial f}{\partial \mathbf{H}}\mathbf{H}^\top = \mathbf{0}$. Thus, while $\mathbf{W} \in \mathbb{R}^{K\times d}$ is trivially of rank at most K, we also have that any stationary $\mathbf{H}$ is of rank at most K (since $\lambda_W \mathbf{W}^\top\mathbf{W} = \lambda_H \mathbf{H}\mathbf{H}^\top$ is of rank at most K).

Denote the following compact SVDs: $\mathbf{W} = \mathbf{U}_W \mathbf{\Sigma}_W \mathbf{V}_W^\top$ and $\mathbf{H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^\top$ (note that $\mathbf{\Sigma}_W, \mathbf{\Sigma}_H \in \mathbb{R}^{K\times K}$ since the SVDs are compact).

At this point we can define the compact SVD of $\mathbf{Y} \in \mathbb{R}^{K\times N}$ as $\mathbf{Y} = \mathbf{U}_Y \mathbf{\Sigma}_Y \mathbf{V}_Y^\top$, and express the objective function for stationary points as

$$\begin{aligned} f(\mathbf{W},\mathbf{H}) &= \frac{1}{2N}\|\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top - \mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top\|_F^2 \\ &= \frac{1}{2N}\|\mathbf{U}_Y^\top\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top\mathbf{V}_Y - \mathbf{\Sigma}_Y\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{\Sigma}_W\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{\Sigma}_H\|_F^2 \end{aligned} \tag{36}$$

where we used the fact that unitary operators do not change the Frobenius norm.

As $\mathbf{\Sigma}_Y \in \mathbb{R}^{K\times K}$ is a diagonal matrix, clearly a global minimizer obeys that $\mathbf{U}_Y^\top\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top\mathbf{V}_Y$ is a diagonal matrix as well.

Now, we first use the specific structure of $\mathbf{Y}$ in our problem. Namely, $\mathbf{Y} = \mathbf{I}_K \otimes \mathbf{1}_n^\top$, and therefore $\mathbf{U}_Y = \mathbf{I}_K$, $\mathbf{\Sigma}_Y = \sqrt{n}\mathbf{I}_K$ and $\mathbf{V}_Y = \frac{1}{\sqrt{n}}\mathbf{I}_K \otimes \mathbf{1}_n$. This implies that $\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n$ is $K\times K$ diagonal. So, necessarily $\mathbf{V}_H = \overline{\mathbf{V}}_H \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n$ for some $K\times K$ orthonormal matrix $\overline{\mathbf{V}}_H$.

The fact that a global minimizer $\mathbf{H}$ can be decomposed to $\mathbf{H} = \mathbf{U}_H\mathbf{\Sigma}_H\overline{\mathbf{V}}_H^\top \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n^\top$ implies its collapse — $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ for some $\overline{\mathbf{H}} \in \mathbb{R}^{d\times K}$. In other words, we proved that (6) is indeed a necessary property of global minimizers. Denoting the compact SVD of $\overline{\mathbf{H}}$ by $\mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\mathbf{V}_{\overline{H}}^\top$, observe that $\mathbf{\Sigma}_H = \sqrt{n}\mathbf{\Sigma}_{\overline{H}}$ (also, $\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top = \mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\mathbf{V}_{\overline{H}}^\top \otimes \mathbf{1}_n^\top$).

By now we have that the objective function (with a slight abuse of notation in the arguments) of the (collapsed) global minimizers is given by (recall $N = Kn$)

$$f(\mathbf{W},\overline{\mathbf{H}}) = \frac{1}{2K}\|\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\overline{\mathbf{V}}_{\overline{H}}^\top - \mathbf{I}_K\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{\Sigma}_W\|_F^2 + \frac{n\lambda_H}{2}\|\mathbf{\Sigma}_{\overline{H}}\|_F^2. \tag{37}$$

It follows that the global minimizers are necessarily aligned, i.e., $\mathbf{W}\overline{\mathbf{H}} = \beta\mathbf{I}_K$ with some constant β, where we used the spectral symmetry of the regularizations and $\mathbf{I}_K$ that needs to be fitted in the first term. Hence

$$\mathbf{W}\overline{\mathbf{H}} = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\overline{\mathbf{V}}_{\overline{H}}^\top = \beta\mathbf{I}_K,$$

which implies that

$$\mathbf{W} = \mathbf{\Sigma}_W\mathbf{R}^\top \in \mathbb{R}^{K\times d} \tag{38}$$

$$\overline{\mathbf{H}} = \mathbf{R}\mathbf{\Sigma}_{\overline{H}} \in \mathbb{R}^{d\times K} \tag{39}$$

for any orthonormal matrix $\mathbf{R} \in \mathbb{R}^{d\times K}$ ($\mathbf{R}^\top\mathbf{R} = \mathbf{I}_K$). Therefore, we have

$$f(\mathbf{W}, \overline{\mathbf{H}}) = \frac{1}{2K}\|\mathbf{\Sigma}_W\mathbf{\Sigma}_{\overline{H}} - \mathbf{I}_K\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{\Sigma}_W\|_F^2 + \frac{n\lambda_H}{2}\|\mathbf{\Sigma}_{\overline{H}}\|_F^2. \tag{40}$$

The symmetry and separability of (40) with respect to the spectral values implies that $\mathbf{\Sigma}_W = \sigma_W\mathbf{I}_k$ and $\mathbf{\Sigma}_{\overline{H}} = \sigma_{\overline{H}}\mathbf{I}_k$. The values of σ_W and $\sigma_{\overline{H}}$ are determined by minimizing the simplified objective (again with a slight abuse of notation)

$$f(\mathbf{W}, \overline{\mathbf{H}}) = \frac{1}{2}(\sigma_W\sigma_{\overline{H}} - 1)^2 + K\frac{\lambda_W}{2}\sigma_W^2 + K\frac{n\lambda_H}{2}\sigma_{\overline{H}}^2. \tag{41}$$

The derivatives are given by

$$\frac{\partial}{\partial\sigma_W}f = \sigma_{\overline{H}}(\sigma_W\sigma_{\overline{H}} - 1) + K\lambda_W\sigma_W = 0, \tag{42}$$

$$\frac{\partial}{\partial\sigma_{\overline{H}}}f = \sigma_W(\sigma_W\sigma_{\overline{H}} - 1) + Kn\lambda_H\sigma_{\overline{H}} = 0, \tag{43}$$

implying that $\lambda_W\sigma_W^2 = n\lambda_H\sigma_{\overline{H}}^2$, which can also be obtained by attaining the AM-GM inequality

$$K\frac{\lambda_W}{2}\sigma_W^2 + K\frac{n\lambda_H}{2}\sigma_{\overline{H}}^2 \geq K\sqrt{n\lambda_H\lambda_W}\sigma_W\sigma_{\overline{H}}.$$

Therefore, setting $\beta = \sigma_W\sigma_{\overline{H}}$, to find the eigenvalues of the minimizers we just need to find $\beta \geq 0$ that minimizes

$$\tilde{f}(\beta) = \frac{1}{2}(\beta - 1)^2 + c\beta, \tag{44}$$

for $c = K\sqrt{n\lambda_H\lambda_W} > 0$. It can be shown that: (i) if $c > 1$ then $\tilde{f}$ is minimized by $\beta^* = 0$ and the minimal value is $\frac{1}{2}$; (ii) if $c \leq 1$ then $\tilde{f}$ is minimized by $\beta^* = 1 - c$ and the minimal value is $c - \frac{1}{2}c^2$.

Summarizing our finding, we have that if $c = K\sqrt{n\lambda_H\lambda_W} > 1$ then the minimizer is $(\mathbf{W}, \mathbf{H}) = (\mathbf{0}, \mathbf{0})$ (because the singular values of the matrices are zero). On the other hand, if $c = K\sqrt{n\lambda_H\lambda_W} \leq 1$ then the minimizers obey $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$, $\mathbf{W} = \sqrt{\frac{n\lambda_H}{\lambda_W}}\overline{\mathbf{H}}^\top$, and

$$\mathbf{W}\overline{\mathbf{H}} = \sigma_W\sigma_{\overline{H}}\mathbf{I}_K = (1-c)\mathbf{I}_K$$

$$\overline{\mathbf{H}}^\top\overline{\mathbf{H}} = \sigma_{\overline{H}}^2\mathbf{I}_K = (1-c)\sqrt{\frac{\lambda_W}{n\lambda_H}}\mathbf{I}_K$$

$$\mathbf{W}\mathbf{W}^\top = \sigma_W^2\mathbf{I}_K = (1-c)\sqrt{\frac{n\lambda_H}{\lambda_W}}\mathbf{I}_K$$

as stated in the theorem.

B. Proof of Theorem 3.2

Proof. First, note that the objective

$$f(\mathbf{W},\mathbf{H},\mathbf{b}) := \frac{1}{2N}\|\mathbf{W}\mathbf{H}+\mathbf{b}\mathbf{1}_N^\top-\mathbf{Y}\|_F^2+\frac{\lambda_W}{2}\|\mathbf{W}\|_F^2+\frac{\lambda_H}{2}\|\mathbf{H}\|_F^2 \tag{45}$$

is convex w.r.t. $\mathbf{b}$, for which there is the following closed-form minimizer (which depends on $\mathbf{W}\mathbf{H}$)

$$\mathbf{b}^* = \frac{1}{N}\left(\mathbf{Y}-\mathbf{W}\mathbf{H}\right)\mathbf{1}_N = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}(\mathbf{y}_k-\mathbf{W}\mathbf{h}_{k,i}). \tag{46}$$

Since $\{\mathbf{y}_k\}$ are one-hot vectors, note that for $k'\in[K]$

$$b^*_{k'} = \frac{n}{N}-\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{w}_{k'}^\top\mathbf{h}_{k,i} = \frac{1}{K}-\mathbf{w}_{k'}^\top\mathbf{h}_G, \tag{47}$$

where $\mathbf{h}_G := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{h}_{k,i}$ is the global feature mean.

The proof is based on lower bounding $f(\mathbf{W},\mathbf{H},\mathbf{b}^*)$ by a sequence of inequalities that hold with equality if and only if (15)-(20) are satisfied. Observe that

$$\begin{aligned}
&\frac{1}{2Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{W}\mathbf{h}_{k,i}+\mathbf{b}^*-\mathbf{y}_k\|_2^2+\frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2+\frac{\lambda_H}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}\|_2^2 \\
&=\frac{1}{2K}\sum_{k'=1}^{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}_{k'}^\top(\mathbf{h}_{k,i}-\mathbf{h}_G)+\frac{1}{K}-1_{k'=k})^2+\frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2+\frac{\lambda_H}{2}\sum_{k=1}^{K}n\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}\|_2^2 \\
&\overset{(b)}{\geq}\frac{1}{2K}\sum_{k'=1}^{K}\sum_{k=1}^{K}\left(\mathbf{w}_{k'}^\top(\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}-\mathbf{h}_G)+\frac{1}{K}-1_{k'=k}\right)^2+\frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2+\frac{\lambda_H}{2}\sum_{k=1}^{K}n\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}\right\|_2^2
\end{aligned} \tag{48}$$

In (b) we used Jensen's inequality, which (due to the strict convexity of $\|\cdot\|^2$) holds with equality iff $\mathbf{h}_{k,1}=\ldots=\mathbf{h}_{k,n}$ for all $k\in[K]$. Indeed, note that the equality condition for (b) is satisfied by (16).

Next, to simplify the notation, let us denote $\mathbf{h}_k := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}$ (note that $\mathbf{h}_G=\frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_k$). Thus, continuing from the last RHS in (48), we have

$$\begin{aligned}
&\frac{1}{2K}\sum_{k'=1}^{K}\sum_{k=1}^{K}\left(\mathbf{w}_{k'}^\top(\mathbf{h}_k-\mathbf{h}_G)+\frac{1}{K}-1_{k'=k}\right)^2+\frac{\lambda_W}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2+\frac{n\lambda_H}{2}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2^2 \\
&=\frac{1}{2K}\sum_{k=1}^{K}\left(\mathbf{w}_k^\top(\mathbf{h}_k-\mathbf{h}_G)-\frac{K-1}{K}\right)^2+\frac{K-1}{2K}\sum_{k'=1}^{K}\frac{1}{K-1}\sum_{k=1,k\neq k'}^{K}\left(\mathbf{w}_{k'}^\top(\mathbf{h}_k-\mathbf{h}_G)+\frac{1}{K}\right)^2 \\
&\quad+\frac{\lambda_W}{2}K\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2^2+\frac{n\lambda_H}{2}K\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2^2
\end{aligned}$$

$$\overset{(c)}{\geq} \frac{1}{2}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k^\top(\mathbf{h}_k-\mathbf{h}_G)-\frac{K-1}{K}\right)^2+\frac{K-1}{2K}\sum_{k'=1}^{K}\left(\frac{1}{K-1}\sum_{k=1,k\neq k'}^{K}\mathbf{w}_{k'}^\top(\mathbf{h}_k-\mathbf{h}_G)+\frac{1}{K}\right)^2 \tag{49}$$
$$+\frac{\lambda_W}{2}K\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2\right)^2+\frac{n\lambda_H}{2}K\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\right)^2$$
$$\overset{(d)}{\geq} \frac{1}{2}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k^\top(\mathbf{h}_k-\mathbf{h}_G)-\frac{K-1}{K}\right)^2+\frac{K-1}{2K}\sum_{k'=1}^{K}\left(\frac{1}{K-1}\sum_{k=1,k\neq k'}^{K}\mathbf{w}_{k'}^\top(\mathbf{h}_k-\mathbf{h}_G)+\frac{1}{K}\right)^2$$
$$+K\sqrt{n\lambda_H\lambda_W}\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2\right)\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\right)$$

In (c) we used Jensen's inequality, which holds with equality iff

$$\mathbf{w}_1^\top(\mathbf{h}_1-\mathbf{h}_G)=\ldots=\mathbf{w}_K^\top(\mathbf{h}_K-\mathbf{h}_G), \tag{50}$$
$$\mathbf{w}_{k'}^\top(\mathbf{h}_{k_1}-\mathbf{h}_G)=\mathbf{w}_{k'}^\top(\mathbf{h}_{k_2}-\mathbf{h}_G),\ \ \forall k_1,k_2\in[K]\setminus k', \tag{51}$$
$$\|\mathbf{w}_1\|_2=\ldots=\|\mathbf{w}_K\|_2, \tag{52}$$
$$\|\mathbf{h}_1\|_2=\ldots=\|\mathbf{h}_K\|_2, \tag{53}$$

which are satisfied when conditions (18) and (20) are satisfied. In (d) we used the AM–GM inequality, i.e., $\frac{a}{2}+\frac{b}{2}\geq\sqrt{ab}$, with $a=\lambda_W\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{w}_k\|_2\right)^2$ and $b=n\lambda_H\left(\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\right)^2$. It holds with equality iff $a=b$, which is satisfied by (20) that implies $\lambda_W\|\mathbf{w}_k\|_2^2=n\lambda_H\|\mathbf{h}_k\|_2^2$.

Now, observe that the first two terms in the last RHS of (49) are invariant to the global mean of $\mathbf{H}$ (since it is subtracted there from $\{\mathbf{h}_k\}$). Therefore, the expression can be further reduced by requiring that $\mathbf{h}_G$ minimizes the term $\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2$. To this end, using the triangle inequality $\|\mathbf{h}_k\|_2\geq\|\mathbf{h}_k-\mathbf{h}_G\|_2-\|\mathbf{h}_G\|_2$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k\|_2\geq\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{h}_k-\mathbf{h}_G\|_2-\|\mathbf{h}_G\|_2,$$

which becomes equality when $\mathbf{h}_G=\mathbf{0}$, as required by condition (17). From (47), this also implies that $\mathbf{b}^*=\frac{1}{K}\mathbf{1}_K$, as required by condition (15).

Next, consider $\mathbf{w}_{k'}^\top\mathbf{h}_k=\|\mathbf{w}_{k'}\|_2\|\mathbf{h}_k\|_2\cos\tilde{\alpha}_{k',k}$, where $\tilde{\alpha}_{k,k'}$ denotes the angle between $\mathbf{w}_{k'}$ and $\mathbf{h}_k$. From (51)-(53) it follows that $\tilde{\alpha}_{k',k}$ is exactly the same for any chosen $k'\in[K]$ and $k\in[K]\setminus k'$. This equiangular property implies that the minimal (most negative) possible value of $\cos\tilde{\alpha}_{k',k}$ is given by $\cos\tilde{\alpha}_{k',k}=-\frac{1}{K-1}$, as we have in the standard simplex ETF (Definition 2.2).

Note that so far all the iff conditions are satisfied by both $(\mathbf{W}^*,\mathbf{H}^*,\mathbf{b}^*=\frac{1}{K}\mathbf{1}_K)$ that satisfy (16)-(20) and the naive $(\mathbf{W}^*,\mathbf{H}^*,\mathbf{b}^*)=(\mathbf{0},\mathbf{0},\frac{1}{K}\mathbf{1}_K)$. Now, it is left to show that if $K\sqrt{n\lambda_H\lambda_W}\leq 1$ then $\mathbf{w}_k$ and $\mathbf{h}_k$ must have the same direction, as implied by (20), and the simplex equiangular property of $\{\mathbf{h}_k^*\}$ and $\{\mathbf{w}_k^*\}$. While for $K\sqrt{n\lambda_H\lambda_W}>1$, we get the naive minimizer.

As all the inequalities used so far are attainable with iff conditions, we can consider now $(\mathbf{W},\mathbf{H})$ that satisfy these conditions to further lower the bound. Specifically, using the symmetry w.r.t. k, and choosing any $k'\neq k$, the last RHS in (49) (with the required $\mathbf{h}_G=\mathbf{0}$) turns into the expression

$$\frac{1}{2}\left(\mathbf{w}_k^\top\mathbf{h}_k-\frac{K-1}{K}\right)^2+\frac{(K-1)}{2}\left(\mathbf{w}_{k'}^\top\mathbf{h}_k+\frac{1}{K}\right)^2+K\sqrt{n\lambda_H\lambda_W}\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2 \tag{54}$$
$$=\frac{1}{2}\left(\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2\cos\alpha-\frac{K-1}{K}\right)^2+\frac{(K-1)}{2}\left(\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2\cos\tilde{\alpha}+\frac{1}{K}\right)^2+K\sqrt{n\lambda_H\lambda_W}\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2,$$

where we used α (resp. $\tilde{\alpha}$) to denote the angle between $\mathbf{w}_k$ and $\mathbf{h}_k$ (resp. $\mathbf{w}'_k$ and $\mathbf{h}_k$), and the necessary condition that $\|\mathbf{w}_k\|_2 = \|\mathbf{w}'_k\|_2$.

Invoking Lemma B.1 with $\beta = \|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2$ and $c = K\sqrt{n\lambda_H\lambda_W}$, we get that if $K\sqrt{n\lambda_H\lambda_W} > 1$ then the minimizer is $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$ (since $\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2 = 0$), and otherwise, the minimizer must obey $\alpha = 0$ and $\tilde{\alpha} = \arccos(-\frac{1}{K-1})$. Therefore, we get the desired results that $\mathbf{w}_k^*$ and $\mathbf{h}_k^*$ must have the same direction and $\mathbf{w}_{k'}^\top \mathbf{h}_k = -\|\mathbf{w}_k\|_2\|\mathbf{h}_k\|_2 \frac{1}{K-1}$ for any $k' \in [K]$ and $k \in [K] \setminus k'$. Together with $\lambda_W\|\mathbf{w}_k\|_2^2 = n\lambda_H\|\mathbf{h}_k\|_2^2$ (which is required to attain equality for AM-GM), we get the necessity of $\mathbf{w}_k^* = \sqrt{n\lambda_H/\lambda_W}\mathbf{h}_k^*$ in (20). Finally, the simplex equiangular property of $\{\mathbf{h}_k^*\}$ (and similarly of $\mathbf{w}_k^*$) follows from

$$\mathbf{h}_{k'}^{*\top}\mathbf{h}_k^* = \sqrt{\frac{\lambda_W}{n\lambda_H}}\mathbf{w}_{k'}^{*\top}\mathbf{h}_k^* = \sqrt{\frac{\lambda_W}{n\lambda_H}}\|\mathbf{w}_{k'}^*\|_2\|\mathbf{h}_k^*\|_2\cos\tilde{\alpha}_{k',k} = \|\mathbf{h}_k^*\|_2^2\cos\tilde{\alpha}_{k',k} = -\|\mathbf{h}_k^*\|_2^2\frac{1}{K-1} \quad \forall k' \neq k$$

where we used the simplex equiangular condition between $\mathbf{w}_{k'}$ and $\mathbf{h}_k$ ($k' \neq k$).

□

Lemma B.1. *Let*

$$\tilde{f}(\alpha, \tilde{\alpha}, \beta) = \frac{1}{2}\left(\beta\cos\alpha - \frac{K-1}{K}\right)^2 + \frac{(K-1)}{2}\left(\beta\cos\tilde{\alpha} + \frac{1}{K}\right)^2 + c\beta, \tag{55}$$

where $\beta \geq 0$, $-\frac{1}{K-1} \leq \cos\tilde{\alpha} \leq 1$ and $c > 0$. Then, (i) if $c > 1$ then $\tilde{f}$ is minimized by $\beta^* = 0$ and the minimal value is $\frac{K-1}{2K}$; (ii) if $c \leq 1$ then $\tilde{f}$ is minimized by $(\alpha^*, \tilde{\alpha}^*, \beta^*) = (0, \arccos(-\frac{1}{K-1}), \frac{(1-c)(K-1)}{K})$ and the minimal value is $\frac{K-1}{K}\left(c - \frac{1}{2}c^2\right)$.

Proof. The proof is based on separately analyzing the cases $\beta = 0$, $0 < \beta < \frac{K-1}{K}$ and $\beta \geq \frac{K-1}{K}$.

For $\beta = 0$, we get objective value of $\frac{(K-1)^2}{2K^2} + \frac{K-1}{2K^2} = \frac{K-1}{2K}$ for any α and $\tilde{\alpha}$. Assuming that $0 < \beta < \frac{K-1}{K}$, clearly, the minimizer of (55) w.r.t. α is only $\alpha^* = 0$ (or other integer multiplications of 2π), and the minimizer of (55) w.r.t. $\tilde{\alpha}$ is $\tilde{\alpha}^* = \arccos(-\frac{1}{K-1})$ (recall the assumption $-\frac{1}{K-1} \leq \cos\tilde{\alpha} \leq 1$). Thus, we have

$$\begin{aligned}\tilde{f}(0, \arccos(\frac{-1}{K-1}), \beta) &= \frac{1}{2}\left(\beta - \frac{K-1}{K}\right)^2 + \frac{(K-1)}{2}\left(-\frac{\beta}{K-1} + \frac{1}{K}\right)^2 + c\beta, \\ &= \frac{1}{2}\frac{K}{K-1}\beta^2 - (1-c)\beta + \frac{1}{2}\frac{K-1}{K}\end{aligned} \tag{56}$$

which is a "smiling" parabola in β, with feasible minimum at $\beta^* = \max\{\frac{(1-c)(K-1)}{K}, 0\}$. This means that if $c > 1$ we get the (feasible) minimum at $(\alpha^*, \tilde{\alpha}^*, \beta^*) = (0, \arccos(\frac{-1}{K-1}), 0)$, for which $\tilde{f}(\alpha^*, \tilde{\alpha}^*, \beta^*) = \frac{K-1}{2K}$. If $c \leq 1$, we get minimum at $(\alpha^*, \tilde{\alpha}^*, \beta^*) = (0, \arccos(\frac{-1}{K-1}), \frac{(1-c)(K-1)}{K})$ with objective value of $\tilde{f}(\alpha^*, \tilde{\alpha}^*, \beta^*) = \frac{1}{2}\frac{K-1}{K}\left(1 - (1-c)^2\right) = \frac{K-1}{K}\left(c - \frac{1}{2}c^2\right)$.

Assuming that $\beta \geq \frac{K-1}{K}$, the first term in (55) is minimized (eliminated) by $\alpha^* = \arccos(\frac{K-1}{K\beta})$, and the second term in (55) is minimized (eliminated) by $\tilde{\alpha}^* = \arccos(\frac{-1}{K\beta})$. Thus, we get $\tilde{f}(\alpha^*, \tilde{\alpha}^*, \beta) = c\beta$, which is minimized by $\beta^* = \frac{K-1}{K}$, and the objective value is $\tilde{f}(\alpha^*, \tilde{\alpha}^*, \beta^*) = c\frac{K-1}{K}$. Since $c > 0$, note that this value is always larger than the minimal value obtained for $\beta < \frac{K-1}{K}$.

To summarize, (i) if $c > 1$ we get the minimizers $\tilde{f}(\alpha^*, \tilde{\alpha}^*, \beta^* = 0) = \frac{K-1}{2K}$; (ii) If $c \leq 1$ we get the minimizer $\tilde{f}(\alpha^* = 0, \tilde{\alpha}^* = \arccos(-\frac{1}{K-1}), \beta^* = \frac{(1-c)(K-1)}{K}) = \frac{K-1}{K}\left(c - \frac{1}{2}c^2\right)$.

□

C. Proof of Theorem 4.1

We are going to connect the minimization of the three-factors objective of (23)

$$f(\mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1) := \frac{1}{2N}\|\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2$$

with two sub-problems that include two-factors objectives. We will use the following lemma from (Zhu et al., 2021) (which slightly generalizes a result from (Srebro, 2004)). In this lemma, $\|\mathbf{Z}\|_*$ denotes the nuclear norm of the matrix $\mathbf{Z}$, i.e., the sum of its singular values.

Lemma C.1 (Lemma A.3 in (Zhu et al., 2021))**.** *For any fixed $\mathbf{Z} \in \mathbb{R}^{K\times N}$ and $\alpha > 0$, we have*

$$\|\mathbf{Z}\|_* = \min_{\mathbf{W},\mathbf{H}\ s.t.\ \mathbf{W}\mathbf{H}=\mathbf{Z}} \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}}\|\mathbf{W}\|_F^2 + \sqrt{\alpha}\|\mathbf{H}\|_F^2\right). \tag{57}$$

Note that the minimizers $\mathbf{W}, \mathbf{H}$ obey $\mathbf{W} = \alpha^{1/4}\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{R}^\top$ and $\mathbf{H} = \alpha^{-1/4}\mathbf{R}\mathbf{\Sigma}^{1/2}\mathbf{V}^\top$, where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the SVD of $\mathbf{Z}$ and $\mathbf{R}$ is any orthonormal matrix of suitable dimensions.

The first sub-problem is derived as follows:

$$\min_{\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1} \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 \tag{58}$$

$$= \min_{\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1,\mathbf{H}\ s.t.\ \mathbf{H}=\mathbf{W}_1\mathbf{H}_1} \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 \tag{59}$$

$$= \min_{\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1,\mathbf{H}\ s.t.\ \mathbf{H}=\mathbf{W}_1\mathbf{H}_1} \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 \tag{60}$$

$$+ \sqrt{\lambda_{W_1}\lambda_{H_1}}\frac{1}{2}\left(\frac{1}{\sqrt{\lambda_{H_1}/\lambda_{W_1}}}\|\mathbf{W}_1\|_F^2 + \sqrt{\lambda_{H_1}/\lambda_{W_1}}\|\mathbf{H}_1\|_F^2\right)$$

$$\geq \min_{\mathbf{W}_2,\mathbf{H}} \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 \tag{61}$$

$$+ \sqrt{\lambda_{W_1}\lambda_{H_1}} \min_{\mathbf{W}_1,\mathbf{H}_1\ s.t.\ \mathbf{W}_1\mathbf{H}_1=\mathbf{H}} \frac{1}{2}\left(\frac{1}{\sqrt{\lambda_{H_1}/\lambda_{W_1}}}\|\mathbf{W}_1\|_F^2 + \sqrt{\lambda_{H_1}/\lambda_{W_1}}\|\mathbf{H}_1\|_F^2\right)$$

$$= \min_{\mathbf{W}_2,\mathbf{H}} f_1(\mathbf{W}_2, \mathbf{H}) := \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{H}\|_* \tag{62}$$

where the last equality follows from Lemma C.1.

With very similar steps, the second sub-problem is stated as:

$$\min_{\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1} \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 \tag{63}$$

$$\geq \min_{\mathbf{W},\mathbf{H}_1} f_2(\mathbf{W}, \mathbf{H}_1) := \frac{1}{2Kn}\|\mathbf{W}\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 + \sqrt{\lambda_{W_2}\lambda_{W_1}}\|\mathbf{W}\|_* \tag{64}$$

Therefore, we can analyze the minimizers of (62) and (64) and translate the results to the minimizers of (23), using the characteristics of the minimizers in Lemma C.1.

Let us start with (62) and denote $\lambda_H = \sqrt{\lambda_{W_1}\lambda_{H_1}}$, i.e.,

$$f_1(\mathbf{W}_2, \mathbf{H}) := \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \lambda_H\|\mathbf{H}\|_*$$

The subdifferential and gradient are given by

$$\frac{\partial f_1}{\partial \mathbf{H}} = \mathbf{W}_2^\top\frac{1}{N}(\mathbf{W}_2\mathbf{H} - \mathbf{Y}) + \lambda_H\partial\|\mathbf{H}\|_*, \tag{65}$$

$$\frac{\partial f_1}{\partial \mathbf{W}_2} = \frac{1}{N}(\mathbf{W}_2\mathbf{H} - \mathbf{Y})\mathbf{H}^\top + \lambda_{W_2}\mathbf{W}_2, \tag{66}$$

where $\partial\|\mathbf{H}\|_* = \{\mathbf{U}_H\mathbf{V}_H^\top + \mathbf{Z},\ \mathbf{Z} \in \mathbb{R}^{d\times N} \mid \mathbf{U}_H^\top\mathbf{Z} = \mathbf{0}, \mathbf{Z}\mathbf{V}_H = \mathbf{0}, \|\mathbf{Z}\| \leq 1\}$ when $\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top$ is the compact SVD of $\mathbf{H} \in \mathbb{R}^{d\times N}$ (i.e., $\mathbf{\Sigma}_H \in \mathbb{R}^{\mathrm{rank}(\mathbf{H})\times\mathrm{rank}(\mathbf{H})}$). (See, e.g., (Watson, 1992; Recht et al., 2010)). From these expressions we have that any stationary point $(\mathbf{W}_2, \mathbf{H})$ of f_1 (i.e., any point for which all the gradients equal zero) obeys

$$\lambda_{W_2}\mathbf{W}_2^\top\mathbf{W}_2 = \lambda_H\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{U}_H^\top \tag{67}$$

which follows from $\mathbf{W}_2^\top\frac{\partial f_1}{\partial \mathbf{W}_2} - \frac{\partial f_1}{\partial \mathbf{H}}\mathbf{H}^\top = \mathbf{0}$, and $(\mathbf{U}_H\mathbf{V}_H^\top + \mathbf{Z})\mathbf{H}^\top = \mathbf{U}_H\mathbf{\Sigma}_H\mathbf{U}_H^\top$. Thus, while $\mathbf{W}_2 \in \mathbb{R}^{K\times d}$ is trivially of rank at most K, we also have that any stationary $\mathbf{H}$ is of rank at most K (since $\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{U}_H^\top$ is of rank at most K).

Thus, let us consider the *compact* SVDs: $\mathbf{W}_2 = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top$ and $\mathbf{H} = \mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top$ (note that $\mathbf{\Sigma}_W, \mathbf{\Sigma}_H \in \mathbb{R}^{K\times K}$ since the SVDs are compact). Denote also the compact SVD of $\mathbf{Y} \in \mathbb{R}^{K\times N}$ as $\mathbf{Y} = \mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$. The objective function for stationary points can be expressed as

$$\begin{aligned} f_1(\mathbf{W}_2, \mathbf{H}) &= \frac{1}{2N}\|\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top - \mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \lambda_H\|\mathbf{H}\|_* \\ &= \frac{1}{2N}\|\mathbf{U}_Y^\top\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top\mathbf{V}_Y - \mathbf{\Sigma}_Y\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{\Sigma}_W\|_F^2 + \lambda_H\|\mathbf{\Sigma}_H\|_* \end{aligned} \tag{68}$$

where we used the fact that unitary operators do not change the Frobenius norm, as well as the fact that the Frobenius and nuclear norms depend only on the singular values.

As $\mathbf{\Sigma}_Y \in \mathbb{R}^{K\times K}$ is a diagonal matrix, clearly a global minimizer obeys that $\mathbf{U}_Y^\top\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top\mathbf{V}_Y$ is a diagonal matrix as well.

Now, we first use the specific structure of $\mathbf{Y}$ in our problem. Namely, $\mathbf{Y} = \mathbf{I}_K \otimes \mathbf{1}_n^\top$, and therefore $\mathbf{U}_Y = \mathbf{I}_K$, $\mathbf{\Sigma}_Y = \sqrt{n}\mathbf{I}_K$ and $\mathbf{V}_Y = \frac{1}{\sqrt{n}}\mathbf{I}_K \otimes \mathbf{1}_n$. This implies that $\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n$ is $K \times K$ diagonal. So, necessarily $\mathbf{V}_H = \overline{\mathbf{V}}_H \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n$ for some $K \times K$ orthonormal matrix $\overline{\mathbf{V}}_H$.

The fact that a global minimizer $\mathbf{H}$ can be decomposed to $\mathbf{H} = \mathbf{U}_H\mathbf{\Sigma}_H\overline{\mathbf{V}}_H^\top \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n^\top$ implies its collapse — $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ for some $\overline{\mathbf{H}} \in \mathbb{R}^{d\times K}$. Denoting the compact SVD of $\overline{\mathbf{H}}$ by $\mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\mathbf{V}_{\overline{H}}^\top$, observe that $\mathbf{\Sigma}_H = \sqrt{n}\mathbf{\Sigma}_{\overline{H}}$ (also, $\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top = \mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\mathbf{V}_{\overline{H}}^\top \otimes \mathbf{1}_n^\top$).

By now we have that the objective function (with a slight abuse of notation in the arguments) of the (collapsed) global minimizers is given by (recall $N = Kn$)

$$f_1(\mathbf{W}_2, \overline{\mathbf{H}}) = \frac{1}{2K}\|\mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\overline{\mathbf{V}}_{\overline{H}}^\top - \mathbf{I}_K\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{\Sigma}_W\|_F^2 + \sqrt{n}\lambda_H\|\mathbf{\Sigma}_{\overline{H}}\|_*. \tag{69}$$

It follows that the global minimizers are necessarily aligned, i.e., $\mathbf{W}_2\overline{\mathbf{H}} = \beta\mathbf{I}_K$ with some constant β, where we used the spectral symmetry of the regularizations and $\mathbf{I}_K$ that needs to be fitted in the first term. Hence

$$\mathbf{W}_2\overline{\mathbf{H}} = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top\mathbf{U}_{\overline{H}}\mathbf{\Sigma}_{\overline{H}}\overline{\mathbf{V}}_{\overline{H}}^\top = \beta\mathbf{I}_K,$$

which implies that

$$\mathbf{W}_2 = \mathbf{\Sigma}_W\mathbf{R}^\top \in \mathbb{R}^{K\times d} \tag{70}$$

$$\overline{\mathbf{H}} = \mathbf{R}\mathbf{\Sigma}_{\overline{H}} \in \mathbb{R}^{d\times K} \tag{71}$$

for any orthonormal matrix $\mathbf{R} \in \mathbb{R}^{d\times K}$ ($\mathbf{R}^\top\mathbf{R} = \mathbf{I}_K$). Therefore, we have

$$f_1(\mathbf{W}_2, \overline{\mathbf{H}}) = \frac{1}{2K}\|\mathbf{\Sigma}_W\mathbf{\Sigma}_{\overline{H}} - \mathbf{I}_K\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{\Sigma}_W\|_F^2 + \sqrt{n}\lambda_H\|\mathbf{\Sigma}_{\overline{H}}\|_*. \tag{72}$$

The symmetry and separability of (72) with respect to the spectral values implies that $\mathbf{\Sigma}_W = \sigma_W\mathbf{I}_k$ and $\mathbf{\Sigma}_{\overline{H}} = \sigma_{\overline{H}}\mathbf{I}_k$. The values of σ_W and $\sigma_{\overline{H}}$ are determined by minimizing the simplified objective (again with a slight abuse of notation and recalling $\lambda_H = \sqrt{\lambda_{W_1}\lambda_{H_1}}$)

$$f_1(\mathbf{W}_2, \overline{\mathbf{H}}) = \frac{1}{2}(\sigma_W\sigma_{\overline{H}} - 1)^2 + K\frac{\lambda_{W_2}}{2}\sigma_W^2 + K\sqrt{n\lambda_{W_1}\lambda_{H_1}}\sigma_{\overline{H}}. \tag{73}$$

The derivatives are given by

$$\frac{\partial}{\partial \sigma_W} f_1 = \sigma_{\overline{H}}(\sigma_W \sigma_{\overline{H}} - 1) + K\lambda_{W_2}\sigma_W = 0, \tag{74}$$

$$\frac{\partial}{\partial \sigma_{\overline{H}}} f_1 = \sigma_W(\sigma_W \sigma_{\overline{H}} - 1) + K\sqrt{n\lambda_{W_1}\lambda_{H_1}} = 0, \tag{75}$$

implying that $\lambda_{W_2}\sigma_W^2 = \sqrt{n\lambda_{W_1}\lambda_{H_1}}\sigma_{\overline{H}}$. Plugging $\sigma_{\overline{H}} = \frac{\lambda_{W_2}\sigma_W^2}{\sqrt{n\lambda_{W_1}\lambda_{H_1}}}$ in (75) we get

$$\lambda_{W_2}\sigma_W^4 - \sqrt{n\lambda_{W_1}\lambda_{H_1}}\sigma_W + Kn\lambda_{W_1}\lambda_{H_1} = 0$$

The value of σ_W can be computed numerically as the positive root of the above 4th degree polynomial (the analytical result is extremely cumbersome) and the same goes for the value of $\sigma_{\overline{H}}$. Yet, even without stating these exact constants we can summarize our findings for (62) as follows. We have shown that the minimizers obey $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$, where $\overline{\mathbf{H}} = \sigma_{\overline{H}}\mathbf{R}$ and $\mathbf{W}_2 = \sigma_W \mathbf{R}^\top$ for some non-negative constants $\sigma_{\overline{H}}, \sigma_W$ (which depend on $K, n, \lambda_{W_2}, \lambda_{W_1}, \lambda_{H_1}$) and any orthonormal matrix $\mathbf{R} \in \mathbb{R}^{d\times K}$. Therefore, $\mathbf{W}_2 \propto \overline{\mathbf{H}}^\top$, and

$$\mathbf{W}_2\overline{\mathbf{H}} \propto \overline{\mathbf{H}}^\top\overline{\mathbf{H}} \propto \mathbf{W}_2\mathbf{W}_2^\top \propto \mathbf{I}_K.$$

Now, since $\mathbf{H} = \sigma_{\overline{H}}\mathbf{R} \otimes \mathbf{1}_n^\top$, from Lemma C.1 we know that the minimal objective value of (62) is attained by the minimizers $\mathbf{W}_1, \mathbf{H}_1$ of (23) for which we have $\mathbf{W}_1 = \sqrt[4]{\lambda_{H_1}/\lambda_{W_1}}\sqrt{\sigma_H}\mathbf{R}\tilde{\mathbf{R}}^\top$ and $\mathbf{H}_1 = \frac{1}{\sqrt[4]{\lambda_{H_1}/\lambda_{W_1}}}\sqrt{\sigma_H}\tilde{\mathbf{R}} \otimes \mathbf{1}_n^\top$ for any orthonormal matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{d\times K}$. (Note that the last two expressions require the singular value of $\mathbf{H}$, which is $\sigma_H = \sqrt{n}\sigma_{\overline{H}}$).

We conclude that for $d > K$ and $(\mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$ being a (nonzero) global minimizer of (23), we have that $\mathbf{W}_1^*\mathbf{H}_1^*$ collapses to an orthogonal $d \times K$ frame, and $\mathbf{W}_2^{*\top}$ is an orthogonal $d \times K$ matrix that is aligned with $\mathbf{W}_1^*\mathbf{H}_1^*$.

Analyzing the minimizers of (64) by steps which are very similar to those used for (62) yields the following.

The minimizers of (64) obey $\mathbf{H}_1 = \overline{\mathbf{H}}_1 \otimes \mathbf{1}_n^\top$, where $\overline{\mathbf{H}}_1 = \sigma_{\overline{H}_1}\tilde{\mathbf{R}}$ and $\mathbf{W} = \sigma_W\tilde{\mathbf{R}}^\top$ for some non-negative constants $\sigma_{\overline{H}_1}, \sigma_W$ (which depend on $K, n, \lambda_{W_2}, \lambda_{W_1}, \lambda_{H_1}$) and any orthonormal matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{d\times K}$. Therefore, $\mathbf{W} \propto \overline{\mathbf{H}}_1^\top$, and

$$\mathbf{W}\overline{\mathbf{H}}_1 \propto \overline{\mathbf{H}}_1^\top\overline{\mathbf{H}}_1 \propto \mathbf{W}\mathbf{W}^\top \propto \mathbf{I}_K.$$

Now, since $\mathbf{W} = \sigma_W\tilde{\mathbf{R}}^\top$, from Lemma C.1 we know that the minimal objective value of (64) is attained by the minimizers $\mathbf{W}_2, \mathbf{W}_1$ of (23) for which we have $\mathbf{W}_2 = \sqrt[4]{\lambda_{W_1}/\lambda_{W_2}}\sqrt{\sigma_W}\mathbf{R}^\top$ and $\mathbf{W}_1 = \frac{1}{\sqrt[4]{\lambda_{W_1}/\lambda_{W_2}}}\sqrt{\sigma_W}\mathbf{R}\tilde{\mathbf{R}}^\top$ for any orthonormal matrix $\mathbf{R} \in \mathbb{R}^{d\times K}$.

We conclude that for $d > K$ and $(\mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$ being a (nonzero) global minimizer of (23), we have that $\mathbf{H}_1^*$ collapses to an orthogonal $d \times K$ frame, and $(\mathbf{W}_2^*\mathbf{W}_1^*)^\top$ is an orthogonal $d \times K$ matrix that is aligned with $\mathbf{H}_1^*$.

D. On the Within-Class Variability Metric NC1

In this section, we discuss some properties of the within-class variability of the features $\mathbf{H}_1$ and $\mathbf{H}_2 := \mathbf{W}_1\mathbf{H}_1$ for the model in (23). First, let us define the metric NC_1 that is used to measure the within-class variability. Note that this metric is related to the classical Fisher's ratio. For a given (organized) features matrix $\mathbf{H} = [\mathbf{h}_{1,1}, \dots, \mathbf{h}_{1,n}, \mathbf{h}_{2,1}, \dots, \mathbf{h}_{K,n}] \in \mathbb{R}^{d\times Kn}$, denote the per-class and global means as $\overline{\mathbf{h}}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{h}_{k,i}$ and $\overline{\mathbf{h}}_G := \frac{1}{Kn}\sum_{k=1}^k\sum_{i=1}^n \mathbf{h}_{k,i}$, respectively. Define the within-class and between-class $d \times d$ covariance matrices

$$\mathbf{\Sigma}_W(\mathbf{H}) := \frac{1}{Kn}\sum_{k=1}^K\sum_{i=1}^n(\mathbf{h}_{k,i} - \overline{\mathbf{h}}_k)(\mathbf{h}_{k,i} - \overline{\mathbf{h}}_k)^\top,$$

$$\mathbf{\Sigma}_B(\mathbf{H}) := \frac{1}{K}\sum_{k=1}^K(\overline{\mathbf{h}}_k - \overline{\mathbf{h}}_G)(\overline{\mathbf{h}}_k - \overline{\mathbf{h}}_G)^\top.$$

We define the corresponding within-class variability metric as

$$NC_1(\mathbf{H}) := \frac{1}{K}\text{Tr}\left(\mathbf{\Sigma}_W(\mathbf{H})\mathbf{\Sigma}_B^\dagger(\mathbf{H})\right), \tag{76}$$

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

From the definitions above, observe that $\mathbf{\Sigma}_W(\mathbf{H}_2) = \mathbf{W}_1\mathbf{\Sigma}_W(\mathbf{H}_1)\mathbf{W}_1^\top$ and $\mathbf{\Sigma}_B(\mathbf{H}_2) = \mathbf{W}_1\mathbf{\Sigma}_B(\mathbf{H}_1)\mathbf{W}_1^\top$. Therefore,

$$\begin{aligned} NC_1(\mathbf{H}_2) &= \frac{1}{K}\text{Tr}\left(\mathbf{W}_1\mathbf{\Sigma}_W(\mathbf{H}_1)\mathbf{W}_1^\top(\mathbf{W}_1\mathbf{\Sigma}_B(\mathbf{H}_1)\mathbf{W}_1^\top)^\dagger\right) \\ &= \frac{1}{K}\text{Tr}\left(\mathbf{W}_1\mathbf{\Sigma}_W(\mathbf{H}_1)\mathbf{W}_1^\top\mathbf{W}_1^{\top\dagger}\mathbf{\Sigma}_B^\dagger(\mathbf{H}_1)\mathbf{W}_1^\dagger\right) \\ &= \frac{1}{K}\text{Tr}\left(\mathbf{W}_1^\dagger\mathbf{W}_1\mathbf{\Sigma}_W(\mathbf{H}_1)\left(\mathbf{W}_1^\dagger\mathbf{W}_1\right)^\top\mathbf{\Sigma}_B^\dagger(\mathbf{H}_1)\right). \end{aligned} \tag{77}$$

Now, by their definitions, the columns of $\mathbf{\Sigma}_W(\mathbf{H}_1)$ and $\mathbf{\Sigma}_B(\mathbf{H}_1)$ are in the range of $\mathbf{H}_1$. Thus, since $\mathbf{W}_1^\dagger\mathbf{W}_1$ is an orthogonal projection matrix (onto the subspace spanned by the rows of $\mathbf{W}_1$), we have that

$$NC_1(\mathbf{H}_2) = \frac{1}{K}\text{Tr}\left(\mathbf{W}_1^\dagger\mathbf{W}_1\mathbf{\Sigma}_W(\mathbf{H}_1)\left(\mathbf{W}_1^\dagger\mathbf{W}_1\right)^\top\mathbf{\Sigma}_B^\dagger(\mathbf{H}_1)\right) = \frac{1}{K}\text{Tr}\left(\mathbf{\Sigma}_W(\mathbf{H}_1)\mathbf{\Sigma}_B^\dagger(\mathbf{H}_1)\right) = NC_1(\mathbf{H}_1)$$

is guaranteed when there are no columns of $\mathbf{H}_1$ in the null space of $\mathbf{W}_1$. One such case is at initialization, when $\mathbf{W}_1$ is initialized by continuous random distribution and thus its rows span $\mathbb{R}^d$ with probability 1. Moreover, after random initialization, we empirically observed that $\mathbf{H}_1$ and $\mathbf{H}_2$ also have similar NC_1 along gradient-based optimization (see Figure 3), which is due to having similar K dimensional subspaces dominantly spanned by the columns of $\mathbf{H}_1$ and the rows of $\mathbf{W}_1$ (as well as those of $\mathbf{W}_2$). At convergence to the a global minimizer, again it is guaranteed that there are no columns of $\mathbf{H}_1$ in the null space of $\mathbf{W}_1$. Specifically, as demonstrated in the proof of Theorem 4.1, the global minimizers necessarily have that $\mathbf{W}_2^{*\top}$, $\mathbf{W}_1^{*\top}$ and $\mathbf{H}_1^*$ have exactly the same K dimensional range (column space). Briefly, denoting the objective of (23) by f, this follows from $\mathbf{W}_2^*\mathbf{W}_2^{*\top} \propto \mathbf{I}_K$, as well as $\lambda_{W_2}\mathbf{W}_2^{*\top}\mathbf{W}_2^* = \lambda_{W_1}\mathbf{W}_1^*\mathbf{W}_1^{*\top}$ and $\lambda_{W_1}\mathbf{W}_1^{*\top}\mathbf{W}_1^* = \lambda_{H_1}\mathbf{H}_1^*\mathbf{H}_1^{*\top}$, where the last two equalities follow from $\mathbf{W}_1^\top\frac{\partial f}{\partial \mathbf{W}_1} - \frac{\partial f}{\partial \mathbf{H}_1}\mathbf{H}_1^\top = \mathbf{0}$ and $\mathbf{W}_2^\top\frac{\partial f}{\partial \mathbf{W}_2} - \frac{\partial f}{\partial \mathbf{W}_1}\mathbf{W}_1^\top = \mathbf{0}$, respectively.

E. Proof of Theorem 4.2

Let $(\mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$ be a minimizer of (24). Similar to the arguments in the proof of Theorem 4.1, the construction $(\mathbf{W}_2, \mathbf{H}) = (\mathbf{W}_2^*, \mathbf{W}_1^*\mathbf{H}_1^*)$ is also a minimizer of the following sub-problem

$$\tilde{f}_1(\mathbf{W}_2, \mathbf{H}) := \frac{1}{2Kn}\|\mathbf{W}_2\sigma(\mathbf{H}) - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{H}\|_*. \tag{78}$$

We are going to show that the construction $(\mathbf{W}_2, \mathbf{H}) = (\mathbf{W}_2^*, \sigma(\mathbf{W}_1^*\mathbf{H}_1^*))$ is also a minimizer of

$$f_1(\mathbf{W}_2, \mathbf{H}) := \frac{1}{2Kn}\|\mathbf{W}_2\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{H}\|_*. \tag{79}$$

Recall that, following the proof of Theorem 4.1, for a given minimizer of (23), $(\tilde{\mathbf{W}}_2^*, \tilde{\mathbf{W}}_1^*, \tilde{\mathbf{H}}_1^*)$, we have that $(\mathbf{W}_2, \mathbf{H}) = (\tilde{\mathbf{W}}_2^*, \tilde{\mathbf{W}}_1^*\tilde{\mathbf{H}}_1^*)$ minimizes (79). Therefore, this will imply that the orthogonal collapse and alignment properties of $\tilde{\mathbf{W}}_2^*$ and $\tilde{\mathbf{W}}_1^*\tilde{\mathbf{H}}_1^*$, which have been obtained by analyzing $f_1(\mathbf{W}_2, \mathbf{H})$ in Appendix C, carry on to $\mathbf{W}_2^*$ and $\sigma(\mathbf{W}_1^*\mathbf{H}_1^*)$ constructed from global minimizers of (24).

We begin by considering an intermediate objective:

$$\overline{f}_1(\mathbf{W}_2, \mathbf{H}) := \frac{1}{2Kn}\|\mathbf{W}_2\sigma(\mathbf{H}) - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \sqrt{\lambda_{W_1}\lambda_{H_1}}\|\sigma(\mathbf{H})\|_*. \tag{80}$$

Clearly, $\min_{\mathbf{W}_2,\mathbf{H}} f_1(\mathbf{W}_2, \mathbf{H}) \leq \min_{\mathbf{W}_2,\mathbf{H}} \overline{f}_1(\mathbf{W}_2, \mathbf{H})$, since the ReLUs in (80) can be translated to a non-negativity constraint on $\mathbf{H}$ that reduces the feasible set of the minimization problem. Yet, let us show that this inequality is not strict.

Essentially, showing that $\min_{\mathbf{W}_2,\mathbf{H}} f_1(\mathbf{W}_2, \mathbf{H}) = \min_{\mathbf{W}_2,\mathbf{H}} \overline{f}_1(\mathbf{W}_2, \mathbf{H})$ is translated to proving that (79) has a non-negative minimizer. To this end, we use the following properties of the minimizers that has been shown in the proof of Theorem 4.1: $\tilde{\mathbf{H}}^*$ has the structure $\tilde{\mathbf{H}}^* = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ and

$$\tilde{\mathbf{W}}_2^* = \mathbf{\Sigma}_W^*\mathbf{R}^\top \in \mathbb{R}^{K\times d} \tag{81}$$

$$\overline{\mathbf{H}} = \mathbf{R}\mathbf{\Sigma}_{\overline{H}}^* \in \mathbb{R}^{d\times K} \tag{82}$$

where $\mathbf{\Sigma}_W^*, \mathbf{\Sigma}_{\overline{H}}^* \in \mathbb{R}^{K\times K}$ are non-negative diagonal matrices and $\mathbf{R} \in \mathbb{R}^{d\times K}$ can be any orthonormal matrix ($\mathbf{R}^\top\mathbf{R} = \mathbf{I}_K$). (The freedom in $\mathbf{R}$ is due to the fact that the problem can be expressed only in terms of singular values). Now, we can get the existence of the desired non-negative matrices by considering

$$\mathbf{R} = \begin{bmatrix} \mathbf{I}_K \\ \mathbf{0}_{(d-K)\times K} \end{bmatrix},$$

for which

$$\tilde{\mathbf{W}}_2^* = \mathbf{\Sigma}_W^* \begin{bmatrix} \mathbf{I}_K & \mathbf{0}_{K\times(d-K)} \end{bmatrix}$$

$$\tilde{\mathbf{W}}_1^*\tilde{\mathbf{H}}_1^* = \begin{bmatrix} \mathbf{I}_K \\ \mathbf{0}_{(d-K)\times K} \end{bmatrix} \mathbf{\Sigma}_{\overline{H}}^* \otimes \mathbf{1}_n^\top$$

are clearly non-negative.

The above result also implies that the set of minimizers of (80) is a subset of the set of minimizers of (79). Thus, such minimizers carry the property that $\sigma(\mathbf{H}^*)$ is a collapsed orthogonal matrix: $\sigma(\mathbf{H}^*) = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ for some non-negative $\overline{\mathbf{H}} \in \mathbb{R}^{d\times K}$ that obeys $\overline{\mathbf{H}}_2^\top\overline{\mathbf{H}}_2 = \alpha\mathbf{I}_k$ for some positive scalar α.

Now, let us compare $\overline{f}_1(\mathbf{W}_2, \mathbf{H})$ with $\tilde{f}_1(\mathbf{W}_2, \mathbf{H})$. Observe that, intuitively, $\min_{\mathbf{W}_2,\mathbf{H}} \tilde{f}_1(\mathbf{W}_2, \mathbf{H}) \leq \min_{\mathbf{W}_2,\mathbf{H}} \overline{f}_1(\mathbf{W}_2, \mathbf{H})$ because, apparently, negative entries of $\mathbf{H}$ can be used to reduce $\|\mathbf{H}\|_*$. Formally, we have that

$$\begin{aligned}\min_{\mathbf{W}_2,\mathbf{H}} \tilde{f}_1(\mathbf{W}_2, \mathbf{H}) &= \min_{\mathbf{W}_2,\tilde{\mathbf{H}},\mathbf{H}:\sigma(\mathbf{H})=\sigma(\tilde{\mathbf{H}})} \overline{f}_1(\mathbf{W}_2, \tilde{\mathbf{H}}) + \sqrt{\lambda_{W_1}\lambda_{H_1}}(\|\mathbf{H}\|_* - \|\sigma(\tilde{\mathbf{H}})\|_*) \\ &= \min_{\mathbf{W}_2,\tilde{\mathbf{H}}} \overline{f}_1(\mathbf{W}_2, \tilde{\mathbf{H}}) + \min_{\mathbf{H}:\sigma(\mathbf{H})=\sigma(\tilde{\mathbf{H}})} \sqrt{\lambda_{W_1}\lambda_{H_1}}(\|\mathbf{H}\|_* - \|\sigma(\tilde{\mathbf{H}})\|_*)\end{aligned}$$

where the inner minimization is non-positive (observe that it equals zero for the feasible point $\mathbf{H} = \sigma(\tilde{\mathbf{H}})$). Yet, let us show that in fact $\min_{\mathbf{W}_2,\mathbf{H}} \tilde{f}_1(\mathbf{W}_2, \mathbf{H}) = \min_{\mathbf{W}_2,\mathbf{H}} \overline{f}_1(\mathbf{W}_2, \mathbf{H})$.

First, note that (with a slight abuse of notation) both $\overline{f}_1$ and $\tilde{f}_1$ can be expressed as functions of $\mathbf{H}$ alone that differ only in the nuclear norm term. Indeed, both $\overline{f}_1$ and $\tilde{f}_1$ have the same closed-form expression for the minimizer w.r.t. $\mathbf{W}_2$, which is a function of $\sigma(\mathbf{H})$:

$$\mathbf{W}_2(\sigma(\mathbf{H})) = \mathbf{Y}\sigma(\mathbf{H})^\top(\sigma(\mathbf{H})\sigma(\mathbf{H})^\top + N\lambda_{W_2}\mathbf{I}_d)^{-1}.$$

Plugging this expression into $\overline{f}_1$ and $\tilde{f}_1$ gives

$$\overline{f}_1(\mathbf{H}) = \frac{1}{2N}\|\mathbf{W}_2(\sigma(\mathbf{H}))\sigma(\mathbf{H}) - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2(\sigma(\mathbf{H}))\|_F^2 + \sqrt{\lambda_{W_1}\lambda_{H_1}}\|\sigma(\mathbf{H})\|_*, \tag{83}$$

$$\tilde{f}_1(\mathbf{H}) = \frac{1}{2N}\|\mathbf{W}_2(\sigma(\mathbf{H}))\sigma(\mathbf{H}) - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2(\sigma(\mathbf{H}))\|_F^2 + \sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{H}\|_*. \tag{84}$$

Next, recall that the sub-problems $\overline{f}_1(\mathbf{W}_2, \mathbf{H})$ and $\tilde{f}_1(\mathbf{W}_2, \mathbf{H})$ are associated with the 3 factors objectives (23) (as essentially shown above, restricted to a non-negative feasible set) and (24), respectively. Each of these objectives has three squared Frobenius norm regularization terms ($\frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2$) that are equal at the minimizers (the AM-GM inequality holds with equality). Since for the minimizers we have that $\sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{W}_1\mathbf{H}_1\|_* = \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2$ (Lemma C.1 holds with equality), it also holds for them that $\sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{W}_1\mathbf{H}_1\|_* = \lambda_{W_2}\|\mathbf{W}_2\|_F^2$.

Therefore, we have that a minimizer $\mathbf{H}^*$ of (84) obeys $\sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{H}^*\|_* = \lambda_{W_2}\|\mathbf{W}_2(\sigma(\mathbf{H}^*))\|_F^2$, similarly to a minimizer $\tilde{\mathbf{H}}^*$ of (83) that obeys $\sqrt{\lambda_{W_1}\lambda_{H_1}}\|\sigma(\tilde{\mathbf{H}}^*)\|_* = \lambda_{W_2}\|\mathbf{W}_2(\sigma(\tilde{\mathbf{H}}^*))\|_F^2$. Consequently,

$$\min_{\mathbf{H}} \overline{f}_1(\mathbf{H}) = \min_{\mathbf{H}:\sqrt{\lambda_{W_1}\lambda_{H_1}}\|\sigma(\mathbf{H})\|_* = \lambda_{W_2}\|\mathbf{W}_2(\sigma(\mathbf{H}))\|_F^2} \frac{1}{2N}\|\mathbf{W}_2(\sigma(\mathbf{H}))\sigma(\mathbf{H}) - \mathbf{Y}\|_F^2 + \frac{3\lambda_{W_2}}{2}\|\mathbf{W}_2(\sigma(\mathbf{H}))\|_F^2, \tag{85}$$

$$\min_{\mathbf{H}} \tilde{f}_1(\mathbf{H}) = \min_{\mathbf{H}:\sqrt{\lambda_{W_1}\lambda_{H_1}}\|\mathbf{H}\|_* = \lambda_{W_2}\|\mathbf{W}_2(\sigma(\mathbf{H}))\|_F^2} \frac{1}{2N}\|\mathbf{W}_2(\sigma(\mathbf{H}))\sigma(\mathbf{H}) - \mathbf{Y}\|_F^2 + \frac{3\lambda_{W_2}}{2}\|\mathbf{W}_2(\sigma(\mathbf{H}))\|_F^2. \tag{86}$$

The analysis of f_1, which is carried to $\overline{f}_1$, showed that $\overline{f}_1$ is minimized by orthogonally collapsed $\sigma(\mathbf{H})$, i.e., $\sigma(\mathbf{H}) = \sigma(\overline{\mathbf{H}}) \otimes \mathbf{1}_n^\top$ where $\sigma(\overline{\mathbf{H}})^\top\sigma(\overline{\mathbf{H}}) \propto \mathbf{I}_K$. The same symmetry w.r.t. the sample index applies also for $\tilde{f}_1$ and implies collapse $\sigma(\mathbf{H}) = \sigma(\overline{\mathbf{H}}) \otimes \mathbf{1}_n^\top$. Yet, we need to show that even though (86) has a different constraint than (85), flat spectrum of $\sigma(\overline{\mathbf{H}})$ (i.e., orthogonality) is still optimal.

Denote the SVD of $\sigma(\overline{\mathbf{H}}) = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{d\times d}$ and $\mathbf{V} \in \mathbb{R}^{K\times K}$ are orthonormal matrices and $\mathbf{S} \in \mathbb{R}^{d\times K}$ is a rectangular diagonal matrix with $\{s_k\}$ on its diagonal. We have that

$$\mathbf{W}_2 = n\sigma(\overline{\mathbf{H}})^\top(n\sigma(\overline{\mathbf{H}})\sigma(\overline{\mathbf{H}})^\top + N\lambda_{W_2}\mathbf{I}_d)^{-1} = \mathbf{V}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top + K\lambda_{W_2}\mathbf{I}_d)^{-1}\mathbf{U}^\top,$$

$$\mathbf{W}_2\sigma(\mathbf{H}) - \mathbf{Y} = \mathbf{V}\left(\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top + K\lambda_{W_2}\mathbf{I}_d)^{-1}\mathbf{S} - \mathbf{I}_K\right)\mathbf{V}^\top \otimes \mathbf{1}_n^\top = \mathbf{V}\mathrm{diag}\left\{\frac{-K\lambda_{W_2}}{(s_1^2 + K\lambda_{W_2})}, ..., \frac{-K\lambda_{W_2}}{(s_K^2 + K\lambda_{W_2})}\right\}\mathbf{V}^\top \otimes \mathbf{1}_n^\top.$$

Substituting these expressions in the objective of (86) yields

$$\frac{1}{2K}\sum_{k=1}^{K}\frac{(K\lambda_{W_2})^2}{(s_k^2 + K\lambda_{W_2})^2} + \frac{3\lambda_{W_2}}{2}\sum_{k=1}^{K}\frac{s_k^2}{(s_k^2 + K\lambda_{W_2})^2} = \frac{1}{K}\sum_{k=1}^{K} g(s_k^2),$$

where $g(z) := \frac{K\lambda_{W_2}}{2}\left(\frac{K\lambda_{W_2}}{(z+K\lambda_{W_2})^2} + 3\frac{z}{(z+K\lambda_{W_2})^2}\right)$. The second derivative of $g(z)$ is given by $g''(z) = \frac{K\lambda_{W_2}}{2}\left(\frac{6K\lambda_{W_2}}{(z+K\lambda_{W_2})^4} + 3\frac{2z-4K\lambda_{W_2}}{(z+K\lambda_{W_2})^4}\right) = \frac{K\lambda_{W_2}}{2}\frac{6z-6K\lambda_{W_2}}{(z+K\lambda_{W_2})^4}$. Note that $g''(z) > 0$ for $z > K\lambda_{W_2}$. In other words, $g''(s_k^2)$ is locally convex w.r.t. the eigenvalues of $\sigma(\overline{\mathbf{H}})^\top\sigma(\overline{\mathbf{H}})$ in the domain $s_k^2 > K\lambda_{W_2}$. Observe that this is the domain of interest, because substituting $s_k^2 = K\lambda_{W_2}$ in the objective yields a value of $\frac{1}{2}$ (equivalently, $g(K\lambda_{W_2}) = \frac{1}{2}$), which is the value that is obtained by the trivial $\tilde{f}_1(\mathbf{H} = \mathbf{0})$ (equivalently, by plugging zero variables in (24)).

As we showed, in the domain of interest the objective of the constrained formulations is convex w.r.t. eigenvalues of $\sigma(\overline{\mathbf{H}})^\top\sigma(\overline{\mathbf{H}})$. Therefore, by Jensen's inequality, optimal $\mathbf{H}$ of (86) necessarily has the property that it collapses to a matrix $\overline{\mathbf{H}}$ such that $\sigma(\overline{\mathbf{H}})$ has equal singular values, i.e., orthogonality: $\sigma(\overline{\mathbf{H}})^\top\sigma(\overline{\mathbf{H}}) \propto \mathbf{I}_K$.

Finally, Lemma E.1 shows that under this collapse the negative entries of $\mathbf{H}$ in (84) do not contribute to reducing the objective, i.e., $\|\mathbf{H}^*\|_* = \|\sigma(\mathbf{H}^*)\|_*$. Thus, we have equality of the minimal values: $\min_{\mathbf{W}_2,\mathbf{H}} \tilde{f}_1(\mathbf{W}_2,\mathbf{H}) = \min_{\mathbf{W}_2,\mathbf{H}} \overline{f}_1(\mathbf{W}_2,\mathbf{H})$.

From the above equivalences between $\tilde{f}_1$ and f_1 through $\overline{f}_1$, we obtain the desired result that if $(\mathbf{W}_2^*,\mathbf{H}^*)$ is a (global) minimizer of $\tilde{f}_1(\mathbf{W}_2,\mathbf{H})$, then $(\mathbf{W}_2^*,\sigma(\mathbf{H}^*))$ is a (global) minimizer of $f_1(\mathbf{W}_2,\mathbf{H})$.

Lemma E.1. *Let $\mathbf{H} \in \mathbb{R}^{d\times Kn}$ be a non-negative orthogonal collapsed matrix, i.e., $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ for some non-negative $\overline{\mathbf{H}} \in \mathbb{R}^{d\times K}$ that obeys $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} = \alpha\mathbf{I}_k$ for some positive scalar α. Then, $\mathbf{Z}^* = \mathbf{H}$ is a minimizer of*

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \|\mathbf{Z}\|_* \\ \text{s.t.} \quad & \sigma(\mathbf{Z}) = \mathbf{H} \end{aligned} \tag{87}$$

where $\sigma(\cdot)$ is the element-wise ReLU function.

Proof. Let $\mathcal{T}$ be the set of the indices of the strictly possitive entries of $\mathbf{H}$, i.e., $\mathcal{T} := \{(i,j) \,:\, H_{ij} > 0\}$. Let $\mathbf{H}_\mathcal{T}$ denote the set of values of the matrix $\mathbf{H}$ in the indices $\mathcal{T}$, and $\mathcal{T}^c$ be the complementary set of $\mathcal{T}$ (so $\mathbf{H}_{\mathcal{T}^c} = 0$). With this notation, the condition $\sigma(\mathbf{Z}) = \mathbf{H}$ can be expressed as

$$\begin{aligned} &\mathbf{Z}_\mathcal{T} - \mathbf{H}_\mathcal{T} = 0, \\ &\mathbf{Z}_{\mathcal{T}^c} - \mathbf{H}_{\mathcal{T}^c} \leq 0. \end{aligned} \tag{88}$$

That is, (87) is equivalent to the convex program

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \|\mathbf{Z}\|_* \\ \text{s.t.} \quad & \mathbf{Z}_\mathcal{T} - \mathbf{H}_\mathcal{T} = 0 \\ & \mathbf{Z}_{\mathcal{T}^c} - \mathbf{H}_{\mathcal{T}^c} \leq 0. \end{aligned} \tag{89}$$

The Lagrangian of this problem is given by

$$L(\mathbf{Z},\mathbf{B}) = \|\mathbf{Z}\|_* + \mathbf{B}_\mathcal{T}\cdot(\mathbf{Z}_\mathcal{T} - \mathbf{H}_\mathcal{T}) + \mathbf{B}_{\mathcal{T}^c}\cdot(\mathbf{Z}_\mathcal{T} - \mathbf{H}_\mathcal{T}) = \|\mathbf{Z}\|_* + \mathbf{B}\cdot(\mathbf{Z} - \mathbf{H}),$$

where "$\cdot$" denotes the dot product between arguments of the same size and $\mathbf{B}$ is the matrix of dual variables, which obeys the element-wise inequality (dual feasibility constraint)

$$\mathbf{B}_{\mathcal{T}^c} \geq 0.$$

To show that $\mathbf{Z} = \mathbf{H}$ satisfies the KKT conditions we just need to show that

$$\mathbf{0} \in \partial L(\mathbf{H},\mathbf{B}) = \partial\|\mathbf{H}\|_* + \mathbf{B} \tag{90}$$

for $\mathbf{B}$ that obeys $\mathbf{B}_{\mathcal{T}^c} \geq 0$ (the complementary slackness trivially holds since $\mathbf{B}_{\mathcal{T}^c} \odot (\mathbf{H}_{\mathcal{T}^c} - \mathbf{H}_{\mathcal{T}^c}) = \mathbf{0}$). Recall that $\partial\|\mathbf{H}\|_* = \{\mathbf{U}_H\mathbf{V}_H^\top + \mathbf{G},\ \mathbf{G} \in \mathbb{R}^{d\times N} \mid \mathbf{U}_H^\top\mathbf{G} = \mathbf{0}, \mathbf{G}\mathbf{V}_H = \mathbf{0}, \|\tilde{\mathbf{G}}\| \leq 1\}$ when $\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top$ is an SVD of $\mathbf{H} \in \mathbb{R}^{d\times N}$ with $\mathbf{\Sigma}_H \in \mathbb{R}^{\mathrm{rank}(\mathbf{H})\times\mathrm{rank}(\mathbf{H})}$.

In our case, $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ for some non-negative $\overline{\mathbf{H}} \in \mathbb{R}^{d\times K}$ that obeys $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} = \alpha\mathbf{I}_k$ for some positive scalar α. Thus, a compact SVD of $\mathbf{H}$ is given by $\mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^\top = (\frac{1}{\sqrt{\alpha}}\overline{\mathbf{H}})(\sqrt{n\alpha}\mathbf{I}_K)(\mathbf{I}_K \otimes \frac{1}{\sqrt{n}}\mathbf{1}_n)^\top$. Since $\mathbf{U}_H\mathbf{V}_H^\top = \frac{1}{\sqrt{n\alpha}}\overline{\mathbf{H}} \otimes \mathbf{1}_n^\top = \frac{1}{\sqrt{n\alpha}}\mathbf{H}$, we see that a zero subgradient in $\partial L(\mathbf{H},\mathbf{B})$:

$$\mathbf{0} = \mathbf{U}_H\mathbf{V}_H^\top + \mathbf{G} + \mathbf{B} = \frac{1}{\sqrt{n\alpha}}\mathbf{H} + \mathbf{G} + \mathbf{B}$$

is obtained for $\mathbf{G} = \mathbf{0}$ and $\mathbf{B} = -\frac{1}{\sqrt{n\alpha}}\mathbf{H}$. Such $\mathbf{B}$ is also dual feasible, since $\mathbf{B}_{\mathcal{T}^c} = -\frac{1}{\sqrt{n\alpha}}\mathbf{H}_{\mathcal{T}^c} = \mathbf{0}$. This optimality certificate implies that $\mathbf{Z}^* = \mathbf{H}$ is a minimizer of the problem (87). $\square$

F. Proof of Theorem 5.1

As stated in the theorem, we consider (5) with $\lambda_H = \frac{\tilde{\lambda}_H}{n}$:

$$\min_{\mathbf{W},\mathbf{H}} \frac{1}{2Kn}\|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\tilde{\lambda}_H}{2n}\|\mathbf{H}\|_F^2, \tag{91}$$

and denote by $(\mathbf{W}^*, \mathbf{H}^*)$ a global minimizer. From Theorem 3.1 we have that $\mathbf{H}^* = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ and $\mathbf{W}^* = \sqrt{\tilde{\lambda}_H/\lambda_W}\overline{\mathbf{H}}^\top$ for some $\overline{\mathbf{H}} \in \mathbb{R}^{d\times K}$ that obeys $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} = \rho\mathbf{I}_K = (1 - K\sqrt{\tilde{\lambda}_H\lambda_W})\sqrt{\frac{\lambda_W}{\tilde{\lambda}_H}}\mathbf{I}_K = (\sqrt{\frac{\lambda_W}{\tilde{\lambda}_H}} - K\lambda_W)\mathbf{I}_K$.

Note that for any value of n, we have that $(\mathbf{W}, \mathbf{H}) = (\mathbf{W}^*, \mathbf{H}_n^* := \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top)$ is a global minimizer of (91).

We turn to examine (91) for fixed $\mathbf{H}$ and minimization only w.r.t. $\mathbf{W}$. Namely,

$$\hat{\mathbf{W}}_n = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2Kn}\|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2. \tag{92}$$

This strongly convex problem has the following closed-form solution

$$\hat{\mathbf{W}}_n(\mathbf{H}) = \frac{1}{Kn}\mathbf{Y}\mathbf{H}^\top\left(\frac{1}{Kn}\mathbf{H}\mathbf{H}^\top + \lambda_W\mathbf{I}_d\right)^{-1}. \tag{93}$$

Recalling that $\mathbf{Y} = \mathbf{I}_K \otimes \mathbf{1}_n^\top$, for $\mathbf{H} = \mathbf{H}_n^* = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top$ we have that

$$\begin{aligned}\hat{\mathbf{W}}_n(\mathbf{H}_n^*) &= \frac{1}{Kn}(\mathbf{I}_K\overline{\mathbf{H}}^\top \otimes \mathbf{1}_n^\top\mathbf{1}_n)\left(\frac{1}{Kn}(\overline{\mathbf{H}}\overline{\mathbf{H}}^\top \otimes \mathbf{1}_n^\top\mathbf{1}_n) + \lambda_W\mathbf{I}_d\right)^{-1} \\ &= \frac{1}{K}\overline{\mathbf{H}}^\top\left(\frac{1}{K}\overline{\mathbf{H}}\overline{\mathbf{H}}^\top + \lambda_W\mathbf{I}_d\right)^{-1}.\end{aligned} \tag{94}$$

This expression can be simplified as follows

$$\begin{aligned}\hat{\mathbf{W}}_n(\mathbf{H}_n^*) &= \frac{1}{K\lambda_W}\overline{\mathbf{H}}^\top\left(\frac{1}{K\lambda_W}\overline{\mathbf{H}}\overline{\mathbf{H}}^\top + \mathbf{I}_d\right)^{-1} \\ &= \frac{1}{K\lambda_W}\left(\frac{\rho}{K\lambda_W}\mathbf{I}_K + \mathbf{I}_K\right)^{-1}\overline{\mathbf{H}}^\top \\ &= \frac{1}{K\lambda_W + \rho}\overline{\mathbf{H}}^\top,\end{aligned} \tag{95}$$

where the second equality follows from the "push-through identity" and the fact that $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} = \rho\mathbf{I}_K$. Note that, as expected, if we fixed $\mathbf{H}$ to be $\mathbf{H}_n^*$, a global minimizer of the joint optimization w.r.t. $(\mathbf{W}, \mathbf{H})$, then we get $\hat{\mathbf{W}}_n = \mathbf{W}^*$. Indeed, $\hat{\mathbf{W}}_n(\mathbf{H}_n^*) = \frac{1}{K\lambda_W+\rho}\overline{\mathbf{H}}^\top = \frac{1}{\sqrt{\lambda_W/\tilde{\lambda}_H}}\overline{\mathbf{H}}^\top = \mathbf{W}^*$.

Let us turn to examine $\hat{\mathbf{W}}_n$ for $\mathbf{H} = \tilde{\mathbf{H}}_n$ where $\tilde{\mathbf{H}}_n := \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top + \mathbf{E}_n$ with $\mathbf{E}_n \in \mathbb{R}^{d\times Kn}$ whose entries are i.i.d. random variables with zero mean, variance σ_e^2, and finite fourth moment. Hence, $\mathbb{E}\left[\mathbf{E}_n\right] = \mathbf{0}$ and $\mathbb{E}\left[\mathbf{E}_n\mathbf{E}_n^\top\right] = Kn\sigma_e^2\mathbf{I}_d$.

Substituting $\mathbf{H} = \tilde{\mathbf{H}}_n$ in (93), we get

$$\hat{\mathbf{W}}_n(\tilde{\mathbf{H}}_n) = \frac{1}{Kn}\mathbf{Y}\tilde{\mathbf{H}}_n^\top\left(\frac{1}{Kn}\tilde{\mathbf{H}}_n\tilde{\mathbf{H}}_n^\top + \lambda_W\mathbf{I}_d\right)^{-1}. \tag{96}$$

Based on the law of large numbers, as well as the convergence of sample covariance matrices of random variables with finite fourth moment (Vershynin, 2012), we have the following limits

$$\begin{aligned}&\frac{1}{Kn}\mathbf{Y}\tilde{\mathbf{H}}_n^\top = \frac{1}{K}\overline{\mathbf{H}} + \frac{1}{Kn}(\mathbf{I}_K \otimes \mathbf{1}_n^\top)\mathbf{E}_n^\top \xrightarrow[n\to\infty]{a.s.} \frac{1}{K}\overline{\mathbf{H}}, \\ &\frac{1}{Kn}\tilde{\mathbf{H}}_n\tilde{\mathbf{H}}_n^\top = \frac{1}{K}\overline{\mathbf{H}}\overline{\mathbf{H}}^\top + \frac{1}{Kn}(\overline{\mathbf{H}} \otimes \mathbf{1}_n^\top)\mathbf{E}_n^\top + \frac{1}{Kn}\mathbf{E}_n(\overline{\mathbf{H}}^\top \otimes \mathbf{1}_n) + \frac{1}{Kn}\mathbf{E}_n\mathbf{E}_n^\top \xrightarrow[n\to\infty]{a.s.} \frac{1}{K}\overline{\mathbf{H}}\overline{\mathbf{H}}^\top + \sigma_e^2\mathbf{I}_d.\end{aligned} \tag{97}$$

Therefore,

$$\hat{\mathbf{W}}_n(\tilde{\mathbf{H}}_n) \xrightarrow[n\to\infty]{a.s.} \frac{1}{K}\overline{\mathbf{H}}\left(\frac{1}{K}\overline{\mathbf{H}}\overline{\mathbf{H}}^\top + \sigma_e^2\mathbf{I}_d + \lambda_W\mathbf{I}_d\right)^{-1}. \tag{98}$$

Repeating the simplifications of (95) (with $\sigma_e^2 + \lambda_W$ in lieu of λ_W) we get

$$\hat{\mathbf{W}}_n(\tilde{\mathbf{H}}_n) \xrightarrow[n\to\infty]{a.s.} \frac{1}{K(\sigma_e^2+\lambda_W)+\rho}\overline{\mathbf{H}}^\top = \frac{1}{K\sigma_e^2 + \sqrt{\lambda_W/\tilde{\lambda}_H}}\overline{\mathbf{H}}^\top. \tag{99}$$

Comparing (99) with $\mathbf{W}^* = \frac{1}{\sqrt{\lambda_W/\tilde{\lambda}_H}}\overline{\mathbf{H}}^\top$, we get the result that is stated in the theorem:

$$\hat{\mathbf{W}}_n(\tilde{\mathbf{H}}_n) \xrightarrow[n\to\infty]{a.s.} \frac{\sqrt{\lambda_W/\tilde{\lambda}_H}}{K\sigma_e^2 + \sqrt{\lambda_W/\tilde{\lambda}_H}}\mathbf{W}^* = \frac{1}{1+\sigma_e^2 K\sqrt{\tilde{\lambda}_H/\lambda_W}}\mathbf{W}^*.$$

F.1. Intuitive explanation of the result

The intuition that the asymptotic consequence of $\mathbf{E}_n$, i.e., the deviation from "perfectly" collapsed features, will only be some attenuation of $\mathbf{W}^*$ can also be seen from expending the quadratic term in (92) for $\mathbf{H} = \overline{\mathbf{H}} \otimes \mathbf{1}_n^\top + \mathbf{E}_n$ and eliminating the terms that are linear in the zero-mean $\mathbf{E}_n$. Specifically, observe that

$$\begin{aligned}\frac{1}{2Kn}\|\mathbf{W}(\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top + \mathbf{E}_n) - \mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 &= \frac{1}{2Kn}\|(\mathbf{W}\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top - \mathbf{Y}) + \mathbf{W}\mathbf{E}_n\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 \\ = \frac{1}{2Kn}\|\mathbf{W}\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top - \mathbf{Y}\|_F^2 &+ \frac{1}{Kn}\mathrm{Tr}(\mathbf{E}_n^\top\mathbf{W}^\top(\mathbf{W}\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top - \mathbf{Y})) + \frac{1}{2Kn}\|\mathbf{W}\mathbf{E}_n\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2.\end{aligned} \tag{100}$$

Now, suppose we take the limit $n \to \infty$ only in the terms that include $\mathbf{E}_n$, we would get

$$\begin{aligned}&\frac{1}{Kn}\mathrm{Tr}(\mathbf{E}_n^\top\mathbf{W}^\top(\mathbf{W}\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top - \mathbf{Y})) \xrightarrow[n\to\infty]{a.s.} \mathbf{0}, \\ &\frac{1}{2Kn}\|\mathbf{W}\mathbf{E}_n\|_F^2 = \frac{1}{2Kn}\mathrm{Tr}(\mathbf{E}_n\mathbf{E}_n^\top\mathbf{W}^\top\mathbf{W}) \xrightarrow[n\to\infty]{a.s.} \frac{1}{2}\sigma_e^2\mathrm{Tr}(\mathbf{W}^\top\mathbf{W}),\end{aligned} \tag{101}$$

under which (100) can be interpreted as

$$\frac{1}{2Kn}\|\mathbf{W}\overline{\mathbf{H}}\otimes\mathbf{1}_n^\top - \mathbf{Y}\|_F^2 + \frac{\sigma_e^2}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2. \tag{102}$$

This hints that, asymptotically, the minimizer $\hat{\mathbf{W}}$ would be similar to the minimizer that is obtained for the case of $\sigma_e = 0$ (as shown above, this is in fact $\mathbf{W}^*$) up to some scaling.

The above intuition is aligned with the results of Theorem 5.1. Yet, contrary to the proof of the theorem, it does not require having a closed-form expression for the minimizer $\hat{\mathbf{W}}$. Interestingly, this allows us to generalize it to the extended UFMs. Specifically, consider the model in (23) with fixed $\mathbf{H}_1 = \overline{\mathbf{H}}_1 \otimes \mathbf{1}_n^\top + \mathbf{E}_n$, where $(\mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^* = \overline{\mathbf{H}}_1 \otimes \mathbf{1}_n^\top)$ is a global minimizer (as stated in Theorem 4.1). Namely,

$$\frac{1}{2Kn}\|\mathbf{W}_2\mathbf{W}_1(\overline{\mathbf{H}}_1\otimes\mathbf{1}_n^\top + \mathbf{E}_n) - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 \tag{103}$$

Repeating the above heuristic, asymptotically, we may interpret this objective as

$$\frac{1}{2Kn}\|\mathbf{W}_2\mathbf{W}_1\overline{\mathbf{H}}_1\otimes\mathbf{1}_n^\top - \mathbf{Y}\|_F^2 + \frac{\sigma_e^2}{2}\|\mathbf{W}_2\mathbf{W}_1\|_F^2 + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2, \tag{104}$$

which maintains many of the properties of the model analyzed in Theorem 4.1, such as invariance to various orthogonal transformations and the ability to restate the problem as optimization on the singular values of $\mathbf{W}_2, \mathbf{W}_1$ and $\mathbf{H}_1$ (as done in the proof in Appendix C). Again, this hints that, asymptotically, the minimizer $(\hat{\mathbf{W}}_2, \hat{\mathbf{W}}_1)$ would be similar to the minimizer that is obtained for the case without $\mathbf{E}_n$, up to some scaling. While we defer a rigorous study of the effect of fixed features matrix $\mathbf{H}_1$ on the extended UFMs for future research, the discussion here demonstrates the feasibility of this goal.

G. More Numerical Results for the Unconstrained Features Model

In this section, we present more numerical results, for experiments that are similar to those in Section 6 but with different configurations. The definitions of the NC metrics appear in Section 6.

Figure 6 corroborates Theorem 3.1 for $K = 5, d = 20, n = 100$, $\lambda_W = 0.005$ and $\lambda_H = 0.001$ (no bias is used, equivalently $\lambda_b \to \infty$). Both $\mathbf{W}$ and $\mathbf{H}$ are initialized with standard normal distribution and are optimized with plain gradient descent with step-size 0.1.

Figure 7 corroborates Theorem 3.2 for $K = 5, d = 20, n = 100$, $\lambda_W = 0.005$ and $\lambda_H = 0.001$ and $\lambda_b = 0$. All $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{b}$ are initialized with standard normal distribution and are optimized with plain gradient descent with step-size 0.1.

Figure 8 corroborates Theorem 4.1 for $K = 5, d = 20, n = 100$, $\lambda_{W_2} = 0.005$, $\lambda_{W_1} = 0.0025$ and $\lambda_{H_1} = 0.001$ (no bias is used). All $\mathbf{W}_2$, $\mathbf{W}_1$ and $\mathbf{H}_1$ are initialized with standard normal distribution scaled by 0.1 and are optimized with plain gradient descent with step-size 0.1. The metrics are computed for $\mathbf{W} = \mathbf{W}_2$ and $\mathbf{H} = \mathbf{W}_1\mathbf{H}_1$. We also compute NC_1 and NC_2^{OF} for the first layer's features $\mathbf{H} = \mathbf{H}_1$. The collapse of $\mathbf{W}_1\mathbf{H}_1$ and $\mathbf{H}_1$ to OF (demonstrated by NC1 and NC2 converging to zero) is in agreement with Theorems 4.1.

Figure 9 corroborates Theorem 4.2 that considers the nonlinear model in (24). We use $K = 5, d = 20, n = 100$, $\lambda_{W_2} = 0.005$, $\lambda_{W_1} = 0.0025$, and $\lambda_{H_1} = 0.001$ (no bias is used). All $\mathbf{W}_2$, $\mathbf{W}_1$ and $\mathbf{H}_1$ are initialized with standard normal distribution scaled by 0.1, 0.1 and 0.2, respectively, and are optimized with plain gradient descent with step-size 0.1. The metrics are computed for $\mathbf{W} = \mathbf{W}_2$ and $\mathbf{H} = \sigma(\mathbf{W}_1\mathbf{H}_1)$. We also compute NC_1 and NC_2^{OF} for the first layer's features $\mathbf{H} = \mathbf{H}_1$ (as well as for the pre-ReLU $\mathbf{H} = \mathbf{W}_1\mathbf{H}_1$).

Finally, in Figure 10 we show the similarity of the NC metrics that are obtained for the (nonlinear) extended UFM and metrics obtained by a practical well-trained DNN, namely ResNet18 (He et al., 2016) (composed of 4 ResBlocks), trained on CIFAR10 dataset via SGD with learning rate 0.05 (divided by 10 every 40 epochs) and weight decay (L_2 regularization) of 5e-4, MSE loss and no bias in the FC layer.

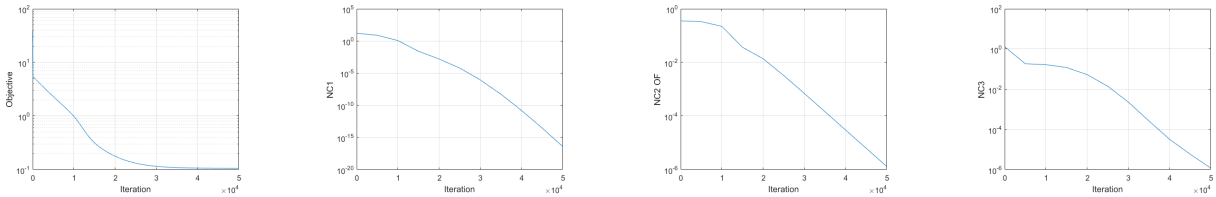

Figure 6. Verification of Theorem 3.1 (MSE loss with no bias). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to OF), and NC3 (alignment between the weights and the features).

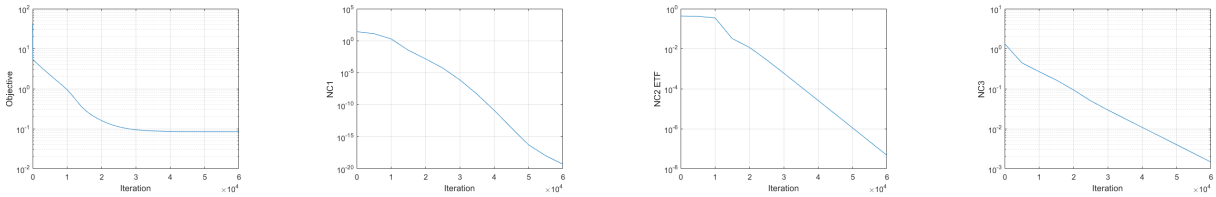

Figure 7. Verification of Theorem 3.2 (MSE loss with unregularized bias). From left to left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to simplex ETF), and NC3 (alignment between the weights and the features).

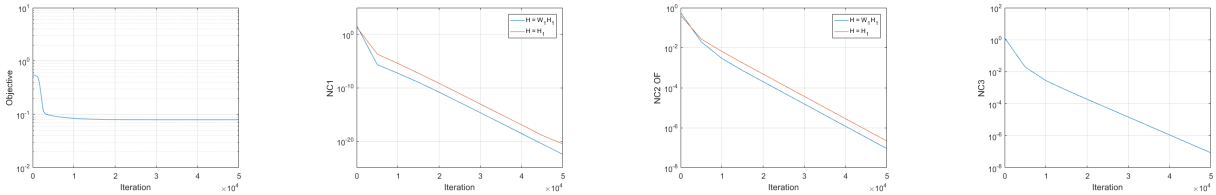

Figure 8. Verification of Theorem 4.1 (two levels of features). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to OF), and NC3 (alignment between the weights and the features).

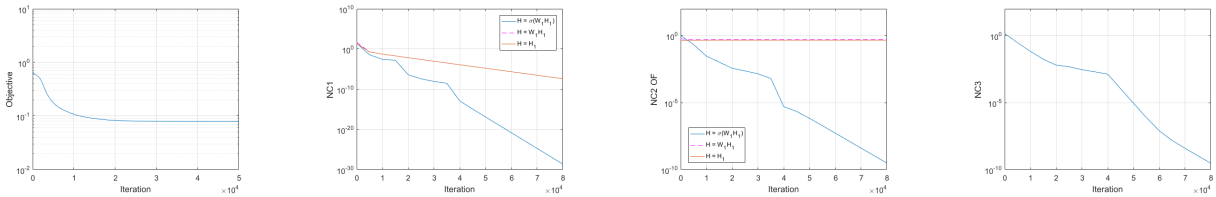

Figure 9. Verification of Theorem 4.2 (two levels of features with ReLU activation). From left to right: the objective value, NC1 (within-class variability), NC2 (similarity of the features to OF), and NC3 (alignment between the weights and the features).

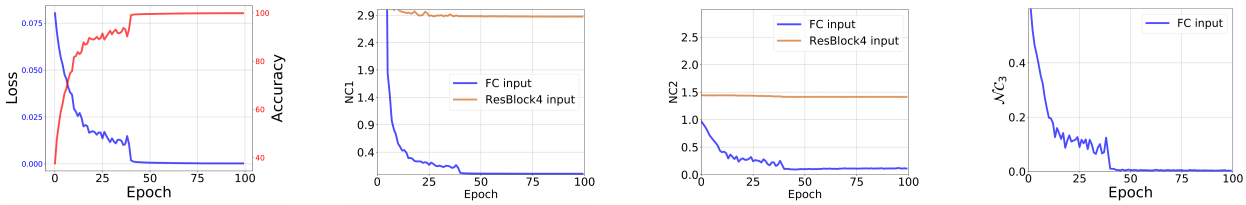

Figure 10. NC metrics for ResNet18 trained on CIFAR10 with MSE loss, weight decay, and no bias. From left to right: training's objective value and accuracy, NC1 (within-class variability), NC2 (similarity of the centered features to simplex ETF), and NC3 (alignment between the weights and the features).