

Neural Collapse under Gradient Flow on Shallow ReLU Networks for Orthogonally Separable Data

Hancheng Min* Zhihui Zhu† René Vidal‡

*Institute of Natural Sciences & School of Mathematical Sciences, Shanghai Jiao Tong University

†Computer Science and Engineering, Ohio State University

‡Electrical and Systems Engineering & Department of Radiology, University of Pennsylvania

hanchmin@sjtu.edu.cn, zhu.3400@osu.edu, vidalr@upenn.edu

Abstract

Among many mysteries behind the success of deep networks lies the exceptional discriminative power of their learned representations as manifested by the intriguing Neural Collapse (NC) phenomenon, where simple feature structures emerge at the last layer of a trained neural network. Prior works on the theoretical understandings of NC have focused on analyzing the optimization landscape of matrix-factorization-like problems by considering the last-layer features as unconstrained free optimization variables and showing that their global minima exhibit NC. In this paper, we show that gradient flow on a two-layer ReLU network for classifying orthogonally separable data provably exhibits NC, thereby advancing prior results in two ways: First, we relax the assumption of unconstrained features, showing the effect of data structure and nonlinear activations on NC characterizations. Second, we reveal the role of the implicit bias of the training dynamics in facilitating the emergence of NC.

1 Introduction

Among many mysteries behind the success of deep learning lies the exceptional discriminative power of neural networks as manifested by the intriguing *Neural Collapse* (NC) phenomenon [1], where simple feature structures emerge in the last layer of a trained network. The NC phenomenon is typically characterized by the following three properties (see the top plot in Figure 1):

1. *Intra-class variability collapse*: The last-layer feature vectors of the data from the same class collapse into a singleton;
2. *Maximal separation of the class means*: The *class means*, i.e., the mean feature vectors for each class, are maximally separated;
3. *Self-duality*: The classifier weights align with the class means.

Prior works on the theoretical understandings of NC have focused on analyzing the optimization landscape of matrix factorization-like problems by considering the last-layer features as unconstrained free optimization variables [2–9], showing their global minima exhibit NC. Extensions of this so-called *Unconstrained Feature Model* (UFM) include adding nonlinearity and additional hidden layers [7, 8, 10, 11], studying local convergence of gradient-based optimization algorithms around global

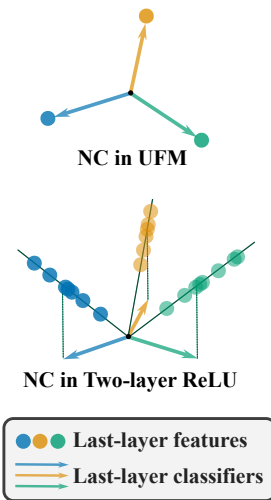


Figure 1: Visualization of the NC phenomenon.

*Correspondance to Hancheng Min

minima [2, 12, 13]. Until recently, Jacot et al. [14] show convergence towards NC in wide networks. Therefore, the question of how training a neural network leads to NC has remained underexplored.

At the center of this problem lies the fact that practical neural networks are typically overparameterized, i.e., their number of parameters is several orders of magnitude larger than the number of available training examples. As a consequence, there are infinitely many parameter choices that can perfectly fit the data, and most critically, they do not necessarily correspond to a trained network that exhibits NC. However, NC is observed even for networks trained without explicit regularization [3], such as weight decay (often associated with the emergence of NC). This suggests a close relationship between the NC phenomenon and the *implicit bias* [15] of training algorithms.

Paper contributions. In this paper, we investigate this relationship between NC and implicit bias by analyzing the dynamics of gradient flow (GF) on a two-layer ReLU network for classification problems, focusing on orthogonally separable data; i.e., any pair of input data with the same or different labels is positively or negatively correlated, respectively. We make the following contributions:

1. In Section 3 we present Theorem 1, which shows that GF with small initialization provably converges to solutions that exhibit NC and provides precise NC characterizations for the trained network, as illustrated in the bottom plot of Figure 1. Compared to prior works for the unconstrained feature model, our results highlight the role of input data and ReLU nonlinearity in determining the NC characteristics. The former causes *intra-class directional collapse* instead of collapse to a singleton, i.e., the feature vectors of the data from the same class collapse into a one-dimensional subspace. The latter leads to *orthogonal class means*, instead of maximally separated class means, and a *projected self-duality* where the classifier weights align with the projected class means (the projection is obtained by subtracting the global mean of all features).
2. In Section 4.1, through our proof of Theorem 1 for the case of binary classification, we explain how the implicit bias of GF facilitates the emergence of NC as training proceeds. Using results on the implicit bias of GF [16–21], we show that in the early phase of training the neurons’ directional alignment [16–20] with the training data makes inter-class features mutually orthogonal. Then, during the late phase of training, such inter-class separation subsequently promotes intra-class directional collapse due to the asymptotic max-margin bias [21] of GF.
3. In Section 4.2, in our proof sketch of Theorem 1 for the case of multi-class classification, we extend the aforementioned results on implicit bias for binary problems to multi-class problems. We make technical contributions in addressing new challenges that arise in the dynamic analysis due to the multi-dimensional network output and the cross-entropy loss.

In summary, our work bridges the theoretical analysis of NC and implicit bias of GF by drawing explicit connections between the two. Moreover, we further advance the theoretical understandings for both by highlighting the role of input data and nonlinear activations in NC characterization and by addressing the challenges in analyzing the implicit bias of GF in multi-class problems.

Notations. We denote the Euclidean norm of a vector \mathbf{x} by $\|\mathbf{x}\|$ and its i -th entry by $[\mathbf{x}]_i$, denote the inner product between vectors \mathbf{x} and \mathbf{y} by $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ and write $\mathbf{x} \geq \mathbf{0}$ or $\mathbf{x} > \mathbf{0}$ if all the entries of \mathbf{x} are non-negative or positive, respectively. For an $n \times m$ matrix \mathbf{A} , we let $\|\mathbf{A}\|_F$ denote the Frobenius norm of \mathbf{A} . For a scalar-valued or matrix-valued function of time, $\mathbf{F}(t)$, we let $\frac{d}{dt}\mathbf{F}(t)$ denote its time derivative. We define $\mathbf{1}$ to be the vector of all ones, whose dimension will be clear from the context. We let \mathbf{I}_n be the identity matrix of order n and sometimes drop the subscript if its order is clear from the context. We let $[N] := \{1, \dots, N\}$ and let \mathbb{S}^{D-1} be the unit-sphere in \mathbb{R}^D .

2 Preliminaries

Orthogonally separable data. We consider a classification problem on a dataset $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ of size n , where each data point $\mathbf{x}_i \in \mathbb{R}^D$ is associated with its label $\mathbf{y}_i \in \mathbb{R}^{d_y}$, and the number of unique elements in $\{\mathbf{y}_i\}_{i=1}^n$ determines the number of classes K . Throughout this paper, we assume:

Assumption 1 (Orthogonal separability). *Any pair of data with the same (different) label(s) are positively (negatively) correlated, i.e., $\exists 0 < \mu_s \leq 1$ and $0 < \mu_d \leq \frac{1}{\sqrt{K-1}}$ such that $\forall 1 \leq i, j \leq n$,*

$$\left\langle \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|} \right\rangle \geq \mu_s, \text{ if } \mathbf{y}_i = \mathbf{y}_j, \quad \left\langle \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|} \right\rangle \leq -\mu_d, \text{ if } \mathbf{y}_i \neq \mathbf{y}_j. \quad (1)$$

²No dataset can satisfy orthogonal separability with $\mu_d > \frac{1}{\sqrt{K-1}}$.

Two-layer ReLU network. We are interested in solving this classification problem by training a width- h two-layer ReLU network $f(\cdot; \theta) : \mathbb{R}^D \rightarrow \mathbb{R}^{d_y}$, parametrized by $\theta := \{W, V\} \in \mathbb{R}^{D \times h} \times \mathbb{R}^{d_y \times h}$ with $W = [w_1, \dots, w_h]$ and $V = [v_1, \dots, v_h]$, and defined as

$$f(x; \theta) = V \sigma(W^\top x) = \sum_{j=1}^h v_j \sigma(\langle w_j, x \rangle), \quad (2)$$

where $\sigma(\cdot) = \max\{\cdot, 0\}$ is the element-wise ReLU activation function. We consider networks with width $h \geq K$; we call (w_j, v_j) the j -th neuron pair in the network, w_j its input neuron weight and v_j its output neuron weight. Moreover, we let $\phi_\theta(x) = [\sigma(\langle w_1, x \rangle), \sigma(\langle w_2, x \rangle), \dots, \sigma(\langle w_h, x \rangle)]^\top \in \mathbb{R}^h$ be the last-layer feature of x , and $V = [v_1, v_2, \dots, v_h] \in \mathbb{R}^{d_y \times h}$ denote the last-layer classifier. Note that we have considered bias-free ReLU networks; see the remark in the Appendix A for extending the results to networks with biases.

Gradient flow with small initialization. Given some $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}_{\geq 0}$ such that $\ell(y_i, \hat{y}_i)$ (expressions shown in later sections) measures the discrepancy between the actual label y_i and a predicted label \hat{y}_i , we let $\mathcal{L}(\theta) = \sum_{i=1}^n \ell(y_i, f(x_i; \theta))$ be the loss function. We train the network via gradient flow (GF), which can be viewed as gradient descent with infinitesimal step size:

$$\frac{d}{dt} \theta \in -\partial_\theta \mathcal{L}(\theta), \quad (3)$$

where ∂_θ denotes the Clarke sub-differential [22] operator. We study solutions (or trajectories) $\theta(t)$, $t \geq 0$ that satisfies (3) almost everywhere. We assume that the initialization $\theta(0)$ satisfies

Assumption 2 (ϵ -small and balanced initialization). *The initialization $\theta(0) = \{w_j(0), v_j(0)\}_{j=1}^h$ satisfies the following: there exists an initialization shape $\{w_{j0}, v_{j0}\}_{j=1}^h$ with $w_{j0}, v_{j0} \neq 0, \forall j$ and an initialization scale $\epsilon > 0$ such that $\forall j, w_j(0) = \epsilon w_{j0}, v_j(0) = \epsilon v_{j0}, \|w_{j0}\| = \|v_{j0}\|$.*

Aside from the two assumptions, we will introduce additional assumptions in different training scenarios when their respective settings become clear. For now, we shall remark on these two.

Remark 1. While the data assumption is strong, there are two main reasons for considering it: First, we investigate NC in shallow networks, whose single hidden layer has limited expressive power of collapsing features, thus one shall study more structured data, as also noted by Hong and Ling [23]. Second, as it will become clearer in Section 4, the emergence of NC is closely related to the asymptotic convergence of the network weights, whose precise characterization is limited to cases with structurally simple data [18, 20, 24]. Nonetheless, as we show in Section 5, simple real data satisfies orthogonal separability approximately, leading to NC characters that match our theorem.

Remark 2. Under the assumption of balanced initialization, we have $\|w_j(0)\| = \|v_j(0)\|, \forall j$, and this balance is maintained throughout the GF trajectories, i.e., $\|w_j(t)\| = \|v_j(t)\|, \forall t, \forall j$ [25]. This assumption of balanced initialization has been common in prior works of this type [18, 20, 24] for the sake of tractable analysis. Our experiments in Section 5 do not require balanced initialization.

3 Main result: Neural Collapse under GF on Two-layer ReLU Networks

Our main result shows that under small initialization, with some additional assumptions on the initialization shape, GF provably converges to neural collapse solutions on orthogonally separable data. Our results are presented for both binary classification and multi-class classification problems:

- **Case one: Binary classification:** We consider binary ($K = 2$) data with scalar ± 1 labels, i.e. $d_y = 1$ and $y_i \in \{-1, +1\}, \forall i$. Accordingly, the two-layer ReLU network $f(x; \theta)$ has a scalar output \hat{y} . The loss function can be either the exponential loss $\ell(y, \hat{y}) = \exp(-y\hat{y})$ or the logistic loss $\ell(y, \hat{y}) = 2 \log(1 + \exp(-y\hat{y}))$. For this case, we use plain font to suggest that label y and network output $\hat{y} = f(x; \theta)$ are scalars. Moreover, we define the index sets $\mathcal{I}_+ := \{i : y_i = +1\}$ and $\mathcal{I}_- := \{i : y_i = -1\}$ for ± 1 -class data respectively.
- **Case two: Multi-class classification:** We consider multi-class ($K > 2$) data with one-hot labels, i.e., $d_y = K$, and $y_i \in \{e_1, \dots, e_K\}$, where e_k is the k -th column of the identity matrix I_K . Accordingly, the two-layer ReLU network has its output $\hat{y} = f(x; \theta) \in \mathbb{R}^K$. The loss function is the Cross-Entropy (CE) loss $\ell(y, \hat{y}) = -\sum_{k=1}^K [y]_k \log \frac{\exp([\hat{y}]_k)[y]_k}{\sum_{l=1}^K \exp([\hat{y}]_l)}$. Moreover, we define the index sets $\mathcal{I}_k := \{i : y_i = e_k\}, \forall k \in [K]$ for data from each class.

Main result. Our main theorem follows. Note that our theorem *requires additional assumptions on the data and initialization shape that vary depending on the case*, thus we feel it is better to introduce and explain them alongside technical discussions on the convergence analysis in later sections.

Theorem 1 (NC of GF on Two-layer ReLU Networks). *Given orthogonally separable data (Assumption 1), ϵ -small and balanced initialization (Assumption 2) for a sufficiently small ϵ , and some additional case-dependent assumptions on the data and initialization shape, the limit $\bar{\theta} := \lim_{t \rightarrow \infty} \frac{\theta(t)}{\|\theta(t)\|_F}$ exists for any solution $\theta(t), t \geq 0$ to (3). Moreover, for the limit $\bar{\theta} = \{\bar{\mathbf{W}}, \bar{\mathbf{V}}\}$, we have: $\exists \bar{\phi}_k \in \mathbb{S}^{h-1}, k \in \mathcal{K}$, where $\mathcal{K} := \{+, -\}$ (**case one**) or $\mathcal{K} := [K]$ (**case two**), such that*

1. (**Intra-class directional collapse**) *The last-layer features of the training data satisfy that*

$$\phi_{\bar{\theta}}(\mathbf{x}_i) = \langle s_k \mathbf{u}_k, \mathbf{x}_i \rangle \cdot \bar{\phi}_k, \forall i \in \mathcal{I}_k, k \in \mathcal{K}, s_k = \sqrt{\gamma_k^{-1} / (2 \sum_{k \in \mathcal{K}} \gamma_k^{-1})}; \quad (4)$$

where $\gamma_k = \max_{\mathbf{u} \in \mathbb{S}^{D-1}} \min_{i \in \mathcal{I}_k} \langle \mathbf{x}_i, \mathbf{u} \rangle$, $k \in \mathcal{K}$ is the maximum margin achieved exclusively for class- k data and \mathbf{u}_k is the corresponding max-margin direction;

2. (**Orthogonal class means**) *The class means directions $\bar{\phi}_k, k \in \mathcal{K}$ satisfy that*

$$\bar{\phi}_k \geq \mathbf{0}, \quad \langle \bar{\phi}_k, \bar{\phi}_{k'} \rangle = 0, \quad \forall k, k' \in \mathcal{K}, k \neq k'; \quad (5)$$

3. (**Projected self-duality**) *The last-layer classifier satisfies that*

$$\bar{\mathbf{V}} = s_+ \bar{\phi}_+^\top - s_- \bar{\phi}_-^\top \text{ (case one) or } \bar{\mathbf{V}} = \sqrt{\frac{K}{K-1}} \left(\mathbf{I} - \frac{1}{K} \mathbf{1}\mathbf{1}^\top \right) [s_1 \bar{\phi}_1, \dots, s_K \bar{\phi}_K]^\top \text{ (case two)}. \quad (6)$$

Note that GF on positively homogeneous networks with classification losses drives the network weights to diverge to infinity [26, 27]. It suffices to study the properties of the asymptotic weight direction $\bar{\theta}$ since we have $f(\cdot; \theta) = \|\theta\|_F^2 f(\cdot; \frac{\theta}{\|\theta\|_F})$ due to the positive homogeneity of f w.r.t. θ . The following remarks compare the NC characterizations in Theorem 1 with those in prior works.

NC in two-layer ReLU. Theorem 1 shows the following NC characters at the late stage of training:

- (**Intra-class directional collapse**) Unlike previous works [2–8], which study unconstrained feature models and show that features collapse to class means with equal length, our work addresses a more realistic and challenging setting involving input data. In this setting, the limited expressiveness of two-layer ReLU networks may prevent exact collapse to a singleton. Nevertheless, the result in (4) shows a *directional collapse* in the sense that all data points in the k -th class \mathcal{I}_k have their last-layer features $\phi_{\bar{\theta}}(\mathbf{x}_i)$ collapse into a one-dimensional subspace spanned by $\bar{\phi}_k$, though the features may have varying lengths. Consequently, the intra-class variability at the last layer is determined by the variability of projections $\{\langle s_k \mathbf{u}_k, \mathbf{x}_i \rangle\}_i$, a significant reduction compared to the variability of the original data $\{\mathbf{x}_i\}_i$. Moreover, if the features are normalized to unit norm (e.g., by applying RMSnorm [28]), they collapse exactly to their corresponding class means, shedding light on the role of the normalization layer in the neural collapse phenomenon.
- (**Orthogonal class means**) The result (5) suggests that the class-mean features are orthogonal to each other, forming a non-negative orthogonal frame when normalized. The orthogonal structure, rather than a simplex Equiangular Tight Frame (simplex ETF)³, arises because the features are always non-negative due to ReLU—but any orthogonal frame can be transformed into a simplex ETF by removing its global mean. This finding aligns with results from the unconstrained features model using ReLU as the activation [7].
- (**Projected self-duality**) In the case of binary classification, the classifier $\bar{\mathbf{V}}$ converges to $s_+ \bar{\phi}_+ - s_- \bar{\phi}_-$, which yields a maximum margin, as we will show in Section 4.1. For the case of multi-class classification, (6) implies that $\bar{\mathbf{V}} \bar{\mathbf{V}}^\top = \frac{K}{K-1} \left(\mathbf{I} - \frac{1}{K} \mathbf{1}\mathbf{1}^\top \right) \bar{\Phi} \bar{\Phi}^\top \left(\mathbf{I} - \frac{1}{K} \mathbf{1}\mathbf{1}^\top \right)$, where $\bar{\Phi} = [s_1 \bar{\phi}_1, \dots, s_K \bar{\phi}_K]$. Since $\bar{\Phi} \bar{\Phi}^\top$ is a diagonal matrix with positive diagonals, $\bar{\mathbf{V}}$ forms a (scaled) simplex ETF, thereby achieving maximum margin. In particular, when the diagonal scales $s_k, k \in \mathcal{K}$ are all equal, $\bar{\mathbf{V}}$ becomes an exact simplex ETF, and each classifier converges to the corresponding *projected class mean* (projection is obtained by subtracting the global mean), up to a scaling factor—achieving self-duality between features and classifiers weights[1].

³A K -simplex ETF in \mathbb{R}^h is a collection of points specified by the columns of $\tilde{\mathbf{E}} = \sqrt{\frac{K}{K-1}} \mathbf{P} \left(\mathbf{I} - \frac{1}{K} \mathbf{1}\mathbf{1}^\top \right)$, where $\mathbf{P} \in \mathbb{R}^{h \times K}$ and $\mathbf{P}^\top \mathbf{P} = \mathbf{I}$.

Convergence of GF/GD to NC. Prior works on the convergence of gradient-based methods towards NC consider the mean squared loss [2, 12–14]; Besides, additional conditions such as initialization close to a global optimum [13], weight decay regularization [2, 13, 14], or large width [14] are needed. Compared with these works, we study the convergence under the cross-entropy loss without explicit regularization or width-overparametrization, showing that NC happens under a broader class of problems. Moreover, our results highlight the role of implicit bias of the training algorithm, which we shall discuss next.

4 Detailed Discussions: Connecting Neural Collapse with Implicit Bias of GF

4.1 Proof of Neural Collapse in Binary Classification

In this section, we provide a proof of Theorem 1 for binary classification of orthogonally separable data. Recall that \mathcal{I}_+ and \mathcal{I}_- denote the index sets for data with positive and negative labels, respectively. Let $\mathcal{N}_+(t) := \{j \in [h] : \text{sign}(v_j(t)) = +1\}$ and $\mathcal{N}_-(t) := \{j \in [h] : \text{sign}(v_j(t)) = -1\}$ denote the index set of neurons whose last-layer weights $v_j(t)$ at time t has positive and negative signs, respectively. Under Assumption 2, we have that $\text{sign}(v_j(t)) = \text{sign}(v_j(0))$ [18, 20], thus $\mathcal{N}_+(t) \equiv \mathcal{N}_+(0)$, $\mathcal{N}_-(t) \equiv \mathcal{N}_-(0)$, $\forall t$, and we conveniently let $\mathcal{N}_+ := \mathcal{N}_+(0)$ and $\mathcal{N}_- := \mathcal{N}_-(0)$.

Alignment phase of the GF. During the early phase of training, often referred to as *alignment phase*, several works [16–18, 20, 24, 29, 11, 30] have shown that the norm of the neuron weights, which is initially of scale $\mathcal{O}(\epsilon)$, remains small (of scale $\mathcal{O}(\epsilon^{1/2})$) for an extended period of time of length $\Theta(\log \frac{1}{\epsilon})$. As a result, one focuses on understanding the directional dynamics of the input neuron weights during this phase, which can be approximated as follows $\forall j \in [h]$:

$$\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} = \text{sign}(v_j) \Pi_{\mathbf{w}_j^\perp} \left(\sum_{i=1}^n \xi_{ij} \mathbf{x}_i y_i + \mathcal{O}(\epsilon) \right), \text{ for some } \xi_{ij} \in \partial_z \sigma(z)|_{z=\langle \mathbf{x}_i, \mathbf{w}_j \rangle}, \quad (7)$$

where $\Pi_{\mathbf{w}_j^\perp} = (\mathbf{I} - \mathbf{w}_j \mathbf{w}_j^\top / \|\mathbf{w}_j\|^2)$ defines the projection onto the subspace orthogonal to \mathbf{w}_j . If one neglects the $\mathcal{O}(\epsilon)$ term, the dynamics $\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$, $j \in [h]$ are decoupled. The dynamic behavior of $\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ critically depends on the stationary points of (7), which we shall address next.

The following discussions assume the $\mathcal{O}(\epsilon)$ error term is zero (only for the sake of the explanation here, the error terms are appropriately handled in the analyses). First of all, the directions $\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}$ that render $\xi_{ij} = 0, \forall i \in [n]$ are trivial stationary points of (7), and they form the “dead region” for the neuron as all the activations to the data are zero (as its name suggest, neuron weights within dead region have zero gradient thus receive no update along GF). Next, the stationary points with some $\xi_{ij} \neq 0$ are often called *extremal vectors* [16, 17, 29], and the analyses in the work of Phuong and Lampert [17], Min et al. [20] have suggested that for binary orthogonally separable data the only extremal vectors are the class mean directions: $\bar{\mathbf{x}}_+$ and $\bar{\mathbf{x}}_-$. More importantly, for neurons with $j \in \mathcal{N}_+$, $\bar{\mathbf{x}}_+$ is an attractor, and $\bar{\mathbf{x}}_-$ is a repeller⁴ (the opposite for $j \in \mathcal{N}_-$). Therefore, by following (7), the neurons weights with $j \in \mathcal{N}_+$ either fall into the dead region, or converge in direction to the average direction $\bar{\mathbf{x}}_+$ of the positive class; while those with $j \in \mathcal{N}_-$ either fall into the dead region or converge to $\bar{\mathbf{x}}_-$.

Transient analysis: inter-class separation via alignment dynamics of neurons. Based on the discussions above, the convergence analysis requires a non-degenerate initialization shape $\{\mathbf{w}_{j0}\}_{j=1}^h$ such that 1) *no input neuron weight is initialized to align with the repeller for (7)*, since moving away from the repeller can take a long time that cannot be quantified; and 2) *there must exist at least one neuron weight per class that is guaranteed to converge to the average direction of that class*, avoiding the uninteresting case of all neuron weights entering the dead region.

Assumption 3 (Non-degenerate initialization). *Let $\bar{\mathbf{x}}_+ = \frac{\mathbf{x}_+}{\|\mathbf{x}_+\|}$ and $\bar{\mathbf{x}}_- = \frac{\mathbf{x}_-}{\|\mathbf{x}_-\|}$, where $\mathbf{x}_+ := \sum_{i \in \mathcal{I}_+} \mathbf{x}_i$ and $\mathbf{x}_- := \sum_{i \in \mathcal{I}_-} \mathbf{x}_i$. The initialization shape $\{\mathbf{w}_{j0}\}_{j=1}^h$ satisfies that $\mathcal{N}_+, \mathcal{N}_- \neq \emptyset$ and*

$$\max_{i \in \mathcal{I}_+, j \in \mathcal{N}_+} \langle \mathbf{x}_i, \mathbf{w}_{j0} \rangle > 0, \quad \max_{i \in \mathcal{I}_-, j \in \mathcal{N}_-} \langle \mathbf{x}_i, \mathbf{w}_{j0} \rangle > 0; \quad (8)$$

$$\max_{j \in \mathcal{N}_+} \left\langle \bar{\mathbf{x}}_-, \frac{\mathbf{w}_{j0}}{\|\mathbf{w}_{j0}\|} \right\rangle < 1, \quad \max_{j \in \mathcal{N}_-} \left\langle \bar{\mathbf{x}}_+, \frac{\mathbf{w}_{j0}}{\|\mathbf{w}_{j0}\|} \right\rangle < 1. \quad (9)$$

⁴Roughly speaking, an attractor or a repeller is a stationary point that has the flow around its neighborhood pointing towards or against it, respectively.

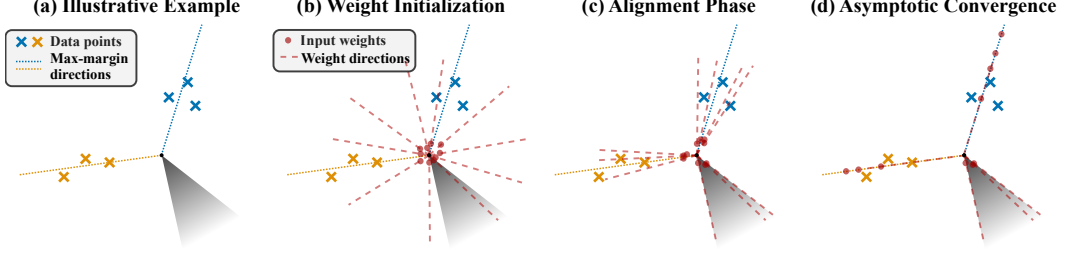


Figure 2: Convergence and implicit bias of GF in two-layer ReLU networks: (a) An example of orthogonally separable data (Assumption 1), gray region indicates the “dead region” for neuron weights: $\{w : \langle w, x_i \rangle \leq 0, \forall i \in [n]\}$; (b) Weight initialization following Assumption 2, input neuron weights have small norm and random directions; (c) During alignment phase, inter-class separation is achieved through directional alignment between input neuron weights and data points, as described in (10); (d) Asymptotically, neuron weights diverge to infinity while their directions align with the class-wise max-margin directions, as described in (12).

Given a non-degenerate initialization⁵, whenever the neuron weights converge to the vicinity of their respective attracting class averages, they stay close to the class averages for the rest of the training, leading to the following inter-class separation, due to the orthogonally separability of the data:

Claim (Inter-class separation via alignment, based the analyses from Phuong and Lampert [17], Min et al. [20]). *Given orthogonally separable data (Assumption 1), ϵ -small, balanced and non-degenerate initialization (Assumptions 2, 3) for a sufficiently small ϵ , for any solution $\theta(t)$, $t \geq 0$ to (3), $\exists T^*$ and $\emptyset \neq \tilde{\mathcal{N}}_+ \subseteq \mathcal{N}_+$ and $\emptyset \neq \tilde{\mathcal{N}}_- \subseteq \mathcal{N}_-$ such that $\forall t \geq T^*$, we have*

$$\mathbf{W}_+^\top(t) \mathbf{x}_i \begin{cases} > 0, & \forall i \in \mathcal{I}_+ \\ \leq 0, & \forall i \in \mathcal{I}_- \end{cases}, \quad \mathbf{W}_-^\top(t) \mathbf{x}_i \begin{cases} \leq 0, & \forall i \in \mathcal{I}_- \\ > 0, & \forall i \in \mathcal{I}_+ \end{cases}, \text{ and } \mathbf{W}_c^\top(t) \mathbf{x}_i \leq 0, \forall i \in [n], \quad (10)$$

where $\mathbf{W}_+(t) := [\mathbf{w}_j(t)]_{j \in \tilde{\mathcal{N}}_+}$, $\mathbf{W}_-(t) := [\mathbf{w}_j(t)]_{j \in \tilde{\mathcal{N}}_-}$, and $\mathbf{W}_c(t) := [\mathbf{w}_j(t)]_{j \in [h] - (\tilde{\mathcal{N}}_+ \cup \tilde{\mathcal{N}}_-)}$.

As a result, the **inter-class separation** $\langle \phi_{\theta(t)}(\mathbf{x}_i), \phi_{\theta(t)}(\mathbf{x}_{i'}) \rangle = 0, \forall i \in \mathcal{I}_+, i' \in \mathcal{I}_-$ holds $\forall t \geq T^*$.

We refer to Figure 2c for an illustration of the weight-data alignment that achieves inter-class separation. As shown in this claim, all neuron weights with index j outside $\tilde{\mathcal{N}}_+ \cup \tilde{\mathcal{N}}_-$ will stay within the dead region after T^* and can be disregarded in the subsequent analysis. Therefore, without loss of generality, we assume until the end of Section 4.1 that $\tilde{\mathcal{N}}_+ \cup \tilde{\mathcal{N}}_- = [h]$ and reorder the indices such that $\tilde{\mathcal{N}}_+ = [|\tilde{\mathcal{N}}_+|]$ and $\tilde{\mathcal{N}}_- = [h] - [|\tilde{\mathcal{N}}_+|]$. To see that (10) indeed suggests inter-class separation characterized by last-layer features being mutually orthogonal, notice that $\forall t \geq T^*$, we have

$$\phi_{\theta(t)}(\mathbf{x}_i) \stackrel{(\forall i \in \mathcal{I}_+)}{=} \begin{bmatrix} \sigma(\mathbf{W}_+^\top(t) \mathbf{x}_i) \\ \sigma(\mathbf{W}_-^\top(t) \mathbf{x}_i) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_+^\top(t) \mathbf{x}_i \\ \mathbf{0} \end{bmatrix}, \quad \phi_{\theta(t)}(\mathbf{x}_i) \stackrel{(\forall i \in \mathcal{I}_-)}{=} \begin{bmatrix} \sigma(\mathbf{W}_+^\top(t) \mathbf{x}_i) \\ \sigma(\mathbf{W}_-^\top(t) \mathbf{x}_i) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{W}_-^\top(t) \mathbf{x}_i \end{bmatrix}.$$

Asymptotic analysis: intra-class directional collapse and self-duality via max-margin bias. Now with the inter-class separation described in (10), we are ready to study the asymptotic convergence of the weights through the lens of max-margin bias. In particular, notice that given inter-class separation (10), we rewrite the loss as (where we define $\mathbf{V}_+ := [\mathbf{v}_j(t)]_{j \in \mathcal{N}_+}$ and $\mathbf{V}_- := [\mathbf{v}_j(t)]_{j \in \mathcal{N}_-}$):

$$\mathcal{L}(\theta) = \sum_{i=1}^n \ell(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i; \theta)) = \sum_{i \in \mathcal{I}_+} \ell(\mathbf{y}_i, \mathbf{V}_+ \mathbf{W}_+^\top \mathbf{x}_i) + \sum_{i \in \mathcal{I}_-} \ell(\mathbf{y}_i, \mathbf{V}_- \mathbf{W}_-^\top \mathbf{x}_i). \quad (11)$$

This suggests that after T^* , the GF on $\{\mathbf{W}_+, \mathbf{V}_+\}$ is fully decoupled from that on $\{\mathbf{W}_-, \mathbf{V}_-\}$, allowing one to study them separately. Moreover, each of the flows corresponds to training a *two-layer linear network* on positively correlated data with the same label, whose asymptotic convergence of the weight directions has been characterized in the work of Phuong and Lampert [17], mainly based on the analysis from Ji and Telgarsky [21] on the max-margin bias of GF in linear networks.

To be precise, for the GF on losses of the form $\sum_{i \in \mathcal{I}_+} \ell(\mathbf{y}_i, \mathbf{V} \mathbf{W}^\top \mathbf{x}_i)$, Ji and Telgarsky [21] show that as time $t \rightarrow \infty$, both \mathbf{V} and \mathbf{W}^\top diverge to infinity and their limiting directions exist. In

⁵ Min et al. [20] has shown that the non-degeneracy is satisfied with high probability when the input weight shapes are randomly initialized.

the case of (11), this means $\theta = \{\mathbf{W}_+, \mathbf{V}_+, \mathbf{W}_-, \mathbf{V}_-\}$ diverge to infinity, and $\lim_{t \rightarrow \infty} \frac{\theta(t)}{\|\theta(t)\|} := \bar{\theta} = \{\bar{\mathbf{W}}_+, \bar{\mathbf{V}}_+, \bar{\mathbf{W}}_-, \bar{\mathbf{V}}_-\}$. Moreover, they show that the limiting directions satisfies the following alignment condition $\bar{\mathbf{V}}_+ \bar{\mathbf{W}}_+^\top \propto \mathbf{u}_+$, the class-wise max-margin direction we have defined in Theorem 1 and balancedness condition $\bar{\mathbf{V}}_+^\top \bar{\mathbf{V}}_+ = \bar{\mathbf{W}}_+^\top \bar{\mathbf{W}}_+$ (similar for $\bar{\mathbf{W}}_-, \bar{\mathbf{V}}_-$, with \mathbf{u}_+ replaced by \mathbf{u}_-). It was first found in the work of Phuong and Lampert [17] that the only time these two conditions are satisfied is when $\bar{\mathbf{W}}_+^\top$ has rank 1 and the top left and right singular vectors align with $\bar{\mathbf{V}}_+$ and \mathbf{u}_+ , respectively. We show that this necessarily implies NC. The formal results are:

Claim (Directional collapse and self-duality via max-margin bias, based on the analyses from Phuong and Lampert [17], Ji and Telgarsky [21]). *Given Assumptions 1,2,&3, the limit $\bar{\theta} := \lim_{t \rightarrow \infty} \frac{\theta(t)}{\|\theta(t)\|_F}$ exists for any solution $\theta(t), t \geq 0$ to (3). For the limiting direction $\bar{\theta} = \{\bar{\mathbf{W}}_+, \bar{\mathbf{V}}_+, \bar{\mathbf{W}}_-, \bar{\mathbf{V}}_-\}$, $\exists \mathbf{g}_+ \in \mathbb{S}^{|\mathcal{N}_+|-1}, \mathbf{g}_- \in \mathbb{S}^{|\mathcal{N}_-|-1}$, such that*

$$\bar{\mathbf{W}}_+ = s_+ \mathbf{u}_+ \mathbf{g}_+^\top, \bar{\mathbf{V}}_+ = s_+ \mathbf{g}_+^\top, \bar{\mathbf{W}}_- = s_- \mathbf{u}_- \mathbf{g}_-^\top, \bar{\mathbf{V}}_- = -s_- \mathbf{g}_-^\top, \quad (12)$$

where s_+ and s_- are defined in Theorem 1. As a result, we have the **intra-class directional collapse**

$$\phi_{\bar{\theta}}(\mathbf{x}_i) = \begin{bmatrix} \langle s_+ \mathbf{u}_+, \mathbf{x}_i \rangle \mathbf{g}_+ \\ \mathbf{0} \end{bmatrix} = \langle s_+ \mathbf{u}_+, \mathbf{x}_i \rangle \begin{bmatrix} \mathbf{g}_+ \\ \mathbf{0} \end{bmatrix} := \langle s_+ \mathbf{u}_+, \mathbf{x}_i \rangle \cdot \bar{\phi}_+, \forall i \in \mathcal{I}_+, \quad (13)$$

$$\phi_{\bar{\theta}}(\mathbf{x}_i) = \begin{bmatrix} \mathbf{0} \\ \langle s_- \mathbf{u}_-, \mathbf{x}_i \rangle \mathbf{g}_- \end{bmatrix} = \langle s_- \mathbf{u}_-, \mathbf{x}_i \rangle \begin{bmatrix} \mathbf{0} \\ \mathbf{g}_- \end{bmatrix} := \langle s_- \mathbf{u}_-, \mathbf{x}_i \rangle \cdot \bar{\phi}_-, \forall i \in \mathcal{I}_-, \quad (14)$$

and the **self-duality** between the last-layer classifier weight and the last-layer feature

$$\bar{\mathbf{V}} = [\bar{\mathbf{V}}_+ \quad \mathbf{0}] + [\mathbf{0} \quad \bar{\mathbf{V}}_-] = s_+ [\mathbf{g}_+^\top \quad \mathbf{0}] - s_- [\mathbf{0} \quad \mathbf{g}_-^\top] = s_+ \bar{\phi}_+ - s_- \bar{\phi}_-. \quad (15)$$

Note that the scaling factors s_+, s_- are determined based on the results in the work of Lyu and Li [27] that the limiting $\bar{\theta}$ must satisfy another max-margin problem defined on the entire dataset.

Connecting NC with implicit bias of GF. In summary, we bridge the theoretical analysis of NC and the implicit bias of GF closer by showing how the latter facilitates the emergence of NC along GF. Notably, the inter-class separation is achieved by the directional alignment of neuron weights thanks to the small initialization scale. This resonates with a large amount of work [31–41] that identifies the small initialization as the active learning or feature learning regime, allowing simple weight and hidden feature structures to arise during the early phase of GF, which otherwise cannot be achieved if initialized in the so-called lazy regime [42–44]. Then, we have shown how the asymptotic max-margin bias promotes the intra-class directional collapse and self-duality after inter-class separation, where the key observation is that the max-margin bias often makes the weights asymptotically converge in direction to low-rank matrices, leading to low-dimensional projections that significantly reduce the variability within the input data. We note that prior work [3] studies the max-margin bias in UFM, while ours considers such bias in ReLU networks.

4.2 Proof Sketch of Neural Collapse in Multi-class Classification

As shown in the last section, the proof of the NC characterization for binary classification in Theorem 1 follows from existing results on the implicit bias of GF [17, 20, 21]. However, to understand similar NC characterizations in the case of multi-class classification, one needs to extend the implicit bias analyses to multi-class problems, which prior work rarely does. In this section, we provide a proof sketch of Theorem 1 for multi-class classification, emphasizing the additional challenges it brings to the convergence analysis by considering a multi-dimensional network output and the cross-entropy loss, and discussing our contributions in addressing these challenges.

Weight alignment in multi-class problems. Recall that in binary problems, the directional dynamics are studied only for the input weights $\mathbf{w}_j, j \in [h]$, because the output weights v_j are scalars whose sign (i.e. the “direction” of the scalar) remains the same as its initialization. However, for multi-class problems, each v_j becomes a K -dimensional vector, whose directional dynamics are non-trivial. Indeed, we show that (See Appendix C.2) during the early phase of GF, we have

$$\frac{d}{dt} \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} = \sqrt{\frac{K-1}{K}} \Pi_{\mathbf{w}_j^\perp} \left(\sum_{i=1}^n \xi_{ij} \left\langle \tilde{\mathbf{E}} \mathbf{y}_i, \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right\rangle \mathbf{x}_i + \mathcal{O}(\epsilon) \right), \quad (16)$$

$$\frac{d}{dt} \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} = \sqrt{\frac{K-1}{K}} \Pi_{\mathbf{v}_j^\perp} \left(\sum_{i=1}^n \xi_{ij} \left\langle \mathbf{x}_i, \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\rangle \tilde{\mathbf{E}} \mathbf{y}_i + \mathcal{O}(\epsilon) \right), \quad (17)$$

where $\Pi_{w_j^\perp}, \Pi_{v_j^\perp}$ and ξ_{ij} are defined similarly as in (7), and $\tilde{E} = \sqrt{\frac{K}{K-1}} (\mathbf{I} - \frac{1}{K} \mathbb{1}\mathbb{1}^\top)$. We let \tilde{e}_k be the k -th column of \tilde{E} and call it the *pseudo-label* of class k , as opposed to the one-hot label e_k .

From (16)(17) (still, we exclude the $\mathcal{O}(\epsilon)$ error terms for discussions), we see that although the dynamics $\{\frac{d}{dt} \frac{w_j}{\|w_j\|}, \frac{d}{dt} \frac{v_j}{\|v_j\|}\}$, $j \in [h]$ are decoupled among neuron pairs, the directional dynamics of each neuron pair, now concerning both input and output weights, are described by a Riemannian flow on $\mathbb{S}^{D-1} \times \mathbb{S}^{K-1}$ that is highly nonlinear (due to the ξ_{ij}) and has non-trivial interactions between the input and output weights. This is the major challenge to the convergence analysis of GF under small initialization for multi-class problems. For our purpose, we discuss the alignment of the neuron weights through the lens of stationary points.

Aside from trivial stationary points that correspond to the dead regions for the input neuron weight, $\{\bar{x}_k, \tilde{e}_k\}$, $k \in [K]$, where $\bar{x}_k = \sum_{i \in \mathcal{I}_k} x_i / \|\sum_{i \in \mathcal{I}_k} x_i\|$, are attractors of (16)(17). To give a rough explanation, assume $\frac{v_j}{\|v_j\|} = \tilde{e}_k$, i.e. perfect alignment between output weights and the pseudo-label always holds, then one can write (16) as

$$\frac{d}{dt} \frac{w_j}{\|w_j\|} = \sqrt{\frac{K-1}{K}} \Pi_{w_j^\perp} \left(\sum_{i \in \mathcal{I}_k} \xi_{ij} \underbrace{\langle \tilde{e}_k, \tilde{e}_k \rangle}_{=1} x_i \right) + \sqrt{\frac{K-1}{K}} \Pi_{w_j^\perp} \left(\sum_{k' \neq k} \sum_{i \in \mathcal{I}_{k'}} \xi_{ij} \underbrace{\langle \tilde{e}_{k'}, \tilde{e}_k \rangle}_{<0} x_i \right),$$

resulting on a flow that pushes $\frac{w_j}{\|w_j\|}$ towards (against) the directions of data points in the k -th class (other classes), and eventually towards \bar{x}_k when sufficient alignment with the k -th class has led to $\xi_{ij} = \mathbb{1}_{i \in \mathcal{I}_k}$. Similarly, by assuming $\frac{w_j}{\|w_j\|} = \bar{x}_k$, we can write (17) as $\frac{d}{dt} \frac{v_j}{\|v_j\|} = \sqrt{\frac{K-1}{K}} \Pi_{v_j^\perp} (\sum_{i \in \mathcal{I}_k} \langle x_i, \bar{x}_k \rangle \tilde{e}_k)$, pushing the output weight toward the pseudo-label \tilde{e}_k .

Transient analysis: inter-class separation via alignment dynamics of neurons. Based on the discussion above, there is the *region of attraction* (ROA) for each attractor $\{\bar{x}_k, \tilde{e}_k\}$ such that all the neuron weights initialized within the ROA are guaranteed to converge to $\{\bar{x}_k, \tilde{e}_k\}$ via (16)(17). Moreover, the boundaries between two ROAs of different (class average)-(pseudo-label) pairs together form an invariant set that does not converge to any of the attractors $\{\bar{x}_k, \tilde{e}_k\}$, where there exist saddle points of (16)(17). The exact generalization of Assumption 3 to the multi-class case should assume that no weight direction falls on the aforementioned boundaries and each ROA contains at least one neuron pair. However, finding an analytic expression for the boundaries is a challenging problem by itself and far beyond the scope of this paper. Instead, our analysis identifies an invariant subset for each ROA, thus by initializing within those invariant subsets, we guarantee the directional convergence of the neuron weights to $\{\bar{x}_k, \tilde{e}_k\}$, $k \in [K]$, which implies inter-class separation.

Formally, we let $\mathcal{I}_k^w = \{i \in \mathcal{I}_k : \langle x_i, \frac{w}{\|w\|} \rangle > 0\}$ denote the index set for class- k data points that activates the input neuron weight w , let $\mathcal{A}_k^w = \sum_{i \in \mathcal{I}_k^w} \langle x_i, \frac{w}{\|w\|} \rangle$ be the *aggregate alignment with the k -th class* of the input neuron weight w , and let $\mathcal{B}_k^v = \langle \tilde{e}_k, \frac{v}{\|v\|} \rangle$ be the *alignment with the k -th pseudo-label* of the output neuron v . Then we define the following:

Assumption 4 (Semi-local initialization). *The initialization shape $\{w_{j0}, v_{j0}\}_{j=1}^h$ satisfies that \exists a partition $\{\mathcal{N}_k, k \in [K]\}$ of $[h]$, such that $\forall k \in [K]$, we have*

$$|\mathcal{I}_k^{w_{j0}}|^2 > \sum_{k' \neq k} |\mathcal{I}_{k'}^{w_{j0}}|^2, \quad \mathcal{A}_k^{w_{j0}} > 2 \sum_{k' \neq k} \mathcal{A}_{k'}^{w_{j0}}, \quad \mathcal{B}_k^{v_{j0}} \geq 1 - \frac{1}{2(K-1)}, \quad \forall j \in \mathcal{N}_k. \quad (18)$$

Given some additional assumption on the orthogonal separability of the data (see below), the condition in (18) defines an invariant subset of the ROA of the attractor $\{\bar{x}_k, \tilde{e}_k\}$: any neuron weights initialized to satisfy (18) remains to do so during the alignment phase of GF, while getting attracted by $\{\bar{x}_k, \tilde{e}_k\}$, leading to the desired inter-class separation:

Proposition 1 (Inter-class separation in multi-class problems). *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 $X_{\max} := \max_{i \in [n]} \|x_i\|$ and $X_{\min} := \min_{i \in [n]} \|x_i\|$, ϵ -small, balanced and semi-local initialization (Assumptions 2, 4) for a sufficiently small ϵ , for any solution $\theta(t)$, $t \geq 0$ to (3), $\exists T^*$ such that $\forall t \geq T^*$, we have $\mathbf{W}_k(t)^\top x_i \begin{cases} > 0, & \forall i \in \mathcal{I}_k, \forall i \in [n], k \in [K], \\ \leq 0, & \forall i \notin \mathcal{I}_k \end{cases}$ where $\mathbf{W}_k(t) := [w_j(t)]_{j \in \mathcal{N}_k}$. As a result, the *inter-class separation* $\langle \phi_{\theta(t)}(x_i), \phi_{\theta(t)}(x_{i'}) \rangle = 0, \forall i \in \mathcal{I}_k, i' \in \mathcal{I}_{k'}, k \neq k'$ holds $\forall t \geq T^*$.*

Asymptotic analysis: intra-class directional collapse and projected self-duality via max-margin bias. Once the inter-class separation is achieved, we can again decompose the loss function as $\mathcal{L}(\theta) = \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \ell_{\text{CE}}(\mathbf{y}_i, \mathbf{V}_k \mathbf{W}_k^\top \mathbf{x}_i)$, where $\mathbf{V}_k = [\mathbf{v}_j]_{j \in \mathcal{N}_k}$, thus it suffices to study the GF on $\{\mathbf{W}_k, \mathbf{V}_k\}$ for training a two-layer linear network on exclusively on class- k data for each $k \in [K]$ with cross-entropy loss. Another major technical contribution of our analysis is to extend the max-margin results in the work of Phuong and Lampert [17], Ji and Telgarsky [21] to the multi-class problems (albeit under a special case that all data have the same label and are positively correlated), leading to the following asymptotic characterization of GF:

Proposition 2 (Variability collapse and self-duality in multi-class problems). *Let $K > 2$. Given Assumptions 1, 2, & 4, the limit $\bar{\theta} := \lim_{t \rightarrow \infty} \frac{\theta(t)}{\|\theta(t)\|_F}$ exists for any solution $\theta(t), t \geq 0$ to (3). For the limiting direction $\bar{\theta} = \{\bar{\mathbf{W}}_k, \bar{\mathbf{V}}_k\}_{k=1}^K, \exists \mathbf{g}_k \in \mathbb{S}^{|\mathcal{N}_k|-1}$, such that*

$$\bar{\mathbf{W}}_k = s_k \mathbf{u}_k \mathbf{g}_k^\top, \quad \bar{\mathbf{V}}_k = s_k \tilde{\mathbf{e}}_k \mathbf{g}_k^\top, \quad (19)$$

where $s_k, k \in [K]$ are defined in Theorem 1. As a result, the **intra-class directional collapse** (4) and **projected self-duality** (6) hold.

Limitations of current analysis. Aside from the orthogonal separability of the data, for which we have made remarks in Section 2, the convergence analysis for multi-class problems requires a stricter separability condition ($\frac{X_{\max}^2}{X_{\min}^2 \mu_d \mu_s^2} < 2K - 3$), as shown in Proposition 1. This assumption is required to show that the subsets of the parameter space defined in Assumption 4 are invariant under GF. We believe such an assumption is not needed in practice, but our limited understanding of the ROAs and their invariant subsets has led to this additional technical condition to ensure directional convergence. Future research on better characterizations of the ROAs and their invariant sets will naturally relax or even potentially remove this requirement. The additional assumption (Assumption 4) on the initialization shape for multi-class problems is another limitation of our analysis for the early phase of GF, requiring all neuron weights to have decent alignment with one of the (class-average)-(pseudo-label) pair. Relaxing such an assumption necessitates a careful in-depth analysis of the neuron weight alignment dynamics shown in (16)(17), for which we have discussed the underlying challenges, and we leave it as an important future research direction. Nonetheless, we would like point out that: First, our result on asymptotic convergence of the weights is applicable whenever one can show that inter-class separation happens at sometime during the GF, and our transient analysis simply provides one condition under which the separation is guaranteed to happen; Moreover, the semi-local initialization can be satisfied if, instead of random initialization, one initializes all the neuron pair shapes $\{\mathbf{w}_{j0}, \mathbf{v}_{j0}\}_{j=1}^h$ by drawing uniformly from the (data)-(pseudo-label) pairs $\{\mathbf{x}_i, \tilde{\mathbf{E}}\mathbf{y}_i\}_{i=1}^n$, which is a practically possible initialization scheme.

5 Numerical Experiments

We conduct experiments primarily for the purpose of validating our theoretical results. We first train a two-layer ReLU network for classifying three MNIST [45] digits and visualize the neuron weights alignment at the end of the training, thereby showing the NC characterizations in Theorem 1. Next, based on our remarks on intra-class directional collapse, our Theorem 1 suggests that proper normalization layers such as RMSNorm [28] can potentially lead to a more significant level of NC, and we conduct some preliminary experiments with ResNet [46] to verify this conjecture.

Validating Theorem 1 in MNIST digits classification. We train a two-layer ReLU network to classify three MNIST digits $\{0, 1, 2\}$. The experimental details are in Appendix B.1. Figure 3 visualizes the training results: First, we show that the dataset of MNIST digits, centered by the mean digit of the entire dataset, approximately satisfies the orthogonally separable assumption. Then, we visualized the neuron pairs, showing their respective alignment with the data and the pseudo-labels. Moreover, we visualize the top 3 principal components of the last-layer feature of the digits, together with the classifiers, whose structure matches the NC characterizations in Theorem 1.

Experiments on the role of normalization layers on NC. Next, we train a modified ResNet18 (by replacing the final linear classifier by a two-layer ReLU classifier) on MNIST and CIFAR10 [47] datasets. In addition, we add a normalization layer (Identity/None, LayerNorm [48], or RMSNorm [28]) before the ReLU classifier and vary the methods for normalization. The experimental details are in Appendix B.2. Figure 4 reports (repeated for 5 runs; mean(line) and std(shade) are

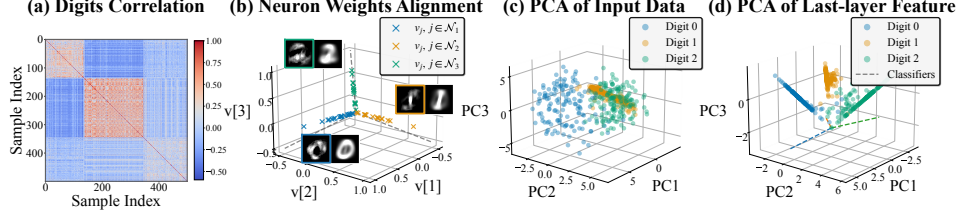


Figure 3: Validating Theorem 1 in classifying MNIST digits $\{0, 1, 2\}$. (a) Normalized correlation matrix of subsampled 500 MNIST digits; (b) (For the trained network) visualization of output neuron weights (as crosses, gray dashed line represents \tilde{e}_k directions for references), the average input neuron weights (as grayscale image, surrounded by colored box), and the average of the digits (for comparison, next to the neuron weights); (c) PCA of raw digits data \mathbf{X} , keeping the top 3 principal components (1000 points visualized) results in a $\sim 61\%$ relative approximation error for \mathbf{X} ; (d) (For the trained network) PCA of last-layer feature $\phi_\theta(\mathbf{X})$ and classifiers (rows of \mathbf{V}), keeping the top 3 components (1000 points visualized) results in a $\sim 0.2\%$ relative approximation error for $\phi_\theta(\mathbf{X})$.

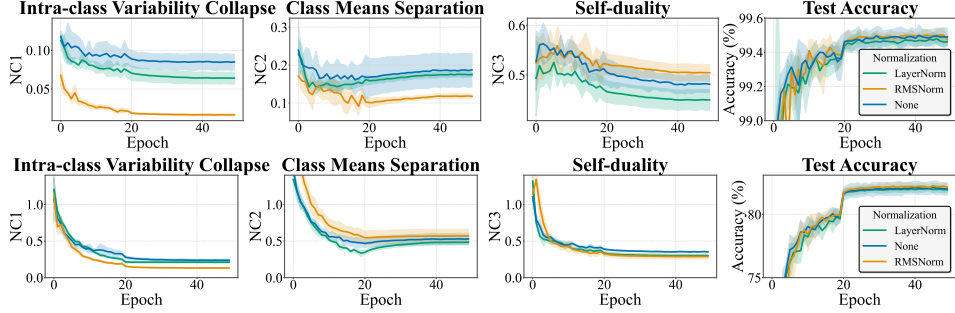


Figure 4: Measuring NC in trained modified ResNet18 on MNIST (top) and CIFAR10 (bottom)

reported) the evolutions over training 50 epochs of the metrics NC1, NC2 and NC3 that measure the NC characteristics in Theorem 1 (lower value implies more prominent NC; definitions in Appendix B.2) at the last-layer of the ReLU classifier. Notably, using the RMSNorm layer significantly improves the intra-class directional collapse, as we conjectured in the remark for Theorem 1, suggesting potential practical value in using RMSNorm layers for promoting NC.

6 Conclusion

In this paper, we investigated the connection between NC and the implicit bias of GF through a convergence analysis of GF on two-layer ReLU networks for orthogonally separable data and showed that the implicit bias of GF facilitates the emergence of NC along the GF trajectory. Future work includes relaxing the assumptions on the data and initialization, and extending the convergence analysis to understand the emergence of NC in deeper networks; For example, similar early weight directional alignment and asymptotic max-margin bias have been studied in prior works [30, 26, 27], following the same high-level proof and utilizing these existing results on alignment and max-margin for deep networks might extend the current work to a more practical setting.

Acknowledgments and Disclosure of Funding

This work is primarily done when H. Min was a postdoc at the University of Pennsylvania. The authors acknowledge the support of the NSF under grants 2031985 and IIS-2312840, the Simons Foundation under grant 814201, the ONR MURI Program under grant 503405-78051, and the University of Pennsylvania Startup Funds.

References

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

- [2] Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):11, 2022.
- [3] Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. In *ICLR*, 2022.
- [4] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *NeurIPS*, 2021.
- [5] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *PNAS*, 118(43):e2103091118, 2021.
- [6] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *ICML*, 2022.
- [7] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *ICML*, 2022.
- [8] Peter S  kenf  k, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. *NeurIPS*, 2023.
- [9] Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin G Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. In *ICML*, 2024.
- [10] Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogeneous classifiers and the role of weight decay. In *ICASSP*, 2022.
- [11] Peng Wang, Xiao Li, Can Yaras, Zhihui Zhu, Laura Balzano, Wei Hu, and Qing Qu. Understanding deep representation learning via layerwise feature compression and discrimination. *arXiv preprint arXiv:2311.02960*, 2023.
- [12] Mengjia Xu, Akshay Rangamani, Qianli Liao, Tomer Galanti, and Tomaso Poggio. Dynamics in deep classifiers trained with the square loss: Normalization, low rank, neural collapse, and generalization bounds. *Research*, 6:0024, 2023.
- [13] Peng Wang, Huikang Liu, Can Yaras, Laura Balzano, and Qing Qu. Linear convergence analysis of neural collapse with unconstrained features. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.
- [14] Arthur Jacot, Peter S  kenf  k, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. In *ICLR*, 2025.
- [15] Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023.
- [16] Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.
- [17] Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In *ICLR*, 2021.
- [18] Etienne Boursier, Loucas Pullaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In *NeurIPS*, 2022.
- [19] Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *ICML*, 2024.
- [20] Hancheng Min, Enrique Mallada, and Ren   Vidal. Early neuron alignment in two-layer relu networks with small initialization. In *ICLR*, 2024.
- [21] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *ICLR*, 2019.

- [22] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.
- [23] Wanli Hong and Shuyang Ling. Beyond unconstrained features: Neural collapse for shallow neural networks with general data. *arXiv preprint arXiv:2409.01832*, 2024.
- [24] Dmitry Chistikov, Matthias Englert, and Ranko Lazic. Learning a neuron by a shallow reLU network: Dynamics and implicit bias for correlated inputs. In *NeurIPS*, 2023.
- [25] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *NeurIPS*, 2018.
- [26] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *NeurIPS*, 2020.
- [27] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*, 2019.
- [28] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.
- [29] Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *JMLR*, 2025.
- [30] Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *TMLR*, 2024.
- [31] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In *ICLR*, 2014.
- [32] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *NeurIPS*, 2017.
- [33] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *NeurIPS*, 2019.
- [34] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *COLT*, 2020.
- [35] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *JMLR*, 22(71):1–47, 2021.
- [36] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *NeurIPS*, 2021.
- [37] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.
- [38] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*, pages 18422–18462. PMLR, 2022.
- [39] Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization. In *AISTATS*, 2025.
- [40] Zhenyu Zhu, Fanghui Liu, and Volkan Cevher. How gradient descent balances features: A dynamical analysis for two-layer neural networks. In *ICLR*, 2025.
- [41] Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B Simon, Michael R DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. *arXiv preprint arXiv:2506.06489*, 2025.

- [42] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.
- [43] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *NeurIPS*, 2019.
- [44] Vignesh Kothapalli and Tom Tirer. Can kernel methods explain how the data affects neural collapse? *TMLR*, 2025.
- [45] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016.
- [47] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical Report.
- [48] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [49] Samuel Burer and Renato D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Math. Program.*, 103(3):427–444, July 2005.
- [50] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *ICLR*, 2018.
- [51] Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *ICML*, 2021.
- [52] Hancheng Min, René Vidal, and Enrique Mallada. On the convergence of gradient flow on multi-layer linear models. In *ICML*, 2023.
- [53] Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In *AISTATS*, 2023.
- [54] Arthur Castello B de Oliveira, Milad Siami, and Eduardo D Sontag. Dynamics and perturbations of overparameterized linear neural networks. In *IEEE CDC*, 2023.
- [55] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *NeurIPS*, 2019.
- [56] Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free over-parametrized asymmetric matrix factorization. *SIAM Journal on Mathematics of Data Science*, 5(3):723–744, 2023.
- [57] Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *ICML*, 2023.
- [58] Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *NeurIPS*, 2024.
- [59] Hancheng Min and René Vidal. Understanding incremental learning with closed-form solution to gradient flow on overparametrized matrix factorization. In *IEEE CDC*, 2025.
- [60] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.
- [61] Yuanhe Zhang, Fanghui Liu, and Yudong Chen. Lora-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *ICML*, 2025.

- [62] Uijeong Jang, Jason D Lee, and Ernest K Ryu. Lora training in the ntk regime has no spurious local minima. In *ICML*, 2024.
- [63] Junsu Kim, Jaeyeon Kim, and Ernest K. Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (but it probably won’t fail). In *ICML*, 2025.
- [64] Salma Tarmoun, Guilherme França, Benjamin D Haeffele, and René Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *ICML*, 2021.
- [65] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR*, 2019.
- [66] Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.
- [67] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS*.
- [68] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *NeurIPS*, 2019.
- [69] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *NeurIPS*, 2019.
- [70] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *COLT*, 2019.
- [71] Lenaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *NeurIPS*, 2018.
- [72] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *COLT*, 2022.
- [73] Matus Telgarsky. Feature selection and low test error in shallow low-rotation relu networks. In *ICLR*, 2023.
- [74] Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *COLT*, 2022.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our abstract and introduction clearly state our claims and underlying assumptions.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations in Remark 1 and 2 and paragraph **Limitations of current analysis**.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Our assumptions are stated in the main paper, and we provide proofs of theorems in our technical appendices

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide experimental details in Appendices B.1 and B.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We attach the code in the supplemental material

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental details in the Appendices B.1 and B.2

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: See Figure 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The experiments are for validating our theoretical findings and do not require heavy computer resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: the research follows the Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: In this theory paper, there is no potential societal consequence that we feel must be specifically highlighted.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: the paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: original papers that produced the datasets are cited

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: the paper does not release new assets

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No LLMs used for writing this paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Additional Remarks and Related Works

We make the following remark on ReLU networks with biases

Extension to ReLU nets with biases When considering two-layer ReLU networks with biases, since $\sigma(\langle w, x \rangle + b) = \sigma(\langle [w; b], [x; 1] \rangle)$, adding a bias term effectively adds one homogenous coordinate to the entire dataset. Therefore, our results still hold if the augmented dataset satisfies the orthogonal separability condition. Notably, homogenous coordinate increases data correlation: $\langle [x_i; 1], [x_j; 1] \rangle = \langle x_i, x_j \rangle + 1$, thus it mainly affects the negative correlation between data points with different labels. In the case when $\min_i \|x_i\| \gg 1$, the orthogonal separability (among augmented data $\{[x_i; 1]\}_{i=1}^n$) still holds with the bias term.

A.1 Additional related works

Matrix factorization in deep learning theory A major part of the research efforts in the theoretical understandings of deep learning is through tractable mathematical problems, where similar phenomena can manifest to those observed in deep learning practice. Among these problems, matrix factorization [49] is an important one, which has been studied for understanding the convergence rate of gradient descent on neural networks [50–54], the implicit bias of neural network training algorithms [32, 55, 21, 31, 33, 37, 36, 56–59], and, as we already mentioned in the introduction, the NC phenomena [2–9]. More recently, the prevalence of LoRA [60] in practice has motivated many works on its theoretical properties [39, 61–63] through analyzing matrix factorization problems. While matrix factorization problems offer many valuable insights into various aspects of deep learning, they generally neglect the role of training data in these problems. For example, the convergence analysis often assumes input data with isotropic covariance [50, 64], and as we have discussed, the analysis of NC often assumes the last-layer features as an unconstrained optimization variable [2–4]. Given that NC is characterizing the ability of neural networks to map input data to structured latent features, it is also important to consider the role of input data. Indeed, our work highlights how the input data structure induces a directional collapse rather than a singleton collapse.

Gradient descent on shallow neural networks Theoretical properties of the gradient descent algorithms on shallow networks have been studied in different learning regimes and from various perspectives. Earlier works [65, 66] concern the convergence of gradient descent in the so-called kernel regime, where under specific settings (large network width, random weight initialization with large variance, etc.) the linearization around network weight initialization holds valid throughout training [42, 67]. However, limitations of such “lazy regime” training [43] are identified through some specific student-teacher learning settings [68, 69]. This motivates the study of convergence in small initialization settings, often called *active* or *feature learning regime* [34]. From a dynamics perspective, in the infinite width limit with proper weight initialization scaling, the weight evolution during training can be characterized by some mean-field dynamics [70, 71]; In the finite width with vanishing weight initialization, the early training phase can be characterized by the directional alignment between the weights and the input data [16, 29, 19]. From a generalization perspective, learning in the feature learning regime can enjoy many advantages, such as sample efficiency [72, 73] or benign overfitting [74]. Our work studies the convergence of ReLU networks in the feature learning regime and contributes to this line of work in the following regards: First, prior works primarily concern the convergence of the weights, while our result discusses its implications on the learned last-layer feature. Second, from a technical perspective, our result addresses several challenges emerging from considering a multi-class problem with the cross-entropy loss.

B Experimental Details

B.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 3(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 $\mathcal{N}_k = \{j \in [h] : k = \arg \max_{k'} \langle \bar{\mathbf{e}}_{k'}, \mathbf{v}_j \rangle\}$, 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 3(b).

B.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 4.

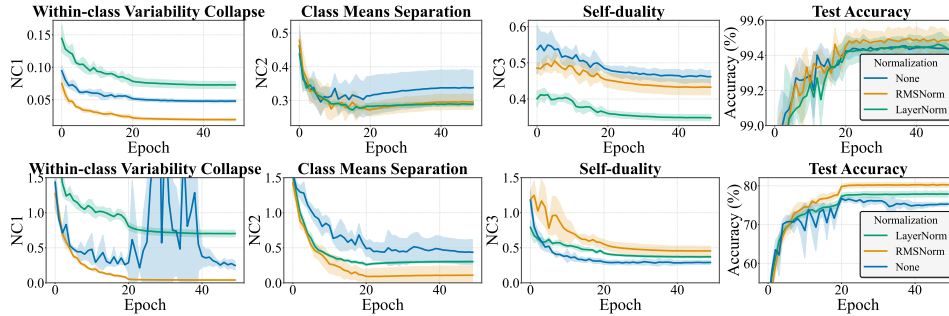


Figure 5: Measuring NC in trained modified ResNet50 on MNIST (top) and CIFAR10 (bottom)

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 5.

C Neural Alignment under Multi-class Orthogonally Separable Data

C.1 Basics on neuron dynamics under multi-class problems

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

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

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

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

It will become clear soon that it is convenient to decompose the weight dynamics into those of the 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)

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

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} .

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

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

Proof. First of all, we have

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

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

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

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

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

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

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

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

C.2 Analyzing neuron dynamics during alignment phase

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

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

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

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 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:

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

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

Proof of Lemma 2. From Section C.1, we have

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

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

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

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

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 be shown by contradiction:

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

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

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

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

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

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

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

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

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

where the last inequality is due to $\epsilon \leq \frac{\sqrt{K}}{16X_{\max} \sqrt{h}}$. This leads to a contradiction. Therefore, one must have $T \geq \tau_{j^*} \geq \frac{1}{4nX_{\max}} \log \left(\frac{1}{\sqrt{h}\epsilon} \right)$. This finishes the proof. \square

Proof of Lemma 3. We have shown in Section C.1 that

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

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

C.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

$$\mathcal{A}_k := \sum_{i \in \mathcal{I}_k^{\mathbf{w}}} \alpha_i = \underbrace{\sum_{i \in \mathcal{I}_k : \alpha_i > 0} \alpha_i}_{\text{we mostly use this notation for clarity}}, \quad (\text{alignment between input neuron and } k\text{-th class})$$

$$\mathcal{B}_k := \left\langle \tilde{\mathbf{e}}_k, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle. \quad (\text{alignment between output neuron and } k\text{-th class})$$

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:*

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

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 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 $|\mathcal{I}_k|$, and $|\mathcal{I}_{k'}^{\mathbf{w}_j}|$ decreases to 0. When the initialization scale ϵ is sufficiently small so that T is large, this Lemma will show that $T_{j,k}^*$ is finite, and we will provide an upper bound.

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

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

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

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

C.3.1 Proof of Lemma 4

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

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

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

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

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

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 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 (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

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

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

Auxiliary 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

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 $\|\sum_{i \in \mathcal{I}} z_i\| \leq \sqrt{\sum_{i \in \mathcal{I}} \|z_i\|^2}$.

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

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

Proof. Notice that

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

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 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 contradiction. We let its lowest value be $\zeta > 0$. This finishes the proof. \square

The proof. Now we are ready to prove Lemma 4.

Proof of Lemma 4. We define the following:

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

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 contradiction. **Note: In the proof we will use " $\stackrel{(*)}{\geq}$ " to represent an inequality that holds when ϵ is sufficiently small.**

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

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 $\tau_1 \leq \min\{T_{j,k}^*, T\}$, then we have the following derivation

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

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 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 $\min\{\tau_1, \tau_2, \tau_3\} = \tau_1$.

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

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 we still have $\mathcal{B}_k \geq 1 - \frac{1}{2(K-1)} > 0$, and by Lemma 6, we also have $\mathcal{B}_{k'} \leq 2(1 - \mathcal{B}_k) - \frac{1}{K-1} \leq 0$.

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

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

and for non-target classes, we have

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

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

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

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

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

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

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 $\tau_3 \leq \min\{T_{j,k}^*, T\}$. At $t = \tau_3$, either of the following must happen:

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

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

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

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} & \left. \frac{d}{dt} \alpha_i \right|_{\alpha_i=0} \\ &= \sqrt{\frac{K-1}{K}} \left\langle \mathbf{x}_i, \left(\sum_{i'=1}^n \xi_{i'} \beta_{i'} \mathbf{x}_{i'} \right) \right\rangle + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{B}_k \left\langle \mathbf{x}_i, \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \right\rangle + \sum_{k' \neq k} \mathcal{B}_{k'} \left\langle \mathbf{x}_i, \sum_{i' \in \mathcal{I}_{k'}, \alpha_{i'} \geq 0} \xi_{i'} \mathbf{x}_{i'} \right\rangle \right) + \mathcal{O}(\epsilon) \\ &= \sqrt{\frac{K-1}{K}} \left(\mathcal{B}_k \left\langle \mathbf{x}_i, \sum_{i \in \mathcal{I}_k, \alpha_i \geq 0} \xi_i \mathbf{x}_i \right\rangle + \left\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'} \right\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)$$

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 scenarios, $\min\{\tau_1, \tau_2, \tau_3\} > \min\{T_{j,k}^*, T\}$ can not be true under the case when $\min\{\tau_1, \tau_2, \tau_3\} = \tau_3$. The proof is complete once we add the derivations for 47.

Complete the proof. Lastly, it remains to prove (47), which comes from the following derivations: For the first term,

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

For the second term,

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

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,

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

□

C.3.2 Proof of Lemma 5

We will use the following lemma:

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 $\mathbf{p} \geq \mathbf{0}$, $[\mathbf{p}]_1 = 0$, and $\langle \mathbf{p}, \mathbf{1} \rangle = 1$, we have

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

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,

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

□

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

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

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

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

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

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

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

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

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

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

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 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}$$

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

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

contradicting the definition of T_1 .

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

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

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

From (54), we have $\forall t \leq T_2$

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

Therefore, by the Fundamental Theorem of Calculus, we have

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

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 .

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

C.4 Proof of Proposition 1

As we have discussed in Appendix C.3, it suffices to prove Proposition 3.

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

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

and

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

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

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

Then by the Fundamental Theorem of Calculus, we have

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

which is a contradiction, knowing that \mathcal{A}_k cannot exceed $X_{\max}|\mathcal{I}_k|$. Therefore, we must have $\min\{T_{j,k}^*, T\} = T_{j,k}^*$ 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 follows Lemma 5. \square

D Asymptotic Convergence Analysis under Multi-class Orthogonally Separable Data

D.1 Basic results upon inter-class separation

With the loss decomposition upon inter-class separation, for which we have shown to persist after $T^* = \max_{j,k} T_{j,k}^*$,

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

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

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

The following basic results can be obtained from [21, 27]

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

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

D.2 Proof of Proposition 2

Our proof of Proposition 2 follows the same strategy as those in [17, 21], with the major difference being that we are handling cross-entropy loss, in which we provide an extension of Lemma 2.11 in [21], stated as Lemma 13. Lemma 13 is central to our proof.

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 $\gamma \geq \mu_s X_{\min}$.*

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, $\forall i \in \mathcal{I}_k$

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

□

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

Proof. This result is from Lemma 2.10 in [21]. Note that the referenced Lemma requires an additional assumption that the support vectors of $\mathbf{x}_i, i \in \mathcal{I}_k$ span the ambient space, but the authors of [21] have commented that this condition can be relaxed to the case that the span of support vectors is the span of $\mathbf{x}_i, i \in \mathcal{I}_k$, which is true here given the positive correlations between $\mathbf{x}_i, i \in \mathcal{I}_k$. □

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 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 $\text{tr}((\mathbf{e}_k \mathbb{1}_n^T - \hat{\mathbf{Y}})^\top \Theta^\top \Pi_{\bar{\mathbf{u}}_{\infty}}^\perp \mathbf{X}) \leq 0$, where $\hat{\mathbf{Y}} = \text{SoftmaxCol}(\Theta^\top \mathbf{X})$.*

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

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

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

For the k -th summand, let

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

we have

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

One can upper bound these two terms separately as follows:

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

and for the second term,

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

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

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

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

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

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, then

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

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

Proof of Proposition 2. Without loss of generality, we prove the case of $k = 1$. The existence $\{\bar{\mathbf{W}}_1, \bar{\mathbf{V}}_1\}$ is by [27]. We first show that $\bar{\mathbf{W}}_1 \propto \mathbf{u}_1 \mathbf{g}_1^\top$, which is equivalent to the statement that $\frac{\|\Pi_{\bar{\mathbf{u}}_1}^\perp \bar{\mathbf{W}}_1\|_F}{\|\bar{\mathbf{W}}_1\|_F} = 0$, and we prove by contradiction. Suppose $\frac{\|\Pi_{\bar{\mathbf{u}}_1}^\perp \bar{\mathbf{W}}_1\|_F}{\|\bar{\mathbf{W}}_1\|_F} > 0$, which necessarily implies that $\exists \rho > 0$ such that $\frac{\|\Pi_{\bar{\mathbf{u}}_1}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top\|_F}{\|\bar{\mathbf{W}}_1 \bar{\mathbf{W}}_1^\top\|_F} = \rho$, then for any $\epsilon > 0, M > 0$ exists $T_{\epsilon, M} > 0$ such that $\forall t \geq T_{\epsilon, M}$, 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 choice of ϵ and M later.

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

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

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

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

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, $\|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \bar{\mathbf{W}}_1 \bar{\mathbf{V}}_1^T (\mathbf{e}_1 - \mathbf{e}_k)\| = 0, \forall k \neq 1$, which can not happen. Since $\mathbf{e}_1 - \mathbf{e}_k, k \neq 1$ spans a $k - 1$ -dimensional subspace 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 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 4 (these columns must have at least $\frac{\sqrt{2}}{2}$ cosine alignment with $\bar{\mathbf{e}}_1$).

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

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

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 sufficiently large to apply Lemma 13 so that $\frac{d}{dt} \|\Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{W}_1\|_F^2 = 2\text{tr} \left((\mathbf{e}_1 \mathbb{1}_n^T - \hat{\mathbf{Y}})^\top \mathbf{V}_1 \mathbf{W}_1^\top \Pi_{\bar{\mathbf{u}}_\infty}^\perp \mathbf{X} \right) \leq 0$.

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 that $\bar{\mathbf{W}}_1 \propto \mathbf{u}_1 \mathbf{g}_1^\top$.

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 to show that $\bar{\mathbf{v}}_1 = \tilde{\mathbf{e}}_1$, which is proved by again contradiction.

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 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 (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 $\max_{j \in \mathcal{N}_1} \|\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} - \bar{\mathbf{v}}_1\| \leq \epsilon_1$ and $\max_{i \in [n]} \|[0 \quad -\exp(\tilde{z}_{i2}) \quad \cdots \quad -\exp(\tilde{z}_{iK})]/\sum_{k>1} \exp(\tilde{z}_{ik}) - \mathbf{e}_{k^*}\| \leq \epsilon_2$. Therefore, for some $j \in [h]$ and $t > T_{\epsilon_1, \epsilon_2}$, we have, from (55)

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

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 $\Theta(\log(t))$. Therefore (74) suggests the divergence of $\mathcal{B}_j^{\mathbf{v}_j}$, a contradiction.

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 are determined by the fact that $\bar{\mathbf{W}}_k, \bar{\mathbf{V}}_k$ must be a KKT point of (66). \square