

---

# Appendix of Paper "How Gradient Descent Separates Data with Neural Collapse: A Layer-Peeled Perspective"

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Elements of optimization

2 In this section, we introduce some basic definition and theory about optimization. In the following  
3 discussion, we consider a standard form inequality constrained optimization problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} f(x) \\ \text{s.t. } g_i(x) \leq 0, i = 1, 2, \dots, n \end{aligned} \quad (1)$$

4 In addition, we assume all of those functions  $f$  and  $g_i$  are twice differentiable. A point  $x \in \mathbb{R}^d$  is  
5 said to be feasible if and only if it satisfies all of the constraints in (1), i.e.  $g_i(x) \leq 0, i = 1, 2, \dots, n$ .  
6 And the Lagrangian of problem (1) is defined as following:

$$L(x, \lambda) = f(x) + \sum_{i=1}^n \lambda_i g_i(x) \quad (2)$$

### 7 A.1 Karush Kuhn Tucker Conditions

8 Now let's first introduce the Karush Kuhn Tucker (KKT) point and approximate KKT point. Here we  
9 follows the definition of  $(\epsilon, \delta)$ -KKT point as in [5].

10 **Definition A.1** (Definition of KKT point). A feasible point is said to be KKT point of problem (1) if  
11 there exist  $\lambda_1, \lambda_2, \dots, \lambda_n \geq 0$  such that the following Karush Kuhn Tucker (KKT) conditions hold:

12 1.  $\nabla f(x) + \sum_{i=1}^n \lambda_i \nabla g_i(x) = 0$

13 2.  $\lambda_i g_i(x) = 0, \forall i = 1, 2, \dots, n$

14 **Definition A.2** (Definition of  $(\epsilon, \delta)$ -KKT point).  $\forall \epsilon, \delta > 0$ , a feasible point is said to be  $(\epsilon, \delta)$ -KKT  
15 point of problem (1) if there exist  $\lambda_1, \lambda_2, \dots, \lambda_n \geq 0$  such that:

16 1.  $\|\nabla f(x) + \sum_{i=1}^n \lambda_i \nabla g_i(x)\| \leq \epsilon$

17 2.  $\lambda_i g_i(x) \geq -\delta, \forall i = 1, 2, \dots, n$

18 Generally speaking, KKT conditions may be not necessary for global optimum. We need some  
19 additional regular conditions to obtain the make it necessary, for example, as shown in [1] we can  
20 require the problem satisfy the following Mangasarian-Fromovitz constraint qualification (MFCQ):

21 **Definition A.3** (Mangasarian-Fromovitz constraint qualification (MFCQ) ). For a feasible point  $x$  of  
22 (1), problem (1) is said to satisfy (MFCQ) at  $x$  if there exist a vector  $v \in \mathbb{R}^d$  such that:

$$\langle \nabla_x g_i(x), v \rangle > 0, \forall i \in [n] \quad (3)$$

Moreover, when MFCQ holds we can build a connection between approximate KKT point and KKT point, see detailed proof in [1]:

**Theorem A.1** (Relationship between Approximate KKT point and KKT point). *Let  $\{x_k \in \mathbb{R}^d : k \in \mathbb{N}\}$  be a sequence of feasible points of (1),  $\{\epsilon_k > 0 : k \in \mathbb{N}\}$  and  $\{\delta_k > 0 : k \in \mathbb{N}\}$  be two sequences.  $x_k$  is an  $(\epsilon_k, \delta_k)$ -KKT point for every  $k$ , and  $\epsilon_k \rightarrow 0, \delta_k \rightarrow 0$ . If  $x_k \rightarrow x$  as  $k \rightarrow +\infty$  and MFCQ holds at  $x$ , then  $x$  is a KKT point of (P).*

## B Omitted proofs from Section 3.1

Recall our ULPM problem:

$$\min_{\mathbf{W}, \mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}) = - \sum_{k=1}^K \sum_{i=1}^n \log \left( \frac{\exp(\mathbf{w}_k^\top \mathbf{h}_{k,i})}{\sum_{j=1}^K \exp(\mathbf{w}_j^\top \mathbf{h}_{k,i})} \right) \quad (4)$$

Let's review some basic notation defined in the main body. Let  $s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i}, \forall k \in [K], i \in [n], j \in [K]$ , the margin of a single feature  $\mathbf{h}_{k,i}$  is defined to be  $q_{k,i}(\mathbf{W}, \mathbf{H}) := \min_{j \neq k} s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \max_{j \neq k} \mathbf{w}_j^\top \mathbf{h}_{k,i}$ . We define the neural collapse margin of entire dataset as  $q_{\min} = q_{\min}(\mathbf{W}, \mathbf{H}) = \min_{k \in [1,K], i \in [1,n]} q_{k,i}(\mathbf{W}, \mathbf{H})$ . Then an immediate result is the relationship between neural collapse margin and neural collapse.

**Lemma B.1** (Neural Collapse Margin as an Indicator of Neural Collapse). *The neural collapse margin is always smaller than*

$$q_{\min}(\mathbf{W}, \mathbf{H}) \leq \frac{\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2}{2(K-1)\sqrt{n}}$$

and  $(\mathbf{W}, \mathbf{H})$  must satisfies the neural collapse conditions when the inequality above is reduced to an equality.

*Proof.* First we can find that the margin will not change if we minus a vector  $a$  for all  $w_j$ , so we can first denote  $\tilde{\mathbf{w}}_i = \mathbf{w}_i - \frac{1}{n} \sum_{i=1}^K \mathbf{w}_i$  and then we have  $\mathbf{w}_k^\top \mathbf{h}_{k,i} - \max_{j \neq k} \mathbf{w}_j^\top \mathbf{h}_{k,i} = \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} - \max_{j \neq k} \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i} \geq q_{\min}(\mathbf{W}, \mathbf{H})$ , that is:

$$\tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} - \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i} \geq q_{\min}(\mathbf{W}, \mathbf{H}), \forall j \neq k \in [K], i \in [n] \quad (5)$$

Note that  $\sum_{j=1}^K \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i} = 0$  then sum this inequality over  $j$  we have:

$$(K-1)\tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} - \sum_{j \neq k} \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i} = K\tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} \geq (K-1)q_{\min}(\mathbf{W}, \mathbf{H}), \forall k \in [K], i \in [n] \quad (6)$$

By Cauchy inequality, we have:

$$\frac{1}{2} \left( \frac{1}{\sqrt{n}} \|\tilde{\mathbf{w}}_k\|_2^2 + \sqrt{n} \|\mathbf{h}_{k,i}\|_2^2 \right) \geq \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} \geq \frac{K-1}{K} q_{\min}(\mathbf{W}, \mathbf{H}) \quad (7)$$

Sum (7) over  $k$  and  $i$  we have:

$$\frac{1}{2} \sqrt{n} (\|\tilde{\mathbf{W}}\|_F^2 + \|\mathbf{H}\|_F^2) \geq n(K-1)q_{\min}(\mathbf{W}, \mathbf{H}) \quad (8)$$

On the other hands, we know that:

$$\|\tilde{\mathbf{W}}\|_F^2 = \sum_{i=1}^K \|\mathbf{w}_i - \frac{1}{n} \sum_{i=1}^K \mathbf{w}_i\|_2^2 \leq \sum_{i=1}^K \|\mathbf{w}_i\|_2^2 = \|\mathbf{W}\|_F^2 \quad (9)$$

Then we can conclude that:

$$q_{\min}(\mathbf{W}, \mathbf{H}) \leq \frac{\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2}{2(K-1)\sqrt{n}} \quad (10)$$

as desired. When the equality holds, first we have  $\|\tilde{\mathbf{W}}\|_F^2 = \|\mathbf{W}\|_F^2$ . The equality holds if and only if  $\frac{1}{n} \sum_{i=1}^K w_i = 0, \tilde{w}_i = w_i$ . Take it back into (8), then we must have all of the equality holds in (7) and (5), which give us:

$$\mathbf{w}_k = \sqrt{n} \mathbf{h}_{k,i}, \|\mathbf{w}_k\|_2^2 = n \|\mathbf{h}_{k,i}\|_2^2 = \frac{\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2}{2K} \quad (11)$$

Take this into (5) we have:

$$\mathbf{h}_{k,i} = \mathbf{h}_{k,i'}, \mathbf{h}_{k,i}^\top \mathbf{h}_{j,i'} = \mathbf{w}_k^\top \mathbf{w}_j = -\frac{\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2}{2K(K-1)\sqrt{n}}, \quad (12)$$

which implies neural collapse conditions.  $\square$

Now let's turn to the training dynamics, our main theorem about training convergence is as following:

**Theorem B.1.** *For problem (4), let  $(\mathbf{W}(t), \mathbf{H}(t))$  be the path of gradient flow at time  $t$ , if there exist a time  $t_0$  such that  $\mathcal{L}_{CE}(\mathbf{W}(t_0), \mathbf{H}(t_0)) < \log 2$ , then any limit point of  $\{(\hat{\mathbf{H}}(t), \hat{\mathbf{W}}(t)) := (\frac{\mathbf{H}(t)}{\sqrt{\|\mathbf{W}(t)\|_2^2 + \|\mathbf{H}(t)\|_2^2}}, \frac{\mathbf{W}(t)}{\sqrt{\|\mathbf{W}(t)\|_2^2 + \|\mathbf{H}(t)\|_2^2}})\}$  is along the direction of an Karush-Kuhn-Tucker (KKT) point of the following minimum-norm separation problem:*

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{H}} \quad & \frac{1}{2} \|\mathbf{W}\|_F^2 + \frac{1}{2} \|\mathbf{H}\|_F^2 \\ \text{s.t.} \quad & \forall k \neq j \in [K], i \in [n], \quad \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i} \geq 1. \end{aligned} \quad (13)$$

To prove Theorem B.1, we need to introduce some additional notations. Since the  $\mathbf{W}$  and  $\mathbf{H}$  are all optimization variable here, we denote  $\theta = \text{vec}(\mathbf{W}, \mathbf{H})$  as the whole parameter for simplicity and all of previous function can be defined on  $\theta$  by matching the corresponding parameter. Denote  $\rho = \|\theta\|$  as the norm of  $\theta$  and  $\tilde{\gamma} = \frac{-\log(e^{\mathcal{L}(\theta)} - 1)}{\rho^2}$ . Now we can state our first lemma to show how does training dynamics of gradient flow on ULPM objective (4) related to a KKT point of (13).

**Lemma B.2.** *If there exist a time  $t_0$  such that  $\mathcal{L}(\theta(t_0)) < \log 2$ , then for any  $t > t_0$   $\tilde{\theta} := \theta/q_{\min}(\theta)^{1/2}$  is a  $(\epsilon, \delta)$ -approximate KKT point of the following minimum-norm separation problem. More precisely, we have*

$$\epsilon = \sqrt{\frac{2(1 - \beta(t))}{\tilde{\gamma}(t)}}, \delta = \frac{K^2(K-1)n}{2\tilde{\gamma}(t)q_{\min}(t)}$$

where:

$$\beta = \langle \frac{\theta}{\|\theta\|_2}, \frac{d\theta}{dt} / \|\frac{d\theta}{dt}\|_2 \rangle$$

is the angle between  $\theta$  and its corresponding gradient.

*Proof.* The training dynamics is given by gradient flow:

$$\frac{d\theta}{dt} = -\frac{\partial \mathcal{L}(\theta)}{\partial \theta} \quad (14)$$

Then by the chain rule we have:

$$-\frac{d\mathcal{L}(\theta)}{dt} = -\frac{\partial \mathcal{L}}{\partial \theta} \frac{d\theta}{dt} = \left\| \frac{d\theta}{dt} \right\|_2^2 \quad (15)$$

It indicates that the loss function  $\mathcal{L}$  is monotonically decreasing. If  $\mathcal{L}(\theta(t_0)) < \log 2$ , we have  $\mathcal{L}(\theta(t)) < \log 2, \forall t > t_0$ . On the other hand,

$$\mathcal{L}(\theta(t)) = \sum_{k=1}^K \sum_{i=1}^n \log(1 + \sum_{j \neq k} e^{-s_{k,i,j}(t)}) \geq \log(1 + \exp(-q_{\min}(t))) \quad (16)$$

which gives us  $q_{\min}(t) > 0, \forall t > t_0$ .

Let  $g = \frac{d\theta}{dt}$ , note that we can rewrite the ULPM objective function (4) as  $\mathcal{L}(\theta) = \sum_{k=1}^K \sum_{i=1}^n \log(1 + \sum_{j \neq k} e^{-s_{k,i,j}})$ . By the chain rule and the gradient flow equation we have

$$g = -\frac{\partial \mathcal{L}}{\partial \theta} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} g_{k,i,j},$$

68 where  $g_{k,i,j}$  is the gradient of  $s_{k,i,j}$ , i.e.  $g_{k,i,j} = \nabla_{\theta} s_{k,i,j}(\theta)$ . Now let  $\tilde{g}_{k,i,j} = g_{k,i,j}/q_{min}^{1/2} =$   
69  $\nabla_{\theta} s_{k,i,j}(\tilde{\theta})$  and construct  $\lambda_{k,i,j} = \frac{\rho}{\|g\|_2} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}}$ , we only need to show:

$$\|\tilde{\theta} - \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \lambda_{k,i,j} \tilde{g}_{k,i,j}\|_2^2 \leq \frac{1-\beta}{\tilde{\gamma}} \quad (17)$$

70

$$\sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \lambda_{k,i,j} (s_{k,i,j}(\tilde{\theta}) - 1) \leq \frac{K^2(K-1)n}{2eq_{min}\tilde{\gamma}} \quad (18)$$

To prove (17), we only need to compute (Recall that  $\tilde{\theta} = \theta/q_{min}(\theta)^{1/2}$ ):

$$\|\tilde{\theta} - \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq y_n} \lambda_{k,i,j} \tilde{g}_{k,i,j}\|_2^2 = \frac{\rho^2}{q_{min}} \left\| \frac{\theta}{\|\theta\|_2} - \frac{g}{\|g\|_2} \right\|_2^2 = \frac{\rho^2}{q_{min}} (2 - 2\beta)$$

71 Note that:

$$\tilde{\gamma} = \frac{-\log(e^{\mathcal{L}(\theta)} - 1)}{\rho^2}, \quad \mathcal{L}(\theta) = \sum_{n=1}^N \log(1 + \sum_{j \neq y_n} e^{-s_{n,j}}) \geq \log(1 + \exp(-q_{min})) \quad (19)$$

72 Then we have the following inequality:

$$\tilde{\gamma} \leq \frac{q_{min}}{\rho^2} \quad (20)$$

73 take this back into (19) we have (17) as desired.

74

75 To prove (18), first by our construction:

$$\sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \lambda_{k,i,j} (s_{k,i,j}(\tilde{\theta}) - 1) = \frac{\rho}{q_{min}\|g\|_2} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} (s_{k,i,j} - q_{min}) \quad (21)$$

76 Note that  $\|g\|_2 \geq \left\langle g, \frac{\theta}{\|\theta\|_2} \right\rangle = \frac{1}{\rho} \langle g, \theta \rangle$  and  $\langle g_{k,i,j}, \theta \rangle = 2s_{k,i,j}$  since  $s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i}$ ,  
77 we have:

$$\begin{aligned} \|g\|_2 &\geq \frac{1}{\rho} \langle g, \theta \rangle = \frac{1}{\rho} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} \langle g_{k,i,j}, \theta \rangle \\ &= \frac{2}{\rho} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} s_{k,i,j} \\ &\geq \frac{2}{\rho} \frac{q_{min}}{K} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} e^{-s_{k,i,j}} \quad (\text{since } s_{k,i,j} \geq q_{min} > 0 \text{ and } e^{-s_{k,i,l}} \leq 1) \\ &\geq \frac{2}{\rho} \frac{q_{min}}{K} e^{-q_{min}} \end{aligned} \quad (22)$$

78 Take this inequality back into the (21) we have:

$$\begin{aligned}
\sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \lambda_{k,i,j} (s_{k,i,j}(\tilde{\theta}) - 1) &\leq \frac{K\rho^2}{2q_{\min}^2} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{q_{\min} - s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} (s_{k,i,j} - q_{\min}) \\
&\leq \frac{K\rho^2}{2q_{\min}^2} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} e^{q_{\min} - s_{k,i,j}} (s_{k,i,j} - q_{\min}) \\
&\leq \frac{K}{2q_{\min}\tilde{\gamma}} \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} e^{q_{\min} - s_{k,i,j}} (s_{k,i,j} - q_{\min}) \\
&\leq \frac{K^2(K-1)n}{2eq_{\min}\tilde{\gamma}}
\end{aligned} \tag{23}$$

79 Where the last inequality is obtained from the fact  $xe^{-x} \leq \frac{1}{e}, \forall x > 0$ , which can be proved by some  
80 elementary calculus.  $\square$

81 Based on the Lemma B.2, we have shown that the  $(\mathbf{W}, \mathbf{H})$  will be a  $(\epsilon, \delta)$ -KKT point, if we can  
82 show  $(\epsilon, \delta)$  converge to zero, then by Theorem A.1 we know the limit point will be along the direction  
83 of a KKT point. Ignoring the constant term, we only have to show how does  $\tilde{\gamma}(t), \beta(t)$  and  $q_{\min}(t)$   
84 evolves with time t. Now we provide the following lemmas to illustrate the dynamics of them. The  
85 first lemma aims at proving that the norm of parameter  $\rho(t)$  and  $\tilde{\gamma}(t)$  is monotonically increasing.

86 **Lemma B.3.** *If there exist  $t_0$  such that  $\mathcal{L}(\theta(t_0)) < \log 2$ , then  $\forall t > t_0$  we have:*

$$\frac{d\rho^2}{dt} > 0, \frac{d\tilde{\gamma}}{dt} \geq 0 \tag{24}$$

87 *Proof.* We can disentangle the whole training dynamics into the following two parts:

- 88 • the radial part:  $v := \hat{\theta} \hat{\theta}^\top \frac{d\theta}{dt}$ ,
- 89 • the tangent part:  $u = (I - \hat{\theta} \hat{\theta}^\top) \frac{d\theta}{dt}$ .

90 First analyze the radial part, by chain rule:  $\|v\|_2 = |\hat{\theta}^\top \frac{d\theta}{dt}| = |\frac{1}{\rho} \langle \theta, \frac{d\theta}{dt} \rangle| = |\frac{1}{\rho} \frac{1}{2} \frac{d\rho^2}{dt}|$ . For  $\frac{d\rho^2}{dt}$ , we  
91 have the following equation:

$$\frac{1}{2} \frac{d\rho^2}{dt} = \left\langle \theta, \frac{d\theta}{dt} \right\rangle = 2 \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} s_{k,i,j} \tag{25}$$

92 where the last equality holds by equation (22). Then when  $t > t_0$ , we have shown that  $q_{\min}(t) \geq 0$ ,  
93 combine this with the fact  $s_{k,i,j} \geq q_{\min}$  we obtain the first inequality in (24)

$$\begin{aligned}
\frac{1}{2} \frac{d\rho^2}{dt} &= 2 \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} s_{k,i,j} \\
&\geq 2 \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} q_{\min} \geq 0
\end{aligned} \tag{26}$$

94 Next, we aim to prove the monotonicity of  $\tilde{\gamma}(t)$ , we compute the derivative of  $\tilde{\gamma}(t)$ :

$$\tilde{\gamma} = \frac{-\log(e^{\mathcal{L}(\theta)} - 1)}{\rho^2}, \quad \frac{d}{dt} \log \tilde{\gamma} = \frac{d}{dt} (\log(-\log(e^{\mathcal{L}(\theta)} - 1)) - 2 \log \rho) \tag{27}$$

95

$$\frac{d}{dt} \log(-\log(e^{\mathcal{L}(\theta)} - 1)) = \frac{1}{\log(e^{\mathcal{L}(\theta)} - 1)} \frac{e^{\mathcal{L}(\theta)}}{e^{\mathcal{L}(\theta)} - 1} \frac{d\mathcal{L}(\theta)}{dt} \geq -\frac{d\mathcal{L}(\theta)}{dt} \frac{1}{q_{\min}} \frac{e^{\mathcal{L}(\theta)}}{e^{\mathcal{L}(\theta)} - 1} \tag{28}$$

96 Recall we have:

$$\begin{aligned}
\frac{1}{2} \frac{d\rho^2}{dt} &= 2 \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} s_{k,i,j} \\
&\geq 2 \sum_{k=1}^K \sum_{i=1}^n \frac{\sum_{j \neq k} e^{-s_{k,i,j}}}{1 + \sum_{j \neq k} e^{-s_{k,i,j}}} q_{min} \\
&= \sum_{k=1}^K \sum_{i=1}^n \log(1 + \sum_{j \neq k} e^{-s_{k,i,j}}) \frac{1}{\log(1 + \sum_{j \neq k} e^{-s_{k,i,j}})} \frac{\sum_{j \neq k} e^{-s_{k,i,j}}}{1 + \sum_{j \neq k} e^{-s_{k,i,j}}} q_{min} \quad (29) \\
&\geq 2 \sum_{k=1}^K \sum_{i=1}^n \log(1 + \sum_{j \neq k} e^{-s_{k,i,j}}) \frac{e^{\mathcal{L}(\theta)} - 1}{\mathcal{L}(\theta) e^{\mathcal{L}(\theta)}} q_{min} \\
&= 2 \frac{e^{\mathcal{L}(\theta)} - 1}{e^{\mathcal{L}(\theta)}} q_{min}
\end{aligned}$$

97 Then second last line is because the definition of th loss function  $\mathcal{L}(\theta) = \sum_{k=1}^K \sum_{i=1}^n \log(1 +$   
98  $\sum_{j \neq k} e^{-s_{k,i,j}}) \geq \log(1 + \sum_{j \neq k} e^{-s_{k,i,j}})$  and the monotonicity of  $\frac{e^x - 1}{xe^x}$  (in fact,  $\frac{d}{dx} \frac{e^x - 1}{xe^x} =$   
99  $\frac{e^{-x}(x - e^x + 1)}{x^2} \leq 0, \forall x > 0$ ).

100 As a result, we notice that

$$\frac{1}{2} \frac{d}{dt} \log \tilde{\gamma}(t) \geq - \left( \frac{1}{2} \frac{d\rho^2}{dt} \right)^{-1} \frac{d\mathcal{L}}{dt} - \frac{d}{dt} \log \rho \quad (30)$$

At the same time, we notice that  $\|v\|_2^2 = \frac{1}{\rho^2} \left( \frac{1}{2} \frac{d\rho^2}{dt} \right)^2 = \frac{1}{2} \frac{d\rho^2}{dt} \cdot \frac{d}{dt} \log \rho$  on the one hand, and by the chain rule:

$$\frac{d}{dt} \hat{\theta} = \frac{1}{\rho^2} \left( \rho \frac{d\theta}{dt} - \frac{d\rho}{dt} \theta \right) = \frac{1}{\rho^2} \left( \rho \frac{d\theta}{dt} - \left( \hat{\theta}^\top \frac{d\theta}{dt} \right) \theta \right) = \frac{u}{\rho}$$

Combine this with radial term:

$$- \frac{d\mathcal{L}}{dt} = \left\| \frac{d\theta}{dt} \right\|_2^2 = \|v\|_2^2 + \|u\|_2^2 = \frac{1}{2} \frac{d\rho^2}{dt} \cdot \frac{d}{dt} \log \rho + \rho^2 \left\| \frac{d\hat{\theta}}{dt} \right\|_2^2$$

101 Dividing  $\frac{1}{2} \frac{d\rho^2}{dt}$  on both sides, we have

$$- \frac{d\mathcal{L}}{dt} \cdot \left( \frac{1}{2} \frac{d\rho^2}{dt} \right)^{-1} = \frac{d}{dt} \log \rho + \left( \frac{d}{dt} \log \rho \right)^{-1} \left\| \frac{d\hat{\theta}}{dt} \right\|_2^2 \quad (31)$$

$$\frac{d}{dt} \log \rho + \left( \frac{d}{dt} \log \rho \right)^{-1} \left\| \frac{d\hat{\theta}}{dt} \right\|_2^2 \leq - \frac{d\mathcal{L}(\theta)}{dt} \frac{1}{q_{min}} \frac{e^{\mathcal{L}(\theta)}}{2(e^{\mathcal{L}(\theta)} - 1)} \quad (32)$$

102 Now by equation (27) and (28) we obtain:

$$\frac{1}{2} \frac{d}{dt} \log \tilde{\gamma} \geq - \frac{d\mathcal{L}(\theta)}{dt} \frac{1}{q_{min}} \frac{e^{\mathcal{L}(\theta)}}{2(e^{\mathcal{L}(\theta)} - 1)} - \frac{d}{dt} \log \rho \geq \left( \frac{d}{dt} \log \rho \right)^{-1} \left\| \frac{d\hat{\theta}}{dt} \right\|_2^2 \quad (33)$$

103 By (26) we know the  $\rho$  is monotonically increasing, we have  $\frac{d}{dt} \log \rho > 0$  and then we get the second  
104 inequality in (24)  $\square$

105 Lemma B.3 gives us the monotonicity of  $\tilde{\gamma}$ , note that since the loss function  $\mathcal{L}(\theta(t_0)) < \log 2$ , we  
106 have  $\tilde{\gamma}(t) \geq \tilde{\gamma}(t_0) > 0$ , then we can treat  $\tilde{\gamma}(t)$  in Lemma B.2 as a positive constant. The remaining  
107 work is to show  $q_{min}(t)$  grows to infinity and  $\beta(t) \rightarrow 1$ . To show  $q_{min}(t) \rightarrow \infty$ , it's equivalent to  
108 show  $\mathcal{L}(t) \rightarrow 0$  and we have the following lemma:

109 **Lemma B.4.** *If there exist  $t_0$  such that  $\mathcal{L}(\theta(t_0)) < \log 2$ , then  $\mathcal{L}(\theta) \rightarrow 0$  and  $q_{\min}(\theta) \rightarrow \infty$  as*  
 110  *$t \rightarrow \infty$*

*Proof.* By (15) and (25), the evolution of loss function  $\mathcal{L}(\theta)$  can be written as:

$$\frac{d\mathcal{L}(\theta)}{dt} = -\left\|\frac{d\theta}{dt}\right\|_2^2 \leq -\left\langle \frac{d\theta}{dt}, \frac{\theta}{\|\theta\|_2} \right\rangle^2 = -(2 \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} s_{k,i,j})^2$$

111 Combine it with (16), (26) and (29) we have:

$$2 \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} s_{k,i,j} \geq 2 \frac{e^{\mathcal{L}(\theta)} - 1}{e^{\mathcal{L}(\theta)}} q_{\min} \geq -2 \frac{e^{\mathcal{L}(\theta)} - 1}{e^{\mathcal{L}(\theta)}} \log(e^{\mathcal{L}(\theta)} - 1) \quad (34)$$

112 which indicates:

$$\frac{d\mathcal{L}(\theta)}{dt} \leq -4 \left( \frac{e^{\mathcal{L}(\theta)} - 1}{e^{\mathcal{L}(\theta)}} \log(e^{\mathcal{L}(\theta)} - 1) \right)^2 \quad (35)$$

113 Since we have shown that  $\mathcal{L}$  is monotonically decreasing, then if the  $\mathcal{L}$  doesn't decrease to zero, it must  
 114 stay larger than a positive number  $C_1 > 0$  and we know that  $\exists C_2, C_3 > 0$  such that  $C_2 < e^{\mathcal{L}(\theta)} <$   
 115  $C_3 < \log 2, \forall t > t_0$  which further implies  $\exists C_4 > 0$  such that  $(e^{\mathcal{L}(\theta)} - 1) \log(e^{\mathcal{L}(\theta)} - 1) < -C_4$   
 116 Take them back into (36), we must have:

$$\frac{d\mathcal{L}(\theta)}{dt} \leq -\frac{4}{e^{\mathcal{L}(\theta)}} C_4^2 \leq -C_4^2 \quad (36)$$

117 Then we know that the loss function will exponentially decrease to zero and contradicts with previous  
 118 assumption. Thus we must have  $\mathcal{L}(\theta) \rightarrow 0$  and combine this with  $q_{\min} \geq -\log(e^{\mathcal{L}(\theta)} - 1)$  we know  
 119  $q_{\min}(\theta) \rightarrow \infty$  as desired.  $\square$

120 To bound  $\beta(t)$ , we first need an useful lemma to bound the changes of the direction of  $\theta$ .

121 **Lemma B.5.** *If there exist  $t_0$  such that  $\mathcal{L}(\theta(t_0)) < \log 2$ , then for any  $t > t_0$*

$$\left\| \frac{d\hat{\theta}}{dt} \right\| \leq \frac{1}{\tilde{\gamma}(t_0)} \frac{d}{dt} \log \rho \quad (37)$$

122 *Proof.* First we know that:

$$\left\| \frac{d\hat{\theta}}{dt} \right\|_2 = \frac{1}{\rho} \left\| (I - \hat{\theta} \hat{\theta}^\top) \frac{d\theta}{dt} \right\|_2 \leq \frac{1}{\rho} \left\| \frac{d\theta}{dt} \right\| \quad (38)$$

123

$$\left\| \frac{d\theta}{dt} \right\|_2 = \left\| \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} g_{k,i,j} \right\| \leq \sum_{k=1}^K \sum_{i=1}^n \sum_{j \neq k} \frac{e^{-s_{k,i,j}}}{1 + \sum_{l \neq k} e^{-s_{k,i,l}}} \|g_{k,i,j}\| \quad (39)$$

124 Recall our  $g_{k,i,j} = \frac{\partial s_{k,i,j}}{\partial \theta}$  and  $s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i}$ , we can find that  $\|g_{k,i,j}\| \leq 2\rho$ . On the  
 125 other hand, Combine it with (25) and (20) we have:

$$\left\| \frac{d\hat{\theta}}{dt} \right\|_2 \leq \frac{1}{\rho} \left\| \frac{d\theta}{dt} \right\| \leq \frac{1}{2q_{\min}} \frac{d\rho^2}{dt} = \frac{\rho^2}{q_{\min}} \frac{d}{dt} \log \rho \leq \frac{1}{\tilde{\gamma}} \frac{d}{dt} \log \rho \leq \frac{1}{\tilde{\gamma}(t_0)} \frac{d}{dt} \log \rho \quad (40)$$

126 as desired. Where the second inequality holds by multiple  $\frac{\rho}{q_{\min}}$  on the right hand of (39) and the  
 127 formulation of  $\frac{d\rho^2}{dt}$  in equation (25), and the last inequality holds since we have shown that  $\tilde{\gamma}(t)$  is  
 128 monotonically increasing in (B.3) and  $\tilde{\gamma}(t_0) > 0$  since  $\mathcal{L}(t_0) < \log 2$ .  $\square$

129 Let's turn back to  $\beta$ , though we can't show directly that it increase to one, we can find a sequence of  
 130 time  $\{t_m\}$  for each limit point such that  $\beta(t_m) \rightarrow 1$

131 **Lemma B.6.** *If there exist  $t_0$  such that  $\mathcal{L}(\theta(t_0)) < \log 2$ , then for every limit point  $\bar{\theta}$  of  $\{\hat{\theta}(t) : t \geq 0\}$ ,*  
 132 *there exists a sequence of time  $\{t_m > 0 : m \in \mathbb{N}\}$  such that  $t_m \rightarrow \infty, \hat{\theta}(t_m) \rightarrow \bar{\theta}$ , and  $\beta(t_m) \rightarrow 1$*

133 *Proof.* Recall in equation (33) we have shown that:

$$\frac{d}{dt} \log \tilde{\gamma} \geq 2 \left( \frac{d}{dt} \log \rho \right)^{-1} \left\| \frac{d\hat{\theta}}{dt} \right\|_2^2 \quad (41)$$

Since  $\frac{d}{dt} \log \rho = \frac{1}{\rho} \frac{d\rho}{dt} = \frac{1}{2\rho^2} \frac{d\rho^2}{dt} = \frac{1}{\rho^2} \langle \theta, \frac{d\theta}{dt} \rangle$  and:

$$\frac{d\hat{\theta}}{dt} = \frac{d}{dt} \frac{\theta}{\|\theta\|} = \frac{1}{\rho^2} \left( \rho \frac{d\theta}{dt} - \frac{1}{\rho} \theta \theta^\top \frac{d\theta}{dt} \right) = \frac{1}{\rho} (I - \hat{\theta} \hat{\theta}^\top) \frac{d\theta}{dt}$$

134 Plug them into (41) we have:

$$\frac{d}{dt} \log \tilde{\gamma} \geq 2 \frac{\left\| \frac{d\theta}{dt} \right\|_2^2 - \left\langle \hat{\theta}, \frac{d\theta}{dt} \right\rangle^2}{\left\langle \hat{\theta}, \frac{d\theta}{dt} \right\rangle^2} \frac{d}{dt} \log \rho = 2(\beta^{-2} - 1) \frac{d}{dt} \log \rho \quad (42)$$

135 For any  $t_2 > t_1 > t_0$ , integrate both sides from time  $t_2$  to  $t_1$  we have:  $\log \tilde{\gamma}(t_2) - \log \tilde{\gamma}(t_1) \geq$   
 136  $2 \int_{t_1}^{t_2} (\beta(t)^{-2} - 1) \frac{d}{dt} \log \rho dt$ . By the continuity of  $\beta$  we know there exist a time  $t^*$  such that:

$$\begin{aligned} \log \tilde{\gamma}(t_2) - \log \tilde{\gamma}(t_1) &\geq 2 \int_{t_1}^{t_2} (\beta(t)^{-2} - 1) \frac{d}{dt} \log \rho dt \\ &= 2(\beta(t^*)^{-2} - 1) \int_{t_1}^{t_2} \frac{d}{dt} \log \rho dt \\ &= 2(\beta(t^*)^{-2} - 1) (\log \rho(t_2) - \log \rho(t_1)) \end{aligned} \quad (43)$$

137 By (20) we know that  $\tilde{\gamma} \leq \frac{q_{min}}{\rho^2}$  and the right hand is bounded, and the  $\tilde{\gamma}$  is monotonically increasing,  
 138 then there exist  $\tilde{\gamma}_\infty$  such that  $\tilde{\gamma}(t) \uparrow \tilde{\gamma}_\infty$ .  
 139 Now we are ready to construct the sequence of  $t_m$ , first take a sequence of  $\{\epsilon_m > 0, m \in \mathbb{N}\}$  such  
 140 that  $\epsilon_m \rightarrow 0$ . We construct  $t_m$  by induction, suppose we have already find  $t_1 < t_2 < \dots < t_{m-1}$   
 141 satisfy our requirement, since  $\bar{\theta}$  is a limit point of  $\{\hat{\theta}(t) : t > 0\}$ , then we can find a time  $s_m$  such  
 142 that:

$$\|\hat{\theta}(s_m) - \bar{\theta}\| \leq \epsilon_m, \quad \log \frac{\tilde{\gamma}_\infty}{\tilde{\gamma}(s_m)} \leq \epsilon_m^3 \quad (44)$$

143 By the monotonicity and continuity of  $\rho$  we can find a time  $s'_m$  such that  $\log \rho(s'_m) - \log \rho(s_m) \leq \epsilon_m$ .  
 144 Take  $t_2 = s'_m, t_1 = s_m$  in (43), there exist a time  $t_m$  such that:

$$2(\beta(t_m)^{-2} - 1) \leq \frac{\log \tilde{\gamma}(t_2) - \log \tilde{\gamma}(t_1)}{\log \rho(t_2) - \log \rho(t_1)} \leq \epsilon_m^2 \quad (45)$$

145 on the other hand, by Lemma B.5 we have:

$$\begin{aligned} \|\hat{\theta}(t_m) - \bar{\theta}\| &\leq \|\hat{\theta}(s_m) - \bar{\theta}\| + \|\hat{\theta}(s_m) - \hat{\theta}(t_m)\| \\ &\leq \epsilon_m + \frac{1}{\tilde{\gamma}(t_0)} (\log \rho(t_m) - \log \rho(s_m)) \leq (1 + \frac{1}{\tilde{\gamma}(t_0)}) \epsilon_m \end{aligned} \quad (46)$$

146 Note that  $\langle \theta, \frac{d\theta}{dt} \rangle > 0$ , then by definition we know  $\beta > 0$ . Combine equation (45) and (46) we have  
 147  $\beta(t_m) \rightarrow 1$  and  $\hat{\theta}(t_m) \rightarrow \bar{\theta}$  as desired.  $\square$

148 Now we are ready to prove the Theorem B.1:

*Proof.* By Lemma B.2, we know that once  $t > t_0$ , then  $(\mathbf{W}(t), \mathbf{H}(t))/q_{min}(\mathbf{W}(t), \mathbf{H}(t))$  is an  $(\sqrt{\frac{2(1-\beta(t))}{\tilde{\gamma}(t)}}, \frac{K^2(K-1)n}{2\tilde{\gamma}(t)q_{min}(t)})$ -approximate KKT point. We have shown that  $\tilde{\gamma}(t) > \tilde{\gamma}(t_0) > 0$  in Lemma B.3,  $q_{min} \rightarrow \infty$  in Lemma B.4. And from Lemma B.6, for any limit point  $(\bar{\mathbf{W}}, \bar{\mathbf{H}})$  of  $\{(\hat{\mathbf{H}}(t), \hat{\mathbf{W}}(t)) := (\frac{\mathbf{H}(t)}{\sqrt{\|\mathbf{W}(t)\|_2^2 + \|\mathbf{H}(t)\|_2^2}}, \frac{\mathbf{W}(t)}{\sqrt{\|\mathbf{W}(t)\|_2^2 + \|\mathbf{H}(t)\|_2^2}})\}$ , there exists a sequence of time  $\{t_m > 0 : m \in \mathbb{N}\}$  such that  $t_m \rightarrow \infty, \beta(t_m) \rightarrow 1$  and  $(\hat{\mathbf{H}}(t_m), \hat{\mathbf{W}}(t_m)) \rightarrow (\bar{\mathbf{W}}, \bar{\mathbf{H}})$ . Then



$(\bar{\mathbf{W}}, \bar{\mathbf{H}})$  is along the direction of a limit point of a sequence of  $(\epsilon, \delta)$ -approximate KKT point with  $\epsilon, \delta \rightarrow 0$ . On the other hand, we can verify that the problem (13) satisfies MCFQ (A.3) by simply setting  $v = \theta$ , then:

$$\langle \nabla s_{k,i,j}, \theta \rangle = 2s_{k,i,j} \geq 0$$

149 Now by Theorem A.1 we know  $(\bar{\mathbf{W}}, \bar{\mathbf{H}})$  is along the direction of a KKT point of problem (13)  $\square$

150 Theorem B.1 characterize the convergence behaviour of gradient flow, under separable conditions the  
151 limit point is along the direction of a KKT point of (13), next we show that the global minimum of  
152 (13) must satisfies neural collapse conditions

153 **Theorem B.2.** *Every global optimum of the minimum-norm separation problem (13) is also a KKT*  
154 *point and it satisfies the neural collapse conditions.*

155 *Proof.* Since we have shown that the problem (13) satisfy MCFQ, then the KKT conditions are  
156 necessary for global optimality, we only need to show the global optimum satisfy neural collapse  
157 conditions. First the constraints in (13) can be transformed to be a single constraint by the definition  
158 of neural collapse margin:

$$\forall k \neq j \in [K], i \in [n], \quad \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i} \geq 1. \Leftrightarrow q_{\min}(\mathbf{W}, \mathbf{H}) \geq 1 \quad (47)$$

159 Note that the neural collapse margin is homogeneous:

$$q_{\min}(\alpha \mathbf{W}, \alpha \mathbf{H}) = \alpha^2 q_{\min}(\mathbf{W}, \mathbf{H}), \forall \alpha \in \mathbb{R} \quad (48)$$

160 Then for any point  $(\mathbf{W}, \mathbf{H})$  satisfies  $q_{\min}(\mathbf{W}, \mathbf{H}) > 0$ , after an appropriate scaling  $\alpha$ ,  
161  $(\alpha \mathbf{W}, \alpha \mathbf{H}), \forall \alpha^2 \geq 1/q_{\min}(\mathbf{W}, \mathbf{H})$  is feasible for (13). Take optimum among all scaling fac-  
162 tor  $\alpha$  we know the minimum norm is attained if and only if  $\alpha^2 = 1/q_{\min}(\mathbf{W}, \mathbf{H})$ . And the optimum  
163 norm is:

$$\frac{1}{2} \|\alpha \mathbf{W}\|_F^2 + \frac{1}{2} \|\alpha \mathbf{H}\|_F^2 = \frac{1}{2q_{\min}(\mathbf{W}, \mathbf{H})} (\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) \quad (49)$$

164 Then by lemma B.1 we have:

$$\frac{1}{2q_{\min}(\mathbf{W}, \mathbf{H})} (\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) \geq 2(K-1)\sqrt{n} \quad (50)$$

165 And the global optimum is attained only when  $(\mathbf{W}, \mathbf{H})$  satisfies neural collapse conditions  $\square$

## 166 C Omitted proofs from Section 3.2

167 Let's first finish the computation of the motivating example:

168 **Example C.1** (A Motivating Example). Consider the case when  $K = 4, n = 1$ , let  $(\mathbf{W}, \mathbf{H})$  be the  
169 following point:

$$\mathbf{W} = \mathbf{H} = C \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \quad (51)$$

170 Verify this  $(\mathbf{W}, \mathbf{H})$  can classify all of the features perfectly is trivial since:

$$\mathbf{W}\mathbf{H} = 2C^2 \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \quad (52)$$

171 It's straightforward to verify it is along the direction of a KKT point of the minimum-norm separation  
172 problem (13) by our construction of  $\Lambda$ : (Note that the  $\mathbf{W}, \mathbf{H}$  should be normalized by dividing  
173  $2\sqrt{2}C$ )

$$\Lambda = \begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} \quad (53)$$

174 Now it only remains to show  $\forall \epsilon > 0$ , we can find  $(\mathbf{W}', \mathbf{H}')$  such that:

$$\begin{aligned} \|\mathbf{W}'\|_F^2 &= \|\mathbf{W}\|_F^2, \|\mathbf{H}'\|_F^2 = \|\mathbf{H}\|_F^2, \\ \|\mathbf{W}' - \mathbf{W}\|_F^2 + \|\mathbf{H}' - \mathbf{H}\|_F^2 &< \epsilon, \mathcal{L}(\mathbf{W}', \mathbf{H}') \leq \mathcal{L}(\mathbf{W}, \mathbf{H}) \end{aligned} \quad (54)$$

175 Without loss of generality, we only compute the case when  $C = 1$  for simplicity and one can easily  
176 generalize it to any  $C \neq 0$  by similar strategy. Here the  $(\mathbf{W}', \mathbf{H}')$  is constructed as below,

$$\begin{aligned} \mathbf{W}' &= \sqrt{\frac{1}{1+2\alpha^2}} \begin{bmatrix} 1+\alpha & -1+\alpha & \alpha & \alpha \\ -1+\alpha & 1+\alpha & \alpha & \alpha \\ -\alpha & -\alpha & 1-\alpha & -1-\alpha \\ -\alpha & -\alpha & -1-\alpha & 1-\alpha \end{bmatrix} \\ \mathbf{H}' &= \sqrt{\frac{1}{1+2\alpha^2}} \begin{bmatrix} 1+\alpha & -1+\alpha & -\alpha & -\alpha \\ -1+\alpha & 1+\alpha & -\alpha & -\alpha \\ \alpha & \alpha & 1-\alpha & -1-\alpha \\ \alpha & \alpha & -1-\alpha & 1-\alpha \end{bmatrix} \end{aligned} \quad (55)$$

177 Note that when  $\alpha = -1$  we have  $(\mathbf{W}', \mathbf{H}') = (\mathbf{W}, \mathbf{H})$ . First we can compute:

$$\mathbf{W}'\mathbf{H}' = \frac{1}{1+2\alpha^2} \begin{bmatrix} 2+4\alpha^2 & 4\alpha^2-2 & -4\alpha^2 & -4\alpha^2 \\ 4\alpha^2-2 & 2+4\alpha^2 & -4\alpha^2 & -4\alpha^2 \\ -4\alpha^2 & -4\alpha^2 & 2+4\alpha^2 & 4\alpha^2-2 \\ -4\alpha^2 & -4\alpha^2 & 4\alpha^2-2 & 2+4\alpha^2 \end{bmatrix} \quad (56)$$

178 and:

$$\mathcal{L}(\mathbf{W}', \mathbf{H}') = -4 \log \frac{e^2}{e^2 + e^{\frac{2\alpha^2-1}{1+2\alpha^2}} + 2e^{-\frac{2\alpha^2}{1+2\alpha^2}}} \quad (57)$$

179 Our aim is to show  $\forall \epsilon > 0$ , there exist  $\alpha$  such that  $|\alpha| < \epsilon$  and  $\mathcal{L}(\mathbf{W}', \mathbf{H}') < \mathcal{L}(\mathbf{W}, \mathbf{H})$ , if that is  
180 true, since  $\|\mathbf{W}'\|_F = \|\mathbf{W}\|_F, \|\mathbf{H}'\|_F = \|\mathbf{H}\|_F$  and  $\|\mathbf{W}' - \mathbf{W}\|_F^2 + \|\mathbf{H}' - \mathbf{H}\|_F^2 \rightarrow 0$  as  $\alpha \rightarrow 0$ ,  
181 the requirement in (56) holds immediately. By the monotonicity of  $\mathcal{L}(\mathbf{W}', \mathbf{H}')$ , it's sufficient to  
182 show that  $f(\alpha) \triangleq e^{\frac{2\alpha^2-1}{1+2\alpha^2}} + 2e^{-\frac{2\alpha^2}{1+2\alpha^2}} < f(0)$ . Then take the derivative of  $f(\alpha)$  we have:

$$f'(\alpha) = e^{\frac{4\alpha^2-2}{1+2\alpha^2}} \left( \frac{8\alpha}{1+2\alpha^2} - \frac{8\alpha(2\alpha^2-1)}{(1+2\alpha^2)^2} \right) + 2e^{-\frac{4\alpha^2}{1+2\alpha^2}} \left( \frac{16\alpha^3}{(1+2\alpha^2)^2} - \frac{8\alpha}{1+2\alpha^2} \right) \quad (58)$$

183

$$\begin{aligned} f''(\alpha) &= e^{\frac{4\alpha^2-2}{1+2\alpha^2}} \left( \frac{8\alpha}{1+2\alpha^2} - \frac{8\alpha(2\alpha^2-1)}{(1+2\alpha^2)^2} \right)^2 + 2e^{-\frac{4\alpha^2}{1+2\alpha^2}} \left( \frac{16\alpha^3}{(1+2\alpha^2)^2} - \frac{8\alpha}{1+2\alpha^2} \right)^2 \\ &\quad + e^{\frac{4\alpha^2-2}{1+2\alpha^2}} \left( -\frac{64\alpha^2}{(1+2\alpha^2)^2} + \frac{64(2\alpha^2-1)\alpha^2}{(1+2\alpha^2)^3} + \frac{8}{1+2\alpha^2} - \frac{8(2\alpha^2-1)}{(1+2\alpha^2)^2} \right) \\ &\quad + 2e^{-\frac{4\alpha^2}{1+2\alpha^2}} \left( \frac{80\alpha^2}{(1+2\alpha^2)^2} - \frac{8}{1+2\alpha^2} - \frac{128\alpha^4}{(1+2\alpha^2)^3} \right) \end{aligned} \quad (59)$$

184 Now we can find that  $f'(0) = 0$  and  $f''(0) = 16(\frac{1}{e^2} - 1) < 0$ . Since the function  $f(\alpha)$  is continuous  
185 twice differentiable, we can conclude that  $\forall \epsilon > 0$  we can find  $\alpha$  such that  $(\mathbf{W}', \mathbf{H}')$  satisfy our  
186 requirement in (56).  $\square$

187 Now we can first characterize the global optimum of our ULPM objective by some similar strategy as  
188 used in proving Lemma B.1

189 **Theorem C.1.** *The optimal value of loss function (4) on a sphere is attained (i.e.  $\mathcal{L}(\mathbf{W}, \mathbf{H}) \leq$   
190  $\mathcal{L}(\mathbf{W}', \mathbf{H}')$ ),  $\forall \|\mathbf{W}'\|_F^2 + \|\mathbf{H}'\|_F^2 = \|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2$  if only if the  $(\mathbf{W}, \mathbf{H})$  satisfies neural collapse  
191 conditions and  $\|\mathbf{W}\|_F = \|\mathbf{H}\|_F$ .*

192 *Proof.* Again we rewrite the ULPM objective by introducing  $s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i}$ :

$$\mathcal{L}(\mathbf{W}, \mathbf{H}) = \sum_{k=1}^K \sum_{i=1}^n \log(1 + \sum_{j \neq k} \exp(-s_{k,i,j})) \quad (60)$$

193 In addition, we can find that centralizing  $\mathbf{w}_i$  doesn't change the value of  $s_{k,i,j}$ . Let  $\tilde{\mathbf{w}}_i = \mathbf{w}_i -$   
 194  $\frac{1}{K} \sum_{k=1}^K \mathbf{w}_k, \forall i \in [K]$ , then  $s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i} = \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} - \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i}$  and  $\sum_{i=1}^K \tilde{\mathbf{w}}_i = 0$ .  
 195 First by the strict convexity of  $e^x$  and Jensen Inequality:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \mathbf{H}) &\geq \sum_{k=1}^K \sum_{i=1}^n \log(1 + (K-1) \exp(\frac{1}{K-1} \sum_{j \neq k} -s_{k,i,j})) \\ &= \sum_{k=1}^K \sum_{i=1}^n \log(1 + (K-1) \exp(-\frac{K \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i}}{K-1})) \end{aligned} \quad (61)$$

196 Where the last equality is obtained from:

$$\sum_{j \neq k} s_{k,i,j} = \sum_{j \neq k} \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} - \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i} = (K-1) \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} - \sum_{j \neq k} \tilde{\mathbf{w}}_j^\top \mathbf{h}_{k,i} = K \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i} \quad (62)$$

197 Now again by the strict convexity of  $\log(1 + (K-1) \exp(-x))$  and Jensen inequality, we have:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \mathbf{H}) &\geq \sum_{k=1}^K \sum_{i=1}^n \log(1 + (K-1) \exp(-\frac{K \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i}}{K-1})) \\ &\geq nK \log(1 + (K-1) \exp(-\frac{1}{n(K-1)} \sum_{k=1}^K \sum_{i=1}^n \tilde{\mathbf{w}}_k^\top \mathbf{h}_{k,i})) \\ &\geq nK \log(1 + (K-1) \exp(-\frac{1}{2n(K-1)} \sum_{k=1}^K \sum_{i=1}^n \frac{1}{\sqrt{n}} \|\tilde{\mathbf{w}}_k\|^2 + \sqrt{n} \|\mathbf{h}_{k,i}\|^2)) \\ &\geq nK \log(1 + (K-1) \exp(-\frac{1}{2\sqrt{n}(K-1)} (\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2))) \end{aligned} \quad (63)$$

198 Where the last inequality holds since  $\|\mathbf{W}\|_F^2 = \sum_{k=1}^K \|\mathbf{w}_k\|^2 \geq \sum_{k=1}^K \|\mathbf{w}_k\|^2 - \frac{1}{K} \|\sum_{k=1}^K \mathbf{w}_k\|^2 =$   
 199  $\sum_{k=1}^K \|\tilde{\mathbf{w}}_k\|^2$ .

200 When all of the above inequality reduce to equality, we must have:

- 201 1.  $\sum_{i=1}^K \mathbf{w}_i = 0, \tilde{\mathbf{w}}_i = \mathbf{w}_i$  (the last inequality in (63))
- 202 2.  $\mathbf{w}_k = \sqrt{n} \mathbf{h}_{k,i}, \forall i \in [n]$  (the third inequality in (63))
- 203 3.  $\|\mathbf{w}_k\| = \|\mathbf{w}_{k'}\|, \|\mathbf{h}_{k,i}\| = \|\mathbf{h}_{k',j}\|, \forall k, k' \in [K], i, j \in [n]$  (the second inequality in (63))
- 204 4.  $s_{k,i,j} = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \mathbf{w}_j^\top \mathbf{h}_{k,i} = \frac{K}{K-1} \mathbf{w}_k^\top \mathbf{h}_{k,i}, \forall k, j \in [K], i \in [n]$  (the first inequality in  
 205 (61))

206 These four conditions are exactly equivalent to neural collapse conditions and  $\|\mathbf{W}\|_F = \|\mathbf{H}\|_F$   $\square$

207 The global optimality is not enough to illustrate how does gradient flow converge to neural collapse  
 208 since there may exist some bad local minimum. We will provide the following second order analysis  
 209 to eliminate spurious local minimum. First define the cross entropy loss on a matrix  $\mathbf{Z} \in \mathbb{R}^{K \times nK}$ :

$$L(\mathbf{Z}) = \sum_{k=1}^K \sum_{i=1}^n -\log \frac{e^{z_{k,i,j}}}{\sum_{l=1}^K e^{z_{k,i,l}}} \quad (64)$$

210 where  $z_{k,i,j}$  denote the  $j$ -th row and  $(k-1)K + i$ -th column elements of  $\mathbf{Z}$ . Then we have  
 211  $\mathcal{L}(\mathbf{W}, \mathbf{H}) = L(\mathbf{W}\mathbf{H})$ . Now compute the gradient of  $L(\mathbf{Z})$  to each element:

$$\begin{aligned} \frac{\partial L(\mathbf{Z})}{\partial z_{k,i,k}} &= -1 + \frac{z_{k,i,k}}{\sum_{l=1}^K e^{z_{k,i,l}}} \\ \frac{\partial L(\mathbf{Z})}{\partial z_{k,i,j}} &= \frac{z_{k,i,j}}{\sum_{l=1}^K e^{z_{k,i,l}}}, \forall k \neq j \end{aligned} \quad (65)$$

212 If  $u \in \mathbb{R}^K$  satisfies  $u^\top \nabla L(Z) = 0$ , denote  $u_p$  as the maximum element of  $u$ , then we have:

$$\begin{aligned} 0 &= u_p \frac{\partial L(\mathbf{Z})}{\partial z_{p,i,p}} + \sum_{q \neq p} u_q \frac{\partial L(\mathbf{Z})}{\partial z_{p,i,q}} = u_p \left( -1 + \frac{z_{p,i,p}}{\sum_{l=1}^K e^{z_{p,i,l}}} \right) + \sum_{q \neq p} u_q \frac{z_{p,i,q}}{\sum_{l=1}^K e^{z_{p,i,l}}} \\ &= - \sum_{q \neq p} (u_p - u_q) \frac{z_{p,i,q}}{\sum_{l=1}^K e^{z_{p,i,l}}} \leq 0 \end{aligned} \quad (66)$$

213 Where the last inequality holds if and only if  $u_q = u_p, \forall q \in [K]$ . Which indicates that the rank of  
214  $\nabla L(Z)$  is  $K - 1$  and  $u^\top \nabla L(Z) = 0 \Leftrightarrow u = \mathbf{1}$ . Again we introduce the definition of tangent space:

215 **Definition C.1** (tangent space). The tangent space of  $(\mathbf{W}, \mathbf{H})$  is defined to be a set of directions that  
216 are orthogonal to  $(\mathbf{W}, \mathbf{H})$ :

$$\mathcal{T}(\mathbf{W}, \mathbf{H}) = \{ \Delta \mathbf{W} \in \mathbb{R}^{K \times d}, \Delta \mathbf{H} \in \mathbb{R}^{d \times nK} : \langle \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{W} \rangle + \langle \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{H} \rangle = 0 \} \quad (67)$$

217 Now we are ready to state our result about the landscape of ULPM in the tangent space:

218 **Theorem C.2.** If  $(\mathbf{W}, \mathbf{H})$  is not the optimal solutions in Theorem C.1 and  $q_{\min}(\mathbf{W}, \mathbf{H}) > 0$ , then  
219  $\exists (\Delta \mathbf{W}, \Delta \mathbf{H}) \in \mathcal{T}(\mathbf{W}, \mathbf{H}), M > 0$  such that

$$\forall 0 < \delta < M, \mathcal{L}(\mathbf{W} + \delta \Delta \mathbf{W}, \mathbf{H} + \delta \Delta \mathbf{H}) \leq \mathcal{L}(\mathbf{W}, \mathbf{H}) \quad (68)$$

220 . Further more, it implies that  $\forall \epsilon > 0, \exists (\mathbf{W}', \mathbf{H}')$  such that:

$$\begin{aligned} \|\mathbf{W}'\|_F^2 + \|\mathbf{H}'\|_F^2 &= \|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2, \\ \|\mathbf{W}' - \mathbf{W}\|_F^2 + \|\mathbf{H}' - \mathbf{H}\|_F^2 &< \epsilon, \mathcal{L}(\mathbf{W}', \mathbf{H}') \leq \mathcal{L}(\mathbf{W}, \mathbf{H}) \end{aligned} \quad (69)$$

*Proof.* First compute the gradient of ULPM objective (4), by the chain rule we have:

$$\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}) = \nabla L(\mathbf{W}\mathbf{H})\mathbf{H}^\top, \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}) = \mathbf{W}^\top \nabla L(\mathbf{W}\mathbf{H})$$

221 If there exist a vector  $(\Delta \mathbf{W}, \Delta \mathbf{H}) \in \mathcal{T}(\mathbf{W}, \mathbf{H})$  such that  $\langle \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{W} \rangle +$   
222  $\langle \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{H} \rangle \neq 0$ , moreover we can assume  $\langle \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{W} \rangle +$   
223  $\langle \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{H} \rangle < 0$  since we can take the negative direction if the formula is greater than  
224 zero, then by Taylor expansion:

$$\mathcal{L}(\mathbf{W} + \delta \Delta \mathbf{W}, \mathbf{H} + \delta \Delta \mathbf{H}) = \mathcal{L}(\mathbf{W}, \mathbf{H}) + \delta \langle \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{W} \rangle + \delta \langle \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{H} \rangle + O(\delta^2) \quad (70)$$

225 we know that  $(\Delta \mathbf{W}, \Delta \mathbf{H})$  satisfies our requirement.

226

Now let's discuss the case when:

$$\langle \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{W} \rangle + \langle \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \Delta \mathbf{H} \rangle = 0, \forall (\Delta \mathbf{W}, \Delta \mathbf{H}) \in \mathcal{T}(\mathbf{W}, \mathbf{H})$$

227 by definition of  $\mathcal{T}(\mathbf{W}, \mathbf{H})$ , it contains all vectors that are orthogonal to  $(\mathbf{W}, \mathbf{H})$ , so  
228  $(\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}))$  is parallel to  $(\mathbf{W}, \mathbf{H})$ , that is, there exist  $\lambda$  such that:

$$\nabla L(\mathbf{W}\mathbf{H})\mathbf{H}^\top = \lambda \mathbf{W}, \mathbf{W}^\top \nabla L(\mathbf{W}\mathbf{H}) = \lambda \mathbf{H} \quad (71)$$

229 Further more, from equation (25) and  $\frac{d\theta}{dt} = -\frac{\partial \mathcal{L}}{\partial \theta}$ , we know that  $q_{\min}(\mathbf{W}, \mathbf{H}) > 0$  implies the  
230 inner product of  $(\mathbf{W}, \mathbf{H})$  and  $(\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{H}), \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}))$  must be negative and thus  $\lambda < 0$   
231 If there doesn't exist  $(\Delta \mathbf{W}, \Delta \mathbf{H})$  satisfies the requirement, we know that for any feasible curve  
232  $\phi(t) = (\mathbf{W}(t), \mathbf{H}(t))$  with  $\phi(0) = (\mathbf{W}, \mathbf{H})$  on the sphere  $\mathcal{S} = \{(\mathbf{W}', \mathbf{H}') : \|\mathbf{W}'\|_F^2 + \|\mathbf{H}'\|_F^2 =$   
233  $\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2\}$ , we know that  $t = 0$  admits the local minimum of  $\mathcal{L}(\phi(t))$  and thus:

$$0 \leq \frac{d^2}{dt^2} \mathcal{L}(\phi(t))|_{t=0} = \phi'(0)^T \nabla^2 \mathcal{L}(\mathbf{W}, \mathbf{H}) \phi'(0) + \nabla \mathcal{L}(\mathbf{W}, \mathbf{H}) \phi''(0) \quad (72)$$

234 On the other hand, since the curve lies on the sphere  $\mathcal{S}$ , denote  $h(\mathbf{W}, \mathbf{H}) = \|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2$ , then  
235  $h(\phi(t))$  must stay as a constant, take twice derivative we have:

$$0 = \frac{d^2}{dt^2} h(\phi(t))|_{t=0} = \phi'(0)^T \nabla^2 h(\mathbf{W}, \mathbf{H}) \phi'(0) + \nabla h(\mathbf{W}, \mathbf{H}) \phi''(0) \quad (73)$$

236 Then sum these two conditions together, we have:

$$0 \leq \frac{d^2}{dt^2}(\mathcal{L}(\phi(t)) - \frac{\lambda}{2}h(\phi(t)))|_{t=0} = \phi'(0)^T \nabla^2(\mathcal{L} - \frac{\lambda}{2}h)(\mathbf{W}, \mathbf{H})\phi'(0) + \nabla(\mathcal{L} - \frac{\lambda}{2}h)(\mathbf{W}, \mathbf{H})\phi''(0) \quad (74)$$

237 By equation (71) we know that  $\nabla(\mathcal{L} - \frac{\lambda}{2}h)(\mathbf{W}, \mathbf{H}) = 0$ . Note that  $\phi'(0) \in \mathcal{T}(\mathbf{W}, \mathbf{H})$  since the  
238 curve lies on  $\mathcal{S}$  and for any  $(\Delta\mathbf{W}, \Delta\mathbf{H}) \in \mathcal{T}(\mathbf{W}, \mathbf{H})$  we can construct a curve  $\phi(t)$  such that  
239  $\phi'(0) = (\Delta\mathbf{W}, \Delta\mathbf{H})$ . Then (74) indicates that  $\forall (\Delta\mathbf{W}, \Delta\mathbf{H}) \in \mathcal{T}(\mathbf{W}, \mathbf{H})$  we have:

$$\begin{aligned} 0 &\leq (\Delta\mathbf{W}, \Delta\mathbf{H})^\top \nabla^2 \mathcal{L}(\mathbf{W}, \mathbf{H})(\Delta\mathbf{W}, \Delta\mathbf{H}) - \frac{\lambda}{2}(\Delta\mathbf{W}, \Delta\mathbf{H})^\top \nabla^2 h(\mathbf{W}, \mathbf{H})(\Delta\mathbf{W}, \Delta\mathbf{H}) \\ &= (\Delta\mathbf{W}, \Delta\mathbf{H})^\top \nabla^2 \mathcal{L}(\mathbf{W}, \mathbf{H})(\Delta\mathbf{W}, \Delta\mathbf{H}) - \lambda(\|\Delta\mathbf{W}\|_F^2 + \|\Delta\mathbf{H}\|_F^2) \end{aligned} \quad (75)$$

240 Since  $\lambda < 0$ , combine the two equations in (71) we know:

$$\lambda \mathbf{W}^\top \mathbf{W} = \mathbf{W}^\top \nabla L(\mathbf{W}\mathbf{H})\mathbf{H}^\top = \lambda \mathbf{H}\mathbf{H}^\top \Rightarrow \mathbf{W}^\top \mathbf{W} = \mathbf{H}\mathbf{H}^\top \quad (76)$$

241 which further implies:

$$\|\mathbf{W}\|_F = \|\mathbf{H}\|_F, \quad \|\mathbf{W}\|_2 = \|\mathbf{H}\|_2 \quad (77)$$

242 On the other hands, we also have (Note that when  $\mathbf{W} = \mathbf{H} = 0$  we must have  $\lambda = 0$ ):

$$\begin{aligned} \nabla L(\mathbf{W}\mathbf{H})\mathbf{H}^\top &= \lambda \mathbf{W} \Rightarrow -\lambda \|\mathbf{W}\|_2 \leq \|\nabla L(\mathbf{W}\mathbf{H})\|_2 \|\mathbf{H}\|_2 \\ &\Rightarrow -\lambda \leq \|\nabla L(\mathbf{W}\mathbf{H})\|_2 \end{aligned}$$

243 Now when  $-\lambda < \|\nabla L(\mathbf{W}\mathbf{H})\|_2$ , we can show that it will contradict with (75): We have shown that  
244 the rank of  $\nabla L(\mathbf{Z})$  is  $K - 1$ , so by (71) and (76) there exist a vector  $\mathbf{a}$  such that  $\mathbf{W}\mathbf{a} = \mathbf{H}^\top \mathbf{a} = 0$ ,  
245 let  $u$  and  $v$  are the left and right singular vectors corresponding to the largest singular value of  
246  $\nabla L(\mathbf{W}\mathbf{H})$ , construct  $\Delta\mathbf{W} = u\mathbf{a}^\top$ ,  $\Delta\mathbf{H} = -\mathbf{a}v^\top$ , then  $(\Delta\mathbf{W}, \Delta\mathbf{H}) \in \mathcal{T}(\mathbf{W}, \mathbf{H})$  and:

$$\begin{aligned} &(\Delta\mathbf{W}, \Delta\mathbf{H})^\top \nabla^2 \mathcal{L}(\mathbf{W}, \mathbf{H})(\Delta\mathbf{W}, \Delta\mathbf{H}) - \lambda(\|\Delta\mathbf{W}\|_F^2 + \|\Delta\mathbf{H}\|_F^2) \\ &= (\mathbf{W}\Delta\mathbf{H} + \Delta\mathbf{W}\mathbf{H})^\top \nabla^2 L(\mathbf{W}\mathbf{H})(\mathbf{W}\Delta\mathbf{H} + \Delta\mathbf{W}\mathbf{H}) + 2\langle \nabla L(\mathbf{W}\mathbf{H}), \Delta\mathbf{W}\Delta\mathbf{H} \rangle - \lambda(\|\Delta\mathbf{W}\|_F^2 + \|\Delta\mathbf{H}\|_F^2) \\ &\leq 2\|\mathbf{a}\|_2^2(-\lambda - u^\top \nabla L(\mathbf{W}\mathbf{H})v) < 0 \end{aligned}$$

247 Then it only remains to analyze the  $-\lambda = \|\nabla L(\mathbf{W}\mathbf{H})\|_2$  cases, construct another convex optimiza-  
248 tion problem:

$$\min_{\mathbf{Z}} L(\mathbf{Z}) - \lambda \|\mathbf{Z}\|_* \quad (78)$$

249 suppose  $\mathbf{Z}$  has SVD  $\mathbf{Z} = \mathbf{U}\Sigma\mathbf{V}^\top$ , as we know that the subgradient of  $\|\mathbf{Z}\|_*$  can be written as (see  
250 [6] for a proof):

$$\partial \|\mathbf{Z}\|_* = \left\{ \mathbf{U}\mathbf{V}^\top + \mathbf{W}, \mathbf{W} \in \mathbb{R}^{K \times nK} \mid \mathbf{U}^\top \mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\|_2 \leq 1 \right\} \quad (79)$$

251 On the other hand, we know that:

$$\begin{aligned} \mathbf{H}^\top \mathbf{H}\mathbf{H}^\top \mathbf{H} &= \mathbf{H}^\top \mathbf{W}^\top \mathbf{W}\mathbf{H} = \mathbf{V}\Sigma^2\mathbf{V}^\top \\ \mathbf{W}\mathbf{W}^\top \mathbf{W}\mathbf{W}^\top &= \mathbf{W}\mathbf{H}\mathbf{H}^\top \mathbf{W}^\top = \mathbf{U}\Sigma^2\mathbf{U}^\top \end{aligned} \quad (80)$$

252 which indicates that  $\mathbf{H}^\top \mathbf{H} = \mathbf{V}\Sigma\mathbf{V}^\top$  and  $\mathbf{W}\mathbf{W}^\top = \mathbf{U}\Sigma\mathbf{U}^\top$ . Combine them with (71) we have:

$$\begin{aligned} \nabla L(\mathbf{W}\mathbf{H})\mathbf{H}^\top \mathbf{H} &= \lambda \mathbf{W}\mathbf{H} \Leftrightarrow \nabla L(\mathbf{W}\mathbf{H})\mathbf{V}\Sigma\mathbf{V}^\top = \lambda \mathbf{U}\Sigma\mathbf{V}^\top \\ &\Leftrightarrow \nabla L(\mathbf{W}\mathbf{H})\mathbf{V} = \lambda \mathbf{U} \\ \mathbf{W}\mathbf{W}^\top \nabla L(\mathbf{W}\mathbf{H}) &= \lambda \mathbf{W}\mathbf{H} \Leftrightarrow \mathbf{U}\Sigma\mathbf{U}^\top \nabla L(\mathbf{W}\mathbf{H}) = \lambda \mathbf{U}\Sigma\mathbf{V}^\top \\ &\Leftrightarrow \mathbf{U}^\top \nabla L(\mathbf{W}\mathbf{H}) = \lambda \mathbf{V}^\top \end{aligned} \quad (81)$$

253 Note that  $-\lambda = \|\nabla L(\mathbf{W}\mathbf{H})\|_2$ , then by (79) we know that  $-\nabla L(\mathbf{W}\mathbf{H}) \in -\lambda \partial \|\mathbf{W}\mathbf{H}\|_*$ . Then  
254 by the strict convexity of (85) we know  $\mathbf{W}\mathbf{H}$  is the global minimum of it. In addition, we have  
255  $\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2 = 2\text{tr}(\Sigma^2) = 2\|\mathbf{W}\mathbf{H}\|_*$ . In addition, previous works [2] have shown that:

$$\|\mathbf{Z}\|_* = \min_{\mathbf{Z}=\mathbf{W}\mathbf{H}} \frac{1}{2} (\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) \quad (82)$$

which is equivalent to:

$$\|\mathbf{W}\mathbf{H}\|_* \leq \frac{1}{2}(\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) \quad (83)$$

Now for any  $(\mathbf{W}', \mathbf{H}')$ , we know that:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \mathbf{H}) - \frac{\lambda}{2}(\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) &= L(\mathbf{W}\mathbf{H}) - \lambda\|\mathbf{W}\mathbf{H}\|_* \leq L(\mathbf{W}'\mathbf{H}') - \lambda\|\mathbf{W}'\mathbf{H}'\|_* \\ &\leq \mathcal{L}(\mathbf{W}', \mathbf{H}') - \frac{\lambda}{2}(\|\mathbf{W}'\|_F^2 + \|\mathbf{H}'\|_F^2) \end{aligned} \quad (84)$$

which indicates  $(\mathbf{W}, \mathbf{H})$  must attained global minimum of the following optimization problem:

$$\min_{\mathbf{W}, \mathbf{H}} \mathcal{L}(\mathbf{W}, \mathbf{H}) - \frac{\lambda}{2}(\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) \quad (85)$$

If  $(\mathbf{W}, \mathbf{H})$  doesn't satisfies neural collapse conditions, by Theorem C.1 we know there exists another point  $(\mathbf{W}', \mathbf{H}')$  such that  $\mathcal{L}(\mathbf{W}', \mathbf{H}') < \mathcal{L}(\mathbf{W}, \mathbf{H})$  and  $\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2 = \|\mathbf{W}'\|_F^2 + \|\mathbf{H}'\|_F^2$  thus:

$$\mathcal{L}(\mathbf{W}, \mathbf{H}) - \frac{\lambda}{2}(\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2) > \mathcal{L}(\mathbf{W}', \mathbf{H}') - \frac{\lambda}{2}(\|\mathbf{W}'\|_F^2 + \|\mathbf{H}'\|_F^2) \quad (86)$$

which contradicts with the global optimality of  $(\mathbf{W}, \mathbf{H})$ , thus  $(\mathbf{W}, \mathbf{H})$  must satisfy all of the neural collapse conditions and we finish the proof.  $\square$

## D Experiment details and additional results

In addition to the experiments in Section 4, we also train a ResNet18 [3] on FashionMNIST, a VGG-13 and another ResNet18 on CIFAR-10 dataset [4]. All the networks are trained for 500 epochs, using a stochastic gradient descent with learning rate 0.01, momentum 0.3, batch size 128 and in particular, without weight decay. The results are plotted in Figure 4, Figure 5 and Figure 6. Again, we observe that in all three experiments, after 100 epochs, the variation of norms become small after 500 epochs; the with-in class variation decreases at rate  $O(1/\log(t))$ ; the cosines between pairs of last layer features and that of the classifiers converge to the equiangular state with maximum angles at rate  $O(1/\log(t))$ ; The distance between normalized centered classifier and normalized last layer feature decreases at rate  $O(1/\log(t))$ . All the experiments are run in Python (version 3.6.9) on Google Colab.

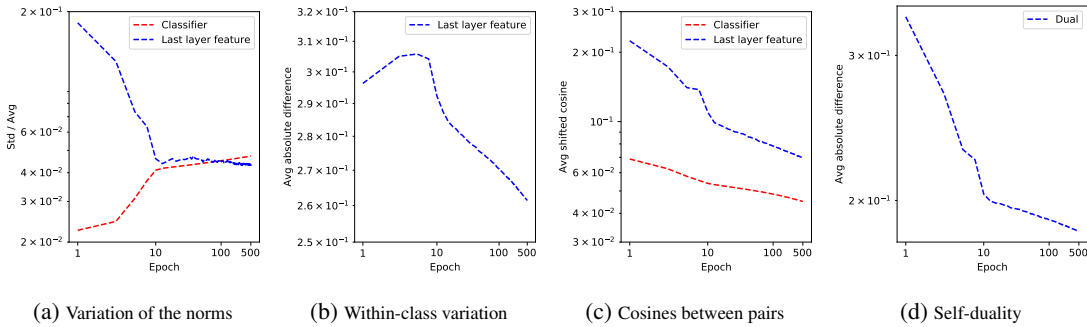


Figure 4: Training ResNet18 without weight decay on FashionMNIST. The scale of the axes are set to be the same as that in Figure 3. The patterns of the curves are also similar to those in Figure 3.

## References

- [1] J. Dutta, K. Deb, Rupesh Tulshyan, and Ramnik Arora. Approximate kkt points and a proximity measure for termination. *Journal of Global Optimization*, 56:1463–1499, 2013.
- [2] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

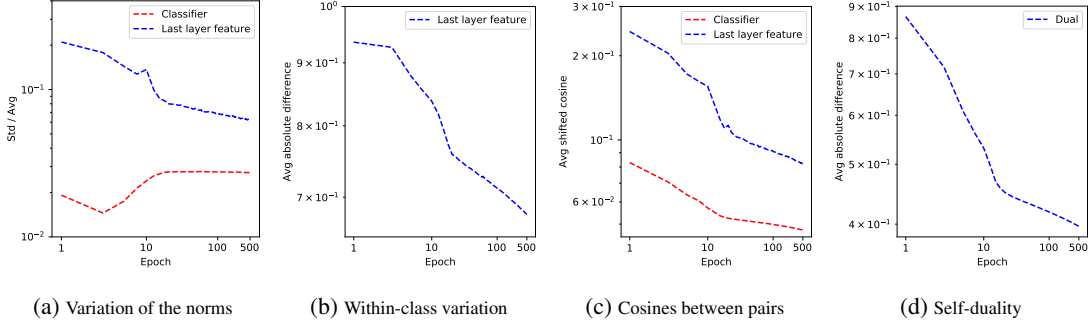


Figure 5: Training VGG-13 without weight decay on CIFR-10. The scale of the axes are set to be the same as that in Figure 3.

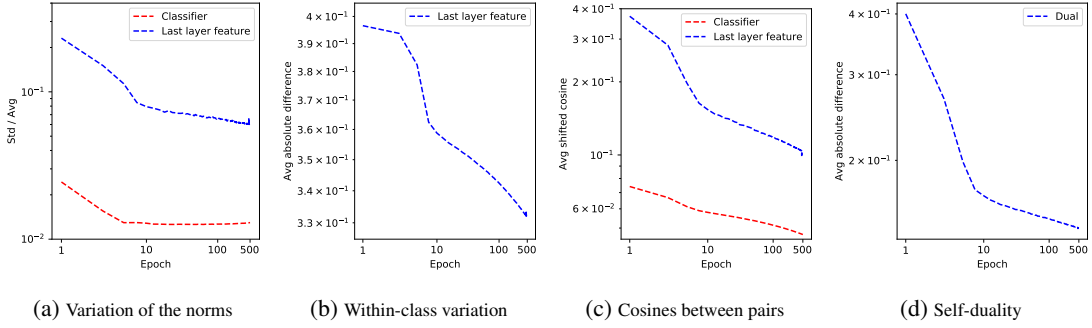


Figure 6: Training ResNet18 without weight decay on CIFAR-10. The scale of the axes are set to be the same as that in Figure 3.

- 281 [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
282 recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,  
283 pages 770–778, 2016.
- 284 [4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.  
285 2009.
- 286 [5] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural  
287 networks. *arXiv preprint arXiv:1906.05890*, 2019.
- 288 [6] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra*  
289 *and its applications*, 170:33–45, 1992.