On the Optimization Landscape of Neural Collapse under MSE Loss: Global Optimality with Unconstrained Features

Jinxin Zhou[* 1] Xiao Li[* 2] Tianyu Ding[3] Chong You[4] Qing Qu[2] Zhihui Zhu[1]

Abstract

When training deep neural networks for classification tasks, an intriguing empirical phenomenon has been widely observed in the last-layer classifiers and features, where (i) the class means and the last-layer classifiers all collapse to the vertices of a Simplex Equiangular Tight Frame (ETF) up to scaling, and (ii) cross-example within-class variability of last-layer activations collapses to zero. This phenomenon is called Neural Collapse (NC), which seems to take place regardless of the choice of loss functions. In this work, we justify NC under the mean squared error (MSE) loss, where recent empirical evidence shows that it performs comparably or even better than the *de-facto* cross-entropy loss. Under a simplified unconstrained feature model, we provide the first global landscape analysis for vanilla nonconvex MSE loss and show that the (only!) global minimizers are neural collapse solutions, while all other critical points are strict saddles whose Hessian exhibit negative curvature directions. Furthermore, we justify the usage of rescaled MSE loss by probing the optimization landscape around the NC solutions, showing that the landscape can be improved by tuning the rescaling hyperparameters. Finally, our theoretical findings are experimentally verified on practical network architectures. The source code is available at https://github.com/jinxinzhou/neural-collapse-MSE.

[*]Equal contribution [1]Department of Electrical & Computer Engineering, University of Denver, Denver, USA [2]Department of Electrical & Computer Engineering, University of Michigan, Ann Arbor, USA [3]Microsoft, Redmond, USA [4]Google Research, New York City, USA. Correspondence to: Qing Qu <qingqu@umich.edu>, Zhihui Zhu <zhihui.zhu@du.edu>.

1. Introduction

Despite the dramatic success of modern deep neural networks (DNNs) across engineering and sciences (Goodfellow et al., 2016; Krizhevsky et al., 2012; LeCun et al., 2015; Senior et al., 2020) that we have witnessed in the past decade, the practice of deep learning has yet been shrouded with mysteries, ranging from the design of appropriate network architectures (Martens et al., 2021; Qi et al., 2020) to the generalization and robustness properties (Madry et al., 2018; Nakkiran et al., 2019; Yang et al., 2020) of the learned networks. For instance, even the right choice of training loss function has not been thoroughly justified. For classification problems, although the cross entropy (CE) loss is the standard choice for network training, recent work (Hui & Belkin, 2021) demonstrated with extensive experiments that DNNs trained with mean-squared error (MSE) loss achieve on par or even better performance compared to those of the CE loss.

Towards demystifying DNN, a recent interesting line of work (Ergen & Pilanci, 2021; Fang et al., 2021; Graf et al., 2021a; Han et al., 2022; Mixon et al., 2020; Papyan, 2020; Papyan et al., 2020; Zhu et al., 2021) studied and characterized the learned deep representations during the terminal phase of training, where several intriguing phenomena have been discovered. In particular, recent seminal work of (Han et al., 2022; Papyan et al., 2020) empirically demonstrated that last-layer features and classifiers of a trained DNN exhibit the following *Neural collapse* ($\mathcal{NC}$) property:

($\mathcal{NC}$1) **Variability collapse:** the individual features of each class concentrate to their class-means.

($\mathcal{NC}$2) **Convergence to simplex ETF:** the class-means have the same length and are maximally distant; they form a Simplex Equiangular Tight Frame (ETF).

($\mathcal{NC}$3) **Convergence to self-duality:** the last-layer linear classifiers perfectly match their class-means.

($\mathcal{NC}$4) **Simple decision rule:** the last-layer classifier is equivalent to a Nearest Class-Center decision rule.

It has been empirically demonstrated that the $\mathcal{NC}$ persists across the range of canonical classification problems with the CE loss. These results imply that deep networks are

essentially learning maximally separable features between classes, and a max-margin classifier in the last layer upon these learned features, touching the ceiling in terms of the training performance. Later work theoretically investigated the $\mathcal{NC}$ based on a simplified assumption of the so-called *unconstrained feature model* (Mixon et al., 2020) or *layer-peeled model* (Fang et al., 2021), where the features are viewed as free optimization variables. The underlying reasoning is that modern deep networks are often highly overparameterized with the capacity of learning any representations (Cybenko, 1989; Hornik, 1991; Lu et al., 2017; Shaham et al., 2018), so that the last-layer features can approximate, or interpolate, any point in the feature space. Under the unconstrained feature model, the work (Fang et al., 2021; Graf et al., 2021a; Ji et al., 2021; Lu & Steinerberger, 2020; Mixon et al., 2020; Tirer & Bruna, 2022; Weinan & Wojtowytsch, 2020) showed that the $\mathcal{NC}$ solutions are the only global optimal solution for nonconvex training losses under different settings. However, given the nonconvexity of the problem, even under the unconstrained feature model these global optimality results do guarantee that the $\mathcal{NC}$ solutions can be efficiently achieved. This has been further resolved by the recent work (Zhu et al., 2021), showing that the CE loss function enjoys a benign global optimization landscape under the unconstrained feature model. It shows that every saddle point is a strict saddle with negative curvature, so that the CE loss can be efficiently optimized to the $\mathcal{NC}$ solution regardless of the nonconvexity.

It should be noted that the $\mathcal{NC}$ phenomenon is not *solely* pertinent to the particular choice of the CE loss. It has been recently reported (Han et al., 2022), that DNNs trained with the MSE loss also exhibit very similar $\mathcal{NC}$ phenomena but with even *faster* collapse in terms of training epochs and with better (adversarial) robustness. In the meanwhile, the MSE loss is not only appealing for its algebraic simplicity, but it also demonstrates on-par or even better generalization performances compared to the CE loss, as reported by recent line of work (Hui & Belkin, 2021). However, the theoretical study of MSE loss for $\mathcal{NC}$ is still limited (Han et al., 2022; Mixon et al., 2020; Tirer & Bruna, 2022). Under the unconstrained feature model, their work proved that the continuous gradient flow of the MSE loss converges to $\mathcal{NC}$ solutions. In particular, the work (Mixon et al., 2020) relies on linearizations of the ordinary differential equation by assuming very small initializations, which is not well aligned with the practice of deep learning where the weights are usually initialized with non-negligible magnitudes such as by the Kaiming initialization (He et al., 2015). Because the choice of the loss function without balanced weight decay, the analysis in (Han et al., 2022) only focuses on the renormalized features and studies the continually renormalized gradient flow.[1] Moreover, in practice deep networks are

usually trained using iterative algorithms such as stochastic gradient descent (SGD) with nontrivial stepsizes, rather than using the continuous gradient flows. The work (Graf et al., 2021b; Poggio & Liao, 2020a;b; Rangamani et al., 2021) study deep homogeneous classification networks (without bias terms but beyond the unconstrained features model) trained with MSE loss, stochastic gradient descent, and weight decay. In particular, the solutions satisfying the so-called symmetric quasi-interpolation assumption are proved to obey $\mathcal{NC}$ properties, but the properties of other solutions are not investigated (Graf et al., 2021b; Rangamani et al., 2021).

As far as we know, the work closest to ours is the concurrent work (Tirer & Bruna, 2022). Under similar unconstrained feature models, the work studies the global optimality condition of $\mathcal{NC}$ for the MSE loss for both two-layer and three layer networks, but *not* the global optimization landscape. Additionally, it studies special cases of the MSE loss with either no bias term, or no weight decay on the bias term. In comparison, our work not only study the MSE loss under more general setting with bias term included, but also shows the strict saddle property of the benign nonconvex landscape.

Contributions. In this work, we provide a thorough analysis of nerual network by examining its last-layer features. In particular, we work under the unconstrained feature model to characterize the *global* optimization landscape of overparameterized neural networks trained with the MSE loss. Our contributions can be highlighted as follows.

- **Characterization of global solutions.** We provide a mathematical characterization of *all* the global solutions for the last layer features and classifier, showing that they satisfy the $\mathcal{NC}$ properties with certain choices of regularization parameters. This is in contrast to previous work (Han et al., 2022; Mixon et al., 2020) which only characterize the solutions that are produced by a particular optimization algorithm (i.e., gradient flow). Moreover, these work only consider cases that the feature dimension is larger than the number of classes, while our analysis covers all choices of feature dimension.

- **Benign global landscape.** We prove that the loss function is a *strict saddle function* (Ge et al., 2015; Sun et al., 2015; Zhang et al., 2020), where every critical point is either a global solution or a *strict saddle point* with negative curvature. This implies that there is *no* spurious local minimizer on the optimization landscape. Hence, our work is distinguished from previous work (Ergen & Pilanci, 2021; Fang et al., 2021; Graf et al., 2021a; Lu & Steinerberger,

---

[1] The model used in (Han et al., 2022) imposes a weight decay on the classifier, but not on the features. Thus, without renormalization, the weights of the classifier will converge to zero while the features will blow up.

2020; Mixon et al., 2020; Tirer & Bruna, 2022; Weinan & Wojtowytsch, 2020) that only characterizes global minimizers. The benign global landscape implies that any method that can escape strict saddle points (e.g. stochastic gradient descent) converges to a global solution that exhibits $\mathcal{NC}$ (see Section 4).

- **Understanding the *rescaled* MSE.** In practice, rescaling the MSE loss (see Section 2.2) is empirically demonstrated to be critical for obtaining competitive performance compared to the CE loss particularly when the number of classes is large (Demirkaya et al., 2020; Hui & Belkin, 2021). We show empirically that the $\mathcal{NC}$ exhibits for rescaled MSE as well. To understand the benefit of the rescaling, we provide a visualization of the optimization landscape w.r.t. unconstrained features, showing that rescaling aligns the gradient direction to be perpendicular to the decision boundary between classes hence may facilitate the convergence of gradient based algorithms to more discriminative features.

Compared to the recent global landscape analysis for the CE loss (Zhu et al., 2021), our result implies that both losses learn similar $\mathcal{NC}$ features and classifiers when $d \geq K$. Hence, from the $\mathcal{NC}$ perspective, this work provides a theoretical explanation for the observations in (Hui & Belkin, 2021) that the DNN trained by the MSE loss achieves on par performance compared to that trained with the CE loss. Additionally, it should be noted that there are several major differences between our result and (Zhu et al., 2021). First, the work of (Zhu et al., 2021) only studied the setting where the feature dimension d is larger than the number of classes K, while we characterized the global optimality for both the cases of $d < K$ and $d \geq K$. We observe dramatically different performance for DNN learned by CE and MSE when $d < K$. Second, for the MSE loss, we showed that the bias term plays an important role[2] for the solution to be $\mathcal{NC}$, while for CE loss the $\mathcal{NC}$ solution can be achieved without bias terms.

2. The Problem Setup

The goal of deep learning is to learn a multi-layer nonlinear mapping $\psi(\cdot) : \mathbb{R}^m \mapsto \mathbb{R}^K$, that is able to fit the training data and generalize. More precisely, a deep neural network classifier can be generally written as

$$\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{W}\phi_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{b}, \quad (1)$$

where $\phi_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^m \mapsto \mathbb{R}^d$ is the feature mapping, on top of which is the linear classifier $(\boldsymbol{W}, \boldsymbol{b})$. $\phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is usually referred to as the *representation* or *feature* of the input $\boldsymbol{x}$ learned from the network. For convenience, we use $\boldsymbol{\theta}$ to

---

[2]For the MSE loss, when there is no bias term, the features (and classifier) that minimize the loss function form orthonormal matrices instead of Simplex ETFs when $d \geq K$.

denote the network parameters in the feature mapping, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b}\}$ to denote *all* the network parameters. In this way, the function implemented by a neural network classifier can also be expressed as a linear classifier acting upon $\phi_{\boldsymbol{\theta}}(\boldsymbol{x})$.

In this work, we focus on learning deep networks for multi-class classification tasks (say, with K classes), where the class label of a sample $\boldsymbol{x}_{k,i}$ in the k-th class is given by a one-hot vector $\boldsymbol{y}_k \in \mathbb{R}^K$ with only the kth entry equal to unity ($1 \leq k \leq K$). Throughout the paper, we study the setting where the number of training samples in each class is balanced, i.e., each class has n training samples. Let $N = Kn$. During the training phase, the task is then to learn the parameters $\boldsymbol{\Theta}$ so that the output of the model on an input sample $\boldsymbol{x}_{k,i}$ approximates the corresponding output $\boldsymbol{y}$ (i.e. $\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i}) \approx \boldsymbol{y}_k$). To quantify this approximation, it can be done by optimizing a simple MSE loss as follows

$$\min_{\boldsymbol{\Theta}} \ \frac{1}{2N} \sum_{k=1}^{K} \sum_{i=1}^{n} \|\psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i}) - \boldsymbol{y}_k\|_2^2 \ + \ \frac{\lambda}{2} \|\boldsymbol{\Theta}\|_F^2, \quad (2)$$

where $\lambda > 0$ is the regularization parameter (a.k.a., the weight decay parameter).

2.1. Basic Problem Formulation Based on Unconstrained Feature Models

Analyzing deep networks is a tremendously difficult task mainly due to the nonlinear interactions between a large number of layers. Nonetheless, as argued by a line of recent work (Cybenko, 1989; Hornik, 1991; Lu et al., 2017; Shaham et al., 2018) that modern deep networks are often highly overparameterized to approximate any continuous function, it motivates us to simplify the analysis by treating the last-layer features as *free* optimization variables $\boldsymbol{h}_{k,i} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i}) \in \mathbb{R}^d$. Such a simplification is called *unconstrained feature model* (Mixon et al., 2020) (or *layer-peeled model* in (Fang et al., 2021)), which simplifies the study of the last-layer representations of the network. To simplify the notation, let us denote

$$\boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}^1 & \boldsymbol{w}^2 & \cdots & \boldsymbol{w}^K \end{bmatrix}^{\top} \in \mathbb{R}^{K \times d},$$
$$\boldsymbol{H} := \begin{bmatrix} \boldsymbol{H}_1 & \boldsymbol{H}_2 & \cdots & \boldsymbol{H}_K \end{bmatrix} \in \mathbb{R}^{d \times N}, \text{ and}$$
$$\boldsymbol{Y} := \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \cdots & \boldsymbol{Y}_K \end{bmatrix} \in \mathbb{R}^{K \times N},$$

where $\boldsymbol{w}^k$ is a row vector of $\boldsymbol{W}$, $\boldsymbol{H}_k := \begin{bmatrix} \boldsymbol{h}_{k,1} & \cdots & \boldsymbol{h}_{k,n} \end{bmatrix} \in \mathbb{R}^{d \times n}$ contains all the k-th class features, and $\boldsymbol{Y}_k := \begin{bmatrix} \boldsymbol{y}_k & \cdots & \boldsymbol{y}_k \end{bmatrix} \in \mathbb{R}^{K \times n}$ for all $k = 1, 2, \cdots, K$. Based on the unconstrained feature model, we consider a slight variant of (2), given by

$$\min_{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) := \left\{ \frac{1}{2N} \left\| \boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\boldsymbol{1}_N^{\top} - \boldsymbol{Y} \right\|_F^2 \right.$$
$$\left. + \frac{\lambda_{\boldsymbol{W}}}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2} \|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2^2 \right\}, \quad (3)$$

where $\lambda_W, \lambda_H, \lambda_b > 0$ are the penalties for W, H, and b, respectively.

Here, because we treat the last-layer feature H as a free optimization variable, we put the weight decay on W and H, which is different from the practice that the weight decay is enforced on all the network parameters Θ as shown in (2). Nonetheless, as discussed in (Zhu et al., 2021), this idealization is reasonable since the energy of the features (i.e., $\|H\|_F$) can indeed be upper bounded by the energy of the weights at every layer if the inputs are bounded (which holds in practice), implying that the norm of H is *implicitly* penalized by penalizing the norm of Θ. Additionally, for the CE loss, the experiments in (Zhu et al., 2021) show on-par performance for the two types of weight decay. Thus, we expect similar performances for the MSE loss.

On the other hand, the experiments in (Graf et al., 2021a; Zhu et al., 2021) conducted on random labels imply that the strong assumption of unconstrained feature model is reasonable for explaining $\mathcal{NC}$ during the training phase: when the network (1) is highly overparameterized, the learned network in practice will fit to the random labels and neural collapse, regardless of the input. Moreover, as we shall see in the following sections, both theory and experiments demonstrate that such simplification preserves the core properties of last-layer classifiers and features—the $\mathcal{NC}$ phenomenon.

2.2. Rescaled MSE Loss under Unconstrained Features

On the other hand, it should be noted that, when training with the vanilla formulation of the MSE loss (2), empirically good performances are reported *only* when the number of classes is small (e.g., CIFAR10 (Krizhevsky et al., 2009) with $K < 100$). When training for a large number of classes such as ImageNet (Deng et al., 2009), to achieve better performance *rescaling* is often needed (Demirkaya et al., 2020; Hui & Belkin, 2021). Intuitively, the basic idea is to rescale the MSE loss (3) by a pair of positive scalars (α, M),

$$\min_{W,H,b} \frac{1}{2N} \left\| \Omega_\alpha^{\odot 1/2} \odot \left(WH + b\mathbf{1}^\top - MY \right) \right\|_F^2$$
$$+ \frac{\lambda_W}{2} \|W\|_F^2 + \frac{\lambda_H}{2} \|H\|_F^2 + \frac{\lambda_b}{2} \|b\|_2^2, \quad (4)$$

so that we can put more emphasize on training the correct class. Here, $\odot$ denotes the entry-wise Hadamard product, $\Omega^{\odot 1/2}$ means taking square root for each element, and $\Omega_\alpha = \begin{bmatrix} \omega_1 \mathbf{1}_n^\top & \cdots & \omega_K \mathbf{1}_n^\top \end{bmatrix}$, where $\omega_k \in \mathbb{R}^K$ with entries $\omega_{ki} = \begin{cases} \alpha, & i = k, \\ 1, & \text{otherwise.} \end{cases}$ In comparison to (Han et al., 2022; Mixon et al., 2020; Tirer & Bruna, 2022), our work not only studies $\mathcal{NC}$ under the vanilla setting (3) but also investigates the more practical rescaled version of the MSE loss (4). In particular, in Section 3.3, we provide geometric intuitions on why rescaling would be a better choice for loss

design. We will corroborate our reasoning via experiments on practical network training in Section 4.

3. Main Theoretical Results

In this section, we present our study on global optimality conditions as well as geometric properties of the nonconvex (rescaled) MSE loss under the unconstrained feature model.

3.1. Global Optimality Conditions

First, we study the nonconvex MSE loss (3) by characterizing its global solutions under different settings of the feature and class dimensions. We show that the only global solutions of (3) are neural collapsing, satisfying the $\mathcal{NC}$ properties introduced at the beginning of Section 1.

Theorem 3.1 (Global Optimality Conditions). *Let* $(W^\star, H^\star, b^\star)$ *be a global minimizer of the vanilla MSE loss in (3). Let* $\overline{H}^\star = \begin{bmatrix} \overline{h}_1^\star & \cdots \overline{h}_K^\star \end{bmatrix}$*, with* $\overline{h}_k^\star$ *being the mean of the k-th class features. Then,* $(W^\star, H^\star, b^\star)$ *satisfies the following properties:*

- *If* $\lambda_W \lambda_H < \frac{1}{NK}$*, then* $(W^\star, H^\star)$ *satisfies* $\mathcal{NC}1$ *and* $\mathcal{NC}3$ *as*

$$h_{k,i}^\star = \overline{h}_k^\star, \ \sqrt{\frac{\lambda_W}{\lambda_H n}} w^{\star k} = \overline{h}_k^\star, \quad \forall \, k \in [K], \ i \in [n].$$

Otherwise, if $\lambda_W \lambda_H \geq \frac{1}{NK}$*, then* $W^\star = 0, H^\star = 0$*.*

- *If* $\lambda_W \lambda_H < \frac{1}{NK}$*, then* $\overline{H}^\star$ *further obeys the following properties ($\mathcal{NC}2$) for different d:*

  1. *If* $d < K - 1$*: we have* $\overline{H}^{\star\top} \overline{H}^\star = C_1 \mathcal{P}_d(I - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)$*, where* $\mathcal{P}_d(M)$ *denotes the best rank-d approximating of* M*;*

  2. *If* $d = K - 1$*: we have* $\overline{H}^{\star\top} \overline{H}^\star = C_2(I - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)$*;*

  3. *If* $d \geq K$*: we have* $\overline{H}^{\star\top} \overline{H}^\star =$

$$\begin{cases} C_3 \left(I - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top \right), & \text{if } \lambda_b \leq \frac{\sqrt{KN\lambda_W\lambda_H}}{1 - \sqrt{KN\lambda_W\lambda_H}} \\ C_4\big(I - \frac{\sqrt{n\lambda_W\lambda_H}}{\lambda_b(1 - \sqrt{KN\lambda_W\lambda_H})}\mathbf{1}_K\mathbf{1}_K^\top \big), & \text{otherwise} \end{cases}$$
$$(5)$$

  where $\frac{\sqrt{n\lambda_W\lambda_H}}{\lambda_b(1 - \sqrt{KN\lambda_W\lambda_H})} \leq \frac{1}{K}$ *in the second case since* $\lambda_b \geq \frac{\sqrt{KN\lambda_W\lambda_H}}{1 - \sqrt{KN\lambda_W\lambda_H}}$*.*

Here, C_1*,* C_2*,* C_3*, and* C_4 *are some positive numerical constants that depend on* $\lambda_W, \lambda_H, \lambda_b$*.*

- *The bias satisfies* $b^\star = b^\star \mathbf{1}_K$ *with* $b^\star \leq \frac{1}{K}$ *given by:*

  1. *If* $d < K$*: we have* $b^\star = \frac{1}{K(\lambda_b + 1)}$*;*

  2. *Otherwise,* $b^\star = \begin{cases} \frac{1}{K(\lambda_b + 1)}, & \lambda_b \leq \frac{\sqrt{KN\lambda_W\lambda_H}}{1 - \sqrt{KN\lambda_W\lambda_H}}, \\ \frac{\sqrt{n\lambda_W\lambda_H}}{\lambda_b}, & \text{otherwise.} \end{cases}$

In particular, when $\lambda_b \to 0$*, we have* $b^\star \to \frac{1}{K}$*; when* $\lambda_b \to \infty$*, we have* $b^\star \to 0$*.*

We postpone the detailed proof to Appendix B. In the following, we discuss the implications of Theorem 3.1 in detail.

- **Implications on the choice of the feature dimension d.**
  As we observe from Theorem 3.1, for the MSE loss (3), any global solution always exhibits variability collapse ($\mathcal{NC}1$) and self-duality ($\mathcal{NC}3$). However, the convergence of class means to simplex ETF ($\mathcal{NC}2$) critically depends on the feature dimension d. When $d \geq K - 1$, for proper choices of $\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}$, and $\lambda_{\boldsymbol{b}}$, the global configuration of the class mean $\overline{\boldsymbol{H}}^{\star}$ is always a simplex ETF. In particular, when $d = K - 1$, the simplex ETF configuration even does not depend on $\lambda_{\boldsymbol{b}}$. On the other hand, if $d < K - 1$, our theory implies that the global solution for $\overline{\boldsymbol{H}}^{\star}$ is only the best rank-d approximation of the simplex ETF, where the class-means of the each class are neither having equal length nor being maximally pairwise-distanced. This result is consistent with the fact that K vectors in $\mathbb{R}^d$ cannot form a K-Simplex ETF if $K > d - 1$, and supports the practice of learning overparameterized network for choosing $d \geq K$.[3]

- **Comparison to the CE loss.** For the CE loss under the unconstrained feature model, when $d \geq K$ recent work (Zhu et al., 2021) showed that any global solution satisfies all three $\mathcal{NC}$ properties regardless of choices of the weight decay parameters (i.e., $\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}$, and $\lambda_{\boldsymbol{b}}$). Moreover, the bias term there becomes zero. In contrast, Theorem 3.1 shows that the solution with the MSE loss is dependent upon choice of regularization parameters $\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}, \lambda_{\boldsymbol{b}}$ and that the class mean $\overline{\boldsymbol{H}}^{\star}$ may not be a simplex ETF. Moreover, the bias term is essential to achieve simplex ETF solutions for MSE loss. Without the bias term (i.e., $\lambda_{\boldsymbol{b}} \to \infty$), (5) implies that the class mean $\overline{\boldsymbol{H}}^{\star}$ becomes an orthonormal matrix even when $d \geq K$. Thus, the analysis of global optimality conditions for the MSE loss is more complicated than for the CE loss[4].

- **Comparison to previous work (Han et al., 2022; Mixon et al., 2020).** As discussed in Section 1, the previous work (Han et al., 2022; Mixon et al., 2020) only characterize the solutions to (3) that are produced by a particular optimization algorithm (i.e., gradient flow) and under specific cases such as $\lambda_{\boldsymbol{b}} \to 0$ and the feature dimension is larger than the number of classes. In contrast, we characterize

---

[3]For example, the dimension of the features of a ResNet (He et al., 2016) is typically set to $d = 512$ for CIFAR10 (Krizhevsky et al., 2009), a dataset with $K = 10$ classes. This dimension grows to $d = 2048$ for ImageNet (Deng et al., 2009), a dataset with $K = 1000$ classes.

[4]The proof of Theorem 3.1 is also dramatically different to the one for CE loss in Zhu et al. (2021): the latter mainly shows that $\mathcal{NC}$ solutions have small objective value than others since $\mathcal{NC}$ solutions are the only global minimizers, while the proof of Theorem 3.1 directly analyzes the global minimizers for different scenarios.

the global optimality conditions for the MSE loss (3) and our analysis covers all choices of feature dimension and weight decay parameters.

- **Extension to the rescaled MSE.** Although our current analysis is only for the vanilla MSE loss (3), we expect that similar global optimality results should also hold for the rescaled version (4). This has been corroborated by our experimental results in Section 4. Notice that if we fix $\alpha = 1$ in (4), the analysis only with large M is simple and remain the same as Theorem 3.1. However, dealing with both α and M requires extra technicalities, that we leave for future work.

3.2. Characterizations of The Benign Global Landscape

Theorem 3.1 implies that the (only!) global minimizers to (3) are those satisfying $\mathcal{NC}$ properties. However, the MSE loss function is *nonconvex*, hence it is not obvious whether the benign global solutions can be *efficiently* achieved even under the unconstrained feature model. To deal with this challenge, in the following we further investigate the global optimization landscape of (3). By leveraging recent advances on nonconvex optimization (Ge et al., 2015; Qu et al., 2020a;b; Sun et al., 2015; 2016; 2018; Zhang et al., 2020), we first show that our nonconvex MSE loss (3) without bias term is a *strict saddle* function that every non-global critical point is a saddle point with negative curvature (i.e., its Hessian has at least one negative eigenvalue).

Theorem 3.2. (Benign landscape for MSE without bias term) *The following MSE loss without bias term*

$$\frac{1}{2N} \|\boldsymbol{W}\boldsymbol{H} - \boldsymbol{Y}\|_F^2 + \frac{\lambda_{\boldsymbol{W}}}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2} \|\boldsymbol{H}\|_F^2$$

is a strict saddle function with no spurious local minimum. That is, any of its critical point is either a global minimizer, or it is a strict saddle point whose Hessian has a strictly negative eigenvalue.

We postpone the proof to Appendix B (see Lemma B.1). By viewing $\boldsymbol{W}$ and $\boldsymbol{H}$ as two factors of a matrix $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}$, the formulation in (3) is closely related to nonconvex low-rank matrix problems (Bhojanapalli et al., 2016; Chi et al., 2019; Ge et al., 2016; 2017; Haeffele & Vidal, 2015; Li et al., 2019a;b) with the Burer-Moneirto factorization approach (Burer & Monteiro, 2003). In particular, the work (Ciliberto et al., 2017; Li et al., 2019a) studied a similar problem with $\lambda_{\boldsymbol{W}} = \lambda_{\boldsymbol{H}}$, but only for particular choices of d: d is either required to be exactly the rank of the solution of the corresponding convex problem (Li et al., 2019a), or relatively large in (Ciliberto et al., 2017). In contrast, our Theorem 3.2 characterizes the benign landscape for all choices of feature dimension.

The following result establishes global optimization landscape of the MSE loss (3).

Theorem 3.3. (Benign landscape for MSE loss (3)) *As-*

sume that the feature dimension d is larger than the number of classes K. The nonconvex MSE loss function $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ in (3) *is a strict saddle function.*

This result is similar to that of (Zhu et al., 2021, Theorem 3.2), which showed that the unconstrained feature model with CE loss is a strict saddle function. The high level proof idea for Zhu et al. (2021) is to construct the negative curvature direction for saddle points in the null space of $\boldsymbol{W} \in \mathbb{R}^{K \times d}$, inspired by the recent advances on matrix factorization problems (Bhojanapalli et al., 2016; Chi et al., 2019; Ciliberto et al., 2017; Ge et al., 2016; 2017; Haeffele & Vidal, 2015; Li et al., 2019a;a;b). We refer to (Zhu et al., 2021) for the detailed discussion on the connection and differences between the two problems. Because the proof in (Zhu et al., 2021) actually holds more generally for any smooth convex loss function with weight decay, the same technique also offers a proof for Theorem 3.3 (and potentially can extend Theorem 3.3 for the rescaled MSE in (4)). Here, it should be noted that we make the assumption $d > K$ so that the null space of $\boldsymbol{W} \in \mathbb{R}^{K \times d}$ always exists. However, we believe the strict saddle property holds for any d and leave it as future work.

As a consequence, if $\boldsymbol{H}$ is a free optimization variable, this implies that the global solutions of the training problem in (3) can be efficiently found by many first-order and second-order optimization methods (Bottou et al., 2018). In particular, (stochastic) gradient descent with random initialization is guaranteed (Ge et al., 2015; Lee et al., 2016) to almost surely find a global minimizer for strict saddle functions with no spurious local minima, which is the case for our problem (3). In comparison, existing results on MSE loss (Han et al., 2022; Mixon et al., 2020) only studied the trajectory of gradient flows (3) on either the linear terms (Mixon et al., 2020) or the central path component (Han et al., 2022), which is insufficient to explain/guarantee efficient, global convergence of iterative optimization algorithms.

3.3. Delving Deeper into The Optimization Landscape: Why Rescaling (4) Helps?

While our global landscape analysis for the vanilla MSE loss (3) in Section 3.2 implies that a gradient based algorithm converges to global $\mathcal{NC}$ solutions *asymptotically* (Lee et al., 2016), it did not characterize the rate of convergence – in other words, how fast an optimization method converges. Often around the global solutions (i.e., the simplex ETF), we expect that the landscape has certain regularity condition which measures how
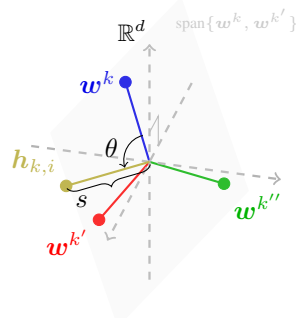

Figure 1. Visualization method.

well-aligned between the negative gradient direction and the direction towards the global solution. Thus, the regularity conditions in turn will characterize how fast a gradient based method converges. For better understanding the regularity properties and algorithmic convergences, we use visualization techniques to visualize the optimization landscape of MSE losses around the global ETFs solutions. In particular, our visualization sheds light on (*i*) why training with vanilla MSE loss performs worse than that of the CE loss, and (*ii*) how the rescaling techniques in Section 2.2 improves the performance of the MSE loss.

Even under the unconstrained feature model, visualization of the MSE loss landscape could still be difficult, which is due to the fact that the variables $\boldsymbol{H}, \boldsymbol{W}$, and $\boldsymbol{b}$ are all high-dimensional. Here, we further simplify the problem by assuming $\boldsymbol{b} = \boldsymbol{0}$ and that $\boldsymbol{W}$ is at the global optimum and forms a simplex ETF. Thus, we can examine the landscape only with respect to (w.r.t.) the feature vectors $\boldsymbol{h}_{k,i}$ for the kth class. Although $\boldsymbol{h}_{k,i} \in \mathbb{R}^d$ is still high-dimensional for large d, we plot the optimization landscape by restricting $\boldsymbol{h}_{k,i}$ to a 2D plane spanned by $\{\boldsymbol{w}^k, \boldsymbol{w}^{k'}\}$, where $\boldsymbol{w}^k$ is the classifier for the kth class and $k' \neq k$ can be chosen arbitrarily because the simplex ETF is invariant to rotations. Finally, we visualize the landscape using the polar coordinates, where the s-axis denotes the ℓ_2 norm of $\boldsymbol{h}_{k,i}$ and the θ-axis denotes the angle between $\boldsymbol{h}_{k,i}$ and $\boldsymbol{w}^k$ (see Figure 1 for an illustration). The predicted membership for $\boldsymbol{h}_{k,i}$ is determined by θ and is invariant to s. Hence, larger gradient along the θ direction may help with learning more discriminative features. See Appendix C for a formal explanation. This design choice allows us to examine the gradient in directions co-linear to (i.e., with varying s) and perpendicular to (i.e., with varying θ) the decision boundary separately.

In Figure 2, the visualizations of landscapes of different loss functions are provided. As we observe from Figure 2a, the landscape of vanilla MSE loss is steep w.r.t. s while it is flat w.r.t. θ. Because the size of θ determines the closeness to the right class, this implies that optimizing the vanilla MSE loss will take a longer time to converge to a desired solution with $\theta \approx 0$. In contrast, the landscape of CE loss in Figure 2b is steeper w.r.t. θ than w.r.t. s in a large region where $s > 1$ and $\theta < 1.5$. This difference of the landscapes around the global solutions potentially explains why CE is a preferred choice than the vanilla MSE, given that the features $\boldsymbol{h}_{k,i}$ would converge faster to the simplex ETF solutions via optimizing the CE loss. Nonetheless, the issue with the vanilla MSE can be mitigated via the rescaling approach that we discussed in Section 2.2.

To intuitively explain the effect of the rescaling approach for MSE, we note that the gradient of the rescaled MSE (denoted by f_{RMSE}) w.r.t. a feature vector $\boldsymbol{h}_k$ from k-th
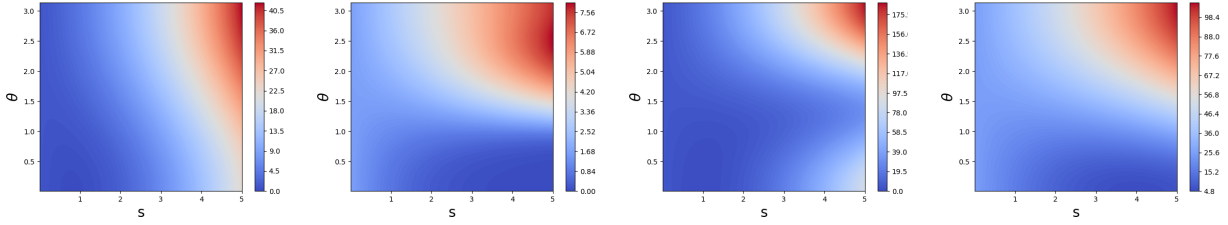
(a) Vanilla MSE ($\alpha = 1$, $M = 1$) (b) Cross Entropy (c) Rescaled MSE ($\alpha = 5$, $M = 1$) (d) Rescaled MSE ($\alpha = 1$, $M = 5$)

Figure 2. **Visualization of optimization landscape with different losses.** We fix $\boldsymbol{W}$ as a simplex ETF and illustrate the landscape only w.r.t. a feature $\boldsymbol{h}_{k,i}$. For each plot, the s-axis denotes $\|\boldsymbol{h}_{k,i}\|_2$, and the θ-axis denotes the angle $\arccos\left(\langle \boldsymbol{h}_{k,i}, \boldsymbol{w}^k \rangle\right)$.

class is

$$\nabla_{\boldsymbol{h}_k} f_{RMSE} = \alpha(\boldsymbol{w}^{k\top}\boldsymbol{h} - M)\boldsymbol{w}^k + \sum_{j \neq k}(\boldsymbol{w}^{j\top}\boldsymbol{h})\boldsymbol{w}^j,$$

which indicates that α and M control the direction of $\boldsymbol{h}_k$ moves to. This is captured visually in Figure 2. Likewise, the gradient of CE loss (denoted by f_{CE}) is

$$\nabla_{\boldsymbol{h}_k} f_{CE} = \beta(\sum_{j \neq k} e^{\boldsymbol{w}^{j\top}\boldsymbol{h}})\boldsymbol{w}^k + \beta \sum_{j \neq k} e^{\boldsymbol{w}^{j\top}\boldsymbol{h}}\boldsymbol{w}^j,$$

where $\beta = 1/\sum_{j=1}^K e^{\boldsymbol{w}^{j\top}\boldsymbol{h}}$. In the case when $\boldsymbol{h}_k$ is close to $\boldsymbol{w}^k$, $e^{\boldsymbol{w}^{j\top}\boldsymbol{h}} \approx 1 + \boldsymbol{w}^{j\top}\boldsymbol{h}$ for $j \neq k$ and thus

$$\nabla_{\boldsymbol{h}_k} f_{CE} \approx \beta(\boldsymbol{w}^{k\top}\boldsymbol{h} - K)\boldsymbol{w}^k + \beta \sum_{j \neq k}(\boldsymbol{w}^{j\top}\boldsymbol{h})\boldsymbol{w}^j$$

since $\sum_j \boldsymbol{w}^j = 0$. We can see that increasing M makes $\nabla_{\boldsymbol{h}_k} f_{RMSE}$ more similar to $\nabla_{\boldsymbol{h}_k} f_{CE}$. As shown in Figures 2c and 2d, the rescaled MSE loss (4) (with large M, in particular), leads to a "better" optimization landscape similar to that of the CE loss. Therefore, through studying the $\mathcal{NC}$ and corresponding optimization landscapes, our work provides intuitive explanations on (*i*) the incompetence of the vanilla MSE loss (3), and (*ii*) the effectiveness of rescaling (4) for classification tasks.
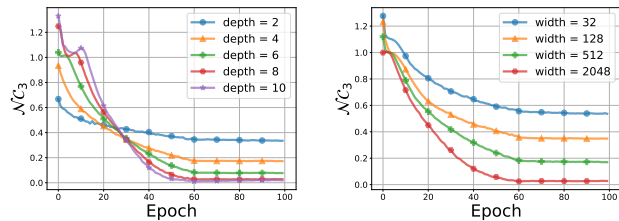
4. Experiments

Figure 3. Learning curves of $\mathcal{NC}_3$ for Multilayer Perceptron (with different depth and width) on CIFAR10 trained with MSE loss.

In this section, we conduct experiments to validate our findings from Section 3 on practical networks and standard datasets. We first introduce new metrics to better evaluate how well the $\mathcal{NC}$ properties are satisfied in practical neural networks, in addition to the ones used in (Papyan et al., 2020; Zhu et al., 2021). Second, we verify our theoretical results in Section 3.1 by showing that the $\mathcal{NC}$ phenomena

are algorithmic independent. Third, by a similar experiment as in (Zhu et al., 2021), we show that we could fix the last layer weights as a Simplex ETF while achieving comparable generalization performances as explicitly training the classifier. Finally, we examine our findings in Section 3.3 that the rescaling factors in the rescaled MSE loss is beneficial for forming benign optimization landscapes. For the details of the experimental setup, we refer readers to the Appendix A.

New metrics for evaluating $\mathcal{NC}$. To evaluate the $\mathcal{NC}$ properties of well-trained neural networks, we adopt the same $\mathcal{NC}_1$, $\mathcal{NC}_2$ and $\mathcal{NC}_3$ metrics as Papyan et al. (2020); Zhu et al. (2021), which measure the within-class variability of $\boldsymbol{H}$, the convergence of $\boldsymbol{W}$ to a simplex ETF, and the self-duality between $\boldsymbol{H}$ and $\boldsymbol{W}$; see Appendix A for the details. To better measure $\mathcal{NC}$, this paper also introduces the following two metrics:

- **Numerical rank.** The $\mathcal{NC}_1$ metric measures the variability collapse through the between-class and within-class covariance matrices, which does not directly reveal the dimensionality of the features spanned for each class. Ideally, when $\mathcal{NC}$ happens, for each class the feature dimension should collapse to one. To measure the dimensionality, we introduce a new metric that we call it *numerical rank*, denoted by $\widetilde{\text{rank}}(\boldsymbol{H}) := \frac{1}{K}\sum_{k=1}^K \frac{\|\boldsymbol{H}_k\|_*^2}{\|\boldsymbol{H}_k\|_F^2}$. Here, $\|\cdot\|_*$ represents the nuclear norm (Recht et al., 2010) (i.e., the sum of singular values), while the Frobenius norm $\|\cdot\|_F$ in the denominator serves as a normalization factor. The metric is evaluated by averaging over all the classes. Our metric is inspired by the numeral sparsity (defined as $\|\boldsymbol{a}\|_1^2/\|\boldsymbol{a}\|_2^2$ for $\boldsymbol{a} \in \mathbb{R}^n$) that serves as a stable measure for sparsity of vectors (Lopes, 2013). For our numerical rank, we expect that the smaller $\widetilde{\text{rank}}(\boldsymbol{H})$ is, the more collapsed the features are to their class means.

- **Cosine margin.** All current metrics measure $\mathcal{NC}$ from a panoramic view, and do not quantify the behavior of individual features. We introduce a metric based on the consine margin of individual features. From the explanation in Section 3.3, neural network determines the class member by the direction of features rather than its length. Thus, we define the cosine margin for each sample as $CM_{k,i} = \cos\theta_{k,i;k} - \max_{j \neq k}\cos\theta_{k,i;j}$, where
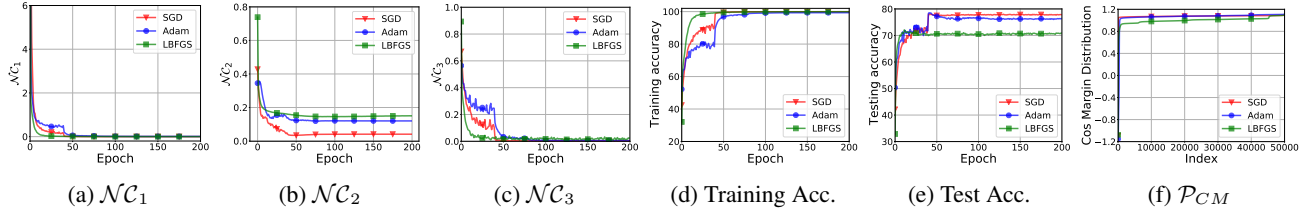
(a) $\mathcal{NC}_1$ (b) $\mathcal{NC}_2$ (c) $\mathcal{NC}_3$ (d) Training Acc.   (e) Test Acc.   (f) $\mathcal{P}_{CM}$

Figure 4. **Illustration of $\mathcal{NC}$, training and test accuracy and cosine margin across different training algorithms** with ResNet18 on CIFAR10. The networks are trained without data augmentation.
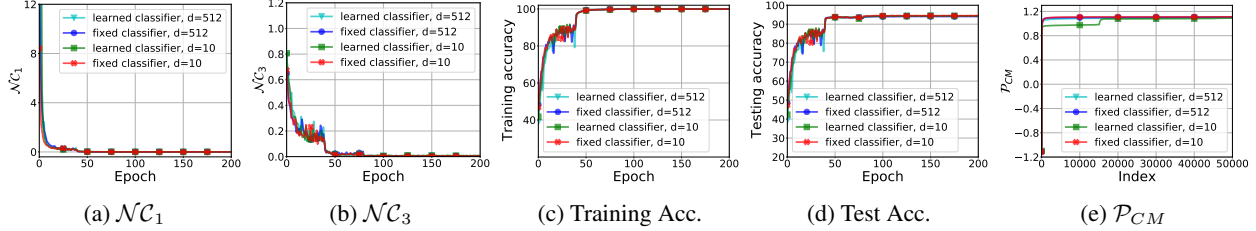

(a) $\mathcal{NC}_1$ (b) $\mathcal{NC}_3$ (c) Training Acc.   (d) Test Acc.   (e) $\mathcal{P}_{CM}$

Figure 5. **Comparison of the performances on learned vs. fixed last-layer classifiers.** We compare within-class variation collapse $\mathcal{NC}_1$, self-duality $\mathcal{NC}_3$, training accuracy, test accuracy and cosine margin distribution $\mathcal{P}_{CM}$ on fixed and learned classifier on CIFAR10-ResNet18 with data augmentation. The network is trained by SGD optimizer.

$\cos\theta_{k,i;j} = \frac{\langle \boldsymbol{w}^j - \boldsymbol{w}_G, \boldsymbol{h}_{k,i} - \boldsymbol{h}_G \rangle}{\|\boldsymbol{w}^j - \boldsymbol{w}_G\|_2 \|\boldsymbol{h}_{k,i} - \boldsymbol{h}_G\|_2}$ represents the cosine of the angle between the feature $\boldsymbol{h}_{k,i}$ and the j-th classifier $\boldsymbol{w}^j$, $\boldsymbol{h}_G$ denotes the global mean of all the features, and $\boldsymbol{w}_G$ denotes the mean of all the rows in $\boldsymbol{W}$. Recall that $\boldsymbol{h}_{k,i} \in \mathbb{R}^d$ denotes the feature of i-th sample in the k-th class and $\boldsymbol{w}^j \in \mathbb{R}^d$ denotes the j-th row of the linear classifier weight $\boldsymbol{W} \in \mathbb{R}^{K \times d}$. We sort the cosine margins over the training dataset in the ascending order and denote the resulted distribution as $\mathcal{P}_{CM}$. We note that a similar metric has been explored by the work (Banburski et al., 2021) as an alternative for the probability margin.[5]

The influence of the width and depth of neural networks on the $\mathcal{NC}$ phenomenon. Our landscape analysis is built on the unconstrained feature model, which is based on the assumption that the network has large capacity. Here, we conduct experiments on multilayer perceptron networks for CIFAR10 to illustrate the effect of network capacity on the degree of $\mathcal{NC}$. From Figure 3, we observe that the more expressive (wider/deeper) the network is (the more parameters of the network), the more severe $\mathcal{NC}$ occurs.

The prevalence of $\mathcal{NC}$ across different optimization algorithms. The benign landscape for optimization of neural networks with vanilla MSE loss suggests the existence of $\mathcal{NC}$ regardless of specific choice of the optimizer. We validate this result by training ResNet18 on CIFAR10 with vanilla MSE loss, using three different optimization algorithms: SGD, Adam and L-BFGS. As shown in Figure 4, $\mathcal{NC}_1, \mathcal{NC}_2$, and $\mathcal{NC}_3$ converge to zero as training progresses, regardless of algorithm used. Similar to the observation for the CE loss in (Zhu et al., 2021), although all algorithms lead to $\mathcal{NC}$ solutions, networks trained with

different algorithms have notably different generalization performances.[6] We find the cosine distribution $\mathcal{P}_{CM}$ consistently aligns with the test accuracy, the more and higher. This may due to the fact that different training methods have different converge rate during the terminal phase of training, and it further lead to different distribution of features.

Improving network efficiency via fixing classifiers as simplex ETFs. In Theorem 3.1, when $d \geq K - 1$ and the weight decay terms are properly chosen, we showed that the optimal classifier for the vanilla MSE loss is a simplex ETF. This implies that we can *(i)* fix the last-layer classifier as a simplex ETF, and *(ii)* reduce the feature dimension $d = K$. By doing so, we substantially reduce the number of trainable parameters without sacrificing the generalization performance as shown in Figure 5.

Choice of the feature dimension d. On the other hand, Theorem 3.1 shows that the optimal class means $\overline{\boldsymbol{H}}^\star$ form a simplex ETF only when $d \geq K - 1$. If $d < K - 1$, then the global solution $\overline{\boldsymbol{H}}^\star$ is only the best rank-d approximation of the simplex ETF, where the class-means of the each class neither have equal length nor are maximally distant. To demonstrate its effect, we run experiments on the CIFAR10 dataset using vanilla MSE loss and ResNet18, with both $d < K - 1$ and $d \geq K - 1$. As shown in Figure 6, even though all cases exhibit $\mathcal{NC}$, choosing $d \geq K - 1$ is crucial for fitting the training data and generalization to test data. This is also corroborated by observing $\mathcal{P}_{CM}$, which shows that more training samples lie on the decision boundary (i.e., $CM_{k,i} = 0$) as d decreases in the range of $d < K - 1$. As shown in Figure 6(e), this is in sharp contrast to CE loss

---

[5]The probability margin cannot be adopted here because probability is not well-defined as softmax is not used in the MSE loss.

[6]L-BFGS with strong Wolfe line-search strategy may result in quite small stepsize at the terminal phase of training. We think that L-BFGS with proper diminishing stepsize can improve the generalization ability.

(a) $\mathcal{NC}_1$ (MSE)   (b) $\mathcal{P}_{CM}$ (MSE)   (c) Training Acc. (MSE)   (d) Test Acc. (MSE)   (e) Test Acc. (CE)
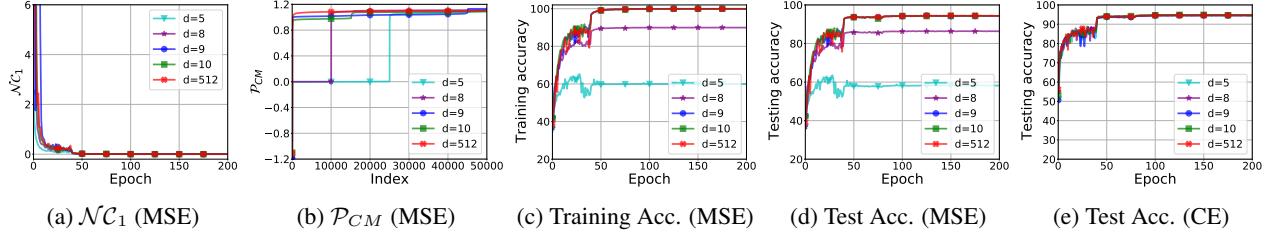
Figure 6. **Comparison of the performances on networks with different feature dimensions d for MSE and CE losses.** We compare within-class variation collapse $\mathcal{NC}_1$, cosine margin distribution $\mathcal{P}_{CM}$, training accuracy, and test accuracy on learned classifier with different feature dimension d on CIFAR10 using ResNet18 with data augmentation. The network is trained by the SGD optimizer.

(a) $\mathcal{NC}_1$ (b) $\mathcal{NC}_3$ (c) $\mathcal{P}_{CM}$ (d) $\widetilde{\text{rank}}$ (e) Train Acc.   (f) Test Acc.

(g) $\mathcal{NC}_1$ (h) $\mathcal{NC}_3$ (i) $\mathcal{P}_{CM}$ (j) $\widetilde{\text{rank}}$ (k) Train Acc.   (l) Test Acc.
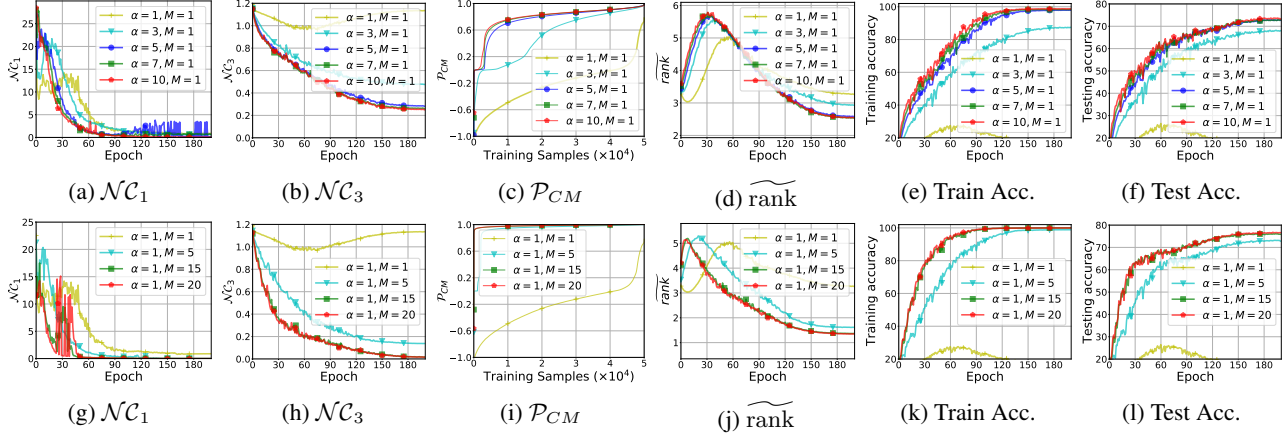
Figure 7. **Effects of rescaling parameters α and M.** Experiments are conducted on the miniImageNet dataset (MIN) with a ResNet18 backbone. Top row shows the result of varying α with fixed M. Bottom row shows the result of varying M with fixed α.

which produces similar performance for different d. Note that all the existing work on CE loss (Fang et al., 2021; Graf et al., 2021a; Papyan et al., 2020; Zhu et al., 2021) only study the case when $d \geq K$. In the Appendix, we visually compare the features learned by CE and MSE, but we leave the thorough analysis for CE loss as future work.

Experiments of the rescaled MSE loss. In Section 3.3, we argued through landscape visualization that rescaling improves the optimization landscape for the MSE loss around the global solutions. Here, we corroborate our findings via experiments, showing that rescaling of MSE indeed leads to better $\mathcal{NC}$ and hence better optimization landscapes. In particular, we empirically examine the effect of the two rescaling parameters (α, M) on the $\mathcal{NC}$ phenomenon and the generalization performance. In Figure 7, we run experiments on the miniImageNet (Vinyals et al., 2016) with ResNet18. We notice that when one scaling factor is fixed, the other parameter has a positive correlation with the degree of $\mathcal{NC}$ as well as the training and test performances. This observation is well-aligned with our analysis in Section 3.3.

5. Conclusion

In this work, we provide a global landscape analysis for deep neural networks trained via the MSE loss, under the unconstrained feature model. Our theoretical results reveal that all global solutions exhibit the $\mathcal{NC}$ phenomenon, and

that the global landscape is benign in the sense that it does not have spurious local minimizers. Such results extend the scope where $\mathcal{NC}$ provably occurs with the MSE loss, which was restricted to neural networks trained via particular and unrealistic algorithms in prior work (Han et al., 2022; Mixon et al., 2020). More broadly, our results extend the scope of the "prevalence of neural collapse" in the seminal work (Papyan et al., 2020), which was restricted to neural networks trained via the CE loss. Combined with the results in (Zhu et al., 2021), the prevalence of neural collapse now subsumes (at least) that deep neural networks trained for classification tasks with both CE and MSE losses exhibit neural collapse, regardless of the training algorithm (as long as it can escape strict saddle points) and network architecture (as long as it is sufficiently expressive).

References

Banburski, A., De La Torre, F., Pant, N., Shastri, I., and Poggio, T. Distribution of classification margins: Are all data equal? *arXiv preprint arXiv:2107.10199*, 2021.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3880–3888, 2016.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Ciliberto, C., Stamos, D., and Pontil, M. Reexamining low rank matrix factorization for trace norm regularization. *arXiv preprint arXiv:1706.08934*, 2017.

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

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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. Layer-peeled model: Toward understanding well-trained deep neural networks. *arXiv preprint arXiv:2101.12699*, 2021.

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

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242. PMLR, 2017.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

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

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Neural collapse in deep homogeneous claaifiers and the role of weight decay. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021b.

Haeffele, B. D. and Vidal, R. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

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.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

Hui, L. and Belkin, M. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hsFN92eQEla.

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.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

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

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

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Li, Q., Zhu, Z., and Tang, G. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019a.

Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z., and Zhao, T. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019b.

Lopes, M. E. Estimating unknown sparsity in compressed sensing. *arXiv preprint arXiv:1204.4227*, 2013.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2017.

Lu, J. and Steinerberger, S. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: a view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6232–6240, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Martens, J., Ballard, A., Desjardins, G., Swirszcz, G., Dalibard, V., Sohl-Dickstein, J., and Schoenholz, S. S. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. *arXiv preprint arXiv:2110.01765*, 2021.

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

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

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.

Poggio, T. and Liao, Q. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020a.

Poggio, T. and Liao, Q. Implicit dynamic regularization in deep networks. Technical report, Center for Brains, Minds and Machines (CBMM), 2020b.

Qi, H., You, C., Wang, X., Ma, Y., and Malik, J. Deep isometric learning for visual recognition. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7824–7835. PMLR, 13–18 Jul 2020.

Qu, Q., Zhai, Y., Li, X., Zhang, Y., and Zhu, Z. Geometric analysis of nonconvex optimization landscapes for overcomplete learning. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=rygixkHKDH.

Qu, Q., Zhu, Z., Li, X., Tsakiris, M. C., Wright, J., and Vidal, R. Finding the sparsest vectors in a subspace: Theory, algorithms, and applications. *arXiv preprint arXiv:2001.06970*, 2020b.

Rangamani, A., Xu, M., Banburski, A., Liao, Q., and Poggio, T. Dynamics and neural collapse in deep classifiers trained with the square loss. *Technical report, Center for Brains, Minds and Machines (CBMM)*, 2021.

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.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Shaham, U., Cloninger, A., and Coifman, R. R. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, 44(3):537–557, 2018.

Strohmer, T. and Heath Jr, R. W. Grassmannian frames with applications to coding and communication. *Applied and computational harmonic analysis*, 14(3):257–275, 2003.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18 (5):1131–1198, 2018.

Tirer, T. and Bruna, J. Extended unconstrained features model for exploring deep neural collapse. *arXiv preprint arXiv:2202.08087*, 2022.

Vinyals, O., Blundell, C., Lillicrap, T. P., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. In *NIPS*, 2016.

Weinan, E. and Wojtowytsch, S. On the emergence of tetrahedral symmetry in the final and penultimate layers of neural network classifiers. *arXiv preprint arXiv:2012.05420*, 2020.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

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

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

Notations and Organizations. For a scalar function $f(\boldsymbol{Z})$ with a variable $\boldsymbol{Z} \in \mathbb{R}^{K \times N}$, its Hessian can be represented by a bilinear form defined via $[\nabla^2 f(\boldsymbol{Z})](\boldsymbol{A}, \boldsymbol{B}) = \sum_{i,j,k,\ell} \frac{\partial^2 f(\boldsymbol{Z})}{\partial z_{ij} z_{k\ell}} a_{ij} b_{k\ell}$ for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{K \times N}$, which avoids representing the Hessian as a tensor, or vectorizing the variable $\boldsymbol{Z}$. We will use the bilinear form for the Hessian throughout the Appendix. Now we give the formal definition of Simplex ETF.

Definition .1 (K-Simplex ETF (Papyan et al., 2020; Strohmer & Heath Jr, 2003)). *A standard Simplex ETF is a collection of points in $\mathbb{R}^K$ specified by the columns of*

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

where $\boldsymbol{I}_K \in \mathbb{R}^{K \times K}$ is the identity matrix, and $\boldsymbol{1}_K \in \mathbb{R}^K$ is the all ones vector.

As in (Fang et al., 2021; Papyan et al., 2020), in this paper we consider general Simplex ETF as a collection of points in $\mathbb{R}^d$ specified by the columns of $\sqrt{\frac{K}{K-1}} \boldsymbol{P} \left(\boldsymbol{I}_K - \frac{1}{K} \boldsymbol{1}_K \boldsymbol{1}_K^\top \right)$, where (i) when $d \geq K$, $\boldsymbol{P} \in \mathbb{R}^{d \times K}$ is an orthonormal matrix, i.e., $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_K$, and (ii) when $d = K - 1$, $\boldsymbol{P}$ is chosen such that $\begin{bmatrix} \boldsymbol{P}^\top & \frac{1}{\sqrt{K}} \boldsymbol{1}_K \end{bmatrix}$ is an orthonormal matrix.

The appendix is organized as follows. In Appendix A, we describe the datasets, network architectures and training settings. In Appendix B, we provide a detailed proof for Theorem 3.1, analyzing the global minimizers to our regularized MSE loss. Finally, in Appendix C we provide additional details for obtaining the visualization of rescaled MSE and CE losses presented in Section 3.3.

A. Datasets, Network Architectures and Training, and Three Neural Collapse Measures

In Section 4, we conduct experiments on CIFAR10 (Krizhevsky et al., 2009) and miniImageNet (Vinyals et al., 2016) datasets. We note that for miniImageNet dataset, since we are not doing few-shot learning where the work (Vinyals et al., 2016) primarily considers, we split the total 60000 images into training set (50000 images) and validation set (10000 images) such that both training and validation set include the full 100 classes. All images from the datasets are normalized by their mean and variance channel-wise. We use the ResNet18 (He et al., 2016) architecture throughout all the experiments. For CIFAR10, we use the same experiment setting in (Zhu et al., 2021) except the replacement of CE loss by standard MSE loss for fair comparison. Specifically, we train ResNet18 for 200 epochs with three different optimizers: SGD, Adam and LBFGS. For SGD, the initial learning rate and momentum are set to 0.05 and 0.9, respectively. For Adam, the initial learning rate, β_1 and $beta_2$ are set to 0.001, 0.9 and 0.999, respectively. We decay the learning rate by 0.1 every 40 epochs for SGD and Adam. We use LBFGS with an initial learning rate of 0.01 and strong Wolfe line search strategy for subsequent iterations. Without explicitly mentioned, we use the weight decay of 5×10^{-4} and the same data augmentation in (Zhu et al., 2021) for all experiments on CIFAR10. For miniImageNet, we use the rescaled MSE loss as described in Section 2.2 with the SGD optimizer with an initial learning rate 0.01, momentum 0.9 and weight decay 0.001. We use a Cosine Annealing Warm Restarts (Loshchilov & Hutter, 2017) learning rate scheduler where the number of epochs before the first restart is set as 200 and the minimum learning rate is 0.0001.

Three $\mathcal{NC}$ measures $\mathcal{NC}_1$-$\mathcal{NC}_3$ (Papyan et al., 2020; Zhu et al., 2021) For the sake of completeness, we describe the three $\mathcal{NC}$ measures $\mathcal{NC}_1$-$\mathcal{NC}_3$ (Papyan et al., 2020; Zhu et al., 2021) used in Section 4. Towards that end, first define the global mean of the last-layer features $\{\boldsymbol{h}_{k,i}\}$ as $\boldsymbol{h}_G = \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \boldsymbol{h}_{k,i}$ and the class mean as $\overline{\boldsymbol{h}}_k = \frac{1}{n} \sum_{i=1}^n \boldsymbol{h}_{k,i}$ ($1 \leq k \leq K$).

- $\mathcal{NC}_1$. We measure the within-class variability collapse by

$$\mathcal{NC}_1 := \frac{1}{K} \operatorname{trace} \left(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^\dagger \right), \tag{6}$$

where $\boldsymbol{\Sigma}_W := \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \left(\boldsymbol{h}_{k,i} - \overline{\boldsymbol{h}}_k \right) \left(\boldsymbol{h}_{k,i} - \overline{\boldsymbol{h}}_k \right)^\top \in \mathbb{R}^{d \times d}$ denotes the within-class covariance of the features, $\boldsymbol{\Sigma}_B := \frac{1}{K} \sum_{k=1}^K \left(\overline{\boldsymbol{h}}_k - \boldsymbol{h}_G \right) \left(\overline{\boldsymbol{h}}_k - \boldsymbol{h}_G \right)^\top \in \mathbb{R}^{d \times d}$ represents the between-class covariance, and $\boldsymbol{\Sigma}_B^\dagger$ denotes the pseudo inverse of $\boldsymbol{\Sigma}_B$.
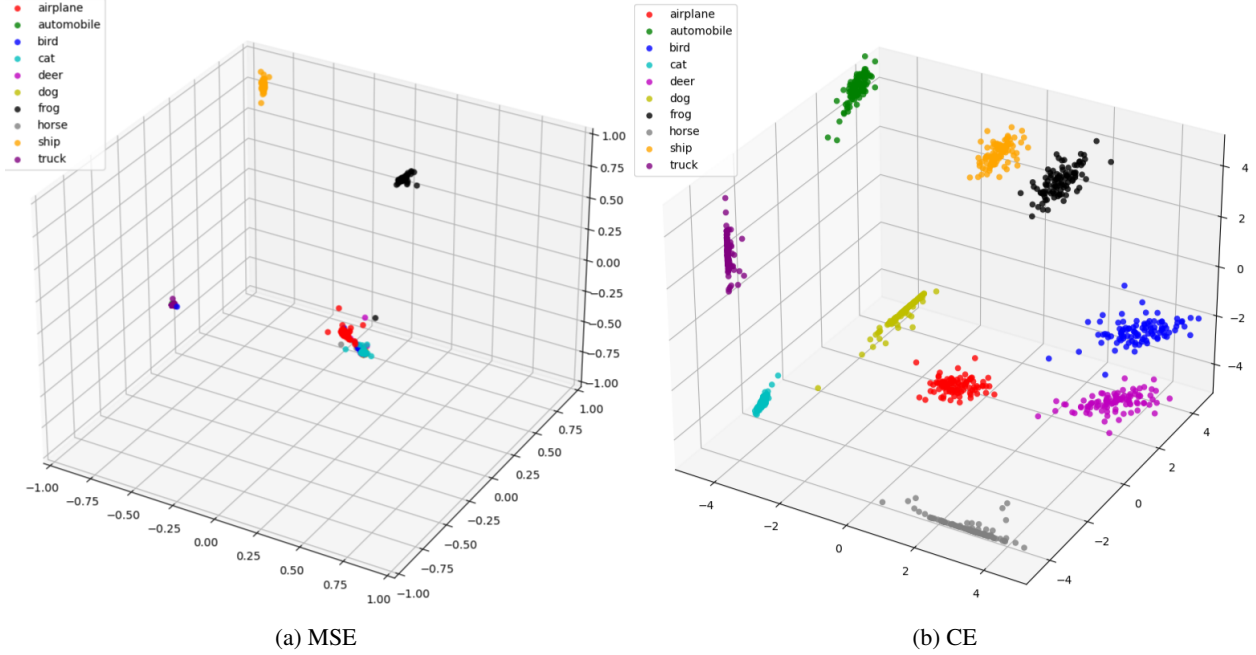
(a) MSE   (b) CE

Figure 8. **Visual comparison of features learned by MSE and CE losses with feature dimension** $d = 3$**.** We compare the training feature distribution by setting the feature dimension $d = 3$ for ResNet18 and training it with CIFAR10. The network is trained by the SGD optimizer.

- $\mathcal{NC}_2$. We measure the onvergence of the learned classifier $\boldsymbol{W} \in \mathbb{R}^{K \times d}$ (for $d \geq K - 1$) to a Simplex ETF by

$$\mathcal{NC}_2 := \left\| \frac{\boldsymbol{WW}^\top}{\|\boldsymbol{WW}^\top\|_F} - \frac{1}{\sqrt{K-1}} \left(\boldsymbol{I}_K - \frac{1}{K} \boldsymbol{1}_K \boldsymbol{1}_K^\top \right) \right\|_F, \tag{7}$$

where the Simplex ETF and $\boldsymbol{WW}^\top$ are rescaled to have unit energy (in Frobenius norm).

- $\mathcal{NC}_2$. For $d \geq K - 1$, we measure the convergence to self-duality between the learned features $\boldsymbol{H}$ and the learned classifier $\boldsymbol{W}$ via

$$\mathcal{NC}_3 := \left\| \frac{\boldsymbol{W}\overline{\boldsymbol{H}}}{\|\boldsymbol{W}\overline{\boldsymbol{H}}\|_F} - \frac{1}{\sqrt{K-1}} \left(\boldsymbol{I}_K - \frac{1}{K} \boldsymbol{1}_K \boldsymbol{1}_K^\top \right) \right\|_F, \tag{8}$$

where $\overline{\boldsymbol{H}} := \begin{bmatrix} \overline{\boldsymbol{h}}_1 - \boldsymbol{h}_G & \cdots & \overline{\boldsymbol{h}}_K - \boldsymbol{h}_G \end{bmatrix} \in \mathbb{R}^{d \times K}$ are the centered class-means.

Visual comparison of features learned by MSE and CE losses with feature dimension $d = 3$**.**   To visualize the learned features, we set the feature dimension $d = 3$ for ResNet18 and train it with CIFAR10. Figure 8 display the learned features with MSE loss and CE loss on randomly selected 100 training samples for each class. We observe that the features learned by CE loss is more diverse and discriminative than MSE loss.

B. Proof of Theorem 3.1

In this part of appendices, we prove Theorem 3.1 in Section 3 that we restate as follows.

Theorem B.1 (Global Optimality Condition). *Let* $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ *be a global minimizer of*

$$\min_{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}} \ f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) := \frac{1}{2N} \left\| \boldsymbol{WH} + \boldsymbol{b}\boldsymbol{1}^\top - \boldsymbol{Y} \right\|_F^2 + \frac{\lambda_W}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2} \|\boldsymbol{H}\|_F^2 + \frac{\lambda_b}{2} \|\boldsymbol{b}\|_2^2. \tag{9}$$

Then $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ *satisfies:*

($\mathcal{NC}1,3$) If $\lambda_W \lambda_H < \frac{1}{NK}$, then $(W^\star, H^\star)$ satisfies $\mathcal{NC}1$ and $\mathcal{NC}3$ as

$$h_{k,i}^\star = \overline{h}_k^\star, \quad \sqrt{\frac{\lambda_W}{\lambda_H n}} w^{\star k} = \overline{h}_k^\star, \quad \forall\, k \in [K],\ i \in [n].$$

Otherwise, if $\lambda_W \lambda_H \geq \frac{1}{NK}$, then $W^\star = 0$ and $H^\star = 0$.

($\mathcal{NC}2$) If $\lambda_W \lambda_H < \frac{1}{NK}$, then $\overline{H}^\star$ further obeys the following properties for different d:

1. If $d < K - 1$: we have $\overline{H}^{\star\top} \overline{H}^\star = C_1 \mathcal{P}_d(I - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top)$, where $\mathcal{P}_d(M)$ denotes the best rank-d approximating of M;

2. If $d = K - 1$: we have $\overline{H}^{\star\top} \overline{H}^\star = C_2(I - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top)$;

3. If $d \geq K$: we have

$$\overline{H}^{\star\top} \overline{H}^\star = \begin{cases} C_3 \left(I - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top\right), & \lambda_b \leq \frac{\sqrt{KN\lambda_W \lambda_H}}{1 - \sqrt{KN\lambda_W \lambda_H}}, \\ C_4 \left(I - \frac{\sqrt{n\lambda_W \lambda_H}}{\lambda_b(1 - \sqrt{KN\lambda_W \lambda_H})} \mathbf{1}_K \mathbf{1}_K^\top\right), & \text{otherwise}, \end{cases}$$

where $\frac{\sqrt{n\lambda_W \lambda_H}}{\lambda_b(1 - \sqrt{KN\lambda_W \lambda_H})} \leq \frac{1}{K}$ in the second case since $\lambda_b \geq \frac{\sqrt{KN\lambda_W \lambda_H}}{1 - \sqrt{KN\lambda_W \lambda_H}}$.

Here, C_1, C_2, C_3, and C_4 are some positive numerical constants that depend on $\lambda_W, \lambda_H, \lambda_b$.

(Bias) The bias satisfies $b^\star = b^\star \mathbf{1}_K$ with $b^\star \leq \frac{1}{K}$ given by:

1. If $d < K$: we have $b^\star = \frac{1}{K(\lambda_b + 1)}$;

2. If $d \geq K$: we have $b^\star = \begin{cases} \frac{1}{K(\lambda_b + 1)}, & \lambda_b \leq \frac{\sqrt{KN\lambda_W \lambda_H}}{1 - \sqrt{KN\lambda_W \lambda_H}}, \\ \frac{\sqrt{n\lambda_W \lambda_H}}{\lambda_b}, & \text{otherwise}. \end{cases}$

In particular, when $\lambda_b \to 0$, we have $b^\star \to \frac{1}{K}$; when $\lambda_b \to \infty$, we have $b^\star \to 0$.

B.1. Main Proof

Proof of Theorem B.1. We first characterize the solutions (W, H) in terms of b. Denote by $\widetilde{Y} = Y - b\mathbf{1}^\top$ and let $\widetilde{Y} = U\Sigma V^\top = \sum_{i=1}^K \sigma_i u_i v_i^\top$ be its SVD, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K \geq 0$ are the singular values. For convenience, we denote by $\widetilde{\lambda} = N\sqrt{\lambda_W \lambda_H}$. By Lemma B.1, we know

$$f(W, H, b) \geq \frac{\lambda_b}{2} \|b\|_2^2 + \frac{1}{N} \cdot \begin{cases} \sum_{i=1}^K \frac{1}{2}\left(\sigma_i - \left[\sigma_i - \widetilde{\lambda}\right]_+\right)^2 + \widetilde{\lambda}\left[\sigma_i - \widetilde{\lambda}\right]_+, & d \geq K \\ \sum_{i=1}^d \frac{1}{2}\left(\sigma_i - \left[\sigma_i - \widetilde{\lambda}\right]_+\right)^2 + \widetilde{\lambda}\left[\sigma_i - \widetilde{\lambda}\right]_+ + \sum_{i=d+1}^K \frac{1}{2}\sigma_i^2, & d < K \end{cases} \tag{10}$$

where the inequality becomes an equality when $WH = \sum_{i=1}^{\min(d,K)} \left[\sigma_i - \sqrt{\lambda_W \lambda_H}\right]_+ u_i v_i^\top$.

Noting that the singular values σ_i also depend on b, to minimize the right hand side (RHS) of (10) in terms of b, we first rewrite each term involving the singular value as

$$\frac{1}{2}\left(\sigma_i - \left[\sigma_i - \widetilde{\lambda}\right]_+\right)^2 + \widetilde{\lambda}\left[\sigma_i - \widetilde{\lambda}\right]_+ = \begin{cases} \frac{1}{2}\sigma_i^2, & \sigma_i \leq \widetilde{\lambda}, \\ \widetilde{\lambda}\sigma_i - \frac{1}{2}\widetilde{\lambda}^2, & \sigma_i \geq \widetilde{\lambda}, \end{cases} \tag{11}$$

where for both cases it increases as σ_i increases. Thus, for any b with the same energy, say c, minimizing the RHS of (10) is equivalent to minimizing the singular values σ_i. With this in mind, we now show that if $b^\star$ is a minimizer to RHS of (10), then $\|b^\star\| \leq \frac{1}{\sqrt{K}}$. By Lemma B.2, we know for any b we have $\sigma_2 = \sigma_3 = \cdots = \sigma_{K-1} = \sqrt{n}$ and $\sigma_1 \geq \sqrt{n}$ (see (25)). On the other hand, when $b = \frac{1}{K}\mathbf{1}$, we have $\sigma_1 = \sigma_2 = \cdots = \sigma_{K-1} = \sqrt{n}$ and $\sigma_K = 0$, which are the smallest possible singular values that can be achieved. Thus, considering the weight decay term on (10), the minimizer $b^\star$ must satisfy $\|b^\star\| \leq \|\frac{1}{K}\| = \frac{1}{\sqrt{K}}$.

Therefore, we only need to optimize over b with $\|b\| = c \leq \frac{1}{\sqrt{K}}$. It this case, it follows from Lemma B.2 that $\sigma_2 = \cdots = \sigma_{K-1} = \sqrt{n}$, $\sigma_1 \geq \sqrt{n}$, $\sigma_K \geq \sqrt{n}\left(1 - \sqrt{K}c\right)$, and both inequalities become equalities *if and only if* $b = \frac{c}{\sqrt{K}}\mathbf{1}$. The

remaining is to optimize the RHS of (10) in terms of σ_K which depends on c. By (10) and (11), this problem reduces to

$$\min_{0 \le c \le \frac{1}{\sqrt{K}}} \frac{\lambda_b}{2} c^2 + \frac{n}{2N} \left(1 - \sqrt{K}c\right)^2 \tag{12}$$

if $d < K$, and otherwise reduces to

$$\begin{cases} \min_{0 \le c \le \frac{1}{\sqrt{K}}} \frac{\lambda_b}{2} c^2 + \frac{\widetilde{\lambda}}{N} \left(\sqrt{n} \left(1 - \sqrt{K}c\right) - \frac{1}{2}\widetilde{\lambda}\right), & \sqrt{n} \left(1 - \sqrt{K}c\right) \ge \widetilde{\lambda} \\ \min_{0 \le c \le \frac{1}{\sqrt{K}}} \frac{\lambda_b}{2} c^2 + \frac{n}{2N} \left(1 - \sqrt{K}c\right)^2, & \sqrt{n} \left(1 - \sqrt{K}c\right) \le \widetilde{\lambda} \end{cases} \tag{13}$$

We now consider the two cases as follows:

1. Case I: $d < K$. In this case, the problem (12) achieves its minimum at $c^\star = \frac{1}{\sqrt{K}(\lambda_b+1)}$.

2. Case II: $d \ge K$. In this case, when $c \ge \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$, problem (13) becomes (12), and thus its minimum among $c \ge \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$ is $c^\star = \max\left(\frac{1}{\sqrt{K}(\lambda_b+1)}, \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)\right)$. On the other hand, when $c \le \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$, the problem (13) is also a quadratic function on c and achieves its minimum among $c \le \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$ is $c^\star = \min\left(\frac{\widetilde{\lambda}}{\lambda_b\sqrt{N}}, \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)\right)$.

   We now find the minimum value among these two cases. When $\frac{1}{\sqrt{K}(\lambda_b+1)} \ge \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$, i.e., $\lambda_b \le \frac{\sqrt{KN\lambda_W\lambda_H}}{1-\sqrt{KN\lambda_W\lambda_H}}$, we have $\frac{\widetilde{\lambda}}{\lambda_b\sqrt{N}} \ge \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$, which together with the form of the two quadratic functions implies that the minimum is achieved when $c^\star = \frac{1}{\sqrt{K}(\lambda_b+1)}$. On the other hand, when $\frac{1}{\sqrt{K}(\lambda_b+1)} < \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$, i.e., $\lambda_b > \frac{\sqrt{KN\lambda_W\lambda_H}}{1-\sqrt{KN\lambda_W\lambda_H}}$, we have $\frac{\widetilde{\lambda}}{\lambda_b\sqrt{N}} < \frac{1}{\sqrt{K}}\left(1 - \frac{\widetilde{\lambda}}{\sqrt{n}}\right)$, which together with the form of the two quadratic functions implies that the minimum is achieved when $c^\star = \frac{\widetilde{\lambda}}{\lambda_b\sqrt{N}} = \frac{\sqrt{N\lambda_W\lambda_H}}{\lambda_b}$. Thus, we can also conclude that $c^\star \to 0$ when $\lambda_b \to \infty$ and $c^\star \to \frac{1}{\sqrt{K}}$ when $\lambda_b \to 0$.

The proof is completed by invoking Lemma B.3 to characterize $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$. $\qquad\square$

B.2. Supporting Lemmas

We first characterize the following balance property between $\boldsymbol{W}$ and $\boldsymbol{H}$ for any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of our loss function:

Lemma B.1. *For any K, d, N, and $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^{K \times N}$ with SVD given by $\widetilde{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{i=1}^{K} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_K \ge 0$ are the singular values, the following problem*

$$\min_{\boldsymbol{W} \in \mathbb{R}^{K \times d}, \boldsymbol{H} \in \mathbb{R}^{d \times N}} \xi(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2} \left\| \boldsymbol{W}\boldsymbol{H} - \widetilde{\boldsymbol{Y}} \right\|_F^2 + \frac{\lambda_W}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2} \|\boldsymbol{H}\|_F^2 \tag{14}$$

is a strict saddle function with no spurious local minimizer, in the sense that

- *Any local minimizer $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ of (14) is a global minimizer of (14), with the following form*

$$\boldsymbol{W}^\star \boldsymbol{H}^\star = \sum_{i=1}^{\min(d,K)} \eta_i \boldsymbol{u}_i \boldsymbol{v}_i^\top,$$

  where we let $\eta_i(\lambda_W, \lambda_H) := \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+$. Correspondingly, the minimal objective value of (14) is

$$\xi_\star = \begin{cases} \sum_{i=1}^{K} \frac{1}{2}(\sigma_i - \eta_i)^2 + \sqrt{\lambda_W\lambda_H}\,\eta_i, & d \ge K \\ \sum_{i=1}^{d} \frac{1}{2}(\sigma_i - \eta_i)^2 + \sqrt{\lambda_W\lambda_H}\,\eta_i + \sum_{i=d+1}^{K} \sigma_i^2, & d < K \end{cases}. \tag{15}$$

- *Any critical point $(\boldsymbol{W}, \boldsymbol{H})$ of (14) that is not a local minimizer is a strict saddle with negative curvature, i.e. the Hessian at this critical point has at least one negative eigenvalue.*

Proof of Lemma B.1 . By definition, any critical point (W, H) of (14) satisfies the following:

$$\nabla_W \xi(W, H) = (WH - \widetilde{Y})H^\top + \lambda_W W = 0,$$
$$\nabla_H \xi(W, H) = W^\top(WH - \widetilde{Y}) + \lambda_H H = 0.$$

By left multiplying the first equation by $W^\top$ on both sides and then right multiplying second equation by $H^\top$ on both sides and combining the equations together, we obtain

$$\lambda_W W^\top W = \lambda_H H H^\top. \tag{16}$$

This further gives

$$\frac{\lambda_W}{\lambda_H} W W^\top W + \lambda_W W = \widetilde{Y} H^\top,$$
$$\frac{\lambda_H}{\lambda_W} H^\top H H^\top + \lambda_H H^\top = \widetilde{Y}^\top W. \tag{17}$$

In the following, without loss of generality, we assume that the critical point (W, H) satisfying the above equations has the form

$$W = \begin{bmatrix} \widehat{W} & \mathbf{0} \end{bmatrix}, \quad H = \begin{bmatrix} \widehat{H} \\ \mathbf{0} \end{bmatrix} \tag{18}$$

where the columns of $\widehat{W}$ are orthogonal and the rows of $\widehat{H}$ are orthogonal, and the zeros $\mathbf{0}$ in W and H might or might not exist depending on the rank of W and H. The underlying reasoning is that, for any W satisfying (17), the Gram-Schmidt process implies that we can always orthogonalize W by an orthonormal matrix $R \in \mathbb{R}^{d \times d}$ (i.e., $R^\top R = RR^\top = I$), such that $\widetilde{W} = WR = \begin{bmatrix} \widehat{W} & \mathbf{0} \end{bmatrix}$. On the other hand, let $\widetilde{H} = R^\top H$. Because $\lambda_W W^\top W = \lambda_H H H^\top$, we have $\lambda_W \widetilde{W}^\top \widetilde{W} = \lambda_H \widetilde{H} \widetilde{H}^\top$, which implies that the rows of $\widetilde{H}$ are also orthogonal. Therefore, multiply R on both sides of (17), we always have

$$\frac{\lambda_W}{\lambda_H} \widetilde{W} \widetilde{W}^\top \widetilde{W} + \lambda_W \widetilde{W} = \widetilde{Y} \widetilde{H}^\top, \quad \frac{\lambda_H}{\lambda_W} \widetilde{H}^\top \widetilde{H} \widetilde{H}^\top + \lambda_H \widetilde{H}^\top = \widetilde{Y}^\top \widetilde{W}.$$

Thus, we can verify that $(\widetilde{W}, \widetilde{H})$ is also a critical point with $\widetilde{W}\widetilde{H} = WH$ and has the same Hessian information as (W, H). Thus, without the loss of generality, we can assume orthogonal (W, H) in the form (18), but with possible zero columns.

Form of the global solutions. Based on the orthogonalization, we further decompose (17) for all $i = 1, \cdots, d$ columns of W as

$$\left(\frac{\lambda_W}{\lambda_H} \|w_i\|^2 + \lambda_W \right) w_i = \widetilde{Y} h^i,$$
$$\left(\frac{\lambda_H}{\lambda_W} \|h^i\|^2 + \lambda_H \right) h^i = \widetilde{Y}^\top w_i, \tag{19}$$

which implies that either (i) $w_i = \mathbf{0}$ and $h^i = \mathbf{0}$, or (ii) w_i, h^i are the (scaled) left and right singular vectors of $\widetilde{Y}$. In particular, when $w_i \neq \mathbf{0}$ and $h_i \neq \mathbf{0}$, then by (16), it gives

$$\|h^i\|^2 = \frac{\lambda_W}{\lambda_H} \|w_i\|^2. \tag{20}$$

By further plugging the equation above into (19), it gives

$$\left(\sqrt{\frac{\lambda_W}{\lambda_H}} \|w_i\|^2 + \sqrt{\lambda_W \lambda_H} \right) \frac{w_i}{\|w_i\|} = \widetilde{Y} \frac{h^i}{\|h^i\|},$$
$$\left(\sqrt{\frac{\lambda_W}{\lambda_H}} \|w_i\|^2 + \sqrt{\lambda_W \lambda_H} \right) \frac{h^i}{\|h^i\|} = \widetilde{Y}^\top \frac{w_i}{\|w_i\|}. \tag{21}$$

Thus, when $\boldsymbol{w}_i \neq \boldsymbol{0}$ and $\boldsymbol{h}_i \neq \boldsymbol{0}$, we conclude that $\sqrt{\frac{\lambda_W}{\lambda_H}}\|\boldsymbol{w}_i\|^2 + \sqrt{\lambda_W\lambda_H}$ is a singular value of $\widetilde{\boldsymbol{Y}}$, say σ_{i_j}, and $\frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$ and $\frac{\boldsymbol{h}^i}{\|\boldsymbol{h}^i\|}$ are the corresponding left and right singular vectors, respectively. In other words, when $\boldsymbol{w}_i \neq \boldsymbol{0}$ and $\boldsymbol{h}_i \neq \boldsymbol{0}$, then

$$\sigma_{i_j} = \sqrt{\frac{\lambda_W}{\lambda_H}}\|\boldsymbol{w}_i\|^2 + \sqrt{\lambda_W\lambda_H}, \quad \boldsymbol{u}_{i_j} = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}, \quad \boldsymbol{v}_{i_j} = \frac{\boldsymbol{h}^i}{\|\boldsymbol{h}^i\|} \tag{22}$$

for some i_j such that $\sigma_{i_j} > \sqrt{\lambda_W\lambda_H}$. Together with (20), it further implies that

$$\boldsymbol{w}_i\boldsymbol{h}^{i\top} = \|\boldsymbol{w}_i\|_2^2 \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|_2}\frac{\boldsymbol{h}^{i\top}}{\|\boldsymbol{w}_i\|_2} = \sqrt{\frac{\lambda_W}{\lambda_H}}\|\boldsymbol{w}_i\|_2^2 \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|_2}\frac{\boldsymbol{h}^{i\top}}{\|\boldsymbol{h}^i\|_2} = \left(\sigma_{i_j} - \sqrt{\lambda_W\lambda_H}\right)\boldsymbol{u}_{i_j}\boldsymbol{v}_{i_j}^\top.$$

Next, we discuss global minimizers and global function values in two cases: (*i*) $d \geq K$, and (*ii*) $d < K$. For both cases, based on the above results, we can write

$$\boldsymbol{W}\boldsymbol{H}^\top = \sum_{i=1}^d \boldsymbol{w}_i\boldsymbol{h}^{i\top} = \sum_{\boldsymbol{w}_i\neq\boldsymbol{0},\boldsymbol{h}^i\neq\boldsymbol{0}}\left(\sigma_{i_j} - \sqrt{\lambda_W\lambda_H}\right)\boldsymbol{u}_{i_j}\boldsymbol{v}_{i_j}^\top + \sum_{\boldsymbol{w}_i=\boldsymbol{0} \text{ and } \boldsymbol{h}^i=\boldsymbol{0}}\boldsymbol{w}_i\boldsymbol{h}^{i\top}$$

$$= \sum_{\boldsymbol{w}_i\neq\boldsymbol{0},\boldsymbol{h}^i\neq\boldsymbol{0}}\left(\sigma_{i_j} - \sqrt{\lambda_W\lambda_H}\right)\boldsymbol{u}_{i_j}\boldsymbol{v}_{i_j}^\top.$$

Case I: $d \geq K$. In this case, given the rank of $\boldsymbol{W}$ is at most K, we know that the minimum is achieved when

$$\boldsymbol{W}^\star\boldsymbol{H}^\star = \sum_{i=1}^K \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ \boldsymbol{u}_i\boldsymbol{v}_i^\top$$

with $\sigma_i \geq \sqrt{\lambda_W\lambda_H}$ for all $i = 1, \cdots, K$. In this case, we have

$$\xi_\star = \frac{1}{2}\sum_{i=1}^K \left(\left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ - \sigma_i\right)^2 + \frac{\lambda_W}{2}\sum_{i=1}^d \|\boldsymbol{w}_i\|_2^2 + \frac{\lambda_H}{2}\sum_{i=1}^d \|\boldsymbol{h}^i\|_2^2$$

$$= \frac{1}{2}\sum_{i=1}^K \left(\left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ - \sigma_i\right)^2 + \lambda_W\sum_{i=1}^d \|\boldsymbol{w}_i\|_2^2$$

$$= \frac{1}{2}\sum_{i=1}^K \left(\left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ - \sigma_i\right)^2 + \sqrt{\lambda_H\lambda_W}\sum_{i=1}^K \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+,$$

where for the second and third equality, we used (20) and (22), respectively.

Case II: $d < K$. In this case, we know that the minimum is achieved when

$$\boldsymbol{W}^\star\boldsymbol{H}^\star = \sum_{i=1}^d \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ \boldsymbol{u}_i\boldsymbol{v}_i^\top$$

with $\sigma_i \geq \sqrt{\lambda_W\lambda_H}$ for all $i = 1, \cdots, d$. Similarly, we have

$$\xi_\star = \frac{1}{2}\sum_{i=1}^d \left(\left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ - \sigma_i\right)^2 + \sqrt{\lambda_H\lambda_W}\sum_{i=1}^d \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ + \sum_{i=d+1}^K \sigma_i^2,$$

where the extra term $\sum_{i=d+1}^K \sigma_i^2$ is coming from the singular values of $\widehat{\boldsymbol{Y}}$ and the decomposition of $\frac{1}{2}\left\|\boldsymbol{W}\boldsymbol{H} - \widetilde{\boldsymbol{Y}}\right\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2$.

In summary, the minimum function value is obtained when

$$\boldsymbol{W}^\star\boldsymbol{H}^\star = \sum_{i=1}^{\min\{d,K\}} \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+ \boldsymbol{u}_i\boldsymbol{v}_i^\top = \sum_{i=1}^{\min\{d,K\}} \eta_i\boldsymbol{u}_i\boldsymbol{v}_i^\top, \tag{23}$$

with $\eta_i(\lambda_W, \lambda_H) := \left[\sigma_i - \sqrt{\lambda_W\lambda_H}\right]_+$, and the minimum function value is attained as in (15).

Showing negative curvature for strict saddles. In the remaining part, we show those critical point $(\boldsymbol{W}, \boldsymbol{H})$ that does not satisfy the condition in (23) are strict saddle points, by showing that the Hessian of (14) has negative eigenvalues. First, we

derive the directional Hessian of (14), which has the following form

$$2[\nabla^2 \xi(\boldsymbol{W}, \boldsymbol{H})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) \;=\; \|\boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{H} + \boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{H}}\|_F^2 + 2\left\langle \boldsymbol{W}\boldsymbol{H} - \widetilde{\boldsymbol{Y}}, \boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{\Delta}_{\boldsymbol{H}} \right\rangle$$
$$+ \lambda_{\boldsymbol{W}} \|\boldsymbol{\Delta}_{\boldsymbol{W}}\|_F^2 + \lambda_{\boldsymbol{H}} \|\boldsymbol{\Delta}_{\boldsymbol{H}}\|_F^2 \,. \tag{24}$$

Given that a critical point $(\boldsymbol{W}, \boldsymbol{H})$ is not a global minimizer, then (23) is not satisfied. This implies that there must exist a singular value of $\widehat{\boldsymbol{Y}}$ with $\sigma_j > \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}$, which cannot be not covered by any $(\boldsymbol{w}_i, \boldsymbol{h}_i)$ in the sense that $\boldsymbol{w}_j \boldsymbol{h}^{j\top} \neq (\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}})\boldsymbol{u}_j \boldsymbol{v}_j^\top$ for some j. We now discuss this situation separately in two cases: (i) $d \geq K$, and (ii) $d < K$.

Case I: $d \geq K$. In this case, since each column of $\boldsymbol{W}$ is either zero or corresponds to the left singular vectors of $\widetilde{\boldsymbol{Y}}$, it implies that the column space of $\boldsymbol{W}$ has a non-trivial null space, i.e., there must exist a unit vector $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\boldsymbol{W}\boldsymbol{\alpha} = \boldsymbol{0}$. Since $\lambda_{\boldsymbol{W}} \boldsymbol{W}^\top \boldsymbol{W} = \lambda_{\boldsymbol{H}} \boldsymbol{H}\boldsymbol{H}^\top$, we also have $\boldsymbol{\alpha}^\top \boldsymbol{H} = \boldsymbol{0}$. With this property, for the index j with $\boldsymbol{w}_j \boldsymbol{h}^{j\top} \neq (\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}})\boldsymbol{u}_j \boldsymbol{v}_j^\top$, we construct $\boldsymbol{\Delta}_{\boldsymbol{W}} = \left(\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}\right)^{1/4} \boldsymbol{u}_j \boldsymbol{\alpha}^\top, \boldsymbol{\Delta}_{\boldsymbol{H}} = \left(\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}\right)^{1/4} \boldsymbol{\alpha} \boldsymbol{v}_j^\top$. Given that $\boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{H} = \boldsymbol{0}$ and $\boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{H}} = \boldsymbol{0}$

$$\|\boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{H} + \boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{H}}\|_F^2 \;=\; 0,$$
$$\left\langle \boldsymbol{W}\boldsymbol{H} - \widetilde{\boldsymbol{Y}}, \boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{\Delta}_{\boldsymbol{H}} \right\rangle \;=\; -\sigma_j,$$
$$\lambda_{\boldsymbol{W}} \|\boldsymbol{\Delta}_{\boldsymbol{W}}\|_F^2 + \lambda_{\boldsymbol{H}} \|\boldsymbol{\Delta}_{\boldsymbol{H}}\|_F^2 \;=\; 2\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}.$$

Plugging this into the Hessian (24), it gives

$$2[\nabla^2 \xi(\boldsymbol{W}, \boldsymbol{H})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) \;=\; -2\sigma_j + 2\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}} \;=\; -2(\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}) \;<\; 0.$$

This implies that there exists a negative curvature for the Hessian, and the saddle point must be strict saddle.

Case II: $d < K$. Recall from (18) and (22) that $\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}} \boldsymbol{W}^\top \boldsymbol{W}$ is a diagonal matrix with the values of diagonal entry from $\left\{ [\sigma_1 - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}]_+, \ldots, [\sigma_K - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}]_+, 0 \right\}$, but here it excludes $[\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}]_+$ which equals $\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}$ by our assumption. Thus, $\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}} \boldsymbol{W}^\top \boldsymbol{W}$ has at least one diagonal entry which is strictly smaller than $\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}$. Now let $\boldsymbol{\alpha} \in \mathbb{R}^d$ be the eigenvector associated with the smallest eigenvalue of $\boldsymbol{W}^\top \boldsymbol{W}$, so that

$$\nu := \sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}} \boldsymbol{\alpha}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\alpha} < \sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}.$$

Since $\lambda_{\boldsymbol{W}} \boldsymbol{W}^\top \boldsymbol{W} = \lambda_{\boldsymbol{H}} \boldsymbol{H}\boldsymbol{H}^\top$, we also have $\sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}} \boldsymbol{\alpha}^\top \boldsymbol{H}^\top \boldsymbol{H} \boldsymbol{\alpha} = \nu$. With this property, we construct $\boldsymbol{\Delta}_{\boldsymbol{W}} = \left(\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}\right)^{1/4} \boldsymbol{u}_j \boldsymbol{\alpha}^\top, \boldsymbol{\Delta}_{\boldsymbol{H}} = \left(\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}\right)^{1/4} \boldsymbol{\alpha} \boldsymbol{v}_j^\top$, which satisfies

$$\|\boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{H} + \boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{H}}\|_F^2 = \sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}} \boldsymbol{\alpha}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\alpha} + \sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}} \boldsymbol{\alpha}^\top \boldsymbol{H}^\top \boldsymbol{H} \boldsymbol{\alpha} = 2\nu,$$
$$\left\langle \boldsymbol{W}\boldsymbol{H} - \widetilde{\boldsymbol{Y}}, \boldsymbol{\Delta}_{\boldsymbol{W}}\boldsymbol{\Delta}_{\boldsymbol{H}} \right\rangle = -\sigma_j,$$
$$\lambda_{\boldsymbol{W}} \|\boldsymbol{\Delta}_{\boldsymbol{W}}\|_F^2 + \lambda_{\boldsymbol{H}} \|\boldsymbol{\Delta}_{\boldsymbol{H}}\|_F^2 = 2\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}.$$

Plugging this into the Hessian quadratic form gives

$$2[\nabla^2 \xi(\boldsymbol{W}, \boldsymbol{H})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) = 2\nu - 2\sigma_j + 2\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}} = -2(\sigma_j - \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}} - \nu) < 0.$$

Therefore, we prove $(\boldsymbol{W}, \boldsymbol{H})$ is a strict saddle for both cases. This completes the proof. $\qquad\square$

Lemma B.2. *Assume the number of training samples in each class is balanced, i.e., $n = n_1 = \cdots = n_K$, and let $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \cdots & \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_K \end{bmatrix} \in \mathbb{R}^{K \times nK}$ be the matrix that contains the one-hot vectors for all the training samples. Then $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{b}\boldsymbol{1}^\top$ has at least $K - 2$ singular values being $\sqrt{n}$. The rest of the two singular values, without loss of generality, denoted by σ_1 and σ_K, depend on $\boldsymbol{b}$. Then, we have the following lower bounds for σ_1 and σ_K.*

1. *For any b, the largest singular value σ_1 can be lower bounded by*

$$\sigma_1 \geq \sqrt{n} \max\left(\sqrt{1 + K\left(\|b\|^2 - \frac{1}{K}\left(\mathbf{1}^\top b\right)^2 \right)}, |1 - \mathbf{1}^\top b| \right). \tag{25}$$

2. *For any b on the sphere $\{b \in \mathbb{R}^K : \|b\| = c\}$ with $c \leq \frac{1}{\sqrt{K}}$, we have*

$$\sigma_1 \geq \sqrt{n}, \quad \sigma_K \geq \sqrt{n}\left(1 - \sqrt{K}c\right) \tag{26}$$

and both inequalities become equalities if and only if $b = \frac{c}{\sqrt{K}}\mathbf{1}$.

Proof of Lemma B.2. To study the singular values of Y, it is equivalent to look at the eigenvalues of the Gram matrix of $\widehat{Y}^\top$:

$$G = \widetilde{Y}\widetilde{Y}^\top = \left(Y - b\mathbf{1}^\top\right)\left(Y - b\mathbf{1}^\top\right)^\top = n\left(I - b\mathbf{1}^\top - \mathbf{1}b^\top + Kbb^\top\right).$$

If b is aligned with $\mathbf{1}$, i.e., they live in the same line, then $-b\mathbf{1}^\top - \mathbf{1}b^\top + Kbb^\top$ is a rank-1 matrix and G has $K - 1$ eignevalues being n and the rest eigenvalue being $n\left(1 - \mathbf{1}^\top b\right)^2$. On the other hand, if b is not aligned with $\mathbf{1}$, then $-b\mathbf{1}^\top - \mathbf{1}b^\top + Kbb^\top$ is a rank-2 matrix and G has $K - 2$ eigenvalues being n. In this case, the rest of the two eigenvalues, denoted by π_1 and π_K, correspond to the eigenvectors within the subspace spanned by $\mathbf{1}$ and b.

To estimate the largest eigenvalues π_1, we construct two orthonormal vectors within this subspace spanned by $\mathbf{1}$ and b and compute the corresponding Rayleigh quotient. Specifically, we first compute the Rayleigh quotient along the direction $\mathbf{1}$ as

$$\frac{\mathbf{1}^\top G\mathbf{1}}{\mathbf{1}^\top\mathbf{1}} = \frac{n}{K}\left(K - 2K\mathbf{1}^\top b + K\left(\mathbf{1}^\top b\right)^2\right) = n\left(1 - 2\mathbf{1}^\top b + \left(\mathbf{1}^\top b\right)^2\right) = n\left(1 - \mathbf{1}^\top b\right)^2.$$

Use Gram-Schmidt orthonormalization to obtain the other direction as $a = b - \frac{1}{K}\mathbf{1}^\top b\mathbf{1}$, which gives the following Rayleigh quotient:

$$\frac{a^\top Ga}{a^\top a} = \frac{n}{\|a\|^2}\left(\|a\|^2 + K\left(\|b\|^2 - \frac{1}{K}\left(\mathbf{1}^\top b\right)^2\right)^2\right) = n + nK\left(\|b\|^2 - \frac{1}{K}\left(\mathbf{1}^\top b\right)^2\right),$$

where the last equality follows because $\|a\|^2 = \|b\|^2 - \frac{1}{K}\left(\mathbf{1}^\top b\right)^2$. Thus, by the min-max theorem (i.e., Courant–Fischer–Weyl min-max principle), we have

$$\pi_1 \geq \max\left(\frac{\mathbf{1}^\top G\mathbf{1}}{\mathbf{1}^\top\mathbf{1}}, \frac{a^\top Ga}{a^\top a}\right) \geq \max\left(n\left(1 - \mathbf{1}^\top b\right)^2, n + nK\left(\|b\|^2 - \frac{1}{K}\left(\mathbf{1}^\top b\right)^2\right)\right) \geq n,$$

where the last inequality becomes an inequality *if and only if* b is a scaled version of the vector $\mathbf{1}$, i.e., $b = \frac{\|b\|}{\sqrt{K}}\mathbf{1}$.

To obtain a lower bound for π_K whenever $\|b\| \leq \frac{1}{\sqrt{K}}$, we again use the the min-max theorem as

$$\frac{1}{n}\pi_K \geq \min_{\|u\|=1}\frac{1}{n}u^\top Gu = \min_{\|u\|=1} 1 - 2u^\top b\mathbf{1}^\top u + K(u^\top b)^2$$

$$\geq \min_{\|u\|=1} 1 - 2\sqrt{K}\left|u^\top b\right| + K(u^\top b)^2$$

$$\geq 1 - 2\sqrt{K}\|b\| + K\|b\|^2 = \left(1 - \sqrt{K}\|b\|\right)^2,$$

where the first inequality achieves equality when u is restricted to the subspace spanned by $\mathbf{1}$ and b, the second inequality becomes an equality only when $u = 1/\sqrt{K}$ and $u^\top b \geq 0$ or $u = -1/\sqrt{K}$ and $u^\top b \leq 0$, and the last inequality achieves equality if and only if u is aligned with b, i.e., $\left|u^\top b\right| = \|b\|$. Thus, for any b on the sphere $\{b \in \mathbb{R}^K : \|b\| = c\}$ with $c \leq \frac{1}{\sqrt{K}}$, π_K achieves its minimum possible value $n\left(1 - \sqrt{K}\|b\|\right)^2$ if and only if $b = \pm\frac{c}{\sqrt{K}}\mathbf{1}$. This completes the proof.

$\square$

Lemma B.3. *Assume the number of training samples in each class is balanced, i.e., $n = n_1 = \cdots = n_K$, and let $Y = \begin{bmatrix} y_1 & \cdots & y_1 & y_2 & \cdots & y_K \end{bmatrix} \in \mathbb{R}^{K \times nK}$ be the matrix that contains the one-hot vectors for all the training*

samples. Suppose $b^\star \leq \frac{1}{K}$. Then any global minimizer $(W^\star, H^\star)$ of

$$\min_{W,H} \frac{1}{2N} \|WH + b^\star \mathbf{1}\mathbf{1}^\top - Y\|_F^2 + \frac{\lambda_W}{2} \|W\|_F^2 + \frac{\lambda_H}{2} \|H\|_F^2. \tag{27}$$

satisfies the self-duality

$$h_{k,i}^\star = \sqrt{\frac{\lambda_W}{\lambda_H n}} w^{\star k}, \quad \forall\, k \in [K],\ i \in [n].$$

Moreover, if $\lambda_W \lambda_H \geq \frac{1}{NK}$, then $W^\star = 0$ and $H^\star = 0$. On the other hand, if $\lambda_W \lambda_H < \frac{1}{NK}$, $(W^\star, H^\star)$ further obeys the following properties for different d:

1. $d < K - 1$: $W^\star W^{\star\top} \sim \mathcal{P}_d(I - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)$ where $\mathcal{P}_d$ denotes the best rank-d approximating and $A \sim B$ means that there is a constant c such that $A = cB$;

2. $d = K - 1$: In this case, $W^\star W^{\star\top} \sim I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$;

3. $d \geq K$ and $b^\star \geq \frac{1}{K} - \sqrt{n\lambda_W\lambda_H}$: $W^\star W^{\star\top} \sim I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$;

4. $d \geq K$ and $b^\star < \frac{1}{K} - \sqrt{n\lambda_W\lambda_H}$: $W^\star W^{\star\top} \sim I - \frac{b^\star}{1 - K\sqrt{n\lambda_W\lambda_H}}\mathbf{1}_K\mathbf{1}_K^\top$;

Proof of Lemma B.3. For convenience, let $\mathbf{1}_{K \times L}$ represents an all-ones matrix of size $K \times L$. Since $Y - b^\star \mathbf{1}_{K \times nK}$ contains many repeated columns, we first consider $\overline{Y} = I_K - b^\star \mathbf{1}_{K \times K}$ that contains the non-repeated columns of $Y - b^\star \mathbf{1}_{K \times nK}$. Let $\overline{Y} = U\overline{\Sigma}U^\top$ be the eigenvalue decomposition, where $U \in \mathbb{R}^{K \times K}$ is an orthonormal matrix and $\overline{\Sigma} \in \mathbb{R}^{K \times K}$ is a diagonal matrix with eigenvalues $\overline{\sigma}_1 \geq \cdots \overline{\sigma}_K$ along the diagonals. Since $b^\star \leq \frac{1}{K}$, the eigenvalues are given by $\overline{\sigma}_1 = \cdots = \overline{\sigma}_{K-1} = 1 \geq \overline{\sigma}_K = 1 - b^\star K$, and the eigenvector corresponding to $\overline{\sigma}_K$ is $u_K = \frac{1}{\sqrt{K}}\mathbf{1}$, which implies that $[U]_{K-1}[U]_{K-1}^\top = I - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$, where $[U]_r$ means a $K \times r$ submtraix of U by taking the first r columns.

Let $\Sigma = \sqrt{n}\overline{\Sigma}$ and $V^\top = \frac{1}{\sqrt{n}}\begin{bmatrix} u^1 & \cdots & u^1 & u^2 & \cdots & u^K \end{bmatrix} \in \mathbb{R}^{K \times nK}$ that repeats the rescaled version of the column of U n times so that $V^\top V = U^\top U = I$. By noting the relation between $Y - b^\star \mathbf{1}_{K \times nK}$ and $\overline{Y}$, we know $U\Sigma V^\top$ is the SVD of $Y - b^\star \mathbf{1}_{K \times nK}$. When $\lambda_W \lambda_H \geq \frac{1}{NK}$, by applying Lemma B.1 and Lemma B.2, we conclude that $W^\star = 0$ and $H^\star = 0$ since $\sqrt{n} - N\sqrt{\lambda_W\lambda_H} \leq 0$. We now assume $\lambda_W \lambda_H < \frac{1}{NK}$ and utilize Lemma B.1 and Lemma B.2 again for the following cases:

1. $d < K - 1$: In this case, we have

$$W^\star = \sqrt{\frac{\lambda_H}{\lambda_W}}\left(\sqrt{n} - N\sqrt{\lambda_W\lambda_H}\right)^{1/2} U(:, 1:d)R,$$

$$H^\star = \sqrt{\frac{\lambda_W}{\lambda_H}}\left(\sqrt{n} - N\sqrt{\lambda_W\lambda_H}\right)^{1/2} R^\top V(:, 1:d)^\top, \forall R \in \mathbb{R}^{d \times d}, R^\top R = I.$$

   Thus, $h_{k,i}^\star = \sqrt{\frac{\lambda_W}{\lambda_H n}}w^{\star k}$ and $W^\star W^{\star\top} \sim U(:, 1:d)U(:, 1:d)^\top = \mathcal{P}_d(I - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)$.

2. $d = K - 1$: In this case, we have

$$W^\star = \sqrt{\frac{\lambda_H}{\lambda_W}}\left(\sqrt{n} - N\sqrt{\lambda_W\lambda_H}\right)^{1/2} U(:, 1:K-1)R,$$

$$H^\star = \sqrt{\frac{\lambda_W}{\lambda_H}}\left(\sqrt{n} - N\sqrt{\lambda_W\lambda_H}\right)^{1/2} R^\top V(:, 1:K-1)^\top, \forall R \in \mathbb{R}^{(K-1)\times(K-1)}, R^\top R = I.$$

   Thus, $h_{k,i}^\star = \sqrt{\frac{\lambda_W}{\lambda_H n}}w^{\star k}$ and $W^\star W^{\star\top} \sim [U]_{K-1}[U]_{K-1}^\top = I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$.

3. $d = K$: In this case, we have

$$W^\star = \sqrt{\frac{\lambda_H}{\lambda_W}}U\left[\Sigma - N\sqrt{\lambda_W\lambda_H}\right]_+^{1/2} R,$$

$$H^\star = \sqrt{\frac{\lambda_W}{\lambda_H}}R^\top \left[\Sigma - N\sqrt{\lambda_W\lambda_H}\right]_+^{1/2} V^\top, \forall R \in \mathbb{R}^{K \times K}, R^\top R = I$$

Thus, $\boldsymbol{h}_{k,i}^{\star} = \sqrt{\frac{\lambda_W}{\lambda_H n}} \boldsymbol{w}^{\star k}$. Moreover, if $N\sqrt{\lambda_W \lambda_H} \geq \sqrt{n}(1 - b^\star K)$, i.e., $b^\star \geq \frac{1}{K} - \sqrt{n\lambda_W \lambda_H}$, then $\boldsymbol{W}^\star \boldsymbol{W}^{\star\top} \sim$
$[\boldsymbol{U}]_{K-1}[\boldsymbol{U}]_{K-1}^\top = \boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top$. On the other hand, if $b^\star < \frac{1}{K} - \sqrt{n\lambda_W \lambda_H}$, then

$$\boldsymbol{W}^\star \boldsymbol{W}^{\star\top} \sim \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top - K\sqrt{n\lambda_W \lambda_H}\boldsymbol{U}\boldsymbol{U}^\top = \overline{\sigma}_K = \boldsymbol{I} - b^\star\boldsymbol{1}_K\boldsymbol{1}_K^\top - N\sqrt{\lambda_W \lambda_H}\boldsymbol{I}$$

$$= (1 - K\sqrt{n\lambda_W \lambda_H})\boldsymbol{I} - b^\star\boldsymbol{1}_K\boldsymbol{1}_K^\top \sim \boldsymbol{I} - \frac{b^\star}{1 - K\sqrt{n\lambda_W \lambda_H}}\boldsymbol{1}_K\boldsymbol{1}_K^\top.$$

4. $d > K$: In this case, we have

$$\boldsymbol{W}^\star = \sqrt{\frac{\lambda_H}{\lambda_W}}\left[\boldsymbol{U}\left[\boldsymbol{\Sigma} - N\sqrt{\lambda_W \lambda_H}\right]_+^{1/2} \quad \boldsymbol{0}\right]\boldsymbol{R},$$

$$\boldsymbol{H}^\star = \sqrt{\frac{\lambda_W}{\lambda_H}}\boldsymbol{R}^\top\left[\begin{matrix}\left[\boldsymbol{\Sigma} - N\sqrt{\lambda_W \lambda_H}\right]_+^{1/2}\boldsymbol{V}^\top \\ \boldsymbol{0}\end{matrix}\right], \forall \boldsymbol{R} \in \mathbb{R}^{d\times d}, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}.$$

One can verify that $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ satisfies the same properties as in the case of $d = K$.

$\square$

C. Visualization of Optimization Landscape

C.1. Details of the Visualization Technique

We provide the technical details on how the visualization in Section 3.3 is obtained.

The following result expresses the output of the classifier layer for a feature vector $\boldsymbol{h}$ as a function of the norm of $\boldsymbol{h}$ and its angle to a classifier weight vector $\boldsymbol{w}^k$.

Proposition C.1. *Given any $d \geq K - 1 > 1$, take the classifier weights $\boldsymbol{W}, \boldsymbol{b}$ to be such that $\boldsymbol{W}$ is an arbitrary K-Simplex ETF (see Definition .1) and $\boldsymbol{b} = \boldsymbol{0}$. Take any $k, k' \in \{1, \ldots, K\}$, and consider a vector $\boldsymbol{h}$ on the two-dimensional plane $\text{span}\{\boldsymbol{w}^k, \boldsymbol{w}^{k'}\}$ parameterized in the polar coordinate system with polar axis being $\boldsymbol{w}^k$. Denote s and θ the radial and angular (in radians) coordinates of $\boldsymbol{h}$, respectively (positive angular direction of the polar coordinate system is taken so that $\boldsymbol{w}^{k'}$'s angular coordinate is in $(0, \pi)$). We have*

- *The feature $\boldsymbol{h}$ can be expressed as a linear combination of $\boldsymbol{w}^k$ and $\boldsymbol{w}^{k'}$:*

$$\boldsymbol{h} = s\left(\frac{\sin\theta}{\sqrt{K^2 - 2K}} + \cos\theta\right)\boldsymbol{w}^k + s(K-1)\frac{\sin\theta}{\sqrt{K^2 - 2K}}\boldsymbol{w}^{k'}; \tag{28}$$

- *The output of the classifier layer $(\boldsymbol{W}, \boldsymbol{b})$ is given by*

$$\langle \boldsymbol{w}^{k''}, \boldsymbol{h}\rangle + b_{k''} = \begin{cases} s\cos\theta, & \text{if } k'' = k; \\ s\frac{\sqrt{K^2-2K}}{K-1}\sin\theta - \frac{s}{K-1}\cos\theta, & \text{if } k'' = k'; \\ -s\sqrt{\frac{K}{K-2}}\frac{1}{K-1}\sin\theta - \frac{s}{K-1}\cos\theta, & \text{otherwise.} \end{cases} \tag{29}$$

Note that (29) is invariant to the arbitrary rotation in K-Simplex ETF.

We omit the proof to Proposition C.1 as it can be obtained via simple algebra.

Based on Proposition C.1, we can obtain the (rescaled) MSE and CE losses as a function of (s, θ). Assuming that $\boldsymbol{h}$ belongs to class k, the rescaled MSE loss defined in (4) w.r.t. $\boldsymbol{h}$ is given by

$$\text{Loss}_{\text{MSE}}(\boldsymbol{h}; \alpha, M) = \frac{\alpha}{2}\left(\langle \boldsymbol{w}^k, \boldsymbol{h}\rangle + b_k - M\right)^2 + \frac{1}{2}\sum_{k''\neq k}\left(\langle \boldsymbol{w}^{k''}, \boldsymbol{h}\rangle + b_{k''} - 1\right)^2, \tag{30}$$

where α, M are rescaling parameters. Plugging in the results in (29), we obtain

$$\text{Loss}_{\text{MSE}}(s, \theta; \alpha, M) = \frac{\alpha}{2} \cdot (s \cos\theta - M)^2 + \frac{s^2}{2} \cdot \left(\frac{\sqrt{K^2 - 2K} \sin\theta - \cos\theta}{K - 1} \right)^2$$

$$+ \frac{s^2}{2} \cdot (K - 2) \cdot \left(\frac{\sqrt{\frac{K}{K-2}} \sin\theta + \cos\theta}{K - 1} \right)^2. \quad (31)$$

Similarly, we may obtain the CE loss as

$$\text{Loss}_{\text{CE}}(s, \theta) = -\log \left(\frac{e^{s \cos\theta}}{e^{s \cos\theta} + e^{s \frac{\sqrt{K^2 - 2K} \sin\theta - \cos\theta}{K - 1}} + (K - 2) e^{-s \frac{\sqrt{\frac{K}{K-2}} \sin\theta + \cos\theta}{K - 1}}} \right). \quad (32)$$

Figure 2 is obtained by plotting the loss functions in (31) and (32).

C.2. Visualization of the Gradient Vector Field

We consider the regime of $K \to \infty$ in which the rescaled MSE loss (31) becomes

$$\lim_{K \to \infty} \text{Loss}_{\text{MSE}}(s, \theta; \alpha, M) = \frac{\alpha}{2} (s \cos\theta - M)^2 + \frac{1}{2} s^2 \sin^2\theta. \quad (33)$$
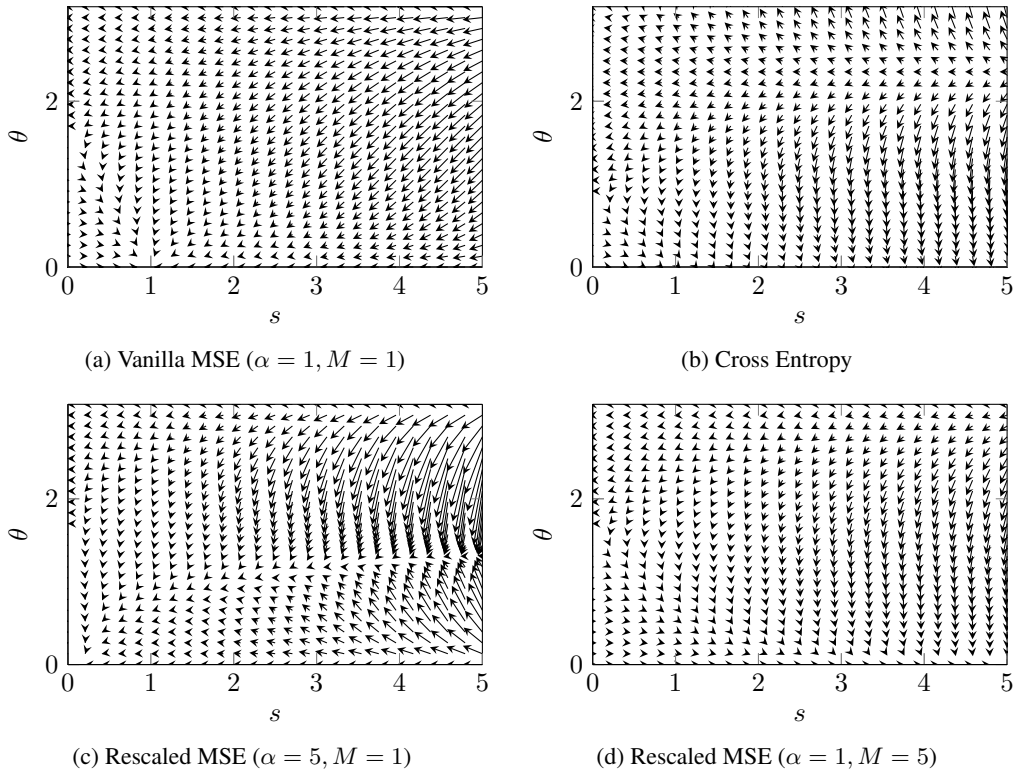
Taking the derivative w.r.t. s and θ, we obtain

$$\frac{\partial}{\partial s} \lim_{K \to \infty} \text{Loss}_{\text{MSE}}(s, \theta; \alpha, M) = s + (\alpha - 1)s \cos^2\theta - \alpha M \cos\theta,$$

$$\frac{\partial}{\partial \theta} \lim_{K \to \infty} \text{Loss}_{\text{MSE}}(s, \theta; \alpha, M) = \alpha M s \sin\theta - (\alpha - 1)s^2 \sin\theta \cos\theta. \quad (34)$$

Similarly, we may obtain the gradient for CE as

$$\frac{\partial}{\partial s} \lim_{K \to \infty} \text{Loss}_{\text{CE}}(s, \theta; \alpha, M) = \frac{e^{\sin\theta}(\sin\theta - \cos\theta)}{e^{\sin\theta} + e^{\cos\theta}},$$

$$\frac{\partial}{\partial \theta} \lim_{K \to \infty} \text{Loss}_{\text{CE}}(s, \theta; \alpha, M) = \frac{s e^{\sin\theta}(\sin\theta + \cos\theta)}{e^{\sin\theta} + e^{\cos\theta}}. \quad (35)$$

In Figure 9, we visualize the gradient of MSE (in (34)) and CE (in (35)) losses by plotting their gradient vector fields. It shows that rescaling of the MSE loss by either increasing M or increasing α helps to align the gradient along the direction of minimizing θ. Recall that θ determines the classifier's prediction of the class membership for $\boldsymbol{h}$ while s is irrelevant.

When restricting our attention to a feature $\boldsymbol{h}$ with $\theta = \frac{\pi}{2}$, the gradient w.r.t. s and θ becomes s and $\alpha M s$, respectively. Here, increasing the rescaling parameters α or M in the range of $(1, \infty)$ has the effect of increasing the component of the gradient along the θ direction while keeping the component along the s direction fixed.

(a) Vanilla MSE ($\alpha = 1, M = 1$)

(b) Cross Entropy

(c) Rescaled MSE ($\alpha = 5, M = 1$)

(d) Rescaled MSE ($\alpha = 1, M = 5$)

Figure 9. **Visualization of the gradient vector fields with different losses.** We fix $\boldsymbol{W}$ as a simplex ETF and illustrate the landscape only w.r.t. a feature $\boldsymbol{h}_{k,i}$. For each plot, the s-axis denotes $\|\boldsymbol{h}_{k,i}\|_2$, and the θ-axis denotes the angle $\arccos\left(\langle \boldsymbol{h}_{k,i}, \boldsymbol{w}^k \rangle\right)$. The arrows point to gradient descent directions with length proportional to the gradient norm.