Linear Loss Classification: Efficient Training Through Neural Collapse

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Logistic loss is widely used for classification, as neural networks trained by gradient descent (GD) on this loss exhibit strong generalization. This success is commonly attributed to inductive biases, such as the directional convergence of the last-layer classifier toward the max-margin solution for a given feature representation, and neural collapse (NC), a terminal-phase phenomenon characterized by structural simplification of last-layer features and classifiers. However, they can emerge slowly due to exponentially decaying gradients. In this work, we introduce *linear loss* $l(u) = -u$, which eliminates gradient decay and leads to faster training dynamics. Under this loss, GD no longer directionally converges to the max-margin solution, but instead aligns with the difference between class means. While this may appear suboptimal, once NC occurs, the two directions become closely aligned, mitigating the discrepancy. Empirically, we demonstrate that GD with the linear loss, combined with appropriate normalizations, induces NC, and that both NC and generalization occur faster than with logistic loss. On the theoretical side, we prove that two-layer neural networks with ReLU activation trained with GD on linear loss exhibit directional NC for orthogonally separable data. Together, these results suggest that the linear loss, despite its simplicity and deviation from standard classification losses, can be sufficient to induce NC and thereby achieve strong generalization faster than other losses.

1. Introduction

Gradient-based methods are used extensively to solve modern machine learning problems, including classification tasks. Among various classification losses, the logistic loss (cross-entropy loss) is the mostly used, as neural networks trained by gradient descent on this loss exhibit remarkably strong generalization performance. This success is often attributed to inductive biases induced by such training dynamics. For instance, Soudry et al. [13], Ji and Telgarsky [6] and Chizat and Bach [3] showed that the last-layer classifier converges to a particular solution that generalizes well, namely, the max-margin direction for separable datasets. Moreover, Papyan et al. [11] identified an empirical phenomenon termed *neural collapse*, characterized by the structural simplification of last-layer features and classifiers during the terminal phase of training. This simplified structure often leads to strong generalization performance and improved adversarial robustness.

Despite these advantages, a key limitation of the logistic loss is that the emergence of inductive biases is slowed by the exponentially decaying gradients. In particular, Soudry et al. [13] showed that while the loss decreases at a rate $\mathcal{O}(1/t)$, but the classifier converges toward the max-margin direction only logarithmically, at a rate $\mathcal{O}(1/\log t)$. To accelerate this convergence, Ji and Telgarsky [7] and Zhang et al. [14] proposed using an adaptive step size inversely proportional to the current

loss value. This approach yields an $\mathcal{O}(1/t)$ convergence rate for the classifier itself, but requires full-batch computations and is therefore not suitable for large-scale datasets.

In this paper, we introduce the linear loss $l(u) = -u$, which does not suffer from gradient decay. At first glance, this loss may appear unsuitable for classification, as it is unbounded below and its minimization leads to a divergent trajectory. However, this does not pose a fundamental issue: minimizing the logistic loss also results in a divergent yet directionally meaningful trajectory. Another apparent drawback is that the resulting last-layer classifier does not exhibit implicit bias toward the max-margin direction. Nevertheless, we show that the linear loss is well suited for classification and can lead to more efficient training than the logistic loss. In particular, we demonstrate that neural networks trained with the linear loss exhibit NC, including the collapse of within-class variance of last-layer features. As feature representations within each class concentrate toward a single point, the role of the implicit bias of the last-layer classifier becomes less critical. Most importantly, our empirical results show that ReLU networks, trained with the linear loss, combined with appropriate normalization techniques, exhibit NC more rapidly than those trained with the logistic loss, leading to the faster convergence to strong generalization performance.

1.1. Our Contribution

- In Section 3, we introduce the linear loss and analyze its properties, demonstrating its suitability as a classification loss.
- In Section 4, we provide empirical evidence that neural collapse (NC) emerges under the linear loss when combined with appropriate normalization techniques. Our results further present that linear loss induces a faster emergence of NC and accelerates convergence to strong generalization performance.
- In Section 5, we theoretically show that the linear loss induces NC under the unconstrained feature model (UFM), where the last-layer features are treated as free variables. Going beyond this restrictive assumption, we prove that two-layer ReLU networks exhibit directional collapse of feature representations, without relying on the UFM.

2. Backgrounds of Neural Collapse

Recent theoretical and empirical studies suggest that continuing the training even after achieving perfect training accuracy can further improve model performance. A notable observation in this regime is *neural collapse*(NC) [11], which refers to the collapse of within-class variability of last-layer features, along with a structured alignment of features and classifiers.

To formally describe NC, we consider training a deep neural network on a dataset containing C classes with N examples per class. Let $\boldsymbol{x}_{i,c}$ denote the ith example of class c, and let $\boldsymbol{h}_{i,c}$ be its last-layer feature obtained by passing $\boldsymbol{x}_{i,c}$ through all layers except the last layer. The last-layer classifier predicts labels according to the decision rule $\arg\max_{c'} \langle \boldsymbol{w}_{c'}, \boldsymbol{h}_{i,c} \rangle + \boldsymbol{b}_{c'}$. We define the class means and the global mean of features as $\boldsymbol{\mu}_c := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}_{i,c}$ and $\boldsymbol{\mu}_G = \frac{1}{C}\sum_{c=1}^{C} \boldsymbol{\mu}_c$, respectively. We further define the within-class covariance and between-class covariance as:

$$\boldsymbol{\Sigma}_W := \frac{1}{CN}\sum_{c=1}^{C}\sum_{i=1}^{N}(\boldsymbol{h}_{i,c} - \boldsymbol{\mu}_c)(\boldsymbol{h}_{i,c} - \boldsymbol{\mu}_c)^\top, \quad \boldsymbol{\Sigma}_B := \frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top$$

At the terminal phase of training (TPT), which begins once the training accuracy reaches 100%, neural collapse (NC) is said to occur when the last-layer features and classifiers exhibit the following properties:

- **(NC1) Within-class variability collapse: $\boldsymbol{\Sigma}_W \to \mathbf{0}$**
- **(NC2) Convergence to simplex ETF[1]:** For any pair of distinct classes $1 \le c_1 \ne c_2 \le C$, $\|\boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_G\| - \|\boldsymbol{\mu}_{c_2} - \boldsymbol{\mu}_G\| \to 0$ and $\cos(\boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{c_2} - \boldsymbol{\mu}_G) \to -\frac{1}{C-1}$
- **(NC3) Convergence to self-duality:** For any class $1 \le c \le C$, $\frac{\boldsymbol{w}_c}{\|\boldsymbol{w}_c\|} - \frac{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|} \to \mathbf{0}$.

3. The Linear Loss

In the remainder of this paper, we focus on binary classification. For a neural network $f(\cdot;\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, and a dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, the empirical risk is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} l(y_i f(\boldsymbol{x}_i, \boldsymbol{\theta})) = \sum_{i=1}^{n} -y_i f(\boldsymbol{x}_i, \boldsymbol{\theta}), \tag{1}$$

where we introduce the linear loss $l(u) = -u$.

This loss is motivated by prior work on exponentially growing adaptive step sizes on the logistic loss [7, 14]. While these approaches compensate for gradient decay by increasing the step size, we instead consider a loss function that inherently avoids such decay. The linear loss is also a suitable classification loss, as it is Bayes consistent (see Section A): minimizing it recovers Bayes' optimal decision rule. Although this loss function is unbounded below and its minimization leads to a divergent trajectory, this does not pose a fundamental issue. Indeed, minimizing the logistic loss also yields a divergent yet directionally meaningful trajectory [13].

Note that $\arg\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^\top(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$, where $\boldsymbol{\mu}_\pm$ are mean vectors of positive/negative classes. This implies that classifier converge in direction toward the mean difference direction $\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$, instead of the max-margin one. This observation suggests that linear loss automatically satisfies NC3. Moreover, maximizing $\boldsymbol{w}^\top(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$ can also be viewed as the maximization of between-class variance. Although this is not exactly the same as Fisher's linear discriminant [5], which simultaneously maximizes between-class variance and minimizes within-class variance, we argue that our linear loss plays a similar role once NC1, the shrinkage of within-class variance occurs.

4. Numerical Experiments: NC under Linear Loss

We conducted experiments to evaluate the effectiveness of the linear loss. We train a three-layer neural network to classify digits 0 and 8 of the MNIST dataset, and compare the evolution of neural collapse (NC) measures and test accuracy under the linear loss and logistic loss.

We find that weight normalization is particularly effective for several reasons. Since the linear loss does not saturate, gradient norms grow much faster than under logistic loss, leading to numerical instability. Moreover, the loss is homogeneous with respect to network weights, so a gradient

1. Simplex equiangular tight frame (ETF) is the collection of vectors that have same length and maximally pairwise-distanced.

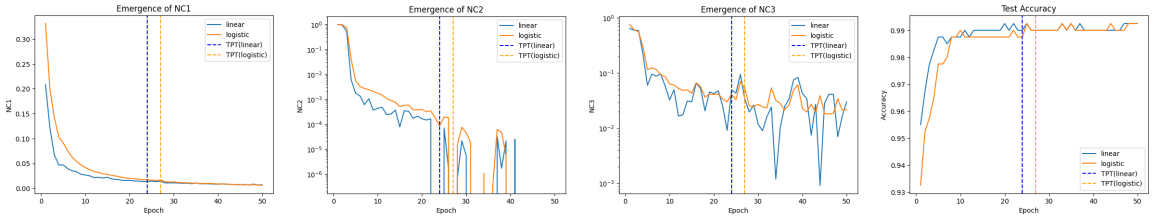

Figure 1: Test accuracy and NC metrics of linear/logistic loss

step on a normalized network can be interpreted as a small-step update on the unnormalized network. This suggests that normalization does not distort the learning dynamics. In our experiments, we normalize the weights of each neuron to have a fixed norm. Detailed experimental setups and definitions of NC measures are provided in Section B.

Figure 1 shows the evolution of NC metrics and test accuracy for networks trained with the linear and logistic losses. While both losses achieve similar final test accuracy, the network trained with the linear loss reaches this performance faster. Moreover, each NC phenomenon emerges more rapidly under the linear loss.

5. Theoretical Analysis of NC under Linear Loss

In this section, we provide theoretical guarantees for the emergence of NC under the linear loss. We also explain that empirical benefits of batch and weight normalization under the linear loss from a theoretical perspective.

5.1. NC on Unconstrained Feature Model

The unconstrained feature model (UFM) is an abstraction of deep neural networks that treats the last-layer feature as free variable [10]. This assumption is well justified in overparametrized regimes, where models can interpolate arbitrary training data.

Under the UFM, we show that the last-layer classifier and features converge in direction under gradient flow. Furthermore, they satisfy NC2, NC3 and a weaker form of NC1: $\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger} \to \mathbf{0}$ (see Section C.1). We further show that weight normalization facilitates the complete collapse of within-class variance, thereby inducing full NC (see Section C.2). These results suggest that NC can arise under the linear loss, when the model is sufficiently overparametrized.

5.2. NC on Two-Layer ReLU Network

Going beyond the restrictive UFM assumption, we consider a two-layer ReLU network:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{v}^T \sigma(\boldsymbol{W}^\top \boldsymbol{x}) = \sum_{j=1}^{h} v_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle), \quad \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} -y_i f(\boldsymbol{x}_i, \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{v} = [v_1, \cdots, v_h]^\top$ and $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_h]^\top$, trained with the linear loss. We denote the feature mapping as $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}) := [\sigma(\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle), \cdots, \sigma(\langle \boldsymbol{w}_h, \boldsymbol{x}\rangle)]^\top$, and the trajectory of weights under gradient flow as $\boldsymbol{v}(t), \boldsymbol{W}(t)$. We impose the following mild assumptions.

Assumption 1 (Orthogonal separability) *There exist positive number* $0 < \mu \leq 1$ *such that for any* $1 \leq i, j \leq n$, $\cos(y_i \boldsymbol{x}_i, y_j \boldsymbol{x}_j) \geq \mu$.

Assumption 2 (Balanced Initialization) $\|\boldsymbol{w}_j(0)\| = |v_j(0)|, \ \forall 1 \leq j \leq h$.

Assumption 3 (Non-extreme initialization) *If* $v_j(0)$ *is positive,* $\cos(\boldsymbol{w}_j(0), \boldsymbol{x}_-) < 1$ *where* $\boldsymbol{x}_- = \sum_{y_i=-1} \boldsymbol{x}_i$. *Conversely, if* $v_j(0)$ *is negative,* $\cos(\boldsymbol{w}_j(0), \boldsymbol{x}_+) < 1$ *where* $\boldsymbol{x}_+ = \sum_{y_i=+1} \boldsymbol{x}_i$.

Now we formally state the neural collapse of a two-later ReLU network trained by gradient flow with linear loss.

Theorem 4 (NC of GF on a two-layer ReLU network trained with linear loss) *Under Assumptions 1, 2 and 3,* $\bar{\boldsymbol{v}} = \lim_{t\to\infty} \frac{\boldsymbol{v}(t)}{\|\boldsymbol{v}(t)\|}$, $\bar{\boldsymbol{W}} = \lim_{t\to\infty} \frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|}$ *exist. Moreover, there exist nonnegative constants* $s_+(t)$, $s_-(t)$ *and unit vectors* $\boldsymbol{u}_+(t), \boldsymbol{u}_-(t)$ *such that*

- *(Directional collapse of features) The feature mappings of each class satisfy*

$$\boldsymbol{\phi}_{\boldsymbol{W}(t),\boldsymbol{v}(t)}(\boldsymbol{x}_i) - \langle s_+(t)\boldsymbol{x}_+, \boldsymbol{x}_i\rangle \boldsymbol{u}_+(t) \underset{t\to\infty}{\to} \boldsymbol{0}, \quad \forall \boldsymbol{x}_i \text{ s.t. } y_i = +1$$

$$\boldsymbol{\phi}_{\boldsymbol{W}(t),\boldsymbol{v}(t)}(\boldsymbol{x}_i) - \langle s_-(t)\boldsymbol{x}_-, \boldsymbol{x}_i\rangle \boldsymbol{u}_-(t) \underset{t\to\infty}{\to} \boldsymbol{0}, \quad \forall \boldsymbol{x}_i \text{ s.t. } y_i = -1.$$

  Moreover, the limits $\bar{\boldsymbol{u}}_+ = \lim_{t\to\infty} \boldsymbol{u}_+(t)$, $\bar{\boldsymbol{u}}_- = \lim_{t\to\infty} \boldsymbol{u}_-(t)$ *exist.*

- *(Orthogonal class means)* $\langle \bar{\boldsymbol{u}}_+, \bar{\boldsymbol{u}}_- \rangle = 0$
- *(Alignment of classifier and features)* $\boldsymbol{v}(t) = s_+(t)\boldsymbol{u}_+(t) - s_-(t)\boldsymbol{u}_-(t)$

There are several caveats about this theorem. First, two-layer ReLU networks do not exhibit 'complete' neural collapse, as its features lie in a one-dimensional subspace, instead of shrinking to a single point. This is not a flaw of linear loss: it comes from the limited expressiveness of the shallow structure. Two-layer ReLU networks trained by logistic loss also experience such 'directional' collapse [9]. Second, unlike logistic loss where $\lim_{t\to\infty} s_+(t)/s_-(t)$ exists, the linear loss would not allow this limit. More concretely, for linear loss, $s_+(t) = \Theta(e^{\|\boldsymbol{x}_+\|t})$, $\quad s_-(t) = \Theta(e^{\|\boldsymbol{x}_-\|t})$. This may cause severe asymmetry between two classes when $\|\boldsymbol{x}_+\| \neq \|\boldsymbol{x}_-\|$, although $\bar{\boldsymbol{v}}$, $\bar{\boldsymbol{W}}$ still exist in that case. Nevertheless, we claim that such an asymmetry can be prevented in a deep network equipped with batch normalization, since batch normalization maintains the balance between $\|\boldsymbol{x}_+\|$ and $\|\boldsymbol{x}_-\|$. We also argue that the directional collapse can also appear under weight normalization, and in that case the restriction in magnitudes prevents the severe asymmetry $\max\{s_+/s_-, s_-/s_+\} \to \infty$(Section D.4).

6. Conclusion and Future Directions

In this paper, we introduced the linear loss $l(u) = -u$ and investigated its benefits in binary classification. We showed that the linear loss is theoretically well grounded, exhibiting both Bayes consistency and the emergence of neural collapse (NC). Empirically, we demonstrated that it accelerates the emergence of NC and leads to faster convergence to strong generalization performance. Taken together, these results suggest that the linear loss provides a simple and effective alternative to standard classification losses. Our analysis is currently limited to binary classification and shallow fully connected networks. Extending these results to multi-class settings and broader architectures, such as convolutional neural networks, is an important direction for future work.

References

[1] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17:1205–1223, 01 2007. doi: 10.1137/050644641.

[2] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs, 2026. URL https://arxiv.org/abs/2206.00939.

[3] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss, 2020. URL https://arxiv.org/abs/2002.04486.

[4] Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced, 2018. URL https://arxiv.org/abs/1806.00900.

[5] R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x.

[6] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/ji19a.html.

[7] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis, 2020. URL https://arxiv.org/abs/1906.04540.

[8] Hancheng Min, Enrique Mallada, and René Vidal. Early neuron alignment in two-layer relu networks with small initialization, 2024. URL https://arxiv.org/abs/2307.12851.

[9] Hancheng Min, Zhihui Zhu, and René Vidal. Neural collapse under gradient flow on shallow relu networks for orthogonally separable data, 2025. URL https://arxiv.org/abs/2510.21078.

[10] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *CoRR*, abs/2011.11619, 2020. URL https://arxiv.org/abs/2011.11619.

[11] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, September 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117. URL http://dx.doi.org/10.1073/pnas.2015509117.

[12] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *CoRR*, abs/1602.07868, 2016. URL http://arxiv.org/abs/1602.07868.

[13] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data, 2024. URL https://arxiv.org/abs/1710.10345.

[14] Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. Gradient descent converges arbitrarily fast for logistic regression via large and adaptive stepsizes. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=cufJbug7du.

Appendix A. Proof of Bayes Consistency of Linear Loss

Define $p_i = P(y_i = 1|\boldsymbol{x}_i)$. Then the expected risk can be written as

$$\begin{aligned} R_l(f) &= E_{\boldsymbol{x}\in\mathcal{X}, y\in\mathcal{Y}}[-yf(\boldsymbol{x})] && (3)\\ &= \sum_i [P(\boldsymbol{x} = \boldsymbol{x}_i, y = 1) \times (-f(\boldsymbol{x}_i)) + P(\boldsymbol{x} = \boldsymbol{x}_i, y = -1) \times f(\boldsymbol{x}_i)] && (4)\\ &= \sum_i P(\boldsymbol{x} = \boldsymbol{x}_i)[p_i \times (-f(\boldsymbol{x}_i)) + (1 - p_i) \times f(\boldsymbol{x}_i)] && (5)\\ &= \sum_i P(\boldsymbol{x} = \boldsymbol{x}_i)(1 - 2p_i)f(\boldsymbol{x}_i) && (6) \end{aligned}$$

This value is minimized when

$$f(\boldsymbol{x}_i) \to \begin{cases} +\infty & \text{if } p_i > \frac{1}{2} \\ -\infty & \text{if } p_i < \frac{1}{2} \end{cases}, \tag{7}$$

or equivalently

$$\operatorname{sgn}(f(\boldsymbol{x}_i)) = \begin{cases} +1 & \text{if } P(y_i = 1|\boldsymbol{x}_i) > P(y_i = -1|\boldsymbol{x}_i) \\ -1 & \text{if } P(y_i = 1|\boldsymbol{x}_i) < P(y_i = -1|\boldsymbol{x}_i) \end{cases}, \tag{8}$$

which is exactly Bayes optimal decision rule.

Appendix B. Detailed Setup for Numerical Experiment

B.1. Problem Setup

We consider a binary classification of digit 0 and digit 8 in MNIST dataset. We use only 2000 train examples and 400 test examples, to focus on TPT.

B.2. Model and Architecture

We train three-layer(two hidden layers) fully connected ReLU network with width 32. Batch normalization is applied before every ReLU activation. We use near-zero initialization, and train both networks by GD with stepsize 0.001. For linear loss, we normalize the norm of each neuron to be a constant(output layer : 0.2, other layers : 0.5).

B.3. NC Metrics

We define several metrics to measure the emergence of NC1∼NC3, based on their definition. We calculate $\operatorname{tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^\dagger)$ to measure NC1, and calculate $\| \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-}{\|\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-\|} \|$ to measure NC3. In binary classification, NC2 is always true, since in that case $\boldsymbol{\mu}_G = \frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)$ and hence automatically $\boldsymbol{\mu}_+ - \boldsymbol{\mu}_G = -(\boldsymbol{\mu}_- - \boldsymbol{\mu}_G)$. However, we check $\langle \frac{\boldsymbol{\mu}_+}{\|\boldsymbol{\mu}_+\|}, \frac{\boldsymbol{\mu}_-}{\|\boldsymbol{\mu}_-\|} \rangle$ as a metric of NC2, since it measure the progress of neural collapse(second part of Theorem 4), and the orthogonality of class means implies the formulation of simplex ETF consists of centered class means, even for the multi-class problem(Min et al. [9]).

Appendix C. Proof of Results in Section 5.1

C.1. Emergence of Weaker Form of NC under UFM

Let $\boldsymbol{H}$ be a collection of $\boldsymbol{h}_i$, which is the feature mapping of $\boldsymbol{x}_i$. Then Our problem becomes

$$\min_{\boldsymbol{H},\boldsymbol{v},b} \mathcal{L}(\boldsymbol{H},\boldsymbol{v},b) = -\sum_{i=1}^{n} y_i(\boldsymbol{v}^\top \boldsymbol{h}_i + b) \tag{9}$$

Let $\mathcal{I}_\pm$ is index set of positive/negative label. Then GF yields

$$\dot{\boldsymbol{v}} = \sum_{i=1}^{n} y_i \boldsymbol{h}_i = \sum_{i\in\mathcal{I}^+} \boldsymbol{h}_i - \sum_{i\in\mathcal{I}^-} \boldsymbol{h}_i, \tag{10}$$

$$\dot{\boldsymbol{h}}_i = y_i \boldsymbol{v} = \begin{cases} \boldsymbol{v} & (i \in \mathcal{I}_+) \\ -\boldsymbol{v} & (i \in \mathcal{I}_-) \end{cases} \tag{11}$$

Then combining Equation (10) and Equation (11) yields

$$\ddot{\boldsymbol{v}} = n\boldsymbol{v}, \boldsymbol{v}(0) = \boldsymbol{v}_0, \dot{\boldsymbol{v}}(0) = \boldsymbol{h}_0 \tag{12}$$

where $\boldsymbol{h}_0 = [\sum_{i\in\mathcal{I}^+} \boldsymbol{h}_i - \sum_{i\in\mathcal{I}^-} \boldsymbol{h}_i]_{t=0}$. The solution is

$$\boldsymbol{v}(t) = \frac{1}{2}\left(\boldsymbol{v}_0 + \frac{\boldsymbol{h}_0}{\sqrt{n}}\right) e^{\sqrt{n}t} + \frac{1}{2}\left(\boldsymbol{v}_0 - \frac{\boldsymbol{h}_0}{\sqrt{n}}\right) e^{-\sqrt{n}t} \tag{13}$$

$$\begin{aligned} \boldsymbol{h}_i(t) &= \boldsymbol{h}_i(0) + y_i \int_0^t \boldsymbol{v}(\tau) d\tau \\ &= \boldsymbol{h}_i(0) + y_i \left[\frac{1}{2\sqrt{n}}\left(\boldsymbol{v}_0 + \frac{\boldsymbol{h}_0}{\sqrt{n}}\right) e^{\sqrt{n}t} - \frac{1}{2\sqrt{n}}\left(\boldsymbol{v}_0 - \frac{\boldsymbol{h}_0}{\sqrt{n}}\right) e^{-\sqrt{n}t}\right] \end{aligned}$$

We emphasize that this is not 'complete NC', because features of each class don't shrink into a single point. However, since each class are pushed indefinitely toward the direction $\pm\left(\boldsymbol{v}_0 + \frac{\boldsymbol{h}_0}{\sqrt{n}}\right)$, within-class variability becomes relatively negligible, compared to between-class varaibility. This yields similar phenomena to NC.

- Since all points in same class move identically, within-class variability remains constant. Nevertheless, between-class variability grows indefinitely, so $\mathrm{tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^\dagger) \to 0$, which is weaker form of NC1.
- In binary classification, NC2 is automatically satisfied, as described in Section B.
- Since $\boldsymbol{v}, \boldsymbol{h}_+$ converges in direction to $\boldsymbol{v}_0 + \frac{\boldsymbol{h}_0}{\sqrt{n}}$ and $\boldsymbol{h}_-$ converges in direction to $-\left(\boldsymbol{v}_0 + \frac{\boldsymbol{h}_0}{\sqrt{n}}\right)$, we can deduce that

  $$\left\| \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{\boldsymbol{h}_+ - \boldsymbol{h}_-}{\|\boldsymbol{h}_+ - \boldsymbol{h}_-\|} \right\| \to 0, \tag{14}$$

  which is exactly NC3.

C.2. Complete NC under UFM with Weight Normalization

We already saw that within-class variance does not vanishes, despite the UFM assumption(extremely overparametrized regime). This is because, When $\boldsymbol{h}_i, \boldsymbol{h}_j$ are in same class, $\boldsymbol{h}_i - \boldsymbol{h}_j$ always remains constant, while $\frac{\|\boldsymbol{h}_i - \boldsymbol{h}_j\|}{\|\boldsymbol{h}_i\|}$ converges to 0. However, if there is weight normalization, norm of $\boldsymbol{h}_i, \boldsymbol{h}_j$ are suppressed while preserving their direction, so $\boldsymbol{h}_i - \boldsymbol{h}_j$ will decrease gradually.

To formalize, suppose that model is overparametrized enough, but the norm of features is limited due to the weight normalization. Then these hidden layers can interpolate any feature $\boldsymbol{h}$ with norm restriction $\|\boldsymbol{h}\| \leq C$. In short, we consider

$$\min_{\boldsymbol{H},\boldsymbol{v}} -\sum_{i=1}^{n} y_i \boldsymbol{v}^\top \boldsymbol{h}_i, \quad \text{s.t. } \|\boldsymbol{v}\|, \|\boldsymbol{h}_i\| \leq 1. \tag{15}$$

In this case we have following differential inclusions :

$$\frac{d\boldsymbol{v}}{dt} \in \sum_{i=1}^{n} y_i \boldsymbol{h}_i - \mathcal{N}_{B_{\boldsymbol{v}}}(\boldsymbol{v}), \quad \frac{d\boldsymbol{h}_i}{dt} \in y_i \boldsymbol{v} - \mathcal{N}_{B_{\boldsymbol{h}}}(\boldsymbol{v}). \tag{16}$$

Here, $B_{\boldsymbol{x}} = \{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leq 1\}$. Since $\mathcal{N}_{B_{\boldsymbol{x}}} = \begin{cases} 0 & \|\boldsymbol{x}\| < 1 \\ \{\lambda \boldsymbol{x} : \lambda \geq 0\} & \|\boldsymbol{x}\| = 1 \end{cases}$, the global minima should satisfy

$$\exists \lambda \geq 0 \text{ s.t. } \lambda \boldsymbol{v} = \sum_{i=1}^{n} y_i \boldsymbol{h}_i, \quad \forall i,\ \exists \mu_i \geq 0 \text{ s.t. } \mu_i \boldsymbol{h}_i = y_i \boldsymbol{v} \tag{17}$$

This implies $\boldsymbol{h}_i = y_i \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$ and $\boldsymbol{v} = \sum_{i=1}^{n} y_i \boldsymbol{h}_i = \boldsymbol{h}_+ - \boldsymbol{h}_-$, which is the complete NC.

Appendix D. Proof of Results in Section 5.2

In this section, we prove Theorem 4. The proofs, especially those in Section D.2, are based on Min et al. [8], Min et al. [9], whose idea is roughly as follows.

- (Gradient flow analysis) By calculating $\frac{d}{dt}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, one can prove the following: if the initial sign of v_j is positive/negative, then $\boldsymbol{w}_j$ will be attracted by positively/negatively labeled data.
- (Early neural alignment) Hence, if the initial sign of v_j is positive/negative, then $\boldsymbol{w}_j$ will only activate positively/negatively labeled data after a sufficient time.
- (Asymptotic analysis) As a result, the entire network is decomposed into two linear two-layer networks. Existing asymptotic convergence results for linear networks imply the directional collapse of features and subsequent NC phenomena.

There are two major differences in the analysis of Min et al. [9] and ours. First, surprisingly, our linear loss removes the error term when analyzing $\frac{d}{dt}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$. This not only simplifies the proof about neural alignment, but also accelerates the alignment. Second, since the loss function is different, the asymptotic behaviors of NN become completely different. In Section D.3 we have proved that the linear loss still allows the directional collapse despite those differences.

D.1. Gradient Flow Analysis

Recall that

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = \sum_{i=1}^{n} -y_i \boldsymbol{v}^\top \sigma(\boldsymbol{W}^\top \boldsymbol{x}_i) = \sum_{i=1}^{n} \sum_{j=1}^{h} -y_i v_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x}_i). \tag{18}$$

Hence,

$$\frac{d}{dt}\boldsymbol{w}_j = \partial_{\boldsymbol{w}_j} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = v_j \sum_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle > 0} y_i \boldsymbol{x}_i \tag{19}$$

and

$$\begin{aligned}
\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} &= \frac{1}{\|\boldsymbol{w}_j\|}\left(\boldsymbol{I} - \frac{\boldsymbol{w}_j \boldsymbol{w}_j^\top}{\|\boldsymbol{w}_j\|^2}\right)\left(v_j \sum_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle > 0} y_i \boldsymbol{x}_i\right) \\
&= \operatorname{sgn}(v_j)\left(\boldsymbol{I} - \frac{\boldsymbol{w}_j \boldsymbol{w}_j^\top}{\|\boldsymbol{w}_j\|^2}\right) \sum_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle > 0} y_i \boldsymbol{x}_i
\end{aligned}$$

Where the last equality is due to the balanced initialization : under balanced initialization, the balance is maintained throughout the GF training (Du et al. [4]). We also emphasize that the sign of v_j does not change throughout the training (Boursier et al. [2]).

This formula gives us intuition about the change of input neuron $\boldsymbol{w}_j$. If $v_j(t) > 0$, then $\boldsymbol{w}_j$ moves toward $\sum_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle > 0} y_i \boldsymbol{x}_i$. As a result, $\boldsymbol{w}_j$ is attracted by positively labeled data and repelled by negatively labeled data. In contrast, if $v_j(t) < 0$, $\boldsymbol{w}_j$ is repelled by positively labeled data and attracted by negatively labeled data. It is known that $\operatorname{sgn}(v_j)$ does not change over time if initialization is balanced(Boursier et al. [2]), the input neurons are separated into two groups, according to the sign of the corresponding output neuron v_j. This observation is a key of neural alignment. The quantity$\sum_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle > 0} y_i \boldsymbol{x}_i$ will be appear multiple time, so we denote it by $\boldsymbol{x}_a(\boldsymbol{w}_j)$ for simplification.

D.2. Early Neuron Alignment

Throughout this section, we need several notations for simplification. We first separate datapoints and output neurons according to their sign.

$$\mathcal{I}_+ = \{i \in [n] \mid y_i = +1\}, \quad \mathcal{I}_- = \{i \in [n] \mid y_i = -1\} \tag{20}$$

$$\mathcal{N}_+(t) = \{j \in [h] \mid v_j(t) > 0\}, \quad \mathcal{N}_-(t) = \{j \in [h] \mid v_j(t) < 0\} \tag{21}$$

Note that in balanced initialization, the sign of $v_j(t)$ does not change over time, so we simply denote $\mathcal{N}_+ = \mathcal{N}_+(0)$, $\mathcal{N}_- = \mathcal{N}_-(0)$ and consider only those sets. We additionally use the following notation:

$$x_{\min} := \min_{1 \le i \le n} \|\boldsymbol{x}_i\|, \quad \boldsymbol{x}_+ := \sum_{i \in \mathcal{I}_+} \boldsymbol{x}_i, \quad \boldsymbol{x}_- := \sum_{i \in \mathcal{I}_-} \boldsymbol{x}_i \tag{22}$$

We also define

$$\mathcal{S}_+ = \{\boldsymbol{w} \mid \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0 \Leftrightarrow y_i = +1,\ \forall i\}, \quad \mathcal{S}_- = \{\boldsymbol{w} \mid \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0 \Leftrightarrow y_i = -1,\ \forall i\} \tag{23}$$

and

$$\mathcal{S}_{dead} = \{\boldsymbol{w} \mid \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle < 0,\ \forall i\} \tag{24}$$

Here are a few remarks about definitions above.

- It is easy to see that $\mathcal{S}_+$, $\mathcal{S}_-$, $\mathcal{S}_{dead}$ are cones except for the absence of the origin. Moreover, $\mathrm{Int}(\mathcal{S}_+) = -\mathrm{Int}(\mathcal{S}_-)$.
- If $y_i = +1$, then $\boldsymbol{x}_i \in \mathcal{S}_+$ and $-\boldsymbol{x}_i \in \mathcal{S}_-$ due to orthogonal separability. Similarly, if $y_i = -1$, $\boldsymbol{x}_i \in \mathcal{S}_-$ and $-\boldsymbol{x}_i \in \mathcal{S}_+$. In particular, $\boldsymbol{x}_a(\boldsymbol{w}) \in \mathcal{S}_+,\ -\boldsymbol{x}_a(\boldsymbol{w}) \in \mathcal{S}_-$ for any $\boldsymbol{w}$

Now we give our first important observation.

Claim 1. $\boldsymbol{x}_a(\boldsymbol{w}_j)$ will change its value only finitely many times.

To prove this claim, we introduce several lemmas.

Lemma 5 *Let $\boldsymbol{x}_r \in \mathbb{S}^{D-1}$ be a fixed reference vector. Then*

$$\frac{d}{dt}\cos(\boldsymbol{x}_r, \boldsymbol{w}_j) = \mathrm{sgn}(v_j)(\cos(\boldsymbol{x}_r, \boldsymbol{x}_a(\boldsymbol{w}_j)) - \cos(\boldsymbol{x}_r, \boldsymbol{w}_j)\cos(\boldsymbol{w}_j, \boldsymbol{x}_a(\boldsymbol{w}_j)))\|\boldsymbol{x}_a(\boldsymbol{w}_j)\| \tag{25}$$

This is a direct consequence of the result in Section D.1.

Lemma 6 (Lemma 11 of Min et al. [8])

$$\|\boldsymbol{x}_a(\boldsymbol{w})\| \geq \sqrt{\mu} n_a(\boldsymbol{w}) x_{\min}, \tag{26}$$

where $n_a(\boldsymbol{w}) = |\{i \in [n] \mid \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0\}|$.

Proof Let $\mathcal{I}_a(\boldsymbol{w}) = \{i \in [n] \mid \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0\}$. Then

$$\begin{aligned}
\|\boldsymbol{x}_a(\boldsymbol{w})\| &= \|\sum_{i \in \mathcal{I}_a(\boldsymbol{w})} y_i \boldsymbol{x}_i\| \\
&= \sqrt{\sum_{i \in \mathcal{I}_a(\boldsymbol{w})} \|y_i \boldsymbol{x}_i\|^2 + \sum_{i,j \in \mathcal{I}_a(\boldsymbol{w}), i \neq j} \langle y_i \boldsymbol{x}_i, y_j \boldsymbol{x}_j \rangle} \\
&= \sqrt{\sum_{i \in \mathcal{I}_a(\boldsymbol{w})} y_i^2 \|\boldsymbol{x}_i\|^2 + \sum_{i,j \in \mathcal{I}_a(\boldsymbol{w}), i \neq j} \|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\| \langle \frac{y_i \boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{y_j \boldsymbol{x}_j}{\|\boldsymbol{x}_j\|} \rangle} \\
&\geq \sqrt{\sum_{i \in \mathcal{I}_a(\boldsymbol{w})} \|\boldsymbol{x}_i\|^2 + \sum_{i,j \in \mathcal{I}_a(\boldsymbol{w}), i \neq j} \mu \|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|} \\
&= \sqrt{n_a(\boldsymbol{w}) x_{\min}^2 + \mu n_a(\boldsymbol{w})(n_a(\boldsymbol{w}) - 1) x_{\min}^2} \\
&= \sqrt{n_a(\boldsymbol{w}) + \mu n_a(\boldsymbol{w})(n_a(\boldsymbol{w}) - 1)} x_{\min} \\
&= \sqrt{\mu n_a(\boldsymbol{w})^2 + (1 - \mu) n_a(\boldsymbol{w})} x_{\min} \\
&\geq \sqrt{\mu} n_a(\boldsymbol{w}) x_{\min}
\end{aligned}$$

■

Proof of Claim 1. It is trivial that $\boldsymbol{w}_j \in \mathcal{S}_{dead}$ does not change $\boldsymbol{x}_a(\boldsymbol{w}_j)(=0)$. Suppose that $j \in \mathcal{N}_+$ and $\boldsymbol{w}_j \notin \mathcal{S}_{dead}$. From Lemma 5, we obtain

$$\frac{d}{dt} y_i \langle \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} \rangle |_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle = 0} = \langle \frac{y_i \boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{\boldsymbol{x}_a(\boldsymbol{w}_j)}{\|\boldsymbol{x}_a(\boldsymbol{w}_j)\|} \rangle \|\boldsymbol{x}_a(\boldsymbol{w}_j)\|. \tag{27}$$

By orthogonal separability, $\langle \frac{y_i \boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{\boldsymbol{x}_a(\boldsymbol{w}_j)}{\|\boldsymbol{x}_a(\boldsymbol{w}_j)\|} \rangle \geq \mu$. Hence, by Lemma 6,

$$\frac{d}{dt} y_i \langle \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} \rangle |_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle = 0} \geq \mu^{3/2} n_a(\boldsymbol{w}) x_{\min} > 0 \tag{28}$$

This implies two things. First, if $y_i = +1$ and $\boldsymbol{x}_i$ is activated, then it cannot be deactivated(cannot cross the boundary $\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle = 0$) since $\frac{d}{dt} \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle > 0$ at the boundary. Likewise, if $y_i = -1$ and $\boldsymbol{x}_i$ is deactivated, then it cannot be activated(cannot cross the boundary $\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle = 0$) since $\frac{d}{dt} \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle < 0$ at the boundary. Hence, the only possibilities that change $\boldsymbol{x}_a(\boldsymbol{w}_j)$ are the activation of a positively labeled datapoint or the deactivation of a negatively labeled datapoint. This means that the number of changes of $\boldsymbol{x}_a(\boldsymbol{w}_j)$ is bounded by n. The case $j \in \mathcal{N}_-$ can be proved in a similar manner. ■

Next, we give the second important observation.

Claim 2. $\boldsymbol{x}_a(\boldsymbol{w}_j)$ will change its value within a finite amount of time.

To prove the claim, we need several technical lemmas.

Lemma 7 (Formalization of nondegenerate initialization) *There exists $\zeta > 0$ s.t.*

$$\max_{j \in \mathcal{N}_+} \cos(\boldsymbol{w}_j(0), \boldsymbol{x}_-) \leq \sqrt{1-\zeta}, \quad \max_{j \in \mathcal{N}_-} \cos(\boldsymbol{w}_j(0), \boldsymbol{x}_+) \leq \sqrt{1-\zeta} \tag{29}$$

Lemma 8 (Lemma 9 of Min et al. [8]) *Let $\boldsymbol{x}_1$ be a positive linear combination of $y_i \boldsymbol{x}_i$s, and let $\boldsymbol{x}_2 \in \mathcal{S}_+^c \cap \mathcal{S}_-^c$. Then there exist $\xi > 0$ s.t.*

$$\sup |\cos(\boldsymbol{x}_1, \boldsymbol{x}_2)| \leq \sqrt{1-\xi}. \tag{30}$$

Proof of Claim 2. We focus on the case when $j \in \mathcal{N}_+$. We already saw that $\boldsymbol{x}_a(\boldsymbol{w}_j)$ does not change when $\boldsymbol{w}_j \in \mathcal{S}_+ \cup \mathcal{S}_{dead}$. Hence, we only need to consider two cases : $\boldsymbol{w}_j \in \mathcal{S}_-$ or $\boldsymbol{w}_j \in \mathcal{S}_+^c \cap \mathcal{S}_-^c \cap \mathcal{S}_{dead}^c$.

Suppose that $\boldsymbol{w}_j \in \mathcal{S}_-$. Then $\boldsymbol{x}_a(\boldsymbol{w}_j) = -\boldsymbol{x}_-$ by the definition of $\mathcal{S}_-$. Let $t_1 > 2/(\zeta \sqrt{\mu} x_{\min})$, and assume that $\boldsymbol{x}_a(\boldsymbol{w}_j(t_0)) = \boldsymbol{x}_a(\boldsymbol{w}_j(t_0 + t_1))$. Then we choose $\boldsymbol{x}_r = \boldsymbol{x}_-$ and use Lemma 5, which yields

$$\frac{d}{dt} \cos(\boldsymbol{w}_j(t), \boldsymbol{x}_-) = -(1 - \cos^2(\boldsymbol{w}_j(t), \boldsymbol{x}_-)) \|\boldsymbol{x}_-\| \tag{31}$$

This formula, together with Lemma 6 and Lemma 7, shows that for any $t \in [t_0, t_0 + t_1]$,

$$\frac{d}{dt} \cos(\boldsymbol{w}_j(t), \boldsymbol{x}_-) \leq -\zeta n_a(\boldsymbol{w}_j(t)) \|\boldsymbol{x}_-\| \leq -\zeta \sqrt{\mu} x_{\min}. \tag{32}$$

Now, by the fundamental theorem of calculus,

$$\begin{aligned}\cos(\boldsymbol{w}_j(t_0+t_1),\boldsymbol{x}_-) &= \cos(\boldsymbol{w}_j(t_0),\boldsymbol{x}_-) + \int_{t_0}^{t_0+t_1} \frac{d}{dt}\cos(\boldsymbol{w}_j(t_0+\tau),\boldsymbol{x}_-)d\tau \\ &\leq \cos(\boldsymbol{w}_j(t_0),\boldsymbol{x}_-) + \int_{t_0}^{t_0+t_1} -\zeta\sqrt{\mu}x_{\min}d\tau \\ &< \cos(\boldsymbol{w}_j(t_0),\boldsymbol{x}_-) - 2 \\ &\leq -1\end{aligned}$$

This is a contradiction, so we can deduce that $\boldsymbol{x}_a(\boldsymbol{w}_j)$ will change within a finite amount of time t_1.

Now suppose that $\boldsymbol{w}_j \in \mathcal{S}_+^c \cap \mathcal{S}_-^c \cap \mathcal{S}_{dead}^c$. Let $t_2 > 2/(\xi\sqrt{\mu}x_{\min})$, and assume that $\boldsymbol{x}_a(\boldsymbol{w}_j(t_0)) = \boldsymbol{x}_a(\boldsymbol{w}_j(t_0+t_2))$. Then for $t \in [t_0, t_0+t_2]$, we can use Lemma 5 with $\boldsymbol{x}_r = \boldsymbol{x}_a(\boldsymbol{w}_j(t_0))$ since $\boldsymbol{x}_a(\boldsymbol{w}_j)$ is fixed in that time interval. Lemma 5 yields

$$\frac{d}{dt}\cos(\boldsymbol{w}_j(t),\boldsymbol{x}_a(\boldsymbol{w}_j(t_0))) = (1-\cos^2(\boldsymbol{w}_j(t),\boldsymbol{x}_a(\boldsymbol{w}_j(t_0))))\|\boldsymbol{x}_a(\boldsymbol{w}_j(t_0))\| \tag{33}$$

Similarly, Lemma 6 and Lemma 8 implies that

$$\frac{d}{dt}\cos(\boldsymbol{w}_j(t),\boldsymbol{x}_a(\boldsymbol{w}_j(t_0))) \geq \xi\sqrt{\mu}x_{\min}, \quad \forall t \in [t_0, t_0+t_2] \tag{34}$$

Then FTC yields contradiction in a similar manner, so we can deduce that $\boldsymbol{x}_a(\boldsymbol{w}_j)$ will change within a finite amount of time t_2. The proof for the case $j \in \mathcal{N}_-$ is similar.

Now, the combination of claim 1 and claim 2 directly implies the following theorem.

Theorem 9 (Early Neural Alignment) *Let* $T = \frac{2n}{\sqrt{\mu}x_{\min}} \times \frac{1}{\min\{\zeta,\xi\}}$. *Then following statements hold :*

- *If* $j \in \mathcal{N}_+$, *then either* $\boldsymbol{w}_j(T) \in \mathcal{S}_+$ *or* $\boldsymbol{w}_j(T) \in \mathcal{S}_{dead}$, *and remains permanently in one of those cones. Moreover, if* $\max_{i\in\mathcal{I}_+}\langle \boldsymbol{w}_j(0), \boldsymbol{x}_i\rangle > 0$, *then* $\boldsymbol{w}_j(T) \in \mathcal{S}_+$.
- *If* $j \in \mathcal{N}_-$, *then either* $\boldsymbol{w}_j(T) \in \mathcal{S}_-$ *or* $\boldsymbol{w}_j(T) \in \mathcal{S}_{dead}$, *and remains permanently in one of those cones. Moreover, if* $\max_{i\in\mathcal{I}_-}\langle \boldsymbol{w}_j(0), \boldsymbol{x}_i\rangle > 0$, *then* $\boldsymbol{w}_j(T) \in \mathcal{S}_-$.

D.3. Asymptotic Behavior

If $\boldsymbol{w}_j \in \mathcal{S}_{dead}$, there is nothing we need to do. So assume that $\{\boldsymbol{w}_j\}$ are separated into $\mathcal{S}_+$ and $\mathcal{S}_-$ according to the sign of the output neuron. The proof is similar if there exists a dead neuron. From Section D.2, we can deduce that

$$\begin{aligned}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) &= \sum_{j=1}^{h}\sum_{i=1}^{n} -y_i v_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x}_i) \\ &= -\sum_{j\in\mathcal{N}_+}\sum_{i\in\mathcal{I}_+} v_j \boldsymbol{w}_j^\top \boldsymbol{x}_i + \sum_{j\in\mathcal{N}_-}\sum_{i\in\mathcal{I}_-} v_j \boldsymbol{w}_j^\top \boldsymbol{x}_i \\ &= -\sum_{i\in\mathcal{I}_+} \boldsymbol{v}_+^\top \boldsymbol{W}_+^\top \boldsymbol{x}_i + \sum_{i\in\mathcal{I}_-} \boldsymbol{v}_-^\top \boldsymbol{W}_-^\top \boldsymbol{x}_i \\ &= -\boldsymbol{v}_+^\top \boldsymbol{W}_+^\top \boldsymbol{x}_+ - \boldsymbol{v}_-^\top \boldsymbol{W}_-^\top (-\boldsymbol{x}_-)\end{aligned}$$

where $\boldsymbol{W}_{\pm} = [\boldsymbol{w}_j]_{j\in\mathcal{N}_{\pm}}$, $\boldsymbol{v}_{\pm} = [v_j]^{\top}_{j\in\mathcal{N}_{\pm}}$. In other words, neurons in $\{\boldsymbol{w}_j, v_j\}_{j\in\mathcal{N}_+}$ are fully decoupled from those in $\{\boldsymbol{w}_j, v_j\}_{j\in\mathcal{N}_-}$, and the dynamics of each set of neurons can be fully described by analyzing the behavior of two-layer linear networks trained by linear loss.

Proof We concentrate only on positive class here, since the proof for negative class is almost the same. For simplicity, we omit the subscript $\cdot_+$. Thus, we are considering loss of the following form :

$$\hat{\mathcal{L}}(\boldsymbol{v}, \boldsymbol{W}) = -\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x}. \tag{35}$$

If we train $\boldsymbol{v}, \boldsymbol{W}$ by GF, then

$$\frac{d}{dt}\boldsymbol{v} = \boldsymbol{W}^{\top}\boldsymbol{x}, \quad \frac{d}{dt}\boldsymbol{W} = \boldsymbol{x}\boldsymbol{v}^{\top} \tag{36}$$

and hence

$$\frac{d}{dt}\|\boldsymbol{v}\| = \frac{\boldsymbol{v}^{\top}\frac{d\boldsymbol{v}}{dt}}{\|\boldsymbol{v}\|} = \frac{\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x}}{\|\boldsymbol{v}\|}, \tag{37}$$

$$\frac{d}{dt}\|\boldsymbol{W}\| = \frac{\mathrm{tr}(\boldsymbol{W}^{\top}\frac{d\boldsymbol{W}}{dt})}{\|\boldsymbol{W}\|} = \frac{\mathrm{tr}(\boldsymbol{W}^{\top}\boldsymbol{x}\boldsymbol{v}^{\top})}{\|\boldsymbol{W}\|} = \frac{\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x}}{\|\boldsymbol{W}\|}. \tag{38}$$

Hence,

$$\frac{d}{dt}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \frac{\boldsymbol{W}^{\top}\boldsymbol{x}}{\|\boldsymbol{v}\|} - \frac{(\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x})\boldsymbol{v}}{\|\boldsymbol{v}\|^3}, \quad \frac{d}{dt}\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|} = \frac{\boldsymbol{x}\boldsymbol{v}^{\top}}{\|\boldsymbol{W}\|} - \frac{(\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x})\boldsymbol{W}}{\|\boldsymbol{W}\|^3}. \tag{39}$$

Now, we consider the scalar $c = \frac{\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|} = \frac{\boldsymbol{x}^{\top}\boldsymbol{W}\boldsymbol{v}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|}$ and its time derivative dc/dt.

$$\begin{aligned}
\frac{d}{dt}c =& \boldsymbol{x}^{\top}\left(\frac{d}{dt}\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}\right)\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} + \boldsymbol{x}^{\top}\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}\left(\frac{d}{dt}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}\right) \\
=& \left[\frac{\boldsymbol{x}^{\top}\boldsymbol{x}\boldsymbol{v}^{\top}\boldsymbol{v}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|} - \frac{\boldsymbol{x}^{\top}(\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x})\boldsymbol{W}\boldsymbol{v}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|^3}\right] + \left[\frac{\boldsymbol{x}^{\top}\boldsymbol{W}\boldsymbol{W}^{\top}\boldsymbol{x}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|} - \frac{\boldsymbol{x}^{\top}\boldsymbol{W}(\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x})\boldsymbol{v}}{\|\boldsymbol{v}\|^3\|\boldsymbol{W}\|}\right] \\
=& \frac{1}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|^3}\left[\|\boldsymbol{x}\|^2\|\boldsymbol{v}\|^2\|\boldsymbol{W}\|^2 - (\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x})^2\right] \\
&+ \frac{1}{\|\boldsymbol{v}\|^3\|\boldsymbol{W}\|}\left[\|\boldsymbol{v}\|^2\|\boldsymbol{W}^{\top}\boldsymbol{x}\|^2 - (\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x})^2\right] \\
\geq& 0
\end{aligned}$$

by Cauchy-Schwarz inequality. Moreover, the final equality holds only when $\boldsymbol{v} \parallel \boldsymbol{W}^{\top}\boldsymbol{x}$ and $\boldsymbol{v}^{\top}\boldsymbol{W}^{\top}\boldsymbol{x} = \|\boldsymbol{v}\|\|\boldsymbol{W}\|\|\boldsymbol{x}\|$, which implies $\boldsymbol{W} \parallel \boldsymbol{x}\boldsymbol{v}^{T}$. Hence, LaSalle principle yields the one-rank collapse

$$\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|} \to \frac{\boldsymbol{x}\boldsymbol{v}^{\top}}{\|\boldsymbol{x}\|\|\boldsymbol{v}\|}. \tag{40}$$

Now, it suffices to show that $\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$, $\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}$ actually have limits. To do this, we consider the quantity

$$\boldsymbol{a}(t) = \int_0^t \boldsymbol{v}(s)ds. \tag{41}$$

Then we have

$$\boldsymbol{W}(t) = \boldsymbol{W}(0) + \boldsymbol{x}\boldsymbol{a}(t)^\top \to \dot{\boldsymbol{v}} = \boldsymbol{W}(0)^\top \boldsymbol{x} + \|\boldsymbol{x}\|^2 \boldsymbol{a}. \tag{42}$$

Since $\dot{\boldsymbol{a}} = \boldsymbol{v}$, we have the following ODE about $\boldsymbol{a}$:

$$\ddot{\boldsymbol{a}} - \|\boldsymbol{x}\|^2 \boldsymbol{a} = \boldsymbol{W}(0)^\top \boldsymbol{x}, \quad \boldsymbol{a}(0) = 0, \quad \dot{\boldsymbol{a}}(0) = \boldsymbol{v}(0). \tag{43}$$

The general solution is

$$\boldsymbol{a}(t) = -\frac{\boldsymbol{W}(0)^\top \boldsymbol{x}}{\|\boldsymbol{x}\|^2} + e^{\|\boldsymbol{x}\|t}\boldsymbol{u}_+ + e^{-\|\boldsymbol{x}\|t}\boldsymbol{u}_- \tag{44}$$

where

$$\boldsymbol{u}_+ = \frac{1}{2}\left(\frac{\boldsymbol{W}(0)^\top \boldsymbol{x}}{\|\boldsymbol{x}\|^2} + \frac{\boldsymbol{v}(0)}{\|\boldsymbol{x}\|}\right), \quad \boldsymbol{u}_+ = \frac{1}{2}\left(\frac{\boldsymbol{W}(0)^\top \boldsymbol{x}}{\|\boldsymbol{x}\|^2} - \frac{\boldsymbol{v}(0)}{\|\boldsymbol{x}\|}\right) \tag{45}$$

and hence

$$\boldsymbol{v}(t) = \|\boldsymbol{x}\|e^{\|\boldsymbol{x}\|t}\boldsymbol{u}_+ - \|\boldsymbol{x}\|e^{-\|\boldsymbol{x}\|t}\boldsymbol{u}_- \tag{46}$$

This implies that $\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$ has limit $\frac{\boldsymbol{u}_+}{\|\boldsymbol{u}_+\|}$, and the one-rank collapse also suggests that $\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}$ has limit too. Finally, Theorem 4 is the direct consequence of this observation.

■

D.4. The Effect of Weight Normalization in Asymptotic Behavior

In this subsection, we show that the directional collapse described in Theorem 4 can also occur under weight normalization. Assume that we decompose each v_i and $\boldsymbol{w}_i$ by magnitude and direction as Salimans and Kingma [12] does, but we fix their magnitudes. Then our loss function become

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = -\sum_{j=1}^{h} \frac{v_j}{|v_j|}\frac{\boldsymbol{w}_j^\top}{\|\boldsymbol{w}_j\|}\boldsymbol{x}_a(\boldsymbol{w}_j) \tag{47}$$

This function is 0-homogeneous, so $|v_j|$, $\|\boldsymbol{w}_j\|$ actually does not change by Euler's homogeneous function theorem. Moreover, by calculation,

$$\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} = \frac{1}{|\boldsymbol{v}_j|\|\boldsymbol{w}_j\|} \times \operatorname{sgn}(v_j)\left(\boldsymbol{I} - \frac{\boldsymbol{w}_j\boldsymbol{w}_j^\top}{\|\boldsymbol{w}_j\|^2}\right)\boldsymbol{x}_a(\boldsymbol{w}_j) \tag{48}$$

under a balanced initialization, which is a constant multiple of that of unnormalized GF. Hence, we can directly use the result of Section D.2 to prove neural alignment.

After a neural alignment, we will consider the GF dynamics of

$$\hat{\mathcal{L}}(\boldsymbol{v}, \boldsymbol{W}) = -\frac{\boldsymbol{v}^\top \boldsymbol{W}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|}, \tag{49}$$

where the subscript $\pm$ is omitted. In this case, $\|\boldsymbol{v}\|, \|\boldsymbol{W}\|$ is constant again, and

$$\frac{d}{dt}\boldsymbol{v} = \frac{1}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|}\left(\boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|^2}\right)\boldsymbol{W}^\top \boldsymbol{x} = \frac{1}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|}\boldsymbol{W}^\top \boldsymbol{x} - \frac{\boldsymbol{v}^\top \boldsymbol{W}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|^3\|\boldsymbol{W}\|}\boldsymbol{v}, \tag{50}$$

$$\frac{d}{dt}\boldsymbol{W} = \frac{1}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|}\boldsymbol{x}\boldsymbol{v}^\top - \frac{\boldsymbol{v}^\top \boldsymbol{W}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|\|\boldsymbol{W}\|^3}\boldsymbol{W}. \tag{51}$$

Then by a similar way to Section D.3, one can prove that $c = \boldsymbol{v}^\top \boldsymbol{W}^\top \boldsymbol{x}$ is an increasing function and $dc/dt = 0$ only when $\boldsymbol{W} \parallel \boldsymbol{x}\boldsymbol{v}^\top$. Then LaSalle principle yields the one-rank collapse :

$$\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|} \to \frac{\boldsymbol{x}\boldsymbol{v}^\top}{\|\boldsymbol{x}\|\|\boldsymbol{v}\|} \tag{52}$$

In this case we don't have closed form solution of $\boldsymbol{v}$ and $\boldsymbol{W}$. However, since $\hat{\mathcal{L}}$ is semialgebraic in $\{\boldsymbol{v}, \boldsymbol{W} \mid \boldsymbol{v} \neq \boldsymbol{0}, \boldsymbol{W} \neq \boldsymbol{0}\}$, the trajectory is finite, and hence $\boldsymbol{v}$, $\boldsymbol{W}$ actually converges(Bolte et al. [1]) to a single point.