Inducing Neural Collapse in Imbalanced Learning: Do We Really Need a Learnable Classifier at the End of Deep Neural Network?
(Appendix)

A Proof for Theorem 1

Note that the constraint constrained optimization problem of Eq. (5) in our case is separable. We consider the k-th ($1 \leq k \leq K$ and $K \geq 2$) problem as:

$$\begin{aligned} \min_{\mathbf{H}} \quad & \frac{1}{N}\sum_{i=1}^{n_k} \mathcal{L}_{CE}(\mathbf{h}_{k,i}, \mathbf{W}^*), \\ s.t. \quad & ||\mathbf{h}_{k,i}||^2 \leq E_H,\ 1 \leq i \leq n_k. \end{aligned} \tag{17}$$

where $\mathbf{W}^*$ is the fixed ETF classifier. The problem above is convex as the objective is a sum of affine functions and log-sum-exp functions with convex constraints. We have the Lagrange function as:

$$\tilde{L} = \frac{1}{N}\sum_{i=1}^{n_k} -\log\frac{\exp(\mathbf{h}_{k,i}^T\mathbf{w}_k^*)}{\sum_{j=1}^{K}\exp(\mathbf{h}_{k,i}^T\mathbf{w}_j^*)} + \sum_{i=1}^{n_k}\mu_i\left(||\mathbf{h}_{k,i}||^2 - E_H\right), \tag{18}$$

where μ_i is the Lagrange multiplier. We have its gradient with respect to $\mathbf{h}_{k,i}$ as:

$$\frac{\partial \tilde{L}}{\partial \mathbf{h}_{k,i}} = -\frac{(1-p_k)\,\mathbf{w}_k^*}{N} + \frac{1}{N}\sum_{j\neq k}^{K} p_j\mathbf{w}_j^* + 2\mu_i\mathbf{h}_{k,i},\ 1 \leq i \leq n_k. \tag{19}$$

First we consider the case when $\mu_i = 0$. $\partial\tilde{L}/\partial\mathbf{h}_{k,i} = 0$ gives the following equation:

$$\sum_{j\neq k}^{K} p_j\mathbf{w}_j^* = \sum_{j\neq k}^{K} p_j\mathbf{w}_k^*. \tag{20}$$

Multiplying $\mathbf{w}_k^*$ by both sides of the equation, we should have:

$$\frac{K}{K-1}\sum_{j\neq k}^{K} p_j = 0, \tag{21}$$

which contradicts with $p_j > 0, \forall 1 \leq j \leq K$ when the ℓ_2 norm of $\mathbf{h}_{k,i}$ is constrained and $\mathbf{W}^*$ has a fixed ℓ_2 norm. So we have $\mu_i > 0$ and according to the KKT condition:

$$||\mathbf{h}_{k,i}||^2 = E_H, \tag{22}$$

Then we have the equation:

$$\frac{\partial \tilde{L}}{\partial \mathbf{h}_{k,i}^*} = \frac{1}{N}\sum_{j\neq k}^{K} p_j(\mathbf{w}_j^* - \mathbf{w}_k^*) + 2\mu_i\mathbf{h}_{k,i}^* = 0, \tag{23}$$

where $\mathbf{h}_{k,i}^*$ is the optimal solution of $\mathbf{h}_{k,i}$. Multiplying $\mathbf{w}_{j'}^*$ $(j' \neq k)$ by both sides of Eq. (23), we get:

$$E_W p_{j'}\left(1 + \frac{1}{K-1}\right) + 2N\mu_i\langle\mathbf{h}_{k,i}^*, \mathbf{w}_{j'}^*\rangle = 0. \tag{24}$$

Since $p_{j'} > 0$ and $K-1 > 0$, we have $\langle\mathbf{h}_{k,i}^*, \mathbf{w}_{j'}^*\rangle < 0$. Then for any pair $j, j' \neq k$, we have:

$$\frac{p_j}{p_{j'}} = \frac{\exp(\langle\mathbf{h}_{k,i}^*, \mathbf{w}_j^*\rangle)}{\exp(\langle\mathbf{h}_{k,i}^*, \mathbf{w}_{j'}^*\rangle)} = \frac{\langle\mathbf{h}_{k,i}^*, \mathbf{w}_j^*\rangle}{\langle\mathbf{h}_{k,i}^*, \mathbf{w}_{j'}^*\rangle}. \tag{25}$$

Considering that the function $f(x) = \exp(x)/x$ is monotonically increasing when $x < 0$, we have :

$$\langle\mathbf{h}_{k,i}^*, \mathbf{w}_j^*\rangle = \langle\mathbf{h}_{k,i}^*, \mathbf{w}_{j'}^*\rangle = C,\ p_j = p_{j'} = p,\ \forall j, j' \neq k, \tag{26}$$

where C and p are constants. From Eq. (24), we have:

$$p = \frac{1-K}{K}\cdot\frac{2N\mu_i C}{E_W}, \tag{27}$$

$$1 - p_k = (K-1)p = \frac{(1-K)(K-1)}{K} \cdot \frac{2N\mu_i C}{E_W}, \tag{28}$$

and

$$1 - p_k + p = (1-K) \cdot \frac{2N\mu_i C}{E_W}. \tag{29}$$

From Eq. (23), we have:

$$\mathbf{h}_{k,i}^{*} = \frac{1}{2N\mu_i}\left[(1-p_k)\,\mathbf{w}_k^{*} - \sum_{j\neq k}^{K} p_j \mathbf{w}_j^{*}\right]. \tag{30}$$

The ETF classifier defined in Eq. (1) satisfies that $\sum_i^K \mathbf{w}_i^{*} = 0$. Given that $p_j = p, \forall j \neq k$ and Eq. (29), we have:

$$\begin{aligned} \mathbf{h}_{k,i}^{*} &= \frac{1}{2N\mu_i}\,(1 - p_k + p)\,\mathbf{w}_k^{*} \\ &= \frac{(1-K)\,C}{E_W}\mathbf{w}_k^{*}, \end{aligned} \tag{31}$$

which indicates that $\mathbf{h}_{k,i}^{*}$ has the same direction as $\mathbf{w}_k^{*}$. Its length has been given in Eq. (22). Then we have:

$$C = \langle \mathbf{h}_{k,i}^{*}, \mathbf{w}_j^{*} \rangle = -\frac{\sqrt{E_H E_W}}{K-1},\ \forall j \neq k, \tag{32}$$

and

$$\mathbf{h}_{k,i}^{*} = \sqrt{\frac{E_H}{E_W}}\mathbf{w}_k^{*}, \tag{33}$$

which is equivalent to Eq. (7) and concludes the proof. □

B Proof for Theorem 2

We would like to show that the $\eta_{\mathbf{h}}$ of $\mathcal{L}_{DR}$ is always smaller than that of $\mathcal{L}_{CE}$ given $\mathbf{h}^t$ is close to $\mathbf{h}^*$.

For the DR loss:

Since $||\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}|| \leq \delta$ and $\mathbf{h}_{k,i}^{*}$ is aligned with $\mathbf{w}_k^{*}$, we let $||\mathbf{h}_{k,i}^{0}||^2 = E_H$ and $\cos\angle(\mathbf{h}_{k,i}^{0}, \mathbf{w}_k^{*}) \geq 0$ for any $k \in [1, K], i \in [1, n_k]$. For the DR loss in (14), the projected SGD takes the following step at time $t+1$ with the sample i in the class k:

$$\mathbf{h}_{k,i}^{t+1} = \mathrm{Proj}_{\mathrm{E_H}}\left(\mathbf{h}_{k,i}^{t} - \gamma\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{h}_{k,i}}\right) = \mathrm{Proj}_{\mathrm{E_H}}\left(\mathbf{h}_{k,i}^{t} - \gamma\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)\mathbf{w}_k^{*}\right), \tag{34}$$

where $\mathrm{Proj}_{\mathrm{E_H}}$ is the orthogonal projection onto the ball $\{\mathbf{h} : ||\mathbf{h}||^2 \leq \mathrm{E_H}\}$. Suppose that at the t-th iteration $||\mathbf{h}_{k,i}^{t}||^2 = E_H$ and $\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) \geq 0$, and one has:

$$\begin{aligned} \left\|\mathbf{h}_{k,i}^{t} - \gamma\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{h}_{k,i}}\right\|^2 &= E_H - 2\sqrt{E_H E_W}\gamma\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) \\ &\quad + \gamma^2 E_W\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)^2 \\ &\geq E_H, \end{aligned}$$

which means $||\mathbf{h}_{k,i}^{t+1}||^2 = E_H$. It is also easy to identify $\cos\angle(\mathbf{h}_{k,i}^{t+1}, \mathbf{w}_k^{*}) \geq 0$. So for all time $t \geq 0$ in the sequence from $\mathbf{h}_{k,i}^{0}$ to $\mathbf{h}_{k,i}^{*}$, we have $||\mathbf{h}_{k,i}^{t}||^2 = E_H$ and $\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) \geq 0$.

By the non-expansiveness of projection, one has the following convergence:

$$\begin{aligned} \left\|\mathbf{h}_{k,i}^{t+1} - \mathbf{h}_{k,i}^{*}\right\|^2 &\leq \left\|\mathbf{h}_{k,i}^{t} - \gamma\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)\mathbf{w}_k^{*} - \mathbf{h}_{k,i}^{*}\right\|^2 \\ &= \left\|\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}\right\|^2 - 2\gamma\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)\left\langle \mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}, \mathbf{w}_k^{*}\right\rangle \\ &\quad + \gamma^2 E_W\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)^2 \\ &\overset{a}{=} 2E_H\left(1 - \cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*})\right) - 2\gamma\sqrt{E_W E_H}\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)^2 \\ &\quad + \gamma^2 E_W\left(\cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}) - 1\right)^2, \end{aligned} \tag{35}$$

where $\overset{a}{=}$ holds because $||\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}||^2 = 2E_H(1 - \cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}))$. When $\gamma = \frac{\sqrt{E_H}}{\sqrt{E_W}}$, we have:

$$\begin{aligned} \left\|\mathbf{h}_{k,i}^{t+1} - \mathbf{h}_{k,i}^{*}\right\|^2 &\leq 2E_H\left(1 - \cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*})\right) - E_H\left(1 - \cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*})\right)^2 \\ &= \frac{1 + \cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*})}{2}\left\|\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}\right\|^2. \end{aligned} \tag{36}$$

Then we get that the $\eta_{\mathbf{h}}$-regularity number of the DR loss is:

$$\eta_{\mathbf{h}}^{(DR)} = \frac{1 + \cos\angle(\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*})}{2}.$$

For the CE loss:

On the other hand, for the CE loss in (3), the projected SGD takes the following step at time $t+1$ with the sample i in the class k:

$$\mathbf{h}_{k,i}^{t+1} = \mathrm{Proj}_{E_H}\left(\mathbf{h}_{k,i}^{t} - \gamma\frac{\partial\mathcal{L}_{CE}}{\partial\mathbf{h}_{k,i}}\right) = \mathrm{Proj}_{E_H}\left(\mathbf{h}_{k,i}^{t} + \gamma\left(1 - p_k\right)\mathbf{w}_k^{*} - \gamma\sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right).$$

By the non-expansiveness of projection, one has the following convergence:

$$\begin{aligned} \left\|\mathbf{h}_{k,i}^{t+1} - \mathbf{h}_{k,i}^{*}\right\|^2 &\leq \left\|\mathbf{h}_{k,i}^{t} + \gamma\left(1 - p_k\right)\mathbf{w}_k^{*} - \gamma\sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*} - \mathbf{h}_{k,i}^{*}\right\|^2 \\ &= \left\|\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}\right\|^2 - 2\gamma\left(p_k - 1\right)\left\langle\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}, \mathbf{w}_k^{*}\right\rangle + \gamma^2\left(p_k - 1\right)^2 E_W \\ &\quad + 2\gamma\left\langle\mathbf{h}_{k,i}^{*} - \mathbf{h}_{k,i}^{t}, \sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\rangle \\ &\quad - 2\gamma\left\langle(1 - p_k)\gamma\mathbf{w}_k^{*}, \sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\rangle + \left\|\gamma\sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\|^2 \\ &= \left\|\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}\right\|^2 + \gamma^2\left(p_k - 1\right)^2 E_W + \left\|\gamma\sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\|^2 + M, \end{aligned} \tag{37}$$

where we have:

$$\begin{aligned} M &= 2\gamma(1 - p_k)\left\langle\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}, \mathbf{w}_k^{*}\right\rangle - 2\gamma\left\langle(1 - p_k)\gamma\mathbf{w}_k^{*}, \sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\rangle + 2\gamma\left\langle\mathbf{h}_{k,i}^{*} - \mathbf{h}_{k,i}^{t}, \sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\rangle \\ &= 2\gamma(1 - p_k)\left(\left\langle\mathbf{h}_{k,i}^{t} - \mathbf{h}_{k,i}^{*}, \mathbf{w}_k^{*}\right\rangle + \frac{\gamma E_W}{K - 1}\sum_{j\neq k}^{K} p_j\right) + 2\gamma\left\langle\mathbf{h}_{k,i}^{*} - \mathbf{h}_{k,i}^{t}, \sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*}\right\rangle \\ &= 2\gamma(1 - p_k)\left(\left\langle\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}\right\rangle - \sqrt{E_H E_W} + \frac{\gamma E_W\left(1 - p_k\right)}{K - 1}\right) - 2\gamma\mathbf{h}_{k,i}^{t}\sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*} - \frac{2\gamma\left(1 - p_k\right)}{K - 1}\sqrt{E_H E_W} \\ &= 2\gamma(1 - p_k)\left\langle\mathbf{h}_{k,i}^{t}, \mathbf{w}_k^{*}\right\rangle - 2\gamma\mathbf{h}_{k,i}^{t}\sum_{j\neq k}^{K} p_j\mathbf{w}_j^{*} - 2\gamma\left(1 - p_k\right)\frac{K}{K - 1}\sqrt{E_H E_W} + \frac{2\gamma^2\left(1 - p_k\right)^2 E_W}{K - 1}. \end{aligned} \tag{38}$$

Since $\sum_i^K \mathbf{w}_i^* = 0$ for an ETF classifier, and we have assumed in Theorem 2 that $p_i = p_j = \frac{1-p_k}{K-1}$, $\forall i, j \neq k$, then we have:

$$\begin{aligned}
&2\gamma(1-p_k)\left\langle \mathbf{h}_{k,i}^t, \mathbf{w}_k^* \right\rangle - 2\gamma \mathbf{h}_{k,i}^t \sum_{j\neq k}^K p_j \mathbf{w}_j^* \\
=&-2\gamma \left\langle \mathbf{h}_{k,i}^t, (1-p_k)\sum_{j\neq k}^K \mathbf{w}_j^* + \sum_{j\neq k}^K p_j \mathbf{w}_j^* \right\rangle \\
\overset{a}{=}&-2\gamma \left\langle \mathbf{h}_{k,i}^t, \frac{K(1-p_k)}{K-1}\sum_{j\neq k}^K \mathbf{w}_j^* \right\rangle \\
=&-\frac{2K}{K-1}(1-p_k)\gamma\sqrt{E_H E_W}\left(\sum_{j\neq k}^K \cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_j^*\right)\right) \\
\overset{b}{=}&\frac{2K}{K-1}(1-p_k)\gamma\sqrt{E_H E_W}\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right),
\end{aligned} \tag{39}$$

where $\overset{a}{=}$ follows from the assumption that $p_i = p_j = \frac{1-p_k}{K-1}$, and $\overset{b}{=}$ holds because $\left\langle \mathbf{h}_{k,i}^t, \sum_i^K \mathbf{w}_i^* \right\rangle = 0$. Then we have :

$$\begin{aligned}
M &= \frac{2K}{K-1}(1-p_k)\gamma\sqrt{E_H E_W}\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right) - 2\gamma(1-p_k)\frac{K}{K-1}\sqrt{E_H E_W} + \frac{2\gamma^2(1-p_k)^2 E_W}{K-1} \\
&= -2(1-p_k)\frac{K}{K-1}\gamma\sqrt{E_H E_W}\left(1-\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right)\right) + \frac{2\gamma^2(1-p_k)^2 E_W}{K-1},
\end{aligned} \tag{40}$$

and

$$\begin{aligned}
&\left\|\mathbf{h}_{k,i}^{t+1} - \mathbf{h}_{k,i}^*\right\|^2 \\
\leq&\left\|\mathbf{h}_{k,i}^t - \mathbf{h}_{k,i}^*\right\|^2 - \frac{2K\sqrt{E_H E_W}}{K-1}\gamma(1-p_k)\left(1-\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right)\right) \\
&+\gamma^2(p_k-1)^2 E_W + \left\|\gamma\sum_{j\neq k}^K p_j \mathbf{w}_j^*\right\|^2 + \frac{2\gamma^2(1-p_k)^2 E_W}{K-1} \\
=&2E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right) - E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)^2 \\
&+E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)^2 - \frac{2K\sqrt{E_H E_W}}{K-1}\gamma(1-p_k)\left(1-\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right)\right) \\
&+\gamma^2(p_k-1)^2 E_W + \frac{\gamma^2(1-p_k)^2 E_W}{(K-1)^2} + \frac{2\gamma^2(1-p_k)^2 E_W}{K-1} \\
=&2E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right) - E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)^2 \\
&+E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)^2 - \frac{2K\sqrt{E_H E_W}}{K-1}\gamma(1-p_k)\left(1-\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right)\right) \\
&+\left(\frac{\gamma(1-p_k)K}{K-1}\right)^2 E_W.
\end{aligned} \tag{41}$$

Let $\gamma(1-p_k)\frac{K}{K-1} = s_k$, and we consider the problem:

$$\min_{s_k} \quad E_H\left(1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)^2 - 2s_k\sqrt{E_H E_W}\left(1-\cos\angle\left(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*\right)\right) + s_k^2 E_W. \tag{42}$$

When $s_k = \sqrt{\frac{E_H}{E_W}} \left(1 - \cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)$, *i.e.*, $\gamma = \frac{K-1}{K}\sqrt{\frac{E_H}{E_W}} \frac{1-\cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)}{1-p_k}$, the objective in Eq. (42) is minimized. Following Eq. (41) we get the optimal bound for CE as:

$$\begin{aligned}\left\|\mathbf{h}_{k,i}^{t+1} - \mathbf{h}_{k,i}^*\right\|^2 &\leq 2E_H\left(1 - \cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right) - E_H\left(1 - \cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)\right)^2 \\ &= \frac{1 + \cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)}{2}\left\|\mathbf{h}_{k,i}^t - \mathbf{h}_{k,i}^*\right\|^2,\end{aligned} \tag{43}$$

which is the same as DR. However, in this case γ varies with $\mathbf{h}_{k,i}^t$ and cannot be a constant as defined in Definition 2. For any fixed learning rate γ, the objective in Eq. (42) is larger than 0. So we have the $\eta_{\mathbf{h}}$-regularity number of the CE loss:

$$\eta_{\mathbf{h}}^{(CE)} \geq \frac{1 + \cos\angle(\mathbf{h}_{k,i}^t, \mathbf{w}_k^*)}{2} = \eta_{\mathbf{h}}^{(DR)}, \tag{44}$$

and conclude the proof. □

C Implementation Details

In implementations, we train a backbone network with our proposed ETF classifier and DR loss. For small datasets such as CIFAR, SVHN, and STL, we simply perform an ℓ_2 normalization for the features output from the backbone network, which means $\sqrt{E_H} = 1$. For large datasets, such as ImageNet, ℓ_2 normalization will induce training instability due to the large dimensionality. We add an ℓ_2 regularization term of the output features instead, to ensure a constrained feature length.

Our analytical work has shown that using a fixed ETF classifier does not suffer from the imbalanced gradient *w.r.t* classifier. In implementations, we also consider the gradient *w.r.t* the backbone network parameters $\mathbf{W}_{1:L-1}$:

$$\begin{aligned}\frac{1}{N}\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{W}_{1:L-1}} &= \frac{1}{N}\frac{\partial \mathbf{H}}{\partial \mathbf{W}_{1:L-1}}\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{H}} \\ &= \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\partial \mathbf{h}_{k,i}}{\partial \mathbf{W}_{1:L-1}}\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{h}_{k,i}}.\end{aligned}$$

Then we have:

$$\begin{aligned}\frac{1}{N}\left\|\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{W}_{1:L-1}}\right\|_2 &\leq \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\|\frac{\partial \mathbf{h}_{k,i}}{\partial \mathbf{W}_{1:L-1}}\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{h}_{k,i}}\right\|_2 \\ &\leq \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\|\frac{\partial \mathbf{h}_{k,i}}{\partial \mathbf{W}_{1:L-1}}\right\|_2 \cdot \left\|\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{h}_{k,i}}\right\|_2 \\ &\leq \frac{2}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sigma_{\max}\sqrt{E_{\mathbf{w}_k}},\end{aligned}$$

where $\sigma_{\max}$ denotes the largest singular value of the Jacobian $\frac{\partial \mathbf{h}_{k,i}}{\partial \mathbf{W}_{1:L-1}}, \forall 1 \leq k \leq K, 1 \leq i \leq n_k$, $\sqrt{E_{\mathbf{w}_k}}$ is the length of the k-th classifier vector, and $\left\|\frac{\partial \mathcal{L}_{DR}}{\partial \mathbf{h}_{k,i}}\right\|_2 \leq 2\sqrt{E_{\mathbf{w}_k}}$ by Eq. (15). Although we cannot realize a balanced gradient *w.r.t* $\mathbf{W}_{1:L-1}$ among classes, we can balance the upper bound of its gradient norm of each class by controlling $\sqrt{E_{\mathbf{w}_k}}$. In implementations, we set $\sqrt{E_{\mathbf{w}_k}} = \frac{N}{Kn_k}$, which is equivalent to performing a weighted loss function on different classes. In experiments, we also compare our method with the weighted CE loss for fair comparison.

D Datasets and Training Details

We conduct long-tailed classification experiments on the four datasets, CIFAR-10, CIFAR-100, SVHN, and STL-10, with two architectures, ResNet-32 and DenseNet with a depth of 150, a growth rate of 12, and a reduction of 0.5. All models are trained with the same training setting. Concretely, we train for 200 epochs, with an initial learning rate of 0.1, a batchsize of 128, a momentum of

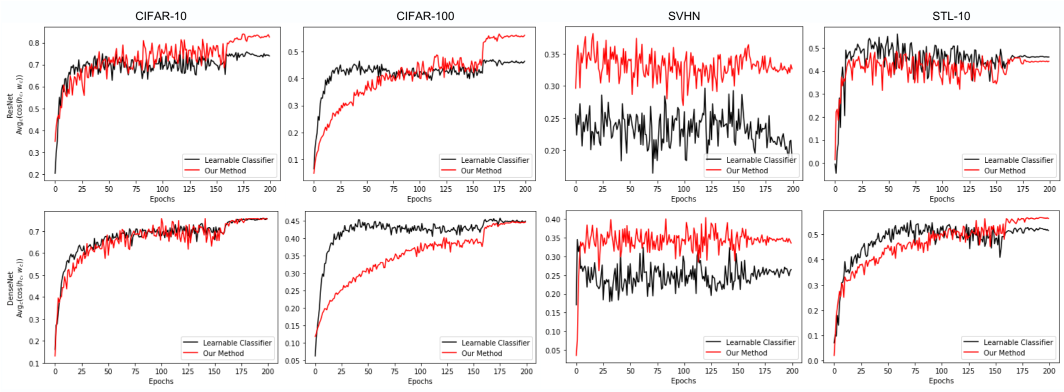


Figure 4: Averages of $\cos\angle(\mathbf{h}_c - \mathbf{h}_g, \mathbf{w}_c), \forall 1 \leq c \leq K$, where $\mathbf{h}_g$ is the global mean, with (red) and without (black) our method, using ResNet (up) and DenseNet (bottom) on four datasets. The models are trained on CIFAR-100 with an imbalance ratio of 0.02.

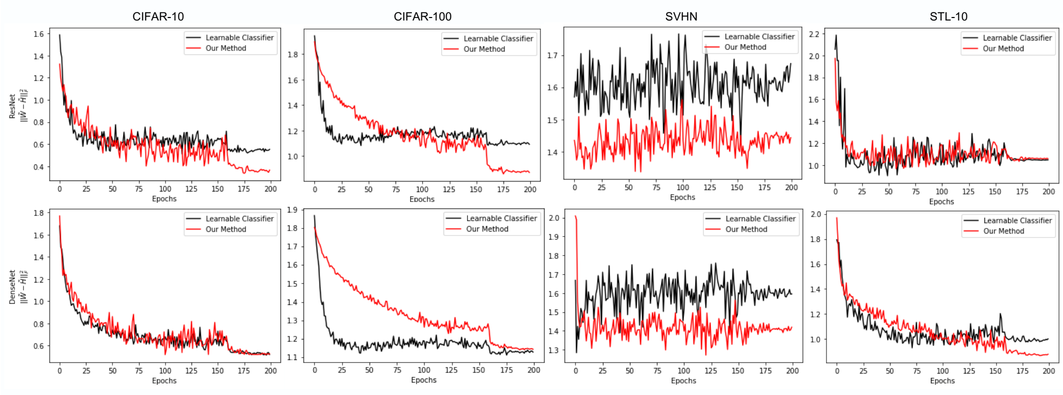


Figure 5: Statistics of $||\tilde{\mathbf{W}} - \tilde{\mathbf{H}}||_F^2$ during training, where $\tilde{\mathbf{W}} = \mathbf{W}/||\mathbf{W}||_F^2$, $\tilde{\mathbf{H}} = \bar{\mathbf{H}}/||\bar{\mathbf{H}}||_F^2$, and $\bar{\mathbf{H}} = [\mathbf{h}_c - \mathbf{h}_g : c = 1, \cdots, K]$, using ResNet (up) and DenseNet (bottom) on four datasets. The models are trained on CIFAR-100 with an imbalance ratio of 0.02

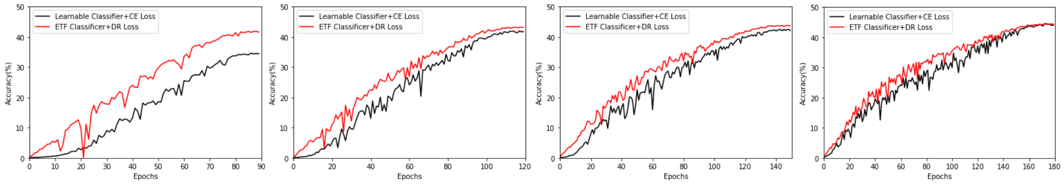
0.9, and a weight decay of $2e-4$. The learning rate is divided by 10 at epoch 160 and 180. The hyper-parameter in the β distribution used for Mixup is set as 1.0 when Mixup is used. We use the code released by [48] to produce the imbalanced datasets. The numbers of training samples are decayed exponentially among classes. We adopt the standard data normalization and augmentation for the four datasets.

We also conduct long-tailed classification experiments on ImageNet-LT with ResNet-50. We train all models for 90, 120, 150, and 180 epochs, with a batchsize of 1024 among 8 NVIDIA A-100 GPUs. Following [48], we use the SGD optimizer with a momentum of 0.9 and a weight decay of $5e-4$. The initial learning is 0.1 and decays following the cosine annealing schedule.

We conduct fine-grained classification experiments on CUB-200-2011 with ResNet-34, ResNet-50, and ResNet-101. The ResNet backbone networks are pre-trained on ImageNet. We train for 300 epochs on CUB-200-2011 with a batchsize of 64 and an initial learning rate of 0.04, which is dropped by 0.1 at epoch 90, 180, and 270. The standard data normalization and augmentation are adopted. Other training settings are the same as the long-tailed experiments.

E Additional Empirical and Experimental Results

Empirical Results. We provide more empirical results that have been discussed in Section 5.1. We calculate the averages of $\cos\angle(\mathbf{h}_c - \mathbf{h}_g, \mathbf{w}_c), \forall 1 \leq c \leq K$, and $||\tilde{\mathbf{W}} - \tilde{\mathbf{H}}||_F^2$ in Figure 4 and 5. It reveals that the model using our method generally has a higher $\cos\angle(\mathbf{h}_c - \mathbf{h}_g, \mathbf{w}_c)$ and a lower



(a) (b) (c) (d)

Figure 6: Top-1 accuracy curves on ImageNet-LT using the ResNet-50 backbone with "learnable classifier + CE loss" and our proposed "ETF classifier + DR loss" for (a) 90, (b) 120, (c) 150, and (d) 180 epochs of training.

$||\tilde{\mathbf{W}} - \tilde{\mathbf{H}}||_F^2$, which indicates that the feature means and classifier vectors of the same class are better aligned. We observe no advantage of ResNet on STL-10 and DenseNet on CIFAR-100 in Figure 4 and 5. In Table 2, we see that the two cases are right the failure cases, which shows consistency between neural collapse convergence and classification performance.

Accuracy Curves of Long-tailed Classification on ImageNet-LT. As shown in Figure 6, we depict the accuracy curves in training of the long-tailed classification results in Table 3. It reveals that the model using our proposed ETF classifier and DR loss converges faster than the traditional learnable classifier with the CE loss. Especially when we train for less epochs, the accuracy curves of our "ETF classifier + DR loss" are less affected, while those of "learnable classifier + CE loss" are deteriorated and converge slowly. Thus the superiority in performance of our method is more remarkable when training for limited epochs. It can be explained by the fact that our method directly has the classifier in its optimality and optimizes the features towards the neural collapse solution, while the learnable classifier with the CE loss needs a sufficient training process to separate classifier vectors of different classes. So our method can be preferred when fast convergence or limited training time is required.

F Limitations and Societal Impacts

Limitations. The limitations of this study may include: (1) The benefits of our proposed ETF classifier and DR loss are mainly analyzed for the case of imbalanced training. But our methods are general and applicable to all classification problems. The advantages and disadvantages of our methods for other machine learning areas, such as label noise learning and few-shot learning, are not discussed in this study, and deserve our future work. (2) In a neural network with a learnable classifier, the lengths, *i.e.,* ℓ_2-norms, of both features and classifier vectors are usually increasing as training. But for our fixed classifier, the lengths of the classifier vectors are fixed during training. As shown in Eq. (12) (for CE loss) and Eq. (15) (for DR loss), the gradient norm with respect to feature is decided by the lengths of classifier vectors. When feature is in a large length, its gradient using our fixed classifier may have a limited step size. So an adaptive mechanism to adjust the length of the ETF classifier may be more favorable and further improve the performance.

Societal Impacts. Our study proposes a new paradigm for neural network classification. It enjoys some theoretical benefits for imbalanced training, which is an important topic in machine learning. Our method is a general technique for network training, but is not related to any potential impact on privacy, public health, fairness, and other societal issues. Besides, our proposed ETF classifier and DR loss actually reduce the computation cost compared with the traditional learnable classifier with the CE loss, so will not introduce extra environment burden.

G Detailed Comparison with Some Studies

As suggested by a reviewer of this paper, we compare with [5, 30, 50] in more details.

The objective of LPM studied in [5] is also the CE loss with constraints of feature and classifier. They prove that (1) neural collapse is the global optimality of this objective when training on a balanced dataset; (2) in imbalanced training, neural collapse will be broken, and the prototypes of minor classes will be merged, which explains the difficulty of imbalanced training.

We also study the objective of CE loss with feature and classifier constraints. As a comparison, (1) We prove that neural collapse can also be the global optimality for imbalanced training as long as the classifier is fixed as an ETF (Theorem 1); (2) We analyze from the gradient perspective and show that the broken neural collapse in imbalanced training is caused by the imbalanced magnitude of gradients of the CE loss (Remark 2). We also show that the "pull-push" mechanism is crucial for the emergence of neural collapse in the CE loss in balanced training (Remark 3); (3) Inspired by the analyses, we propose a new loss function with a provable advantage over the CE loss (Theorem 2).

The objective of LPM studied in [50] is the CE loss with regularizations of feature and classifier. They prove that: (1) neural collapse is the global optimality of this objective when training on a balanced dataset; (2) despite being nonconvex, the landscape of the objective in this case is benign, which means that there is no spurious local minimum, so gradient-based optimization method can easily escape from the strict saddle points to look for the global minimizer.

In contrast, the objective we consider is the CE loss with constraints of feature and classifier, instead of regularizations. We mainly consider neural collapse in imbalanced learning, and accordingly propose a new loss function, which are different from the two results in [50].

[29, 30] show that fixing a learnable classifier as the vertices of regular polytopes, including d-simplex, d-cube, and d-orthoplex, helps to learn stationary and maximally separated features. It does not harm the performance, and in many cases improves the performance. It also brings faster speed of convergence and reduces the model parameters. Their spirit of learning maximally separated features is very similar to neural collapse. Compared with [29, 30], we prove that neural collapse can be the global optimality of the CE loss even in imbalanced learning based on the LPM analytical tool. We also propose a new loss function with a provable advantage over the CE loss.