

A Experimental Details

A.1 Experimental details on classifying MNIST digits

Preprocessing data. We first preprocess the training data, i.e. digits $\{0, 1, 2\}$ by centering: $\mathbf{x}_i \leftarrow \mathbf{x}_i - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = \sum_{i \in [n]} \mathbf{x}_i / n$ is the global mean image of the entire training data. Then we have plotted the normalized correlation matrix $\left[\left\langle \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, \frac{\mathbf{x}_{i'}}{\|\mathbf{x}_{i'}\|} \right\rangle \right]_{i, i' \in [n]}$ of the centered data, showing in Figure 2(a) that two data points of the same digit are likely to have a positive correlation and those different digits are likely to have negative correlations. This suggests that the orthogonality separability assumption in Assumption 1 is approximately satisfied.

Training. Given the centered data, for a two-layer ReLU network (2) of width-50, we initialize all entries of the network weights with i.i.d. Gaussians with variance 10^{-6} . Then we run SGD of batch size 1000 with learning rate 0.1 for 50 epochs. For the trained network, we visualize the output neuron weights $\mathbf{v}_j, j \in [h]$ and determine the \mathcal{N}_k by letting $\mathcal{N}_k = \{j \in [h] : k = \arg \max_{k'} \langle \mathbf{e}_{k'}, \mathbf{v}_j \rangle\}$, then also visualize the average direction of the input neuron weights $\frac{\sum_{j \in \mathcal{N}_k} \mathbf{w}_j}{\|\sum_{j \in \mathcal{N}_k} \mathbf{w}_j\|}$ for each group \mathcal{N}_k , as shown in Figure 2(b).

A.2 Experimental details on normalization layers

Modified ResNet. We take the ResNet18 and ResNet50 implementations (The first conv layer is modified to accommodate MNIST and CIFAR10 input sizes) in Pytorch and replace the final linear classifier with a two-layer ReLU network of width-1000, and also add a normalization layer (Identity/None, LayerNorm, or RMSNorm) between the classifier and the feature extractor. The initialization follows the Pytorch default.

Training. For each choice of (model: ResNet18, ResNet50)-(Dataset: MNIST, CIFAR10), we repeat 5 runs (with different random seeds) of SGD of batch size 128 and learning rate 0.1 (for ResNet18) and 0.02 (for ResNet50) with momentum 0.95 for 50 epochs; and for every 20 epochs, we reduce the learning rate to 0.1 of its current value. We plot the NC metrics and test accuracy against training epochs in Figure 3.

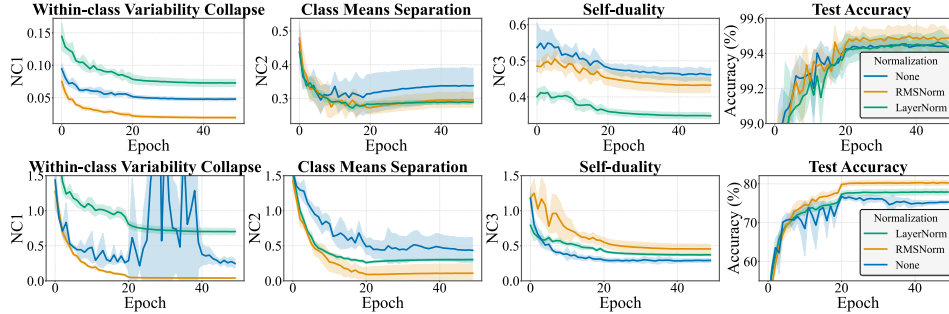


Figure 4: Measuring NC in trained modified ResNet50 on MNIST (top) and CIFAR10 (bottom)

NC metrics. The NC metrics follow those used in prior works except for the projected self-duality. Given the class means $\phi_k = \frac{\sum_{i \in \mathcal{I}_k} \phi_\theta(\mathbf{x}_i)}{|\mathcal{I}_k|}$ and global mean $\bar{\phi} = \sum_{k=1}^K \phi_k / K$, NC1 is defined to be the ratio between intra-class variance and the inter-class variance $\frac{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \|\phi_\theta(\mathbf{x}_i) - \phi_k\|^2 / |\mathcal{I}_k|}{\sum_{k=1}^K \|\phi_k - \bar{\phi}\|^2}$. NC2 is defined to be the proximity of the gram matrix of the class mean directions to the identity matrix $\|\frac{\mathbf{G}}{\|\mathbf{G}\|_F} - \frac{1}{\sqrt{K}} \mathbf{I}\|_F$, where $\mathbf{G} = [\bar{\phi}_1 \ \cdots \ \bar{\phi}_K]^\top [\bar{\phi}_1 \ \cdots \ \bar{\phi}_K]$. NC3 is defined to be the proximity of $\mathbf{V} \bar{\Phi}^\dagger$ to an identity matrix $\|\frac{\mathbf{V} \bar{\Phi}^\dagger}{\|\mathbf{V} \bar{\Phi}^\dagger\|_F} - \frac{1}{\sqrt{K-1}} \tilde{\mathbf{E}}\|_F$.

In the main paper, we have only provided the plot for ResNet18. We show the plot for ResNet50 in Figure 4.

834 B Neural Alignment under Multi-class Orthogonally Separable Data

835 B.1 Basics on neuron dynamics under multi-class problems

836 The differential inclusion $\dot{\theta} \in -\nabla_{\theta} \mathcal{L}(\theta)$ gives rise to the following characterization of the time
837 derivatives of neuron weights $\forall j \in [h]$:

$$\frac{d}{dt} \mathbf{w}_j = \sum_{i=1}^n \xi_{ij} \langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \mathbf{v}_j \rangle \mathbf{x}_i, \quad (20)$$

$$\frac{d}{dt} \mathbf{v}_j = \sum_{i=1}^n \xi_{ij} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \langle \mathbf{x}_i, \mathbf{w}_j \rangle, \quad (21)$$

$$\text{where } \xi_{ij} \begin{cases} = 1, & \langle \mathbf{x}_i, \mathbf{w}_j \rangle > 0 \\ \in [0, 1], & \langle \mathbf{x}_i, \mathbf{w}_j \rangle = 0, \ \hat{\mathbf{y}}_i = \text{Softmax}(\mathbf{f}(\mathbf{x}_i; \theta)) \\ = 0, & \langle \mathbf{x}_i, \mathbf{w}_j \rangle < 0 \end{cases}$$

838 It will become clear soon that it is convenient to decompose the weight dynamics into those of the
839 weight norm and of the weight direction, for which we use the balancedness that $\|\mathbf{w}_j\| \equiv \|\mathbf{v}_j\|, \forall j$:

(weight norm dynamics)

$$\begin{aligned} \frac{d}{dt} \|\mathbf{w}_j\|^2 \left(\text{also } \frac{d}{dt} \|\mathbf{v}_j\|^2 \right) &= 2 \sum_{i=1}^n \xi_{ij} \langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \mathbf{v}_j \rangle \langle \mathbf{x}_i, \mathbf{w}_j \rangle \\ &= 2 \sum_{i=1}^n \xi_{ij} \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \|\mathbf{w}_j\|^2; \end{aligned} \quad (22)$$

(input neuron angular dynamics)

$$\begin{aligned} \frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} &= \Pi_{\mathbf{w}_j}^{\perp} \frac{1}{\|\mathbf{w}_j\|} \sum_{i=1}^n \xi_{ij} \langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \mathbf{v}_j \rangle \mathbf{x}_i \\ &= \Pi_{\mathbf{w}_j}^{\perp} \sum_{i=1}^n \xi_{ij} \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i; \end{aligned} \quad (23)$$

(output neuron angular dynamics)

$$\begin{aligned} \frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} &= \Pi_{\mathbf{v}_j}^{\perp} \frac{1}{\|\mathbf{v}_j\|} \sum_{i=1}^n \xi_{ij} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \langle \mathbf{x}_i, \mathbf{w}_j \rangle \\ &= \Pi_{\mathbf{v}_j}^{\perp} \sum_{i=1}^n \xi_{ij} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle (\mathbf{y}_i - \hat{\mathbf{y}}_i), \end{aligned} \quad (24)$$

840 where $\Pi_{\mathbf{u}}^{\perp} := \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^{\top}}{\|\mathbf{u}\|^2} \right)$ denote the project matrix onto the orthogonal complement of \mathbf{u} .

841 By inspecting (22)(23)(24), we note that the dynamic of each neuron pair $(\mathbf{w}_j, \mathbf{v}_j)$ is almost de-
842 coupled from each other except for the interaction through $\hat{\mathbf{y}}_i, i \in [n]$. Interestingly, at the early
843 phase of the GF, (we will show that) the norm of the weights remains close to zero, resulting in
844 $\hat{\mathbf{y}}_i \simeq \frac{1}{K} \mathbf{1}, \forall i \in [n]$ thus fully decouples the neuron pair dynamics, the precise statement on such an
845 approximation $\hat{\mathbf{y}}_i \simeq \frac{1}{K} \mathbf{1}$ is as follow:

846 **Lemma 1.** $\|\hat{\mathbf{y}}_i - \frac{1}{K} \mathbf{1}\| \leq \frac{8}{\sqrt{K}} \|\mathbf{f}(\mathbf{x}_i; \theta)\|$ whenever $\|\mathbf{f}(\mathbf{x}_i; \theta)\| \leq \frac{1}{4}$.

847 *Proof.* First of all, we have

$$|\exp(z) - 1| = \max\{\exp(z) - 1, 1 - \exp(z)\} = \begin{cases} \exp(|z|) - 1, & z \geq 0 \\ 1 - \exp(-|z|), & z < 0 \end{cases}.$$

848 We always have $1 - \exp(-|z|) \leq |z|$. Moreover, whenever $|z| \leq 1$, we have $\exp(|z|) - 1 \leq 2|z|$.
849 Therefore, we conclude that

$$|\exp(z) - 1| \leq 2|z|, \ \forall |z| \leq 1. \quad (25)$$

850 With (25), whenever $\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\| \leq 1$, we have

$$\max_k |\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_k) - 1| \leq 2 \max_k |[\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_k| \leq 2\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\|. \quad (26)$$

851 Now we bound $\|\hat{\mathbf{y}}_i - \frac{1}{K}\mathbf{1}\|$ using entrywise bound. Notice that

$$\begin{aligned} |[\hat{\mathbf{y}}_i]_k - \frac{1}{K}| &= \left| \frac{\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_k)}{\sum_{k'=1}^K \exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_{k'})} - \frac{1}{K} \right| \\ &= \left| \frac{1 + (\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_k) - 1)}{K + \sum_{k'=1}^K (\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_{k'}) - 1)} - \frac{1}{K} \right| \\ &= \left| \frac{K(\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_k) - 1) - \sum_{k'=1}^K (\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_{k'}) - 1)}{K(K + \sum_{k'=1}^K (\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_{k'}) - 1))} \right| \\ &\leq \frac{K|\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_k) - 1| + \sum_{k'=1}^K |\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_{k'}) - 1|}{K(K - \sum_{k'=1}^K |\exp([\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})]_{k'}) - 1|)} \\ &\stackrel{(26)}{\leq} \frac{4K\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\|}{K(K - 2K\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\|)} \stackrel{(\|\mathbf{f}\| \leq \frac{1}{4})}{\leq} \frac{8}{K}\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\|. \end{aligned} \quad (27)$$

852 Finally, we have $\|\hat{\mathbf{y}}_i - \frac{1}{K}\mathbf{1}\| \leq \sqrt{K} \max_k |[\hat{\mathbf{y}}_i]_k - \frac{1}{K}| \leq \frac{8}{\sqrt{K}}\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\|$. \square

853 B.2 Analyzing neuron dynamics during alignment phase

854 In this section, we show the formal statements for the alignment dynamics we have introduced in
855 (16)(17). During the early phase of the GF training, the norms of the weights remain small (Lemma
856 2), leading to an approximate alignment dynamics in Lemma 3, which will be crucial for subsequent
857 analysis.

858 **Lemma 2.** *Given some balanced, ϵ -small initialization $\boldsymbol{\theta}(0)$ with $\epsilon \leq \frac{\sqrt{K}}{16X_{\max}\sqrt{h}}$, any solution $\boldsymbol{\theta}(t)$
859 to the GF dynamics (3) satisfies that $\forall t \leq \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon} := T$,*

$$\|\mathbf{w}_j(t)\|^2 = \|\mathbf{v}_j(t)\|^2 \leq \frac{\epsilon}{\sqrt{h}}, \quad \forall j \in [h], \quad \|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta}(t))\| \leq 2\epsilon X_{\max}\sqrt{h}. \quad (28)$$

860 The alignment phase refers to the training phase until $T = \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon}$. With Lemma 2, we can
861 approximate the angular dynamics $\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ and $\frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$ throughout the alignment phase as follow:

862 **Lemma 3.** *Given some balanced, ϵ -small initialization $\boldsymbol{\theta}(0)$ with $\epsilon \leq \frac{1}{16X_{\max}\sqrt{h}}$, any solution $\boldsymbol{\theta}(t)$
863 to the GF dynamics (3) satisfies that $\forall t \leq \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon} := T$,*

$$\begin{aligned} \left\| \frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} - \Pi_{\mathbf{w}_j}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i=1}^n \xi_{ij} \left\langle \tilde{\mathbf{E}}\mathbf{y}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right) \right\| &\leq \frac{16}{\sqrt{K}} \epsilon n X_{\max} \sqrt{h}, \quad \forall j \in [h], \\ \left\| \frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} - \Pi_{\mathbf{v}_j}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i=1}^n \xi_{ij} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \tilde{\mathbf{E}}\mathbf{y}_i \right) \right\| &\leq \frac{16}{\sqrt{K}} \epsilon n X_{\max} \sqrt{h}, \quad \forall j \in [h] \end{aligned}$$

865 *Proof of Lemma 2* From Section B.1 we have

$$\frac{d}{dt} \|\mathbf{w}_j\|^2 = 2 \sum_{i=1}^n \xi_{ij} \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \|\mathbf{w}_j\|^2 \quad (29)$$

866 Let $T := \inf\{t : \max_i |\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta}(t))| > 2\epsilon X_{\max}\sqrt{h}\}$, then $\forall t \leq T, j \in [h]$, we have

$$\begin{aligned} \frac{d}{dt} \|\mathbf{w}_j\|^2 &= 2 \sum_{i=1}^n \xi_{ij} \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \|\mathbf{w}_j\|^2 \\ &\leq 2 \sum_{i=1}^n |\xi_{ij}| \left| \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \right| \left| \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \right| \|\mathbf{w}_j\|^2 \end{aligned}$$

$$\begin{aligned}
&\leq 2 \sum_{i=1}^n \|\mathbf{y}_i - \hat{\mathbf{y}}_i\| \|\mathbf{x}_i\| \|\mathbf{w}_j\|^2 \\
&\leq 2 \sum_{i=1}^n \left(\|\mathbf{y}_i - \frac{1}{K} \mathbf{1}\| + \|\frac{1}{K} \mathbf{1} - \hat{\mathbf{y}}_i\| \right) X_{\max} \|\mathbf{w}_j\|^2 \\
&\leq 2 \sum_{i=1}^n \left(\sqrt{\frac{K-1}{K}} + \frac{8}{\sqrt{K}} \|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\| \right) X_{\max} \|\mathbf{w}_j\|^2 \\
&\leq 2 \sum_{i=1}^n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) \|\mathbf{w}_j\|^2 \leq 2n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) \|\mathbf{w}_j\|^2 \quad (30)
\end{aligned}$$

867 Let $\tau_j := \inf\{t : \|\mathbf{w}_j(t)\|^2 > \frac{\epsilon}{\sqrt{h}}\}$, and let $j^* := \arg \min_j \tau_j$. Then $\tau_{j^*} = \min_j \tau_j \leq T$, which can
868 be shown by contradiction:

869 Suppose $\tau_{j^*} > T$, then at $t = T < \tau_{j^*}$, by the definition of τ_{j^*} , we have $\max_j \|\mathbf{w}_j\|^2 \leq \frac{\epsilon}{\sqrt{h}}$,
870 and by the definition of T and the continuity of $\boldsymbol{\theta}(t)$ w.r.t. t , $\exists i^* \in [n]$ such that $|\mathbf{f}(\mathbf{x}_{i^*}; \boldsymbol{\theta}(T))| =$
871 $2\epsilon X_{\max} \sqrt{h}$, therefore,

$$\begin{aligned}
2\epsilon X_{\max} \sqrt{h} = |\mathbf{f}(\mathbf{x}_{i^*}; \boldsymbol{\theta}(T))| &= \left| \sum_{j \in [h]} \xi_{i^*j} \mathbf{v}_j \langle \mathbf{w}_j, \mathbf{x}_{i^*} \rangle \right| \\
&\leq \sum_{j \in [h]} |\xi_{i^*j}| \|\mathbf{v}_j\| \|\mathbf{w}_j\| \|\mathbf{x}_{i^*}\| \\
&\leq \sum_{j \in [h]} X_{\max} \|\mathbf{w}_j\|^2 \leq h X_{\max} \max_{j \in [h]} \|\mathbf{w}_j\|^2, \quad (31)
\end{aligned}$$

872 which suggests that $\max_{j \in [h]} \|\mathbf{w}_j\|^2 \geq \frac{2\epsilon}{\sqrt{h}}$, a contradiction.

873 Now for $t \leq \tau_{j^*} \leq T$, we have

$$\frac{d}{dt} \|\mathbf{w}_{j^*}\|^2 \leq 2n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) \|\mathbf{w}_{j^*}\|^2. \quad (32)$$

874 By Grönwall's inequality, we have $\forall t \leq \tau_{j^*}$

$$\|\mathbf{w}_{j^*}(t)\|^2 \leq \exp \left(2n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) t \right) \|\mathbf{w}_{j^*}(0)\|^2 \leq \exp \left(2n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) t \right) \epsilon^2.$$

875 Suppose $\tau_{j^*} < \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon}$, then by the continuity of $\|\mathbf{w}_{j^*}(t)\|^2$, we have

$$\begin{aligned}
\frac{\epsilon}{\sqrt{h}} = \|\mathbf{w}_{j^*}(\tau_{j^*})\|^2 &\leq \exp \left(2n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) \tau_{j^*} \right) \epsilon^2 \\
&< \exp \left(2n \left(X_{\max} + \frac{16\epsilon X_{\max}^2 \sqrt{h}}{\sqrt{K}} \right) \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon} \right) \epsilon^2 \\
&\leq \exp \left(\left(\frac{1}{2} + \frac{8\epsilon X_{\max} \sqrt{h}}{\sqrt{K}} \right) \log \frac{1}{\sqrt{h}\epsilon} \right) \epsilon^2 \\
&\leq \exp \left(\log \frac{1}{\sqrt{h}\epsilon} \right) \epsilon^2 = \frac{\epsilon}{\sqrt{h}},
\end{aligned}$$

876 where the last inequality is due to $\epsilon \leq \frac{\sqrt{K}}{16X_{\max} \sqrt{h}}$. This leads to a contradiction. Therefore, one must

877 have $T \geq \tau_{j^*} \geq \frac{1}{4nX_{\max}} \log \left(\frac{1}{\sqrt{h}\epsilon} \right)$. This finishes the proof. \square

878 *Proof of Lemma 3* We have shown in Section B.1 that

$$\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} = \Pi_{\mathbf{w}_j}^\perp \left(\sum_{i=1}^n \xi_{ij} \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right). \quad (33)$$

879 Therefore, $\forall \leq T$,

$$\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} = \Pi_{\mathbf{w}_j}^\perp \left(\sum_{i=1}^n \xi_{ij} \left\langle \mathbf{y}_i - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right)$$

$$\begin{aligned}
&= \Pi_{\mathbf{w}_j}^\perp \left(\sum_{i=1}^n \xi_{ij} \left\langle \underbrace{\left(\mathbf{y}_i - \frac{1}{K} \mathbb{1} \right)}_{=\sqrt{\frac{K-1}{K}} \tilde{\mathbf{E}} \mathbf{y}_i} + \left(\frac{1}{K} \mathbb{1} - \hat{\mathbf{y}}_i \right), \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right) \\
&= \Pi_{\mathbf{w}_j}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i=1}^n \xi_{ij} \left\langle \tilde{\mathbf{E}} \mathbf{y}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right) + \Pi_{\mathbf{w}_j}^\perp \left(\sum_{i=1}^n \xi_{ij} \left\langle \left(\frac{1}{K} \mathbb{1} - \hat{\mathbf{y}}_i \right), \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\left\| \frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} - \Pi_{\mathbf{w}_j}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i=1}^n \xi_{ij} \left\langle \tilde{\mathbf{E}} \mathbf{y}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right) \right\| \\
&= \left\| \Pi_{\mathbf{w}_j}^\perp \left(\sum_{i=1}^n \xi_{ij} \left\langle \left(\frac{1}{K} \mathbb{1} - \hat{\mathbf{y}}_i \right), \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i \right) \right\| \\
&\leq \sum_{i=1}^n |\xi_{ij}| \|\mathbf{x}_i\| \left\| \frac{1}{K} \mathbb{1} - \hat{\mathbf{y}}_i \right\| \stackrel{(\text{Lemma 1})}{\leq} \frac{8}{\sqrt{K}} n X_{\max} \|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\| \stackrel{(\text{Lemma 2})}{\leq} \frac{16}{\sqrt{K}} \epsilon n X_{\max}^2 \sqrt{h}, \quad (34)
\end{aligned}$$

where we note that applying Lemma 1 requires $\|\mathbf{f}(\mathbf{x}_i; \boldsymbol{\theta})\| \leq \frac{1}{4}$, which is guaranteed by Lemma 2 and our choice $\epsilon \leq \frac{1}{16 X_{\max} \sqrt{h}}$. We have shown the approximation error bound for $\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$. A similar bound can be derived for $\frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$. \square

B.3 Neural alignment under multi-class orthogonally separable data

Sufficient statement for Proposition 1 It is easy to check that the following proposition is sufficient for Proposition 1 to hold.

Proposition 3 (Sufficient statement for Proposition 1). *Let $K > 2$. Given orthogonally separable data (Assumption 1) with $\frac{X_{\max}^2}{X_{\min}^2 \mu_d \mu_s^2} < 2K - 3$, where $X_{\max} := \max_{i \in [n]} \|\mathbf{x}_i\|$ and $X_{\min} := \min_{i \in [n]} \|\mathbf{x}_i\|$, ϵ -small, balanced and semi-local initialization (Assumptions 2, 4) for a sufficiently small ϵ , for any solution $\boldsymbol{\theta}(t), t \geq 0$ to (3) and any $j \in \mathcal{N}_k, k \in [K]$, define*

$$T_{j,k}^* = \inf \{ t \geq 0 : |\mathcal{I}_k^{\mathbf{w}_j}| = |\mathcal{I}_k|, |\mathcal{I}_{k'}^{\mathbf{w}_j}| = 0, \forall k' \neq k \}, \quad (35)$$

then $\forall t \geq T_{j,k}^*$, we have $|\mathcal{I}_k^{\mathbf{w}_j}| = |\mathcal{I}_k|, |\mathcal{I}_{k'}^{\mathbf{w}_j}| = 0, \forall k' \neq k$, thus $\mathbf{w}_k(t)^\top \mathbf{x}_i \begin{cases} > 0, & \forall i \in \mathcal{I}_k \\ < 0, & \forall i \notin \mathcal{I}_k \end{cases}, \forall i \in [n]$.

Therefore, we can study the dynamic behavior of each neuron pair individually, for convenience, let $j \in \mathcal{N}_k$, and we drop the index j .

For a neuron pair (\mathbf{w}, \mathbf{v}) , we have defined the following:

$$\begin{aligned}
\alpha_i &:= \left\langle \mathbf{x}_i, \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\rangle, & (\text{alignment between input neuron and } i\text{-th data}) \\
\beta_i &:= \left\langle \tilde{\mathbf{E}} \mathbf{y}_i, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle, & (\text{alignment between output neuron and } i\text{-th label}) \\
\mathcal{I}_k^{\mathbf{w}} &:= \{i \in \mathcal{I}_k : \alpha_i > 0\}, & (\text{number of active data points in } k\text{-th class})
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{A}_k &:= \sum_{i \in \mathcal{I}_k^{\mathbf{w}}} \alpha_i = \underbrace{\sum_{i \in \mathcal{I}_k : \alpha_i > 0} \alpha_i}_{\text{we mostly use this notation for clarity}}, & (\text{alignment between input neuron and } k\text{-th class}) \\
\mathcal{B}_k &:= \left\langle \tilde{\mathbf{e}}_k, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle. & (\text{alignment between output neuron and } k\text{-th class})
\end{aligned}$$

Overview of the proof of Proposition 1 First, we utilize the alignment dynamics in Lemma 3 to show that (Recall that $T = \frac{1}{4n X_{\max}} \log \frac{1}{\sqrt{h\epsilon}}$)

Lemma 4. *Given a neuron $(\mathbf{w}_j, \mathbf{v}_j), j \in \mathcal{N}_k$, during the alignment phase $t \leq \min\{T_{j,k}^*, T\}$, the following holds:*

- 900 1. $\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$ remains close to its target pseudo-label: $\mathcal{B}_k \geq 1 - \frac{1}{2(K-1)}$;
 901 2. $\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ remains close to its target class: $\mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'} \geq \mathcal{A}_k(0) - 2 \sum_{k' \neq k} \mathcal{A}_{k'}(0)$;
 902 3. Neuron $\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ does not deactivate data from target class, nor activate data from non-target class:
 903 $|\mathcal{I}_k^{\mathbf{w}}| \geq |\mathcal{I}_k^{\mathbf{w}(0)}|$ and $|\mathcal{I}_{k'}^{\mathbf{w}}| \leq |\mathcal{I}_{k'}^{\mathbf{w}(0)}|, \forall k' \neq k$.

904 The characterizations in [4] suggest that the neuron weight directions $\{\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}\}$ remains close
 905 to the attractor $\{\bar{\mathbf{x}}_k, \bar{\mathbf{e}}_k\}$, and as the weights move closer to the attractor, $|\mathcal{I}_k^{\mathbf{w}_j}|$ increases to $|\mathcal{I}_k|$,
 906 and $|\mathcal{I}_{k'}^{\mathbf{w}_j}|$ decreases to 0. When the initialization scale ϵ is sufficiently small so that T is large, this
 907 Lemma will show that $T_{j,k}^*$ is finite, and we will provide an upper bound.

908 Then the following lemma shows that the desired property for neuron $(\mathbf{w}_j, \mathbf{v}_j)$ still holds after $T_{j,k}^*$.

909 **Lemma 5.** for any neuron $(\mathbf{v}_j, \mathbf{w}_j), j \in \mathcal{N}_k$, we have $\forall t > T_{j,k}^*$:

- 910 1. $\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$ remains close to its target pseudo-label: $\mathcal{B}_k^{\mathbf{v}_j} \geq \frac{\sqrt{2}}{2}$;
 911 2. $\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ is exclusively activated by data from its target class: $|\mathcal{I}_k^{\mathbf{w}_j}| = N_k, |\mathcal{I}_{k'}^{\mathbf{w}_j}| = 0, \forall k' \neq k$.

912 The remaining parts of this section are dedicated to proving these two Lemmas. The next section will
 913 formally prove Proposition [3], thereby proving Proposition [1].

914 B.3.1 Proof of Lemma [4]

915 **Basic dynamics.** The main proof concerns the time derivatives of the alignment to classes
 916 $\frac{d}{dt} \mathcal{A}_k, \frac{d}{dt} \mathcal{B}_k$. With Lemma [3], we have their approximations during the alignment phase:

$$\begin{aligned} \frac{d}{dt} \mathcal{A}_k &= \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \frac{d}{dt} \alpha_i \\ &= \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \left\langle \mathbf{x}_i, \frac{d}{dt} \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\rangle \\ &= \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \left\langle \mathbf{x}_i, \Pi_{\mathbf{w}}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i'=1}^n \xi_{i'} \left\langle \tilde{\mathbf{E}} \mathbf{y}_{i'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \mathbf{x}_{i'} \right) \right\rangle + \mathcal{O}(\epsilon) \end{aligned} \quad (36)$$

$$\begin{aligned} &= \sqrt{\frac{K-1}{K}} \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \left\langle \mathbf{x}_i, \Pi_{\mathbf{w}}^\perp \left(\sum_{i'=1}^n \xi_{i'} \beta_{i'} \mathbf{x}_{i'} \right) \right\rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \sum_{i': \alpha_{i'} \geq 0} (\langle \mathbf{x}_i, \xi_{i'} \mathbf{x}_{i'} \rangle - \alpha_i \xi_{i'} \alpha_{i'}) \beta_{i'} + \mathcal{O}(\epsilon) \end{aligned} \quad (37)$$

$$\begin{aligned} &= \sqrt{\frac{K-1}{K}} \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \sum_{1 \leq k' \leq K} \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} (\langle \mathbf{x}_i, \xi_{i'} \mathbf{x}_{i'} \rangle - \alpha_i \xi_{i'} \alpha_{i'}) \beta_{i'} + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} (\langle \mathbf{x}_i, \xi_{i'} \mathbf{x}_{i'} \rangle - \alpha_i \xi_{i'} \alpha_{i'}) \mathcal{B}_{k'} + \mathcal{O}(\epsilon) \end{aligned} \quad (38)$$

$$\begin{aligned} &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \mathcal{B}_{k'} \left(\sum_{i \in \mathcal{I}_k: \alpha_i > 0} \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} (\langle \mathbf{x}_i, \xi_{i'} \mathbf{x}_{i'} \rangle - \alpha_i \xi_{i'} \alpha_{i'}) \right) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \mathcal{B}_{k'} \left(\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle \right. \\ &\quad \left. - \left(\sum_{i \in \mathcal{I}_k: \alpha_i > 0} \alpha_i \right) \left(\sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \alpha_{i'} \right) \right) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \mathcal{B}_{k'} \left(\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle - \mathcal{A}_k \mathcal{A}_{k'} \right) + \mathcal{O}(\epsilon), \end{aligned} \quad (39)$$

917 where (39) uses the simple fact that $\sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \alpha_{i'} = \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} > 0} \alpha_{i'} = \mathcal{A}_{k'}$, (38) uses the
 918 fact that $\beta_{i'} = \left\langle \tilde{\mathbf{E}} \mathbf{y}_{i'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle = \left\langle \tilde{\mathbf{e}}_{k'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle$ if $i' \in \mathcal{I}_{k'}$, (37) uses the fact that $\xi_{i'} = 0$ if $\alpha_{i'} < 0$, and
 919 (36) uses Lemma [3] with the $\mathcal{O}(\epsilon)$ term being

$$\sum_{i \in \mathcal{I}_k: \alpha_i > 0} \left\langle \mathbf{x}_i, \frac{d}{dt} \frac{\mathbf{w}}{\|\mathbf{w}\|} - \Pi_{\mathbf{w}}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i'=1}^n \xi_{i'} \left\langle \tilde{\mathbf{E}} \mathbf{y}_{i'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \mathbf{x}_{i'} \right) \right\rangle,$$

whose norm can be upper bounded as follows:

$$\begin{aligned} & \left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \left\langle \mathbf{x}_i, \frac{d}{dt} \frac{\mathbf{w}}{\|\mathbf{w}\|} - \Pi_{\mathbf{w}}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i'=1}^n \xi_{i'} \left\langle \tilde{\mathbf{E}} \mathbf{y}_{i'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \mathbf{x}_{i'} \right) \right\rangle \right\| \\ & \leq \sum_{i=1}^n \|\mathbf{x}_i\| \left\| \frac{d}{dt} \frac{\mathbf{w}}{\|\mathbf{w}\|} - \Pi_{\mathbf{w}}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i'=1}^n \xi_{i'} \left\langle \tilde{\mathbf{E}} \mathbf{y}_{i'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \mathbf{x}_{i'} \right) \right\| \leq \frac{16}{\sqrt{K}} \epsilon n^2 X_{\max}^3 \sqrt{h}. \end{aligned}$$

and

$$\begin{aligned} \frac{d}{dt} \mathcal{B}_k &= \left\langle \tilde{\mathbf{e}}_k, \frac{d}{dt} \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \\ &= \left\langle \tilde{\mathbf{e}}_k, \Pi_{\mathbf{v}}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i=1}^n \xi_i \left\langle \mathbf{x}_i, \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\rangle \tilde{\mathbf{E}} \mathbf{y}_i \right) \right\rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left\langle \tilde{\mathbf{e}}_k, \Pi_{\mathbf{v}}^\perp \left(\sum_{1 \leq k' \leq K} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \alpha_i \tilde{\mathbf{E}} \mathbf{y}_i \right) \right\rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left\langle \tilde{\mathbf{e}}_k, \Pi_{\mathbf{v}}^\perp \left(\sum_{1 \leq k' \leq K} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \alpha_i \tilde{\mathbf{e}}_{k'} \right) \right\rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \left\langle \tilde{\mathbf{e}}_k, \Pi_{\mathbf{v}}^\perp \tilde{\mathbf{e}}_{k'} \right\rangle \alpha_i + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} (\langle \tilde{\mathbf{e}}_k, \tilde{\mathbf{e}}_{k'} \rangle - \mathcal{B}_k \mathcal{B}_{k'}) \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \alpha_i + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \mathcal{A}_{k'} (\langle \tilde{\mathbf{e}}_k, \tilde{\mathbf{e}}_{k'} \rangle - \mathcal{B}_k \mathcal{B}_{k'}) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k (1 - \mathcal{B}_k^2) + \sum_{k' \neq k} \mathcal{A}_{k'} \left(-\frac{1}{K-1} - \mathcal{B}_k \mathcal{B}_{k'} \right) \right) + \mathcal{O}(\epsilon), \end{aligned} \quad (40)$$

where multiple facts used to derive (39) are also used here, and in (40), the $\mathcal{O}(\epsilon)$ has its norm upper bounded by $\frac{16}{\sqrt{K}} \epsilon n X_{\max}^2 \sqrt{h}$.

Axiuillary Lemmas. The following lemmas will be needed.

Lemma 6. Given $\mathcal{B}_k, k = 1, \dots, K$ defined for a single neuron pair (\mathbf{w}, \mathbf{v}) , we have

$$-2(1 - \mathcal{B}_k) - \frac{1}{K-1} \leq \mathcal{B}_{k'} \leq 2(1 - \mathcal{B}_k) - \frac{1}{K-1}, \forall k, k' \text{ with } k' \neq k.$$

Proof. With the following basic derivation

$$\mathcal{B}_{k'} = \left\langle \tilde{\mathbf{e}}_{k'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \left\langle \tilde{\mathbf{e}}_{k'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} - \tilde{\mathbf{e}}_k + \tilde{\mathbf{e}}_k \right\rangle = \left\langle \tilde{\mathbf{e}}_{k'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} - \tilde{\mathbf{e}}_k \right\rangle - \frac{1}{K-1},$$

the desired result comes from the fact that $\left| \left\langle \tilde{\mathbf{e}}_{k'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} - \tilde{\mathbf{e}}_k \right\rangle \right| \leq \left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} - \tilde{\mathbf{e}}_k \right\| = 2(1 - \mathcal{B}_k)$. \square

Lemma 7. Given a dataset that satisfies Assumption 1 then the following is true:

• $\forall k$ and some $a_i, \forall i \in \mathcal{I}_k$, we have

$$\left\| \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i \right\| \geq \sqrt{\mu_s} \sum_{i \in \mathcal{I}_k} a_i X_{\min}; \quad (41)$$

• $\forall k \neq k'$ and some $a_i, b_{i'} \geq 0, \forall i \in \mathcal{I}_k, i' \in \mathcal{I}_{k'}$, we have

$$\left\langle \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}} b_{i'} \mathbf{x}_{i'} \right\rangle \leq -\mu_d \left\| \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i \right\| \left\| \sum_{i' \in \mathcal{I}_{k'}} b_{i'} \mathbf{x}_{i'} \right\|. \quad (42)$$

Proof. For the first inequality,

$$\begin{aligned} \left\| \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i \right\| &= \sqrt{\sum_{i \in \mathcal{I}_k} a_i^2 \|\mathbf{x}_i\|^2 + \sum_{i \neq j} a_i a_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle} \\ &\geq \sqrt{\sum_{i \in \mathcal{I}_k} a_i^2 \mu_s \|\mathbf{x}_i\|^2 + \sum_{i \neq j} a_i a_j \mu_s \|\mathbf{x}_i\| \|\mathbf{x}_j\|} \\ &\geq \sqrt{\mu_s} \sqrt{\sum_{i \in \mathcal{I}_k} a_i^2 + \sum_{i \neq j} a_i a_j} X_{\min} = \sqrt{\mu_s} \sum_{i \in \mathcal{I}_k} a_i X_{\min}. \end{aligned}$$

For the second inequality,

$$\begin{aligned} \left\langle \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}} b_{i'} \mathbf{x}_{i'} \right\rangle &= \sum_{i \in \mathcal{I}_k, i' \in \mathcal{I}_{k'}} a_i b_{i'} \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle \leq -\mu_d \sum_{i \in \mathcal{I}_k, i' \in \mathcal{I}_{k'}} a_i b_{i'} \|\mathbf{x}_i\| \|\mathbf{x}_{i'}\| \\ &\leq -\mu_d \left\| \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i \right\| \left\| \sum_{i' \in \mathcal{I}_{k'}} b_{i'} \mathbf{x}_{i'} \right\|. \end{aligned}$$

\square

933 **Lemma 8.** Given $\{z_i, i \in \mathcal{I}\}$ with $\langle z_i, z_j \rangle \leq 0, \forall i, j \in \mathcal{I}, i \neq j$, then $\|\sum_{i \in \mathcal{I}} z_i\| \leq \sqrt{\sum_{i \in \mathcal{I}} \|z_i\|^2}$.

934 *Proof.* $\|\sum_{i \in \mathcal{I}} z_i\| = \sqrt{\sum_{i \in \mathcal{I}} \|z_i\|^2 + \sum_{i \neq j} \langle z_i, z_j \rangle} \leq \sqrt{\sum_{i \in \mathcal{I}} \|z_i\|^2}$. \square

935 **Lemma 9.** Given a dataset that satisfies Assumption [7](#), $\exists 0 < \zeta < 1$ such that $\forall k \in [K]$ and
 936 $\forall w \in \{z : 0 < |\mathcal{I}_k^z| < |\mathcal{I}_k|\}$, we have $\|\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i\|^2 - \mathcal{A}_k^2 \geq \mu_s X_{\min}^2 \zeta$.

937 *Proof.* Notice that

$$\begin{aligned} \|\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i\|^2 - \mathcal{A}_k^2 &= \|\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i\|^2 \left(1 - \left\langle \frac{w}{\|w\|}, \frac{\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i}{\|\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i\|} \right\rangle^2\right) \\ &\stackrel{(\text{Lemma } \textcolor{red}{7})}{\geq} \mu_s X_{\min}^2 \left(1 - \left\langle \frac{w}{\|w\|}, \frac{\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i}{\|\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i\|} \right\rangle^2\right). \end{aligned} \quad (43)$$

938 However, the nonnegative quantity $\left(1 - \left\langle \frac{w}{\|w\|}, \frac{\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i}{\|\sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i\|} \right\rangle^2\right)$ can not be zero: Suppose it
 939 is zero, then $w \propto \pm \sum_{i \in \mathcal{I}_k: \alpha_i > 0} x_i$, which corresponds to either $|\mathcal{I}_k^z| = 0$ or $|\mathcal{I}_k^z| = |\mathcal{I}_k|$, a
 940 contradiction. We let its lowest value be $\zeta > 0$. This finishes the proof. \square

941 **The proof.** Now we are ready to prove Lemma [4](#).

942 *Proof of Lemma [4](#)* We define the following:

$$\begin{aligned} \tau_1 &= \inf \left\{ t \geq 0 : \mathcal{B}_k < 1 - \frac{1}{2(K-1)} \right\}, \\ \tau_2 &= \inf \left\{ t \geq 0 : \mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'} < \mathcal{A}_k(0) - 2 \sum_{k' \neq k} \mathcal{A}_{k'}(0) \right\}, \\ \tau_3 &= \inf \left\{ t \geq 0 : |\mathcal{I}_k^w| < |\mathcal{I}_k^{w(0)}| \text{ or } |\mathcal{I}_k^w| > |\mathcal{I}_k^{w(0)}| \text{ for some } k' \right\}. \end{aligned}$$

943 Then it suffices to show that $\min\{\tau_1, \tau_2, \tau_3\} \geq \min\{T_{j,k}^*, T\}$, for which we prove them by contra-
 944 diction. **Note: In the proof we will use " \geq^* " to represent an inequality that holds when ϵ is**
 945 **sufficiently small.**

946 **Case 1:** $\min\{\tau_1, \tau_2, \tau_3\} = \tau_1$.

947 At τ_1 , by the continuity of \mathcal{B}_k , we must have $\mathcal{B}_k(\tau_1) = 1 - \frac{1}{2(K-1)}$. Suppose $\tau_1 \leq \min\{T_{j,k}^*, T\}$,
 948 then we have the following derivation

$$\begin{aligned} &\left. \frac{d}{dt} \mathcal{B}_k \right|_{t=\tau_1} \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k(1 - \mathcal{B}_k^2) + \sum_{k' \neq k} \mathcal{A}_{k'} \left(-\frac{1}{K-1} - \mathcal{B}_k \mathcal{B}_{k'} \right) \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(\text{Lemma } \textcolor{red}{6})}{\geq} \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k(1 - \mathcal{B}_k^2) + \sum_{k' \neq k} \mathcal{A}_{k'} \left(-\frac{1}{K-1} - \mathcal{B}_k \left(2(1 - \mathcal{B}_k) - \frac{1}{K-1} \right) \right) \right) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k(1 - \mathcal{B}_k^2) - 2 \sum_{k' \neq k} \mathcal{A}_{k'} \left(\frac{1 - \mathcal{B}_k}{2(K-1)} + \mathcal{B}_k(1 - \mathcal{B}_k) \right) \right) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\left(\mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'} \right) (1 - \mathcal{B}_k^2) + 2 \sum_{k' \neq k} \mathcal{A}_{k'} \left(1 - \mathcal{B}_k^2 - \frac{1 - \mathcal{B}_k}{2(K-1)} - \mathcal{B}_k(1 - \mathcal{B}_k) \right) \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(t=\tau_1)}{\geq} \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'} \right) \frac{1}{K-1} + 2 \sum_{k' \neq k} \mathcal{A}_{k'} \left(\frac{1}{2(K-1)} \left(1 - \frac{1}{2(K-1)} \right) \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(\tau_2 \geq \tau_1)}{\geq} \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k(0) - 2 \sum_{k' \neq k} \mathcal{A}_{k'}(0) \right) \frac{1}{K-1} + \mathcal{O}(\epsilon) \\ &\stackrel{(\textcolor{red}{40}), \tau_1 \leq T}{\geq} \sqrt{\frac{K-1}{K}} \left(\mathcal{A}_k(0) - 2 \sum_{k' \neq k} \mathcal{A}_{k'}(0) \right) \frac{1}{K-1} - \frac{16}{\sqrt{K}} \epsilon n X_{\max}^2 \sqrt{h} \stackrel{(*)}{\geq} 0. \end{aligned} \quad (44)$$

949 The definition of τ_1 suggests that \mathcal{B}_k must drop below $1 - \frac{1}{2(K-1)}$ right after $t = \tau_1$, which contradicts
 950 that $\frac{d}{dt}\mathcal{B}_k|_{t=\tau_1} \geq 0$. Therefore $\min\{\tau_1, \tau_2, \tau_3\} > \min\{T_{j,k}^*, T\}$ can not be true under the case when
 951 $\min\{\tau_1, \tau_2, \tau_3\} = \tau_1$.

952 **Case 2:** $\min\{\tau_1, \tau_2, \tau_3\} = \tau_2$.

953 Again, we derive a contradiction by supposing $\tau_2 \leq \min\{T^*, T\}$. Since $\min\{\tau_1, \tau_2, \tau_3\} = \tau_2$, at τ_2
 954 we still have $\mathcal{B}_k \geq 1 - \frac{1}{2(K-1)} > 0$, and by Lemma 6 we also have $\mathcal{B}_{k'} \leq 2(1 - \mathcal{B}_k) - \frac{1}{K-1} \leq 0$.

955 Starting from (39) restricted to $t = \tau_2$, we have for the target class,

$$\begin{aligned}
 & \left. \frac{d}{dt} \mathcal{A}_k \right|_{t=\tau_2} \\
 &= \sqrt{\frac{K-1}{K}} \sum_{1 \leq k' \leq K} \mathcal{B}_{k'} \left(\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle - \mathcal{A}_k \mathcal{A}_{k'} \right) + \mathcal{O}(\epsilon) \\
 &= \sqrt{\frac{K-1}{K}} \underbrace{\mathcal{B}_k}_{\geq 0} \left(\left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i \right\|^2 - \mathcal{A}_k^2 + \underbrace{\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_k: \alpha_{i'} = 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle}_{\geq 0} \right) \\
 &\quad + \sqrt{\frac{K-1}{K}} \sum_{k' \neq k} \underbrace{\mathcal{B}_{k'}}_{\leq 0} \left(\underbrace{\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle}_{\leq 0} - \underbrace{\mathcal{A}_k \mathcal{A}_{k'}}_{\geq 0} \right) + \mathcal{O}(\epsilon), \\
 &\geq \sqrt{\frac{K-1}{K}} \mathcal{B}_k \left(\left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i \right\|^2 - \mathcal{A}_k^2 \right) + \mathcal{O}(\epsilon), \tag{45}
 \end{aligned}$$

956 and for non-target classes, we have

$$\begin{aligned}
 & \left. \frac{d}{dt} \sum_{k' \neq k} \mathcal{A}_{k'} \right|_{t=\tau_2} \\
 &= \sqrt{\frac{K-1}{K}} \sum_{k' \neq k} \sum_{1 \leq k'' \leq K} \mathcal{B}_{k''} \left(\left\langle \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle - \mathcal{A}_{k'} \mathcal{A}_{k''} \right) + \mathcal{O}(\epsilon) \\
 &= \sqrt{\frac{K-1}{K}} \sum_{k' \neq k} \mathcal{B}_k \left(\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle - \mathcal{A}_{k'} \mathcal{A}_k \right) \\
 &\quad + \sqrt{\frac{K-1}{K}} \sum_{k' \neq k} \sum_{k'' \neq k} \mathcal{B}_{k''} \left(\left\langle \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle - \mathcal{A}_{k'} \mathcal{A}_{k''} \right) + \mathcal{O}(\epsilon) \\
 &= \sqrt{\frac{K-1}{K}} \sum_{k' \neq k} \mathcal{B}_k \left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle \\
 &\quad + \sqrt{\frac{K-1}{K}} \left\langle \sum_{k' \neq k} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i, \sum_{k'' \neq k} \mathcal{B}_{k''} \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle \\
 &\quad + \sqrt{\frac{K-1}{K}} \sum_{k' \neq k} \left(-\mathcal{B}_k \mathcal{A}_{k'} \mathcal{A}_k - \sum_{k'' \neq k} \mathcal{B}_{k''} \mathcal{A}_{k'} \mathcal{A}_{k''} \right) + \mathcal{O}(\epsilon) \tag{46}
 \end{aligned}$$

$$\leq -\sqrt{\frac{K-1}{K}} \left(\mu_d \mu_s X_{\min}^2 \mathcal{B}_k \sum_{k'=k} |\mathcal{I}_{k'}^w|^2 + \frac{2}{K-1} X_{\max}^2 \sum_{k'=k} |\mathcal{I}_{k'}^w|^2 \right) + \mathcal{O}(\epsilon), \tag{47}$$

957 The last step to get (47) is to upper bound the three terms in (46) separately, which we defer to the
 958 end of this proof. Combining (45)(47), and recalling the upper bound on the norm of the $\mathcal{O}(\epsilon)$ terms,
 959 we have

$$\begin{aligned}
 & \left. \frac{d}{dt} \left(\mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'} \right) \right|_{t=\tau_2} \\
 &\geq \sqrt{\frac{K-1}{K}} \mathcal{B}_k \left(\left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i \right\|^2 - \mathcal{A}_k^2 \right) \\
 &\quad 2 \left(\mu_d \mu_s X_{\min}^2 \mathcal{B}_k - \frac{2}{K-1} X_{\max}^2 \right) \sum_{k'=k} |\mathcal{I}_{k'}^w|^2 - 32\sqrt{K}\epsilon n^2 X_{\max}^3 \sqrt{h} \\
 &\geq \sqrt{\frac{K-1}{K}} \mathcal{B}_k \left(\left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i \right\|^2 - \mathcal{A}_k^2 \right) \\
 &\quad 2 \underbrace{\left(\mu_d \mu_s X_{\min}^2 \left(1 - \frac{1}{2(K-1)} \right) - \frac{2}{K-1} X_{\max}^2 \right)}_{\geq 0} \sum_{k'=k} |\mathcal{I}_{k'}^w|^2 - 32\sqrt{K}\epsilon n^2 X_{\max}^3 \sqrt{h}
 \end{aligned}$$

$$\geq \sqrt{\frac{K-1}{K}} \left(1 - \frac{1}{2(K-1)}\right) \mu_s X_{\min}^2 \zeta - \frac{32}{\sqrt{K}} \epsilon n^2 X_{\max}^3 \sqrt{h} \stackrel{(*)}{\geq} 0. \quad (48)$$

960 The definition of τ_2 suggests that $\mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'}$ must drop below $\mathcal{A}_k(0) - 2 \sum_{k' \neq k} \mathcal{A}_{k'}(0)$ right
 961 after $t = \tau_2$, which contradicts that $\frac{d}{dt} \left(\mathcal{A}_k - 2 \sum_{k' \neq k} \mathcal{A}_{k'} \right) \Big|_{t=\tau_2} \geq 0$. Therefore $\min\{\tau_1, \tau_2, \tau_3\} >$
 962 $\min\{T_{j,k}^*, T\}$ can not be true under the case when $\min\{\tau_1, \tau_2, \tau_3\} = \tau_2$.

963 **Case 3:** $\min\{\tau_1, \tau_2, \tau_3\} = \tau_3$. Finally, it remains to exclude the case when $\min\{\tau_1, \tau_2, \tau_3\} = \tau_3$ and
 964 $\tau_3 \leq \min\{T_{j,k}^*, T\}$. At $t = \tau_3$, either of the following must happen:

- 965 1. $\exists i \in \mathcal{I}_k$ such that $\alpha_i = 0$ and $\frac{d}{dt} \alpha_i < 0$;
- 966 2. $\exists i \in \mathcal{I}_{k'}$ for some $k' \neq k$ such that $\alpha_i = 0$ and $\frac{d}{dt} \alpha_i > 0$;

967 However, at $t = \tau_3$, $\forall i \in \mathcal{I}_k$, we have

$$\begin{aligned} & \frac{d}{dt} \alpha_i \Big|_{\alpha_i=0} \\ &= \left\langle \mathbf{x}_i, \frac{d}{dt} \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\rangle \\ &\stackrel{(\text{Lemma 5})}{=} \left\langle \mathbf{x}_i, \Pi_{\mathbf{w}}^\perp \left(\sqrt{\frac{K-1}{K}} \sum_{i'=1}^n \xi_{i'} \left\langle \tilde{\mathbf{E}} \mathbf{y}_{i'}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \mathbf{x}_{i'} \right) \right\rangle + \mathcal{O}(\epsilon) \\ &\stackrel{(\alpha_i=0)}{=} \sqrt{\frac{K-1}{K}} \langle \mathbf{x}_i, (\sum_{i'=1}^n \xi_{i'} \beta_{i'} \mathbf{x}_{i'}) \rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{B}_k \langle \mathbf{x}_i, \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \rangle + \sum_{k' \neq k} \underbrace{\mathcal{B}_{k'} \langle \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \rangle}_{\leq 0} \right) + \mathcal{O}(\epsilon) \\ &\geq \sqrt{\frac{K-1}{K}} \mathcal{B}_k \langle \mathbf{x}_i, \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \rangle + \mathcal{O}(\epsilon) \\ &\stackrel{(\text{Lemma 7})}{\geq} \sqrt{\frac{K-1}{K}} \left(1 - \frac{1}{2(K-1)}\right) |\mathcal{I}_k(0)| \mu_s X_{\min}^2 - \frac{16}{\sqrt{K}} \epsilon n X_{\max}^3 \sqrt{h} \stackrel{(*)}{\geq} 0, \end{aligned} \quad (49)$$

968 therefore it can not be that $\exists i \in \mathcal{I}_k$ such that $\alpha_i = 0$ and $\frac{d}{dt} \alpha_i < 0$. Next, $\forall i \in \mathcal{I}_{k'}, k' \neq k$, we have

$$\begin{aligned} & \frac{d}{dt} \alpha_i \Big|_{\alpha_i=0} \\ &= \sqrt{\frac{K-1}{K}} \langle \mathbf{x}_i, (\sum_{i'=1}^n \xi_{i'} \beta_{i'} \mathbf{x}_{i'}) \rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{B}_k \langle \mathbf{x}_i, \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \rangle + \sum_{k' \neq k} \mathcal{B}_{k'} \langle \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \rangle \right) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{B}_k \langle \mathbf{x}_i, \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \rangle + \langle \mathbf{x}_i, \sum_{k' \neq k} \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \mathcal{B}_{k'} \xi_{i'} \mathbf{x}_{i'} \rangle \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(\text{Lemma 7})}{\leq} \sqrt{\frac{K-1}{K}} \left(-\mu_d \mathcal{B}_k \|\mathbf{x}_i\| \left\| \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \right\| + \|\mathbf{x}_i\| \left\| \sum_{k' \neq k} \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \mathcal{B}_{k'} \xi_{i'} \mathbf{x}_{i'} \right\| \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(\text{Lemma 7})}{\leq} \sqrt{\frac{K-1}{K}} \left(-\mu_d \mu_s \mathcal{B}_k X_{\min}^2 |\mathcal{I}_k| + \|\mathbf{x}_i\| \left\| \sum_{k' \neq k} \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \mathcal{B}_{k'} \xi_{i'} \mathbf{x}_{i'} \right\| \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(\text{Lemma 8})}{\leq} \sqrt{\frac{K-1}{K}} \left(-\mu_d \mu_s \mathcal{B}_k X_{\min}^2 |\mathcal{I}_k| + \|\mathbf{x}_i\| \sqrt{\sum_{k' \neq k} \left\| \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \mathcal{B}_{k'} \xi_{i'} \mathbf{x}_{i'} \right\|^2} \right) + \mathcal{O}(\epsilon) \\ &\leq \sqrt{\frac{K-1}{K}} \left(-\mu_d \mu_s \mathcal{B}_k X_{\min}^2 |\mathcal{I}_k| + X_{\max}^2 \sqrt{\sum_{k' \neq k} |\mathcal{B}_{k'}|^2 |\mathcal{I}_{k'}|^2} \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(\tau_3 \leq \tau_1)}{\leq} \sqrt{\frac{K-1}{K}} \left(-\mu_d \mu_s \left(1 - \frac{1}{2(K-1)}\right) X_{\min}^2 |\mathcal{I}_k| + \frac{2}{K-1} X_{\max}^2 \sqrt{\sum_{k' \neq k} |\mathcal{I}_{k'}|^2} \right) + \mathcal{O}(\epsilon) \\ &\stackrel{(t=\tau_3)}{\leq} \sqrt{\frac{K-1}{K}} |\mathcal{I}_k(0)| \left(-\mu_d \mu_s \left(1 - \frac{1}{2(K-1)}\right) X_{\min}^2 + \frac{2}{K-1} X_{\max}^2 \right) - \frac{16}{\sqrt{K}} \epsilon n X_{\max}^3 \sqrt{h} \stackrel{(*)}{\leq} 0 \end{aligned} \quad (50)$$

969 therefore it can not be that $\exists i \in \mathcal{I}_{k'}, k' \neq k$ such that $\alpha_i = 0$ and $\frac{d}{dt}\alpha_i > 0$. By excluding both
 970 scenarios, $\min\{\tau_1, \tau_2, \tau_3\} > \min\{T_{j,k}^*, T\}$ can not be true under the case when $\min\{\tau_1, \tau_2, \tau_3\} = \tau_3$.
 971 The proof is complete once we add the derivations for [47](#).

972 **Complete the proof.** Lastly, it remains to prove [\(47\)](#), which comes from the following derivations:
 973 For the first term,

$$\begin{aligned}
 & \sum_{k'=k} \mathcal{B}_k \left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle \\
 & \leq \sum_{k'=k} \underbrace{\mathcal{B}_k}_{\geq 0} \left(\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} > 0} \mathbf{x}_{i'} \right\rangle + \underbrace{\left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} = 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle}_{\leq 0} \right) \\
 & \leq \sum_{k'=k} \mathcal{B}_k \left\langle \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} > 0} \mathbf{x}_{i'} \right\rangle \\
 & \stackrel{(\text{Lemma } \textcolor{red}{7})}{\leq} -\mu_d \sum_{k'=k} \mathcal{B}_k \left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i \right\| \left\| \sum_{i' \in \mathcal{I}_{k'}: \alpha_{i'} > 0} \mathbf{x}_{i'} \right\| \\
 & \stackrel{(\text{Lemma } \textcolor{red}{7})}{\leq} -\mu_d \mu_s X_{\min}^2 \sum_{k'=k} \mathcal{B}_k |\mathcal{I}_k^w| |\mathcal{I}_{k'}^w| \\
 & \stackrel{(\tau_2 \leq \tau_3)}{\leq} -\mu_d \mu_s X_{\min}^2 \mathcal{B}_k \sum_{k'=k} |\mathcal{I}_{k'}^w|^2.
 \end{aligned}$$

974 For the second term,

$$\begin{aligned}
 & \left\langle \sum_{k' \neq k} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i, \sum_{k'' \neq k} \mathcal{B}_{k''} \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle \\
 & = - \left\langle \sum_{k' \neq k} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i, \sum_{k'' \neq k} \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} (-\mathcal{B}_{k''}) \xi_{i'} \mathbf{x}_{i'} \right\rangle \\
 & \leq \left\| \sum_{k' \neq k} \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i \right\| \left\| \sum_{k'' \neq k} \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} (-\mathcal{B}_{k''}) \xi_{i'} \mathbf{x}_{i'} \right\| \\
 & \stackrel{(\text{Lemma } \textcolor{red}{8})}{\leq} \sqrt{\sum_{k' \neq k} \left\| \sum_{i \in \mathcal{I}_{k'}: \alpha_i > 0} \mathbf{x}_i \right\|^2} \sqrt{\sum_{k'' \neq k} (-\mathcal{B}_{k''})^2 \left\| \sum_{i' \in \mathcal{I}_{k''}: \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\|^2} \\
 & \stackrel{(\text{Lemma } \textcolor{red}{6})}{\leq} \sqrt{\sum_{k' \neq k} |\mathcal{I}_{k'}^w|^2 X_{\max}^2} \sqrt{\left(\frac{2}{K-1}\right)^2 \sum_{k'' \neq k} |\mathcal{I}_{k''}^w|^2 X_{\max}^2} \\
 & \leq \frac{2}{K-1} X_{\max}^2 \sum_{k'=k} |\mathcal{I}_{k'}^w|^2,
 \end{aligned} \tag{51}$$

975 where [\(51\)](#) uses that $-\frac{2}{K-1} \leq \mathcal{B}_{k'} \leq 0, \forall k' \neq k$ by Lemma [6](#). And for the last term,

$$\begin{aligned}
 & \sum_{k' \neq k} \left(-\mathcal{B}_k \mathcal{A}_{k'} \mathcal{A}_k - \sum_{k'' \neq k} \mathcal{B}_{k''} \mathcal{A}_{k'} \mathcal{A}_{k''} \right) \\
 & \stackrel{(t=\tau_2)}{\leq} \sum_{k' \neq k} \mathcal{A}_{k'} \left(-2\mathcal{B}_k \sum_{k'' \neq k} \mathcal{A}_{k''} - \sum_{k'' \neq k} \mathcal{B}_{k''} \mathcal{A}_{k''} \right) \\
 & = \sum_{k' \neq k} \mathcal{A}_{k'} \sum_{k'' \neq k} \mathcal{A}_{k''} (-2\mathcal{B}_k - \mathcal{B}_{k''}) \\
 & \stackrel{(\tau_2 \leq \tau_1)}{\leq} \sum_{k' \neq k} \mathcal{A}_{k'} \sum_{k'' \neq k} \mathcal{A}_{k''} \left(-2 + \frac{3}{K-1} \right) \leq 0.
 \end{aligned}$$

976 □

977 **B.3.2 Proof of Lemma [5](#)**

978 We will use the following lemma:

979 **Lemma 10.** For any $\bar{\mathbf{v}} \in \mathbb{S}^{K-1}$ such that $\langle \bar{\mathbf{v}}, \tilde{\mathbf{e}}_1 \rangle = \beta \in [0, 1]$, then $\forall \mathbf{p}$ such that $\mathbf{p} \geq \mathbf{0}$, $[\mathbf{p}]_1 = 0$,
 980 and $\langle \mathbf{p}, \mathbf{1} \rangle = 1$, we have

$$\left| \frac{1}{K-1} \sum_{k>1} [\bar{\mathbf{v}}]_k - \langle \mathbf{p}, \bar{\mathbf{v}} \rangle \right| \leq \sqrt{1 - \beta^2}, \tag{52}$$

981 *Proof.* First of all, since $\min_{k>1} [\bar{\mathbf{v}}]_k \leq \langle \mathbf{p}, \bar{\mathbf{v}} \rangle \leq \max_{k>1} [\bar{\mathbf{v}}]_k$, we know that

$$\begin{aligned}
 \left| \frac{1}{K-1} \sum_{k>1} [\bar{\mathbf{v}}]_k - \langle \mathbf{p}, \bar{\mathbf{v}} \rangle \right| & \leq \max \left\{ \left| \frac{\sum_{k>1} [\bar{\mathbf{v}}]_k}{K-1} - \min_{k>1} [\bar{\mathbf{v}}]_k \right|, \left| \frac{\sum_{k>1} [\bar{\mathbf{v}}]_k}{K-1} - \max_{k>1} [\bar{\mathbf{v}}]_k \right| \right\} \\
 & \leq \left\| \frac{\sum_{k>1} [\bar{\mathbf{v}}]_k}{K-1} \mathbf{1} - [\bar{\mathbf{v}}]_{2:K} \right\|_{\infty} \leq \left\| \frac{\sum_{k>1} [\bar{\mathbf{v}}]_k}{K-1} \mathbf{1} - [\bar{\mathbf{v}}]_{2:K} \right\|.
 \end{aligned}$$

Now given that $\langle \bar{\mathbf{v}}, \tilde{\mathbf{e}}_1 \rangle = \beta$, we can write $\bar{\mathbf{v}} = \beta \tilde{\mathbf{e}}_1 + \sqrt{1 - \beta^2} \mathbf{y}^\perp$, where $\mathbf{y}^\perp \in \mathbb{S}^{K-1}$ and $\mathbf{y} \perp \tilde{\mathbf{e}}_1$.
Therefore,

$$\begin{aligned} \left\| \frac{\sum_{k>1} [\bar{\mathbf{v}}]_k}{K-1} \mathbf{1} - [\bar{\mathbf{v}}]_{2:K} \right\| &= \left\| \beta \left(\frac{\sum_{k>1} [\tilde{\mathbf{e}}_1]_k}{K-1} \mathbf{1} - [\tilde{\mathbf{e}}_1]_{2:K} \right) + \sqrt{1 - \beta^2} \left(\frac{\sum_{k>1} [\mathbf{y}^\perp]_k}{K-1} \mathbf{1} - [\mathbf{y}^\perp]_{2:K} \right) \right\| \\ &= \sqrt{1 - \beta^2} \left\| \frac{\sum_{k>1} [\mathbf{y}^\perp]_k}{K-1} \mathbf{1} - [\mathbf{y}^\perp]_{2:K} \right\| \\ &= \sqrt{1 - \beta^2} \left\| \left(I - \frac{1}{K-1} \mathbf{1} \mathbf{1}^\top \right) [\mathbf{y}^\perp]_{2:K} \right\| \leq \sqrt{1 - \beta^2} \|\mathbf{y}^\perp\|_{2:K} \leq \sqrt{1 - \beta^2}. \end{aligned}$$

984

□

985 *Proof of Lemma 5* Without loss of generality, we prove this lemma for $k = 1$. We define

$$\begin{aligned} T_1 &= \inf\{t > T_{j,k}^* : \mathcal{B}_k < \frac{\sqrt{2}}{2}\}, \\ T_2 &= \inf\{t > T_{j,k}^* : |\mathcal{I}_k^{w_j}| \neq |\mathcal{I}_k| \text{ or } |\mathcal{I}_{k'}^{w_j}| \neq 0\}. \end{aligned}$$

986 We need to show that $\min\{T_1, T_2\} = \infty$. We derive a contradiction by assuming it is finite.

987 **Case one:** $\min\{T_1, T_2\} = T_1$ is finite

988 Assuming $\min\{T_1, T_2\} = T_1$ is finite, our primary focus is the angular dynamics of $\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}, \forall j \in \mathcal{N}_1$,

$$\frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} = \Pi_{\mathbf{v}_j}^\perp \sum_{i \in \mathcal{I}_1} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle (\mathbf{e}_1 - \hat{\mathbf{y}}_i),$$

989 and in particular those of its alignment with pseudo-label $\tilde{\mathbf{e}}_1$,

$$\begin{aligned} \frac{d}{dt} \mathcal{B}_j^{v_j} &= \left\langle \tilde{\mathbf{e}}_1, \frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \\ &= \left\langle \tilde{\mathbf{e}}_1, \Pi_{\mathbf{v}_j}^\perp \sum_{i \in \mathcal{I}_1} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle (\mathbf{e}_1 - \hat{\mathbf{y}}_i) \right\rangle \\ &= \sum_{i \in \mathcal{I}_1} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \left\langle \tilde{\mathbf{e}}_1, \Pi_{\mathbf{v}_j}^\perp (\mathbf{e}_1 - \hat{\mathbf{y}}_i) \right\rangle. \end{aligned} \quad (53)$$

990 We shall focus on the term $\left\langle \tilde{\mathbf{e}}_1, \Pi_{\mathbf{v}_j}^\perp (\mathbf{e}_1 - \hat{\mathbf{y}}_i) \right\rangle$. For each $i \in \mathcal{I}_1$, we let $z_{ik} = [\mathbf{VW}\mathbf{x}_i]_k =$

991 $\left[\sum_{j \in \mathcal{N}_1} \mathbf{v}_j \mathbf{w}_j^\top \mathbf{x}_i \right]_k$, then

$$\begin{aligned} \mathbf{e}_1 - \hat{\mathbf{y}}_i &= \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{\exp(z_{i1})}{\sum_{k=1}^K \exp(z_{ik})} \\ \frac{\exp(z_{i2})}{\sum_{k=1}^K \exp(z_{ik})} \\ \vdots \\ \frac{\exp(z_{iK})}{\sum_{k=1}^K \exp(z_{ik})} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{k>1} \exp(z_{ik})}{\sum_{k=1}^K \exp(z_{ik})} \\ -\frac{\exp(z_{i2})}{\sum_{k=1}^K \exp(z_{ik})} \\ \vdots \\ -\frac{\exp(z_{iK})}{\sum_{k=1}^K \exp(z_{ik})} \end{bmatrix} \\ &\stackrel{(z_{ik} - z_{i1} := \tilde{z}_{ik})}{=} \begin{bmatrix} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \\ \frac{\exp(\tilde{z}_{i2})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \\ \vdots \\ -\frac{\exp(\tilde{z}_{iK})}{\sum_{k>1} \exp(\tilde{z}_{ik})} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \\ -\frac{1}{K-1} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \\ \vdots \\ -\frac{1}{K-1} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{\sum_{k>1} \exp(\tilde{z}_{ik})} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{-\exp(\tilde{z}_{i2}) + \frac{1}{K-1} \sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \\ \vdots \\ \frac{-\exp(\tilde{z}_{iK}) + \frac{1}{K-1} \sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \end{bmatrix} \\ &= \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} \tilde{\mathbf{e}}_1 + \begin{bmatrix} 0 \\ \frac{-\exp(\tilde{z}_{i2})}{\sum_{k>1} \exp(\tilde{z}_{ik})} + \frac{1}{K-1} \\ \vdots \\ \frac{-\exp(\tilde{z}_{iK})}{\sum_{k>1} \exp(\tilde{z}_{ik})} + \frac{1}{K-1} \end{bmatrix} \right), \end{aligned} \quad (54)$$

992 thus we have

$$\begin{aligned}
& \left\langle \tilde{\mathbf{e}}_1, \Pi_{\mathbf{v}_j}^\perp(\mathbf{e}_1 - \hat{\mathbf{y}}_i) \right\rangle \\
&= \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left\langle \tilde{\mathbf{e}}_1, \left(I - \frac{\mathbf{v}_j \mathbf{v}_j^\top}{\|\mathbf{v}_j\|^2} \right) \left(\sqrt{\frac{K}{K-1}} \tilde{\mathbf{e}}_1 + \begin{bmatrix} 0 \\ \frac{-\exp(\tilde{z}_{i2})}{\sum_{k>1} \exp(\tilde{z}_{ik})} + \frac{1}{K-1} \\ \vdots \\ \frac{-\exp(\tilde{z}_{iK})}{\sum_{k>1} \exp(\tilde{z}_{ik})} + \frac{1}{K-1} \end{bmatrix} \right) \right\rangle \\
&= \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} (1 - (\mathcal{B}_j^{\mathbf{v}_j})^2) + \left\langle \tilde{\mathbf{e}}_1, \left(I - \frac{\mathbf{v}_j \mathbf{v}_j^\top}{\|\mathbf{v}_j\|^2} \right) \begin{bmatrix} 0 \\ \frac{-\exp(\tilde{z}_{i2})}{\sum_{k>1} \exp(\tilde{z}_{ik})} + \frac{1}{K-1} \\ \vdots \\ \frac{-\exp(\tilde{z}_{iK})}{\sum_{k>1} \exp(\tilde{z}_{ik})} + \frac{1}{K-1} \end{bmatrix} \right\rangle \right) \\
&= \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} (1 - (\mathcal{B}_j^{\mathbf{v}_j})^2) - \mathcal{B}_j^{\mathbf{v}_j} \left(\frac{1}{K-1} \sum_{k>1} \left[\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right]_k - \sum_{k>1} \frac{\exp(\tilde{z}_{ik}) \left[\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right]_k}{\sum_{k>1} \exp(\tilde{z}_{ik})} \right) \right) \\
&\stackrel{(\text{Lemma } \textcolor{red}{10})}{\geq} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} (1 - (\mathcal{B}_j^{\mathbf{v}_j})^2) - \mathcal{B}_j^{\mathbf{v}_j} \sqrt{1 - (\mathcal{B}_j^{\mathbf{v}_j})^2} \right) \\
&= \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \sqrt{1 - (\mathcal{B}_j^{\mathbf{v}_j})^2} \left(\sqrt{\frac{K}{K-1}} \sqrt{1 - (\mathcal{B}_j^{\mathbf{v}_j})^2} - \mathcal{B}_j^{\mathbf{v}_j} \right), \tag{55}
\end{aligned}$$

993 from which we see that at $t = T_1$, we have

$$\frac{d}{dt} \mathcal{B}_j^{\mathbf{v}_j} \Big|_{\mathcal{B}_j^{\mathbf{v}_j} = \frac{\sqrt{2}}{2}} \geq \underbrace{\sum_{i \in \mathcal{I}_1} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle}_{\geq 0} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z})} \sqrt{\frac{1}{2}} \left(\sqrt{\frac{K}{K-1}} \sqrt{\frac{1}{2}} - \frac{\sqrt{2}}{2} \right) > 0, \tag{56}$$

994 contradicting the definition of T_1 .

995 **Case two:** $\min\{T_1, T_2\} = T_2$ is finite

996 Assuming $\min\{T_1, T_2\} = T_2$ is finite, we shall focus on the time interval $[T_{j,k}^*, T_2]$, when we have

$$\frac{d}{dt} \mathbf{w}_j = \sum_{i \in \mathcal{I}_1} \langle \mathbf{e}_1 - \hat{\mathbf{y}}_i, \mathbf{v}_j \rangle \mathbf{x}_i = \sum_{i \in \mathcal{I}_1} \left\langle \mathbf{e}_1 - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \|\mathbf{v}_j\| \mathbf{x}_i \tag{57}$$

997 From [\(54\)](#), we have $\forall t \leq T_2$

$$\begin{aligned}
\left\langle \mathbf{e}_1 - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle &= \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z})} \left(\sqrt{\frac{K}{K-1}} \mathcal{B}_j^{\mathbf{v}_j} + \left(\frac{1}{K-1} \sum_{k>1} \left[\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right]_k - \sum_{k>1} \frac{\exp(\tilde{z}_{ik}) \left[\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right]_k}{\sum_{k>1} \exp(\tilde{z}_{ik})} \right) \right) \\
&\stackrel{(\text{Lemma } \textcolor{red}{10})}{\geq} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z})} \left(\sqrt{\frac{K}{K-1}} \mathcal{B}_j^{\mathbf{v}_j} - \sqrt{1 - (\mathcal{B}_j^{\mathbf{v}_j})^2} \right) \\
&\stackrel{(T_1 \geq T_2)}{\geq} \frac{\sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z})} \left(\sqrt{\frac{K}{K-1}} \frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2} \right) \geq 0. \tag{58}
\end{aligned}$$

998 Therefore, by the Fundamental Theorem of Calculus, we have

$$\mathbf{w}_j(T_2) = \mathbf{w}_j(T_{j,k}^*) + \underbrace{\sum_{i \in \mathcal{I}_1} \left(\int_{T_{j,k}^*}^{T_2} \left\langle \mathbf{e}_1 - \hat{\mathbf{y}}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \|\mathbf{v}_j\| \right) \mathbf{x}_i}_{\geq 0}, \tag{59}$$

999 which ensures that $|\mathcal{I}_k^{\mathbf{w}_j(T_2)}| = |\mathcal{I}_k|$ and $|\mathcal{I}_{k'}^{\mathbf{w}_j(T_2)}| = 0$, contradicting to the definition of T_2 .

1000 Therefore, the proof is finished by the fact that $\min\{T_1, T_2\}$ cannot be finite. \square

1001 **B.4 Proof of Proposition [1](#)**

1002 As we have discussed in Appendix [B.3](#) it suffices to prove Proposition [3](#).

1003 *Proof of Proposition 3* We have shown that before $\min\{T_{j,k}^*, T\}$, the properties of the weights in
 1004 Lemma 4 hold. We consider a sufficiently small ϵ such that

$$\frac{16}{\sqrt{K}} \epsilon n^2 X_{\max}^3 \sqrt{h} \leq \frac{1}{2} \sqrt{\frac{K-1}{K}} \left(1 - \frac{1}{2(K-1)}\right) \mu_s X_{\min}^2 \zeta, \quad (60)$$

1005 and

$$\frac{2X_{\max}|\mathcal{I}_k|}{\sqrt{\frac{K-1}{K} \left(1 - \frac{1}{2(K-1)}\right) \mu_s X_{\min}^2 \zeta}} \leq \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon} \quad (61)$$

1006 Then we show that $\min\{T_{j,k}^*, T\} = T_{j,k}^*$ by contradiction: Suppose that $T \leq T_{j,k}^*$, then during $[0, T]$,
 1007 we have, from (45),

$$\begin{aligned} \frac{d}{dt} \mathcal{A}_k &\geq \sqrt{\frac{K-1}{K}} \mathcal{B}_k \left(\left\| \sum_{i \in \mathcal{I}_k: \alpha_i > 0} \mathbf{x}_i \right\|^2 - \mathcal{A}_k^2 \right) + \mathcal{O}(\epsilon), \\ &\geq \sqrt{\frac{K-1}{K}} \left(1 - \frac{1}{2(K-1)}\right) \mu_s X_{\min}^2 \zeta - \frac{16}{\sqrt{K}} \epsilon n^2 X_{\max}^3 \sqrt{h} \\ &\stackrel{(60)}{\geq} \frac{1}{2} \sqrt{\frac{K-1}{K}} \left(1 - \frac{1}{2(K-1)}\right) \mu_s X_{\min}^2 \zeta. \end{aligned} \quad (62)$$

1008 Then by the Fundamental Theorem of Calculus, we have

$$\mathcal{A}_k(T) \geq \mathcal{A}(0) + \frac{T}{2} \sqrt{\frac{K-1}{K}} \left(1 - \frac{1}{2(K-1)}\right) \mu_s X_{\min}^2 \zeta \stackrel{(61)}{\geq} \mathcal{A}(0) + X_{\max}|\mathcal{I}_k|, \quad (63)$$

1009 which is a contradiction, knowing that \mathcal{A}_k cannot exceed $X_{\max}|\mathcal{I}_k|$. Therefore, we must have
 1010 $\min\{T_{j,k}^*, T\} = T_{j,k}^*$ and $T_{j,k}^* \leq T = \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon}$ is finite. Then the rest of the Proposition 3
 1011 follows Lemma 5. \square

1012 C Asymptotic Convergence Analysis under Multi-class Orthogonally 1013 Separable Data

1014 C.1 Basic results upon inter-class separation

1015 With the loss decomposition upon inter-class separation, for which we have shown to persist after

$$1016 T^* = \max_{j,k} T_{j,k}^*,$$

$$\mathcal{L}(\theta) = \sum_{k=1}^K \mathcal{L}_{\text{CE}}(\mathbf{Y}_k, \mathbf{V}_k \mathbf{W}_k^\top \mathbf{X}_k) = \sum_{k=1}^K \sum_{i=1}^{n_k} \ell_{\text{CE}}(\mathbf{y}_{k,i}, \mathbf{V}_k \mathbf{W}_k^\top \mathbf{x}_{k,i}), \quad (64)$$

1017 It suffices to study the following GF on $\sum_{i=1}^{n_k} \ell_{\text{CE}}(\mathbf{y}_{k,i}, \mathbf{V}_k \mathbf{W}_k^\top \mathbf{x}_{k,i})$:

$$\begin{aligned} \dot{\mathbf{W}}_k &= \mathbf{X}_k (\mathbf{Y}_k - \hat{\mathbf{Y}}_k)^\top \mathbf{V}_k \\ \dot{\mathbf{V}}_k &= (\mathbf{Y}_k - \hat{\mathbf{Y}}_k) \mathbf{X}_k^\top \mathbf{W}_k \\ \text{where } \hat{\mathbf{Y}}_k &= \text{SoftmaxCol}(\mathbf{V}_k \mathbf{W}_k^\top \mathbf{X}_k) \end{aligned} \quad (65)$$

1018 The following basic results can be obtained from [18, 24]

- 1019 1. $\|\mathbf{W}_k\|_F, \|\mathbf{V}_k\| \rightarrow \infty$; $\bar{\mathbf{W}}_k, \bar{\mathbf{V}}_k$ exist;
- 1020 2. $\bar{\mathbf{V}}_k^\top \bar{\mathbf{V}}_k - \bar{\mathbf{W}}_k^\top \bar{\mathbf{W}}_k = 0$;
- 1021 3. $\bar{\mathbf{W}}_k, \bar{\mathbf{V}}_k$ is a KKT point of

$$\min_{\mathbf{W}_k, \mathbf{V}_k} \|\mathbf{W}_k\|_F^2 + \|\mathbf{V}_k\|_F^2, \quad \text{s.t. } [\mathbf{V}_k \mathbf{W}_k^\top \mathbf{x}_i]_k - [\mathbf{V}_k \mathbf{W}_k^\top \mathbf{x}_i]_l \geq 1, \forall i \in \mathcal{I}_k, \forall l \neq k \quad (66)$$

1022 C.2 Proof of Proposition 2

1023 Our proof of Proposition 2 follows the same strategy as those in [15, 18], with the major difference
1024 being that we are handling cross-entropy loss, in which we provide an extension of Lemma 2.11
1025 in [18], stated as Lemma 13. Lemma 13 is central to our proof.

1026 **Lemma 11.** Let $\gamma := \min_{1 \leq k \leq K} \gamma_k$, where $\gamma_k := \min_{i \in \mathcal{I}_k} \langle \bar{\mathbf{u}}_{\infty,k}, \mathbf{x}_i \rangle$, then $\gamma \geq \mu_s X_{\min}$.

1027 *Proof.* For any $1 \leq k \leq K$, $\bar{\mathbf{u}}_{\infty,k} = \sum_{i \in \mathcal{I}_k} a_i \mathbf{x}_i$, for some $a_i \geq 0$, then immediately we have,
1028 $\forall i \in \mathcal{I}_k$

$$\langle \bar{\mathbf{u}}_{\infty,k}, \mathbf{x}_i \rangle = \langle \sum_{i' \in \mathcal{I}_k} a_{i'} \mathbf{x}_{i'}, \mathbf{x}_i \rangle \geq \mu_s \|\sum_{i' \in \mathcal{I}_k} a_{i'} \mathbf{x}_{i'}\| \|\mathbf{x}_i\| \geq \mu_s X_{\min}. \quad (67)$$

1029 □

1030 **Lemma 12.** $\gamma^\perp := \min_{k \in [K]} \min_{\|\xi\|=1, \xi \perp \bar{\mathbf{u}}_k} \max_{i \in \mathcal{I}_k} \langle \xi, \mathbf{x}_i \rangle > 0$

1031 *Proof.* This result is from Lemma 2.10 in [18]. Note that the referenced Lemma requires an additional
1032 assumption that the support vectors of $\mathbf{x}_i, i \in \mathcal{I}_k$ span the ambient space, but the authors of [18] have
1033 commented that this condition can be relaxed to the case that the span of support vectors is the span
1034 of $\mathbf{x}_i, i \in \mathcal{I}_k$, which is true here given the positive correlations between $\mathbf{x}_i, i \in \mathcal{I}_k$. □

1035 **Lemma 13.** Given some $\Theta = [\theta_1, \dots, \theta_K] \in \mathbb{R}^{D \times K}$ and some $1 \leq k \leq K$. If it
1036 holds that $\exists k' \neq k, (\theta_k - \theta_{k'})^\top \bar{\mathbf{u}}_{\infty,k} > 0$ and $\|\Pi_{\bar{\mathbf{u}}_{\infty}}^\perp(\theta_k - \theta_{k'})\|$ sufficiently large, then
1037 $\text{tr}((\mathbf{e}_k \mathbf{1}_n^T - \hat{\mathbf{Y}})^\top \Theta^\top \Pi_{\bar{\mathbf{u}}_{\infty}}^\perp \mathbf{X}) \leq 0$, where $\hat{\mathbf{Y}} = \text{SoftmaxCol}(\Theta^\top \mathbf{X})$.

1038 *Proof.* It suffices to prove the case when $k = 1$ (We discuss the others at the end of the proof). We
1039 start by the following derivations:

$$\begin{aligned} & \text{tr}((\mathbf{e}_1 \mathbf{1}_n^T - \hat{\mathbf{Y}})^\top \Theta^\top \Pi_{\bar{\mathbf{u}}_{\infty}}^\perp \mathbf{X}) \\ &= \sum_{i=1}^n \left\langle [\mathbf{e}_1 \mathbf{1}_n^T - \hat{\mathbf{Y}}]_{:,i}, [\Theta^\top \Pi_{\bar{\mathbf{u}}_{\infty}}^\perp \mathbf{X}]_{:,i} \right\rangle \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \langle \mathbf{e}_1 - \hat{\mathbf{y}}_i, \boldsymbol{\Theta}^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \rangle \\
&= \sum_{i=1}^n \left((1 - \hat{y}_{i1}) \boldsymbol{\theta}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i + \sum_{k \neq 1} (-\hat{y}_{ik}) \boldsymbol{\theta}_k^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \right) \\
&= \sum_{i=1}^n \frac{((\sum_{k=1}^K \exp(\boldsymbol{\theta}_k^\top \mathbf{x}_i)) - \exp(\boldsymbol{\theta}_1^\top \mathbf{x}_i)) \boldsymbol{\theta}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i + \sum_{k \neq 1} (-\exp(\boldsymbol{\theta}_k^\top \mathbf{x}_i)) \boldsymbol{\theta}_k^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{\sum_{k=1}^K \exp(\boldsymbol{\theta}_k^\top \mathbf{x}_i)} \\
&= \sum_{i=1}^n \frac{\sum_{k \neq 1} \exp(\boldsymbol{\theta}_k^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{\sum_{k=1}^K \exp(\boldsymbol{\theta}_k^\top \mathbf{x}_i)} \\
&= \sum_{i=1}^n \frac{\sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i)} \\
&= \sum_{k \neq 1} \sum_{i=1}^n \frac{\exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i)}. \tag{68}
\end{aligned}$$

1040 For the k -th summand, let

$$i_k^* = \arg \max_i (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i, \tag{69}$$

1041 we have

$$\begin{aligned}
&\sum_{i=1}^n \frac{\exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i)} \\
&\leq - \frac{\exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_{i_k^*}) (-\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*}}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_{i_k^*})} \\
&\quad + \sum_{i: (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \geq 0} \frac{\exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i)}.
\end{aligned}$$

1042 One can upper bound these two terms separately as follows:

$$\begin{aligned}
&- \frac{\exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_{i_k^*}) (-\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*}}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_{i_k^*})} \\
&\leq -\frac{1}{K} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_{i_k^*}) (-\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*} \\
&= -\frac{1}{K} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty} \mathbf{x}_{i_k^*}) \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*}) (-\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*} \\
&\leq -\frac{1}{K} \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*}) (-\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_{i_k^*} \\
&\leq -\frac{1}{K} \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \exp(\gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)\|) \gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)\|, \tag{70}
\end{aligned}$$

1043 and for the second term,

$$\begin{aligned}
&\sum_{i: (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \geq 0} \frac{\exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i}{1 + \sum_{k \neq 1} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i)} \\
&\leq \sum_{i: (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \geq 0} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \\
&\leq \sum_{i: (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \geq 0} \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty} \mathbf{x}_i) \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \\
&\leq \sum_{i: (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \geq 0} \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \exp(-(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i) (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \\
&\leq \sum_{i: (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{x}_i \geq 0} \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \frac{1}{e} \\
&\leq \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \frac{n}{e} \tag{71}
\end{aligned}$$

1044 Therefore, putting (68)(70)(71) together, we have

$$\begin{aligned}
&\text{tr} \left((\mathbf{e}_1 \mathbb{1}_n^T - \hat{\mathbf{Y}})^\top \boldsymbol{\Theta}^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right) \\
&\leq \sum_k \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \left(\frac{n}{e} - \frac{1}{K} \exp(\gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)\|) \gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)\| \right) \\
&= \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})^\top \bar{\mathbf{u}}_\infty) \left(\frac{n}{e} - \frac{1}{K} \exp(\gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})\|) \gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})\| \right) \\
&\quad + \sum_{k \neq k'} \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty) \frac{n}{e}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})^\top \bar{\mathbf{u}}_\infty)}{\sum_{k \neq k'} \exp(-\gamma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k)^\top \bar{\mathbf{u}}_\infty)} \left(\frac{n}{e} - \frac{1}{K} \exp(\gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})\|) \gamma^\perp \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})\| \right) + \frac{n}{e} \\
&\leq 0,
\end{aligned}$$

1045 when $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{k'})\|$ is sufficiently large.

1046 If $k \neq 1$, consider the permutation matrix $\mathbf{P}_{1 \leftrightarrow k}$ that swap the 1-st and k -th rows/columns of a matrix,
1047 then

$$\begin{aligned}
&\text{tr} \left((\mathbf{e}_k \mathbb{1}_n^T - \hat{\mathbf{Y}}_k)^\top \boldsymbol{\Theta}^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right) \\
&= \text{tr} \left((\mathbf{e}_k \mathbb{1}_n^T - \hat{\mathbf{Y}}_k)^\top \mathbf{P}_{1 \leftrightarrow k} \boldsymbol{\Theta}^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right) \\
&= \text{tr} \left((\mathbf{P}_{1 \leftrightarrow k} \mathbf{e}_k \mathbb{1}_n^T - \mathbf{P}_{1 \leftrightarrow k} \hat{\mathbf{Y}}_k)^\top \mathbf{P}_{1 \leftrightarrow k} \boldsymbol{\Theta}^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right) \\
&= \text{tr} \left((\mathbf{e}_1 \mathbb{1}_n^T - \text{SoftmaxCol}(\mathbf{P}_{1 \leftrightarrow k} \boldsymbol{\Theta}^\top \mathbf{X}))^\top \mathbf{P}_{1 \leftrightarrow k} \boldsymbol{\Theta}^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right).
\end{aligned}$$

1048 Following the derivations for the case $k = 1$ gives the desired result. \square

1049 *Proof of Proposition 2* Without loss of generality, we prove the case of $k = 1$. The existence
1050 $\{\bar{\mathbf{W}}_1, \bar{\mathbf{V}}_1\}$ is by [24]. We first show that $\bar{\mathbf{W}}_1 \propto \mathbf{u}_1 \mathbf{g}_1^\top$, which is equivalent to the statement that
1051 $\frac{\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1\|_F}{\|\bar{\mathbf{W}}_1\|_F} = 0$, and we prove by contradiction. Suppose $\frac{\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1\|_F}{\|\bar{\mathbf{W}}_1\|_F} > 0$, which necessarily implies
1052 that $\exists \rho > 0$ such that $\frac{\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top\|_F}{\|\bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top\|_F} = \rho$, then for any $\epsilon > 0, M > 0$ exists $T_{\epsilon, M} > 0$ such that
1053 $\forall t \geq T_{\epsilon, M}$, we have $\left\| \frac{\bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^\top}{\|\bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^\top\|_F} - \frac{\mathbf{W}_1 \mathbf{V}_1^\top}{\|\mathbf{W}_1 \mathbf{V}_1^\top\|_F} \right\| \leq \epsilon$ and $\|\mathbf{W}_1 \mathbf{V}_1^\top\|_F \geq M$. We will make clear the
1054 choice of ϵ and M later.

1055 Consider the time derivative of $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{W}\|_F^2$:

$$\begin{aligned}
\frac{d}{dt} \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{W}_1\|_F^2 &= 2\text{tr} \left(\mathbf{W}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \frac{d}{dt} \mathbf{W}_1 \right) \\
&= 2\text{tr} \left(\mathbf{W}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} (\mathbf{e}_1 \mathbb{1}_n^T - \hat{\mathbf{Y}})^\top \mathbf{V} \right) \\
&= 2\text{tr} \left((\mathbf{e}_1 \mathbb{1}_n^T - \hat{\mathbf{Y}})^\top \mathbf{V}_1 \mathbf{W}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right)
\end{aligned}$$

1056 We would like to use the result in Lemma 13 so we should examine:

$$\begin{aligned}
\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T\|_F^2 &= \text{tr}(\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T \bar{\mathbf{V}}_1 \bar{\mathbf{W}}_1^\top) \\
&= \text{tr}(\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top \bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top) = \rho \|\bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top\|_F > 0.
\end{aligned} \tag{72}$$

1057 Therefore, $\exists \delta > 0, k \neq 1$ such that $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T (\mathbf{e}_1 - \mathbf{e}_k)\|^2 = \delta$, otherwise, $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T (\mathbf{e}_1 -$
1058 $\mathbf{e}_k)\| = 0, \forall k \neq 1$, which can not happen. Since $\mathbf{e}_1 - \mathbf{e}_k, k \neq 1$ spans a $k - 1$ -dimensional subspace
1059 orthogonal to $\frac{\mathbb{1}}{\sqrt{K}}$, the projection of $\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T$ onto this subspace is zero suggests $\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T$ is
1060 rank-1 and all columns of $\bar{\mathbf{V}}_k$ are aligned with $\frac{\mathbb{1}}{\sqrt{K}}$, which contradicts our alignment result in Lemma
1061 4 (these columns must have at least $\frac{\sqrt{2}}{2}$ cosine alignment with $\bar{\mathbf{e}}_1$).

1062 Then $\forall t \geq T_{\epsilon, M}$, and for the k such that $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T (\mathbf{e}_1 - \mathbf{e}_k)\|^2 = \delta$, we have

$$\begin{aligned}
&\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{W}_1 \mathbf{V}_1^T (\mathbf{e}_1 - \mathbf{e}_k)\|^2 \\
&= \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \frac{\mathbf{W}_1 \mathbf{V}_1^T}{\|\mathbf{W}_1 \mathbf{V}_1^T\|_F} (\mathbf{e}_1 - \mathbf{e}_k)\|^2 \|\mathbf{W}_1 \mathbf{V}_1^T\|_F^2 \\
&\geq \left(\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \frac{\bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T}{\|\bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T\|_F} (\mathbf{e}_1 - \mathbf{e}_k)\| - \sqrt{2}\epsilon \right)^2 \|\mathbf{W}_1 \mathbf{V}_1^T\|_F^2 \\
&\geq \left(\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \frac{\bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T}{\|\bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T\|_F} (\mathbf{e}_1 - \mathbf{e}_k)\| - \sqrt{2}\epsilon \right)^2 M^2
\end{aligned} \tag{73}$$

1063 Choose sufficiently small ϵ and sufficiently large M , we ensure that $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{W}_1 \mathbf{V}_1^T (\mathbf{e}_1 - \mathbf{e}_k)\|$ is suffi-
1064 ciently large to apply Lemma 13 so that $\frac{d}{dt} \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{W}_1\|_F^2 = 2\text{tr} \left((\mathbf{e}_1 \mathbb{1}_n^T - \hat{\mathbf{Y}})^\top \mathbf{V}_1 \mathbf{W}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right) \leq 0$.

1065 On the other hand, $\|\mathbf{W}_k\|_F^2 \rightarrow \infty$, contradicting our assumption that $\frac{\|\Pi_{\mathbf{u}_1}^\perp \bar{\mathbf{W}}_1\|_F}{\|\bar{\mathbf{W}}_1\|_F} > 0$. This proves
 1066 that $\bar{\mathbf{W}}_1 \propto \mathbf{u}_1 \mathbf{g}_1^\top$.

1067 By balancedness $\bar{\mathbf{V}}_1^\top \bar{\mathbf{V}}_1 - \bar{\mathbf{W}}_1^\top \bar{\mathbf{W}}_1 = 0$, we know that $\bar{\mathbf{V}}_1 \propto \bar{\mathbf{v}}_1 \mathbf{g}_1^\top$ for some $\bar{\mathbf{v}}_1 \in \mathbb{S}^{K-1}$. It remains
 1068 to show that $\bar{\mathbf{v}}_1 = \tilde{\mathbf{e}}_1$, which is proved by again contradiction.

1069 Suppose $\bar{\mathbf{v}}_1 \neq \tilde{\mathbf{e}}_1$, then $\exists k^*$ such that $[\bar{\mathbf{v}}_1]_{k^*} \geq [\bar{\mathbf{v}}_1]_k, \forall k \neq k^*, k \neq 1$, and not all equalities can be
 1070 obtained. As a results, consider $[0 \quad -\exp(\tilde{z}_{i2}) \quad \cdots \quad -\exp(\tilde{z}_{iK})]/\sum_{k>1} \exp(\tilde{z}_{ik})$ that appeared in
 1071 (55), it converges to $\mathbf{e}_{k^*}, \forall i \in [n]$. Based on this, for any $\epsilon_1, \epsilon_2, \exists T_{\epsilon_1, \epsilon_2}$ such that $\forall t > T_{\epsilon_1, \epsilon_2}$, we have
 1072 $\max_{j \in \mathcal{N}_1} \|\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} - \bar{\mathbf{v}}_1\| \leq \epsilon_1$ and $\max_{i \in [n]} \|[0 \quad -\exp(\tilde{z}_{i2}) \quad \cdots \quad -\exp(\tilde{z}_{iK})]/\sum_{k>1} \exp(\tilde{z}_{ik}) -$
 1073 $\mathbf{e}_{k^*}\| \leq \epsilon_2$. Therefore, for some $j \in [h]$ and $t > T_{\epsilon_1, \epsilon_2}$, we have, from (55)

$$\begin{aligned}
 & \frac{d}{dt} \mathcal{B}_j^{\mathbf{v}_j} \\
 &= \left\langle \tilde{\mathbf{e}}_1, \frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \\
 &= \left\langle \tilde{\mathbf{e}}_1, \Pi_{\mathbf{v}_j}^\perp \sum_{i \in \mathcal{I}_1} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle (\mathbf{e}_1 - \hat{\mathbf{y}}_i) \right\rangle \\
 &= \sum_{i \in \mathcal{I}_1} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \left\langle \tilde{\mathbf{e}}_1, \Pi_{\mathbf{v}_j}^\perp (\mathbf{e}_1 - \hat{\mathbf{y}}_i) \right\rangle \\
 &= \sum_{i \in \mathcal{I}_1} \frac{\left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} (1 - (\mathcal{B}_j^{\mathbf{v}_j})^2) - \mathcal{B}_j^{\mathbf{v}_j} \left(\frac{1}{K-1} \sum_{k>1} \left[\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right]_k - \sum_{k>1} \frac{\exp(\tilde{z}_{ik}) \left[\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right]_k}{\sum_{k>1} \exp(\tilde{z}_{ik})} \right) \right) \\
 &\geq \sum_{i \in \mathcal{I}_1} \frac{\left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} (1 - (\mathcal{B}_j^{\mathbf{v}_j})^2) - \mathcal{B}_j^{\mathbf{v}_j} \left([\bar{\mathbf{v}}_1]_{k^*} - \frac{\sum_{k>1} [\bar{\mathbf{v}}_1]_k}{K-1} - \epsilon_2 - 2\epsilon_1 \right) \right) \\
 & \quad (\epsilon_1, \epsilon_2 \text{ sufficiently small}) \\
 &\geq \sum_{i \in \mathcal{I}_1} \frac{\left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \sum_{k>1} \exp(\tilde{z}_{ik})}{1 + \sum_{k>1} \exp(\tilde{z}_{ik})} \left(\sqrt{\frac{K}{K-1}} (1 - (\mathcal{B}_j^{\mathbf{v}_j})^2) \right). \tag{74}
 \end{aligned}$$

1074 The right-hand side of (74) is positive and $\Theta(\frac{1}{t})$, by the fact that weight $\|\mathbf{W}_1\|, \|\mathbf{V}_1\|$ grow at a rate
 1075 $\Theta(\log(t))$. Therefore (74) suggests the divergence of $\mathcal{B}_j^{\mathbf{v}_j}$, a contradiction.

1076 Finally, we have shown $\bar{\mathbf{W}}_1 \propto \mathbf{u}_1 \mathbf{g}_1^\top$ and $\bar{\mathbf{V}}_1 \propto \tilde{\mathbf{e}}_1 \mathbf{g}_1^\top$, and the same for other k . The choices of s_k
 1077 are determined by the fact that $\bar{\mathbf{W}}_k, \bar{\mathbf{V}}_k$ must be a KKT point of (66). \square