arXiv:2210.02192v2 [cs.LG] 8 Oct 2022

Are All Losses Created Equal: A Neural Collapse Perspective

Jinxin Zhou
Ohio State University
zhou.3820@osu.edu

Chong You
Google Research
cyou@google.com

Xiao Li
University of Michigan
xlxiao@umich.edu

Kangning Liu
New York University
kl3141@nyu.edu

Sheng Liu
New York University
shengliu@nyu.edu

Qing Qu
University of Michigan
qingqu@umich.edu

Zhihui Zhu*
Ohio State University
zhu.3440@osu.edu

Abstract

While cross entropy (CE) is the most commonly used loss function to train deep neural networks for classification tasks, many alternative losses have been developed to obtain better empirical performance. Among them, which one is the best to use is still a mystery, because there seem to be multiple factors affecting the answer, such as properties of the dataset, the choice of network architecture, and so on. This paper studies the choice of loss function by examining the last-layer features of deep networks, drawing inspiration from a recent line work showing that the global optimal solution of CE and mean-square-error (MSE) losses exhibits a *Neural Collapse* ($\mathcal{NC}$) phenomenon. That is, for sufficiently large networks trained until convergence, (i) all features of the same class collapse to the corresponding class mean and (ii) the means associated with different classes are in a configuration where their pairwise distances are all equal and maximized. We extend such results and show through global solution and landscape analyses that a broad family of loss functions including commonly used label smoothing (LS) and focal loss (FL) exhibits $\mathcal{NC}$. Hence, all relevant losses (i.e., CE, LS, FL, MSE) produce equivalent features on training data. In particular, based on the *unconstrained feature model* assumption, we provide either the global landscape analysis for LS loss or the local landscape analysis for FL loss and show that the (only!) global minimizers are $\mathcal{NC}$ solutions, while all other critical points are strict saddles whose Hessian exhibit negative curvature directions either in the global scope for LS loss or in the local scope for FL loss near the optimal solution. The experiments further show that $\mathcal{NC}$ features obtained from all relevant losses (i.e., CE, LS, FL, MSE) lead to largely identical performance on test data as well, provided that the network is sufficiently large and trained until convergence. The source code is available at https://github.com/jinxinzhou/nc_loss.

1 Introduction

Loss function is an indispensable component in the training of deep neural networks (DNNs). While cross-entropy (CE) loss is one of the most popular choices for classification tasks, studies over the past few years have suggested many improved versions of CE that bring better empirical performance. Some notable examples include label smoothing (LS) [1] where one-hot label is replaced by a smoothed label, focal loss (FL) [2] which puts more emphasis on hard misclassified samples and

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

reduces the relative loss on the already well-classified samples, and so on. Aside from CE and its variants, the mean squared error (MSE) loss which was typically used for regression tasks is recently demonstrated to have a competitive performance when compared to CE for classification tasks [3].

Despite the existence of many loss functions there is however a lack of consensus as to which one is the best to use, and the answer seems to depend on multiple factors such as properties of the dataset, choice of network architecture, and so on [4]. In this work, we aim to understand the effect of loss function in classification tasks from the perspective of characterizing the last-layer features and classifier of a DNN trained under different losses. Our study is motivated by a sequence of recent work that identify an intriguing *Neural Collapse* ($\mathcal{NC}$) phenomenon in trained networks, which refers to the following properties of the last-layer features and classifier:

(i) ***Variability Collapse*:** all features of the same class collapse to the corresponding class mean.

(ii) ***Convergence to Simplex ETF*:** the means associated with different classes are in a Simplex Equiangular Tight Frame (ETF) configuration where their pairwise distances are all equal and maximized.

(iii) ***Convergence to Self-duality*:** the class means are ideally aligned with the last-layer linear classifiers.

(iv) ***Simple Decision Rule*:** the last-layer classifier is equivalent to a Nearest Class-Center decision rule.

This $\mathcal{NC}$ phenomena is first discovered by Papyan et al. [5, 6] under canonical classification problems trained with the CE loss. Following with the CE loss, Han et al. [7] recently reported that DNNs trained with MSE loss for classification problems also exhibit similar $\mathcal{NC}$ phenomena. 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. The intriguing empirical observation motivated a surge of theoretical investigation [7–22], mostly under a simplified *unconstrained feature model* [10] or *layer-peeled model* [12] that treats the last-layer features of each samples before the final classifier as free optimization variables. Under the simplified unconstrained feature model, it has been proved that the $\mathcal{NC}$ solution is the only global optimal solution for the CE and MSE losses which are also proved to have benign global landscape, explaining why the global $\mathcal{NC}$ solution can be obtained.

Contributions. While previous work provide thorough analysis for $\mathcal{NC}$ under CE and MSE losses, the theoretical analysis beyond CE and MSE losses is still limited, and their work only focus on one specific loss without a general format. In this paper, we consider a broad family of loss functions that includes CE and some other popular loss functions such as LS and FL as special cases. Under the *unconstrained feature model*, we theoretically demonstrate in Section 3 that the $\mathcal{NC}$ solution is the only global optimal solution to the family of loss functions. Moreover, we provide a global landscape analysis, showing that the LS loss function is a strict saddle function and FL loss function is a local strict saddle function [23–25]. A (local) strict saddle function is a function for which every critical point is either a global solution or a strict saddle point with negative curvature (locally). Hence, our result suggests that any optimizer can escape strict saddle points and converge to the global solution responding to $\mathcal{NC}$ for LS and FL. As far as we know, this paper is the first work that conducts global optimal solution and benign optimization landscape analysis beyond the scope of CE and MSE losses.

Our theoretical results explained above have important implications for understanding the role of loss function in training DNNs for classification tasks. Because all losses lead to $\mathcal{NC}$ solutions, their corresponding features are equivalent up to a rotation of the feature space. In other words, our analysis provides a theoretical justification for the following claim:

> *All losses (i.e., CE, LS, FL, MSE) lead to largely identical features on* ***training*** *data by large DNNs and sufficiently many iterations.*

We also provide an experimental verification of this claim through experiments in Section 4.1.

While $\mathcal{NC}$ reveals that all losses are equivalent at training time, it does not have a direct implication for the features associated with test data as well as the generalization performance [26]. In particular, a recent work [27] shows empirically that $\mathcal{NC}$ does not occur for the features associated with test data. Nonetheless, we show through empirical evidence that for large DNNs, $\mathcal{NC}$ on training

data well predicts the test performance. In particular, our empirical study in Section 4.2 shows the following:

*All losses (CE, LS, FL, MSE) lead to largely identical performance on **test** data by large DNNs.*

Our conclusion that all losses are created equal appears to go against existing evidence on the advantages of some losses over the others. Here we emphasize that our conclusion has an important premise, namely the neural network has sufficient approximation power and the training is performed for sufficiently many iterations. Hence, our conclusion implies that the better performance with particular choices of loss functions comes as a result that the training does not produce a globally optimal (i.e., $\mathcal{NC}$) solution. In such cases different losses lead to different solutions on the training data, and correspondingly different performance on test data. Such an understanding may provide important practical guidance on what loss to choose in different cases (e.g., different model sizes and different training time budgets), as well as for the design of new and better losses in the future. We note that our conclusion is based on natural accuracy, rather than model transferability or robustness, which is worth additional efforts to exploit and is left as future work.

2 The Problem Setup

A typical deep neural network $\Psi(\cdot) : \mathbb{R}^D \mapsto \mathbb{R}^K$ consists of a multi-layer nonlinear compositional feature mapping $\Phi(\cdot) : \mathbb{R}^D \mapsto \mathbb{R}^d$ and a linear classifier $(\boldsymbol{W}, \boldsymbol{b})$, which can be generally expressed as

$$\Psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) \;=\; \boldsymbol{W}\Phi_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{b}, \tag{1}$$

where we use $\boldsymbol{\theta}$ to represent the network parameters in the feature mapping and $\boldsymbol{W} \in \mathbb{R}^{K\times d}$ and $\boldsymbol{b} \in \mathbb{R}^K$ to represent the linear classifier's weight and bias, respectively. Therefore, *all* the network parameters are the set of $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b}\}$. For the input $\boldsymbol{x}$, the output of the feature mapping $\Phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is usually termed as the *representation* or *feature* learned from the network.

With an appropriate loss function, the parameters $\boldsymbol{\Theta}$ of the whole network are optimized to learn the underlying relation from the input sample $\boldsymbol{x}$ to their corresponding target $\boldsymbol{y}$ so that the output of the network $\Psi_{\boldsymbol{\Theta}}(\boldsymbol{x})$ approximates the corresponding target, i.e. $\Psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) \approx \boldsymbol{y}$ in term of the expectation over a distribution $\mathcal{D}$ of input-output data pairs $(\boldsymbol{x}, \boldsymbol{y})$. While it is hard to get access to the ground-truth distribution $\mathcal{D}$ in most cases, one can approximate the distribution $\mathcal{D}$ through sampling enough data pairs i.i.d. from $\mathcal{D}$. In this paper, we study the multi-class balanced classification tasks with K class and n samples per class, where we use the one-hot vector $\boldsymbol{y}_k \in \mathbb{R}^K$ with unity only in k-th entry $(1 \le k \le K)$ to denote the label of the i-th sample $\boldsymbol{x}_{k,i} \in \mathbb{R}^D$ in the k-th class. We then learn the parameters $\boldsymbol{\Theta}$ via minimizing the following empirical risk over the total $N = nK$ training samples

$$\min_{\boldsymbol{\Theta}} \; \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n} \mathcal{L}\left(\Psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i}), \boldsymbol{y}_k\right) \;+\; \frac{\lambda}{2}\left\|\boldsymbol{\Theta}\right\|_F^2, \tag{2}$$

where $\lambda > 0$ is the regularization parameter (a.k.a., the weight decay parameter[2]) and $\mathcal{L}\left(\Psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i}), \boldsymbol{y}_k\right)$ is a predefined loss function that appropriately measures the difference between the output $\Psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i})$ and the target $\boldsymbol{y}_k$. Some common loss functions used for training deep neural networks will be specified in the next section.

2.1 Commonly Used Training Losses

In this subsection, we first present four common loss functions for classification task. To simplify the notation, let $\boldsymbol{z} = \boldsymbol{W}\Phi_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{b}$ denote the network's output ("logit") vector for the input $\boldsymbol{x}$. Assume $\boldsymbol{z}$ belongs to the k-th class. Also let $\boldsymbol{y}_k^{\text{smooth}} = (1-\alpha)\boldsymbol{y}_k + \frac{\alpha}{K}\boldsymbol{1}_K$ denote the smoothed targets of k-th class, where $0 \le \alpha < 1$ and $\boldsymbol{1}_K \in \mathbb{R}^K$ is a vector with all entries equal to one. We will use z_ℓ, $y_{k,\ell}$ and y_ℓ^{smooth} to denote the ℓ-th entry of $\boldsymbol{z}$, $\boldsymbol{y}_k$ and $\boldsymbol{y}_k^{\text{smooth}}$, respectively, where $y_{k,k}^{\text{smooth}} = 1 - \frac{K-1}{K}\alpha$ and $y_{k,\ell}^{\text{smooth}} = \frac{\alpha}{K}$ for $k \ne \ell$.

[2]Without weight decay, the features and classifiers will tend to blow up for CE and many other losses.

Cross entropy (CE) is perhaps the most common loss for multi-class classification in deep learning. It measures the distance between the target distribution $\boldsymbol{y}_k$ and the network output distribution obtained by applying the softmax function on $\boldsymbol{z}$, resulting in the following expression

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) \;=\; -\log\left(\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}\right). \tag{3}$$

Focal loss (FL) [2] is first proposed to deal with the extreme foreground-background class imbalance in dense object detection, which adaptively focuses less on the well-classified samples. Recent work [28, 29] reports that focal loss also improves calibration and automatically forms curriculum learning in multi-class classification setting. Letting $\gamma \geq 0$ denote the tunable focusing parameter, the focal loss can be expressed as:

$$\mathcal{L}_{\mathrm{FL}}(\boldsymbol{z}, \boldsymbol{y}_k) \;=\; -\left(1-\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}\right)^{\gamma}\log\left(\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}\right). \tag{4}$$

Label smoothing (LS) [1] replaces the hard targets in CE with smoothed targets $\boldsymbol{y}_k^{\text{smooth}}$ obtained from mixing the original targets $\boldsymbol{y}_k$ with a uniform distribution over all entries $\mathbf{1}_K$. Experiments in [30, 31] find that classification models trained with label smoothing have better calibration and generalization. Denoting by $0 \leq \alpha \leq 1$ the tunable smoothing parameter, the label smoothing loss function can be formulated as:

$$\mathcal{L}_{\mathrm{LS}}(\boldsymbol{z}, \boldsymbol{y}_k) \;=\; -\sum_{\ell=1}^{K} y_{k,\ell}^{\text{smooth}}\log\left(\frac{\exp(z_\ell)}{\sum_{j=1}^{K}\exp(z_j)}\right). \tag{5}$$

When $\alpha = 0$, the above label smoothing loss reduces to the CE loss.

Mean square error (MSE) is often used for regression but not classification task. The recent work [3] shows that classification networks trained with MSE loss achieve on par performance compared to those trained with the CE loss. Throughout our paper, we use the rescaled MSE version [3]:

$$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{z}, \boldsymbol{y}_k) \;=\; \kappa(z_k-\beta)^2 + \sum_{\ell\neq k}^{K} z_\ell^2, \tag{6}$$

where $\kappa > 0$ and $\beta > 0$ are hyperparameters.

2.2 Problem Formulation Based on Unconstrained Feature Models

Because of the interaction between a large number of nonlinear layers in the feature mapping $\Phi_{\boldsymbol{\theta}}$, it is tremendously challenging to analyze the optimization of deep neural networks. To simplify the difficulty of deep neural network analysis, a series of recent works of theoretically studying $\mathcal{NC}$ phenomenon use a so-called *unconstrained feature model* (or *layer-peeled model* in [12]) which treats the last-layer features as *free* optimization variables $\boldsymbol{h} = \Phi(\boldsymbol{x}) \in \mathbb{R}^d$. The reason behind the *unconstrained feature model* is that modern highly overparameterized deep networks are able to approximate any continuous functions [32–35] and the characterization of $\mathcal{NC}$ are only related with the last layer features. We adopt the same approach and study the effects of different training losses on the last-layer representations of the network under the unconstrained feature model. For convenient, let us denote

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

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

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}\left(\boldsymbol{W}\boldsymbol{h}_{k,i}+\boldsymbol{b}, \boldsymbol{y}_k\right) + \frac{\lambda_{\boldsymbol{W}}}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2, \tag{7}$$

where $\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}, \lambda_{\boldsymbol{b}} > 0$ are the penalty parameters for $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{b}$, respectively.

By viewing the last-layer feature $\boldsymbol{H}$ as a free optimization variable, the simplified objective function (7) consider the weight decay about $\boldsymbol{W}$ and $\boldsymbol{H}$, which is slightly different from practice that the weight decay is imposed on all the network parameters $\boldsymbol{\Theta}$ as shown in (2). Nonetheless, the underlying rationale is that the weight decay on $\boldsymbol{\Theta}$ *implicitly* penalizes the energy of the features (i.e., $\|\boldsymbol{H}\|_F$) [16].

As $\mathcal{NC}$ phenomena for the learned features and classifiers is first discovered for neural networks trained with the CE loss [5], the CE loss has been mostly studied through the above simplified unconstrained feature model [8,9,11,12,16] to understand the $\mathcal{NC}$ phenomena. The work [7,10,14, 22] also studied the MSE loss, but the analysis there shows the solutions of the learned features and classifiers depend crucially on the bias term, while for CE loss with or without the bias term have no effect on the learned features and classifiers under the unconstrained feature model. The other losses such as focal loss and label smoothing have been less studied, though they are widely employed in practice to obtain better performance. This will be the subject of next section.

3 Understanding Loss Functions Through Unconstrained Features Model

In this section, we study the effect of different loss functions through the unconstrained features model. We will first present a contrastive property for general loss function $\mathcal{L}_{\mathrm{GL}}$ in Definition 1. We will then study the global optimality conditions in terms of the learned features and classifiers as well as geometric properties for (7) with such a general loss function $\mathcal{L}_{\mathrm{GL}}$.

3.1 A Contrastive Property for the Loss Functions

In this paper, we aim to provide a unified analysis for different loss functions. Towards that goal, we first present some common properties behind the CE, FL and FL to motivate the discussion. Taking CE as an example, we can lower bound it by

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) \geq \log\left(1 + (K-1)\exp\left(\frac{\sum_{j\neq k}(z_j - z_k)}{K-1}\right)\right) = \phi_{\mathrm{CE}}\left(\sum_{j\neq k}(z_j - z_k)\right) \tag{8}$$

where $\phi_{\mathrm{CE}}(t) = \log\left(1 + (K-1)\exp\left(\frac{t}{K-1}\right)\right)$, and the inequality achieves equality when $z_j = z_{j'}$ for all $j, j' \neq k$. This requirement is reasonable because the commonly used losses treat all the outputs except for the k-th output $\boldsymbol{z}$ identically. Since ϕ_{CE} is an increasing function, minimizing the CE loss $\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k)$ is equivalent to maximizing $(K-1)z_k - \sum_{j\neq K} z_j$, which contrasts the k-th output z_k simultaneously to all the other outputs z_j for all $j \neq k$. Thus, we call (8) as a *contrastive property*. Maximizing $(K-1)z_k - \sum_{j\neq K} z_j$ would lead to a positive (and relatively large) z_k and negative (and relatively small) z_j. In particular, within the unit sphere $\|\boldsymbol{z}\|_2 = 1$, $(K-1)z_k - \sum_{j\neq K} z_j$ achieves its maximizer when $z_k = \sqrt{\frac{K-1}{K}}$ and $z_j = -\sqrt{\frac{1}{K(K-1)}}$ for all $j \neq k$, which satisfies the requirement $z_j = z_{j'}$ for all $j, j' \neq k$. Thus, $z_k = \sqrt{\frac{K-1}{K}}$ and $z_j = -\sqrt{\frac{1}{K(K-1)}}$ is also the global minimizer for ϕ_{CE} within the unit sphere $\|\boldsymbol{z}\|_2 = 1$. As the global minimizer is unique for each class, it encourages intra-class compactness. On the other hand, the minimizers to different classes are maximally distant, promoting inter-class separability.

Motivated by the above discussion, we now introduce the following properties for a general loss function $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$.

Definition 1 (Contrastive property)**.** *We say a loss function $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ satisfies the contrastive property if there exists a function ϕ such that $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ can be lower bounded by*

$$\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k) \geq \phi\left(\sum_{j\neq k}(z_j - z_k)\right) \tag{9}$$

where the equality holds only when $z_j = z'_j$ for all $j, j' \neq k$. Moreover, $\phi(t)$ satisfies

$$t^* = \arg\min_t \phi(t) + c|t| \text{ is unique for any } c > 0, \text{and } t^* \leq 0. \tag{10}$$

In the appendix, we show that CE, FL and LS all satisfy this property. The motivation for (9) follows from the above discussion. In particular, (9) achieves equality when all the outputs except for the k-th one are identical, which holds for common loss functions since those outputs are treated identically. In (10), c is a constant related with the weight decay penalty parameters. By (9), we can find the global minimizer for $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ by minimizing the right hand side since the equality in (9) is achievable. Thus, the requirement of a unique minimizer (10) ensures a unique minimizer for the regularized $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$. This condition can be easily satisfied. For example, $\phi_{\mathrm{CE}}(t)$ defined in (8) for the CE loss is an increasing and strictly convex function and thus has unique minimizer for $\phi_{\mathrm{CE}}(t) + c|t|$. Along the same line, we require a negative minimizer $t^\star$ to ensure that the minimizer for the regularized $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ has k-th entry being its largest entry, which is required to ensure correct prediction since the largest entry predicts the class membership. Therefore, such a condition is generally satisfied by the common losses. For example, $\phi_{\mathrm{CE}}(t)$ is an increasing function and thus must have a non-positive minimizer for $\phi_{\mathrm{CE}}(t) + c|t|$. Finally, we note that the MSE loss is not included since it has different form than others and thus the analysis will be different. But as mentioned above, the MSE loss has been studied in [7, 10, 14, 22].

3.2 Landscape Analysis for the Unconstrained Features Model

We now study the global optimality conditions in terms of the learned features and classifiers as well as geometric properties for the training problem (7) with the general loss function $\mathcal{L}_{\mathrm{GL}}$ satisfying the above contrastive property.

Theorem 1 (Global Optimality Condition)**.** *Assume that the number of classes K is smaller than the feature dimension d, i.e., $K < d$, and the dataset is balanced for each class, $n = n_1 = \cdots = n_K$. Then any global minimizer $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ of f in (7) with a loss function $\mathcal{L}_{\mathrm{GL}}$ satisfying the contrastive property in Definition 1 has following properties:*

$$\|\boldsymbol{w}^\star\|_2 = \left\|\boldsymbol{w}^{\star 1}\right\|_2 = \left\|\boldsymbol{w}^{\star 2}\right\|_2 = \cdots = \left\|\boldsymbol{w}^{\star K}\right\|_2, \quad \text{and} \quad \boldsymbol{b}^\star = b^\star \mathbf{1},$$

$$\boldsymbol{h}^\star_{k,i} = \sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}} n}} \boldsymbol{w}^{\star k}, \quad \forall\, k \in [K],\ i \in [n], \quad \text{and} \quad \overline{\boldsymbol{h}}^\star_i := \frac{1}{K} \sum_{j=1}^{K} \boldsymbol{h}^\star_{j,i} = \mathbf{0}, \quad \forall\, i \in [n],$$

where either $b^\star = 0$ or $\lambda_{\boldsymbol{b}} = 0$, and the matrix $\boldsymbol{W}^{\star\top}$ is in the form of K-simplex ETF structure (see appendix for the formal definition) in the sense that

$$\boldsymbol{W}^{\star\top} \boldsymbol{W}^\star = \|\boldsymbol{w}^\star\|_2^2 \frac{K}{K-1} \left(\boldsymbol{I}_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top \right).$$

Its proof is given in Appendix C. At a high level, we lower bound the general loss function based on the contrastive property (9) and then check the equality conditions hold for the lower bounds. While similar strategy has been used for CE loss [12, 13, 16], our proof is different from previous work in terms of dealing with the nuclear norm and checking the structures of the minimizer per sample and class, enabling the global optimality analysis for general loss functions. Theorem 1 implies that for all the loss functions (e.g., CE, LS, and FL) satisfying the contrastive property, they share similar global solutions with $\mathcal{NC}$ property in the learned features and classifiers.

While Theorem 1 shows that $\mathcal{NC}$ features and classifiers are the only global minimizers to (7), it is not obvious whether local search algorithms (such as gradient descent) can efficiently find these benign global solutions. The reason is that the training problem (7) is nonconvex due to the bilinear form between $\boldsymbol{W}$ and $\boldsymbol{H}$. To address this challenge, we use the recent advances on the geometric analysis for nonconvex optimization [23–25, 36] to guarantee that the global solutions of (7) can be efficiently achieved by iterative algorithms. Towards that goal, we first present the following general results concerning the global landscape for (7).

Theorem 2 (Benign Landscape)**.** *Assume that the feature dimension d is larger than the number of classes K, i.e., $d > K$. Also assume $\mathcal{L}(\boldsymbol{z}, \boldsymbol{y})$ is a convex function in terms of $\boldsymbol{z}$. Then the objective function f in (7) 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.*

This result is similar to [16, Theorem 3.2] which studies the particular CE loss. Though the result in [16] is about the CE loss, we checked its proof and it only uses convexity and smoothness and thus the result can be applied more generally for any smooth convex loss function $\mathcal{L}(\cdot, \boldsymbol{y})$. So we omit the proof of Theorem 2. We note that the geometric analysis is also closed related to nonconvex low-rank matrix problems [37–43] with the Burer-Moneirto factorization approach [44] if one views $\boldsymbol{W}$ and $\boldsymbol{H}$ as two factors of a matrix $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}$. We refer to [16] for more discussions about the connections and differences.

We now exploit Theorem 2 for the label smoothing and focal loss. In the supplementary material, we show that LS is a convex function. Thus, the following result establishes global optimization landscape for the training problem (7) with such a loss.

Corollary 1 (Benign Landscape with LS)**.** *Assume that the feature dimension d is larger than the number of classes K, i.e., $d > K$. Then the objective function f in (7) with LS loss $\mathcal{L}_{\mathrm{LS}}$ is a strict saddle function with no spurious local minimum.*

Unlike LS, focal loss $\mathcal{L}_{\mathrm{FL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ is convex only in a local region rather than the entire space. For example, we can show that $\mathcal{L}_{\mathrm{FL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ is convex within the region $\Omega = \{\boldsymbol{z} \in \mathbb{R}^K : \exp(z_k)/\sum_{j=1}^K \exp(z_j) \geq 0.21\}$. The set Ω contains a relative large region including the global minimizer which has the value $\frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)}$ approaching 1. Thus, we obtain a benign local landscape for the training problem (7) with FL.

Corollary 2 (Benign Landscape with FL)**.** *Assume that the feature dimension d is larger than the number of classes K, i.e., $d > K$. Then the objective function f in (7) with FL loss $\mathcal{L}_{\mathrm{FL}}$ has a benign local landscape: f is a strict saddle function with no spurious local minimum within the region $\{(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) : \boldsymbol{W}\boldsymbol{h}_{k,i} + \boldsymbol{b} \in \Omega,\ 1 \leq k \leq K, 1 \leq i \leq n\}$.*

While Corollary 2 only provides a local benign landscape for the FL, we observe from experiments that gradient descent with random initialization always converges to a global solution with $\mathcal{NC}$ properties for (7). So we expect the training problem in (7) with FL loss has benign landscape in a much larger region. One direction is to show $\mathcal{L}_{\mathrm{FL}}(\cdot, \boldsymbol{y}_k)$ is locally convex in a much larger region Ω, but we leave the thorough investigation to future work. Noting that the CE and MSE losses are also convex, these results imply that (stochastic) gradient descent with random initialization [23, 36] almost surely finds the global solutions of the training problem in (7) with different training losses. This together with Theorem 1 implies that for different losses, gradient descent will always learn similar features and classifiers—those that exhibit the $\mathcal{NC}$ phenomenon.

4 Experiments

We conduct experiments with practical network architectures on standard image classification datasets to study the effect of different loss functions. First, Section 4.1 provides results to show that the $\mathcal{NC}$ phenomena are not restricted to networks trained via the CE and MSE losses. Rather, there is a family of loss functions, and for the purpose of illustration we pick FL and LS as two prominent special cases, that exhibit the same $\mathcal{NC}$ phenomena. Such results verify our theoretical results in Section 3. To demonstrate the implication of $\mathcal{NC}$ for test performance, we present experimental results in Section 4.2 with a varying number of training iterations and a varying width of networks, showing that *all* losses with $\mathcal{NC}$ global optimality have similar performance on the test dataset when the network is sufficiently large and trained long enough.

Before presenting the experiment results, we first introduce our experimental setup, including datasets, network architectures, training procedure, and metrics for measuring $\mathcal{NC}$.

Setup of Loss Function, Network Architecture, Dataset, and Training We focus on the CE, FL, LS and MSE loss functions for which we use $\gamma = 3$ for FL, $\alpha = 0.1$ for LS, and $\kappa = 1$ and $\beta = 15$ for MSE, except otherwise specificed. We train a WideResNet50 network [45] on CIFAR10 and CIFAR100 datasets [46] and a WideResNet18 network on miniImageNet [47] with various widths and number of iterations for image classification using these four different losses.[3] To examine the

[3] Similar results are expected on other architectures and dataset as $\mathcal{NC}$ is observed across a range of architectures and dataset in [5].

effect of model size, we experiment with four versions of WideResNet, denoted as WideResNet-X, where $X \in \{0.25, 0.5, 1, 2\}$ is a multiplier on the width of its corresponding standard WideResNet. Due to the page limit, we put all results on CIFAR100 and miniImageNet in the Appendix. We use standard preprocessing such that images are normalized (channel-wise) by their mean and standard deviation, as well as standard data augmentation. For optimization, we use SGD with momentum 0.9 and an initial learning rate 0.1 decayed by a factor of 0.1 at $\frac{3}{7}$ and $\frac{5}{7}$ of the total number of iterations. Following [28], the norm of gradient is clipped at 2 which can improve performance for all losses. For CIFAR10 and miniImageNet, the weight decay is set to 5×10^4 for all configurations with all losses. For CIFAR100, the weight decay is fine-tuned to achieve best accuracy for every configuration and loss.

Three $\mathcal{NC}$ Metrics $\mathcal{NC}_1$-$\mathcal{NC}_3$ during Network Training We use the same three metrics $\mathcal{NC}_1$-$\mathcal{NC}_3$ for the last-layer features and classifier as in [5,16,22] to measure the first three $\mathcal{NC}$ properties in Section 1. Before we describe these three metrics, let us denote the global mean $\boldsymbol{h}_G$ and k-th class mean $\overline{\boldsymbol{h}}_k$ of last-layer features $\{\boldsymbol{h}_{k,i}\}$ as

$$\boldsymbol{h}_G \;=\; \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\boldsymbol{h}_{k,i}, \quad \overline{\boldsymbol{h}}_k \;=\; \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{h}_{k,i}\;(1\le k\le K).$$

Within-class variability collapse is measured by $\mathcal{NC}_1$ which depicts the relative magnitude of the within-class covariance $\boldsymbol{\Sigma}_W = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\boldsymbol{h}_{k,i}-\overline{\boldsymbol{h}}_k\right)\left(\boldsymbol{h}_{k,i}-\overline{\boldsymbol{h}}_k\right)^\top \in \mathbb{R}^{d\times d}$ w.r.t. the between-class covariance $\boldsymbol{\Sigma}_B = \frac{1}{K}\sum_{k=1}^{K}\left(\overline{\boldsymbol{h}}_k-\boldsymbol{h}_G\right)\left(\overline{\boldsymbol{h}}_k-\boldsymbol{h}_G\right)^\top \in \mathbb{R}^{d\times d}$ of the last-layer features as following:

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

where $\boldsymbol{\Sigma}_B^{\dagger}$ is the pseudo inverse of $\boldsymbol{\Sigma}_B$.

Convergence to simplex ETF is measured by $\mathcal{NC}_2$ which reflects the ℓ_2 distance between the normalized simplex ETF and the normalized $\boldsymbol{W}\boldsymbol{W}^\top$ as following:

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

where $\boldsymbol{W} \in \mathbb{R}^{K\times d}$ is the weight matrix of learned classifier.

Convergence to self-duality is measured by $\mathcal{NC}_3$ which calculates the ℓ_2 distance between the normalized simplex ETF and the normalized $\boldsymbol{W}\overline{\boldsymbol{H}}$ as following:

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

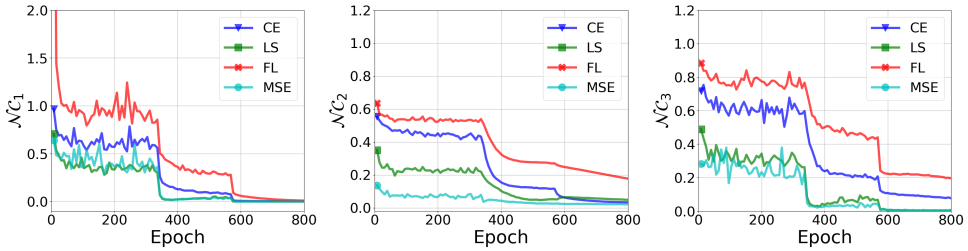
where $\overline{\boldsymbol{H}} = \begin{bmatrix}\overline{\boldsymbol{h}}_1 - \boldsymbol{h}_G & \cdots & \overline{\boldsymbol{h}}_K - \boldsymbol{h}_G\end{bmatrix} \in \mathbb{R}^{d\times K}$ is the centered class-mean matrix.

4.1 Prevalence of $\mathcal{NC}$ Across Varying Training Losses

We show that all loss functions lead to $\mathcal{NC}$ solutions during the terminal phase of training. The results on CIFAR10 using WideResNet50-2 and different loss functions is provided in Figure 1. We consistently observe that all three $\mathcal{NC}$ metrics across different losses converge to a small value as training progresses. This supports our theoretical results in Section 3 that the last-layer features learned under different losses are always maximally linearly separable and perfectly aligned with the linear classifier, and the features and the weight of linear classifier learned by different losses are almost equivalent up to a rotation and a scale of the feature space. The evolution of three $\mathcal{NC}$ metrics across different losses on CIFAR100 is in Appendix A.2.

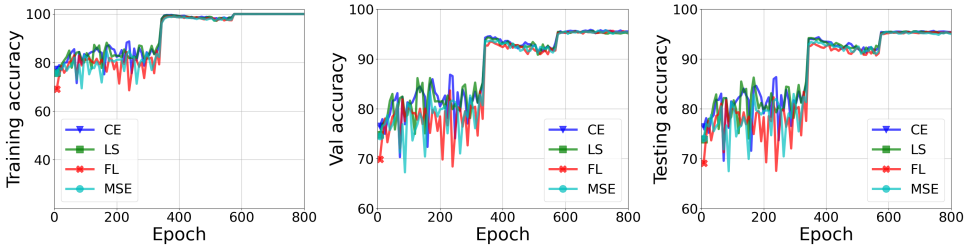
4.2 All Losses Lead to Largely Identical Performance

We show that all loss functions have largely identical performance once the training procedure converges to the $\mathcal{NC}$ global optimality. In Figure 2, we plot the evolution of the training accuracy,



(a) $\mathcal{NC}_1$ (CIFAR10) (b) $\mathcal{NC}_2$ (CIFAR10) (c) $\mathcal{NC}_3$ (CIFAR10)

Figure 1: **The evolution of $\mathcal{NC}$ metrics across different loss functions.** We train the WideResNet50-2 on CIFAR10 dataset for 800 epochs using different loss functions. From left to right: $\mathcal{NC}_1$ (variability collapse), $\mathcal{NC}_2$ (convergence to simplex ETF) and $\mathcal{NC}_3$ (convergence to self-duality).



(a) Train Acc (CIFAR10) (b) Val Acc (CIFAR10) (c) Test Acc (CIFAR10)

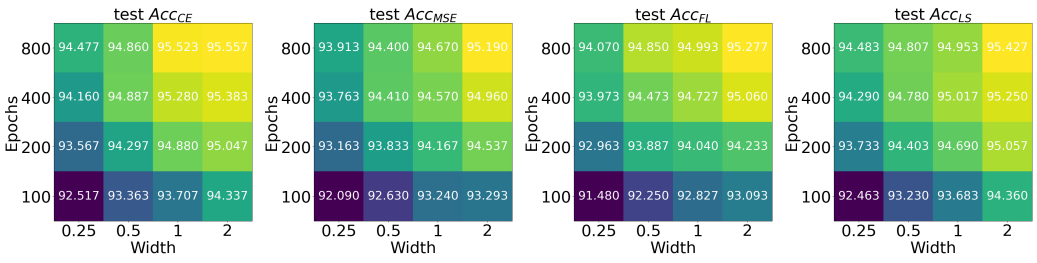
Figure 2: **The evolution of performance across different loss functions.** We train the WideResNet50-2 on CIFAR10 dataset for 800 epochs using different loss functions. From left to right: training accuracy, validation accuracy and test accuracy.

validation accuracy and test accuracy as training progresses, where all losses are optimized on the same WideResNet50-2 architecture and CIFAR10 for 800 epochs. To reduce the randomness, we average the results from 3 different random seeds per width-iteration configuration, and the test accuracy is reported based on the model with best accuracy on validation set, where we organize the validation set by holding out 10 percent data from the training set. The results consistently show that for all cases the training accuracy converges to one hundred percent (reaching to terminal phase), and the validation accuracy and test accuracy are largely the same, as long as the network is trained longer enough and converges to the $\mathcal{NC}$ global solution.

While previous work advocates the advantage of some losses over other others, our experiments show that when conditions between dataset and model allow for SGD to find an $\mathcal{NC}$ solution, all losses we tested produced indistinguishable results. In Figure 3, we plot the average test accuracy of different losses under different pairs of width and iterations. We consistently observe three phenomenon. First, with a fixed number of iterations, increasing the width of network improves the test accuracy for all losses. This is because the wider networks (more over-parameterized) are more powerful to fit the underlying mapping from input data to the targets. Second, with a fixed width of network, increasing the number of iterations improves the test accuracy for all losses. This is because the longer optimization leads the last-layer features and the linear classifer closer to the $\mathcal{NC}$ global solutions. Finally, while there are some unignorable difference between different losses in some width-iteration configurations, the results consistently show that all losses lead to largely identical performance when the network is sufficiently large and trained long enough to achieve a global $\mathcal{NC}$ solution (e.g. width=2 and epochs=800).

5 Conclusion

In this work we provided a theoretical study to extend the scope of $\mathcal{NC}$, a curious phenomenon associated with last-layer features and classifier weight of a classification network, from networks trained with particular losses (i.e., CE and MSE) to those trained via a broad family of loss functions including the popular LS and FL as special cases. Our theory not only establishes $\mathcal{NC}$ as the only global solutions, but also shows a benign optimization landscape that explains why $\mathcal{NC}$ solutions



(a) Test (CE) (b) Test (MSE) (c) Test (FL) (d) Test (LS)

Figure 3: **Illustration of test accuracy across different iterations-width settings.** The figure depicts the test accuracy of various iteration-width configurations for different loss functions on CIFAR10.

are easy to obtain in practice. Such results readily suggest that all relevant losses (i.e., CE, MSE, LS, and FL) produce entirely equivalent features on the training data. Although $\mathcal{NC}$ is an optimization phenomenon pertaining to training data only, we found through experiments that all relevant losses (i.e., CE, MSE, LS, and FL) lead to very similar test performance as well. Such a result may come as a surprise to the common belief that some losses are intrinsically better than the others, and clarify some mystery on how different losses affect the performance.

The family of loss functions considered in this paper by no means is inclusive of all possible loss functions that lead to $\mathcal{NC}$. There are many other popular loss functions, such as center loss [48], large-margin softmax (L-Softmax) loss [49] and many of its variants [50–52], which are all designed with the intuition of encouraging intra-class compactness and inter-class separability between learned features. In addition, many generalized versions of the cross-entropy loss such as those for robust learning under label noise [53–55] and long-tail distribution [56, 57] may have similar property as the vanilla cross-entropy loss. We conjecture that many of them provably produce $\mathcal{NC}$ solutions under unconstrained feature models, while leave a formal justification to future work. Beyond losses for classification task, $\mathcal{NC}$ may also arise with popular losses used in metric learning [58, 59] evidenced by recent study [60]. This means that the observations from this paper, namely all losses lead to largely the same test performance, may apply for all such losses as well.

Loss functions that do not lead to $\mathcal{NC}$. While the study in this paper covers many of the most commonly used loss functions for classification tasks, we note that there are alternative choices in the literature which do not induce $\mathcal{NC}$ features. Many of such losses such as [60–63] are particularly designed to discourage variability collapse and learn diverse features, which are shown to benefit model transferability [64] and robustness.

Acknowledgment

JZ and ZZ acknowledge support from NSF grants CCF-2008460 and CCF-2106881. XL and QQ acknowledge support from U-M START & PODS grants, NSF CAREER CCF 2143904, NSF CCF 2212066, NSF CCF 2212326, and ONR N00014-22-1-2529 grants. SL acknowledges support from NSF NRT-HDR Award 1922658 and Alzheimer's Association grant AARG- NTF-21-848627.

References

[1] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[3] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

[4] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, 25:49–59, 2016.

[5] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

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

[7] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[8] Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

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

[10] Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

[11] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[12] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Layer-peeled model: Toward understanding well-trained deep neural networks. *arXiv e-prints*, pages arXiv–2101, 2021.

[13] Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

[14] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. *arXiv preprint arXiv:2202.08087*, 2022.

[15] Tolga Ergen and Mert Pilanci. Revealing the structure of deep neural networks via convex duality. In *International Conference on Machine Learning*, pages 3004–3014. PMLR, 2021.

[16] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[18] Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

[19] Tomaso Poggio and Qianli Liao. Implicit dynamic regularization in deep networks. Technical report, Center for Brains, Minds and Machines (CBMM), 2020.

[20] Akshay Rangamani, Mengjia Xu, Andrzej Banburski, Qianli Liao, and Tomaso Poggio. Dynamics and neural collapse in deep classifiers trained with the square loss. *Technical report, Center for Brains, Minds and Machines (CBMM)*, 2021.

[21] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Neural collapse in deep homogeneous claaifiers and the role of weight decay. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[22] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. *arXiv preprint arXiv:2203.01238*, 2022.

[23] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[24] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[25] Yuqian Zhang, Qing Qu, and John Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

[26] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[27] Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

[28] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

[29] Leslie N Smith. Cyclical focal loss. *arXiv preprint arXiv:2202.08978*, 2022.

[30] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[31] Blair Chen, Liu Ziyin, Zihao Wang, and Paul Pu Liang. An investigation of how label smoothing affects generalization. *arXiv preprint arXiv:2010.12648*, 2020.

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

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

[34] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: a view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6232–6240, 2017.

[35] Uri Shaham, Alexander Cloninger, and Ronald R Coifman. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, 44(3):537–557, 2018.

[36] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

[37] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[38] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[39] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3880–3888, 2016.

[40] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[41] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[42] Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

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

[44] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[46] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[47] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[49] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016.

[50] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[51] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[52] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[53] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[54] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[55] Jiaheng Wei and Yang Liu. When optimizing f-divergence is robust with label noise. In *International Conference on Learning Representations*, 2021.

[56] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[57] Wittawat Jitkrittum, Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Elm: Embedding and logit margins for long-tail learning. *arXiv preprint arXiv:2204.13208*, 2022.

[58] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.

[59] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[60] Elad Levi, Tete Xiao, Xiaolong Wang, and Trevor Darrell. Reducing class collapse in metric learning with easy positive sampling. *arXiv preprint arXiv:2006.05162*, 2020.

[61] Xu Zhang, Felix X Yu, Sanjiv Kumar, and Shih-Fu Chang. Learning spread-out local feature descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 4595–4603, 2017.

[62] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6450–6458, 2019.

[63] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *arXiv preprint arXiv:2006.08558*, 2020.

[64] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34, 2021.

[65] Jan JM Cuppen. A divide and conquer method for the symmetric tridiagonal eigenproblem. *Numerische Mathematik*, 36(2):177–195, 1980.

[66] N Jakovčević Stor, I Slapničar, and Jesse L Barlow. Forward stable eigenvalue decomposition of rank-one modifications of diagonal matrices. *Linear Algebra and its Applications*, 487:301–315, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the abstract, introduction, main results, and conclusion, we explicitly stat that our results are about unconstrained feature model.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper mainly focuses on understanding the neural collapse phenomena observed in practical neural networks. Based on this understanding, we propose to fix the last layer classifier as a Simplex ETF. So no potential negative societal impact is expected of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We explicitly mention the assumptions in Theorem 1 and Theorem 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] We include all the proofs in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All the results are obtained by averaging the resluts from 3 different random seeds. See Section 4
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite them in Appendix A.
   (b) Did you mention the license of the assets? [Yes] This is mentioned in Appendix A.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] This is discussed in Appendix A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The CIFAR datasets do not contain personally identifiable information or offensive content
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Our research does not involve with participants and screenshots
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Our research does not include any potential participant risks, with links to Institutional Review Board (IRB) approvals
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Our work does not require participants.

Appendices

Organizations and Basic. The appendix is organized as follows. We first introduce the basic definitions and inequalities used throughout the appendices. In Appendix A, we provide more details about the datasets, computational resources, and more experiment results on CIFAR10, CIFAR100 and miniImageNet datasets. In Appendix B, we prove that CE, FL and LS satisfy the contrastive property in Definition 1. In Appendix C, we provide a detailed proof for Theorem 1, showing that the Simplex ETFs are the *only* global minimizers, as long as the loss function satisfies the Definition 1. Finally, in Appendix D, we present the whole proof for Theorem 2 that the FL function is a locally strict saddle function with no spurious local minimizers existing locally and LS function is a globally strict saddle function with no spurious local minimizers existing globally.

Definition 2 (K-Simplex ETF)**.** *A standard Simplex ETF is a collection of points in $\mathbb{R}^K$ specified by the columns of*

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

where $\boldsymbol{I}_K \in \mathbb{R}^{K\times K}$ is the identity matrix, and $\mathbf{1}_K \in \mathbb{R}^K$ is the all ones vector. In the other words, we also have

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

As in [5, 12], in this paper we consider general Simplex ETF as a collection of points in $\mathbb{R}^d$ specified by the columns of $\sqrt{\frac{K}{K-1}}\boldsymbol{P}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$, where $\boldsymbol{P} \in \mathbb{R}^{d\times K}(d \geq K)$ is an orthonormal matrix, i.e., $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_K$.

Lemma 1 (Young's Inequality)**.** *Let p, q be positive real numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then for any $a, b \in \mathbb{R}$, we have*

$$|ab| \;\leq\; \frac{|a|^p}{p} \;+\; \frac{|b|^q}{q},$$

where the equality holds if and only if $|a|^p = |b|^q$. The case $p = q = 2$ is just the AM-GM inequality for $a^2,\ b^2$: $|ab| \leq \frac{1}{2}\left(a^2 + b^2\right)$, where the equality holds if and only if $|a| = |b|$.

The following Lemma extends the standard variational form of the nuclear norm.

Lemma 2. *For any fixed $\boldsymbol{W} \in \mathbb{R}^{K\times d}$, $\boldsymbol{H}_i \in \mathbb{R}^{d\times K}$, $\bar{\boldsymbol{Z}}_i = \boldsymbol{W}\boldsymbol{H}_i \in \mathbb{R}^{K\times K}$ and $\alpha > 0$, we have*

$$\left\|\bar{\boldsymbol{Z}}_i\right\|_* \;\leq\; \frac{1}{2\sqrt{\alpha}}\left(\|\boldsymbol{W}\|_F^2 + \alpha\left\|\boldsymbol{H}_i\right\|_F^2\right). \tag{11}$$

Here, $\left\|\bar{\boldsymbol{Z}}_i\right\|_*$ denotes the nuclear norm of $\bar{\boldsymbol{Z}}_i$:

$$\left\|\bar{\boldsymbol{Z}}_i\right\|_* \;:=\; \sum_{k=1}^{K}\sigma_k(\bar{\boldsymbol{Z}}_i) = \operatorname{trace}(\boldsymbol{\Sigma}), \quad \textit{with} \quad \bar{\boldsymbol{Z}}_i \;=\; \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where $\{\sigma_k\}_{k=1}^K$ denotes the singular values of $\bar{\boldsymbol{Z}}_i$, and $\bar{\boldsymbol{Z}}_i = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the singular value decomposition (SVD) of $\bar{\boldsymbol{Z}}_i$.

Proof of Lemma 2. Let $\bar{\boldsymbol{Z}}_i = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of $\bar{\boldsymbol{Z}}_i$. For any $\boldsymbol{W}\boldsymbol{H}_i = \bar{\boldsymbol{Z}}_i$, we have

$$\begin{aligned}\left\|\bar{\boldsymbol{Z}}_i\right\|_* \;&=\; \operatorname{trace}(\boldsymbol{\Sigma}) \;=\; \operatorname{trace}\left(\boldsymbol{U}^\top\bar{\boldsymbol{Z}}_i\boldsymbol{V}\right) \;=\; \operatorname{trace}\left(\boldsymbol{U}^\top\boldsymbol{W}\boldsymbol{H}_i\boldsymbol{V}\right)\\ &\leq\; \frac{1}{2\sqrt{\alpha}}\left\|\boldsymbol{U}^\top\boldsymbol{W}\right\|_F^2 + \frac{\sqrt{\alpha}}{2}\left\|\boldsymbol{H}_i\boldsymbol{V}\right\|_F^2 \;\leq\; \frac{1}{2\sqrt{\alpha}}\left(\|\boldsymbol{W}\|_F^2 + \alpha\left\|\boldsymbol{H}_i\right\|_F^2\right),\end{aligned}$$

where the first inequality utilize the Young's inequality in Lemma 1 that $|\operatorname{trace}(\boldsymbol{A}\boldsymbol{B})| \leq \frac{1}{2c}\|\boldsymbol{A}\|_F^2 + \frac{c}{2}\|\boldsymbol{B}\|_F^2$ for any $c > 0$ and $\boldsymbol{A}, \boldsymbol{B}$ of appropriate dimensions, and the last inequality follows because $\|\boldsymbol{U}\| = 1$ and $\|\boldsymbol{V}\| = 1$. Therefore, we have

$$\left\|\bar{\boldsymbol{Z}}_i\right\|_* \;\leq\; \frac{1}{2\sqrt{\alpha}}\left(\|\boldsymbol{W}\|_F^2 + \alpha\left\|\boldsymbol{H}_i\right\|_F^2\right).$$

We complete the proof. □

Lemma 3 (Eigenvalues of Diagonal-Plus-Rank-One Matrices)**.** *Let $\tau < 0$, $\boldsymbol{z} \in \mathbb{R}^n$, and $\boldsymbol{D}$ be an $n \times n$ diagonal matrix with diagonals $d_1, \ldots, d_n$. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the diagonal-plus-rank-one matrix $\boldsymbol{D} + \tau \boldsymbol{z}\boldsymbol{z}^\top$.*

- *Case 1: If $d_1 > d_2 > \cdots > d_n$ and $z_i \neq 0$ for all $i = 1, \cdots, n$, then the eigenvalues $\{\lambda_i\}$ are equal to the n roots of the rational function [65, 66]*

$$w(\lambda) = 1 + \tau \boldsymbol{z}^\top (\boldsymbol{D} - \lambda \boldsymbol{I})^{-1} \boldsymbol{z} = 1 + \tau \sum_{j=1}^{n} \frac{z_j^2}{d_j - \lambda},$$

and the diagonals $\{d_i\}$ strictly separate the eigenvalues as following:

$$d_1 > \lambda_1 > d_2 > \lambda_2 > \cdots > d_n > \lambda_n. \tag{12}$$

- *Case 2:If $z_i = 0$ for some i, then d_i is an eigenvalue of $\boldsymbol{D} + \tau \boldsymbol{z}\boldsymbol{z}^\top$ with corresponding eigenvector $\boldsymbol{e}_i$ since*

$$(\boldsymbol{D} + \tau \boldsymbol{z}\boldsymbol{z}^\top)\boldsymbol{e}_i = d_i \boldsymbol{e}_i + \tau \boldsymbol{z} z_i = d_i \boldsymbol{e}_i.$$

The remaining $n-1$ eigenvalues of $\boldsymbol{D} + \tau \boldsymbol{z}\boldsymbol{z}^\top$ are equal to the eigenvalues of the smaller matrix $\boldsymbol{D}' + \tau \boldsymbol{z}'\boldsymbol{z}'^\top$, where $\boldsymbol{D}' \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\boldsymbol{z}' \in \mathbb{R}^{n-1}$ are obtained by removing the i-th rows and columns from $\boldsymbol{D}$ and the i-th element from $\boldsymbol{z}$, respectively. One can repeat this process if $\boldsymbol{z}'$ still has zero element.

- *Case 3: If there are m mutually equal diagonal elements, say $d_{i+1} = \cdots = d_{i+m} = d$, then for any orthogonal $m \times m$ matrix $\boldsymbol{P}$, $\boldsymbol{D} + \tau \boldsymbol{z}\boldsymbol{z}^\top$ has the same eigenvalues as*

$$\boldsymbol{T}\boldsymbol{D}\boldsymbol{T}^\top + \tau(\boldsymbol{T}\boldsymbol{z})(\boldsymbol{T}\boldsymbol{z})^\top = \boldsymbol{D} + \tau \widehat{\boldsymbol{z}}\widehat{\boldsymbol{z}}^\top, \text{ where } \boldsymbol{T} = \begin{bmatrix} \boldsymbol{I}_i & & \\ & \boldsymbol{P} & \\ & & \boldsymbol{I}_{n-i-m} \end{bmatrix}, \widehat{\boldsymbol{z}} = \boldsymbol{T}\widehat{\boldsymbol{z}}.$$

We can then choose $\boldsymbol{P}$ as a Householder transformation such that

$$\boldsymbol{P}\begin{bmatrix} z_{i+1} & z_{i+2} & \cdots & z_{i+m} \end{bmatrix}^\top = \begin{bmatrix} 0 & 0 & \cdots & \sqrt{\sum_{j=i+1}^{i+m} z_j^2} \end{bmatrix}^\top.$$

Thus, according to Case 2, d is an eigenvalue of $\boldsymbol{D} + \tau \widehat{\boldsymbol{z}}\widehat{\boldsymbol{z}}^\top$ repeated $m-1$ times and the remaining eigenvalues can be computed by checking the smaller matrix.

Based on Lemma 3, we can prove the following Lemma.

Lemma 4. *Let $K \geq 3$ and $\boldsymbol{Z} = -\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}\boldsymbol{1}^\top\right) diag\left(\rho_1, \rho_2, \cdots, \rho_K\right)$ with $|\rho_1| \geq |\rho_2| \geq \cdots \geq |\rho_K|$ and $|\rho_1| > 0$. Also let $\sigma_i \geq 0$ be the i-th largest singular value of $\boldsymbol{Z}$. Suppose there exists k with $1 \leq k \leq K-1$ such that*

$$\sigma_1 = \cdots = \sigma_k = \sigma_{\max} > 0 \text{ and } \sigma_{k+1} = \cdots = \sigma_K = 0. \tag{13}$$

Then $|\rho_1|, \cdots, |\rho_K|$ must satisfy either

$$|\rho_1| = |\rho_2| = \cdots = |\rho_K|, \quad \text{with} \quad \sigma_{\max} = |\rho_1|,$$

or

$$\rho_2 = \cdots = \rho_K = 0, \quad \text{with} \quad \sigma_{\max} = \sqrt{\frac{K-1}{K}}|\rho_1|.$$

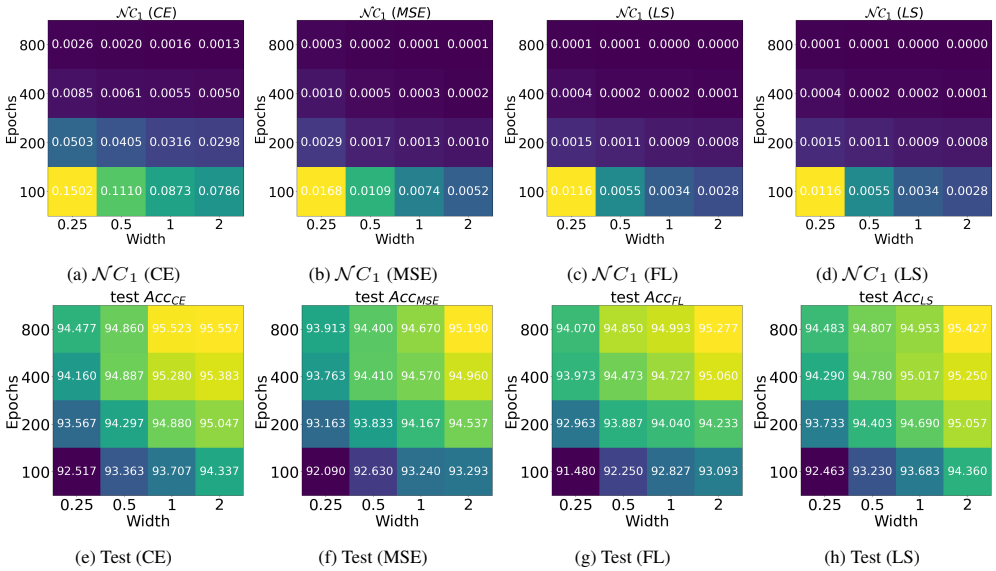
Proof of Lemma 4. Because

$$\begin{aligned} \boldsymbol{Z}^\top \boldsymbol{Z} &= \operatorname{diag}\left(\rho_1, \rho_2, \cdots, \rho_K\right)\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}\boldsymbol{1}^\top\right)\operatorname{diag}\left(\rho_1, \rho_2, \cdots, \rho_K\right) \\ &= \operatorname{diag}\left(\rho_1^2, \rho_2^2, \cdots, \rho_K^2\right) - \frac{1}{K}\boldsymbol{\rho}\boldsymbol{\rho}^\top \end{aligned}$$

where $\boldsymbol{\rho} = \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_K \end{bmatrix}^\top$, $\boldsymbol{Z}^\top \boldsymbol{Z}$ satisfies the form of Diagonal-Plus-Rank-One in Lemma 3 with $\boldsymbol{D} = \operatorname{diag}\left(\rho_1^2, \rho_2^2, \cdots, \rho_K^2\right)$, $\boldsymbol{z} = \boldsymbol{\rho}$ and $\tau = -\frac{1}{K}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_K \geq 0$ denote the n eigenvalues of $\boldsymbol{Z}^\top \boldsymbol{Z}$. Due to $\boldsymbol{1}^\top \boldsymbol{Z} = \boldsymbol{0}^\top$, we can have $\lambda_K = 0$.

- If $|\rho_1| = |\rho_2| = \cdots = |\rho_K|$: we have
$$\rho_1^2 = \lambda_1 = \cdots = \lambda_{K-1} = \rho_K^2 > \lambda_K = 0.$$
Thus, $\sigma_{\max} = \sqrt{\lambda_1} = |\rho_1|$.
- If $|\rho_1| > |\rho_2| = \cdots = |\rho_K| = 0$: according to Case 2 in Lemma 3, we have
$$\lambda_1 = (1 - 1/K)\,\rho_1^2 > \rho_2^2 = \lambda_2 \cdots = \rho_K^2 = \lambda_K = 0.$$
Thus, $\sigma_{\max} = \sqrt{(1-1/K)\,\rho_1^2} = \sqrt{(K-1)/K}|\rho_1|$.
- If $|\rho_1| > |\rho_2| = \cdots = |\rho_K| \neq 0$: according to Case 3 in Lemma 3, we have
$$\lambda_2 \cdots = \lambda_{K-1} = \rho_2^2$$
and the remaining two eigenvalues are the same to those of $\begin{bmatrix}\rho_1^2 & \\ & \rho_K^2\end{bmatrix} + (-\frac{1}{K})\begin{bmatrix}\rho_1 \\ \sqrt{K-1}\rho_K\end{bmatrix}\begin{bmatrix}\rho_1 & \sqrt{K-1}\rho_K\end{bmatrix}$. According to (12) in Lemma 3, we can obtain
$$\rho_1^2 > \lambda_1 > \rho_K^2 > \lambda_K = 0.$$
Combing them together, we can have
$$\rho_1^2 > \lambda_1 > \rho_2^2 = \lambda_2 \cdots = \rho_K^2 > \lambda_K = 0$$
thus, $0 = \lambda_K < \lambda_2 < \lambda_1 = \lambda_{\max}$, which violates the assumption (13).
- If $|\rho_1| = \cdots = |\rho_i| > |\rho_{i+1}| = \cdots = |\rho_K| = 0$ and $1 < i < K$: according to the Case 2 and Case 3 in Lemma 3, we can have
$$\lambda_1 = \cdots = \lambda_{i-1} = \rho_1^2$$
$$\lambda_{i+1} = \cdots = \lambda_K = 0$$
and $0 < \lambda_i = \rho_1^2 - \frac{i}{K}\rho_1^2 < \rho_1^2 = \lambda_{\max}$, which violates the assumption (13).
- If $|\rho_1| = \cdots = |\rho_i| > |\rho_{i+1}| = \cdots = |\rho_K| \neq 0$ and $1 < i < K$: according to Case 3 in Lemma 3, we have
$$\lambda_1 = \cdots = \lambda_{i-1} = \rho_1^2$$
$$\lambda_{i+1} = \cdots = \lambda_{K-1} = \rho_K^2$$
and the remaining two eigenvalues are the same to those of $\boldsymbol{D} = \begin{bmatrix}\rho_1^2 & \\ & \rho_K^2\end{bmatrix} + (-\frac{1}{K})\begin{bmatrix}\sqrt{i}\rho_1 \\ \sqrt{K-i}\rho_K\end{bmatrix}\begin{bmatrix}\sqrt{i}\rho_1 & \sqrt{K-i}\rho_K\end{bmatrix}$. According to (12) in Lemma 3, we can obtain
$$\rho_1^2 = \rho_i^2 > \lambda_i > \rho_K^2 > \lambda_K = 0.$$
Combing them together, we can have
$$\rho_1^2 = \lambda_1 = \cdots = \rho_i^2 > \lambda_i > \rho_{i+1}^2 = \lambda_{i+1} = \cdots = \rho_K^2 > \lambda_K = 0$$
thus, $0 = \lambda_K < \lambda_i < \lambda_1 = \lambda_{\max}$, which violates the assumption (13).
- If $|\rho_1| > |\rho_i| > |\rho_K|$ for some $1 < i < K$: Suppose $|\rho_1| = \cdots = |\rho_m|$, $|\rho_i| = \cdots = |\rho_{i+n-1}|$ and $|\rho_{K-t+1}| = \cdots = |\rho_K|$, where $m < i$, $i+n-1 < K-t+1$ and $m, n, t \geq 1$. According to the (12), Case 2 and Case 3 in Lemma 3, we can find
$$\rho_m^2 > \lambda_m > \rho_i^2 \geq \lambda_{i+n-1} > \rho_K^2 \geq \lambda_K = 0$$
thus, $0 = \lambda_K < \lambda_{i+n-1} < \lambda_m \leq \lambda_{\max}$, which violates the assumption (13).

We complete the proof. □



(a) $\mathcal{NC}_1$ (CE) (b) $\mathcal{NC}_1$ (MSE) (c) $\mathcal{NC}_1$ (FL) (d) $\mathcal{NC}_1$ (LS)

(e) Test (CE) (f) Test (MSE) (g) Test (FL) (h) Test (LS)

Figure 4: **Illustration of $\mathcal{NC}_1$ and test accuracy across different iterations-width configurations.** The figure depicts the $\mathcal{NC}_1$ and test accuracy of various iteration-width configurations for different loss functions on CIFAR10.

A Experiments

In this section, we first describe more details about the datasets and the computational resource used in the paper. Particularly, all CIFAR10, CIFAR100 and miniImageNet are publicly available for academic purpose under the MIT license, and we run all experiments on a single RTX3090 GPU with 24GB memory. Moreover, additional experimental results on CIFAR10, CIFAR100 and miniImageNet are presented in Section A.1, Section A.2, and Section A.3, respectively.

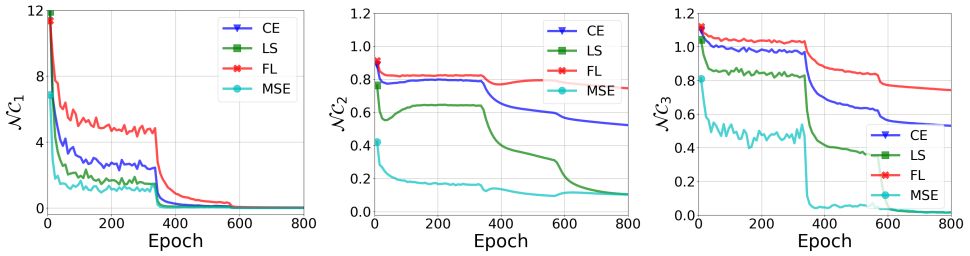
A.1 Additional experimental results on CIFAR10

In Section 4, we present the test accuracy for different losses function across various different iteration-width configurations. Moreover, we further show the $\mathcal{NC}_1$ for different loss functions across different iteration-width configurations , and we reuse the results of test accuracy in Figure 3 for better investigation. The experiment results in Figure 4 consistently show that the value of $\mathcal{NC}_1$ of training WideResNet50-0.25 for 100 epochs is around three orders of magnitude larger than it of training WideResNet50-2 for 800 epochs, which indicates that the previous configuration setting is much less collapsed than the latter one. In terms of test accuracy, the maximal difference across different losses for width $= 0.25$ and epochs $= 100$ configuration is 1.037%, which is larger than 0.36% for width $= 2$ and epochs $= 800$ configuration. These results support our claim that all losses lead to identical performance, as long as the network has sufficient approximation power and the number of optimization is enough for the convergence to the $\mathcal{NC}$ global optimality.

A.2 Additional experimental results on CIFAR100

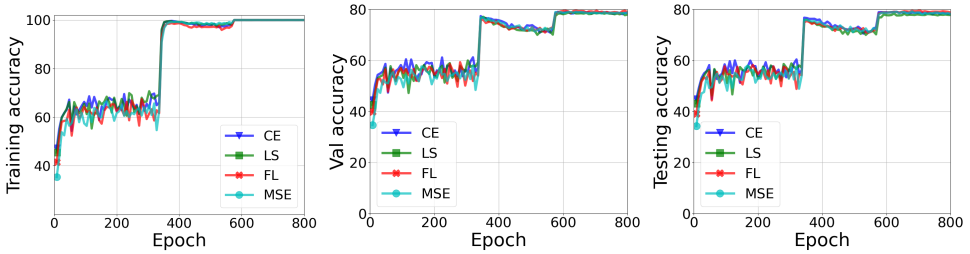
In this parts, we show the additional results on CIFAR100 dataset.

Prevalence of $\mathcal{NC}$ Across Varying Training Losses We show that all loss functions lead to $\mathcal{NC}$ solutions during the terminal phase of training on CIFAR100 dataset. The results on CIFAR100 using WideResNet50-2 and different loss functions is provided in Figure 5. We consistently observe that all three $\mathcal{NC}$ metrics of FL and MSE converge to a small value as training progresses, and metrics of CE and FL still continue to decrease at the last iteration, because CIFAR100 is more difficult than CIFAR10 and requires networks to be optimized longer. The decreasing speed of FL is slowest, which is consistent with our global landscape analysis that FL has benign landscape in



(a) $\mathcal{NC}_1$ (CIFAR100) (b) $\mathcal{NC}_2$ (CIFAR100) (c) $\mathcal{NC}_3$ (CIFAR100)

Figure 5: **The evolution of $\mathcal{NC}$ metrics across different loss functions.** We train the WideResNet50-2 on CIFAR100 dataset for 800 epochs using different loss function. From left to right: NC_1 (variability collapse), NC_2 (convergence to simplex ETF) and NC_3 (convergence to self-duality).



(a) Train Acc (CIFAR100) (b) Val Acc (CIFAR100) (c) Test Acc (CIFAR100)

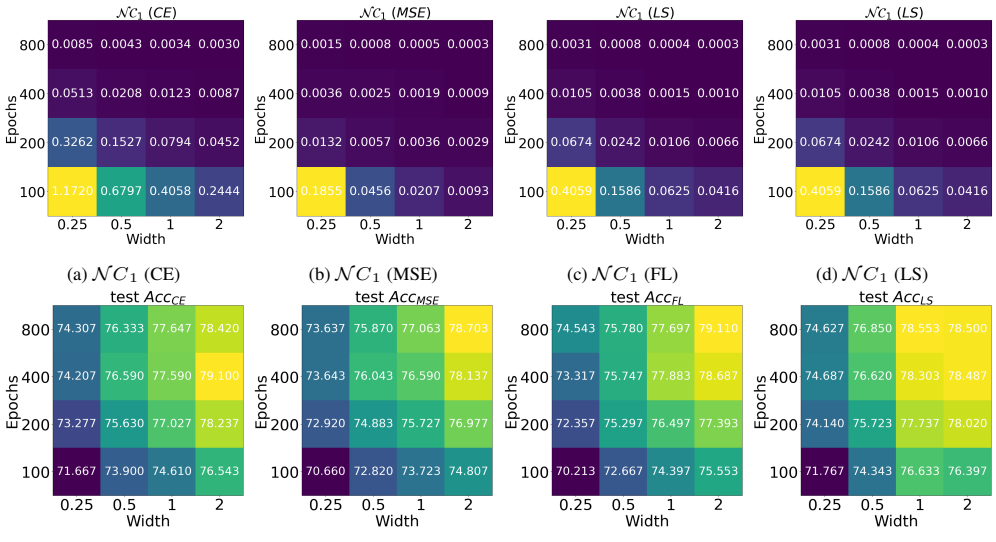
Figure 6: **The evolution of performance across different loss functions.** We train the WideResNet50-2 on CIFAR100 dataset for 800 epochs using different loss function. From left to right: training accuracy, validation accuracy and test accuracy.

the local region near optimality. These results imply that all losses exhibit $\mathcal{NC}$ at the end, regardless of the choice of loss functions.

All Losses Lead to Largely Identical Performance Same as the results on CIFAR10 dataset, the conclusion on CIFAR100 also holds that all loss functions have largely identical performance once the training procedure converges to the $\mathcal{NC}$ global optimality. In Figure 6, we plot the evolution of the training accuracy, validation accuracy and test accuracy with training progressing, where all losses are optimized on the same WideResNet50-2 architecture and CIFAR100 for 800 epochs. To reduce the randomness, we average the results from 3 different random seeds per iteration-width configuration, and the test accuracy is reported based on the model with best accuracy on validation set, where we organize the validation set by holding out 10 percent data from the training set. The results consistently shows that the training accuracy trained by different losses all converge to one hundred percent (reaching to terminal phase), and the validation accuracy and test accuracy across different losses are largely same, as long as the optimization procedure converges to the $\mathcal{NC}$ global solution. In Figure 7, we plot the average $\mathcal{NC}_1$ and test accuracy of different losses under different pairs of width and iterations for CIFAR100 dataset. The three phenomenon mentioned in Section 4.2 also exist on CIFAR100 in most cases. Moreover, the values of $\mathcal{NC}_1$ for width=0.25 and epochs=100 configuration are also around three orders magnitude larger than them for width=2 and epochs=800 configuration and the less collapsed configuration leads to larger difference gap across different loss functions. While there are some small difference between different losses in width $= 2$ and epochs $= 800$ configurations, We guess that it is because CIFAR100 is much harder than CIFAR10 datasets, and network is not sufficiently large and trained not long enough for all losses to achieve a global solution.

A.3 Additional experimental results on miniImageNet

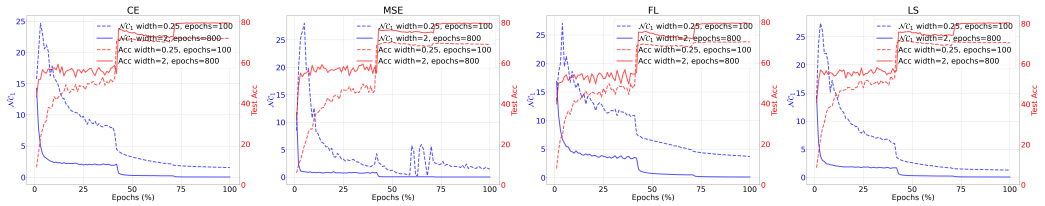
In this parts, we show the additional results on miniImageNet dataset. We trained WideResNet18-0.25 and WideResNet18-2 on miniImageNet for 100 epochs and 800 epochs, respectively. To reduce the randomness, we average the results from 3 different random trials. The $\mathcal{NC}_1$ and test accu-



(a) $\mathcal{NC}_1$ (CE) (b) $\mathcal{NC}_1$ (MSE) (c) $\mathcal{NC}_1$ (FL) (d) $\mathcal{NC}_1$ (LS)

(e) Test (CE) (f) Test (MSE) (g) Test (FL) (h) Test (LS)

Figure 7: **Illustration of $\mathcal{NC}_1$ and test accuracy across different iterations-width configurations.** The figure depicts the $\mathcal{NC}_1$ and test accuracy of various iteration-width configurations for different loss functions on CIFAR100.



(a) $\mathcal{NC}_1$ and Test (CE) (b) $\mathcal{NC}_1$ and Test (MSE) (c) $\mathcal{NC}_1$ and Test (FL) (d) $\mathcal{NC}_1$ and Test (LS)

Figure 8: **The evolution of $\mathcal{NC}_1$ and test accuracy across different loss functions.** We train the WideResNet18-0.25 for 100 epochs and WideResNet18-2 for 800 epochs on miniImageNet using different loss functions.

racy of different loss functions are provided in Figure 8 for comparison. We consistently observe that the $\mathcal{NC}_1$ metric of all losses converges to a small value as training progress, when the neural network has sufficient approximation power and the training is performed for sufficiently many iterations, such as WideResNet18-2 for 800 epochs. Additionally, the conclusion on miniImageNet also holds that all loss functions have largely identical performance once the training procedure converges to the $\mathcal{NC}$ global optimality. Specifically, while the last-iteration test accuracy of training WideResNet18-0.25 for 100 epochs is 0.7195, 0.6915, 0.7020 and 0.7040, respectively, the last-iteration test accuracy of training WideResNet18-2 for 800 epochs is 0.7930, 0.7962, 0.7932 and 0.8020 for CE, MSE, FL and LS, respectively. The experiment results on miniImageNet also support our claim that (i) the test performance may be different across different loss functions when the network is not large enough and is optimized with limited number of iterations, but (ii) the test accuracy across different loss are largely identical, once the networks has sufficient capacity and the training is optimized to converge to the $\mathcal{NC}$ global solution.

B Proof of CE, FL and LS included in GL

In this section, we prove that CE, FL and LS belong to GL in Section B.1, Section B.2 and Section B.3, respectively. Before starting the proof for each loss, let us restate the definition of the GL in Definition 1:

Definition 3 (Contrastive property)**.** *We say a loss function $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ satisfies the contrastive property if there exists a function ϕ such that $\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k)$ can be lower bounded by*

$$\mathcal{L}_{\mathrm{GL}}(\boldsymbol{z}, \boldsymbol{y}_k) \geq \phi\left(\sum_{j \neq k}(z_j - z_k)\right) \tag{14}$$

where the equality holds only when $z_j = z'_j$ for all $j, j' \neq k$. Moreover, $\phi(t)$ satisfies

$$t^* = \arg\min_t \phi(t) + c|t| \text{ is unique for any } c > 0, \text{and } t^* \leq 0. \tag{15}$$

B.1 CE is in GL

In this section, we will show that the CE defined in (3) belongs to the GL defined in Definition 3. First, let us rewrite the CE definition in GL form as following:

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) &= -\log\left(\frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)}\right) = \log\left(1 + \sum_{j \neq k}^K \exp(z_j - z_k)\right) \\ &\geq \log\left(1 + (K-1)\exp\left(\frac{z_j - z_k}{K-1}\right)\right) = \phi_{\mathrm{CE}}\left(\sum_{j \neq k}(z_j - z_k)\right).\end{aligned}$$

where the inequality is due to the $\log$ is an increasing and function and $\exp$ is a strictly convex function, and it achieves equality only when $z_j = z_{j'}$ for all $j, j' \neq k$. Therefore, there exists such a function ϕ_{CE} to lower bound original CE loss $\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k)$ as following:

$$\phi_{\mathrm{CE}}(t) = \log\left(1 + (K-1)\exp\left(\frac{t}{K-1}\right)\right),$$

which satisfies the condition of (14). Next, we will show $\phi_{\mathrm{CE}}(t)$ satisfies the condition (15). The first-order gradient of $\phi_{\mathrm{CE}}(t)$ is following:

$$\nabla\phi_{\mathrm{CE}}(t) = \frac{\exp\left(\frac{t}{K-1}\right)}{1 + (K-1)\exp\left(\frac{t}{K-1}\right)}$$

which is an increasing function and greater than 0 for $t \in \mathbb{R}$. Let denote $\psi_{\mathrm{CE}}(t) = \phi_{\mathrm{CE}}(t) + c|t|$, then

- **When $t \geq 0$:** $\nabla\psi_{\mathrm{CE}}(t) = \nabla\phi_{\mathrm{CE}}(t) + c > 0$, thus the $\psi_{\mathrm{CE}}(t)$ is an increasing function w.r.t. t, and the minimizer is achieved when $t = 0$.
- **When $t \leq 0$:** $\nabla\psi_{\mathrm{CE}}(t) = \nabla\phi_{\mathrm{CE}}(t) - c$, and $\nabla\phi_{\mathrm{CE}}(t)$ is an increasing function, which achieves minimizer when $t = 0$ such that $\nabla\phi_{\mathrm{CE}}(t) = \frac{1}{K}$.
  - if $c \geq \frac{1}{K}$, $\nabla\psi_{\mathrm{CE}}(t) < 0$, and $\psi(t)$ is a decreasing function for $t \leq 0$, and the minimizer is achieved when $t = 0$;
  - if $0 < c \leq \frac{1}{K}$, there exist such t^* such that $\nabla\psi_{\mathrm{CE}}(t) = 0$. When $t < t^*$, $\phi_{\mathrm{CE}}(t)$ is a decreasing function; and when $t^* < t \leq 0$, $\phi_{\mathrm{CE}}(t)$ is an increasing function. Therefore, the minimizer is achieved when $t = t^* < 0$

Combing them together, we can prove that ϕ_{CE} satisfies the condition of (15).

B.2 FL is in GL

In this section, we will show that the FL defined in (4) belongs to the GL defined in Definition 3. let us rewrite the FL definition in GL form as following:

$$\begin{aligned}
\mathcal{L}_{\mathrm{FL}}(\boldsymbol{z}, \boldsymbol{y}_k) &= -\left(1-\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}\right)^{\gamma}\log\left(\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}\right) \\
&= \left(1-\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}\right)^{\gamma}\log\left(\sum_{j=1}^{K}\exp(z_j-z_k)\right) \\
&= \left(1-\frac{1}{1+\sum_{j\neq k}^{K}\exp(z_j-z_k)}\right)^{\gamma}\log\left(1+\sum_{j\neq k}^{K}\exp(z_j-z_k)\right) \\
&= \eta\left(1+\sum_{j\neq k}^{K}\exp(z_j-z_k)\right)
\end{aligned}$$

where the function $\eta(t) = (1-\frac{1}{t})^{\gamma}\log(t)$ is an increasing function for $t \geq 1$ because

$$\nabla\eta(t) = \gamma(\frac{1}{t^2})(1-\frac{1}{t})^{\gamma-1}\log(t) + \frac{1}{t}(1-\frac{1}{t})^{\gamma} > 0$$

Thus, we can find the lower bound function by

$$\begin{aligned}
\mathcal{L}_{\mathrm{FL}}(\boldsymbol{z}, \boldsymbol{y}_k) &\geq \eta\left(1+(K-1)\exp\left(\sum_{j\neq k}^{K}\frac{z_j-z_k}{K-1}\right)\right) \\
&= \eta\left(\xi\left(\sum_{j\neq k}^{K}(z_j-z_k)\right)\right) \\
&= \phi_{\mathrm{FL}}\left(\sum_{j\neq k}^{K}(z_j-z_k)\right)
\end{aligned}$$

where $\phi_{\mathrm{FL}}(t) = \eta(\xi(t))$ and $\xi(t) = 1+(K-1)\exp\frac{t}{K-1} \in [1, K]$, which satisfies the condition of (14). Next, we will show $\phi_{\mathrm{FL}}(t)$ satisfies the condition (15). The first-order gradient of $\phi_{\mathrm{FL}}(t)$ is following:

$$\begin{aligned}
&\nabla_t\psi_{\mathrm{FL}}(t) = \nabla_t\left(\phi_{\mathrm{FL}}(t)+c|t|\right) = \nabla_{\xi(t)}\eta(\xi(t))\,\nabla_t\xi(t) + c\frac{t}{|t|} \\
&= \left(\gamma\left(\frac{1}{\xi(t)}\right)^2\left(1-\frac{1}{\xi(t)}\right)^{\gamma-1}\log(\xi(t)) + \frac{1}{\xi(t)}\left(1-\frac{1}{\xi(t)}\right)^{\gamma}\right)\left(\exp\left(\frac{t}{K-1}\right)\right) + c\frac{t}{|t|} \\
&= \left(\gamma\left(\frac{1}{\xi(t)}\right)^2\left(1-\frac{1}{\xi(t)}\right)^{\gamma-1}\log(\xi(t)) + \frac{1}{\xi(t)}\left(1-\frac{1}{\xi(t)}\right)^{\gamma}\right)\left(\frac{\xi(t)-1}{K-1}\right) + c\frac{t}{|t|} \\
&= \frac{1}{K-1}\underbrace{\frac{(\xi(t)-1)^{\gamma}}{\xi(t)^{\gamma+1}}\left(\xi(t)-1+\gamma\log(\xi(t))\right)}_{\varsigma(\xi(t))\geq 0} + c\frac{t}{|t|}
\end{aligned}$$

Similarly, by chain rule, the second-order derivation is:

$$
\begin{aligned}
&\nabla_t^2\psi(t) = \nabla_t^2\phi(t) = \nabla_{\xi(t)}\varsigma\left(\xi(t)\right)\nabla_t(t) \\
=&(\gamma+1)\frac{1}{\left(\xi(t)\right)^2}(1-\frac{1}{\xi(t)})^{\gamma} \\
&-\frac{\gamma}{\left(\xi(t)\right)^2}(1-\frac{1}{\xi(t)})^{\gamma}\left(\log(\xi(t))-\gamma\frac{\log\left(\xi(t)\right)}{\xi(t)-1}-\gamma\right)\left(\frac{1}{(K-1)^2}(\xi(t)-1)\right) \\
=&\frac{1}{(K-1)^2}\frac{\gamma(\xi(t)-1)^{\gamma+1}}{\left(\xi(t)\right)^{\gamma+2}}\left(\underbrace{-\log(\xi(t))+\gamma\frac{\log(\xi(t))}{\xi(t)-1}+\gamma+\frac{\gamma+1}{\gamma}}_{\vartheta(\xi(t))}\right)
\end{aligned}
$$

- When $t \geq 0$: $\nabla_t\psi_{\text{FL}}(t) = \frac{1}{K-1}\xi_(t) + c \geq 0$, thus the $\psi_{\text{CE}}(t)$ is an increasing function w.r.t. t, and the minimizer is achieved when $x = 0$.

- When $t \leq 0$: $\nabla_t\psi_{\text{FL}}(t) = \frac{1}{K-1}\xi(t) - c \geq 0$. Moreover, we can find $\vartheta(\xi(t))$ is a decreasing function w.r.t. $\xi(t)$ and $\xi(t)$ is an increasing function w.r.t. t, therefore, $\vartheta(\xi(t))$ is a decreasing function w.r.t. t.

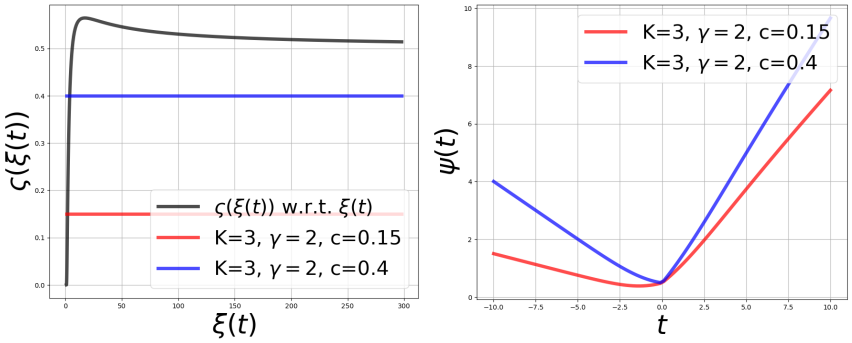
  - If $\vartheta(\xi(0)) = \vartheta(K) \geq 0$, then $\nabla_x^2\psi(x) > 0$ for $x \leq 0$, which means that $\nabla_x\xi(t)$ is an increasing function. Because $\varsigma(\xi(-\infty)) = \varsigma(1) = 0$, here we need to consider two cases(Please refer to Figure 9):

    * if $\varsigma(\xi(0) = \varsigma(K) \leq c(K-1)$, then $\nabla_t\psi_{FL}(t) \geq 0$, that is, $\psi_{FL}(t)$ is a decreasing function. Therefore, the global minimizer is achieved when $x = 0$ (the blue curve in Figure 9).
    * if $\varsigma(\xi(0) = \varsigma(K) \geq c(K-1)$, so $\psi_{FL}(x)$ will first decrease and then increase. Therefore the global minimizer is unique (the red curve in Figure 9).

  - If $\vartheta(\xi(0)) = \vartheta(K) < 0$, then for $t \in [-\infty, t']$, $\nabla_t\psi_{FL}(x)$ is an increasing function w.r.t. t; for $t \in [t', 0)$, $\nabla_t\Phi_{FL}(t)$ is a decreasing function w.r.t. t. Here we need to consider three cases(please refer to Figure 10):

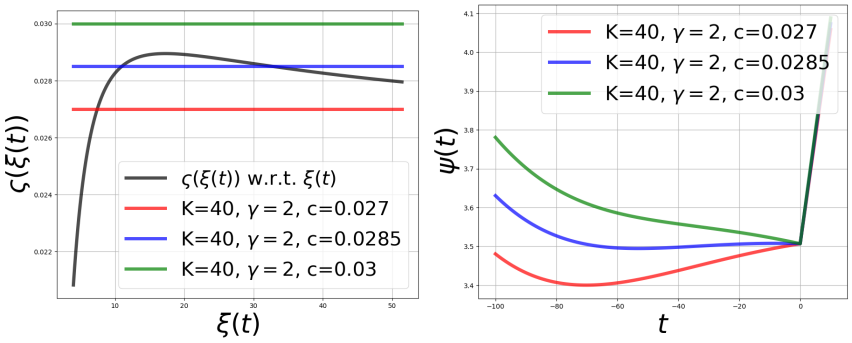
    * if $\varsigma(\xi(t')) \leq c(K-1)$, then $\nabla_t\psi_{FL}(t) \leq 0$, that is, $\psi_{FL}(t)$ is a decreasing function. Therefore, the global minimizer is achieved when $x = 0$ (the green curve in Figure 10).
    * if $\varsigma(\xi(0)) = \varsigma(K) \geq c(K-1)$, so $\psi_{FL}(x)$ will first decrease and then increase. Therefore the global minimizer is unique (the red curve in Figure 10).
    * if $\varsigma(\xi(t')) \geq c(K-1)$ and $\varsigma(\xi(0)) = \varsigma(K) \leq c(K-1)$, then $\nabla_t\psi_{FL}(t) = 0$ has two solutions t_1 and t_2. For $t \in [-\infty, t_1]$, $\psi_{FL}(t)$ is an decreasing function w.r.t. t; for $t \in [t_1, t_2]$, $\Phi_{FL}(t)$ is an increasing function w.r.t. t; and for $t \in [t_2, 0)$, $\psi_{FL}(t)$ is a decreasing function w.r.t. t. The unique minimizer is achieved when either $t = 0$ or $t = t_1$, as long as $\psi_{FL}(0) \neq \psi_{FL}(t_1)$. As for the minor case $\psi_{FL}(0) = \psi_{FL}(t_1)$, it requires carefully chosen penalized parameters, which can be omitted (the blue curve in Figure 10).

In conclusion, for focal loss, $\psi_{FL}(t)$ has a unique minimum in terms of $t \leq 0$, which satisfies the condition of (15).



(a) $\varsigma(\xi(t))$ w.r.t. $\xi(t)$ (b) $\psi(t)$ w.r.t. t

Figure 9: **Illustration of the case of $\vartheta(\xi(0)) \geq 0$, where $c = -K\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}$.**



(a) $\varsigma(\xi(t))$ w.r.t. $\xi(t)$ (b) $\psi(t)$ w.r.t. t

Figure 10: **Illustration of the case of $\vartheta(\xi(0)) < 0$, where $c = K\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}$.**

B.3 LS is in GL

In this section, we will show that the LS defined in (5) belongs to the GL defined in Definition 3. First, let us rewrite the LS definition in GL form as following:

$$\begin{aligned}
\mathcal{L}_{\mathrm{LS}}(\boldsymbol{z}, \boldsymbol{y}_k) &= -\left(1 - \frac{(K-1)\alpha}{K}\right) \log\left(\frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)}\right) - \frac{\alpha}{K}\sum_{\ell\neq k}^K \log\left(\frac{\exp(z_\ell)}{\sum_{j=1}^K \exp(z_j)}\right) \\
&= \left(1 - \frac{(K-1)\alpha}{K}\right) \log\left(\frac{\sum_{j=1}^K \exp(z_j)}{\exp(z_k)}\right) + \frac{\alpha}{K}\sum_{\ell\neq k}^K \log\left(\frac{\sum_{j=1}^K \exp(z_j)}{\exp(z_\ell)}\right) \\
&= \left(1 - \frac{(K-1)\alpha}{K}\right) \log\left(\sum_{j=1}^K \exp(z_j - z_k))\right) + \frac{\alpha}{K}\sum_{\ell\neq k}^K \log\left(\frac{\sum_{j=1}^K \exp(z_j - z_k)}{\exp(z_\ell - z_k)}\right) \\
&= \log\left(\sum_{j=1}^K \exp(z_j - z_k))\right) - \frac{\alpha}{K}\sum_{\ell\neq k}^K (z_\ell - z_k) \\
&\geq \log\left(1 + (K-1)\exp\left(\frac{z_j - z_k}{K-1}\right)\right) - \frac{\alpha}{K}\sum_{\ell\neq k}^K (z_\ell - z_k)
\end{aligned}$$

where the inequality is due to the $\log$ is an increasing and function and $\exp$ is a strictly convex function, and it achieves equality only when $z_j = z_{j'}$ for all $j, j' \neq k$. Therefore, there exists such a function ϕ_{LS} to lower bound original LS loss $\mathcal{L}_{\mathrm{LS}}(\boldsymbol{z}, \boldsymbol{y}_k)$ as following:

$$\phi_{\mathrm{LS}}(t) = \log\left(1 + (K-1)\exp\left(\frac{t}{K-1}\right)\right) - \frac{\alpha}{K}t,$$

which satisfies the condition of (14). Next, we will show $\phi_{\text{LS}}(t)$ satisfies the condition (15). The first-order gradient of $\phi_{\text{LS}}(t)$ is following:

$$\nabla\phi_{\text{LS}}(t) = \frac{\exp\left(\frac{t}{K-1}\right)}{1+(K-1)\exp\left(\frac{t}{K-1}\right)} - \frac{\alpha}{K}$$

Let denote $\psi_{\text{LS}}(t) = \phi_{\text{LS}}(t) + c|t|$, then

- **When** $t \geq 0$**:** $\nabla\psi_{\text{LS}}(t) = \nabla\phi_{\text{LS}}(t) + c > 0$ due to $\nabla\phi_{\text{LS}}(t) \geq 0 \quad \text{for} \quad t > 0$, thus the $\psi_{\text{LS}}(t)$ is an increasing function w.r.t. t, and the minimizer is achieved when $x = 0$.
- **When** $t \leq 0$**:** $\nabla\psi_{\text{LS}}(t) = \nabla\phi_{\text{LS}}(t) - c$, and $\nabla\phi_{\text{LS}}(t)$ is an increasing function, which achieves minimizer when $t = 0$ such that $\phi_{\text{LS}}(t) = \frac{1-\alpha}{K}$.
  - if $c \geq \frac{1-\alpha}{K}$, $\nabla\psi_{\text{LS}}(t) < 0$, and $\psi(t)$ is a decreasing function for $t \leq 0$, and the minimizer is achieved when $t = 0$;
  - if $0 < c \leq \frac{1-\alpha}{K}$, there exist such t^* such that $\nabla\psi_{\text{LS}}(t) = 0$. When $t < t^*$, $\phi_{\text{LS}}(t)$ is a decreasing function; and when $t^* < t \leq 0$, $\phi_{\text{LS}}(t)$ is an increasing function. Therefore, the minimizer is achieved when $t = t^* < 0$

Combing them together, we can prove that ϕ_{LS} satisfies the condition of (14).

C Proof of Theorem 1 for GL

In this part of appendices, we prove Theorem 1 in Section 3 that we restate as follows.

Theorem 3 (Global Optimality Condition of GL)**.** *Assume that the number of classes K is smaller than feature dimension d, i.e., $K < d$, and the dataset is balanced for each class, $n = n_1 = \cdots = n_K$. Then any global minimizer $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ of*

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) := g(\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b}\mathbf{1}^\top) + \frac{\lambda_{\boldsymbol{W}}}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2, \tag{16}$$

with

$$g(\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b}\mathbf{1}^\top) := \sum_{i=1}^{n} g(\boldsymbol{W}\boldsymbol{H}_i+\boldsymbol{b}\mathbf{1}^\top) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}+\boldsymbol{b},\boldsymbol{y}_k); \tag{17}$$

$$\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}+\boldsymbol{b},\boldsymbol{y}_k) = \mathcal{L}(\boldsymbol{z}_{k,i},\boldsymbol{y}_k) \text{ satisfying the the } \textbf{\textit{Contrastive property}} \text{ in Definition 3}; \tag{18}$$

obeys the following

$$\|\boldsymbol{w}^\star\|_2 = \left\|\boldsymbol{w}^{\star 1}\right\|_2 = \left\|\boldsymbol{w}^{\star 2}\right\|_2 = \cdots = \left\|\boldsymbol{w}^{\star K}\right\|_2, \quad \text{and} \quad \boldsymbol{b}^\star = b^\star\mathbf{1},$$

$$\boldsymbol{h}_{k,i}^\star = \sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}} n}}\boldsymbol{w}^{\star k}, \quad \forall\, k\in[K],\ i\in[n], \quad \text{and} \quad \overline{\boldsymbol{h}}_i^\star := \frac{1}{K}\sum_{j=1}^{K}\boldsymbol{h}_{j,i}^\star = \mathbf{0}, \quad \forall\, i\in[n],$$

where either $b^\star = 0$ or $\lambda_{\boldsymbol{b}} = 0$, and the matrix $\boldsymbol{W}^{\star\top}$ is in the form of K-simplex ETF structure defined in Definition 2 in the sense that

$$\boldsymbol{W}^{\star\top}\boldsymbol{W}^\star = \|\boldsymbol{w}^\star\|_2^2\frac{K}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right).$$

C.1 Main Proof

At a high level, we lower bound the general loss function based on the contrastive property (14), then check the equality conditions hold for the lower bounds and these equality conditions ensure that the global solutions $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ are in the form as shown in Theorem 3.

Proof of Theorem 3. First by Lemma 5, Lemma 6 and Lemma 7, we know that any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of f in (16) satisfies

$$\begin{aligned}
&\boldsymbol{W}^\top \boldsymbol{W} = \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \boldsymbol{H}\boldsymbol{H}^\top; \\
&\lambda_{\boldsymbol{H}} \boldsymbol{H}_i = -\boldsymbol{W}^\top \nabla_{\boldsymbol{Z}_i = \boldsymbol{W}\boldsymbol{H}_i}\, g(\boldsymbol{W}\boldsymbol{H}_i + \boldsymbol{b}\mathbf{1}^\top); \\
&\boldsymbol{b} = -\frac{\nabla g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)}{\lambda_{\boldsymbol{b}}}\mathbf{1}.
\end{aligned}$$

For the rest of the proof, let $\boldsymbol{G}_i = \nabla_{\boldsymbol{Z}_i = \boldsymbol{W}\boldsymbol{H}_i}\, g(\boldsymbol{W}\boldsymbol{H}_i + \boldsymbol{b}\mathbf{1}^\top)$ and $\tau = -\frac{\nabla g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)}{\lambda_{\boldsymbol{b}}}$ to simplify the notations, and thus $\|\boldsymbol{H}\|_F^2 = \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \|\boldsymbol{W}\|_F^2$, $\lambda_{\boldsymbol{H}} \boldsymbol{H}_i = -\boldsymbol{W}^\top \boldsymbol{G}_i$ and $\boldsymbol{b} = \tau \mathbf{1}$.

We will first provide a lower bound for the general loss term $g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)$ according to the Definition 3, and then show that the lower bound is attained if and only if the parameters are in the form described in Theorem 3. By Lemma 8, we have

$$\begin{aligned}
f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;&=\; g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top) \;+\; \frac{\lambda_{\boldsymbol{W}}}{2} \|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2} \|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2^2 \\
&\geq\; \phi\left(\rho^\star\right) + K\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star|
\end{aligned}$$

where ϕ is lower bound function satisfying the Definition 3, $\rho^\star = \arg\min_\rho \phi(\rho) + K\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho| \leq 0$. Furthermore, by Lemma 8, we know that $\bar{\boldsymbol{Z}}_i^\star = \boldsymbol{W}^\star \boldsymbol{H}_i^\star = -\rho^\star \left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$, which satisfies the K-simplex ETF structure defined in Definition 2. In Lemma 9, we show the any minimizer $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ of $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ has following properties via check the equality conditions hold for the lower bounds in Lemma 8:

(a) $\|\boldsymbol{w}^\star\|_2 = \left\|\boldsymbol{w}^{\star 1}\right\|_2 = \left\|\boldsymbol{w}^{\star 2}\right\|_2 = \cdots = \left\|\boldsymbol{w}^{\star K}\right\|_2$;

(b) $\boldsymbol{b}^\star = b^\star \mathbf{1}$, where either $b^\star = 0$ or $\lambda_{\boldsymbol{b}} = 0$;

(c) $\overline{\boldsymbol{h}}_i^\star := \frac{1}{K}\sum_{j=1}^K \boldsymbol{h}_{j,i}^\star = \mathbf{0}, \quad \forall\, i \in [n]$, and $\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}} n}} \boldsymbol{w}^{k\star} = \boldsymbol{h}_{k,i}^\star, \quad \forall\, k \in [K],\ i \in [n]$;

(d) $\boldsymbol{W}\boldsymbol{W}^\top = \|\boldsymbol{w}^\star\|_2^2 \frac{K-1}{K}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$;

The proof is complete. □

C.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 5. *Let $\rho = \|\boldsymbol{W}\|_F^2$. Any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) obeys*

$$\boldsymbol{W}^\top \boldsymbol{W} \;=\; \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \boldsymbol{H}\boldsymbol{H}^\top \quad \textit{and} \quad \rho \;=\; \|\boldsymbol{W}\|_F^2 \;=\; \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \|\boldsymbol{H}\|_F^2 . \tag{19}$$

Proof of Lemma 5. By definition, any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) satisfies the following:

$$\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;=\; \nabla_{\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}}\, g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)\boldsymbol{H}^\top + \lambda_{\boldsymbol{W}} \boldsymbol{W} \;=\; \mathbf{0}, \tag{20}$$

$$\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;=\; \boldsymbol{W}^\top \nabla_{\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}}\, g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top) + \lambda_{\boldsymbol{H}} \boldsymbol{H} \;=\; \mathbf{0}. \tag{21}$$

Left multiply the first equation by $\boldsymbol{W}^\top$ on both sides and then right multiply second equation by $\boldsymbol{H}^\top$ on both sides, it gives

$$\begin{aligned}
\boldsymbol{W}^\top \nabla_{\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}}\, g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)\boldsymbol{H}^\top \;&=\; -\lambda_{\boldsymbol{W}} \boldsymbol{W}^\top \boldsymbol{W}, \\
\boldsymbol{W}^\top \nabla_{\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}}\, g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)\boldsymbol{H}^\top \;&=\; -\lambda_{\boldsymbol{H}} \boldsymbol{H}^\top \boldsymbol{H}.
\end{aligned}$$

Therefore, combining the equations above, we obtain

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

Moreover, we have

$$\rho \;=\; \|\boldsymbol{W}\|_F^2 \;=\; \operatorname{trace}\left(\boldsymbol{W}^\top \boldsymbol{W}\right) \;=\; \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \operatorname{trace}\left(\boldsymbol{H}\boldsymbol{H}^\top\right) \;=\; \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \operatorname{trace}\left(\boldsymbol{H}^\top \boldsymbol{H}\right) \;=\; \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \|\boldsymbol{H}\|_F^2,$$

as desired. □

Next, we characterize the following relationship per group between $\boldsymbol{W}$ and $\boldsymbol{H}_i$ for $i \in [n]$ for any critical $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) satisfies the following:

Lemma 6. *Let $\boldsymbol{G}_i = \nabla_{\boldsymbol{Z}_i = \boldsymbol{W}\boldsymbol{H}_i}\, g(\boldsymbol{W}\boldsymbol{H}_i + \boldsymbol{b}\mathbf{1}^\top)$. Any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) obeys*

$$\boldsymbol{W}^\top \boldsymbol{G}_i \;=\; -\lambda_{\boldsymbol{H}} \boldsymbol{H}_i. \tag{22}$$

Proof of Lemma 5. By definition, any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) satisfies the following:

$$\nabla_{\boldsymbol{H}_i} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;=\; \boldsymbol{W}^\top \nabla_{\boldsymbol{Z}_i = \boldsymbol{W}\boldsymbol{H}_i}\, g(\boldsymbol{W}\boldsymbol{H}_i + \boldsymbol{b}\mathbf{1}^\top) + \lambda_{\boldsymbol{H}} \boldsymbol{H}_i \;=\; \mathbf{0}; \tag{23}$$

$$\boldsymbol{W}^\top \boldsymbol{G}_i \;=\; -\lambda_{\boldsymbol{H}} \boldsymbol{H}_i. \tag{24}$$

as desired. □

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

Lemma 7. *Let $\tau = -\frac{\nabla g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)}{\lambda_{\boldsymbol{b}}}$. Any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) obeys*

$$\boldsymbol{b} \;=\; \tau \mathbf{1}. \tag{25}$$

Proof of Lemma 7. By definition, any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16) satisfies the following:

$$\nabla_{\boldsymbol{b}} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;=\; \nabla\, g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)\mathbf{1} + \lambda_{\boldsymbol{b}} \boldsymbol{b} \;=\; \mathbf{0},$$

$$\boldsymbol{b} \;=\; -\frac{\nabla g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)}{\lambda_{\boldsymbol{b}}} \mathbf{1} = \tau \mathbf{1} \tag{26}$$

as desired. □

Lemma 8. *Let $\boldsymbol{W} = \begin{bmatrix} (\boldsymbol{w}^1)^\top \\ \vdots \\ (\boldsymbol{w}^K)^\top \end{bmatrix} \in \mathbb{R}^{K \times d}$, $\boldsymbol{H} = [\boldsymbol{H}_1 \;\; \boldsymbol{H}_2 \;\; \cdots \;\; \boldsymbol{H}_n] \in \mathbb{R}^{d \times N}$, $\boldsymbol{H}_i = [\boldsymbol{h}_{1,i} \;\; \cdots \;\; \boldsymbol{h}_{K,i}] \in \mathbb{R}^{d \times K}$, $\bar{\boldsymbol{Z}} = \boldsymbol{W}\boldsymbol{H} \in \mathbb{R}^{d \times N}$, $N = nK$, and $\boldsymbol{b} = \tau \mathbf{1}$. Given $g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}_K^\top)$ defined in (17), for any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of (16), it satisfies*

$$f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \geq \phi\left(\rho^\star\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star| \tag{27}$$

$$\bar{\boldsymbol{Z}}^\star = -\rho^\star \left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right) \boldsymbol{I}_K^n \tag{28}$$

where ϕ is lower bound function satisfying the Definition 3, $\rho^\star = \arg\min_\rho \phi(\rho) + K\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho|$, and $\bar{\boldsymbol{Z}}^\star = \boldsymbol{W}^\star \boldsymbol{H}^\star$.

Proof of Lemma 8. With $\bar{\boldsymbol{Z}}_i = \boldsymbol{W}\boldsymbol{H}_i$, and $\left\|\bar{\boldsymbol{Z}}_i\right\|_2 = \sigma_i^{\max}$, we have the following lower bound for $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ as

$$\begin{aligned}
f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) &= g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top) + \frac{\lambda_{\boldsymbol{W}}}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2 \\
&= \sum_{i=1}^n \left(g(\boldsymbol{W}\boldsymbol{H}_i + \boldsymbol{b}\mathbf{1}^\top) + \frac{\lambda_{\boldsymbol{W}}}{2n}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_{\boldsymbol{H}}}{2}\|\boldsymbol{H}_i\|_F^2\right) + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2 \\
&\geq \sum_{i=1}^n \left(g(\bar{\boldsymbol{Z}}_i + \boldsymbol{b}\mathbf{1}^\top) + \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}/n}\left\|\bar{\boldsymbol{Z}}_i\right\|_*\right) + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2 \\
&\geq \sum_{i=1}^n \left(g(\bar{\boldsymbol{Z}}_i + \boldsymbol{b}\mathbf{1}^\top) + \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}/n}\frac{\left\|\bar{\boldsymbol{Z}}\right\|_F^2}{\left\|\bar{\boldsymbol{Z}}_i\right\|_2}\right) + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2 \\
&= \sum_{i=1}^n \left(g(\bar{\boldsymbol{Z}}_i + \boldsymbol{b}\mathbf{1}^\top) + \frac{\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}/n}}{\sigma_i^{\max}}\left\|\bar{\boldsymbol{Z}}_i\right\|_F^2\right) + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2,
\end{aligned}$$

where the first inequality is from Lemma 2, and the second inequality becomes equality only when $\bar{\boldsymbol{Z}}_i \neq \mathbf{0}$ and

$$\begin{aligned} &\forall\, k, \sigma_k(\bar{\boldsymbol{Z}}_i) = \sigma_i^{\max} \text{ or } 0 \\ &\exists\, k, \sigma_k(\bar{\boldsymbol{Z}}_i) \neq 0 \end{aligned} \tag{29}$$

where $\sigma_k(\bar{\boldsymbol{Z}}_i)$ is the k-th singular value of $\bar{\boldsymbol{Z}}_i$. While we only consider $\bar{\boldsymbol{Z}}_i \neq \mathbf{0}$, we will show the $\bar{\boldsymbol{Z}}_i = \mathbf{0}$ can be included in an uniform form as following proof. We can further bound $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ by

$$\begin{aligned} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;&\geq\; \sum_{i=1}^{n} \left(g(\bar{\boldsymbol{Z}}_i + \boldsymbol{b}\mathbf{1}^\top) + \frac{\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}/n}}{\sigma_i^{\max}} \left\|\bar{\boldsymbol{Z}}_i\right\|_F^2 \right) + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2^2 , \\ &\geq \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n} \phi \left(\sum_{j \neq k} \left(\bar{z}_{k,i,j} - \bar{z}_{k,i,k} + \underbrace{b_j - b_k}_{=0} \right) \right) + \sum_{i=1}^{n} \frac{\sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}/n}}{\sigma_i^{\max}} \left\|\bar{\boldsymbol{Z}}_i\right\|_F^2 + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2^2 , \\ &= \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\phi \left(\sum_{j \neq k}^{K} (\bar{z}_{k,i,j} - \bar{z}_{k,i,k}) \right) + \frac{K\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}}{\sigma_i^{\max}} \|\bar{z}_{k,i}\|_2^2 \right) + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2^2 , \end{aligned} \tag{30}$$

where the first inequality is from the first condition (14) of loss function $\mathcal{L}$ and the equality achieves only when $\bar{z}_{k,i,j} = \bar{z}_{k,i,j'}$ for $j \neq k, j' \neq k$, and $b_j - b_k = 0$ is due to Lemma 7. If we denote by $\rho_{k,i} = \sum_{j \neq k}^{K} (\bar{z}_{k,i,j} - \bar{z}_{k,i,k}) / (K-1)$, then

$$\begin{aligned} \|\bar{\boldsymbol{z}}_{k,i}\|_2^2 &= \sum_{j \neq k} \bar{z}_{k,i,j}^2 + \bar{z}_{k,i,k}^2 \\ &\geq (K-1) \left(\sum_{j \neq k} \frac{\bar{z}_{k,i,j}}{K-1} \right)^2 + \bar{z}_{k,i,k} \\ &= (K-1) \left(\sum_{j \neq k} \frac{\bar{z}_{k,i,j} - \bar{z}_{k,i,k}}{K-1} + \bar{z}_{k,i,k} \right)^2 + \bar{z}_{k,i,k} \\ &= (K-1) \left(\rho_{k,i} + \bar{z}_{k,i,k} \right)^2 + \bar{z}_{k,i,k} \\ &\geq \frac{K-1}{K} \rho_{k,i}^2 \end{aligned}$$

where the first inequality achieves equality only when $\bar{z}_{k,i,j} = \bar{z}_{k,i,j'}$ for $j \neq k, j' \neq k$, and the last line achieves equality only when $\bar{z}_{k,i,k} = -\frac{K-1}{K}\rho_{k,i}$, thus $\bar{z}_{k,i,j} = \frac{1}{K}\rho_{k,i}$ for $j \neq k$. Denoting $\boldsymbol{\rho}_i = [\rho_{i,1} \quad \rho_{i,2} \quad \cdots \quad \rho_{i,K}]$ and $\operatorname{diag}(\boldsymbol{\rho}_i)$ is a diagonal matrix using $\boldsymbol{\rho}_i$ as diagonal entries, and supposing $|\rho_1| \geq |\rho_2| > \cdots > |\rho_K|$, we can express $\bar{\boldsymbol{Z}}_i$ as:

$$\bar{\boldsymbol{Z}}_i = -(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)\operatorname{diag}(\boldsymbol{\rho}_i), \tag{31}$$

and we can extend the expression of (30) as following

$$f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) \;\geq\; \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\underbrace{\phi\left(\rho_{k,i}\right) + \frac{(K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}}{\sigma_i^{\max}} \rho_{k,i}^2}_{\psi(\rho_{k,i})} \right) + \frac{\lambda_{\boldsymbol{b}}}{2} \|\boldsymbol{b}\|_2 \tag{32}$$

which is decouplable if we treat the i-th samples per class as a group, thus we only consider the i-th samples per class. In the next part, denote $\rho^\star = \arg\min_\rho \phi(\rho) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho|$.

When $K \geq 3$, according to the $\boldsymbol{Z} = -(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)\operatorname{diag}(\boldsymbol{\rho}_i)$, the condition of (29) and Lemma 4, we know $\boldsymbol{Z}$ has only two possible forms corresponding to two different objective value of $\sum_{k=1}^{K} \psi(\rho_k)$ such that

- $|\rho_1| = |\rho_2| = \cdots = |\rho_K|$: we can have $\sigma_{\max} = |\rho_1|$ and

$$\begin{aligned}\sum_{k=1}^{K}\psi(\rho_k) &= \sum_{k=1}^{K}\left(\phi\left(\rho_k\right) + \frac{(K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}}{\sigma^{\max}}\rho_k^2\right)\\ &= \sum_{k=1}^{K}\left(\phi\left(\rho_k\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho_k|\right)\\ &\geq K\left(\phi\left(\rho^\star\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star|\right)\end{aligned}$$

  where the last line holds equality only when $|\rho_1| = |\rho_2| = \cdots = |\rho_K| = \rho^*$.

- $|\rho_2| = \cdots = |\rho_K| = 0$: we can have $\sigma_{\max} = \sqrt{(K-1)/K}|\rho_1|$ and

$$\begin{aligned}\sum_{k=1}^{K}\psi(\rho_k) &= \phi\left(\rho_1\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}\sqrt{\frac{K}{K-1}}|\rho_1| + (K-1)\phi\left(0\right)\\ &= \phi\left(\rho_1\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho_1|\\ &\quad + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}\left(\sqrt{\frac{K}{K-1}} - 1\right)|\rho_1| + (K-1)\phi\left(0\right)\\ &\geq K\left(\phi\left(\rho^\star\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star|\right)\end{aligned}$$

  where the last line holds equality only when $|\rho_1| = \cdots = |\rho_K| = |\rho^\star| = 0$.

When $K = 2$, according to the Lemma 3, we can calculate $\sigma_{\max} = \sqrt{\frac{\rho_1^2+\rho_2^2}{2}}$, then

$$\begin{aligned}\sum_{k=1}^{2}\psi(\rho_i) &= \phi\left(\rho_1\right) + \phi\left(\rho_2\right) + \frac{(K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}}{\sigma^{\max}}\left(\rho_1^2 + \rho_2^2\right)\\ &= \phi\left(\rho_1\right) + \phi\left(\rho_2\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}\sqrt{2(\rho_1^2+\rho_2^2)}\\ &= \phi\left(\rho_1\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho_1| + \phi\left(\rho_2\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho_2|\\ &\quad + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}\left(\sqrt{2(\rho_1^2+\rho_2^2)} - |\rho_1| - |\rho_2|\right)\\ &\geq 2\left(\phi\left(\rho^\star\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star|\right)\end{aligned}$$

where the last line holds equality only when $|\rho_1| = |\rho_2| = |\rho^\star|$.

Combining them together, for $K \geq 2$, we can further extend the expression of (32) as following

$$\begin{aligned}f(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) &\geq \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\phi\left(\rho_{k,i}\right) + \frac{(K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}}{\sigma_i^{\max}}\rho_{k,i}^2\right) + \frac{\lambda_{\boldsymbol{b}}}{2}\left\|\boldsymbol{b}\right\|_2\\ &\geq \frac{1}{N}\sum_{i=1}^{n}K\left(\phi\left(\rho^\star\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star|\right) + \frac{\lambda_{\boldsymbol{b}}}{2}\left\|\boldsymbol{b}\right\|_2\\ &\geq \phi\left(\rho^\star\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho^\star| \qquad (33)\end{aligned}$$

where the last equation is achieved when $\boldsymbol{b} = \boldsymbol{0}$ or $\lambda_{\boldsymbol{b}} = 0$. According to the condition (15) of loss function $\mathcal{L}$ that the minimizer $\rho^\star$ of $\phi(\rho) + c|\rho|$ is unique for any $c > 0$, and by denoting $\boldsymbol{I}_K^n = [\boldsymbol{I}_K \quad \cdots \quad \boldsymbol{I}_K] \in \mathbb{R}^{K\times nK}$, we have

$$\bar{\boldsymbol{Z}}_i^\star = -\rho^\star\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top\right) \qquad (34)$$

$$\bar{\boldsymbol{Z}}^\star = -\rho^\star\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top\right)\boldsymbol{I}_K^n \qquad (35)$$

as desired. □

Next, we show that the lower bound in (27) is attained if and only if $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ satisfies the following conditions.

Lemma 9. *Under the same assumptions of Lemma 8, the lower bound in (27) is attained for any minimizer $(\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star)$ of (16) if and only if the following hold*

$$\|\boldsymbol{w}^\star\|_2 = \left\|\boldsymbol{w}^{\star 1}\right\|_2 = \left\|\boldsymbol{w}^{\star 2}\right\|_2 = \cdots = \left\|\boldsymbol{w}^{\star K}\right\|_2, \quad \text{and} \quad \boldsymbol{b}^\star = b^\star \mathbf{1},$$

$$\boldsymbol{h}_{k,i}^\star = \sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}} n}} \boldsymbol{w}^{\star k}, \quad \forall\, k \in [K],\ i \in [n], \quad \text{and} \quad \overline{\boldsymbol{h}}_i^\star := \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{h}_{j,i}^\star = \mathbf{0}, \quad \forall\, i \in [n],$$

where either $b^\star = 0$ or $\lambda_{\boldsymbol{b}} = 0$, and the matrix $\boldsymbol{W}^{\star\top}$ is in the form of K-simplex ETF structure (see appendix for the formal definition) in the sense that

$$\boldsymbol{W}^{\star\top}\boldsymbol{W}^\star = \|\boldsymbol{w}^\star\|_2^2 \frac{K}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right).$$

The proof of Lemma 9 utilizes the Lemma Lemma 5, Lemma 6 and Lemma 7, and the conditions (33) and the structure of $\bar{\boldsymbol{Z}}^\star$ (35) during the proof of Lemma 8.

Proof of Lemma 9. From the (35), we know that $\bar{\boldsymbol{Z}}_1^\star = \bar{\boldsymbol{Z}}_2^\star = \cdots = \bar{\boldsymbol{Z}}_n^\star$ and then $\boldsymbol{G}_i^\star = \nabla_{\bar{\boldsymbol{Z}}_i^\star = \boldsymbol{W}^\star \boldsymbol{H}_i^\star} g(\boldsymbol{W}^\star \boldsymbol{H}_i^\star + \boldsymbol{b}\mathbf{1}^\top)$ is equivalent for $i \in [n]$. Let denote $\boldsymbol{G}^\star = \boldsymbol{G}_1^\star = \boldsymbol{G}_2^\star = \cdots = \boldsymbol{G}_n^\star$, the (22) in Lemma 6 can be expressed as:

$$\boldsymbol{W}^{\star\top}\boldsymbol{G}^\star = -\lambda_{\boldsymbol{H}}\boldsymbol{H}_i^\star$$

Therefore, $\tilde{\boldsymbol{H}}^\star = \boldsymbol{H}_1^\star = \boldsymbol{H}_2^\star = \cdots = \boldsymbol{H}_n^\star$, which means the last-layer features from different classes are collapsed to their corresponding class-mean $\boldsymbol{h}_{k,1}^\star = \boldsymbol{h}_{k,2}^\star = \cdots = \boldsymbol{h}_{k,n}^\star$, for $k \in [K]$. Furthermore, $\boldsymbol{H}^\star\boldsymbol{H}^{\star\top} = n\tilde{\boldsymbol{H}}^\star\tilde{\boldsymbol{H}}^{\star\top}$, combining this with (19) in Lemma 5, we know that

$$\lambda_{\boldsymbol{W}}\boldsymbol{W}^{\star\top}\boldsymbol{W}^\star = \lambda_{\boldsymbol{H}}\boldsymbol{H}^\star\boldsymbol{H}^{\star\top} = n\lambda_{\boldsymbol{H}}\tilde{\boldsymbol{H}}^\star\tilde{\boldsymbol{H}}^{\star\top}$$

By denoting $\boldsymbol{W}^\star = \boldsymbol{U}_{\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{W}}\boldsymbol{V}_{\boldsymbol{W}}^\top$ and $\tilde{\boldsymbol{H}}^\star = \boldsymbol{U}_{\tilde{\boldsymbol{H}}}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{H}}}\boldsymbol{V}_{\tilde{\boldsymbol{H}}}^\top$, where $\boldsymbol{U}_{\boldsymbol{W}}$, $\boldsymbol{\Sigma}_{\boldsymbol{W}}$, $\boldsymbol{V}_{\boldsymbol{W}}^\top$ are the left singular vector matrix, singular value matrix, and right singular vector matrix of $\boldsymbol{W}^\star$, respectively; and $\boldsymbol{U}_{\tilde{\boldsymbol{H}}^\star}$, $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{H}}^\star}$, $\boldsymbol{V}_{\tilde{\boldsymbol{H}}^\star}^\top$ are the left singular vector matrix, singular value matrix, and right singular vector matrix of $\tilde{\boldsymbol{H}}$, respectively, we can get

$$\boldsymbol{V}_{\boldsymbol{W}}^\top = \boldsymbol{U}_{\tilde{\boldsymbol{H}}}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \sqrt{\frac{n\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{H}}}$$

Therefore, $\boldsymbol{Z}_i^\star = \boldsymbol{W}^\star\tilde{\boldsymbol{H}}^\star = \sqrt{\frac{\lambda_{\boldsymbol{W}}}{n\lambda_{\boldsymbol{H}}}}\boldsymbol{U}_{\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{W}}^2\boldsymbol{V}_{\tilde{\boldsymbol{H}}}^\top$. According to the $\boldsymbol{Z}_i = -\rho^\star(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)$ in (34) and $\rho^\star \le 0$, which is symmetric, thus, $\boldsymbol{U}_{\boldsymbol{W}} = \boldsymbol{V}_{\tilde{\boldsymbol{H}}}$, $\boldsymbol{W}^\star = \sqrt{\frac{n\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\tilde{\boldsymbol{H}}^{\star\top}$, that is, $\boldsymbol{w}^{\star k} = \sqrt{\frac{n\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\boldsymbol{h}_{k,i}^\star$, $\forall\, k \in [K],\ i \in [n]$ and

$$\begin{aligned}
\boldsymbol{Z}_i^\star &= \sqrt{\frac{\lambda_{\boldsymbol{W}}}{n\lambda_{\boldsymbol{H}}}}\boldsymbol{W}^\star\boldsymbol{W}^{\star\top} = \sqrt{\frac{\lambda_{\boldsymbol{W}}}{n\lambda_{\boldsymbol{H}}}}\boldsymbol{W}^\star\boldsymbol{W}^\star \\
&= -\rho^\star(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) = -\rho^\star(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \\
\boldsymbol{W}^\star &= (\frac{\rho^{\star 2} n\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}})^{\frac{1}{4}}(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \\
\tilde{\boldsymbol{H}}^\star &= (\frac{\rho^{\star 2}\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}})^{\frac{1}{4}}(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)
\end{aligned}$$

Therefore,

$$\left\|\boldsymbol{w}^{\star 1}\right\|_2 = \left\|\boldsymbol{w}^{\star 2}\right\|_2 = \cdots = \left\|\boldsymbol{w}^{\star K}\right\|_2$$

$$\overline{\boldsymbol{h}}_i^\star := \frac{1}{K}\sum_{j=1}^{K}\boldsymbol{h}_{j,i}^\star = \mathbf{0}, \quad \forall\, i \in [n]$$

where $\overline{\boldsymbol{h}}_i^\star = \sum_{k=1}^{K}(\boldsymbol{h}_{k,i}^\star)$ and according to the condition of (33) and Lemma 7, $\boldsymbol{b}^\star = \mathbf{0}$ or $\lambda_{\boldsymbol{b}} = 0$. □

D Proof of Corollary 1 and Corollary 2

Following Theorem 2, we only need to prove convexity for label smoothing and local convexity for focal loss.

For any output (logit) $\boldsymbol{z} \in \mathbb{R}^K$, define

$$\boldsymbol{p} = \sigma(\boldsymbol{z}) \in \mathbb{R}^K, \text{ where } p_i = \frac{\exp(z_i)}{\sum_{j=1}^K \exp(z_j)}.$$

Let $\boldsymbol{y}^{\text{smooth}} \in \mathbb{R}^K$ be the label vector with $0 \le y_i^{\text{smooth}} \le 1$ and $\sum_i y_i^{\text{smooth}} = 1$. The three loss functions can be written as

$$f(\boldsymbol{z}) = \sum_{i=1}^K y_i^{\text{smooth}} \xi(p_i).$$

Some useful properties:

$$\partial_{z_i} \xi(p_k) = \begin{cases} \xi'(p_k)(p_k - p_k^2), & i = k, \\ -\xi'(p_k) p_k p_i, & i \neq k, \end{cases} \implies \nabla_{\boldsymbol{z}} \xi(p_k) = \xi'(p_k) p_k (\boldsymbol{e}_k - \boldsymbol{p})$$

$$\partial_{z_i} p_k = \begin{cases} p_k - p_k^2, & i = k, \\ -p_k p_i, & i \neq k, \end{cases} \implies \nabla_{\boldsymbol{z}} \boldsymbol{p} = \nabla_{\boldsymbol{z}} \sigma(\boldsymbol{z}) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$$

Therefore, the gradient and Hessian of $f(\boldsymbol{z})$ are given by

$$\nabla f(\boldsymbol{z}) = \sum_{i=1}^K y_i^{\text{smooth}} \nabla_{\boldsymbol{z}} \xi(p_i) = \sum_{i=1}^K y_i^{\text{smooth}} \underbrace{\xi'(p_i) p_i}_{\eta(p_i)} (\mathbf{1}_i - \boldsymbol{p}) \tag{36}$$

$$\nabla^2 f(\boldsymbol{z}) = \nabla(\nabla f(\boldsymbol{z})) = \sum_{i=1}^K y_i^{\text{smooth}} \left(\eta'(p_i) p_i \underbrace{(\mathbf{1}_i - \boldsymbol{p})(\mathbf{1}_i - \boldsymbol{p})^\top}_{\succeq \mathbf{0}} - \eta(p_i) \underbrace{\left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right)}_{\succeq \mathbf{0}} \right)$$

Thus, $\nabla^2 f(\boldsymbol{z})$ is PSD when $\eta(p_i) \le 0$ and $\eta'(p_i) \ge 0$ for all i, i.e.,

$$\xi'(p_i) \le 0, \quad \xi''(p_i) p_i + \xi'(p_i) \ge 0. \tag{37}$$

Now we consider the following cases:

- **CE loss** with $\boldsymbol{y}^{\text{smooth}} = \boldsymbol{e}_k$ and $\xi(t) = -\log(t)$. In this case, $\xi'(p_i) = -\frac{1}{p_i}$ and $\eta(p_i) = \xi'(p_i) p_i = -1$, and thus
  $$\nabla^2 f(\boldsymbol{z}) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \succeq \mathbf{0},$$
  where the inequality can be obtained by the Gershgorin circle theorem.
- **Label smoothing** with $\boldsymbol{y}^{\text{smooth}} = (1-\alpha)\boldsymbol{e}_k + \frac{\alpha}{K}\mathbf{1}$ and $\xi(t) = -\log(t)$. In this case, $\xi'(p_i) = -\frac{1}{p_i}$ and $\eta(p_i) = \xi'(p_i) p_i = -1$, and thus
  $$\nabla^2 f(\boldsymbol{z}) = \sum_{i=1}^K y_i^{\text{smooth}} \left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \succeq \mathbf{0}$$
  since $\sum_{i=1}^K y_i^{\text{smooth}} = 1$.
- **Focal loss** with $\boldsymbol{y}^{\text{smooth}} = \boldsymbol{e}_k$ and $\xi(t) = -(1-t)^\beta \log(t)$. In this case,
  $$\begin{aligned}
  \xi'(p_i) &= \beta(1-p_i)^{\beta-1}\log(p_i) - \frac{(1-p_i)^\beta}{p_i}, \\
  \eta(p_i) &= \xi'(p_i) p_i = \beta p_i (1-p_i)^{\beta-1}\log(p_i) - (1-p_i)^\beta \le 0, \ \forall\, \beta \ge 0, p_i \in [0,1], \\
  \eta'(p_i) &= \beta(1-p_i)^{\beta-1}\log(p_i) - \beta(\beta-1)p_i(1-p_i)^{\beta-2}\log(p_i) + \beta(1-p_i)^{\beta-1} + \beta(1-p_i)^{\beta-1} \\
  &= \beta(1-p_i)^{\beta-2}\left((1-\beta p_i)\log(p_i) + 2(1-p_i)\right) \\
  &\ge \beta(1-p_i)^{\beta-2}\left(\log(p_i) + 2(1-p_i)\right).
  \end{aligned}$$

Thus, $\eta'(p_i) \geq 0$ whenever $0.21 \leq p_i \leq 1$. The Hessian becomes

$$\nabla^2 f(\boldsymbol{z}) = \eta'(p_k) p_k \underbrace{(\boldsymbol{e}_k - \boldsymbol{p})(\boldsymbol{e}_k - \boldsymbol{p})^\top}_{\succeq \mathbf{0}} - \eta(p_k) \underbrace{\left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right)}_{\succeq \mathbf{0}}$$

which is PSD when $0.21 \leq p_k \leq 1$.