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*

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

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

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

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

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

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 $d_1, \dots, d_n$. Let $\lambda_1, \dots, \lambda_n$ 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 $\boldsymbol{Z} = -\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}\boldsymbol{1}^\top\right) \operatorname{diag}(\rho_1, \rho_2, \cdots, \rho_K)$ 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

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

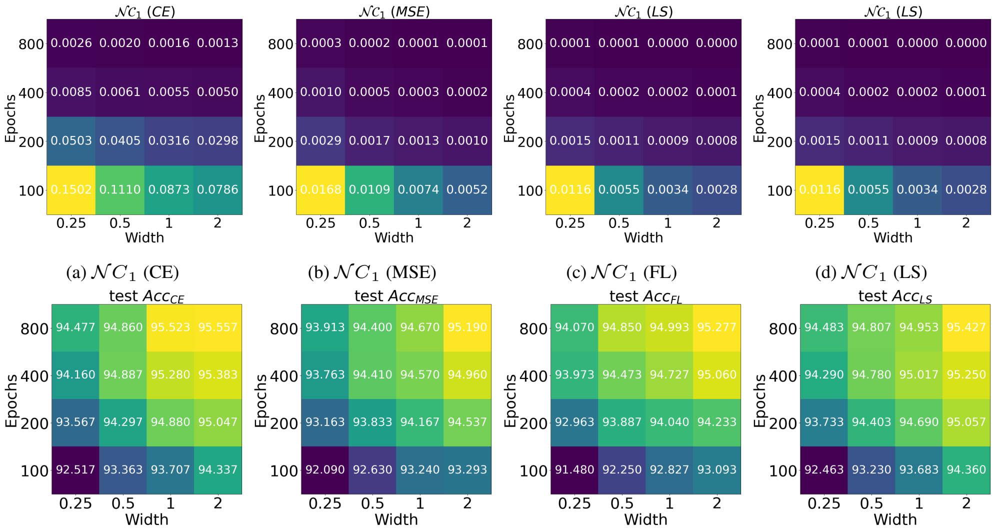
Proof of Lemma 4. Because

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

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
$$\begin{aligned}\lambda_1 = \cdots = \lambda_{i-1} = \rho_1^2 \\ \lambda_{i+1} = \cdots = \lambda_K = 0\end{aligned}$$
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
$$\begin{aligned}\lambda_1 = \cdots = \lambda_{i-1} = \rho_1^2 \\ \lambda_{i+1} = \cdots = \lambda_{K-1} = \rho_K^2\end{aligned}$$
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) $\mathcal{N}C_1$ (CE) (b) $\mathcal{N}C_1$ (MSE) (c) $\mathcal{N}C_1$ (FL) (d) $\mathcal{N}C_1$ (LS)

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

Figure 4: **Illustration of $\mathcal{N}C_1$ and test accuracy across different iterations-width configurations.** The figure depicts the $\mathcal{N}C_1$ 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.

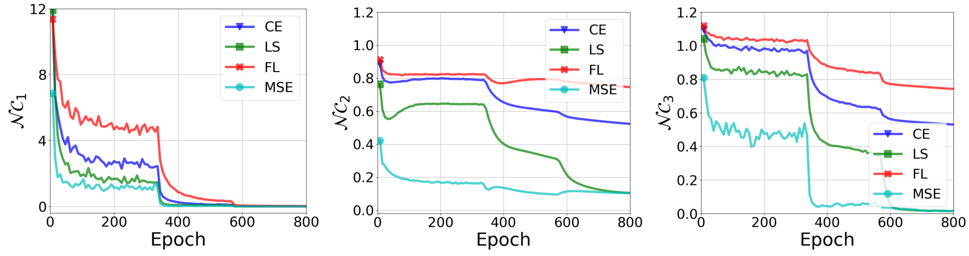
A.1 Additional experimental results on CIFAR10

In Section 4, we present the test accuracy for different losses function across various different iteration-width configurations. Moreover, we further show the $\mathcal{N}C_1$ 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 $\mathcal{N}C_1$ 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

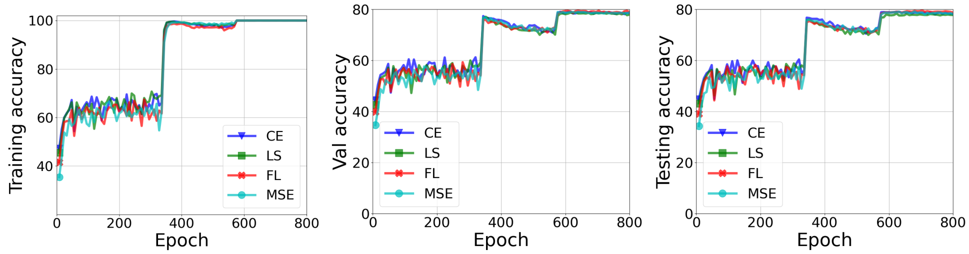
In this parts, we show the additional results on CIFAR100 dataset.

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) $\mathcal{N}C_1$ (CIFAR100) (b) $\mathcal{N}C_2$ (CIFAR100) (c) $\mathcal{N}C_3$ (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)

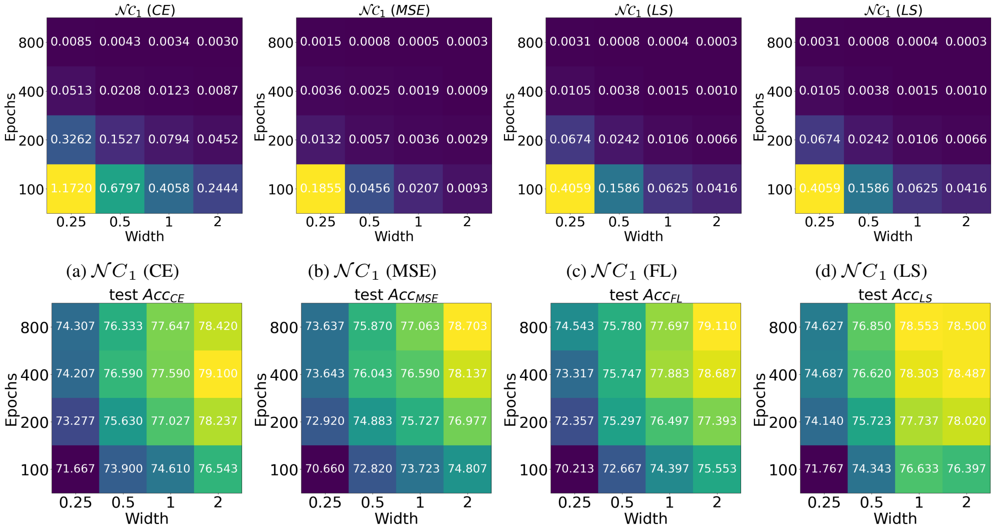
Figure 6: **The evolution of performance across different loss functions.** We train the WideResNet50-2 on CIFAR100 dataset for 800 epochs using different loss function. From left to right: training accuracy, validation accuracy and test accuracy.

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 $\mathcal{N}C_1$ 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 $\mathcal{N}C_1$ 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

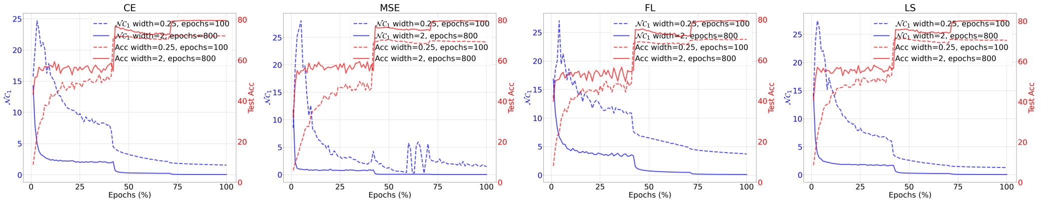
In this parts, we show the additional results on miniImageNet dataset. We trained WideResNet18-0.25 and WideResNet18-2 on miniImageNet for 100 epochs and 800 epochs, respectively. To reduce the randomness, we average the results from 3 different random trials. The $\mathcal{N}C_1$ and test accu-



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

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

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



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

Figure 8: **The evolution of $\mathcal{N}\mathcal{C}_1$ 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 $\mathcal{N}\mathcal{C}_1$ 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 $\mathcal{N}\mathcal{C}$ 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 $\mathcal{N}\mathcal{C}$ 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

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

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 $\phi_{\mathrm{FL}}(t)=\eta\left(\xi\left(t\right)\right)$ 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:

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

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.

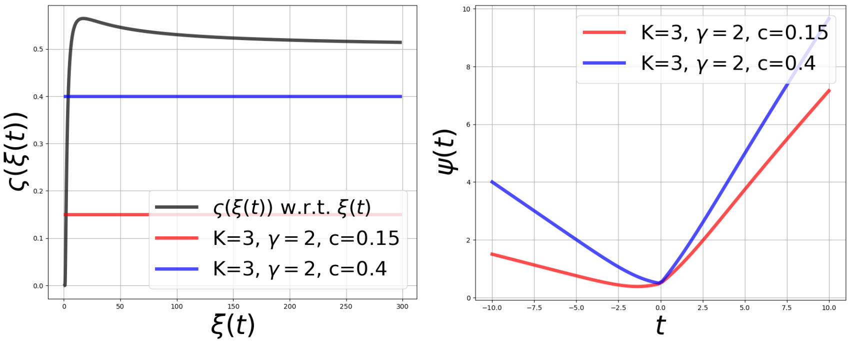
  - If $\vartheta(\xi(0)) = \vartheta(K) \geq 0$, then $\nabla_x^2\psi(x) > 0$ for $x \leq 0$, which means that $\nabla_x\xi(t)$ is an increasing function. Because $\varsigma(\xi(-\infty)) = \varsigma(1) = 0$, here we need to consider two cases(Please refer to Figure 9):

    - 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).

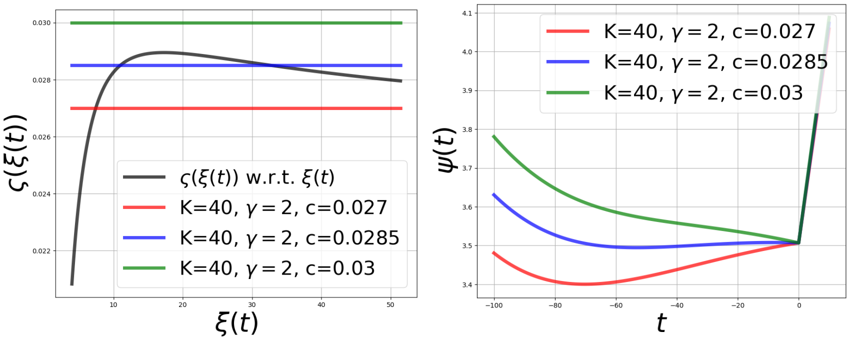
In conclusion, for focal loss, $\psi_{FL}(t)$ has a unique minimum in terms of $t \leq 0$, which satisfies the condition of (15).



(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*

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

with

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

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

obeys the following

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

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

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

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

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

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

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

We will first provide a lower bound for the general loss term $g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\boldsymbol{1}^\top)$ 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

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

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 $\bar{\boldsymbol{Z}}_i^\star = \boldsymbol{W}^\star \boldsymbol{H}_i^\star = -\rho^\star \left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top\right)$, 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) $\boldsymbol{b}^\star = b^\star \boldsymbol{1}$, where either $b^\star = 0$ or $\lambda_{\boldsymbol{b}} = 0$;

(c) $\overline{\boldsymbol{h}}_i^\star := \frac{1}{K}\sum_{j=1}^K \boldsymbol{h}_{j,i}^\star = \boldsymbol{0}, \quad \forall\, i \in [n]$, 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) $\boldsymbol{W}\boldsymbol{W}^\top = \|\boldsymbol{w}^\star\|_2^2 \frac{K-1}{K}\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top\right)$;

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*

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

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

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

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

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

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

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 $\boldsymbol{G}_i = \nabla_{\boldsymbol{Z}_i=\boldsymbol{W}\boldsymbol{H}_i}\, g(\boldsymbol{W}\boldsymbol{H}_i + \boldsymbol{b}\boldsymbol{1}^\top)$. 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:

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

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

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

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

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

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

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

as desired. □

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

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

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

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

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

where the first inequality is from Lemma 2, and the second inequality becomes equality only when $\bar{\boldsymbol{Z}}_i \neq \boldsymbol{0}$ 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 $\bar{\boldsymbol{Z}}_i \neq \boldsymbol{0}$, we will show the $\bar{\boldsymbol{Z}}_i = \boldsymbol{0}$ can be included in an uniform form as following proof. We can further bound $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ by

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

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 $\rho_{k,i} = \sum_{j \neq k}^{K} \left(\bar{z}_{k,i,j} - \bar{z}_{k,i,k} \right) / (K-1)$, then

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

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 $\boldsymbol{\rho}_i = \begin{bmatrix} \rho_{i,1} & \rho_{i,2} & \cdots & \rho_{i,K} \end{bmatrix}$ 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:

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

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

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

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 $\rho^\star = \arg\min_\rho \phi\left(\rho\right) + (K-1)\sqrt{n\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}|\rho|$.

When $K \geq 3$, according to the $\boldsymbol{Z} = -(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top)\operatorname{diag}(\boldsymbol{\rho}_i)$, 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

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

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

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

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

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*

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

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

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

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

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 $\boldsymbol{G}_i^\star = \nabla_{\bar{\boldsymbol{Z}}_i^\star = \boldsymbol{W}^\star \boldsymbol{H}_i^\star} g(\boldsymbol{W}^\star\boldsymbol{H}_i^\star + \boldsymbol{b}\boldsymbol{1}^\top)$ 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 $\boldsymbol{Z}_i = -\rho^\star(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top)$ in (34) and $\rho^\star \leq 0$, 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, $\boldsymbol{w}^{\star k} = \sqrt{\frac{n\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\boldsymbol{h}_{k,i}^\star, \quad \forall\, k \in [K],\ i \in [n]$ and

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

Therefore,

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

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

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, $\boldsymbol{b}^\star = \boldsymbol{0}$ 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:

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

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

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

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

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

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 $0.21 \leq p_k \leq 1$.