Journal of Machine Learning Research 26 (2025) 1-83 Submitted 5/25; Revised 10/25; Published 11/25

Towards Understanding Gradient Flow Dynamics of Homogeneous Neural Networks Beyond the Origin

Akshay Kumar KUMAR511@UMN.EDU
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Jarvis Haupt JDHAUPT@UMN.EDU
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Mahdi Soltanolkotabi

Abstract

Recent works exploring the training dynamics of homogeneous neural network weights under gradient flow with small initialization have established that in the early stages of training, the weights remain small and near the origin, but converge in direction. Building on this, the current paper studies the gradient flow dynamics of homogeneous neural networks with locally Lipschitz gradients, *after* they escape the origin. Insights gained from this analysis are used to characterize the first saddle point encountered by gradient flow after escaping the origin. Also, it is shown that for homogeneous *feed-forward* neural networks, under certain conditions, the sparsity structure emerging among the weights before the escape is preserved after escaping the origin and until reaching the next saddle point.

Keywords: deep learning, implicit regularization, gradient flow, homogeneous neural networks, training dynamics

1. Introduction

Modern deep neural networks have a surprisingly good generalization behavior when trained via gradient-based methods, even when having sufficient capacity to overfit the training set. A widespread belief is that the training algorithm induces *implicit regularization*, which leads to solutions with favorable generalization performance (Soudry et al., 2018). This has motivated several works to study the training dynamics of neural networks in variety of settings including, for example, the *Neural Tangent Kernel* (NTK)/large initialization regime (Jacot et al., 2018; Chizat et al., 2019; Arora et al., 2019b), late phase of training homogeneous networks with classification loss (Lyu and Li, 2020; Ji and Telgarsky, 2020), and linear and non-linear neural networks (Arora et al., 2019a; Ji and Telgarsky, 2019; Timor et al., 2023; Chizat and Bach, 2020; Jacot, 2023; Abbe et al., 2022).

This paper studies the training dynamics of neural networks in the small initialization regime—an important yet not fully understood area. Understanding this regime is particularly important because, for small initialization, neural networks trained via gradient descent operate in the *feature learning* regime, allowing them to learn underlying features

present in the data (Yang and Hu, 2021; Geiger et al., 2020; Mei et al., 2019). Also, smaller initialization has been observed to lead to better generalization in various tasks (Chizat et al., 2019; Geiger et al., 2020). However, the highly non-linear evolution of the weights in this regime poses significant challenges towards understanding the training dynamics.

Recent studies have focused on the early stages of training homogeneous neural networks via gradient flow with small initialization (Maennel et al., 2018; Kumar and Haupt, 2024, 2025; Luo et al., 2021; Atanasov et al., 2022; Boursier and Flammarion, 2025). For such networks, the origin is a critical point, causing the weights to remain near the origin for some time after training begins, provided the initialization is "small". However, interesting structure among the weights begins to emerge even while the weights remain near the origin. In Kumar and Haupt (2024, 2025), it is shown that for sufficiently small initialization, the weights remain small and near the origin for a sufficiently long time, during which they converge in direction towards a Karush-Kuhn-Tucker (KKT) point of a so-called *Neural Correlation Function* (NCF), a phenomenon referred to as early directional convergence. Also, for feed-forward homogeneous neural networks, weights were observed to converge to a KKT point that exhibits sparsity (Kumar and Haupt, 2025; Atanasov et al., 2022; Zhou et al., 2022), and these KKT points were characterized in Kumar and Haupt (2025).

Compared to Kumar and Haupt (2024, 2025), which focus on the gradient flow dynamics of homogeneous neural networks near the origin, this work studies the dynamics of gradient flow after escaping the origin. To understand our contributions, consider the experiment shown in Figure 1, where we train a two-layer neural network with square activation function (a 3-homogeneous network) using gradient descent; more details about the experiment are in the figure caption. The initial weights are small and random, as depicted in Figure 1a. The evolution of the training loss in Figure 1b resembles a piece-wise constant function, alternating between periods of stagnation and sharp decreases. Similar behavior has been observed in previous works for other neural networks (see Section 1.1 for details). In Figure 1c, we plot the weights at iteration i_1 (marked in Figure 1b), just before escaping the origin. The weights remain small, but are larger than the initial weights, and more importantly, they are sparse. This aligns with observations made in Kumar and Haupt (2025), that in the early stages of training, weights tend to converge in direction towards a KKT point of the NCF that exhibits sparsity. Next, we make two key observations about the dynamics after escaping the origin:

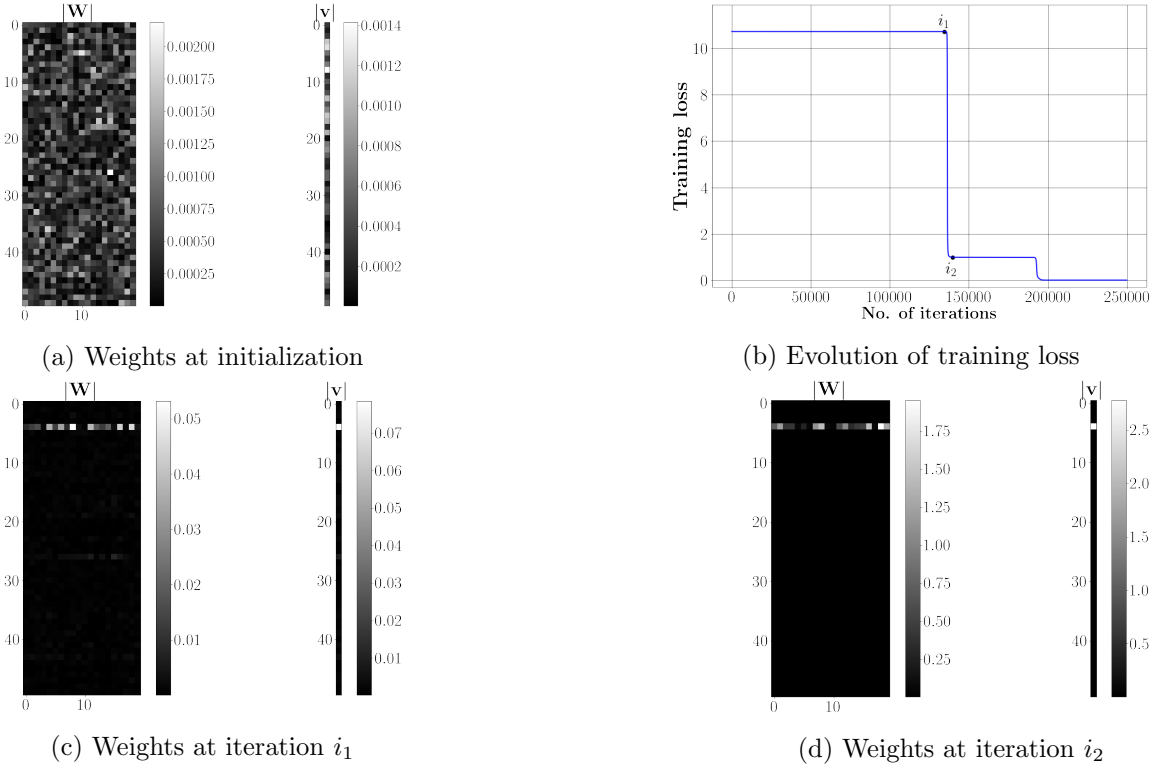
- After gradient descent escapes the origin, the loss rapidly decreases and soon becomes stagnant again, indicating the trajectory of gradient descent is near a saddle point.
- The sparsity structure of the weights is preserved even after escaping the origin. This is evident when comparing Figure 1c with Figure 1d, where the latter depicts the weights at iteration i_2, immediately after reaching the next saddle point.

Our contributions. This paper attempts to explain the above observations for homogeneous neural networks. Our main contributions can be summarized as follows:

- **Gradient Flow Dynamics Post-Escape:** In Theorem 4 and Theorem 10, we describe the gradient flow dynamics of homogeneous neural networks that have locally Lipschitz gradients and order of homogeneity at least two, after escaping the origin. Our results show that for sufficiently small initialization, the gradient flow escapes

along the same path as the gradient flow with small initial weights and initial direction along a KKT point of the NCF. Subsequent corollaries provide a characterization of the saddle point encountered by gradient flow following its escape from the origin.

- **Post-Escape Sparsity in Feed-Forward Networks:** In Section 4, we show that for homogeneous *feed-forward* neural networks, the sparsity structure that emerges among the weights before escaping the origin is preserved post-escape under certain specified conditions. This is achieved by showing that, for KKT points of the NCF, the weights with zero magnitude form a *zero-preserving subset*—a subset of weights that remain zero throughout training under gradient flow, provided they are initialized at zero. This insight is combined with Theorem 4 and Theorem 10 to prove our result.



(a) Weights at initialization

(b) Evolution of training loss

(c) Weights at iteration i_1

(d) Weights at iteration i_2

Figure 1: We train a two-layer neural network with output $\mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x})$, where $\sigma(x) = x^2$, and trainable weights $\mathbf{v} \in \mathbb{R}^{50}, \mathbf{W} \in \mathbb{R}^{50\times 20}$. The training set has 100 points sampled uniformly from the unit sphere in $\mathbb{R}^{20}$. We minimize the square loss with respect to the output of a smaller two-layer neural network with two neurons and square activation. We train using gradient descent with small initial weights, as depicted in panel (a). Panel (b) shows the evolution of loss with iterations. Panels (c) and (d) depict the absolute value of weights at iteration i_1 and i_2 (marked in panel (b)), approximately just before escaping the origin and immediately after reaching the next saddle point, respectively (the gap between them is 5000 iterations). Panels (c) and (d) show that the sparsity structure emerging among the weights before escaping the origin is preserved until reaching the next saddle point.

Note that we assume the neural network to have locally Lipschitz gradients, which excludes ReLU networks. However, our experimental results suggest that some of our findings may extend to ReLU networks as well. Also, our results only capture a segment of the gradient flow dynamics beyond the origin, specifically up to the first saddle point encountered by gradient flow after escaping the origin. More details about this are provided in later sections.

1.1 Related Works

Several works have investigated the training dynamics of neural networks in the small initialization regime. One of the earliest works examined diagonal linear networks, showing that small initialization leads gradient flow to converge towards minimum ℓ_1-norm solutions (Woodworth et al., 2020; Vaskevicius et al., 2019). For linear neural networks, it has been observed that gradient descent with small initialization tends to converge toward low-rank solutions (Gunasekar et al., 2017; Arora et al., 2019a), with rigorous results in some cases (Stöger and Soltanolkotabi, 2021; Jin et al., 2023; Chou et al., 2024; Lawrence et al., 2022). A similar sparsity-inducing effect of small initialization has been observed for non-linear neural networks as well (Chizat et al., 2019), but theoretical results have mostly been established for two-layer neural networks trained on simple data sets (Lyu et al., 2021; Boursier et al., 2022; Wang and Ma, 2023). Recent works on two-layer networks have explored more challenging scenarios (Damian et al., 2022; Abbe et al., 2022; Mousavi-Hosseini et al., 2023), but they often make other assumptions about the training algorithm, such as layer-wise training, the use of explicit regularization like weight decay, etc., whereas we study the dynamics of gradient flow with respect to all the weights and do not add any explicit regularization in the training loss. It is also worth noting that most of the aforementioned works examine the *entire* training process; in contrast, our work describes a segment of the gradient flow dynamics beyond the origin, however, our results hold for a wider class of neural networks.

Another line of work has identified the so-called saddle-to-saddle dynamics in the trajectory of gradient descent when the initialization is small (Jacot et al., 2021; Li et al., 2021). These works have observed that during training, the trajectory of gradient descent passes through a sequence of saddle points. Moreover, the loss curve almost appears like a piece-wise constant function, alternating between being stagnant and decreasing sharply; see Figure 1b for an example. Other works refer to this phenomenon as *incremental learning* (Gidel et al., 2019; Gissin et al., 2020; Razin et al., 2022; Slutzky et al., 2025), since the function learned by the neural network gradually increases in complexity as it moves from one saddle to another. So far, this kind of saddle-to-saddle dynamics has been rigorously established for diagonal linear neural networks (Pesme and Flammarion, 2023; Abbe et al., 2024) and two-layer linear and non-linear neural networks trained with various gradient-based methods under data-related assumptions (Jin et al., 2023; Boursier et al., 2022; Wang and Ma, 2023; Berthier et al., 2024; Abbe et al., 2023), and is conjectured to be true for a wider class of neural networks. Our work, which describes the first saddle point encountered by gradient flow after escaping the origin, can be seen as a step towards establishing it.

Lastly, we highlight the work by Li et al. (2021), which studies the gradient flow dynamics of linear neural networks under small initialization after escaping the origin. While our paper studies a broader class of neural networks, their proof technique has inspired our approach.

1.2 Organization

The paper is organized as follows: Section 2 outlines the problem setup and reviews previous works on early directional convergence. Section 3 present our findings on the gradient flow dynamics after escaping the origin for homogeneous neural networks, with detailed proofs provided in the Appendix. In Section 4, we discuss the implications of these results on the gradient flow dynamics of feed-forward homogeneous neural networks. Finally, Section 5 provides concluding remarks, which is followed by the Appendix.

2. Preliminaries and Problem Setup

The set of natural numbers is denoted by $\mathbb{N}$, and for any $L \in \mathbb{N}$, we let $[L] \coloneqq \{1, 2, \cdots, L\}$. We use $\|\cdot\|_2$ to denote the ℓ_2-norm for a vector, while $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius and spectral norm for a matrix, respectively. The d-dimensional unit-norm sphere is denoted by $\mathbb{S}^{d-1}$. For a non-zero vector $\mathbf{z}$, $\mathbf{z}^\perp \coloneqq \{\mathbf{b} : \mathbf{b}^\top \mathbf{z} = 0, \mathbf{b} \in \mathbb{S}^{d-1}\}$, that is, set of unit-norm vectors orthogonal to $\mathbf{z}$. The ith entry of a vector $\mathbf{z}$ is denoted by z_i. A KKT point of an optimization problem is called a non-negative (positive, zero) KKT point if the objective value at the KKT point is non-negative (positive, zero).

First- and second-order KKT points. We state the first- and second-order KKT conditions for a maximization problem on unit sphere, which can be easily proved. Consider the following optimization problem

$$\max_{\mathbf{w}} f(\mathbf{w}), \text{ s.t. } \|\mathbf{w}\|_2^2 = 1.$$

Suppose $f(\mathbf{w})$ is twice-continuously differentiable near $\mathbf{w}_*$, then

- $\mathbf{w}_*$ is a first-order KKT point if there exists a λ_* such that $\nabla f(\mathbf{w}_*) - 2\lambda_* \mathbf{w}_* = \mathbf{0}$.
- For $\Delta > 0$, we define a first-order KKT point $\mathbf{w}_*$ to be a $\Delta-$second-order KKT point if $2\lambda_* - \mathbf{b}^\top \nabla^2 f(\mathbf{w}_*)\mathbf{b} \geq \Delta$, for all $\mathbf{b} \in \mathbf{w}_*^\perp$.

Homogeneous neural netwoks. For a neural network $\mathcal{H}$, $\mathcal{H}(\mathbf{x}; \mathbf{w})$ denotes its output, where $\mathbf{x} \in \mathbb{R}^d$ is the input and $\mathbf{w} \in \mathbb{R}^k$ is a vector containing all the weights. A neural network $\mathcal{H}$ is referred to as *L-positively homogeneous* if

$$\mathcal{H}(\mathbf{x}; c\mathbf{w}) = c^L \mathcal{H}(\mathbf{x}; \mathbf{w}), \text{ for all } c \geq 0 \text{ and } \mathbf{w} \in \mathbb{R}^k.$$

Assumption 1 *We make the following assumptions on the neural networks considered in this paper: (i) For any fixed $\mathbf{x}$, $\mathcal{H}(\mathbf{x}; \mathbf{w})$ is locally Lipschitz in $\mathbf{w}$ and is definable in some o-minimal structure. (ii) The neural network $\mathcal{H}$ is L-positively homogeneous, for some $L \geq 2$. (iii) The gradient of $\mathcal{H}(\mathbf{x}; \mathbf{w})$ with respect to $\mathbf{w}$, $\nabla \mathcal{H}(\mathbf{x}; \mathbf{w})$, is locally Lipschitz in $\mathbf{w}$.*

The first two conditions are satisfied by deep feed-forward neural networks with homogeneous activation functions like ReLU and polynomial ReLU ($\max(0, x)^p, p \geq 1$). We note that definability in some o-minimal structure is a mild technical assumption and is satisfied by all modern deep neural networks (Ji and Telgarsky, 2020). Assuming locally Lipschitz

gradient is crucial for proving our results rigorously, and while it rules out deep ReLU neural networks, it does include deep linear networks and feed-forward neural networks with polynomial ReLU activation functions ($\max(0,x)^p, p \geq 2$).

Training setup. Let $\{\mathbf{x}_i, y_i\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ be the training data, and define $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, $\mathbf{y} = [y_1, \cdots, y_n]^\top \in \mathbb{R}^n$. Let $\mathcal{H}(\mathbf{X}; \mathbf{w}) = [\mathcal{H}(\mathbf{x}_1; \mathbf{w}), \cdots, \mathcal{H}(\mathbf{x}_n; \mathbf{w})] \in \mathbb{R}^n$ be the vector containing outputs of the neural network, and $\mathcal{J}(\mathbf{X}; \mathbf{w})$ denotes the Jacobian of $\mathcal{H}(\mathbf{X}; \mathbf{w})$ with respect to $\mathbf{w}$. For training, we minimize the following objective:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \ell(\mathcal{H}(\mathbf{x}_i; \mathbf{w}), y_i) = \ell(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}), \tag{1}$$

where $\ell(\cdot,\cdot)$ is the loss function, and $\ell(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n \ell(p_i, q_i)$, for any two vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$. We use $\ell'(\cdot,\cdot)$ to denote the derivative of $\ell(\cdot,\cdot)$ with respect to the first variable, and define $\ell'(\mathbf{p}, \mathbf{q}) = [\ell'(p_1, q_1), \cdots, \ell'(p_n, q_n)]^\top \in \mathbb{R}^n$. Also, $\ell''(\cdot,\cdot)$ denotes the second-order derivative of $\ell(\cdot,\cdot)$ with respect to the first variable.

Assumption 2 *We make the following assumptions on the loss function:*

- *Smoothness: For some $K > 0$, $|\ell''(p,q)| \leq K$, for all $p, q \in \mathbb{R}$.*
- *Convexity: $(\ell'(\mathbf{p}, \mathbf{y}) - \ell'(\mathbf{q}, \mathbf{y}))^\top (\mathbf{p} - \mathbf{q}) \geq 0$, for all $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$.*

The above assumption is satisfied by common loss functions such as square and logistic loss. We minimize the optimization problem in eq. (1) using gradient flow:

$$\dot{\mathbf{w}} = -\nabla \mathcal{L}(\mathbf{w}) = -\mathcal{J}(\mathbf{X}; \mathbf{w})^\top \ell'(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}), \tag{2}$$

and use $\boldsymbol{\psi}(t, \mathbf{w}(0))$ to denote the solution of above differential equation, where $\mathbf{w}(0)$ is the initialization. We aim to study the evolution of $\boldsymbol{\psi}(t, \mathbf{w}(0))$ with time for small initialization.

Early directional convergence. We next briefly discuss the results of Kumar and Haupt (2024, 2025) which study the phenomenon of early directional convergence.

Lemma 1 *The origin is a critical point of the optimization problem in eq.* (1).

Proof If $\mathcal{H}$ is $L-$homogeneous, then $\mathcal{J}(\mathbf{X}; \mathbf{w})$ is $(L-1)$-homogeneous. Since $L \geq 2$, we get $\mathcal{J}(\mathbf{X}; \mathbf{0}) = \mathbf{0}$, which implies $\nabla \mathcal{L}(\mathbf{0}) = \mathbf{0}$. ■

Therefore, if the gradient flow in eq. (2) is initialized near the origin, then it is expected to remain near the origin for some time, before escaping from it. In Kumar and Haupt (2024, 2025), authors study the dynamics of gradient flow with small initialization in the early stages of training and while the gradient flow remains near the origin.

We start with some basic concepts introduced in Kumar and Haupt (2024, 2025). Let $\widetilde{\mathbf{y}} := -\ell'(0, \mathbf{y})$, then the Neural Correlation Function (NCF) is defined as

$$\mathcal{N}(\mathbf{u}) = \widetilde{\mathbf{y}}^\top \mathcal{H}(\mathbf{X}; \mathbf{u}), \tag{3}$$

and the constrained NCF refers to the following optimization problem

$$\max_{\mathbf{u}} \mathcal{N}(\mathbf{u}), \text{ s.t. } \|\mathbf{u}\|_2^2 = 1. \tag{4}$$

Next, consider the (positive) gradient flow of the NCF:

$$\dot{\mathbf{u}} = \nabla \mathcal{N}(\mathbf{u}) = \mathcal{J}(\mathbf{X}; \mathbf{u})^\top \widetilde{\mathbf{y}}. \tag{5}$$

We use $\boldsymbol{\phi}(t, \mathbf{u}(0))$ to denote the solution of above differential equation, where $\mathbf{u}(0)$ is the initialization. Then, as shown in Kumar and Haupt (2024, 2025), for any unit-norm vector $\mathbf{u}_0$,[1] $\boldsymbol{\phi}(t, \mathbf{u}_0)$ satisfies the following condition: (i) either $\boldsymbol{\phi}(t, \mathbf{u}_0)$ converges to the origin, or (ii) $\boldsymbol{\phi}(t, \mathbf{u}_0)/\|\boldsymbol{\phi}(t, \mathbf{u}_0)\|_2$ converges to a non-negative KKT point of the constrained NCF. Also, if the order of homogeneity is strictly greater than two, then $\boldsymbol{\phi}(t, \mathbf{u}_0)/\|\boldsymbol{\phi}(t, \mathbf{u}_0)\|_2$ may converge in finite time.

We next define the notion of stable set for a non-negative KKT point.

Definition 2 *The stable set $\mathcal{S}(\mathbf{u}_*)$ of a non-negative KKT point $\mathbf{u}_*$ of eq. (4) is the set of all unit-norm initializations such that gradient flow in eq. (5) converges in direction to $\mathbf{u}_*$:*

$$\mathcal{S}(\mathbf{u}_*) \coloneqq \left\{ \mathbf{u}_0 \in \mathbb{S}^{k-1} : \frac{\boldsymbol{\phi}(t, \mathbf{u}_0)}{\|\boldsymbol{\phi}(t, \mathbf{u}_0)\|_2} \to \mathbf{u}_* \right\}$$

We now present the main result of Kumar and Haupt (2024, 2025), which describes the dynamics of gradient flow with small initialization in the early stages of training.

Lemma 3 *Suppose $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$, where $\mathbf{w}_*$ is a non-negative KKT point of eq. (4). For any sufficiently small $\epsilon > 0$, there exists T and $\overline{\delta}$ such that the following holds: for any $\delta \in (0, \overline{\delta})$ we have*

$$\|\boldsymbol{\psi}(t, \delta \mathbf{w}_0)\|_2 = O(\delta), \text{ for all } t \in [0, T/\delta^{L-2}], \text{ and } \frac{\boldsymbol{\psi}(T/\delta^{L-2}, \delta \mathbf{w}_0)^\top \mathbf{w}_*}{\|\boldsymbol{\psi}(T/\delta^{L-2}, \delta \mathbf{w}_0)\|_2} = 1 - O(\epsilon).$$

In the above lemma, the initialization is $\delta \mathbf{w}_0$, where $\mathbf{w}_0$ is a vector and $\delta > 0$ is a scalar that controls the scale of initialization. It is shown that, for all sufficiently small initialization, the weights remain small for sufficiently long time and converge in direction towards a non-negative KKT point of the constrained NCF. Also, as discussed earlier, if $\mathbf{w}_0$ does not belong to the stable set of a non-negative KKT point, then $\boldsymbol{\phi}(t, \mathbf{w}_0)$ must converge to the origin. In this case, rather than directional convergence, the weights approximately become zero; see Kumar and Haupt (2024, 2025) for more details.

3. Gradient Flow Dynamics Beyond the Origin

In this section, we describe the gradient flow dynamics of homogeneous neural networks after escaping the origin.

1. The unit-norm assumption is for simplicity, the result will hold for other vectors as well.

3.1 Two-Homogeneous Neural Networks

We begin by considering two-homogeneous neural networks and the following theorem describes the gradient flow dynamics of such neural networks after escaping the origin.

Theorem 4 *Suppose $\mathcal{H}$ is 2-homogeneous, and let $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$, where $\mathbf{w}_*$ is a $\Delta-$second-order positive KKT point of eq. (4), for some $\Delta > 0$. Then, for any fixed $\widetilde{T} \in (-\infty, \infty)$, there exists a $\widetilde{C} > 0$ such that for all sufficiently small $\delta > 0$,*

$$\left\| \boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) - \mathbf{p}(t) \right\|_2 \leq \widetilde{C}\delta^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}, \text{ for all } t \in [-\widetilde{T}, \widetilde{T}], \quad (6)$$

where $T_1 > 0$ is a constant, $\widetilde{\delta} \in (A_2\delta - A_1\delta^3, A_2\delta + A_1\delta^3)$, for some positive constants A_1, A_2, and $b_\delta \in [\kappa_1, \kappa_2]$, for some $\kappa_2 \geq \kappa_1 > 0$, depends on δ. Moreover,

$$\mathbf{p}(t) \coloneqq \lim_{\delta\to 0} \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right),$$

which exists for all $t \in (-\infty, \infty)$, and $\mathbf{p}(t) = \boldsymbol{\psi}(t, \mathbf{p}(0))$, where $\mathbf{p}(0) = \lim_{\delta\to 0} \boldsymbol{\psi}\left(\frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)$ and $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, for some $\eta > 0$.

We begin by explaining the motivation behind defining $\mathbf{p}(t)$. Recall that $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ denotes the solution of the gradient flow of the training loss with initialization $\delta\mathbf{w}_*$. For small δ, $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ will remain small and near the origin for some time after the training begins. Now, loosely speaking, it turns out that $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ would have escaped from the origin after $(\ln(1/\delta)/2\mathcal{N}(\mathbf{w}_*) + O(1))$ time has elapsed, where recall that $\mathcal{N}(\cdot)$ is the NCF. Since our interest is in the dynamics of gradient flow after it escapes the origin, this time is added in $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ while defining $\mathbf{p}(t)$. Taking $\delta \to 0$, gives us the limiting solution of $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ after it escapes from the origin. Therefore, $\mathbf{p}(t)$ can be viewed as the approximate path $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ takes after it escapes from the origin, for all sufficiently small δ.

We also note two key properties of $\mathbf{p}(t)$. First, $\mathbf{p}(t) = \boldsymbol{\psi}(t, \mathbf{p}(0))$, which implies $\mathbf{p}(t)$ is a solution of the gradient flow of the training loss with initialization $\mathbf{p}(0)$. In fact, this follows directly from the definition of $\mathbf{p}(t)$ and continuity of $\psi(\cdot, \cdot)$ with respect to the initialization:

$$\begin{aligned}\mathbf{p}(t) = \lim_{\delta\to 0} \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right) &= \lim_{\delta\to 0} \boldsymbol{\psi}\left(t, \boldsymbol{\psi}\left(\frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right) \\ &= \boldsymbol{\psi}\left(t, \lim_{\delta\to 0} \boldsymbol{\psi}\left(\frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right) = \boldsymbol{\psi}(t, \mathbf{p}(0)).\end{aligned}$$

Another key property is that $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, for some $\eta > 0$. This implies that $\|\mathbf{p}(0)\|_2 \neq 0$, and thus, $\mathbf{p}(0)$ is away from the origin. Now, since the origin is a critical point of the training loss, $\|\mathbf{p}(t)\|_2 \neq 0$, for all finite $t \leq 0$.[2] Also, since loss always decreases under gradient flow, $\mathcal{L}(\mathbf{p}(t)) \leq \mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, for all $t \geq 0$, which implies $\mathbf{p}(t)$ remains away from the origin, for all $t \geq 0$, including at infinity. Therefore, $\mathbf{p}(t)$ has truly escaped from

2. If $\|\mathbf{p}(t_0)\|_2 = 0$, for some finite $t_0 < 0$, then since the origin is a critical point of the training loss, $\|\mathbf{p}(t)\|_2 = 0$, for all $t \geq t_0$, which contradicts $\|\mathbf{p}(0)\|_2 \neq 0$.

the origin. We highlight this property because, if $\mathbf{w}_*$ were a zero KKT point instead of a positive KKT point, that is, $\mathcal{N}(\mathbf{w}_*) = 0$, then $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ may never escape from the origin (see Lemma 33 for an example). Assuming $\mathbf{w}_*$ to be a positive KKT point ensures that $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ eventually escapes from the origin.

We next discuss eq. (6) and its consequences. For small δ, $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ will remain small and near the origin for some time after the training begins. From eq. (6), we observe that, for all sufficiently small δ, $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ would have definitely escaped from the origin after $T_1 + (\ln(1/b_\delta) + \ln(1/\widetilde{\delta}))/2\mathcal{N}(\mathbf{w}_*)$ time has elapsed, since $\mathbf{p}(0)$ is away from the origin and the RHS of eq. (6) is small, for small δ. More importantly, eq. (6) implies that $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ will remain close to $\mathbf{p}(t)$ after escaping the origin for arbitrarily long time, provided the initialization is sufficiently small. Using the definition of $\mathbf{p}(t)$, it can be further said that the trajectory of $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ and $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ will approximately be same after they escape from the origin, for arbitrarily long time, if the initialization is sufficiently small. Therefore, the behavior of the gradient flow after escaping from the origin is primarily determined by the KKT point whose stable set contains the initial direction of the weights.

Note that $\mathbf{p}(t)$ is a solution of the gradient flow of the training loss. Suppose the trajectory of $\mathbf{p}(t)$ is bounded for all $t \geq 0$, and it converge to a saddle point, or more generally, a stationary point of the training loss. Then, in the following corollary, we show that $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ gets close to that saddle point at some time, for all sufficiently small δ.

Corollary 5 *Consider the setting of Theorem 4. Suppose $\mathbf{p}(t)$ is bounded for all $t \geq 0$, and let $\mathbf{p}^* = \lim_{t\to\infty} \mathbf{p}(t)$, where $\nabla\mathcal{L}(\mathbf{p}^*) = 0$ Then, for any sufficiently small $\epsilon > 0$, there exists a time T_ϵ such that for all sufficiently small $\delta > 0$,*

$$\|\boldsymbol{\psi}(T_\delta, \delta\mathbf{w}_0) - \mathbf{p}^*\|_2 \leq \epsilon \text{ and } \|\nabla\mathcal{L}(\boldsymbol{\psi}(T_\delta, \delta\mathbf{w}_0))\|_2 \leq \epsilon,$$

where $T_\delta \coloneqq T_\epsilon + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}$.

The above corollary implies that for all sufficiently small δ, after escaping from the origin, gradient flow get close to the saddle point to which $\mathbf{p}(t)$ converges.

The above corollary assumes $\mathbf{p}(t)$ is bounded and converges to a finite saddle point. This is a reasonable assumption when square loss is used for training. For loss functions such as logistic loss, the trajectory of gradient flow may not be bounded and saddle points of the training loss could be at infinity. The next corollary considers such cases.

Corollary 6 *Consider the setting of Theorem 4. Suppose $\lim_{t\to\infty} \mathbf{p}(t)/\|\mathbf{p}(t)\|_2 = \mathbf{p}^*$ and $\lim_{t\to\infty} \|\mathbf{p}(t)\| = \infty$, such that $\lim_{t\to\infty} \nabla\mathcal{L}(\mathbf{p}(t)) = \mathbf{0}$ and $\lim_{\alpha\to\infty} \nabla\mathcal{L}(\alpha\mathbf{p}^*) = 0$. Then, for any sufficiently small $\epsilon > 0$, there exists a time T_ϵ such that for all sufficiently small $\delta > 0$,*

$$\|\boldsymbol{\psi}(T_\delta, \delta\mathbf{w}_0)\|_2 \geq \frac{1}{2\epsilon}, \frac{\boldsymbol{\psi}(T_\delta, \delta\mathbf{w}_0)^\top \mathbf{p}^*}{\|\boldsymbol{\psi}(T_\delta, \delta\mathbf{w}_0)\|_2} \geq 1 - \epsilon, \text{ and } \|\nabla\mathcal{L}(\boldsymbol{\psi}(T_\delta, \delta\mathbf{w}_0))\|_2 \leq \epsilon,$$

where $T_\delta \coloneqq T_\epsilon + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}$.

In the above corollary, we have assumed that $\mathbf{p}(t)$ "converges" to a saddle point at infinity, in the sense that its norm diverges to infinity, but the direction converges. Under this assumption, $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ gets close to that saddle point at some time, for all sufficiently small

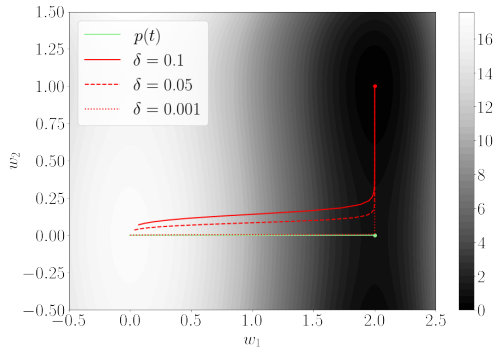


Figure 2: The contour of the loss function in eq. (7) is in the background. The foreground contains evolution of $\boldsymbol{\psi}(t, \delta \mathbf{w}_0)$, for $\delta \in \{0.1, 0.05, 0.001\}$ and $t \in [0, 3]$ (in red), and $\mathbf{p}(t)$, for $t \in [-1, 1]$ (in green). The saddle point at $(2, 0)$ and the global minimum at $(2, 1)$ are marked with green and red dot respectively.

initialization, since its norm becomes large, and it is approximately aligned in direction with the saddle point.

We would like to emphasize that the above corollaries *do not* imply that $\boldsymbol{\psi}(t, \delta \mathbf{w}_0)$ will also converge to the same saddle point to which $\mathbf{p}(t)$ converges. For instance, consider Corollary 5, which shows that $\boldsymbol{\psi}(t, \delta \mathbf{w}_0)$ gets close to the saddle point $\mathbf{p}^*$ at some time. It is possible that $\boldsymbol{\psi}(t, \delta \mathbf{w}_0)$ may eventually escape from this saddle point, in contrast to $\mathbf{p}(t)$ which converges towards $\mathbf{p}^*$. The evolution of gradient flow after escaping this saddle point is not described by the above results, and is an important direction for future research. Thus, in this sense, our work only captures a segment of the gradient flow dynamics beyond the origin, specifically until the first saddle point encountered by gradient flow after escaping the origin. The following two-dimensional example may be helpful in understanding our results better, the complete details of which are in Appendix E.

Example 1 *Suppose* $\mathcal{H}(\mathbf{x}; \mathbf{w}) = w_1^2 x_1 + w_2^2 x_2$, *which is two-homogeneous. The training data has two points* $(1, 0)$ *and* $(0, 1)$, *where the corresponding labels are* 4 *and* 1 *respectively. Assuming square loss, the training loss becomes*

$$\mathcal{L}(\mathbf{w}) = (w_1^2 - 4)^2 + (w_2^2 - 1)^2, \tag{7}$$

and $\mathcal{N}(\mathbf{w}) = 8w_1^2 + 2w_2^2$. *Let* $\mathbf{w}_0 = (1/\sqrt{2}, 1/\sqrt{2})$, *then* $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$, *where* $\mathbf{w}_* = (1, 0)$, *and* $\mathcal{N}(\mathbf{w}_*) = 8$. *For any* $\delta \in (0, 1)$, *we have*

$$\boldsymbol{\psi}(t, \delta \mathbf{w}_0) = \left(\frac{2\delta}{\sqrt{\delta^2 + (8 - \delta^2) e^{-32t}}}, \frac{\delta}{\sqrt{\delta^2 + (2 - \delta^2) e^{-8t}}} \right),$$

$$\boldsymbol{\psi}(t, \delta \mathbf{w}_*) = \left(\frac{2\delta}{\sqrt{\delta^2 + (4 - \delta^2) e^{-32t}}}, 0 \right), \text{ and } \mathbf{p}(t) = \left(\frac{2}{\sqrt{1 + 4e^{-32t}}}, 0 \right).$$

Note that, for any $\gamma \in (0, 1)$,

$$\left\| \boldsymbol{\psi}\left(\frac{(1-\gamma)\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_* \right) \right\|_2 = O(\delta^\gamma), \left\| \boldsymbol{\psi}\left(\frac{(1-\gamma)\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_0 \right) \right\|_2 = O(\delta^\gamma),$$

and for any $\kappa > 0$,

$$\left\| \boldsymbol{\psi}\left(\frac{\ln(\kappa/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right\|_2 = O(1), \left\| \boldsymbol{\psi}\left(\frac{\ln(\kappa/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right)\right\|_2 = O(1),$$

Therefore, after $(O(1) + \ln(1/\delta)/2\mathcal{N}(\mathbf{w}_*))$ *time has elapsed,* $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ *and* $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ *would have escaped from the origin. Next,* $\mathbf{p}(t)$ *and* $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ *converge to different limits since*

$$\lim_{t\to\infty} \mathbf{p}(t) = (2, 0) \text{ and } \lim_{t\to\infty} \boldsymbol{\psi}(t, \delta\mathbf{w}_0) = (2, 1).$$

In Figure 2, we plot the evolution of $\mathbf{p}(t)$ *and* $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$*, for small values of* δ*. For the smallest value of* δ*, observe that* $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ *follows the same path as* $\mathbf{p}(t)$ *all the way until it gets very close to* $(2, 0)$*, which is a saddle point. Then it escapes from it and eventually converges to* $(2, 1)$*. Thus, even though* $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ *gets close to* $(2, 0)$*, it does not converge to it, unlike* $\mathbf{p}(t)$*.*

It is also worth noting that for any $\kappa > 0$,

$$\boldsymbol{\psi}\left(\frac{\ln(\kappa/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) = (O(1), O(\delta^{3/4})), \boldsymbol{\psi}\left(\frac{\ln(\kappa/\delta)}{4} + \frac{\ln(\kappa/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) = (2 - O(\delta^{10}), O(1)).$$

Thus, $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ *escapes from the saddle point at* $(2, 0)$ *after* $(O(1) + \ln(1/\delta)/2\mathcal{N}(\mathbf{w}_*)) + \ln(1/\delta)/4)$ *time has elapsed. This suggests that to analyze the gradient flow dynamics beyond the first saddle point, another* $O(\ln(1/\delta))$ *term needs to be added in the time. This observation may be helpful for future works that attempt to understand the gradient flow dynamics beyond the first saddle point.*

3.1.1 Proof Outline of Theorem 4

We first discuss three important lemmata, which are then used to prove Theorem 4. The first lemma proves the existence of $\mathbf{p}(t)$, for all $t \in (-\infty, \infty)$, by showing $\boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)$ is a Cauchy sequence, and also characterizes $\mathcal{L}(\mathbf{p}(0))$.

Lemma 7 *Consider the setting of Theorem 4, then for any fixed* $t \in (-\infty, \infty)$ *and all sufficiently small* $\delta_2 \geq \delta_1 > 0$*, there exists a* $C > 0$ *such that*

$$\left\| \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta_1)}{2\mathcal{N}(\mathbf{w}_*)}, \delta_1\mathbf{w}_*\right) - \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta_2)}{2\mathcal{N}(\mathbf{w}_*)}, \delta_2\mathbf{w}_*\right)\right\|_2 \leq C\delta_2,$$

implying $\mathbf{p}(t)$ *exists for all* $t \in (-\infty, \infty)$*. Furthermore, let* $\delta_1 \to 0$ *and* $\delta_2 = \delta > 0$*, then*

$$\left\| \mathbf{p}(t) - \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right\|_2 \leq C\delta.$$

Also, $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$*, for some* $\eta > 0$*.*

We next show that $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ gets sufficiently aligned with $\mathbf{w}_*$ while staying near the origin.

Lemma 8 *Consider the setting in Theorem 4, then there exists $T_1, a_1 > 0$ such that for all sufficiently small $\delta > 0$,*

$$\left\| \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) - b_\delta \widetilde{\delta}^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}} \mathbf{w}_* \right\|_2 \leq a_1 \widetilde{\delta}^{\frac{3\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}},$$

where $\widetilde{\delta} \in (A_2\delta - A_1\delta^3, A_2\delta + A_1\delta^3)$, for some positive constants A_1, A_2. Also, $b_\delta \in [\kappa_1, \kappa_2]$, for some $\kappa_2 \geq \kappa_1 > 0$, depends on δ.

If we define $T_{\widetilde{\delta}} \coloneqq T_1 + 4\ln(1/\widetilde{\delta})/(\Delta + 8\mathcal{N}(\mathbf{w}_*))$, then the above lemma implies

$$\left\|\boldsymbol{\psi}\left(T_{\widetilde{\delta}}, \delta\mathbf{w}_0\right)\right\|_2 = O\left(\delta^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}\right), \frac{\boldsymbol{\psi}\left(T_{\widetilde{\delta}}, \delta\mathbf{w}_0\right)^\top \mathbf{w}_*}{\left\|\boldsymbol{\psi}\left(T_{\widetilde{\delta}}, \delta\mathbf{w}_0\right)\right\|_2} = 1 - O\left(\delta^{\frac{2\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}\right),$$

for all sufficiently small $\delta > 0$. Therefore, in the early stages of training, $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ remains small and converges in direction to $\mathbf{w}_*$. Note that, the above equation implies a stronger directional convergence than Lemma 3, since there ϵ is small but fixed (does not depend on δ). However, compared to Lemma 3, here $\mathbf{w}_*$ is assumed to be a *second-order* KKT point.

The following lemma allows us to combine Lemma 8 and Lemma 7 in order to prove Theorem 4. It essentially shows that if initialized near the origin and sufficiently aligned with $\mathbf{w}_*$, then gradient flow remains close to $\mathbf{p}(t)$.

Lemma 9 *Consider the setting in Theorem 4. Suppose there exists a constant $C_1 > 0$ such that for every $\delta > 0$, $\mathbf{a}_\delta$ is a vector that satisfies $\|\mathbf{a}_\delta - \delta\mathbf{w}_*\|_2 \leq C_1\delta^3$. Then, for any fixed $\widetilde{T} \in (-\infty, \infty)$, there exists a constant $C > 0$ such that for all sufficiently small $\delta > 0$,*

$$\left\| \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \mathbf{a}_\delta\right) - \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right) \right\|_2 \leq C\delta, \text{ for all } t \in [-\widetilde{T}, \widetilde{T}].$$

Proof of Theorem 4: From Lemma 8, we know

$$\left\| \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) - b_\delta \widetilde{\delta}^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}} \mathbf{w}_* \right\|_2 \leq a_1 \widetilde{\delta}^{\frac{3\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}},$$

for some $T_1, a_1 > 0$. Define $\bar{\delta} = b_\delta \widetilde{\delta}^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}$, then the above equation implies

$$\left\| \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) - \bar{\delta}\mathbf{w}_* \right\|_2 \leq a_2\bar{\delta}^3,$$

where $a_2 = a_1/\kappa_1^3 \geq a_1/b_\delta^3$. Next, using Lemma 9, for any fixed $\widetilde{T} \in (-\infty, \infty)$, there exists a constant $\widetilde{C}_1 > 0$ such that for all sufficiently small $\delta > 0$,

$$\left\| \boldsymbol{\psi}\left(t + \frac{\ln(1/\bar{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right)\right) - \boldsymbol{\psi}\left(t + \frac{\ln(1/\bar{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \bar{\delta}\mathbf{w}_*\right) \right\|_2 \leq \widetilde{C}_1\bar{\delta},$$

for all $t \in [-\widetilde{T}, \widetilde{T}]$, and from Lemma 7, there exists a $\widetilde{C}_2 > 0$ such that

$$\left\| \mathbf{p}(t) - \boldsymbol{\psi}\left(t + \frac{\ln\left(1/\bar{\delta}\right)}{2\mathcal{N}(\mathbf{w}_*)}, \bar{\delta}\mathbf{w}_*\right)\right\|_2 \leq \widetilde{C}_2\bar{\delta}, \text{ for all } t \in [-\widetilde{T}, \widetilde{T}].$$

Now, since

$$\begin{aligned}
&\boldsymbol{\psi}\left(t + \frac{\ln\left(1/\bar{\delta}\right)}{2\mathcal{N}(\mathbf{w}_*)}, \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right)\right) \\
&= \boldsymbol{\psi}\left(t + \frac{\ln\left(1/\bar{\delta}\right)}{2\mathcal{N}(\mathbf{w}_*)} + T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) \\
&= \boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \left(\frac{\Delta}{2\mathcal{N}(\mathbf{w}_*)} + 4\right)\frac{\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) \\
&= \boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right),
\end{aligned}$$

we have that for all $t \in [-\widetilde{T}, \widetilde{T}]$ and for all sufficiently small $\delta > 0$,

$$\begin{aligned}
&\left\|\boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) - \mathbf{p}(t)\right\|_2 \leq \left\|\boldsymbol{\psi}\left(t + \frac{\ln\left(1/\bar{\delta}\right)}{2\mathcal{N}(\mathbf{w}_*)}, \bar{\delta}\mathbf{w}_*\right) - \mathbf{p}(t)\right\|_2 \\
&+ \left\|\boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_0\right) - \boldsymbol{\psi}\left(t + \frac{\ln\left(1/\bar{\delta}\right)}{2\mathcal{N}(\mathbf{w}_*)}, \bar{\delta}\mathbf{w}_*\right)\right\|_2 \\
&\leq \widetilde{C}_1\bar{\delta} + \widetilde{C}_2\bar{\delta} \leq \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}},
\end{aligned}$$

where $\widetilde{C}$ is a positive constant. The last inequality is true since $\widetilde{\delta} \leq A_2\delta + A_1\delta^3 \leq 2A_2\delta$, for all sufficiently small δ. Thus, the proof is complete. ■

3.2 Deep Homogeneous Neural Networks

We now present our main result describing the dynamics of gradient flow for L-homogeneous neural networks after it escapes from origin, where $L > 2$.

Theorem 10 *Suppose $\mathcal{H}$ is L-homogeneous, where $L > 2$. Let $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$, where $\mathbf{w}_*$ is a $\Delta-$second-order positive KKT point of eq. (4), for some $\Delta > 0$. Then, for any fixed $\widetilde{T} \in (-\infty, \infty)$, there exists a $\widetilde{C} > 0$ such that for all sufficiently small $\delta > 0$,*

$$\begin{aligned}
&\left\|\boldsymbol{\psi}\left(t + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T\right)\widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}, \delta\mathbf{w}_0\right) - \mathbf{p}(t)\right\|_2 \\
&\leq \widetilde{C}\delta^{\frac{\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}, \qquad (8)
\end{aligned}$$

for all $t \in [-\widetilde{T}, \widetilde{T}]$*, where* $T_1 > 0$ *is a constant,* $\widetilde{\delta} \in (A_2\delta - A_1\delta^{L+1}, A_2\delta + A_1\delta^{L+1})$ *for some constants* $A_1, A_2 > 0$*,* $T \geq \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_*)}$*, and* $b_\delta^{L-2} \in \left[\frac{1}{TL(L-2)\mathcal{N}(\mathbf{w}_*)}, 1\right]$ *depends on* δ*. Moreover,*

$$\mathbf{p}(t) := \lim_{\delta \to 0} \boldsymbol{\psi}\left(t + \frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right),$$

which exists for all $t \in (-\infty, \infty)$*, and* $\mathbf{p}(t) = \boldsymbol{\psi}(t, \mathbf{p}(0))$*, where* $\mathbf{p}(0) = \lim_{\delta \to 0} \boldsymbol{\psi}\left(\frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right)$ *and* $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$*, for some* $\eta > 0$*.*

Overall, the result here is similar to two-homogeneous case, except for the fact that it takes longer time to escape from the origin. For small δ, $\boldsymbol{\psi}(t, \delta \mathbf{w}_*)$ would have escaped from the origin after $\frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} + O(1)$ time has elapsed, which is reflected in the definition of $\mathbf{p}(t)$. From eq. (8), we can say that the trajectory of $\boldsymbol{\psi}(t, \delta \mathbf{w}_0)$ and $\boldsymbol{\psi}(t, \delta \mathbf{w}_*)$ will approximately be the same after escaping from the origin, for arbitrarily long time if the initialization is sufficiently small. The proof follows a similar approach to the two-homogeneous case, with key differences arising from the fact that in deep homogeneous networks, the gradient flow of NCF can become unbounded at a finite time. See the proof for more details.

Next, we can use the above theorem to determine the saddle point encountered by gradient flow after escaping from the origin, in a similar way as for two-homogeneous neural networks. We start with the case when $\mathbf{p}(t)$ is bounded.

Corollary 11 *Consider the setting of Theorem 10. Suppose* $\mathbf{p}(t)$ *is bounded for all* $t \geq 0$*, and let* $\mathbf{p}^* = \lim_{t \to \infty} \mathbf{p}(t)$*, where* $\nabla \mathcal{L}(\mathbf{p}^*) = 0$ *Then, for any sufficiently small* $\epsilon > 0$*, there exists a time* T_ϵ *such that for all sufficiently small* $\delta > 0$*,*

$$\|\boldsymbol{\psi}(T_\delta, \delta \mathbf{w}_0) - \mathbf{p}^*\|_2 \leq \epsilon, \|\nabla \mathcal{L}(\boldsymbol{\psi}(T_\delta, \delta \mathbf{w}_0))\|_2 \leq \epsilon,$$

where $T_\delta \coloneqq T_\epsilon + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T\right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}$.

We next consider the case when $\mathbf{p}(t)$ can become unbounded.

Corollary 12 *Consider the setting of Theorem 10. Suppose* $\lim_{t \to \infty} \mathbf{p}(t)/\|\mathbf{p}(t)\|_2 = \mathbf{p}^*$ *and* $\lim_{t \to \infty} \|\mathbf{p}(t)\| = \infty$*, such that* $\lim_{t \to \infty} \nabla \mathcal{L}(\mathbf{p}(t)) = \mathbf{0}$ *and* $\lim_{\alpha \to \infty} \nabla \mathcal{L}(\alpha \mathbf{p}^*) = 0$*. Then, for any sufficiently small* $\epsilon > 0$*, there exists a time* T_ϵ *such that for all sufficiently small* $\delta > 0$*,*

$$\|\boldsymbol{\psi}(T_\delta, \delta \mathbf{w}_0)\|_2 \geq \frac{1}{2\epsilon}, \frac{\boldsymbol{\psi}(T_\delta, \delta \mathbf{w}_0)^\top \mathbf{p}^*}{\|\boldsymbol{\psi}(T_\delta, \delta \mathbf{w}_0)\|_2} \geq 1 - \epsilon, \text{ and } \|\nabla \mathcal{L}(\boldsymbol{\psi}(T_\delta, \delta \mathbf{w}_0))\|_2 \leq \epsilon,$$

where $T_\delta \coloneqq T_\epsilon + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T\right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}$.

Remark 13 *Theorem 4 and Theorem 10 assume* $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$*, where* $\mathbf{w}_*$ *is a second-order positive KKT point of eq.* (4)*. If this is not satisfied, three scenarios arise: (i)* $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$*, where* $\mathbf{w}_*$ *is a first-order positive KKT point, but not second-order, (ii)* $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$*, where* $\mathbf{w}_*$ *is a zero KKT point, or (iii)* $\phi(t, \mathbf{w}_0)$ *converges to the origin. In the last two cases, the gradient flow may not escape the origin. We are unable to handle the first case, and leave it as a future direction. Perhaps this case does not occur typically, since for many problems gradient descent almost surely avoids first-order saddle points (Lee et al., 2016, 2019).*

Remark 14 *Theorem 4 and Theorem 10 implies that the trajectories of $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ and $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ stay close after escaping the origin. One might attempt to prove this using the result of Kumar and Haupt (2024, 2025) stated in Lemma 3, which shows for $L = 2$ that, for small $\epsilon > 0$, there exists a time T such that $\boldsymbol{\psi}(T, \delta\mathbf{w}_0)^\top \mathbf{w}_* / \|\boldsymbol{\psi}(T, \delta\mathbf{w}_0)\|_2 = 1 - O(\epsilon)$, for all small $\delta > 0$. To explore this idea further, suppose $\mathbf{w}_0 = \mathbf{w}_* + \epsilon\mathbf{b}$, where $\mathbf{b} \in \mathbf{w}_*^\perp$, that is, approximate directional convergence holds at the initialization itself. Since $\boldsymbol{\psi}(\cdot, \cdot)$ varies continuously with initialization and $\|\delta\mathbf{w}_0 - \delta\mathbf{w}_*\|_2 = \delta\epsilon$, which is small because δ and ϵ are small, it can be shown that $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ and $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ remain close for some time after the training begins. However, that time may not be long enough to ensure that $\boldsymbol{\psi}(t, \delta\mathbf{w}_*)$ and $\boldsymbol{\psi}(t, \delta\mathbf{w}_0)$ remain close after they escape from the origin, making this approach unsuccessful (see Appendix E for more details).*

4. Implications for Feed-Forward Neural Networks

This section discusses the implications of the above theorems for feed-forward neural network. Specifically, we use the above theorems and the results of Kumar and Haupt (2025), which studies the KKT points of the NCF, to show how sparsity structure is preserved among the weights after gradient flow escapes from the origin.

We first introduce the notion of *zero-preserving subset*, which is a subset of the weights of a neural network that remain zero throughout training, if they are zero at initialization.

Definition 15 *Consider the training setup of Section 2, and recall $\boldsymbol{\psi}(t, \mathbf{w}(0))$ denotes the solution of*

$$\dot{\mathbf{w}} = -\nabla\mathcal{L}(\mathbf{w}),$$

where $\mathbf{w}(0)$ is the initialization. If $\mathbf{w}_z$ is a vector containing subset of entries in $\mathbf{w}$,[3] *then $\boldsymbol{\psi}_{\mathbf{w}_z}(t, \mathbf{w}(0))$ denotes the evolution of weights belonging to $\mathbf{w}_z$. We define $\mathbf{w}_z$ to be a zero-preserving subset of $\mathbf{w}$ under the following condition: if $\|\boldsymbol{\psi}_{\mathbf{w}_z}(0, \mathbf{w}(0))\|_2 = 0$, then $\|\boldsymbol{\psi}_{\mathbf{w}_z}(t, \mathbf{w}(0))\|_2 = 0$, for all $t \in (-\infty, \infty)$.*

Note that there could be a subset of weights which are non-zero at initialization, but become zero at some stage of the training and remain zero after that. There could also be subset of weights which are zero at initialization, but become non-zero during training. Such set of weights do not satisfy the definition of zero-preserving subset. The weights belonging to zero-preserving subset are zero at initialization, and they remain zero throughout training.

It may seem that the training data plays a role in determining the zero-preserving subsets, that is, for a fixed architecture, different training data may lead to different sets of zero-preserving subsets. While this is possible, we show that for some architectures, certain zero-preserving subsets do not change with the training data. In such cases, zero-preserving subsets should be viewed as a property of the architecture, rather than the training data.

For $L \geq 2$, the output of an $L-$layer feed-forward neural network $\mathcal{H}$ is

$$\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L) = \mathbf{W}_L\sigma(\mathbf{W}_{L-1}\cdots\sigma(\mathbf{W}_1\mathbf{x})\cdots), \tag{9}$$

3. Throughout this section, we will abuse the notation slightly and refer to a vector as a set containing its entries and vice-versa.

where $\mathbf{W}_l \in \mathbb{R}^{k_l \times k_{l-1}}$, $k_0 = d$ and $k_L = 1$. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is applied coordinate wise. Note that, if $\sigma(x) = \max(x, \alpha x)^p$, for some $p \in \mathbb{N}$ and $\alpha \in \mathbb{R}$, then the above neural network is homogeneous. We use $\mathbf{W}_l[:, j]$ and $\mathbf{W}_l[j, :]$ to denote the j-th column and j-th row of $\mathbf{W}_l$ respectively. Also, $\mathbf{W}_l[j, :]$ contains incoming weights from the j-th neuron in the l-th layer, and $\mathbf{W}_{l+1}[:, j]$ contains outgoing weights from the same neuron.

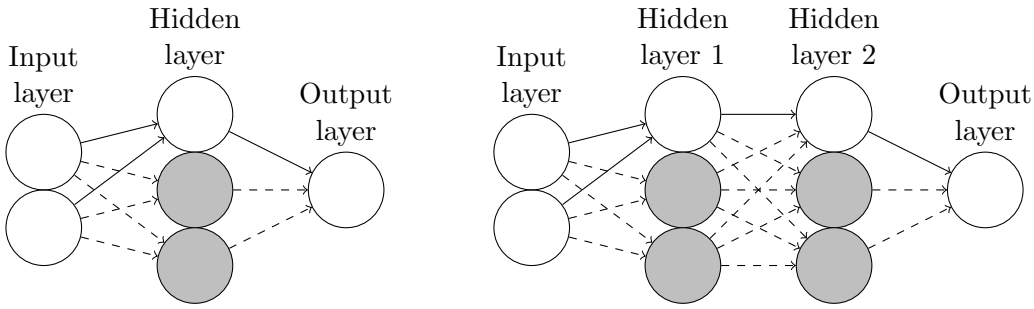
The following lemma describes the zero-preserving subsets of feed-forward neural networks for arbitrary training data.

Lemma 16 *Let $\mathcal{H}$ be an $L-$layer feed-forward neural network as in eq. (9), where $\sigma(\cdot)$ is locally Lipschitz, continuously differentiable*[4] *and $\sigma(0) = 0$. Suppose $\mathbf{w}_z$ is a subset of the weights such that*

$$\mathbf{w}_z = \bigcup_{l=1}^{L-1} \left(\mathbf{W}_l[j, :] \cup \mathbf{W}_{l+1}[:, j], j \in \mathcal{S}_l \right),$$

where $\mathcal{S}_l$ is an arbitrary subset of $[k_l]$, for all $l \in [L-1]$, then $\mathbf{w}_z$ is a zero-preserving subset.

The assumption $\sigma(0) = 0$ is satisfied by all homogeneous activation functions. According to the above lemma, the zero-preserving subset is a collection of rows and columns of the weight matrices such that if $\mathbf{W}_l[j, :] \in \mathbf{w}_z$, then $\mathbf{W}_{l+1}[:, j] \in \mathbf{w}_z$, and vice-versa. Another way to view the zero-preserving subset is to look at the hidden neurons. The zero-preserving subset is formed by combining *all* the incoming and outgoing weights of certain subset of hidden neurons; some examples are provided in Figure 3. Also, note that the zero-preserving subset contains only rows of $\mathbf{W}_1$ and only columns of $\mathbf{W}_L$, whereas it could contain rows and columns of other weight matrices.



(a) Single hidden layer (b) Two hidden layer

Figure 3: The weights corresponding to the dashed arrows, or equivalently, all the incoming and outgoing weights of the hidden neurons in gray, form a zero-preserving subset.

Now, recall that our goal in this section is to show that, for feed-forward homogeneous neural networks trained via gradient flow with small initialization, the sparsity structure emerging among the weights just before escaping the origin is preserved post-escape. Briefly, our approach to achieve this goal is as follows: From Theorem 4 and Theorem 10, we know that for all sufficiently small initializations, gradient flow escapes the origin along the same path as gradient flow with small initial weights and initial direction along a KKT point of

4. We assume differentiability for simplicity, the result can also be proved for ReLU activation.

the constrained NCF. Now, in Kumar and Haupt (2025), KKT points of the constrained NCF for feed-forward homogeneous networks are studied, and for these KKT points, it can be shown that all the rows and columns of the hidden weights with zero magnitude form a zero-preserving subset. Combining this fact with the result of Theorem 4 and Theorem 10, we will show that if the initial weights lie in a stable set of a positive KKT point, then the sparsity structure will approximately be preserved after gradient flow escapes the origin. We next discuss this approach in greater detail.

We start with the following lemma from Kumar and Haupt (2025, Lemma 5). For feed-forward homogeneous networks, it highlights a key property of KKT points of the constrained NCF.

Lemma 17 *Let $\mathcal{H}$ be an $L-$layer feed-forward neural network as in eq. (9), where $\sigma(x) = \max(x, \alpha x)^p$, for some $p \in \mathbb{N}$ and $\alpha \in \mathbb{R}$. Let $\big(\overline{\mathbf{W}}_1, \cdots \overline{\mathbf{W}}_L\big)$ be a positive KKT point of*

$$\max_{\mathbf{W}_1, \cdots, \mathbf{W}_L} \mathcal{N}(\mathbf{W}_1, \cdots, \mathbf{W}_L) \coloneqq \widetilde{\mathbf{y}}^\top \mathcal{H}(\mathbf{X}; \mathbf{W}_1, \cdots, \mathbf{W}_L), \ \text{s.t.} \ \sum_{i=1}^{L} \|\mathbf{W}_i\|_F^2 = 1. \tag{10}$$

Then,

$$\textit{diag}(\overline{\mathbf{W}}_l \overline{\mathbf{W}}_l^\top) = p \cdot \textit{diag}(\overline{\mathbf{W}}_{l+1}^\top \overline{\mathbf{W}}_{l+1}), \ \textit{for all } l \in [L-1],$$

or equivalently, $\|\overline{\mathbf{W}}_l[j,:]\|_2^2 = p\|\overline{\mathbf{W}}_{l+1}[:,j]\|_2^2$, for all $j \in [k_l]$ and $l \in [L-1]$.

In Kumar and Haupt (2025), authors empirically observed that, for feed-forward homogeneous neural networks, the weights tend to converge towards KKT points of the constrained NCF that have many rows and columns of the hidden weights with zero norm, during the early stages of training. In fact, for $p \geq 2$ (polynomial Leaky ReLU), the hidden weights usually had only one row and column with non-zero norm.

Now, using these observations and the above lemma, we describe the zero-preserving subset emerging from the sparsity structure in positive KKT points, and its consequence on dynamics of gradient flow beyond the origin. Since $\|\overline{\mathbf{W}}_l[j,:]\|_2^2 = p\|\overline{\mathbf{W}}_{l+1}[:,j]\|_2^2$, if $\|\overline{\mathbf{W}}_l[j,:]\|_2$ is 0, then $\|\overline{\mathbf{W}}_{l+1}[:,j]\|_2$ is 0, and vice versa. Therefore, all the rows and columns of a KKT point with zero norm will form a zero-preserving subset. Consequently, for gradient flow with initial direction along such KKT points, by the definition of zero-preserving subset, all the rows and columns with zero norm at initialization will remain zero during training, implying the sparsity structure is preserved during training.

Furthermore, from Theorem 4 and Theorem 10, we know that the trajectory of gradient flow with small initialization after escaping the origin is approximately equal to the gradient flow with small initial weights and initial direction along a KKT point of the NCF. Hence, for a feed-forward homogeneous neural network, if direction of the initial weights are in a stable set of a positive KKT point, then the sparsity structure will be approximately preserved after escaping the origin. We formalize this idea in the following theorem, where we use $\mathbf{W}_{1:L}^0 \coloneqq (\mathbf{W}_1^0, \cdots, \mathbf{W}_L^0)$ and $\overline{\mathbf{W}}_{1:L} \coloneqq (\overline{\mathbf{W}}_1, \cdots, \overline{\mathbf{W}}_L)$, to denote the initial direction of the weights and the KKT point of the constrained NCF respectively.

Theorem 18 *Let $\mathcal{H}$ be an $L-$layer feed-forward neural network as in eq. (9), where $\sigma(x) = \max(x, \alpha x)^p$, for some $p \in \mathbb{N}, p \geq 2$ and $\alpha \in \mathbb{R}$, or $p = 1, \alpha = 1$. Suppose $\mathbf{W}_{1:L}^0 \in \mathcal{S}(\overline{\mathbf{W}}_{1:L})$,*

where $\overline{\mathbf{W}}_{1:L}$ *is a* $\Delta-$*second-order positive KKT point of eq.* (10)*. Define* $\overline{\mathcal{N}} \coloneqq \mathcal{N}(\overline{\mathbf{W}}_{1:L})$*, and let* $\mathbf{w}_z$ *be the following subset of the weights:*

$$\mathbf{w}_z = \bigcup_{l=1}^{L-1} \Big(\left\{ \mathbf{W}_l[j,:] : \left\| \overline{\mathbf{W}}_l[j,:] \right\|_2 = 0, j \in [k_l] \right\} \cup \left\{ \mathbf{W}_{l+1}[:,j] : \left\| \overline{\mathbf{W}}_{l+1}[:,j] \right\|_2 = 0, j \in [k_l] \right\} \Big).$$

- *If* $\mathcal{H}$ *is two-homogeneous, then, for all sufficiently small* $\delta > 0$*,*

$$\left\| \boldsymbol{\psi} \left(T^1_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2 \leq b_1 \delta^{\frac{\Delta}{\Delta + 8\overline{\mathcal{N}}}}, \text{ and } \frac{\left\| \boldsymbol{\psi}_{\mathbf{w}_z} \left(T^1_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2}{\left\| \boldsymbol{\psi} \left(T^1_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2} \leq b_2 \delta^{\frac{2\Delta}{\Delta + 8\overline{\mathcal{N}}}}, \tag{11}$$

where $b_1, b_2 > 0$ *are some constants,* $T^1_{\widetilde{\delta}} = T_1 + 4\ln(1/\widetilde{\delta})/(\Delta + 8\overline{\mathcal{N}})$ *and* $T_1, \widetilde{\delta}$ *are the same as defined in Theorem 4. Furthermore, for any fixed* $\widetilde{T} > 0$*, there exists a* $\widetilde{C} > 0$ *such that, for all sufficiently small* $\delta > 0$ *and for all* $t \in [-\widetilde{T}, \widetilde{T}]$*,*

$$\left\| \boldsymbol{\psi}_{\mathbf{w}_z} \left(t + T^2_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2 \leq \widetilde{C} \delta^{\frac{\Delta}{\Delta + 8\overline{\mathcal{N}}}}, \tag{12}$$

where $T^2_{\widetilde{\delta}} = T_1 + \frac{\ln(1/b_\delta)}{2\overline{\mathcal{N}}} + \frac{\ln(1/\widetilde{\delta})}{2\overline{\mathcal{N}}}$*, and* b_δ *is the same as defined in Theorem 4.*

- *If* $\mathcal{H}$ *is* L*-homogeneous, for some* $L > 2$*, then, for all sufficiently small* $\delta > 0$*,*

$$\left\| \boldsymbol{\psi} \left(T^1_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2 \leq b_1 \delta^{\frac{\Delta}{\Delta + 2L^2\overline{\mathcal{N}}}}, \text{ and } \frac{\left\| \boldsymbol{\psi}_{\mathbf{w}_z} \left(T^1_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2}{\left\| \boldsymbol{\psi} \left(T^1_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2} \leq b_2 \delta^{\frac{L\Delta}{\Delta + 2L^2\overline{\mathcal{N}}}}, \tag{13}$$

where $b_1, b_2 > 0$ *are some constants,* $T^1_{\widetilde{\delta}} = \frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}} \left(1 - \widetilde{\delta}^{\frac{2(L-2)L^2\overline{\mathcal{N}}}{2L^2\overline{\mathcal{N}} + \Delta}} \right)$*, and* $T_1, T, \widetilde{\delta}$ *are the same as defined in Theorem 10. Furthermore, for any fixed* $\widetilde{T} > 0$*, there exists a* $\widetilde{C} > 0$ *such that, for all sufficiently small* $\delta > 0$ *and for all* $t \in [-\widetilde{T}, \widetilde{T}]$*,*

$$\left\| \boldsymbol{\psi}_{\mathbf{w}_z} \left(t + T^2_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L} \right) \right\|_2 \leq \widetilde{C} \delta^{\frac{\Delta}{2L^2\overline{\mathcal{N}} + \Delta}}, \tag{14}$$

where $T^2_{\widetilde{\delta}} = \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\overline{\mathcal{N}}} - T \right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\overline{\mathcal{N}} + \Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}$*, and* b_δ *is the same as defined in Theorem 10.*

In the above theorem, $\overline{\mathbf{W}}_{1:L}$ is a positive KKT point of the constrained NCF, and $\mathbf{w}_z$ contain all the rows and columns of $\overline{\mathbf{W}}_{1:L}$ that have zero norm,[5] which is a zero-preserving subset, as discussed above. For two-homogeneous networks, eq. (11) implies that at time $T^1_{\widetilde{\delta}}$, the norm of the weights are small, but weights belonging to $\mathbf{w}_z$ have much smaller magnitude. Therefore, a sparsity structure emerges among the weights before gradient flow escapes the origin. We mainly use Lemma 8 to get eq. (11) which, we recall, implies that during the early stages of training, the weights remain small and converge in direction

5. More precisely, all the rows of $(\overline{\mathbf{W}}_1, \cdots, \overline{\mathbf{W}}_{L-1})$ and all the columns of $(\overline{\mathbf{W}}_2, \cdots, \overline{\mathbf{W}}_L)$ with zero norm.

towards $\overline{\mathbf{W}}_{1:L}$. Since $\overline{\mathbf{W}}_{1:L}$ has a sparsity structure, specifically, weights belonging to $\mathbf{w}_z$ have zero magnitude, therefore, the same sparsity structure would emerge in the weights of the neural network during the early stages of training.

From Theorem 4, we know that gradient flow escapes the origin after $T^2_{\widetilde{\delta}}$ time has elapsed. Thus, eq. (12) implies that weights belonging to $\mathbf{w}_z$ remain small even after gradient flow escapes the origin. Hence, the sparsity structure that emerges among the weights before escaping the origin is preserved post-escape as well. To get eq. (12), we rely on Theorem 4 and $\mathbf{w}_z$ being a zero-preserving subset.

For deep homogeneous networks, similar results are provided in eq. (13) and eq. (14). Additionally, using the above theorem and following the proofs of Corollaries 5, 6, 11, and 12, it can be shown that the sparsity structure is preserved until the first saddle point encountered by gradient flow after escaping the origin.

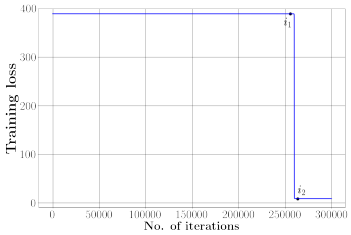
4.1 Numerical Experiments

We next conduct experiments to validate the above theorem, with results presented in Figures 4, 5 and 6. The setting for these experiments is similar to Figure 1, the weights are trained using gradient descent with small initialization and until they escape the origin and reach the next saddle point. Each figure contains both the evolution of training loss and the plot of weights, before escaping the origin and after reaching the next saddle point. The code is available at `github.com/akk0135/beyond_origin_experiments`.

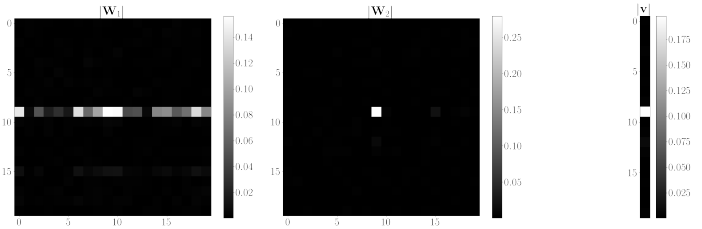
Square activation function. In Figure 1, we observed that the weights of a two-layer network with square activation became sparse before gradient descent escaped the origin, with only one row of $\mathbf{W}$ and $\mathbf{v}$ being non-zero. In accordance with Theorem 18, this sparsity structure was preserved after escaping the origin as well. Similarly, Figure 4 shows that for a three-layer neural network with square activation, before gradient descent escapes the origin, a single row of $\mathbf{W}_1$, and a single entry of $\mathbf{W}_2$, $\mathbf{v}$ are non-zero, and this sparsity structure is preserved even after gradient descent escapes the origin and reaches the next saddle point. Notably, the rows and columns with zero norm form a zero-preserving subset.

ReLU activation function. Although our results exclude ReLU networks, we explore two- and three-layer ReLU networks in Figure 5 and Figure 6. Note that, before escaping the origin, the weights become approximately sparse, but, compared to square activation, a greater number of rows and columns of the weights are non-zero. This aligns with observations from Kumar and Haupt (2025), which noted that for ReLU networks, gradient descent tend to converge towards a KKT point of the constrained NCF with multiple non-zero rows and columns. Next, a close look at the weights seems to suggest that rows and columns of relatively small size stay small, even after the gradient descent escapes the origin and reaches the next saddle point. Also, the rows and columns with relatively small size seems to follow the definition of zero-preserving subset, that is, if the j-th row of $\mathbf{W}_l$ is small, then the j-th column of $\mathbf{W}_{l+1}$ is small, and vice-versa. For example, in the two-layer case, the first two rows of $\mathbf{W}$ and first two columns of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2. In the three-layer case, rows 15-17 of $\mathbf{W}_1$ and columns 15-17 of $\mathbf{W}_2$, and rows 2-5 of $\mathbf{W}_2$ and columns 2-5 of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2.

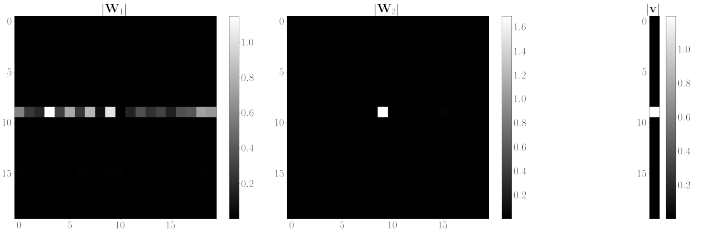
These observations suggest that the sparsity structure is also preserved for ReLU networks, even though our results exclude ReLU activation. Additional experiments are provided in Appendix F. These observations can be useful for future work in this area.



(a) Evolution of training loss with iterations



(b) Weights at iteration i_1



(c) Weights at iteration i_2

Figure 4: We train a three-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))$, where $\sigma(x) = x^2$ (**square activation**), and $\mathbf{v} \in \mathbb{R}^{20}, \mathbf{W}_2, \mathbf{W}_1 \in \mathbb{R}^{20 \times 20}$ are the trainable weights. The sparsity structure is preserved upon escaping from the origin.

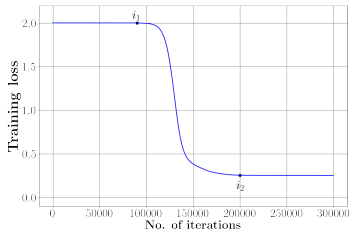
5. Conclusion and Future Directions

This paper studied the gradient flow dynamics of homogeneous neural networks which are trained with small initialization. We showed that for all sufficiently small initializations, the gradient flow escapes along the same path as the gradient flow with small initial weights and initial direction along a KKT point of the NCF. Next, we studied the gradient flow dynamics of feed-forward homogeneous neural networks and showed that the sparsity structure that emerges among the weights before escaping the origin is preserved post-escape.

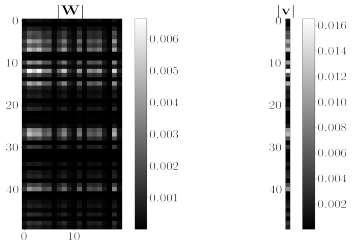
Our work describes a segment of the gradient flow dynamics beyond the origin, specifically up to the first saddle point encountered by gradient flow after escaping the origin. Understanding gradient flow dynamics beyond that saddle point is an important direction for future research; Appendix G provides some observations in this regard. Also, our results hold for neural networks with locally Lipschitz gradient, which excludes ReLU neural networks. Extending our results for such neural networks would be a valuable next step.

Acknowledgments

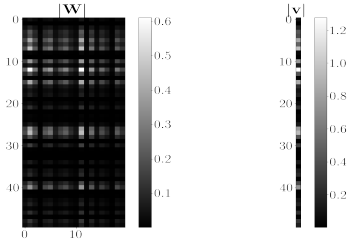
The authors graciously acknowledge resource support from the Minnesota Supercomputing Institute (MSI), and financial support in the form of gift funding from InterDigital.



(a) Evolution of training loss with iterations

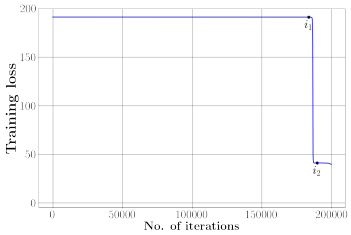


(b) Weights at iteration i_1

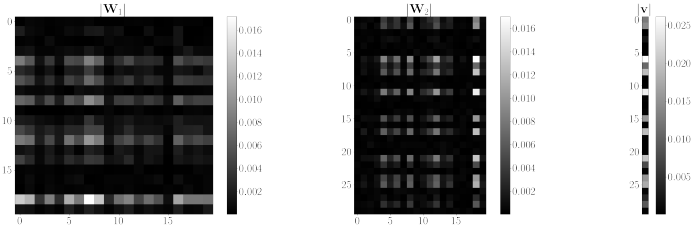


(c) Weights at iteration i_2

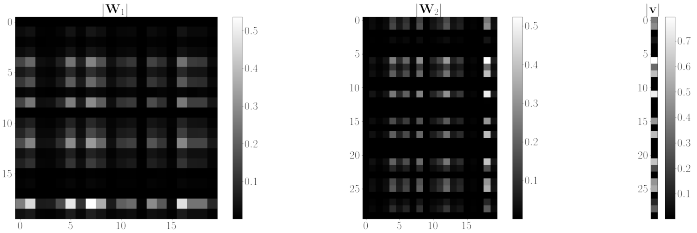
Figure 5: We train a two-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x})$, where $\sigma(x) = \max(x, 0)$ (**ReLU activation**), and $\mathbf{v} \in \mathbb{R}^{50}, \mathbf{W} \in \mathbb{R}^{50\times 20}$ are the trainable weights. As in the previous example, the sparsity structure is preserved upon escaping from the origin.



(a) Evolution of training loss with iterations



(b) Weights at iteration i_1



(c) Weights at iteration i_2

Figure 6: We train a three-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}))$, where $\sigma(x) = \max(x, 0)$ (**ReLU activation**), and $\mathbf{v} \in \mathbb{R}^{30}, \mathbf{W}_2 \in \mathbb{R}^{30\times 20}, \mathbf{W}_1 \in \mathbb{R}^{20\times 20}$ are the trainable weights. This highlights the preservation of sparsity structure after escaping from the origin even for deeper ReLU networks.

Organization of the Appendix: Appendix A contains key lemmata useful to prove our main results. The proofs omitted from Section 3.1, 3.2 and 4 are in Appendix B, C and D, respectively. Appendix E contains additional results and discussion, while Appendix F reports further experiments. Finally, Appendix G presents empirical observations on the training dynamics after escaping the first saddle point.

Appendix A. Key Lemmata

The following lemma, also known as Euler's theorem, states two important properties of homogeneous functions (Lyu and Li, 2020, Theorem B.2).

Lemma 19 *Let $F : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz, differentiable and L-positively homogeneous for some $L > 0$. Then,*

- *For any $\mathbf{x} \in \mathbb{R}^d$ and $c \geq 0$, $\nabla F(c\mathbf{x}) = c^{L-1}F(\mathbf{x})$.*
- *For any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^\top \nabla F(\mathbf{x}) = LF(\mathbf{x})$.*

Some properties of the second-order KKT points of the constrained NCF are derived next.

Lemma 20 *Suppose $\mathcal{H}$ is $L-$homogeneous, where $L \geq 2$. Let $\mathbf{w}_*$ be a $\Delta-$second-order positive KKT point of the constrained NCF in eq.* (4)*. Then,*

$$\nabla \mathcal{N}(\mathbf{w}_*) = L\mathcal{N}(\mathbf{w}_*)\mathbf{w}_*, L\mathcal{N}(\mathbf{w}_*) - \mathbf{b}^\top \nabla^2 \mathcal{N}(\mathbf{w}_*)\mathbf{b} \geq \Delta, \forall \mathbf{b} \in \mathbf{w}_*^\perp, \text{ and} \tag{15}$$

$$\|\nabla^2 \mathcal{N}(\mathbf{w}_*)\|_2 \leq L(L-1)\mathcal{N}(\mathbf{w}_*). \tag{16}$$

Proof For eq. (4), the Lagrangian is

$$L(\mathbf{w}, \lambda) = \mathcal{N}(\mathbf{w}) + \lambda(1 - \|\mathbf{w}\|_2^2). \tag{17}$$

Since $\mathbf{w}_*$ is a second-order KKT point, $\|\mathbf{w}_*\|_2^2 = 1$,

$$\mathbf{0} = \nabla_{\mathbf{w}} L(\mathbf{w}_*, \lambda_*) = \nabla \mathcal{N}(\mathbf{w}_*) - 2\lambda_* \mathbf{w}_*, \tag{18}$$

and

$$\Delta \leq -\mathbf{b}^\top \nabla^2_{\mathbf{w}} L(\mathbf{w}_*, \lambda_*)\mathbf{b} = 2\lambda_* - \mathbf{b}^\top \nabla^2 \mathcal{N}(\mathbf{w}_*)\mathbf{b}, \forall \mathbf{b} \in \mathbf{w}_*^\perp. \tag{19}$$

Multiplying eq. (18) by $\mathbf{w}_*^\top$ from the left, and using Lemma 19 and $\|\mathbf{w}_*\|_2^2 = 1$, we get that

$$\mathbf{0} = L\mathcal{N}(\mathbf{w}_*) - 2\lambda_*.$$

Using the above value of λ_* in eq. (18) and eq. (19) gives us eq. (15). Next, since $\nabla \mathcal{N}(\mathbf{w})$ is $(L-1)$-homogeneous, from Lemma 19, we have

$$\nabla^2 \mathcal{N}(\mathbf{w}_*)\mathbf{w}_* = (L-1)\nabla \mathcal{N}(\mathbf{w}_*).$$

Since $\nabla \mathcal{N}(\mathbf{w}_*) = L\mathcal{N}(\mathbf{w}_*)\mathbf{w}_*$, the above equation implies that $\mathbf{w}_*$ is an eigenvector of $\nabla^2 \mathcal{N}(\mathbf{w}_*)$ with eigenvalue $L(L-1)\mathcal{N}(\mathbf{w}_*)$. Now, since $\nabla^2 \mathcal{N}(\mathbf{w}_*)$ is a symmetric matrix, other eigenvectors would be orthogonal to $\mathbf{w}_*$. However, for all $\mathbf{b} \in \mathbf{w}_*^\perp$, we know

$$\mathbf{b}^\top \nabla^2 \mathcal{N}(\mathbf{w}_*)\mathbf{b} \leq L\mathcal{N}(\mathbf{w}_*) - \Delta \leq L(L-1)\mathcal{N}(\mathbf{w}_*).$$

Hence, $\|\nabla^2 \mathcal{N}(\mathbf{w}_*)\|_2 \leq L(L-1)\mathcal{N}(\mathbf{w}_*)$. ■

A.1 Local Behavior at KKT point

In this subsection, we derive some important inequalities satisfied by the weights near a $\Delta-$second-order KKT point of the constrained NCF.

Lemma 21 *Suppose $\mathcal{H}$ is L-homogeneous, where $L \geq 2$. Let $\mathbf{w}_*$ be a $\Delta-$second-order positive KKT point of eq. (4), then for all sufficiently small $\epsilon > 0$ and $\mathbf{b} \in \mathbf{w}_*^{\perp}$, we have*

$$\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b}) = \mathcal{N}(\mathbf{w}_*) + \frac{\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2), \text{ and} \tag{20}$$

$$\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{w}_* = L\mathcal{N}(\mathbf{w}_*) + \frac{(L-2)\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2). \tag{21}$$

Proof Using Taylor's theorem, for all sufficiently small $\epsilon > 0$, we have

$$\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b}) = \mathcal{N}(\mathbf{w}_*) + \epsilon\mathbf{b}^\top\nabla\mathcal{N}(\mathbf{w}_*) + \frac{\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2).$$

From Lemma 20, we know $\nabla\mathcal{N}(\mathbf{w}_*)^\top\mathbf{b} = L\mathcal{N}(\mathbf{w}_*)\mathbf{w}_*^\top\mathbf{b} = 0$. Using this fact in the above equation gives us eq. (20). Next, note that

$$\begin{aligned}\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{w}_* &= \nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top(\mathbf{w}_* + \epsilon\mathbf{b}) - \epsilon\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{b} \\ &= L\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b}) - \epsilon\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{b},\end{aligned} \tag{22}$$

where the last equality follows from Lemma 19. Now, using eq. (20), we have

$$L\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b}) = L\mathcal{N}(\mathbf{w}_*) + \frac{L\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2).$$

From Taylor's theorem, for all sufficiently small $\epsilon > 0$, we have

$$\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{b} = \nabla\mathcal{N}(\mathbf{w}_*)^\top\mathbf{b} + \epsilon\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon) = \epsilon\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon),$$

where we used $\nabla\mathcal{N}(\mathbf{w}_*)^\top\mathbf{b} = 0$. Combining the above two equalities with eq. (22) gives us

$$\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{w}_* = L\mathcal{N}(\mathbf{w}_*) + \frac{(L-2)\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2),$$

which completes the proof. ■

Lemma 22 *Suppose $\mathcal{H}$ is L-homogeneous, where $L \geq 2$. Let $\mathbf{w}_*$ be a $\Delta-$second-order positive KKT point of eq. (4), then there exists a sufficiently small $\gamma > 0$ such that for any unit-norm vector $\mathbf{w}$ satisfying $\mathbf{w}^\top\mathbf{w}_* \geq 1 - \gamma$, and for all $t_2 \geq t_1 \geq 0$, the following holds:*

$$\left(\nabla\mathcal{N}(t_1\mathbf{w}) - \nabla\mathcal{N}(t_2\mathbf{w}_*)\right)^\top(t_1\mathbf{w} - t_2\mathbf{w}_*) - L(L-1)\mathcal{N}(\mathbf{w}_*)t_2^{L-2}\|t_1\mathbf{w} - t_2\mathbf{w}_*\|_2^2 \leq 0. \tag{23}$$

Proof The proof is trivial if $t_1 = 0$, since in that case

$$\left(\nabla\mathcal{N}(t_1\mathbf{w}) - \nabla\mathcal{N}(t_2\mathbf{w}_*)\right)^\top(t_1\mathbf{w} - t_2\mathbf{w}_*) = \nabla\mathcal{N}(t_2\mathbf{w}_*)^\top(t_2\mathbf{w}_*) = t_2^L L\mathcal{N}(\mathbf{w}_*),$$

and $L(L-1)\mathcal{N}(\mathbf{w}_*)t_2^{L-2}\|t_1\mathbf{w}-t_2\mathbf{w}_*\|_2^2 = L(L-1)\mathcal{N}(\mathbf{w}_*)t_2^L \geq t_2^L L\mathcal{N}(\mathbf{w}_*)$. Therefore, we assume $t_1 > 0$. Now, define $t = t_2/t_1$, then eq. (23) can be written as

$$\left(\nabla\mathcal{N}(\mathbf{w}) - \nabla\mathcal{N}(t\mathbf{w}_*)\right)^\top(\mathbf{w}-t\mathbf{w}_*) - L(L-1)\mathcal{N}(\mathbf{w}_*)t^{L-2}\|\mathbf{w}-t\mathbf{w}_*\|_2^2 \leq 0. \tag{24}$$

We will show that the above inequality is true for all $t \geq 1$ to prove the lemma. We will proceed by considering two cases: $L = 2$ and $L > 2$.

Case 1 ($L = 2$): Since $\mathbf{w}_*$ is a Δ-second-order KKT point, from Lemma 20 we know

$$\nabla\mathcal{N}(\mathbf{w}_*) = 2\mathcal{N}(\mathbf{w}_*)\mathbf{w}_*, \text{ and } \Delta + \mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} \leq 2\mathcal{N}(\mathbf{w}_*), \forall \mathbf{b} \in \mathbf{w}_*^\perp. \tag{25}$$

For the sake of brevity, we define $\lambda_* = 2\mathcal{N}(\mathbf{w}_*)$ and $b_* = \mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b}$ here onward in this proof. Using Lemma 19, homogeneity and $\|\mathbf{w}\|_2 = 1$, the LHS of eq. (24) can be written as

$$\begin{aligned}
&2\mathcal{N}(\mathbf{w}) + 2t^2\mathcal{N}(\mathbf{w}_*) - t\nabla\mathcal{N}(\mathbf{w}_*)^\top\mathbf{w} - t\nabla\mathcal{N}(\mathbf{w})^\top\mathbf{w}_* - 2\mathcal{N}(\mathbf{w}_*)(1+t^2-2t\mathbf{w}^\top\mathbf{w}_*)\\
&= 2\mathcal{N}(\mathbf{w}) + \lambda_* t^2 - \lambda_* t\mathbf{w}_*^\top\mathbf{w} - t\nabla\mathcal{N}(\mathbf{w})^\top\mathbf{w}_* - \lambda_*(1+t^2-2t\mathbf{w}^\top\mathbf{w}_*)\\
&= 2\mathcal{N}(\mathbf{w}) - \lambda_* - t\nabla\mathcal{N}(\mathbf{w})^\top\mathbf{w}_* + \lambda_* t\mathbf{w}_*^\top\mathbf{w},
\end{aligned}$$

where the first equality uses eq. (25), and simplifying the first equality leads to the second. Since $\mathbf{w}$ has unit norm, we define $\mathbf{w} = (\mathbf{w}_* + \epsilon\mathbf{b})/\sqrt{1+\epsilon^2}$, where $\mathbf{b} \in \mathbf{w}_*^\perp$. Putting this choice of $\mathbf{w}$ in the above equation gives us

$$\frac{2\mathcal{N}(\mathbf{w}_*+\epsilon\mathbf{b})}{(1+\epsilon^2)} - \lambda_* - \frac{t\nabla\mathcal{N}(\mathbf{w}_*+\epsilon\mathbf{b})^\top\mathbf{w}_*}{\sqrt{1+\epsilon^2}} + \frac{\lambda_* t\mathbf{w}_*^\top(\mathbf{w}_*+\epsilon\mathbf{b})}{\sqrt{1+\epsilon^2}}.$$

To complete the proof, we will show that the above quantity is less than zero for all sufficiently small $\epsilon > 0$. Using Lemma 21 and $\|\mathbf{w}_*\|_2 = 1$, the above quantity becomes

$$\frac{2\mathcal{N}(\mathbf{w}_*) + \epsilon^2 b_* + o(\epsilon^2)}{(1+\epsilon^2)} - \lambda_* - \frac{2t\mathcal{N}(\mathbf{w}_*) + o(\epsilon^2)}{\sqrt{1+\epsilon^2}} + \frac{\lambda_* t}{\sqrt{1+\epsilon^2}},$$

which can be further simplified to get

$$\frac{\lambda_* + \epsilon^2 b_*}{(1+\epsilon^2)} - \lambda_* + o(\epsilon^2).$$

Since $1/(1+\epsilon^2) = 1 - \epsilon^2 + o(\epsilon^2)$, the above quantity can be written as

$$\lambda_*(1-\epsilon^2) + \epsilon^2 b_* - \lambda_* + o(\epsilon^2) = \epsilon^2(b_* - \lambda_*) + o(\epsilon^2) \leq -\Delta\epsilon^2 + o(\epsilon^2).$$

Since $\Delta > 0$, the proof is complete.

Case 2 ($L > 2$): Since $\mathbf{w}_*$ is a $\Delta-$second-order KKT point, from Lemma 20 we have

$$\nabla\mathcal{N}(\mathbf{w}_*) = L\mathcal{N}(\mathbf{w}_*)\mathbf{w}_*, \text{ and } \Delta + \mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} \leq L\mathcal{N}(\mathbf{w}_*), \forall \mathbf{b} \in \mathbf{w}_*^\perp. \tag{26}$$

For the sake of brevity, we define $\lambda_* = L\mathcal{N}(\mathbf{w}_*)$ and $b_* = \mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b}$ here onward in this proof. Using Lemma 19, homogeneity and $\|\mathbf{w}\|_2 = 1$, the LHS of eq. (24) can be written as

$$\begin{aligned}
&L\mathcal{N}(\mathbf{w}) + Lt^L\mathcal{N}(\mathbf{w}_*) - t^{L-1}\nabla\mathcal{N}(\mathbf{w}_*)^\top\mathbf{w} - t\nabla\mathcal{N}(\mathbf{w})^\top\mathbf{w}_*\\
&- t^{L-2}L(L-1)\mathcal{N}(\mathbf{w}_*)(1+t^2-2t\mathbf{w}^\top\mathbf{w}_*)\\
&= L\mathcal{N}(\mathbf{w}) + \lambda_* t^L - \lambda_* t^{L-1}\mathbf{w}_*^\top\mathbf{w} - t\nabla\mathcal{N}(\mathbf{w})^\top\mathbf{w}_* - \lambda_* t^{L-2}(L-1)(1+t^2-2t\mathbf{w}^\top\mathbf{w}_*)\\
&= L\mathcal{N}(\mathbf{w}) - \lambda_*((L-1)t^{L-2} + (L-2)t^L) - t\nabla\mathcal{N}(\mathbf{w})^\top\mathbf{w}_* + \lambda_*(2L-3)t^{L-1}\mathbf{w}_*^\top\mathbf{w},
\end{aligned}$$

where the first equality uses eq. (26), and simplifying the first equality leads to the second. Define $\mathbf{w} = (\mathbf{w}_* + \epsilon\mathbf{b})/\sqrt{1+\epsilon^2}$, where $\mathbf{b} \in \mathbf{w}_*^\perp$. Putting this choice of $\mathbf{w}$ in the above equation gives us

$$\frac{L\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})}{(1+\epsilon^2)^{L/2}} - \lambda_*((L-1)t^{L-2} + (L-2)t^L) - \frac{t\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top \mathbf{w}_*}{(1+\epsilon^2)^{(L-1)/2}} + \frac{(2L-3)\lambda_* t^{L-1}}{\sqrt{1+\epsilon^2}}.$$

To complete the proof, we will show that the above quantity is less than zero for all sufficiently small $\epsilon > 0$. Using Lemma 21 and $\|\mathbf{w}_*\|_2 = 1$, we further get

$$\frac{\lambda_* + \frac{L\epsilon^2}{2}b_* + o(\epsilon^2)}{(1+\epsilon^2)^{L/2}} - \lambda_*((L-1)t^{L-2} + (L-2)t^L) - \frac{t\lambda_* + \frac{t(L-2)\epsilon^2}{2}b_* + o(\epsilon^2)}{(1+\epsilon^2)^{(L-1)/2}} + \frac{\lambda_*(2L-3)t^{L-1}}{\sqrt{1+\epsilon^2}}.$$

Since $1/(1+\epsilon^2)^k = 1 - k\epsilon^2 + o(\epsilon^2)$, the above quantity can be written as

$$\begin{aligned}
&\lambda_*\left(1 - \frac{L\epsilon^2}{2}\right) + \frac{L\epsilon^2 b_*}{2} - \lambda_*((L-1)t^{L-2} + (L-2)t^L) - t\lambda_*\left(1 - \frac{(L-1)\epsilon^2}{2}\right) \\
&- \frac{t(L-2)b_*\epsilon^2}{2} + \lambda_*(2L-3)t^{L-1}\left(1 - \frac{\epsilon^2}{2}\right) + o(\epsilon^2) \\
&= \lambda_*(1 - (L-1)t^{L-2} - (L-2)t^L - t + (2L-3)t^{L-1}) \\
&+ \frac{\epsilon^2}{2}(-L\lambda_* + Lb_* + t(L-1)\lambda_* - t(L-2)b_* - \lambda_*(2L-3)t^{L-1}) + o(\epsilon^2).
\end{aligned}$$

Now, define $a(t) = (1 + (L-2)t^{L-1} - (L-1)t^{L-2})$, then

$$(1-t)a(t) = 1 - (L-1)t^{L-2} - (L-2)t^L - t + (2L-3)t^{L-1}.$$

Also, note that $a(1) = 0$, and $a'(t) = (L-2)(L-1)(t^{L-2} - t^{L-3}) \geq 0$, for all $t \geq 1$. Hence, $a(t) \geq 0$ and an increasing function, for all $t \geq 1$.

Next, let $b(t) = -L\lambda_* + Lb_* + t(L-1)\lambda_* - t(L-2)b_* - \lambda_*(2L-3)t^{L-1}$. Thus, our goal is to that show for all sufficiently small $\epsilon > 0$,

$$(1-t)a(t)\lambda_* + b(t)\epsilon^2/2 + o(\epsilon^2) \leq 0. \tag{27}$$

We first consider the case when $t \in [1, L/(L-2)]$. In this case, since $a(t) \geq 0$, for all $t \geq 1$, we have $(1-t)a(t) \leq 0$. Next

$$\begin{aligned}
b(t) &= -L\lambda_* + (L - t(L-2))b_* + t(L-1)\lambda_* - \lambda_*(2L-3)t^{L-1} \\
&\leq -L\lambda_* + (L - t(L-2))(\lambda_* - \Delta) + t(L-1)\lambda_* - \lambda_*(2L-3)t^{L-1} \\
&= -t(L-2)\lambda_* + t(L-1)\lambda_* - \lambda_*(2L-3)t^{L-1} - (L - t(L-2))\Delta \\
&\leq t\lambda_* - \lambda_*(2L-3)t^{L-1} \leq t\lambda_*(4-2L) \leq \lambda_*(4-2L),
\end{aligned}$$

where the first inequality follows from eq. (26), and the second uses $\Delta > 0$. The last two inequalities are true since $t \geq 1$. Therefore,

$$(1-t)a(t)\lambda_* + b(t)\epsilon^2/2 + o(\epsilon^2) \leq \lambda_*(4-2L)\epsilon^2 + o(\epsilon^2).$$

Since $L > 2$, the above inequality implies eq. (27) is true for all sufficiently small $\epsilon > 0$.

We next consider $t \geq L/(L-2)$. Define $g(t) = (1-t)a(t)\lambda_* + b(t)\epsilon^2/2$. From the above discussion, we know

$$g(L/(L-2)) \leq \lambda_*(4-2L)\epsilon^2.$$

Now, if we can show that for all sufficiently small $\epsilon > 0$, $g(t) \leq g(L/(L-2))$, for all $t \geq L/(L-2)$, then

$$(1-t)a(t)\lambda_* + b(t)\epsilon^2/2 + o(\epsilon^2) \leq \lambda_*(4-2L)\epsilon^2 + o(\epsilon^2), \text{ for all } t \geq L/(L-2).$$

Thus, eq. (27) is true for all sufficiently small $\epsilon > 0$. To complete the proof, we next show that $g'(t) \leq 0$, for all $t \geq L/(L-2)$ and all sufficiently small $\epsilon > 0$. Note that

$$g'(t) = (1-t)a'(t)\lambda_* - a(t)\lambda_* + b'(t)\epsilon^2/2.$$

Since $a'(t) \geq 0$ and $t \geq 1$, $(1-t)a'(t)\lambda_* \leq 0$. Further, $a'(t) \geq 0$ implies $a(t) \geq a(L/(L-2))$, for all $t \geq L/(L-2)$. Since $a(L/L-2) = (1 + (L/L-2)^{L-2}) \geq 1$, we have that

$$\begin{aligned}
-a(t)\lambda_* + b'(t)\epsilon^2/2 &= -a(t)\lambda_* + ((L-1)\lambda_* - (L-2)b_* - (2L-3)(L-1)\lambda_* t^{L-2})\epsilon^2/2 \\
&\leq -\lambda_* - (L-2)b_*\epsilon^2/2 + (L-1)\lambda_*(1-(2L-3)t^{L-2})\epsilon^2/2 \\
&\leq -\lambda_* - (L-2)b_*\epsilon^2/2 \leq 0,
\end{aligned}$$

where the first inequality uses $a(t) \geq a(L/(L-2)) \geq 1$. The second inequality uses $L > 2$ and $t \geq 1$. The last inequality is true for all sufficiently small $\epsilon > 0$. Thus, $g'(t) \leq 0$. ■

Lemma 23 *Suppose $\mathcal{H}$ is $L-$homogeneous, where $L \geq 2$, and $\mathbf{w}_*$ is a $\Delta-$second-order positive KKT point of eq. (4). Then there exists a sufficiently small $\gamma > 0$ such that for all unit-norm vectors $\mathbf{w}$ satisfying $\mathbf{w}^\top \mathbf{w}_* \geq 1-\gamma$ the following inequalities hold:*

$$\mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{w}) - L\mathcal{N}(\mathbf{w})\mathbf{w}_*^\top \mathbf{w} - \frac{\Delta}{2}\|\mathbf{w}-\mathbf{w}_*\|_2^2 \geq 0, \text{ and} \tag{28}$$

$$\mathcal{N}(\mathbf{w}) - \mathcal{N}(\mathbf{w}_*) + \frac{\Delta}{4}\|\mathbf{w}-\mathbf{w}_*\|_2^2 \leq 0. \tag{29}$$

Proof Let $\mathbf{w} = (\mathbf{w}_* + \epsilon\mathbf{b})/\sqrt{1+\epsilon^2}$, where $\mathbf{b} \in \mathbf{w}_*^\perp$. Putting this $\mathbf{w}$ in eq. (28) gives us

$$\frac{\mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})}{(1+\epsilon^2)^{(L-1)/2}} - \frac{L\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})\mathbf{w}_*^\top(\mathbf{w}_* + \epsilon\mathbf{b})}{(1+\epsilon^2)^{(L+1)/2}} - \frac{\Delta}{2}(2 - 2\mathbf{w}^\top\mathbf{w}_*). \tag{30}$$

To complete the proof, we will show that the above quantity is greater than zero for all sufficiently small $\epsilon > 0$. Note that, for all sufficiently small $\epsilon > 0$,

$$\mathbf{w}^\top\mathbf{w}_* = 1/\sqrt{1+\epsilon^2} = 1 - \epsilon^2/2 + o(\epsilon^2),$$

and from Lemma 21, we have

$$\begin{aligned}
\nabla\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})^\top\mathbf{w}_* &= L\mathcal{N}(\mathbf{w}_*) + \frac{(L-2)\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2) \text{ and} \\
\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b}) &= \mathcal{N}(\mathbf{w}_*) + \frac{\epsilon^2}{2}\mathbf{b}^\top\nabla^2\mathcal{N}(\mathbf{w}_*)\mathbf{b} + o(\epsilon^2).
\end{aligned}$$

Let $\lambda_* = L\mathcal{N}(\mathbf{w}_*)$, $b_* = \mathbf{b}^\top \nabla^2 \mathcal{N}(\mathbf{w}_*)\mathbf{b}$, then eq. (30) can be simplified to get

$$\begin{aligned}
&\lambda_* \left(1 - \frac{(L-1)\epsilon^2}{2}\right) + \frac{(L-2)\epsilon^2}{2} b_* + o(\epsilon^2) - \lambda_* \left(1 - \frac{(L+1)\epsilon^2}{2}\right) - \frac{L\epsilon^2}{2} b_* - \frac{\Delta\epsilon^2}{2} \\
&= \frac{\epsilon^2}{2}\left(-(L-1)\lambda_* + (L-2)b_* + (L+1)\lambda_* - Lb_* - \Delta\right) + o(\epsilon^2) \\
&= \frac{\epsilon^2}{2}\left(2\lambda_* - 2b_* - \Delta\right) + o(\epsilon^2) \geq \Delta\epsilon^2/2 + o(\epsilon^2).
\end{aligned}$$

The last inequality uses $b_* \leq \lambda_* - \Delta$. Since $\Delta > 0$, the above inequality proves eq. (28). Similarly, eq. (29) can be written as

$$\begin{aligned}
&\frac{\mathcal{N}(\mathbf{w}_* + \epsilon\mathbf{b})}{(1+\epsilon^2)^{L/2}} - \mathcal{N}(\mathbf{w}_*) + \frac{\Delta(2 - 2\mathbf{w}^\top \mathbf{w}_*)}{4} \\
&= \mathcal{N}(\mathbf{w}_*)\left(1 - \frac{L\epsilon^2}{2}\right) + \frac{\epsilon^2}{2}\mathbf{b}^\top \nabla^2 \mathcal{N}(\mathbf{w}_*)\mathbf{b} - \mathcal{N}(\mathbf{w}_*) + \frac{\Delta\epsilon^2}{4} + o(\epsilon^2) \\
&= \frac{\epsilon^2}{2}(b_* - L\mathcal{N}(\mathbf{w}_*) + \Delta/2) + o(\epsilon^2) \leq \frac{-\Delta\epsilon^2}{4} + o(\epsilon^2).
\end{aligned}$$

The last inequality uses $b_* \leq \lambda_* - \Delta$. Since $\Delta > 0$, the above inequality proves eq. (29). ■

A.2 Gradient Flow Dynamics of NCF Near a KKT point

In this subsection, we describe the gradient flow dynamics of NCF when initialized near a second-order positive KKT point.

Lemma 24 *Suppose $\mathcal{H}$ is $2-$homogeneous, and $\mathbf{w}_*$ is a $\Delta-$second-order positive KKT point of the constrained NCF in eq. (4). Let $\mathbf{w}(t)$ be the solution of*

$$\dot{\mathbf{w}} = \nabla\mathcal{N}(\mathbf{w}), \mathbf{w}(0) = \mathbf{w}_0,$$

where $\|\mathbf{w}_0\|_2 = 1$. There exists a sufficiently small $\gamma > 0$ such that if $\mathbf{w}_*^\top \mathbf{w}_0 > 1 - \gamma$, then

$$\frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2} \geq 1 - e^{-t\Delta}\gamma, \forall t \geq 0, \text{ and } \mathbf{w}(t) = \mathbf{g}(t)e^{2t\mathcal{N}(\mathbf{w}_*)}, \tag{31}$$

where $\|\mathbf{g}(t)\|_2 \in [\kappa_1, \kappa_2]$, for some $\kappa_2 \geq \kappa_1 > 0$.

Proof We chose a sufficiently small $\gamma > 0$ such that

$$\mathcal{N}(\mathbf{w}_*) \geq \mathcal{N}(\mathbf{w}) > 0 \text{ and } \mathbf{w}_*^\top \nabla\mathcal{N}(\mathbf{w}) - 2\mathcal{N}(\mathbf{w})\mathbf{w}_*^\top \mathbf{w} - \frac{\Delta}{2}\|\mathbf{w} - \mathbf{w}_*\|_2^2 \geq 0, \tag{32}$$

for all unit-norm vector $\mathbf{w}$ satisfying $\mathbf{w}_*^\top \mathbf{w} > 1 - \gamma$. The above conditions are possible since $\mathcal{N}(\mathbf{w}_*) > 0$ and due to Lemma 23.

Now, since $\mathbf{w}_0^\top \mathbf{w}_* > 1 - \gamma$, we have $\mathcal{N}(\mathbf{w}_0) > 0$. From (Kumar and Haupt, 2024, Lemma C.4) , $\|\mathbf{w}(t)\|_2 \geq \|\mathbf{w}_0\|_2 > 0$, for all $t \geq 0$. We next show that $\mathbf{w}_*^\top \mathbf{w}(t)/\|\mathbf{w}(t)\|_2 > 1 - \gamma$,

for all $t \geq 0$. For the sake of contradiction, suppose $\overline{T} > 0$ be the smallest time such that $\mathbf{w}_*^\top \mathbf{w}(\overline{T})/\|\mathbf{w}(\overline{T})\|_2 = 1-\gamma$. Then, for all $t \in [0, \overline{T})$, we have

$$\begin{aligned}
\frac{d}{dt}\left(\frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) &= \mathbf{w}_*^\top \left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_2^2}\right)\frac{\nabla \mathcal{N}(\mathbf{w})}{\|\mathbf{w}\|_2} \\
&= \mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{w})/\|\mathbf{w}\|_2 - 2\mathbf{w}_*^\top \mathbf{w}\mathcal{N}(\mathbf{w})/\|\mathbf{w}\|_2^3 \\
&= \mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{w}/\|\mathbf{w}\|_2) - 2\mathbf{w}_*^\top (\mathbf{w}/\|\mathbf{w}\|_2)\mathcal{N}(\mathbf{w}/\|\mathbf{w}\|_2) \\
&\geq \frac{\Delta}{2}\left\|\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|_2} - \mathbf{w}_*\right\|_2^2 = \Delta\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) \geq 0, \qquad (33)
\end{aligned}$$

where the first inequality follows from eq. (32). The above equation implies

$$\left(\frac{\mathbf{w}_*^\top \mathbf{w}(\overline{T})}{\|\mathbf{w}(\overline{T})\|_2}\right) \geq \left(\frac{\mathbf{w}_*^\top \mathbf{w}(0)}{\|\mathbf{w}(0)\|_2}\right) > 1-\gamma,$$

which leads to a contradiction. Now, since $\mathbf{w}_*^\top \mathbf{w}(t)/\|\mathbf{w}(t)\|_2 > 1-\gamma$, for all $t \geq 0$, from eq. (33) we have

$$\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) \leq e^{-t\Delta}\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(0)}{\|\mathbf{w}(0)\|_2}\right) \leq e^{-t\Delta}\gamma,$$

which proves the first part of eq. (31). Now, if $\mathbf{w}(t) = \mathbf{g}(t)e^{2t\mathcal{N}(\mathbf{w}_*)}$, then

$$\nabla \mathcal{N}(\mathbf{g})e^{2t\mathcal{N}(\mathbf{w}_*)} = \nabla \mathcal{N}(\mathbf{w}) = \dot{\mathbf{w}} = \dot{\mathbf{g}}e^{2t\mathcal{N}(\mathbf{w}_*)} + 2\mathcal{N}(\mathbf{w}_*)\mathbf{g}(t)e^{2t\mathcal{N}(\mathbf{w}_*)},$$

which implies

$$\dot{\mathbf{g}} + 2\mathcal{N}(\mathbf{w}_*)\mathbf{g} = \nabla \mathcal{N}(\mathbf{g}).$$

Multiplying the above equation by $\mathbf{g}^\top$ from the left, we get

$$\frac{1}{2}\frac{d\|\mathbf{g}\|_2^2}{dt} + 2\mathcal{N}(\mathbf{w}_*)\|\mathbf{g}\|_2^2 = 2\mathcal{N}(\mathbf{g}) = 2\mathcal{N}(\mathbf{g}/\|\mathbf{g}\|_2)\|\mathbf{g}\|_2^2,$$

which implies

$$\|\mathbf{g}(t)\|_2^2 = \|\mathbf{g}(0)\|_2^2 e^{4\int_0^t (\mathcal{N}(\mathbf{g}(s)/\|\mathbf{g}(s)\|_2) - \mathcal{N}(\mathbf{w}_*))ds}.$$

Since $\mathcal{N}(\mathbf{g}(t)/\|\mathbf{g}(t)\|_2) = \mathcal{N}(\mathbf{w}(t)/\|\mathbf{w}(t)\|_2) \leq \mathcal{N}(\mathbf{w}_*)$, for all $t \geq 0$, we have

$$\|\mathbf{g}(t)\|_2^2 \leq \|\mathbf{g}(0)\|_2^2 := \kappa_2.$$

Next, since $\mathcal{N}(\mathbf{w})$ is locally Lipschitz, there exists a $\mu > 0$ such that, for all $t \geq 0$,

$$\mathcal{N}(\mathbf{w}_*) - \mathcal{N}(\mathbf{g}(t)/\|\mathbf{g}(t)\|_2) = |\mathcal{N}(\mathbf{w}_*) - \mathcal{N}(\mathbf{g}(t)/\|\mathbf{g}(t)\|_2)| \leq \mu\|\mathbf{g}(t)/\|\mathbf{g}(t)\|_2 - \mathbf{w}_*\|_2,$$

therefore, for $t \geq 0$, using the above equation and first part of eq. (31), we get

$$\begin{aligned}
\int_0^t (\mathcal{N}(\mathbf{g}(s)/\|\mathbf{g}(s)\|_2) - \mathcal{N}(\mathbf{w}_*))\, ds &\geq \int_0^t -\mu\|\mathbf{g}(s)/\|\mathbf{g}(s)\|_2 - \mathbf{w}_*\|_2 ds \\
&= \int_0^t -\mu\sqrt{\|\mathbf{w}(s)/\|\mathbf{w}(s)\|_2 - \mathbf{w}_*\|_2^2}ds \\
&= \int_0^t -\mu\sqrt{2 - 2\mathbf{w}_*^\top \mathbf{w}(s)/\|\mathbf{w}(s)\|_2}ds \\
&\geq -\mu\sqrt{2\gamma}\int_0^t e^{-s\Delta/2}ds \geq -2\mu\sqrt{2\gamma}/\Delta.
\end{aligned}$$

Hence, $\|\mathbf{g}(t)\|_2^2 \geq \|\mathbf{g}(0)\|_2^2 e^{-8\mu\sqrt{2\gamma}/\Delta} := \kappa_1$, which completes the proof. ■

Lemma 25 *Suppose $\mathcal{H}$ is $L-$homogeneous, where $L > 2$, and $\mathbf{w}_*$ is a $\Delta-$second-order positive KKT point of eq. (4). Let $\mathbf{w}(t)$ be the solution of*

$$\dot{\mathbf{w}} = \nabla\mathcal{N}(\mathbf{w}), \mathbf{w}(0) = \mathbf{w}_0,$$

where $\|\mathbf{w}_0\|_2 = 1$. There exists a sufficiently small $\gamma > 0$ such that if $\mathbf{w}_*^\top \mathbf{w}_0 > 1 - \gamma$, then $\mathcal{N}(\mathbf{w}_0) > 0$ and there exists a finite $T > 0$ such that

$$\lim_{t\to T} \|\mathbf{w}(t)\|_2 = \infty, \text{ where } T \in \left[\frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_0)}\right], \text{ and}$$

$$\frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2} \geq 1 - \gamma(1 - t/T)^{\frac{\Delta}{L(L-2)\mathcal{N}(\mathbf{w}_*)}}.$$

Further, for all $t \in [0, T)$, $\|\mathbf{w}(t)\|_2$ is an increasing function and

$$\mathbf{w}(t) = \frac{\mathbf{g}(t)}{(T-t)^{1/(L-2)}},$$

where $\|\mathbf{g}(t)\|_2$ is a decreasing function for all $t \in [0, T)$,

$$\|\mathbf{g}(0)\|_2^{L-2} = T, \text{ and } \lim_{t\to T} \|\mathbf{g}(t)\|_2^{L-2} = 1/(L(L-2)\mathcal{N}(\mathbf{w}_*)).$$

Proof We begin by choosing a sufficiently small $\gamma > 0$ such that

$$\mathcal{N}(\mathbf{w}_*) \geq \mathcal{N}(\mathbf{w}) > 0, \text{ and } \mathbf{w}_*^\top \nabla\mathcal{N}(\mathbf{w}) - L\mathcal{N}(\mathbf{w})\mathbf{w}_*^\top \mathbf{w} - \frac{\Delta}{2}\|\mathbf{w} - \mathbf{w}_*\|_2^2 \geq 0, \tag{34}$$

for all unit norm vector $\mathbf{w}$ satisfying $\mathbf{w}_*^\top \mathbf{w} \geq 1 - \gamma$. The above conditions are possible since $\mathcal{N}(\mathbf{w}_*) > 0$ and due to Lemma 23.

Now, since $\mathbf{w}_0^\top \mathbf{w}_* > 1 - \gamma$, we have $\mathcal{N}(\mathbf{w}_0) > 0$. From (Kumar and Haupt, 2025, Lemma 13), $\|\mathbf{w}(t)\|_2 \geq \|\mathbf{w}_0\|_2 > 0$, for all $t \geq 0$. Next, we show that $\|\mathbf{w}(t)\|_2$ becomes unbounded at some finite time. For this, define

$$\widetilde{\mathcal{N}}(\mathbf{w}) \coloneqq \mathcal{N}\left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right) = \frac{\mathcal{N}(\mathbf{w})}{\|\mathbf{w}\|_2^L}, \text{ then } \nabla\widetilde{\mathcal{N}}(\mathbf{w}) = \left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_2^2}\right)\frac{\nabla\mathcal{N}(\mathbf{w})}{\|\mathbf{w}\|_2^L}.$$

Now, note that

$$\frac{d\widetilde{\mathcal{N}}(\mathbf{w})}{dt} = \dot{\mathbf{w}}^\top \nabla\widetilde{\mathcal{N}}(\mathbf{w}) = \nabla\mathcal{N}(\mathbf{w})\left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_2^2}\right)\frac{\nabla\mathcal{N}(\mathbf{w})}{\|\mathbf{w}\|_2^L} \geq 0,$$

which implies $\widetilde{\mathcal{N}}(\mathbf{w}(t))$ increases with time. Hence,

$$\frac{1}{2}\frac{d\|\mathbf{w}\|_2^2}{dt} = L\mathcal{N}(\mathbf{w}) = L\|\mathbf{w}\|_2^L \widetilde{\mathcal{N}}(\mathbf{w}) \geq L\|\mathbf{w}\|_2^L \mathcal{N}(\mathbf{w}_0) \geq 0, \tag{35}$$

which implies $\|\mathbf{w}(t)\|_2$ is an increasing function, for all $t \geq 0$, and

$$\frac{d\|\mathbf{w}\|_2}{dt} \geq L\|\mathbf{w}\|_2^{L-1}\mathcal{N}(\mathbf{w}_0).$$

Taking $\|\mathbf{w}\|_2^{L-1}$ to the LHS and integrating from 0 to t, we get

$$\frac{1}{L-2}\left(\frac{1}{\|\mathbf{w}(0)\|_2^{L-2}} - \frac{1}{\|\mathbf{w}(t)\|_2^{L-2}}\right) \geq L\mathcal{N}(\mathbf{w}_0)t.$$

Using $\|\mathbf{w}(0)\|_2 = 1$, and simplifying the above equation we get

$$\|\mathbf{w}(t)\|_2^{L-2} \geq \frac{1}{1 - tL(L-2)\mathcal{N}(\mathbf{w}_0)}.$$

The above equation implies that for some $T \leq 1/(L(L-2)\mathcal{N}(\mathbf{w}_0))$, $\|\mathbf{w}(T)\|_2 = \infty$. (We will show $T \geq 1/L(L-2)\mathcal{N}(\mathbf{w}_*)$ later.)

We next derive related results for $\mathbf{g}(t)$. We know

$$\frac{1}{2}\frac{d\|\mathbf{w}\|_2^2}{dt} = L\mathcal{N}(\mathbf{w}) = L\mathcal{N}(\mathbf{w}/\|\mathbf{w}\|_2)\|\mathbf{w}\|_2^L,$$

which implies

$$\frac{1}{\|\mathbf{w}\|_2^{L-1}}\frac{d\|\mathbf{w}\|_2}{dt} = L\mathcal{N}(\mathbf{w}/\|\mathbf{w}\|_2).$$

Integrating both sides from 0 to $t \in (0, T)$, we get

$$\frac{1}{L-2}\left(\frac{1}{\|\mathbf{w}(0)\|_2^{L-2}} - \frac{1}{\|\mathbf{w}(t)\|_2^{L-2}}\right) = \int_0^t L\mathcal{N}(\mathbf{w}(s)/\|\mathbf{w}(s)\|_2)ds.$$

Using $\|\mathbf{w}(0)\|_2 = 1$ and re-arranging the above equation gives us

$$1 - \int_0^t L(L-2)\mathcal{N}(\mathbf{w}(s)/\|\mathbf{w}(s)\|_2)ds = \frac{1}{\|\mathbf{w}(t)\|_2^{L-2}}. \tag{36}$$

Substituting $\mathbf{w}(t) = \mathbf{g}(t)/(T-t)^{1/(L-2)}$, we get

$$\frac{1 - \int_0^t L(L-2)\mathcal{N}(\mathbf{w}(s)/\|\mathbf{w}(s)\|_2)ds}{(T-t)} = \frac{1}{\|\mathbf{g}(t)\|_2^{L-2}}.$$

Note that $\|\mathbf{g}(0)\|_2^{L-2} = T$. Next, we compute $\|\mathbf{g}(T)\|_2$. In the LHS of the above equality, at $t = T$, the denominator is obviously 0, and the numerator is also 0 since, using eq. (36),

$$1 - \int_0^T L(L-2)\mathcal{N}(\mathbf{w}(s)/\|\mathbf{w}(s)\|_2)ds = \lim_{t\to T}\frac{1}{\|\mathbf{w}(t)\|_2^{L-2}} = 0.$$

Therefore, using L'Hopital's rule,

$$\frac{1}{\|\mathbf{g}(T)\|_2^{L-2}} = \frac{\lim_{t\to T} -L(L-2)\mathcal{N}(\mathbf{w}(t)/\|\mathbf{w}(t)\|_2)}{-1} = \lim_{t\to T} L(L-2)\mathcal{N}(\mathbf{w}(t)/\|\mathbf{w}(t)\|_2).$$

We will soon show that $\lim_{t\to T} \mathbf{w}(t)/\|\mathbf{w}(t)\|_2 = \mathbf{w}_*$, which will complete the result for $\|\mathbf{g}(T)\|_2^{L-2}$. We now prove that $\|\mathbf{g}(t)\|_2$ is a decreasing function. For this, define

$$h(t) := \frac{1 - \int_0^t L(L-2)\mathcal{N}(\mathbf{w}(s)/\|\mathbf{w}(s)\|_2)ds}{(T-t)} = \frac{1 - \int_0^t L(L-2)\widetilde{\mathcal{N}}(\mathbf{w}(s))ds}{(T-t)}.$$

We next show that $h'(t) \geq 0$, for all $t \in (0,T)$. This would imply $h(t)$ is increasing and therefore, $\|\mathbf{g}(t)\|_2$ is decreasing, for all for all $t \in (0,T)$. For all $t \in [0,T)$,

$$\begin{aligned} h'(t) &= \frac{-L(L-2)\widetilde{\mathcal{N}}(\mathbf{w}(t))(T-t) + (1 - \int_0^t L(L-2)\widetilde{\mathcal{N}}(\mathbf{w}(s))ds)}{(T-t)^2} \\ &= \frac{1}{(T-t)}\left(-L(L-2)\widetilde{\mathcal{N}}(\mathbf{w}(t)) + h(t)\right). \end{aligned}$$

Note that

$$h(0) = 1/T \geq 0, \text{ and } h'(0) = \frac{1}{T}\left(-L(L-2)\mathcal{N}(\mathbf{w}_0) + \frac{1}{T}\right) \geq 0,$$

where the last inequality follows since $T \leq 1/(L(L-2)\mathcal{N}(\mathbf{w}_0))$. For the sake of contradiction, suppose there exists a $t_1 \in (0,T)$ such that $h'(t_1) = -\epsilon$, for some $\epsilon > 0$. Next, we compute $h''(t)$. Let $a(t)$ denote the numerator of $h'(t)$, then

$$a'(t) = -L(L-2)\nabla\mathcal{N}(\mathbf{w}(t))^\top \left(\mathbf{I} - \frac{\mathbf{w}(t)\mathbf{w}(t)^\top}{\|\mathbf{w}(t)\|_2^2}\right)\frac{\nabla\mathcal{N}(\mathbf{w}(t))}{\|\mathbf{w}\|_2^L} + h'(t).$$

Thus,

$$\begin{aligned} h''(t) &= \frac{a'(t)}{(T-t)} + \frac{a(t)}{(T-t)^2} = \frac{1}{T-t}\left(a'(t) + \frac{a(t)}{(T-t)}\right) = \frac{1}{T-t}\left(a'(t) + h'(t)\right) \\ &= \frac{1}{(T-t)}\left(-L(L-2)\nabla\mathcal{N}(\mathbf{w}(t))^\top \left(\mathbf{I} - \frac{\mathbf{w}(t)\mathbf{w}(t)^\top}{\|\mathbf{w}(t)\|_2^2}\right)\frac{\nabla\mathcal{N}(\mathbf{w}(t))}{\|\mathbf{w}\|_2^L} + 2h'(t)\right) \end{aligned}$$

Since $-L(L-2)\nabla\mathcal{N}(\mathbf{w}(t))^\top \left(\mathbf{I} - \frac{\mathbf{w}(t)\mathbf{w}(t)^\top}{\|\mathbf{w}(t)\|_2^2}\right)\frac{\nabla\mathcal{N}(\mathbf{w}(t))}{\|\mathbf{w}\|_2^L} \leq 0$, the above equation implies

$$h''(t) - \frac{2h'(t)}{(T-t)} \leq 0.$$

From Lemma 29, for any $t \in (t_1, T)$, we get

$$h'(t) \leq h'(t_1)/P(t), \text{ where } P(t) = e^{-\int_{t_1}^t 2/(T-s)ds} = (T-t)^2/(T-t_1)^2.$$

Hence, for any $t \in (t_1, T)$,

$$h'(t) \leq -\epsilon(T-t_1)^2/(T-t)^2.$$

Integrating the above equation from t_1 to $t \in (t_1, T)$, we get

$$h(t) - h(t_1) \leq -\epsilon(T-t_1)^2\left(\frac{1}{T-t} - \frac{1}{T-t_1}\right).$$

Now, if t is chosen sufficiently close to T, then $h(t)$ becomes negative. This is impossible since $h(t) = 1/\|\mathbf{g}(t)\|_2^{L-2}$, leading to a contradiction. Hence, $h'(t) \geq 0$, for all $t \in (0, T)$.

We next show that $\mathbf{w}_*^\top \mathbf{w}(t)/\|\mathbf{w}(t)\|_2 > 1-\gamma$, for all $t \in [0, T)$. For the sake of contradiction, suppose $\overline{T} \in [0, T)$ is the smallest time such that $\mathbf{w}_*^\top \mathbf{w}(\overline{T})/\|\mathbf{w}(\overline{T})\|_2 = 1-\gamma$. Then, for all $t \in [0, \overline{T})$,

$$\begin{aligned}
\frac{d}{dt}\frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2} &= \mathbf{w}_*^\top \left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_2^2}\right)\frac{\nabla \mathcal{N}(\mathbf{w})}{\|\mathbf{w}\|_2} \\
&= \mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{w})/\|\mathbf{w}\|_2 - L\mathbf{w}_*^\top \mathbf{w}\mathcal{N}(\mathbf{w})/\|\mathbf{w}\|_2^3 \\
&= \|\mathbf{w}\|_2^{L-2}\left(\mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{w}/\|\mathbf{w}\|_2) - L\mathbf{w}_*^\top (\mathbf{w}/\|\mathbf{w}\|_2)\mathcal{N}(\mathbf{w}/\|\mathbf{w}\|_2)\right) \\
&\geq \|\mathbf{w}\|_2^{L-2}\frac{\Delta}{2}\left\|\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|_2} - \mathbf{w}_*\right\|_2^2 = \|\mathbf{w}\|_2^{L-2}\Delta\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) \geq 0, \qquad (37)
\end{aligned}$$

where the first inequality follows from eq. (34). The above equation implies

$$\left(\frac{\mathbf{w}_*^\top \mathbf{w}(\overline{T})}{\|\mathbf{w}(\overline{T})\|_2}\right) \geq \left(\frac{\mathbf{w}_*^\top \mathbf{w}(0)}{\|\mathbf{w}(0)\|_2}\right) > 1-\gamma,$$

which leads to a contradiction. Now, from (Kumar and Haupt, 2025, Lemma 2), we know

$$\lim_{t\to T} \mathbf{w}(t)/\|\mathbf{w}(t)\|_2 = \overline{\mathbf{w}},$$

where $\overline{\mathbf{w}}$ is a first-order KKT point of eq. (4). We next show that $\overline{\mathbf{w}} = \mathbf{w}_*$. Since $\overline{\mathbf{w}}$ is a first-order KKT point, from Lemma 20, we know $L\mathcal{N}(\overline{\mathbf{w}})\overline{\mathbf{w}} = \nabla \mathcal{N}(\overline{\mathbf{w}})$. Also, since $\overline{\mathbf{w}}^\top \mathbf{w}^* \geq 1-\gamma$, from eq. (34), we have

$$0 \leq \mathbf{w}_*^\top \nabla \mathcal{N}(\overline{\mathbf{w}}) - L\mathcal{N}(\overline{\mathbf{w}})\mathbf{w}_*^\top \overline{\mathbf{w}} - \frac{\Delta}{2}\|\overline{\mathbf{w}} - \mathbf{w}_*\|_2^2 = -\frac{\Delta}{2}\|\overline{\mathbf{w}} - \mathbf{w}_*\|_2^2,$$

implying $\overline{\mathbf{w}} = \mathbf{w}_*$. Hence, $\lim_{t\to T} \|\mathbf{g}(t)\|_2^{L-2} = 1/(L(L-2)\mathcal{N}(\mathbf{w}_*)) \geq \|\mathbf{g}(0)\|_2^{L-2} = 1/T$. Finally, since $\|\mathbf{g}(t)\|_2$ is a decreasing function, $\|\mathbf{w}(t)\|_2^{L-2} = \|\mathbf{g}(t)\|_2^{L-2}/(T-t) \geq \|\mathbf{g}(T)\|_2^{L-2}/(T-t)$. Let $\alpha := \|\mathbf{g}(T)\|_2^{L-2}$, then, using eq. (37), we have

$$\frac{d}{dt}\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) \leq -\|\mathbf{w}\|_2^{L-2}\Delta\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) \leq -\frac{\alpha\Delta}{(T-t)}\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right).$$

Integrating the above equation from 0 to $t \in (0, T)$, we get

$$\left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|_2}\right) \leq \left(1 - \frac{\mathbf{w}_*^\top \mathbf{w}(0)}{\|\mathbf{w}(0)\|_2}\right) e^{-\int_0^t \frac{\alpha\Delta}{(T-s)}ds} \leq \gamma\left(1 - \frac{t}{T}\right)^{\alpha\Delta},$$

which completes the proof. ■

Lemma 26 *Let $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$. Then, for any fixed $\widetilde{T} > 0$, there exists a $\widetilde{C} > 0$ such that*

$$\|\boldsymbol{\psi}(t, \mathbf{p}) - \boldsymbol{\psi}(t, \mathbf{q})\|_2 \leq \widetilde{C}\|\mathbf{p} - \mathbf{q}\|_2, \textit{for all } t \in [-\widetilde{T}, \widetilde{T}].$$

Proof Let $\mathbf{u}_1(t) = \boldsymbol{\psi}(t, \mathbf{p})$ and $\mathbf{u}_2(t) = \boldsymbol{\psi}(t, \mathbf{q})$. Then,

$$\dot{\mathbf{u}}_1 = -\nabla\mathcal{L}(\mathbf{u}_1), \mathbf{u}_1(0) = \mathbf{p}, \dot{\mathbf{u}}_2 = -\nabla\mathcal{L}(\mathbf{u}_2), \mathbf{u}_2(0) = \mathbf{q}.$$

Since $\mathcal{L}(\cdot)$ has locally Lipshitz gradient, and $\widetilde{T}$ is fixed, we may assume that there exists a μ such that

$$\|\nabla\mathcal{L}(\mathbf{u}_1(t)) - \nabla\mathcal{L}(\mathbf{u}_2(t))\|_2 \leq \mu\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2, \text{ for all } t \in [-\widetilde{T}, \widetilde{T}].$$

Hence, for any $t \in [0, \widetilde{T}]$, we have

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2^2 = -(\mathbf{u}_1 - \mathbf{u}_2)^\top(\nabla\mathcal{L}(\mathbf{u}_1) - \nabla\mathcal{L}(\mathbf{u}_2)) \leq \mu\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2^2.$$

Integrating the above inequality from 0 to $t \in [0, \widetilde{T}]$, we get

$$\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2 \leq e^{\mu t}\|\mathbf{u}_1(0) - \mathbf{u}_2(0)\|_2 \leq e^{\mu\widetilde{T}}\|\mathbf{u}_1(0) - \mathbf{u}_2(0)\|_2. \tag{38}$$

Similarly, for any $t \in [-\widetilde{T}, 0]$, we have

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2^2 = -(\mathbf{u}_1 - \mathbf{u}_2)^\top(\nabla\mathcal{L}(\mathbf{u}_1) - \nabla\mathcal{L}(\mathbf{u}_2)) \geq -\mu\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2^2.$$

Integrating the above inequality from $t \in [-\widetilde{T}, 0]$ to 0, we get

$$\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\|_2 \leq e^{-\mu t}\|\mathbf{u}_1(0) - \mathbf{u}_2(0)\|_2 \leq e^{\mu\widetilde{T}}\|\mathbf{u}_1(0) - \mathbf{u}_2(0)\|_2. \tag{39}$$

Combining eq. (38) and eq. (39) completes the proof. ■

A.3 Auxiliary Lemmata

In this subsection, we provide auxiliary lemmata which are crucial for the proofs.

Lemma 27 *Suppose $\eta_1 > \eta_2 > 0$, then $f_1(t) = \frac{1-\eta_1 t}{1-\eta_2 t}$ and $f_2(t) = \frac{1}{1-\eta_1 t}$ are monotonically decreasing and increasing functions, respectively, on $t \in [0, 1/\eta_1)$.*

Proof The claim follows since, for $t \in [0, 1/\eta_1)$,

$$f_1'(t) = \frac{-\eta_1(1-\eta_2 t) + \eta_2(1-\eta_1 t)}{(1-\eta_2 t)^2} = \frac{\eta_2 - \eta_1}{(1-\eta_2 t)^2} \leq 0, \text{ and } f_2'(t) = \frac{\eta_1}{(1-\eta_1 t)^2} \geq 0.$$

■

Lemma 28 *If $g(t) = b/(1-at)$, for some $a > 0$. Then, for all $t \in [0, 1/a)$, $e^{-\int_0^t g(s)ds} = (1-at)^{b/a}$.*

Proof Since $\int_0^t g(s)ds = -b\ln(1-at)/a$, we have $e^{-\int_0^t g(s)ds} = (1-at)^{b/a}$. ■

Lemma 29 *If $x(t)$ satisfies $\dot{x} \leq g(t)x + h(t)$, then*

$$x(t) \leq \frac{1}{P(t)}\left(x(0) + \int_0^t P(s)h(s)ds\right), \text{ where } P(t) = e^{-\int_0^t g(s)ds}.$$

Proof Multiplying by $e^{-\int_0^t g(s)ds}$ on both sides, we get

$$e^{-\int_0^t g(s)ds}h(t) \geq e^{-\int_0^t g(s)ds}\dot{x} - e^{-\int_0^t g(s)ds}g(t)x = \frac{d}{dt}\left(e^{-\int_0^t g(s)ds}x\right).$$

Let $P(t) = e^{-\int_0^t g(s)ds}$, then integrating both sides from 0 to t, we get

$$\int_0^t P(s)h(s)ds \geq P(t)x(t) - x(0),$$

which can be rearranged to get the desired result. ■

Appendix B. Proofs Omitted from Section 3.1

This section contains the proof of Lemma 7, 8 and 9, which were used to prove Theorem 4.
Proof of Lemma 9: We choose $\gamma > 0$ sufficiently small such that for all unit-norm vector $\mathbf{w}$ satisfying $\mathbf{w}^\top \mathbf{w}_* \geq 1 - \gamma$, we have

$$\mathcal{N}(\mathbf{w}) \leq \mathcal{N}(\mathbf{w}_*), \|\nabla^2\mathcal{N}(\mathbf{w})\|_2 \leq 3\mathcal{N}(\mathbf{w}_*) \text{ and} \tag{40}$$

$$\left(\nabla\mathcal{N}(t_1\mathbf{w}) - \nabla\mathcal{N}(t_2\mathbf{w}_*)\right)^\top (t_1\mathbf{w} - t_2\mathbf{w}_*) \leq 2\mathcal{N}(\mathbf{w}_*)\|t_1\mathbf{w} - t_2\mathbf{w}_*\|_2^2, \forall t_2 \geq t_1 \geq 0. \tag{41}$$

The first inequality follows from Lemma 23. The second inequality holds since, from Lemma 20, $\|\nabla^2\mathcal{N}(\mathbf{w}_*)\|_2 = 2\mathcal{N}(\mathbf{w}_*)$, and $\nabla^2\mathcal{N}(\mathbf{w})$ is continuous in the neighborhood of $\mathbf{w}_*$. The third inequality follows from Lemma 22.

We further define $f(\mathbf{w}) = \mathcal{J}(\mathbf{X};\mathbf{w})^\top(\ell'(\mathcal{H}(\mathbf{X};\mathbf{w}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y}))$. Then, note that

$$\|f(\mathbf{w})\|_2 \leq \|\mathcal{J}(\mathbf{X};\mathbf{w})\|_2\|\ell'(\mathcal{H}(\mathbf{X};\mathbf{w}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y})\|_2 \leq Kn\|\mathcal{J}(\mathbf{X};\mathbf{w})\|_2\|\mathcal{H}(\mathbf{X};\mathbf{w})\|_2,$$

where the first inequality follows from Cauchy-Schwartz inequality, and the second follows from the smoothness of the loss function (see Assumption 2). Note that the RHS in the above inequality is $3-$homogeneous in $\mathbf{w}$. Thus, there exists a $\beta > 0$ such that

$$\|f(\mathbf{w})\|_2 \leq \beta\|\mathbf{w}\|_2^3, \text{ for all } \mathbf{w} \in \mathbb{R}^k. \tag{42}$$

Also, $f(\mathbf{w})$ is continuously differentiable in the neighborhood of $\mathbf{w}_*$. Thus, we may assume that for all vector $\mathbf{w} \in \mathbb{R}^k$ that satisfy $\mathbf{w}_*^\top\mathbf{w}/\|\mathbf{w}\|_2 \geq 1 - \gamma$, we have

$$\begin{aligned}\|\nabla f(\mathbf{w})\|_2 &\leq \sum_{i=1}^n \left\|\nabla^2\mathcal{H}(\mathbf{x}_i;\mathbf{w})\right\|_2 \left|\ell'(\mathcal{H}(\mathbf{x}_i;\mathbf{w}),y_i) - \ell'(0,y_i)\right| + \\ &\qquad \left\|\nabla\mathcal{H}(\mathbf{x}_i;\mathbf{w})\nabla\mathcal{H}(\mathbf{x}_i;\mathbf{w})^\top\right\|_2 \left|\ell''(\mathcal{H}(\mathbf{x}_i;\mathbf{w}),y_i)\right| \\ &\leq K\sum_{i=1}^n \|\nabla^2\mathcal{H}(\mathbf{x}_i;\mathbf{w})\|_2|\mathcal{H}(\mathbf{x}_i;\mathbf{w})| + \|\nabla\mathcal{H}(\mathbf{x}_i;\mathbf{w})\nabla\mathcal{H}(\mathbf{x}_i;\mathbf{w})^\top\|_2,\end{aligned}$$

where the second inequality uses smoothness of the loss function. Note that the final upper bound in the above equation is 2−homogeneous. Thus, there exists a $\zeta > 0$ such that for all vectors $\mathbf{w} \in \mathbb{R}^k$ that satisfy $\mathbf{w}_*^\top \mathbf{w}/\|\mathbf{w}\|_2 \geq 1-\gamma$, we have

$$\|\nabla f(\mathbf{w})\|_2 \leq \zeta \|\mathbf{w}\|_2^2.$$

Furthermore, from the mean value theorem, we have

$$\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\|_2 \leq \|\nabla f(\tilde{\mathbf{w}})\|_2 \|\mathbf{w}_1 - \mathbf{w}_2\|_2 \leq \zeta \max(\|\mathbf{w}_2\|_2^2, \|\mathbf{w}_1\|_2^2)\|\mathbf{w}_1 - \mathbf{w}_2\|_2, \quad (43)$$

where $\mathbf{w}_*^\top \mathbf{w}_1/\|\mathbf{w}_1\|_2, \mathbf{w}_*^\top \mathbf{w}_2/\|\mathbf{w}_2\|_2 \geq 1-\gamma$.

Now, let $\mathbf{q}(t) = \boldsymbol{\psi}(t, \mathbf{a}_\delta)$, $\mathbf{u}(t) = \boldsymbol{\psi}(t, \delta \mathbf{w}_*)$, then $\|\mathbf{q}(0) - \mathbf{u}(0)\|_2 = \|\mathbf{a}_\delta - \delta \mathbf{w}_*\|_2 \leq C_1 \delta^3$, which implies $\|\mathbf{q}(0)\|_2 \leq \delta + C_1\delta^3$. Therefore, for all sufficiently small $\delta > 0$, we have

$$\begin{aligned}\frac{\mathbf{w}_*^\top \mathbf{q}(0)}{\|\mathbf{q}(0)\|_2} &= \frac{\mathbf{w}_*^\top \mathbf{u}(0) + \mathbf{w}_*^\top(\mathbf{q}(0) - \mathbf{u}(0))}{\|\mathbf{q}(0)\|_2} \\ &\geq \frac{\delta - \|\mathbf{q}(0) - \mathbf{u}(0)\|_2}{\|\mathbf{q}(0)\|_2} \geq \frac{\delta - C_1\delta^3}{\|\mathbf{q}_\delta(0)\|_2} \geq \frac{\delta - C_1\delta^3}{\delta + C_1\delta^3} > 1 - \gamma.\end{aligned}$$

Define

$$\overline{T}_1 = \min_{t \geq 0} \left\{ t : \frac{\mathbf{w}_*^\top \mathbf{q}(t)}{\|\mathbf{q}(t)\|_2} = 1 - \gamma \right\}.$$

Recall that $\mathbf{q}(t)$ satisfies

$$\dot{\mathbf{q}} = -\mathcal{J}(\mathbf{X}; \mathbf{q})^\top \ell'(\mathcal{H}(\mathbf{X}; \mathbf{q}), \mathbf{y}) = \nabla \mathcal{N}(\mathbf{q}) - \mathcal{J}(\mathbf{X}; \mathbf{q})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{q}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})).$$

Multiplying the above equation by $\mathbf{q}^\top$ from the left we get

$$\frac{1}{2}\frac{d\|\mathbf{q}\|_2^2}{dt} = 2\mathcal{N}(\mathbf{q}) - 2\mathcal{H}(\mathbf{X}; \mathbf{q})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{q}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})) \leq 2\mathcal{N}(\mathbf{q}) = 2\mathcal{N}(\mathbf{q}/\|\mathbf{q}\|_2)\|\mathbf{q}\|_2^2,$$

where the first equality follows from Lemma 19, and the first inequality is due to convexity of the loss function. The second equality holds since $\mathcal{N}(\mathbf{q})$ is two-homogeneous.

Now, since $\mathbf{w}_*^\top \mathbf{q}(t)/\|\mathbf{q}(t)\|_2 \geq 1-\gamma$ for all $t \in [0, \overline{T}_1]$, from eq. (40), we get

$$\frac{1}{2}\frac{d\|\mathbf{q}\|_2^2}{dt} \leq 2\mathcal{N}(\mathbf{w}_*)\|\mathbf{q}\|_2^2,$$

which implies

$$\|\mathbf{q}(t)\|_2 \leq \|\mathbf{q}(0)\|_2 e^{2t\mathcal{N}(\mathbf{w}_*)}, \text{ for all } t \in [0, \overline{T}_1]. \quad (44)$$

Let $\mathbf{z}(t) = e^{2t\mathcal{N}(\mathbf{w}_*)}\mathbf{w}_*$, which is the solution of

$$\dot{\mathbf{z}} = \nabla \mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_*.$$

Note that, for $t \in [0, \overline{T}_1]$,

$$\begin{aligned}&\left(\nabla \mathcal{N}\left(\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2}\right) - \nabla \mathcal{N}(\mathbf{z}(t))\right)^\top \left(\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right) \\ &= \left(\nabla \mathcal{N}\left(\frac{\|\mathbf{q}(t)\|_2}{\|\mathbf{q}(0)\|_2}\frac{\mathbf{q}(t)}{\|\mathbf{q}(t)\|_2}\right) - \nabla \mathcal{N}(\mathbf{z}(t))\right)^\top \left(\frac{\|\mathbf{q}(t)\|_2}{\|\mathbf{q}(0)\|_2}\frac{\mathbf{q}(t)}{\|\mathbf{q}(t)\|_2} - \mathbf{z}(t)\right) \\ &\leq 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\|\mathbf{q}(t)\|_2}{\|\mathbf{q}(0)\|_2}\frac{\mathbf{q}(t)}{\|\mathbf{q}(t)\|_2} - \|\mathbf{z}(t)\|_2\frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|_2}\right\|^2 = 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|^2, \quad (45)\end{aligned}$$

where the inequality follows from $\|\mathbf{q}(t)\|_2/\|\mathbf{q}(0)\|_2 \le e^{2t\mathcal{N}(\mathbf{w}_*)} = \|\mathbf{z}(t)\|_2$, $\mathbf{w}_*^\top \mathbf{q}(t)/\|\mathbf{q}(t)\|_2 \ge 1-\gamma$ for all $t \in [0, \overline{T}_1]$, and eq. (41). Hence, for $t \in [0, \overline{T}_1]$,

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|_2^2 \\
&= \left(\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right)^\top \left(\frac{\dot{\mathbf{q}}}{\|\mathbf{q}(0)\|_2} - \dot{\mathbf{z}}\right) \\
&= \left(\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right)^\top \left(\frac{\nabla\mathcal{N}(\mathbf{q})}{\|\mathbf{q}(0)\|_2} - \nabla\mathcal{N}(\mathbf{z})\right) - \left(\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right)^\top \frac{f(\mathbf{q})}{\|\mathbf{q}(0)\|_2} \\
&\le 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|_2^2 + \beta\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|_2 \frac{\|\mathbf{q}\|_2^3}{\|\mathbf{q}(0)\|_2},
\end{aligned}$$

where the inequality follows from eq. (45) and eq. (42). The above equation implies

$$\frac{d}{dt}\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|_2 \le 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|_2 + \frac{\beta\|\mathbf{q}\|_2^3}{\|\mathbf{q}(0)\|_2}.$$

Using Lemma 29, we get

$$\begin{aligned}
\left\|\frac{\mathbf{q}(t)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(t)\right\|_2 &\le e^{2t\mathcal{N}(\mathbf{w}_*)}\left(\left\|\frac{\mathbf{q}(0)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(0)\right\|_2 + \int_0^t e^{-2s\mathcal{N}(\mathbf{w}_*)}\beta\|\mathbf{q}(s)\|_2^3/\|\mathbf{q}(0)\|_2 ds\right) \\
&\le e^{2t\mathcal{N}(\mathbf{w}_*)}\left(\left\|\frac{\mathbf{q}(0)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(0)\right\|_2 + \beta\|\mathbf{q}(0)\|_2^2\int_0^t e^{4s\mathcal{N}(\mathbf{w}_*)}ds\right) \\
&\le C_2^2\delta^2 e^{6t\mathcal{N}(\mathbf{w}_*)},
\end{aligned}$$

where the second equality uses eq. (44). In the third inequality, C_2 is some sufficiently large constant, and it follows since, for all sufficiently small $\delta > 0$, $\|\mathbf{q}(0)\|_2 = O(\delta)$ and

$$\left\|\frac{\mathbf{q}(0)}{\|\mathbf{q}(0)\|_2} - \mathbf{z}(0)\right\|_2 = \frac{\|\mathbf{a}_\delta - \|\mathbf{a}_\delta\|_2\mathbf{w}_*\|_2}{\|\mathbf{a}_\delta\|_2} \le \frac{\|\mathbf{a}_\delta - \delta\mathbf{w}_*\|_2 + |\delta - \|\mathbf{a}_\delta\|_2|}{\delta - C_1\delta^3} \le \frac{2C_1\delta^3}{\delta - C_1\delta^3} = O(\delta^2).$$

Define $\tau_q(t) \coloneqq \mathbf{q}(t)/\|\mathbf{q}(0)\|_2 - \mathbf{z}(t)$, then, using the definition of $\overline{T}_1$, we have

$$1-\gamma = \frac{\mathbf{w}_*^\top\mathbf{q}(\overline{T}_1)}{\|\mathbf{q}(\overline{T}_1)\|_2} = \frac{\mathbf{w}_*^\top\mathbf{z}(\overline{T}_1) + \mathbf{w}_*^\top\tau_q(\overline{T}_1)}{\|\mathbf{q}(\overline{T}_1)\|_2/\|\mathbf{q}(0)\|_2} \ge \frac{e^{2\overline{T}_1\mathcal{N}(\mathbf{w}_*)} - \|\tau_q(\overline{T}_1)\|_2}{e^{2\overline{T}_1\mathcal{N}(\mathbf{w}_*)}} \ge 1 - C_2^2\delta^2 e^{4\overline{T}_1\mathcal{N}(\mathbf{w}_*)},$$

which implies $\overline{T}_1 \ge \frac{1}{2\mathcal{N}(\mathbf{w}_*)}\ln\left(\frac{\sqrt{\gamma}}{C_2\delta}\right)$.

Next, we focus on $\mathbf{u}(t)$. Note that $\mathbf{w}_*^\top\mathbf{u}(0)/\|\mathbf{u}(0)\|_2 = 1 > 1-\gamma$, and $\mathbf{u}(t)$ satisfies

$$\dot{\mathbf{u}} = -\mathcal{J}(\mathbf{X};\mathbf{u})^\top\ell'(\mathcal{H}(\mathbf{X};\mathbf{u}),\mathbf{y}) = \nabla\mathcal{N}(\mathbf{u}) - \mathcal{J}(\mathbf{X};\mathbf{u})^\top(\ell'(\mathcal{H}(\mathbf{X};\mathbf{u}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y})).$$

Thus, if we define

$$\overline{T}_2 = \min_{t\ge 0}\left\{t : \frac{\mathbf{w}_*^\top\mathbf{u}(t)}{\|\mathbf{u}(t)\|_2} = 1-\gamma\right\},$$

then, similar to $\mathbf{q}(t)$, we can show, for all $t \in [0, \overline{T}_2]$,

$$\|\mathbf{u}(t)\|_2 \leq \|\mathbf{u}(0)\|_2 e^{2t\mathcal{N}(\mathbf{w}_*)}.$$

Also, there exists a sufficiently large constant C_3 such that

$$\begin{aligned}\left\| \frac{\mathbf{u}(t)}{\|\mathbf{u}(0)\|_2} - \mathbf{z}(t) \right\|_2 &\leq e^{2t\mathcal{N}(\mathbf{w}_*)} \left(\left\| \frac{\mathbf{u}(0)}{\|\mathbf{u}(0)\|_2} - \mathbf{z}(0) \right\|_2 + \int_0^t e^{-2s\mathcal{N}(\mathbf{w}_*)} \beta \|\mathbf{u}(s)\|_2^3 / \|\mathbf{u}(0)\|_2 ds \right) \\ &\leq C_3^2 \delta^2 e^{6t\mathcal{N}(\mathbf{w}_*)},\end{aligned}$$

and $\overline{T}_2 \geq \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\sqrt{\gamma}}{C_3\delta}\right)$.

Next, choose $C_4 > 0$ sufficiently small such that $C_4 < \sqrt{\gamma}\min(1/C_2, 1/C_3)$ and $C_4^2 \leq \mathcal{N}(\mathbf{w}_*)/(\zeta(1+C_1)^2)$. Then, for all $t \in [0, \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{C_4}{\delta}\right)]$,

$$\begin{aligned}&\frac{\mathbf{w}_*^\top \mathbf{u}(t)}{\|\mathbf{u}(t)\|_2} \geq 1-\gamma, \frac{\mathbf{w}_*^\top \mathbf{q}(t)}{\|\mathbf{q}(t)\|_2} \geq 1-\gamma, \text{ and} \\ &\|\mathbf{u}(t)\|_2 \leq \|\mathbf{u}(0)\|_2 e^{2t\mathcal{N}(\mathbf{w}_*)} \leq C_4, \|\mathbf{q}(t)\|_2 \leq \|\mathbf{q}(0)\|_2 e^{2t\mathcal{N}(\mathbf{w}_*)} \leq C_4(1+C_1\delta^2). \end{aligned} \tag{46}$$

Therefore, for $t \in \left[0, \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{C_4}{\delta}\right)\right]$, we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\mathbf{u}(t)-\mathbf{q}(t)\|_2^2 &= (\mathbf{u}-\mathbf{q})^\top(\dot{\mathbf{u}}-\dot{\mathbf{q}}) \\ &= (\mathbf{u}-\mathbf{q})^\top(\nabla\mathcal{N}(\mathbf{u}) - \nabla\mathcal{N}(\mathbf{q})) - (\mathbf{u}-\mathbf{q})^\top(f(\mathbf{u})-f(\mathbf{q})) \\ &\leq 3\mathcal{N}(\mathbf{w}_*)\|\mathbf{u}-\mathbf{q}\|_2^2 + \zeta\|\mathbf{u}-\mathbf{q}\|_2^2 \max(\|\mathbf{q}\|_2^2, \|\mathbf{u}\|_2^2). \\ &\leq (3\mathcal{N}(\mathbf{w}_*) + \zeta C_4^2(1+C_1)^2)\|\mathbf{u}(t)-\mathbf{q}(t)\|_2^2 \\ &\leq 4\mathcal{N}(\mathbf{w}_*)\|\mathbf{u}(t)-\mathbf{q}(t)\|_2^2,\end{aligned}$$

where the first inequality follows from combining mean value theorem with eq. (40) and 0-homogeneity of $\nabla^2\mathcal{N}(\mathbf{w})$, and eq. (43). The second inequality follows from eq. (46) when $\delta \leq 1$. The last inequality uses our assumption on C_4. Therefore,

$$\begin{aligned}\left\| \mathbf{u}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)}\ln\left(\frac{C_4}{\delta}\right)\right) - \mathbf{q}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)}\ln\left(\frac{C_4}{\delta}\right)\right)\right\|_2 &\leq \|\mathbf{u}(0)-\mathbf{q}(0)\|_2 \, e^{\frac{4\mathcal{N}(\mathbf{w}_*)}{2\mathcal{N}(\mathbf{w}_*)}\ln\left(\frac{C_4}{\delta}\right)} \\ &\leq C_1 C_4^2 \delta.\end{aligned}$$

Since $\widetilde{T}$ is fixed, from Lemma 26 and the above inequality, there exists a $C > 0$ such that for all sufficiently small $\delta > 0$, we have

$$\left\| \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)}\ln\left(\frac{1}{\delta}\right), \mathbf{a}_\delta\right) - \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)}\ln\left(\frac{1}{\delta}\right), \delta\mathbf{w}_*\right)\right\|_2 \leq C\delta,$$

for all $t \in [-\widetilde{T}, \widetilde{T}]$, which completes the proof. ■

Proof of Lemma 7: We choose $\gamma > 0$ sufficiently small such that for all unit-norm vectors $\mathbf{w}$ satisfying $\mathbf{w}^\top \mathbf{w}_* \geq 1-\gamma$, we have $\mathcal{N}(\mathbf{w}) \leq \mathcal{N}(\mathbf{w}_*)$ and

$$\left(\nabla \mathcal{N}(t_1\mathbf{w}) - \nabla \mathcal{N}(t_2\mathbf{w}_*)\right)^\top (t_1\mathbf{w} - t_2\mathbf{w}_*) \leq 2\mathcal{N}(\mathbf{w}_*)\|t_1\mathbf{w} - t_2\mathbf{w}_*\|_2^2, \forall t_2 \geq t_1 \geq 0. \quad (47)$$

The first and second inequality follow from Lemma 23 and Lemma 22, respectively. Let $\mathbf{u}_{\delta_1}(t) = \boldsymbol{\psi}\left(t, \delta_1\mathbf{w}_*\right)$ and $\mathbf{u}_{\delta_2}(t) = \boldsymbol{\psi}\left(t, \delta_2\mathbf{w}_*\right)$, where recall that $\delta_2 \geq \delta_1 > 0$. Let $\mathbf{z}(t) = e^{2t\mathcal{N}(\mathbf{w}_*)}\mathbf{w}_*$, which is the solution of

$$\dot{\mathbf{z}} = \nabla \mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_*.$$

Note that $\mathbf{u}_{\delta_1}(0) = \delta_1\mathbf{w}_*$. Define

$$T_1^* = \min_{t \geq 0} \left\{ t : \frac{\mathbf{w}_*^\top \mathbf{u}_{\delta_1}(t)}{\|\mathbf{u}_{\delta_1}(t)\|_2} = 1-\gamma \right\}.$$

Then, as demonstrated in the proof of Lemma 9, we can show

$$\|\mathbf{u}_{\delta_1}(t)\|_2 \leq \|\mathbf{u}_{\delta_1}(0)\|_2 e^{2t\mathcal{N}(\mathbf{w}_*)}.$$

Also, there exists a sufficiently large constant C_2 such that

$$\left\| \frac{\mathbf{u}_{\delta_1}(t)}{\|\mathbf{u}_{\delta_1}(0)\|_2} - \mathbf{z}(t) \right\|_2 \leq C_2^2 \delta_1^2 e^{6t\mathcal{N}(\mathbf{w}_*)}, \text{ for all } t \in [0, T_1^*], \quad (48)$$

and $T_1^* \geq \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\sqrt{\gamma}}{C_2\delta_1}\right)$. Hence, if $\delta_2 \leq \sqrt{\gamma}/C_2$, then $T_1^* \geq \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right)$.

Next, since $\delta_1 \mathbf{z}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right)\right) = \delta_2\mathbf{w}_*$, using eq. (48) we get

$$\begin{aligned}\left\| \mathbf{u}_{\delta_1}\left(\frac{\ln(\delta_2/\delta_1)}{2\mathcal{N}(\mathbf{w}_*)}\right) - \delta_2\mathbf{w}_* \right\|_2 &= \left\| \mathbf{u}_{\delta_1}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right)\right) - \delta_1 \mathbf{z}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right)\right) \right\|_2 \\ &\leq \frac{C_2^2\delta_1^3\delta_2^3}{\delta_1^3} = C_2\delta_2^3. \end{aligned} \quad (49)$$

Now, note that

$$\begin{aligned}\boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{1}{\delta_1}\right), \delta_1\mathbf{w}_*\right) &= \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{1}{\delta_2}\right) + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right), \delta_1\mathbf{w}_*\right) \\ &= \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{1}{\delta_2}\right), \boldsymbol{\psi}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right), \delta_1\mathbf{w}_*\right)\right) \\ &= \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{1}{\delta_2}\right), \mathbf{u}_{\delta_1}\left(\frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\delta_2}{\delta_1}\right)\right)\right), \end{aligned}$$

Combining the above equality with eq. (49) and Lemma 9, we get that for any fixed $t \in (-\infty, \infty)$ and all sufficiently small $\delta_2, \delta_1 > 0$, there exists a constant $C > 0$ such that

$$\left\| \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{1}{\delta_1}\right), \delta_1\mathbf{w}_*\right) - \boldsymbol{\psi}\left(t + \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{1}{\delta_2}\right), \delta_2\mathbf{w}_*\right) \right\|_2 \leq C\delta_2,$$

which implies $\mathbf{p}(t)$ exists for all $t \in (-\infty, \infty)$.

We next prove that $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, for some $\eta > 0$. Let $\mathbf{u}(t) = \boldsymbol{\psi}(t, \delta \mathbf{w}_*)$ and $\mathbf{z}(t) = e^{2t\mathcal{N}(\mathbf{w}_*)}\mathbf{w}_*$, then from eq. (48), there exists a sufficiently large constant B_1 such that

$$\|\mathbf{u}(t)/\delta - \mathbf{z}(t)\|_2 \leq B_1^2 \delta^2 e^{6t\mathcal{N}(\mathbf{w}_*)}, \text{ for all } t \in [0, T_1], \tag{50}$$

where $T_1 \geq \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\sqrt{\gamma}}{B_1 \delta}\right)$.

Define $\alpha \coloneqq \max_{\|\mathbf{w}\|_2 = 1} \|\mathcal{H}(\mathbf{X}, \mathbf{w})\|_2$. From the convexity of the loss function, we know

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &\leq \mathcal{L}(\mathbf{0}) + \ell'(\mathcal{H}(\mathbf{X}, \mathbf{w}), \mathbf{y})^\top \mathcal{H}(\mathbf{X}, \mathbf{w}) \\
&= \mathcal{L}(\mathbf{0}) + (\ell'(\mathcal{H}(\mathbf{X}, \mathbf{w}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y}))^\top \mathcal{H}(\mathbf{X}, \mathbf{w}) + \ell'(\mathbf{0}, \mathbf{y})^\top \mathcal{H}(\mathbf{X}, \mathbf{w}) \\
&\leq \mathcal{L}(\mathbf{0}) + Kn\|\mathcal{H}(\mathbf{X}, \mathbf{w})\|_2^2 - \mathcal{N}(\mathbf{w}) \leq \mathcal{L}(\mathbf{0}) + \alpha^2 Kn\|\mathbf{w}\|_2^4 - \mathcal{N}(\mathbf{w}),
\end{aligned} \tag{51}$$

where the last two inequalities follow from smoothness of the loss function and the definition of α. Now, choose $\epsilon \in (0, 1)$ sufficiently small such that

$$\epsilon \leq \frac{\sqrt{\gamma}}{B_1}, \text{ and } Kn\alpha^2(\sqrt{\epsilon} + \epsilon^{2.5}/B_1^2)^4 \leq \mathcal{N}(\mathbf{w}) - \frac{\mathcal{N}(\mathbf{w}_*)}{2}, \text{ if } \|\mathbf{w} - \mathbf{w}_*\|_2 \leq \epsilon^2 B_1^2,$$

where the second inequality holds true since $\mathcal{N}(\mathbf{w})$ is continuous and $\mathcal{N}(\mathbf{w}_*) > 0$. Note that

$$T_1 \geq \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\sqrt{\gamma}}{B_1 \delta}\right) \geq \frac{1}{2\mathcal{N}(\mathbf{w}_*)} \ln\left(\frac{\epsilon}{\delta}\right) \coloneqq T_2.$$

Hence, using eq. (50), $\|\mathbf{u}(T_2)\|_2 \leq \epsilon + B_1^2 \epsilon^3$. Also, if we define $\tau \coloneqq \mathbf{u}(T_2) - \delta \mathbf{z}(T_2)$, then $\|\tau\|_2 \leq B_1^2 \epsilon^3$. Therefore, using eq. (51), we get

$$\begin{aligned}
\mathcal{L}(\mathbf{u}(T_2)) &\leq \mathcal{L}(\mathbf{0}) + \alpha^2 Kn\|\mathbf{u}(T_2)\|_2^4 - \mathcal{N}(\mathbf{u}(T_2)) \\
&\leq \mathcal{L}(\mathbf{0}) + \alpha^2 Kn(\epsilon + B_1^2\epsilon^3)^4 - \mathcal{N}(\mathbf{u}(T_2)) \\
&= \mathcal{L}(\mathbf{0}) + \epsilon^2(\alpha^2 Kn(\sqrt{\epsilon} + B_1^2\epsilon^{2.5})^4 - \mathcal{N}(\mathbf{u}(T_2)/\epsilon)) \leq \mathcal{L}(\mathbf{0}) - \epsilon^2 \mathcal{N}(\mathbf{w}_*)/2,
\end{aligned}$$

where the last inequality follows from the choice of ϵ and since $\|\mathbf{u}(T_2)/\epsilon - \mathbf{w}_*\|_2 \leq B_1^2 \epsilon^2$. If we define $\eta = \epsilon^2 \mathcal{N}(\mathbf{w}_*)/2$, then the above equation implies, for all sufficiently small $\delta > 0$,

$$\mathcal{L}\left(\boldsymbol{\psi}\left(\frac{\ln(\epsilon) + \ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right)\right) = \mathcal{L}\left(\boldsymbol{\psi}\left(\frac{\ln(\epsilon/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right)\right) \leq \mathcal{L}(\mathbf{0}) - \eta,$$

where note that $\epsilon \in (0, 1)$ and fixed, and thus, η is fixed. Since $\ln(\epsilon) < 0$, and the loss decreases with time, the above equation implies, for all sufficiently small $\delta > 0$,

$$\mathcal{L}\left(\boldsymbol{\psi}\left(\frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right)\right) \leq \mathcal{L}\left(\boldsymbol{\psi}\left(\frac{\ln(\epsilon/\delta)}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right)\right) \leq \mathcal{L}(\mathbf{0}) - \eta.$$

Taking $\delta \to 0$, gives us $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, which completes the proof. ■

Proof of Lemma 8: We choose $\gamma > 0$ sufficiently small such that for all unit-norm vectors $\mathbf{w}$ satisfying $\mathbf{w}^\top \mathbf{w}_* \geq 1 - \gamma$, we have

$$\mathcal{N}(\mathbf{w}) \leq \mathcal{N}(\mathbf{w}_*), \|\nabla^2 \mathcal{N}(\mathbf{w})\|_2 \leq 3\mathcal{N}(\mathbf{w}_*), \text{ and} \tag{52}$$

$$\left(\nabla \mathcal{N}(t_1\mathbf{w}) - \nabla \mathcal{N}(t_2\mathbf{w}_*)\right)^\top (t_1\mathbf{w} - t_2\mathbf{w}_*) \leq 2\mathcal{N}(\mathbf{w}_*)\|t_1\mathbf{w} - t_2\mathbf{w}_*\|_2^2, \forall t_2 \geq t_1 \geq 0. \tag{53}$$

The first inequality follows from Lemma 23. The second inequality holds since, from Lemma 20, $\|\nabla^2 \mathcal{N}(\mathbf{w}_*)\|_2 = 2\mathcal{N}(\mathbf{w}_*)$, and $\nabla^2 \mathcal{N}(\mathbf{w})$ is continuous in the neighborhood of $\mathbf{w}_*$. The third inequality follows from Lemma 22. Define $f(\mathbf{w}) = \mathcal{J}(\mathbf{X}; \mathbf{w})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y}))$, then, as shown in the proof of Lemma 9, there exists a $\beta > 0$ such that

$$\|f(\mathbf{w})\|_2 \leq \beta \|\mathbf{w}\|_2^3, \text{ for all } \mathbf{w} \in \mathbb{R}^k. \tag{54}$$

Let $\mathbf{z}_1(t)$ denote the solution of

$$\dot{\mathbf{z}} = \nabla \mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_0.$$

Since $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$, it follows that $\mathbf{z}_1(t)$ converges to $\mathbf{w}_*$ in direction. Thus, we can assume there exists some time T_γ and a constant B_1 such that

$$\frac{\mathbf{w}_*^\top \mathbf{z}_1(T_\gamma)}{\|\mathbf{z}_1(T_\gamma)\|_2} = 1 - \frac{\gamma^2}{8}, \text{ and } \|\mathbf{z}_1(t)\|_2 \leq B_1, \text{ for all } t \in [0, T_\gamma].$$

Let $\mu_1 \coloneqq \max_{\|\mathbf{w}\|_2 = 1} \mathcal{N}(\mathbf{w})$, and $\mathbf{w}(t) \coloneqq \boldsymbol{\psi}(t, \delta \mathbf{w}_0)$, then $\mathbf{w}(t)$ is the solution of

$$\dot{\mathbf{w}} = \nabla \mathcal{N}(\mathbf{w}) - \mathcal{J}(\mathbf{X}; \mathbf{w})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})), \mathbf{w}(0) = \delta \mathbf{w}_0.$$

Multiplying the above equation by $\mathbf{w}^\top$ from the left and using convexity of $\ell(\cdot, \cdot)$, we get

$$\frac{1}{2}\frac{d\|\mathbf{w}\|_2^2}{dt} = 2\mathcal{N}(\mathbf{w}) - 2\mathcal{H}(\mathbf{X}; \mathbf{w})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})) \leq 2\mathcal{N}(\mathbf{w}) \leq \mu_1 \|\mathbf{w}\|_2^2,$$

which implies $\|\mathbf{w}(t)\|_2 \leq \delta e^{\mu_1 t}$. Now, since $\nabla \mathcal{N}(\mathbf{w})$ is locally Lipschitz, and for all $t \in [0, T_\gamma]$, $\|\mathbf{z}_1(t)\|_2 \leq B_1$ and $\|\mathbf{w}(t)\|_2/\delta \leq e^{\mu_1 T_\gamma} \coloneqq B_2$, there exists a $\mu_2 > 0$ such that

$$\|\nabla \mathcal{N}(\mathbf{w}(t)/\delta) - \nabla \mathcal{N}(\mathbf{z}_1(t))\|_2 \leq \mu_2 \|\mathbf{w}(t)/\delta - \mathbf{z}_1(t)\|_2, \forall t \in [0, T_\gamma].$$

Hence, for all $t \in [0, T_\gamma]$,

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\left\|\frac{\mathbf{w}(t)}{\delta} - \mathbf{z}_1(t)\right\|_2^2 &= \left(\frac{\mathbf{w}}{\delta} - \mathbf{z}_1\right)^\top \left(\frac{\dot{\mathbf{w}}}{\delta} - \dot{\mathbf{z}}_1\right) \\
&= \left(\frac{\mathbf{w}}{\delta} - \mathbf{z}_1\right)^\top (\nabla \mathcal{N}(\mathbf{w}/\delta) - \nabla \mathcal{N}(\mathbf{z}_1)) - \left(\frac{\mathbf{w}}{\delta} - \mathbf{z}_1\right)^\top \frac{f(\mathbf{w})}{\delta} \\
&\leq \mu_2 \left\|\frac{\mathbf{w}}{\delta} - \mathbf{z}_1\right\|_2^2 + \beta \left\|\frac{\mathbf{w}}{\delta} - \mathbf{z}_1\right\|_2 \|\mathbf{w}\|_2^3/\delta,
\end{aligned}$$

which implies

$$\frac{d}{dt}\left\|\frac{\mathbf{w}(t)}{\delta} - \mathbf{z}_1(t)\right\|_2 \leq \mu_2 \left\|\frac{\mathbf{w}}{\delta} - \mathbf{z}_1(t)\right\|_2 + \beta \|\mathbf{w}\|_2^3/\delta.$$

Using Lemma 29, we get

$$\left\|\frac{\mathbf{w}(T_\gamma)}{\delta}-\mathbf{z}_1(T_\gamma)\right\|_2 \le e^{\mu_2 T_\gamma}\int_0^{T_\gamma} e^{-\mu_2 s}\beta\|\mathbf{w}(s)\|_2^3/\delta ds$$
$$\le e^{\mu_2 T_\gamma}\int_0^{T_\gamma}\beta\|\mathbf{w}(s)\|_2^3/\delta ds \le \beta\delta^2 e^{\mu_2 T_\gamma}\int_0^{T_\gamma} e^{3\mu_1 s}ds \le A_1\delta^2, \quad (55)$$

where A_1 is a sufficiently large constant. Let $\tau_1 \coloneqq \mathbf{w}(T_\gamma)/\delta - \mathbf{z}_1(T_\gamma)$, then, for all sufficiently small $\delta > 0$, we have

$$\frac{\mathbf{w}_*^\top \mathbf{w}(T_\gamma)}{\|\mathbf{w}(T_\gamma)\|_2} = \frac{\mathbf{w}_*^\top \mathbf{z}_1(T_\gamma)+\mathbf{w}_*^\top\tau_1}{\|\mathbf{w}(T_\gamma)\|_2/\delta} \ge \frac{(1-\gamma^2/8)\|\mathbf{z}_1(T_\gamma)\|_2 - A_1\delta^2}{\|\mathbf{w}(T_\gamma)\|_2/\delta}$$
$$\ge \frac{(1-\gamma^2/8)\|\mathbf{z}_1(T_\gamma)\|_2 - A_1\delta^2}{\|\mathbf{z}_1(T_\gamma)\|_2 + A_1\delta^2}$$
$$= 1 - \frac{\gamma^2\|\mathbf{z}_1(T_\gamma)\|_2/8 - 2A_1\delta^2}{\|\mathbf{z}_1(T_\gamma)\|_2 + A_1\delta^2} \ge 1-\gamma^2/4.$$

Let $A_2 = \|\mathbf{z}_1(T_\gamma)\|_2$, where note that A_2 is a constant that does not depend on δ. Define $\mathbf{w}_1 = \mathbf{w}(T_\gamma)/\|\mathbf{w}(T_\gamma)\|_2$ and $\widetilde{\delta} = \|\mathbf{w}(T_\gamma)\|_2$. From eq. (55), we know

$$\widetilde{\delta} \in (A_2\delta - A_1\delta^3, A_2\delta + A_1\delta^3),$$

thus, $\widetilde{\delta}$ can be made sufficiently small by choosing δ sufficiently small. Next, let $\mathbf{u}(t) = \boldsymbol{\psi}(t+T_\gamma, \delta\mathbf{w}_0)$. Then,

$$\mathbf{u}(t) = \boldsymbol{\psi}(t+T_\gamma,\delta\mathbf{w}_0) = \boldsymbol{\psi}(t,\boldsymbol{\psi}(T_\gamma,\delta\mathbf{w}_0)) = \boldsymbol{\psi}(t,\mathbf{w}(T_\gamma)) = \boldsymbol{\psi}(t,\widetilde{\delta}\mathbf{w}_1),$$

and $\mathbf{u}(t)$ is the solution of

$$\dot{\mathbf{u}} = \nabla\mathcal{N}(\mathbf{u}) - \mathcal{J}(\mathbf{X};\mathbf{u})^\top(\ell'(\mathcal{H}(\mathbf{X};\mathbf{u}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y})), \mathbf{u}(0) = \widetilde{\delta}\mathbf{w}_1.$$

Since $\mathbf{w}_*^\top\mathbf{u}(0)/\|\mathbf{u}(0)\|_2 = \mathbf{w}_*^\top\mathbf{w}_1 \ge 1-\gamma^2/4 > 1-\gamma$, we define

$$T^* = \min_{t\ge 0}\left\{t : \frac{\mathbf{w}_*^\top\mathbf{u}(t)}{\|\mathbf{u}(t)\|_2} = 1-\gamma\right\}.$$

Now, since $\mathbf{w}_*^\top\mathbf{u}(t)/\|\mathbf{u}(t)\|_2 \ge 1-\gamma$, for all $t\in[0,T^*]$, from eq. (52), we have

$$\frac{1}{2}\frac{d\|\mathbf{u}(t)\|_2^2}{dt} = 2\mathcal{N}(\mathbf{u}) - 2\mathcal{H}(\mathbf{X};\mathbf{u})^\top(\ell'(\mathcal{H}(\mathbf{X};\mathbf{u}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y})) \le 2\mathcal{N}(\mathbf{w}_*)\|\mathbf{u}(t)\|_2^2,$$

which implies

$$\|\mathbf{u}(t)\|_2 \le \|\mathbf{u}(0)\|_2 e^{2t\mathcal{N}(\mathbf{w}_*)} = \widetilde{\delta}e^{2t\mathcal{N}(\mathbf{w}_*)}.$$

Next, let $\mathbf{z}_2(t) = e^{2t\mathcal{N}(\mathbf{w}_*)}\mathbf{w}_*$, which is the solution of

$$\dot{\mathbf{z}} = \nabla\mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_*. \quad (56)$$

Note that, for $t \in [0, T^*]$,

$$\begin{aligned}
&\left(\nabla\mathcal{N}\left(\frac{\mathbf{u}(t)}{\|\mathbf{u}(0)\|_2}\right) - \nabla\mathcal{N}(\mathbf{z}_2(t))\right)^\top \left(\frac{\mathbf{u}(t)}{\|\mathbf{u}(0)\|_2} - \mathbf{z}_2(t)\right) \\
&= \left(\nabla\mathcal{N}\left(\frac{\|\mathbf{u}(t)\|_2}{\|\mathbf{u}(0)\|_2}\frac{\mathbf{u}(t)}{\|\mathbf{u}(t)\|_2}\right) - \nabla\mathcal{N}\left(\mathbf{z}_2(t)\right)\right)^\top \left(\frac{\|\mathbf{u}(t)\|_2}{\|\mathbf{u}(0)\|_2}\frac{\mathbf{u}(t)}{\|\mathbf{u}(t)\|_2} - \mathbf{z}_2(t)\right) \\
&\leq 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\|\mathbf{u}(t)\|_2}{\|\mathbf{u}(0)\|_2}\frac{\mathbf{u}(t)}{\|\mathbf{u}(t)\|_2} - \|\mathbf{z}_2(t)\|_2\frac{\mathbf{z}_2(t)}{\|\mathbf{z}_2(t)\|_2}\right\|^2 = 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{u}(t)}{\|\mathbf{u}(0)\|_2} - \mathbf{z}_2(t)\right\|^2, \quad (57)
\end{aligned}$$

where the inequality follows from $\|\mathbf{u}(t)\|_2/\|\mathbf{u}(0)\|_2 \leq e^{2t\mathcal{N}(\mathbf{w}_*)} = \|\mathbf{z}_2(t)\|_2$, $\mathbf{w}_*^\top \mathbf{u}(t)/\|\mathbf{u}(t)\|_2 \geq 1-\gamma$ for all $t \in [0, T^*]$, and eq. (53). Then, for all $t \in [0, T^*]$, we have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\left\|\frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_2(t)\right\|_2^2 &= \left(\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right)^\top \left(\frac{\dot{\mathbf{u}}}{\widetilde{\delta}} - \dot{\mathbf{z}}_2\right) \\
&= \left(\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right)^\top \left(\nabla\mathcal{N}(\mathbf{u}/\widetilde{\delta}) - \nabla\mathcal{N}(\mathbf{z}_2)\right) - \left(\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right)^\top \frac{f(\mathbf{u})}{\widetilde{\delta}} \\
&\leq 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right\|_2^2 + \beta\left\|\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right\|_2\|\mathbf{u}\|_2^3/\widetilde{\delta}.
\end{aligned}$$

From the above equation, we have

$$\frac{d}{dt}\left\|\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right\|_2 \leq 2\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_2\right\|_2 + \beta\|\mathbf{u}\|_2^3/\widetilde{\delta},$$

which implies

$$\begin{aligned}
\left\|\frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_2(t)\right\|_2 &\leq e^{2t\mathcal{N}(\mathbf{w}_*)}\left(\|\mathbf{w}_1 - \mathbf{w}_*\|_2 + \int_0^t e^{-2s\mathcal{N}(\mathbf{w}_*)}\beta\|\mathbf{u}(s)\|_2^3/\widetilde{\delta}ds\right) \\
&\leq e^{2t\mathcal{N}(\mathbf{w}_*)}\|\mathbf{w}_1 - \mathbf{w}_*\|_2 + \beta\widetilde{\delta}^2 e^{2t\mathcal{N}(\mathbf{w}_*)}\int_0^t e^{4s\mathcal{N}(\mathbf{w}_*)}ds \\
&\leq e^{2t\mathcal{N}(\mathbf{w}_*)}\|\mathbf{w}_1 - \mathbf{w}_*\|_2 + A_3\widetilde{\delta}^2 e^{6t\mathcal{N}(\mathbf{w}_*)},
\end{aligned}$$

where A_3 is a sufficiently large constant. Now, define $\overline{T} = \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}$, then we claim that $\overline{T} < T^*$, for all sufficiently small $\delta > 0$. For the sake of contradiction, let $T^* \leq \overline{T}$. Define $\tau_2(t) = \mathbf{u}(t) - \widetilde{\delta}\mathbf{z}_2(t)$, then

$$\|\tau_2(T^*)\|_2 = \|\mathbf{u}(T^*) - \widetilde{\delta}\mathbf{z}_2(T^*)\|_2 \leq \widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)}\|\mathbf{w}_1 - \mathbf{w}_*\|_2 + A_3\widetilde{\delta}^3 e^{6T^*\mathcal{N}(\mathbf{w}_*)}.$$

Dividing by $\widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)}$ on both sides and using $T^* \leq \overline{T}$, we get

$$\frac{\|\tau_2(T^*)\|_2}{\widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)}} \leq \|\mathbf{w}_1 - \mathbf{w}_*\|_2 + A_3\widetilde{\delta}^2 e^{4\overline{T}\mathcal{N}(\mathbf{w}_*)}.$$

Next, by definition of T^*, we have

$$1 - \gamma = \frac{\mathbf{w}_*^\top \mathbf{u}(T^*)}{\|\mathbf{u}(T^*)\|_2} \geq \frac{\mathbf{w}_*^\top \mathbf{u}(T^*)}{\widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)}} = \frac{\widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)} + \mathbf{w}_*^\top \tau_2(T^*)}{\widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)}} \geq 1 - \frac{\|\tau_2(T^*)\|_2}{\widetilde{\delta}e^{2T^*\mathcal{N}(\mathbf{w}_*)}}.$$

Now, since $\mathbf{w}_1^\top \mathbf{w}_* \geq 1 - \gamma^2/4$,

$$\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 = 2 - 2\mathbf{w}_1^\top \mathbf{w}_* \leq \gamma^2/2.$$

Hence,

$$1 - \gamma \geq 1 - \gamma/\sqrt{2} - A_3\widetilde{\delta}^2 e^{4\overline{T}\mathcal{N}(\mathbf{w}_*)} = 1 - \gamma/\sqrt{2} - A_3\widetilde{\delta}^2/\widetilde{\delta}^{\frac{16\mathcal{N}(\mathbf{w}_*)}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}},$$

which implies

$$\gamma(1 - 1/\sqrt{2}) \leq A_3\widetilde{\delta}^2/\widetilde{\delta}^{\frac{16\mathcal{N}(\mathbf{w}_*)}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} = A_3\widetilde{\delta}^{\frac{2\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}.$$

The above inequality can not be true if δ is chosen sufficiently small, leading to a contradiction. Hence, $\overline{T} < T^*$. Next, let $\mathbf{z}_3(t)$ be the the solution of

$$\dot{\mathbf{z}} = \nabla\mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_1, \tag{58}$$

Since $\mathbf{w}_1^\top \mathbf{w}_* \geq 1 - \gamma^2/4 > 1 - \gamma$, then we may assume without loss of generality that $\gamma > 0$ is sufficiently small such that Lemma 24 is applicable and hence,

$$\frac{\mathbf{w}_*^\top \mathbf{z}_3(t)}{\|\mathbf{z}_3(t)\|_2} \geq 1 - e^{-t\Delta}\gamma, \text{ for all } t \geq 0, \text{ and } \mathbf{z}_3(t) = \mathbf{g}(t)e^{2t\mathcal{N}(\mathbf{w}_*)}, \tag{59}$$

where $\|\mathbf{g}(t)\|_2 \in [\kappa_1, \kappa_2]$, for some $\kappa_2 \geq \kappa_1 > 0$. By definition of $\overline{T}$, we have

$$\frac{\mathbf{w}_*^\top \mathbf{z}_3(\overline{T})}{\|\mathbf{z}_3(\overline{T})\|_2} \geq 1 - \widetilde{\delta}^{\frac{4\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}\gamma, \text{ and } \|\mathbf{z}_3(\overline{T})\|_2 = \frac{\|\mathbf{g}(\overline{T})\|_2}{\widetilde{\delta}^{\frac{8\mathcal{N}(\mathbf{w}_*)}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}},$$

which implies

$$\left\| \frac{\mathbf{z}_3(\overline{T})}{\|\mathbf{z}_3(\overline{T})\|_2} - \mathbf{w}_* \right\|_2^2 = 2 - \frac{2\mathbf{w}_*^\top \mathbf{z}_3(\overline{T})}{\|\mathbf{z}_3(\overline{T})\|_2} \leq 2\widetilde{\delta}^{\frac{4\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}\gamma.$$

Multiplying by $\widetilde{\delta}^2\|\mathbf{z}_3(\overline{T})\|_2^2$ on both sides gives us

$$\begin{aligned}\left\| \widetilde{\delta}\mathbf{z}_3(\overline{T}) - \widetilde{\delta}\|\mathbf{z}_3(\overline{T})\|_2 \mathbf{w}_* \right\|_2^2 &\leq 2\widetilde{\delta}^2\|\mathbf{z}_3(\overline{T})\|_2^2\delta^{\frac{4\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}\gamma \\ &\leq 2\|\mathbf{g}(\overline{T})\|_2^2\gamma\frac{\widetilde{\delta}^2\widetilde{\delta}^{\frac{4\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}}{\widetilde{\delta}^{\frac{16\mathcal{N}(\mathbf{w}_*)}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}} = 2\|\mathbf{g}(\overline{T})\|_2^2\gamma\widetilde{\delta}^{\frac{6\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}.\end{aligned} \tag{60}$$

Next, note that for $t \in [0, \overline{T}]$, $\mathbf{w}_*^\top \mathbf{z}_3(t)/\|\mathbf{z}_3(t)\|_2, \mathbf{w}_*^\top \mathbf{u}(t)/\|\mathbf{u}(t)\|_2 \geq 1 - \gamma$. Since $\nabla^2\mathcal{N}(\mathbf{w})$ is 0-homogeneous, from the mean value theorem and eq. (52), for all $t \in [0, \overline{T}]$, we have

$$\left\| \left(\frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_3(t) \right)^\top \left(\nabla\mathcal{N}(\mathbf{u}(t)/\widetilde{\delta}) - \nabla\mathcal{N}(\mathbf{z}_3(t)) \right) \right\|_2 \leq 3\mathcal{N}(\mathbf{w}_*) \left\| \frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_3(t) \right\|_2^2.$$

Hence, for $t \in [0, \overline{T}]$, we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left\| \frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_3(t) \right\|_2^2 &= \left(\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_3 \right)^\top \left(\frac{\dot{\mathbf{u}}}{\widetilde{\delta}} - \dot{\mathbf{z}}_3 \right) \\ &= \left(\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_3 \right)^\top \left(\nabla\mathcal{N}(\mathbf{u}/\widetilde{\delta}) - \nabla\mathcal{N}(\mathbf{z}_3) \right) - \left(\frac{\mathbf{u}}{\widetilde{\delta}} - \mathbf{z}_3 \right)^\top \frac{f(\mathbf{w})}{\widetilde{\delta}} \\ &\leq 3\mathcal{N}(\mathbf{w}_*) \left\| \frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_3(t) \right\|_2^2 + \beta \left\| \frac{\mathbf{u}(t)}{\widetilde{\delta}} - \mathbf{z}_3(t) \right\|_2 \|\mathbf{u}\|_2^3/\widetilde{\delta}.\end{aligned}$$

From the above equation, we have

$$\frac{d}{dt}\left\|\frac{\mathbf{u}(t)}{\widetilde{\delta}}-\mathbf{z}_3(t)\right\|_2 \leq 3\mathcal{N}(\mathbf{w}_*)\left\|\frac{\mathbf{u}(t)}{\widetilde{\delta}}-\mathbf{z}_3(t)\right\|_2+\beta\|\mathbf{u}(t)\|_2^3/\widetilde{\delta},$$

which implies

$$\begin{aligned}\left\|\frac{\mathbf{u}(t)}{\widetilde{\delta}}-\mathbf{z}_3(t)\right\|_2 &\leq e^{3\mathcal{N}(\mathbf{w}_*)t}\int_0^t e^{-3\mathcal{N}(\mathbf{w}_*)s}\beta\|\mathbf{u}(s)\|_2^3/\widetilde{\delta}ds\\ &\leq \beta\widetilde{\delta}^2 e^{3\mathcal{N}(\mathbf{w}_*)t}\int_0^t e^{3\mathcal{N}(\mathbf{w}_*)s}ds \leq A_4\widetilde{\delta}^2 e^{6t\mathcal{N}(\mathbf{w}_*)},\end{aligned}$$

where A_4 is a sufficiently large constant. Putting $t=\overline{T}$ in the above equation gives us

$$\left\|\mathbf{u}(\overline{T})-\widetilde{\delta}\mathbf{z}_3(\overline{T})\right\|_2 \leq A_4\widetilde{\delta}^3 e^{6\overline{T}\mathcal{N}(\mathbf{w}_*)} = A_4\widetilde{\delta}^3/\widetilde{\delta}^{\frac{24\mathcal{N}(\mathbf{w}_*)}{\Delta+8\mathcal{N}(\mathbf{w}_*)}} = A_4\widetilde{\delta}^{\frac{3\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}.$$

From eq. (60) and the above equation, we get

$$\begin{aligned}\left\|\mathbf{u}(\overline{T})-\widetilde{\delta}\|\mathbf{z}_3(\overline{T})\|_2\mathbf{w}_*\right\|_2 &\leq \left\|\mathbf{u}(\overline{T})-\widetilde{\delta}\mathbf{z}_3(\overline{T})\right\|_2+\left\|\widetilde{\delta}\|\mathbf{z}_3(\overline{T})\|_2\mathbf{w}_*-\widetilde{\delta}\mathbf{z}_3(\overline{T})\right\|_2\\ &\leq \left(A_4+\|\mathbf{g}(\overline{T})\|_2\sqrt{2\gamma}\right)\widetilde{\delta}^{\frac{3\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}.\end{aligned}$$

Also, note that

$$\widetilde{\delta}\|\mathbf{z}_3(\overline{T})\|_2 = \|\mathbf{g}(\overline{T})\|_2\widetilde{\delta}/\widetilde{\delta}^{\frac{8\mathcal{N}(\mathbf{w}_*)}{\Delta+8\mathcal{N}(\mathbf{w}_*)}} = \|\mathbf{g}(\overline{T})\|_2\widetilde{\delta}^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}.$$

Thus, using the above two equations and the definition of $\mathbf{u}(t)$, we get

$$\left\|\boldsymbol{\psi}(\overline{T}+T_\gamma,\delta\mathbf{w}_0)-\|\mathbf{g}(\overline{T})\|_2\widetilde{\delta}^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}\mathbf{w}_*\right\|_2 \leq \left(A_4+\|\mathbf{g}(\overline{T})\|_2\sqrt{2\gamma}\right)\widetilde{\delta}^{\frac{3\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}.$$

Since $\overline{T}$ depends on δ, $\|\mathbf{g}(\overline{T})\|_2$ may also depend on δ, but $\|\mathbf{g}(\overline{T})\|_2 \in [\kappa_1,\kappa_2]$. Hence, for all sufficiently small $\delta>0$,

$$\left\|\boldsymbol{\psi}(\overline{T}+T_\gamma,\delta\mathbf{w}_0)-\|\mathbf{g}(\overline{T})\|_2\widetilde{\delta}^{\frac{\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}}\mathbf{w}_*\right\|_2 \leq C\widetilde{\delta}^{\frac{3\Delta}{\Delta+8\mathcal{N}(\mathbf{w}_*)}},$$

where C is a sufficiently large constant, which completes the proof. ■

Proof of Corollary 5: Since $\mathbf{p}(t)$ is bounded for all $t\geq 0$, there exists a constant $B>0$ such that $\|\mathbf{p}(t)\|_2 \leq B$, for all $t\geq 0$. Moreover, since $\mathcal{L}(\cdot)$ has locally Lipschitz gradient, there exists a constant $\widetilde{A}>0$ such that, if $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \leq 2B$, then

$$\|\nabla\mathcal{L}(\mathbf{w}_1)-\nabla\mathcal{L}(\mathbf{w}_2)\|_2 \leq \widetilde{A}\|\mathbf{w}_1-\mathbf{w}_2\|_2. \tag{61}$$

Since $\mathbf{p}^*=\lim_{t\to\infty}\mathbf{p}(t)$ and $\nabla\mathcal{L}(\mathbf{p}^*)=\mathbf{0}$, for any $\epsilon\in(0,B)$, there exists a T_ϵ such that

$$\|\mathbf{p}(T_\epsilon)-\mathbf{p}_*\|_2 \leq \epsilon/2 \text{ and } \|\nabla\mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 \leq \epsilon/2. \tag{62}$$

Since T_ϵ does not depend on δ, therefore, from Theorem 4, for all sufficiently small $\delta > 0$ and for all $t \in [-T_\epsilon, T_\epsilon]$,

$$\left\| \boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_0\right) - \mathbf{p}(t) \right\|_2 \leq \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon/2.$$

Putting $t = T_\epsilon$ in the above equation, and using eq. (62), we get

$$\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right) - \mathbf{p}^*\|_2 \leq \epsilon,$$

where $T_\delta \coloneqq T_\epsilon + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}$. Next, since $\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2 \leq B + \epsilon/2 \leq 2B$, using eq. (61) and eq. (62), we get

$$\begin{aligned}\|\nabla \mathcal{L}\left(\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\right)\|_2 &\leq \|\nabla \mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 + \|\nabla \mathcal{L}\left(\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\right) - \nabla \mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 \\ &\leq \epsilon/2 + \widetilde{A}\widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon,\end{aligned}$$

where the final inequality is true for all sufficiently small $\delta > 0$, thus completing the proof. ■

Proof of Corollary 6: Since $\lim_{t\to\infty} \mathbf{p}(t)/\|\mathbf{p}(t)\|_2 = \mathbf{p}^*$, $\lim_{t\to\infty} \|\mathbf{p}(t)\|_2 = \infty$, and $\lim_{t\to\infty} \nabla \mathcal{L}(\mathbf{p}(t)) = \mathbf{0}$, therefore, for any $\epsilon \in (0, 1)$, we can choose a T_ϵ such that

$$\mathbf{p}(T_\epsilon)^\top \mathbf{p}^* / \|\mathbf{p}(T_\epsilon)\|_2 \geq 1 - \epsilon/2, \|\mathbf{p}(T_\epsilon)\|_2 \geq 1/\epsilon, \text{ and } \|\nabla \mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 \leq \epsilon/2. \tag{63}$$

Let $B_\epsilon \coloneqq \max_{t\in[0,T_\epsilon]} \|\mathbf{p}(t)\|_2$. Then, since $\mathcal{L}(\cdot)$ has locally Lipschitz gradient, there exists a constant $\widetilde{A}_\epsilon > 0$ such that, if $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \leq B_\epsilon + \epsilon$, it follows that

$$\|\nabla \mathcal{L}(\mathbf{w}_1) - \nabla \mathcal{L}(\mathbf{w}_2)\|_2 \leq \widetilde{A}_\epsilon \|\mathbf{w}_1 - \mathbf{w}_2\|_2, \tag{64}$$

where note that $\widetilde{A}_\epsilon$ depends on ϵ. Since T_ϵ does not depend on δ, from Theorem 4, for all sufficiently small $\delta > 0$ and for all $t \in [-T_\epsilon, T_\epsilon]$,

$$\left\| \boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_0\right) - \mathbf{p}(t) \right\|_2 \leq \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon/2. \tag{65}$$

Putting $t = T_\epsilon$ in the above equation, and using eq. (63), we get, for all sufficiently small δ,

$$\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2 \geq \|\mathbf{p}(T_\epsilon)\|_2 - \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} \geq 1/\epsilon - \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} \geq 1/(2\epsilon),$$

where $T_\delta \coloneqq T_\epsilon + T_1 + \frac{\ln(1/b_\delta)}{2\mathcal{N}(\mathbf{w}_*)} + \frac{\ln(1/\widetilde{\delta})}{2\mathcal{N}(\mathbf{w}_*)}$. We also have

$$\frac{\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)^\top \mathbf{p}^*}{\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2} \geq \frac{\mathbf{p}(T_\epsilon)^\top \mathbf{p}_* - \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}}{\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2} \geq \frac{(1-\epsilon/2)\|\mathbf{p}(T_\epsilon)\|_2 - \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}}{\|\mathbf{p}(T_\epsilon)\|_2 + \widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}}} \geq 1 - \epsilon,$$

where the first inequality uses eq. (65). The second inequality uses eq. (65) and eq. (63). The final inequality is true for all sufficiently small $\delta > 0$. Next, since $\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2 \leq B_\epsilon + \epsilon/2$, using eq. (64) and eq. (63), we get

$$\begin{aligned}\|\nabla \mathcal{L}\left(\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\right)\|_2 &\leq \|\nabla \mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 + \|\nabla \mathcal{L}\left(\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\right) - \nabla \mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 \\ &\leq \epsilon/2 + \widetilde{A}_\epsilon\widetilde{C}\delta^{\frac{\Delta}{\Delta + 8\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon,\end{aligned}$$

where the final inequality is true for all sufficiently small $\delta > 0$, thus completing the proof. ■

Appendix C. Proof Omitted from Section 3.2

We first prove three important lemmata, which are then used to prove Theorem 10.

Lemma 30 *Consider the setting in Theorem 10. Suppose there exists a constant $C_1 > 0$ such that for every $\delta > 0$, $\mathbf{a}_\delta$ is a vector that satisfies $\|\mathbf{a}_\delta - \delta \mathbf{w}_*\|_2 \leq C_1 \delta^{L+1}$. Then, for any fixed $\widetilde{T} \in (-\infty, \infty)$, there exists a constant $C > 0$ such that for all sufficiently small $\delta > 0$,*

$$\left\| \boldsymbol{\psi}\left(t + \frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \mathbf{a}_\delta\right) - \boldsymbol{\psi}\left(t + \frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta \mathbf{w}_*\right) \right\|_2 \leq C\delta, \text{ for all } t \in [-\widetilde{T}, \widetilde{T}].$$

Proof We choose $\gamma > 0$ sufficiently small such that for all unit-norm vector $\mathbf{w}$ satisfying $\mathbf{w}^\top \mathbf{w}_* \geq 1 - \gamma$, we have

$$\mathcal{N}(\mathbf{w}) \leq \mathcal{N}(\mathbf{w}_*), \|\nabla^2 \mathcal{N}(\mathbf{w})\|_2 \leq L(L - 1/2)\mathcal{N}(\mathbf{w}_*), \text{ and} \tag{66}$$

$$(\nabla \mathcal{N}(t_1 \mathbf{w}) - \nabla \mathcal{N}(t_2 \mathbf{w}_*))^\top (t_1 \mathbf{w} - t_2 \mathbf{w}_*) \leq L(L-1)\mathcal{N}(\mathbf{w}_*) t_2^{L-2} \|t_1 \mathbf{w} - t_2 \mathbf{w}_*\|_2^2, \forall t_2 \geq t_1 \geq 0. \tag{67}$$

The first inequality follows from Lemma 23. The second inequality holds since, from Lemma 20, $\|\nabla^2 \mathcal{N}(\mathbf{w}_*)\|_2 = L(L-1)\mathcal{N}(\mathbf{w}_*)$, and $\nabla^2 \mathcal{N}(\mathbf{w})$ is continuous in the neighborhood of $\mathbf{w}_*$. The third inequality follows from Lemma 22.

We further define $f(\mathbf{w}) = \mathcal{J}(\mathbf{X}; \mathbf{w})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y}))$. Then, note that

$$\|f(\mathbf{w})\|_2 \leq \|\mathcal{J}(\mathbf{X}; \mathbf{w})\|_2 \|(\ell'(\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})\|_2 \leq Kn \|\mathcal{J}(\mathbf{X}; \mathbf{w})\|_2 \|\mathcal{H}(\mathbf{X}; \mathbf{w}), \mathbf{y})\|_2,$$

where the first inequality follows from the Cauchy-Schwartz inequality, and the second follows from the smoothness of the loss function (see Assumption 2). Note that the RHS in the above inequality is $(2L-1)-$homogeneous in $\mathbf{w}$. Thus, there exists a $\beta > 0$ such that

$$\|f(\mathbf{w})\|_2 \leq \beta \|\mathbf{w}\|_2^{2L-1}, \text{ for all } \mathbf{w} \in \mathbb{R}^k. \tag{68}$$

Also, $f(\mathbf{w})$ is continuously differentiable in the neighborhood of $\mathbf{w}_*$. Thus, in a similar way as in the proof of Lemma 9, we may assume that for all vectors $\mathbf{w} \in \mathbb{R}^k$ that satisfy $\mathbf{w}_*^\top \mathbf{w} / \|\mathbf{w}\|_2 \geq 1 - \gamma$, we have

$$\|\nabla f(\mathbf{w})\|_2 \leq K \sum_{i=1}^n \|\nabla^2 \mathcal{H}(\mathbf{x}_i; \mathbf{w})\|_2 |\mathcal{H}(\mathbf{x}_i; \mathbf{w})| + \|\nabla \mathcal{H}(\mathbf{x}_i; \mathbf{w}) \nabla \mathcal{H}(\mathbf{x}_i; \mathbf{w})^\top\|_2.$$

Note that the final upper bound in the above equation is $(2L-2)$-homogeneous. Thus, there exists a $\zeta > 0$ such that for all vector $\mathbf{w} \in \mathbb{R}^k$ satisfying $\mathbf{w}_*^\top \mathbf{w} / \|\mathbf{w}\|_2 \geq 1 - \gamma$, we have

$$\|\nabla f(\mathbf{w})\|_2 \leq \zeta \|\mathbf{w}\|_2^{2L-2}.$$

Furthermore, from the mean value theorem, we have

$$\begin{aligned} \|f(\mathbf{w}_1) - f(\mathbf{w}_2)\|_2 &\leq \|\nabla f(\tilde{\mathbf{w}})\|_2 \|\mathbf{w}_1 - \mathbf{w}_2\|_2 \\ &\leq \zeta \max(\|\mathbf{w}_2\|_2^{2L-2}, \|\mathbf{w}_1\|_2^{2L-2}) \|\mathbf{w}_1 - \mathbf{w}_2\|_2, \end{aligned} \tag{69}$$

where $\mathbf{w}_*^\top \mathbf{w}_1 / \|\mathbf{w}_1\|_2, \mathbf{w}_*^\top \mathbf{w}_2 / \|\mathbf{w}_2\|_2 \geq 1 - \gamma$.

Now, let $\mathbf{q}(t) = \boldsymbol{\psi}(t, \mathbf{a}_\delta)$ and $\mathbf{u}(t) = \boldsymbol{\psi}(t, \delta \mathbf{w}_*)$. Then $\|\mathbf{q}(0) - \mathbf{u}(0)\|_2 \leq C_1 \delta^{L+1}$, which implies $\|\mathbf{q}(0)\|_2 \leq \delta + C_1 \delta^{L+1}$. Therefore, for all sufficiently small $\delta > 0$, we have

$$\frac{\mathbf{w}_*^\top \mathbf{q}(0)}{\|\mathbf{q}(0)\|_2} = \frac{\mathbf{w}_*^\top \mathbf{u}(0) + \mathbf{w}_*^\top (\mathbf{q}(0) - \mathbf{u}(0))}{\|\mathbf{q}(0)\|_2} \geq \frac{\delta - C_1 \delta^{L+1}}{\|\mathbf{q}(0)\|_2} \geq \frac{\delta - C_1 \delta^{L+1}}{\delta + C_1 \delta^{L+1}} > 1 - \gamma.$$

Define

$$\overline{T}_1 = \min_{t \geq 0} \left\{ t : \frac{\mathbf{w}_*^\top \mathbf{q}(t)}{\|\mathbf{q}(t)\|_2} = 1 - \gamma \right\}.$$

Recall that $\mathbf{q}(t)$ satisfies

$$\dot{\mathbf{q}} = -\mathcal{J}(\mathbf{X}; \mathbf{q})^\top \ell'(\mathcal{H}(\mathbf{X}; \mathbf{q}), \mathbf{y}) = \nabla \mathcal{N}(\mathbf{q}) - \mathcal{J}(\mathbf{X}; \mathbf{q})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{q}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})).$$

Multiplying the above equation by $\mathbf{q}^\top$ from the left we get

$$\frac{1}{2} \frac{d\|\mathbf{q}\|_2^2}{dt} = L\mathcal{N}(\mathbf{q}) - L\mathcal{H}(\mathbf{X}; \mathbf{q})^\top (\ell'(\mathcal{H}(\mathbf{X}; \mathbf{q}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})) \leq L\mathcal{N}(\mathbf{q}) = L\mathcal{N}(\mathbf{q}/\|\mathbf{q}\|_2)\|\mathbf{q}\|_2^L,$$

where the first equality follows from Lemma 19, the first inequality is due to convexity of the loss function, and the second equality holds since $\mathcal{N}(\mathbf{q})$ is L-homogeneous.

Now, since $\mathbf{w}_*^\top \mathbf{q}(t)/\|\mathbf{q}(t)\|_2 \geq 1 - \gamma$, for all $t \in [0, \overline{T}_1]$, from eq. (66), we get

$$\frac{d\|\mathbf{q}\|_2^2}{dt} \leq L\mathcal{N}(\mathbf{w}_*)\|\mathbf{q}\|_2^L,$$

which implies

$$\|\mathbf{q}(t)\|_2^{L-2} \leq \frac{\|\mathbf{q}(0)\|_2^{L-2}}{1 - t\|\mathbf{q}(0)\|_2^{L-2} L(L-2)\mathcal{N}(\mathbf{w}_*)}, \text{ for all } t \in [0, \overline{T}_1].$$

Define $\mathbf{s}_q(t) = \frac{1}{\delta} \mathbf{q}\left(\frac{t}{\delta^{L-2}}\right)$ and $\eta = (1 + C_1 \delta^L)^{L-2}$, then, for all $t \in [0, \overline{T}_1]$,

$$\|\mathbf{s}_q(t)\|_2^{L-2} \leq \frac{(1 + C_1\delta^L)^{L-2}}{1 - t(1 + C_1\delta^L)^{L-2} L(L-2)\mathcal{N}(\mathbf{w}_*)} = \frac{\eta}{1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*)}, \tag{70}$$

where we used $\|\mathbf{q}(0)\|_2 \leq \delta(1 + C_1\delta^L)$. Also, since

$$\begin{aligned}
\frac{d\mathbf{s}_q}{dt} = \frac{1}{\delta^{L-1}} \dot{\mathbf{q}}\left(\frac{t}{\delta^{L-2}}\right) &= -\frac{1}{\delta^{L-1}} \mathcal{J}\left(\mathbf{X}; \mathbf{q}\left(\frac{t}{\delta^{L-2}}\right)\right)^\top \ell'\left(\mathcal{H}\left(\mathbf{X}; \mathbf{q}\left(\frac{t}{\delta^{L-2}}\right)\right), \mathbf{y}\right) \\
&= -\mathcal{J}\left(\mathbf{X}; \frac{1}{\delta}\mathbf{q}\left(\frac{t}{\delta^{L-2}}\right)\right)^\top \ell'\left(\mathcal{H}\left(\mathbf{X}; \mathbf{q}\left(\frac{t}{\delta^{L-2}}\right)\right), \mathbf{y}\right) \\
&= -\mathcal{J}\left(\mathbf{X}; \mathbf{s}_q(t)\right)^\top \ell'\left(\mathcal{H}\left(\mathbf{X}; \delta \mathbf{s}_q\left(t\right)\right), \mathbf{y}\right),
\end{aligned}$$

where the second equality follows from $(L-1)$-homogeneity of $\mathcal{J}(\mathbf{X}; \mathbf{w})$, we have that $\mathbf{s}_q(t)$ is the solution of

$$\dot{\mathbf{s}} = \nabla \mathcal{N}(\mathbf{s}) - \mathcal{J}(\mathbf{X}; \mathbf{s})^\top (\ell'(\mathcal{H}(\mathbf{X}; \delta\mathbf{s}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})) = \nabla \mathcal{N}(\mathbf{s}) - f(\delta\mathbf{s})/\delta^{L-1}, \mathbf{s}(0) = \mathbf{q}(0)/\delta.$$

Let $\mathbf{z}(t) = \frac{\eta^{\frac{1}{L-2}}\mathbf{w}_*}{(1-t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{1/(L-2)}}$, which is the solution of

$$\dot{\mathbf{z}} = \nabla\mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \eta^{\frac{1}{L-2}}\mathbf{w}_*.$$

Note that, for $t \in [0, \overline{T}_1]$,

$$\begin{aligned}
&\left(\nabla\mathcal{N}\left(\mathbf{s}_q(t)\right) - \nabla\mathcal{N}(\mathbf{z}(t))\right)^\top \left(\mathbf{s}_q(t) - \mathbf{z}(t)\right) \\
&= \left(\nabla\mathcal{N}\left(\|\mathbf{s}_q(t)\|_2 \frac{\mathbf{s}_q(t)}{\|\mathbf{s}_q(t)\|_2}\right) - \nabla\mathcal{N}\left(\mathbf{z}(t)\right)\right)^\top \left(\|\mathbf{s}_q(t)\|_2 \frac{\mathbf{s}_q(t)}{\|\mathbf{s}_q(t)\|_2} - \mathbf{z}(t)\right) \\
&\leq L(L-1)\mathcal{N}(\mathbf{w}_*)\|\mathbf{z}(t)\|_2^{L-2} \left\|\|\mathbf{s}_q(t)\|_2 \frac{\mathbf{s}_q(t)}{\|\mathbf{s}_q(t)\|_2} - \|\mathbf{z}(t)\|_2 \frac{\mathbf{z}(t)}{\|\mathbf{z}(t)\|_2}\right\|^2 \\
&= L(L-1)\mathcal{N}(\mathbf{w}_*)\|\mathbf{z}(t)\|_2^{L-2} \|\mathbf{s}_q(t) - \mathbf{z}(t)\|^2, \qquad (71)
\end{aligned}$$

where the inequality follows since, from eq. (70), $\|\mathbf{s}_q(t)\|_2 \leq \|\mathbf{z}(t)\|_2$, $\mathbf{w}_*^\top \mathbf{q}(t)/\|\mathbf{q}(t)\|_2 \geq 1-\gamma$ for all $t \in [0, \overline{T}_1]$, and from eq. (41). Hence, for $t \in [0, \overline{T}_1]$,

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2^2 &= (\mathbf{s}_q - \mathbf{z})^\top (\dot{\mathbf{s}}_q - \dot{\mathbf{z}}) \\
&= (\mathbf{s}_q - \mathbf{z})^\top (\nabla\mathcal{N}(\mathbf{s}_q) - \nabla\mathcal{N}(\mathbf{z})) - (\mathbf{s}_q - \mathbf{z})^\top f(\delta\mathbf{s}_q)/\delta^{L-1} \\
&\leq L(L-1)\mathcal{N}(\mathbf{w}_*)\|\mathbf{z}(t)\|_2^{L-2} \|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2^2 + \beta\delta^L \|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2 \|\mathbf{s}_q(t)\|_2^{2L-1}.
\end{aligned}$$

Next, since for all $t \in [0, \overline{T}_1]$,

$$\|\mathbf{s}_q(t)\|_2^{L-2} \leq \|\mathbf{z}(t)\|_2^{L-2} = \frac{\eta}{1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*)}, \qquad (72)$$

we have

$$\frac{d}{dt}\|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2 \leq \frac{\eta L(L-1)\mathcal{N}(\mathbf{w}_*)}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))} \|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2 + \beta\delta^L \|\mathbf{z}(t)\|_2^{2L-1}.$$

Using Lemma 29, we get

$$\|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2 \leq \frac{1}{P(t)} \left(\|\mathbf{s}_q(0) - \mathbf{z}(0)\|_2 + \beta\delta^L \int_0^t P(s)\|\mathbf{z}(t)\|_2^{2L-1} ds\right),$$

where $P(t) = e^{-\int_0^t b(s)ds}$ and $b(t) = \frac{\eta L(L-1)\mathcal{N}(\mathbf{w}_*)}{(1-t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))}$. Using Lemma 28, we have

$$P(t) = (1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{(L-1)/(L-2)},$$

which implies

$$\begin{aligned}
\int_0^t P(s)\|\mathbf{z}(s)\|_2^{2L-1} ds &\leq \eta^{\frac{2L-1}{L-2}} \int_0^t (1 - s\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L-1}{L-2} - \frac{2L-1}{L-2}} ds \\
&= \eta^{\frac{2L-1}{L-2}} \int_0^t 1/\left(1 - s\eta L(L-2)\mathcal{N}(\mathbf{w}_*)\right)^{\frac{L}{L-2}} ds \\
&= \eta^{\frac{2L-1}{L-2}} \frac{L-2}{2\eta L(L-2)\mathcal{N}(\mathbf{w}_*)} \left(\frac{1}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{2}{L-2}}} - 1\right) \\
&\leq \frac{\eta^{\frac{L+1}{L-2}}}{2L\mathcal{N}(\mathbf{w}_*)} \left(\frac{1}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{2}{L-2}}}\right).
\end{aligned}$$

Hence, for all $t \in [0, \overline{T}_1]$ and all sufficiently small $\delta > 0$,

$$\begin{aligned}\|\mathbf{s}_q(t) - \mathbf{z}(t)\|_2 &\leq \frac{\|\mathbf{s}_q(0) - \mathbf{z}(0)\|_2}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L-1}{L-2}}} + \frac{\beta\delta^L \eta^{\frac{L+1}{L-2}}}{2L\mathcal{N}(\mathbf{w}_*)} \left(\frac{1}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L+1}{L-2}}}\right) \\ &\leq \frac{C_2\delta_2^L}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L+1}{L-2}}}, \end{aligned} \tag{73}$$

where C_2 is a sufficiently large constant, and the last inequality follows since $(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L+1}{L-2}} \leq (1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L-1}{L-2}}$ and, for all sufficiently small $\delta > 0$,

$$\|\mathbf{s}_q(0) - \mathbf{z}(0)\|_2 = \left\|\frac{\mathbf{a}_\delta}{\delta} - (1 + C_1\delta^L)\mathbf{w}_*\right\|_2 \leq \left\|\frac{\mathbf{a}_\delta}{\delta} - \mathbf{w}_*\right\|_2 + \left\|C_1\delta^L\mathbf{w}_*\right\|_2 = O(\delta^L).$$

Let $\tau_q(t) = \mathbf{s}_q(t) - \mathbf{z}(t)$, and define

$$h_1(t) = \frac{\eta^{1/(L-2)}}{(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))^{1/(L-2)}}, h_2(t) = \frac{1}{(1 - tL(L-2)\mathcal{N}(\mathbf{w}_*))}.$$

By definition of $\overline{T}_1$,

$$\begin{aligned}1 - \gamma = \frac{\mathbf{w}_*^\top \mathbf{s}_q(\overline{T}_1)}{\|\mathbf{s}_q(\overline{T}_1)\|_2} \geq \frac{\mathbf{w}_*^\top \mathbf{s}_q(\overline{T}_1)}{h_1(\overline{T}_1)} = \frac{\mathbf{w}_*^\top \mathbf{z}(\overline{T}_1) + \mathbf{w}_*^\top \tau_q(\overline{T}_1)}{h_1(\overline{T}_1)} \geq \frac{h_1(\overline{T}_1) - \|\tau_p(\overline{T}_1)\|_2}{h_1(\overline{T}_1)} \\ \geq 1 - \frac{C_2\delta^L/\eta^{1/(L-2)}}{\left(1 - \overline{T}_1\eta L(L-2)\mathcal{N}(\mathbf{w}_*)\right)^{\frac{L}{L-2}}},\end{aligned}$$

which implies, for some sufficiently large constant $\widetilde{C}_2$,

$$1 - \eta\overline{T}_1 L(L-2)\mathcal{N}(\mathbf{w}_*) \leq \frac{C_2^{\frac{L-2}{L}}\delta^{L-2}/\eta^{1/L}}{\gamma^{\frac{L-2}{L}}} \leq \widetilde{C}_2\delta^{L-2}.$$

Therefore, there exists a sufficiently large constant C_3 such that

$$\overline{T}_1 \geq \frac{1 - \widetilde{C}_2\delta^{L-2}}{\eta L(L-2)\mathcal{N}(\mathbf{w}_*)} \geq \frac{1 - C_3\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)},$$

where the second inequality follows since, for all sufficiently small $\delta > 0$,

$$\frac{1 - \widetilde{C}_2\delta^{L-2}}{(1 + C_1\delta^L)^{L-2}} \geq \frac{1 - \widetilde{C}_2\delta^{L-2}}{(1 + 2C_1(L-2)\delta^L)} \geq (1 - \widetilde{C}_2\delta^{L-2})(1 - 2C_1(L-2)\delta^L) \geq 1 - C_3\delta^{L-2},$$

where C_3 is a sufficiently large constant. Hence,

$$\frac{\mathbf{w}_*^\top \mathbf{s}_q(t)}{\|\mathbf{s}_q(t)\|_2} \geq 1 - \gamma, \text{ for all } t \in \left[0, \frac{\left(1 - \delta^{L-2}C_3\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\right]. \tag{74}$$

Next, recall $\mathbf{u}(t) = \boldsymbol{\psi}(t, \delta \mathbf{w}_*)$, and define $\mathbf{s}_u(t) = \frac{1}{\delta}\mathbf{u}\left(\frac{t}{\delta^{L-2}}\right)$. Note that

$$\frac{\mathbf{w}_*^\top \mathbf{s}_u(0)}{\|\mathbf{s}_u(0)\|_2} = 1 > 1 - \gamma.$$

Define

$$\overline{T}_3 \coloneqq \min_{t \geq 0} \left\{ t : \frac{\mathbf{w}_*^\top \mathbf{s}_u(t)}{\|\mathbf{s}_u(t)\|_2} = 1 - \gamma \right\}.$$

Then, following in a similar way as in the proof of $\mathbf{s}_q(t)$ above, we can show that there exists a sufficiently large constant C_4 such that, for all sufficiently small $\delta > 0$,

$$\frac{\mathbf{w}_*^\top \mathbf{s}_u(t)}{\|\mathbf{s}_u(t)\|_2} \geq 1 - \gamma, \text{ for all } t \in \left[0, \frac{\left(1 - \delta^{L-2} C_4\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\right], \text{ and}$$

$$\|\mathbf{s}_u(t)\|_2^{L-2} \leq \frac{\|\mathbf{s}_u(0)\|_2^{L-2}}{1 - t\|\mathbf{s}_u(0)\|_2^{L-2} L(L-2)\mathcal{N}(\mathbf{w}_*)} = \frac{1}{1 - tL(L-2)\mathcal{N}(\mathbf{w}_*)}.$$

Now, choose C_5 large enough such that $C_5^{L/(L-2)} > \zeta 2^{(2L-2)/(L-2)}/(L\mathcal{N}(\mathbf{w}_*))$ and $C_5 \geq \max(C_3, C_4)$, and define $\overline{T}_4 \coloneqq \frac{\left(1-\delta^{L-2}C_5\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}$. Then, for all $t \in \left[0, \overline{T}_4\right]$, we have

$$\begin{aligned}
&\mathbf{w}_*^\top \mathbf{s}_q(t)/\|\mathbf{s}_q(t)\|_2, \mathbf{w}_*^\top \mathbf{s}_u(t)/\|\mathbf{s}_u(t)\|_2 \geq 1 - \gamma, \\
&\|\mathbf{s}_u(t)\|_2^{L-2} \leq \frac{1}{1 - tL(L-2)\mathcal{N}(\mathbf{w}_*)} \leq \frac{1}{C_5 \delta^{L-2}}, \text{ and} \\
&\|\mathbf{s}_q(t)\|_2^{L-2} \leq \frac{\eta}{1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*)} = \frac{1}{1/\eta - 1 + C_5\delta^{L-2}} \leq \frac{2}{C_5\delta^{L-2}},
\end{aligned}$$

where the last inequality follows since, for all sufficiently small $\delta > 0$,

$$\frac{1}{\eta} = \frac{1}{(1 + C_1\delta^L)^{L-2}} \geq \frac{1}{1 + 2C_1(L-2)\delta^L} \geq 1 - 2C_1(L-2)\delta^L \geq 1 - C_5\delta^{L-2}/2. \tag{75}$$

Therefore, for $t \in [0, \overline{T}_4]$, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\mathbf{s}_q(t) - \mathbf{s}_u(t)\|_2^2 = (\mathbf{s}_q(t) - \mathbf{s}_u)^\top (\dot{\mathbf{s}}_q - \dot{\mathbf{s}}_u) \\
&= (\mathbf{s}_q(t) - \mathbf{s}_u)^\top (\nabla\mathcal{N}(\mathbf{s}_q) - \nabla\mathcal{N}(\mathbf{s}_u)) - (\mathbf{s}_q - \mathbf{s}_u)^\top (f(\delta\mathbf{s}_q) - f(\delta\mathbf{s}_u))/\delta^{L-1} \\
&\leq (L(L-1/2)\mathcal{N}(\mathbf{w}_*))\|\mathbf{s}_q(t) - \mathbf{s}_u(t)\|_2^2 \max(\|\mathbf{s}_q(t)\|_2^{L-2}, \|\mathbf{s}_u(t)\|_2^{L-2}) + \\
&\quad \zeta\delta^L \|\mathbf{s}_q(t) - \mathbf{s}_u(t)\|_2^2 \max(\|\mathbf{s}_u(t)\|_2^{2L-2}, \|\mathbf{s}_q(t)\|_2^{2L-2}) \\
&\leq \left(L(L-1/2)\mathcal{N}(\mathbf{w}_*) + \frac{\zeta 2^{L/(L-2)}}{C_5^{L/(L-2)}}\right)\|\mathbf{s}_q(t) - \mathbf{s}_u\|_2^2 \max(\|\mathbf{s}_q(t)\|_2^{L-2}, \|\mathbf{s}_u(t)\|_2^{L-2}) \\
&\leq L^2\mathcal{N}(\mathbf{w}_*)\|\mathbf{s}_q(t) - \mathbf{s}_u\|_2^2 \max(\|\mathbf{s}_q(t)\|_2^{L-2}, \|\mathbf{s}_u(t)\|_2^{L-2}),
\end{aligned}$$

where the first inequality follows from eq. (66), $\nabla^2\mathcal{N}(\mathbf{w})$ being $(L-2)$-homogeneous and eq. (69). The second inequality uses $\delta^L \max(\|\mathbf{s}_q(t)\|_2^L, \|\mathbf{s}_u(t)\|_2^L) \leq 2^{L/(L-2)}/C_5^{L/(L-2)}$, for all

$t \in [0, \overline{T}_4]$. The last inequality follows from our assumption on C_5. Now, let $\kappa = L^2\mathcal{N}(\mathbf{w}_*)$, then

$$\frac{d}{dt}\left\|\mathbf{s}_q(t) - \mathbf{s}_u(t)\right\|_2 \leq \frac{\kappa\eta\left\|\mathbf{s}_q(t) - \mathbf{s}_u\right\|_2}{1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*)}.$$

Using Lemma 29, we get

$$\left\|\mathbf{s}_q(t) - \mathbf{s}_u(t)\right\|_2 \leq \frac{1}{P_1(t)}\left(\left\|\mathbf{s}_q(0) - \mathbf{s}_u(0)\right\|_2\right),$$

where $P_1(t) = e^{-\int_0^t b_1(s)ds}$ and $b_1(t) = \frac{\eta\kappa}{(1-t\eta L(L-2)\mathcal{N}(\mathbf{w}_*))}$. From Lemma 28, we know

$$P_1(t) = \left(1 - t\eta L(L-2)\mathcal{N}(\mathbf{w}_*)\right)^{L/(L-2)}.$$

Now, using eq. (75), there exists a constant C_6 such that

$$P_1(\overline{T}_4) = \left(1 - \eta(1 - \delta^{L-2}C_5)\right)^{L/(L-2)} \geq \left(\frac{\eta C_5\delta^{L-2}}{2}\right)^{L/(L-2)} \geq C_6\delta^L.$$

Hence,

$$\left\|\mathbf{s}_q(\overline{T}_4) - \mathbf{s}_u(\overline{T}_4)\right\|_2 \leq \frac{1}{C_6\delta^L}\left(\left\|\mathbf{a}_\delta/\delta - \mathbf{w}_*\right\|_2\right) \leq C_1/C_6,$$

which implies

$$\left\|\boldsymbol{\psi}\left(\frac{\left(1/\delta^{L-2} - C_5\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \mathbf{a}_\delta\right) - \boldsymbol{\psi}\left(\frac{\left(1/\delta^{L-2} - C_5\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right\|_2 \leq C_1\delta/C_6.$$

Since $\widetilde{T}$ is fixed, from Lemma 26 and the above inequality, there exists a $C > 0$ such that for all sufficiently small δ, we have

$$\left\|\boldsymbol{\psi}\left(t + \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\left(\frac{1}{\delta^{L-2}}\right), \mathbf{a}_\delta\right) - \boldsymbol{\psi}\left(t + \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\left(\frac{1}{\delta^{L-2}}\right), \delta\mathbf{w}_*\right)\right\|_2 \leq C\delta,$$

for all $t \in [-\widetilde{T}, \widetilde{T}]$, which completes the proof. ■

Lemma 31 *Consider the setting of Theorem 4. For any fixed $t \in (-\infty, \infty)$ and all sufficiently small $\delta_2 \geq \delta_1 > 0$, there exists a $C > 0$ such that*

$$\left\|\boldsymbol{\psi}\left(t + \frac{1/\delta_1^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta_1\mathbf{w}_*\right) - \boldsymbol{\psi}\left(t + \frac{1/\delta_2^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta_2\mathbf{w}_*\right)\right\|_2 \leq C\delta_2,$$

which implies $\mathbf{p}(t)$ exists for all $t \in (-\infty, \infty)$. Furthermore, let $\delta_1 \to 0$ and $\delta_2 = \delta$, then

$$\left\|\mathbf{p}(t) - \boldsymbol{\psi}\left(t + \frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right\|_2 \leq C\delta.$$

Also, $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, for some $\eta > 0$.

Proof We choose $\gamma > 0$ sufficiently small such that for all unit-norm vectors $\mathbf{w}$ satisfying $\mathbf{w}^\top \mathbf{w}_* \geq 1-\gamma$, we have $\mathcal{N}(\mathbf{w}) \leq \mathcal{N}(\mathbf{w}_*)$, and for all $t_2 \geq t_1 \geq 0$,

$$\left(\nabla \mathcal{N}(t_1\mathbf{w}) - \nabla \mathcal{N}(t_2\mathbf{w}_*)\right)^\top (t_1\mathbf{w} - t_2\mathbf{w}_*) \leq L(L-1)\mathcal{N}(\mathbf{w}_*) t_2^{L-2} \|t_1\mathbf{w} - t_2\mathbf{w}_*\|_2^2.$$

The first and second inequalities follow from Lemma 23 and Lemma 22, respectively. For the sake of brevity, let $\mathbf{u}_1(t) = \boldsymbol{\psi}(t, \delta_1\mathbf{w}_*), \mathbf{u}_2(t) = \boldsymbol{\psi}(t, \delta_2\mathbf{w}_*)$ and $\mathbf{s}_1(t) = \frac{1}{\delta_1}\mathbf{u}_1\left(\frac{t}{\delta_1^{L-2}}\right), \mathbf{s}_2(t) = \frac{1}{\delta_2}\mathbf{u}_2\left(\frac{t}{\delta_2^{L-2}}\right)$, where recall that $\delta_2 \geq \delta_1 > 0$. Also, let

$$\mathbf{z}(t) = \frac{\mathbf{w}_*}{(1 - tL(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{1}{L-2}}},$$

which is the solution of

$$\dot{\mathbf{z}} = \nabla \mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_*.$$

Now, $\mathbf{s}_1(0) = \mathbf{w}_*$. Define

$$T_1^* = \min_{t \geq 0} \left\{ t : \frac{\mathbf{w}_*^\top \mathbf{s}_1(t)}{\|\mathbf{s}_1(t)\|_2} = 1 - \gamma \right\}.$$

Then, following a a similar way as in the proof of Lemma 30 to get eq. (73), eq. (74), and eq. (72), we can show that there exists a $C_1, C_2 > 0$ such that, for all $t \in [0, \frac{\left(1-\delta_1^{L-2}C_1\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}]$,

$$\|\mathbf{s}_1(t) - \mathbf{z}(t)\|_2 \leq \frac{C_2\delta_1^L}{(1 - tL(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L+1}{L-2}}}, \frac{\mathbf{w}_*^\top \mathbf{s}_1(t)}{\|\mathbf{s}_1(t)\|_2} \geq 1 - \gamma, \text{ and} \tag{76}$$

$$\|\mathbf{s}_1(t)\|_2^{L-2} \leq \frac{1}{1 - tL(L-2)\mathcal{N}(\mathbf{w}_*)},$$

for all sufficiently small $\delta_1 > 0$. Now, suppose $\delta_2^{L-2} \leq \frac{1}{C_1}$, then

$$T_2^* \coloneqq \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\left(1 - \frac{\delta_1^{L-2}}{\delta_2^{L-2}}\right) \leq \frac{\left(1 - \delta_1^{L-2}C_1\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}.$$

Next, note that

$$\delta_1 \mathbf{z}\left(T_2^*\right) = \frac{\mathbf{w}_*}{(1 - T_2^* L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{1}{L-2}}} = \frac{\delta_1 \mathbf{w}_*}{(1 - (1 - \delta_1^{L-2}/\delta_2^{L-2}))^{\frac{1}{L-2}}} = \delta_2 \mathbf{w}_*, \text{ and}$$

$$\|\mathbf{s}_1(T_2^*) - \mathbf{z}(T_2^*)\|_2 \leq \left(\frac{C_2\delta_1^L}{(1 - T_2^* L(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L+1}{L-2}}}\right) = \frac{C_2\delta_2^{L+1}}{\delta_1}.$$

Hence,

$$\left\| \boldsymbol{\psi}\left(\frac{1/\delta_1^{L-2} - 1/\delta_2^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta_1\mathbf{w}_*\right) - \delta_2\mathbf{w}_* \right\|_2 = \|\delta_1\mathbf{s}_1(T_2^*) - \delta_1\mathbf{z}(T_2^*)\|_2 \leq C_2\delta_2^{L+1}. \tag{77}$$

Now, note that

$$\boldsymbol{\psi}\left(t+\frac{1/\delta_1^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)},\delta_1\mathbf{w}_*\right)=\boldsymbol{\psi}\left(t+\frac{1/\delta_2^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)},\boldsymbol{\psi}\left(\frac{1/\delta_1^{L-2}-1/\delta_2^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)},\delta_1\mathbf{w}_*\right)\right).$$

Combining the above equality with eq. (77) and Lemma 30, we get that for any fixed $t\in(-\infty,\infty)$, there exist a constant $C>0$ such that

$$\left\|\boldsymbol{\psi}\left(t+\frac{1/\delta_1^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)},\delta_1\mathbf{w}_*\right)-\boldsymbol{\psi}\left(t+\frac{1/\delta_2^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)},\delta_2\mathbf{w}_*\right)\right\|_2\leq C\delta_2,$$

which implies $\mathbf{p}(t)$ exists for all $t\in(-\infty,\infty)$.

We next prove that $\mathcal{L}(\mathbf{p}(0))\leq\mathcal{L}(\mathbf{0})-\eta$, for some $\eta>0$. Let $\mathbf{u}(t)=\boldsymbol{\psi}(t,\delta\mathbf{w}_*)$ and $\mathbf{z}(t)=\mathbf{w}_*/(1-tL(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{1}{L-2}}$. From eq. (76), there exist constants B_1,B_2 such that

$$\left\|\frac{\mathbf{u}(t/\delta^{L-2})}{\delta}-\mathbf{z}(t)\right\|_2\leq\frac{B_2\delta^L}{(1-tL(L-2)\mathcal{N}(\mathbf{w}_*))^{\frac{L+1}{L-2}}},\text{ for all }t\in\left[0,\frac{\left(1-B_1\delta^{L-2}\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\right].\quad(78)$$

Define $\alpha\coloneqq\max_{\|\mathbf{w}\|_2=1}\|\mathcal{H}(\mathbf{X},\mathbf{w})\|_2$. Similar to eq. (51), we have

$$\mathcal{L}(\mathbf{w})\leq\mathcal{L}(\mathbf{0})+Kn\|\mathcal{H}(\mathbf{X},\mathbf{w})\|_2^2-\mathcal{N}(\mathbf{w})\leq\mathcal{L}(\mathbf{0})+\alpha^2Kn\|\mathbf{w}\|_2^{2L}-\mathcal{N}(\mathbf{w}).\quad(79)$$

Now, choose $\epsilon>0$ sufficiently small such that

$$\epsilon\leq\frac{1}{B_1},\text{ and }\alpha^2Kn(\epsilon^{\frac{1}{2(L-2)}}+B_1\epsilon^{\frac{L+1/2}{L-2}})^{2L}\leq\mathcal{N}(\mathbf{w})-\frac{\mathcal{N}(\mathbf{w}_*)}{2},\text{ if }\|\mathbf{w}-\mathbf{w}_*\|_2\leq\epsilon^{\frac{L}{L-2}}B_1,$$

where the second inequality holds true since $\mathcal{N}(\mathbf{w})$ is continuous. Note that

$$\frac{\left(1-B_1\delta^{L-2}\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\geq\frac{\left(1-\delta^{L-2}/\epsilon\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}\coloneqq T_2.$$

Hence, using eq. (78), $\|\mathbf{u}(T_2/\delta^{L-2})\|_2\leq\epsilon^{1/(L-2)}+B_2\epsilon^{(L+1)/(L-2)}$. Also, if we define $\tau\coloneqq\mathbf{u}(T_2/\delta^{L-2})-\delta\mathbf{z}(T_2/\delta^{L-2})$, then, $\|\tau\|_2\leq B_2\epsilon^{(L+1)/(L-2)}$. Therefore, using eq. (79), we get

$$\begin{aligned}
\mathcal{L}\left(\mathbf{u}\left(\frac{T_2}{\delta^{L-2}}\right)\right)&\leq\mathcal{L}(\mathbf{0})+\alpha^2Kn\left\|\mathbf{u}\left(\frac{T_2}{\delta^{L-2}}\right)\right\|_2^{2L}-\mathcal{N}\left(\mathbf{u}\left(\frac{T_2}{\delta^{L-2}}\right)\right),\\
&\leq\mathcal{L}(\mathbf{0})+\alpha^2Kn(\epsilon^{1/(L-2)}+B_2\epsilon^{(L+1)/(L-2)})^{2L}-\mathcal{N}\left(\mathbf{u}\left(\frac{T_2}{\delta^{L-2}}\right)\right)\\
&=\mathcal{L}(\mathbf{0})+\epsilon^{\frac{L}{L-2}}\left(\alpha^2Kn(\epsilon^{\frac{1}{2(L-2)}}+B_1\epsilon^{\frac{L+1/2}{L-2}})^{2L}-\mathcal{N}\left(\frac{1}{\epsilon^{\frac{1}{L-2}}}\mathbf{u}\left(\frac{T_2}{\delta^{L-2}}\right)\right)\right)\\
&\leq\mathcal{L}(\mathbf{0})-\epsilon^{\frac{L}{L-2}}\frac{\mathcal{N}(\mathbf{w}_*)}{2},
\end{aligned}$$

where the last inequality follows from the choice of ϵ and since $\|\mathbf{u}(T_2/\delta^{L-2})/\epsilon^{1/(L-2)}-\mathbf{w}_*\|_2=\tau/\epsilon^{1/(L-2)}\leq B_1\epsilon^{L/(L-2)}$. Let $\eta=\epsilon^{L/(L-2)}\mathcal{N}(\mathbf{w}_*)/2$, then, from the above equation,

$$\mathcal{L}\left(\boldsymbol{\psi}\left(\frac{\left(1/\delta^{L-2}-\epsilon\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)},\delta\mathbf{w}_*\right)\right)\leq\mathcal{L}(\mathbf{0})-\eta,$$

where note that $\epsilon > 0$ and fixed, and thus, η is fixed. Since the loss decreases with time, the above equation implies

$$\mathcal{L}\left(\boldsymbol{\psi}\left(\frac{1/\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right) \leq \mathcal{L}\left(\boldsymbol{\psi}\left(\frac{\left(1/\delta^{L-2}-\epsilon\right)}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \delta\mathbf{w}_*\right)\right) \leq \mathcal{L}(\mathbf{0}) - \eta.$$

Taking $\delta \to 0$, gives us $\mathcal{L}(\mathbf{p}(0)) \leq \mathcal{L}(\mathbf{0}) - \eta$, which completes the proof. ■

Lemma 32 *Consider the setting in Theorem 10. There exists $T_1, a_1 \geq 0$ such that for all sufficiently small $\delta > 0$,*

$$\left\|\boldsymbol{\psi}\left(\frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}}\left(1 - \widetilde{\delta}^{\frac{2(L-2)L^2\mathcal{N}(\mathbf{w}_*)}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}\right), \delta\mathbf{w}_0\right) - b_\delta \widetilde{\delta}^{\frac{\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}\mathbf{w}_*\right\|_2 \leq a_1 \widetilde{\delta}^{\frac{(L+1)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}},$$

where $\widetilde{\delta} \in (A_2\delta - A_1\delta^{L+1}, A_2\delta + A_1\delta^{L+1})$, for some positive constants A_2, A_1. Also, $T \geq \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_*)}$, and $b_\delta^{L-2} \in \left[\frac{1}{TL(L-2)\mathcal{N}(\mathbf{w}_*)}, 1\right]$ depends on δ.

Proof Choose $\alpha \in (0,1)$ sufficiently small such that

$$0 < \frac{(L(L-1)\mathcal{N}(\mathbf{w}_*) + \alpha)(1+\alpha)^{L-2}}{L(L-2)(\mathcal{N}(\mathbf{w}_*) - \alpha)} + 1 < \frac{2L-1}{L-2}. \tag{80}$$

Next, we choose $\gamma > 0$ sufficiently small such that for all unit-norm vector $\mathbf{w}$ satisfying $\mathbf{w}^\top\mathbf{w}_* \geq 1-\gamma$, we have

$$\mathcal{N}(\mathbf{w}_*) - \alpha \leq \mathcal{N}(\mathbf{w}) \leq \mathcal{N}(\mathbf{w}_*), \|\nabla^2\mathcal{N}(\mathbf{w})\|_2 \leq L(L-1)\mathcal{N}(\mathbf{w}_*) + \alpha, \tag{81}$$

$$\mathbf{w}_*^\top\nabla\mathcal{N}(\mathbf{w}) - L\mathcal{N}(\mathbf{w})\mathbf{w}_*^\top\mathbf{w} - \frac{\Delta}{2}\|\mathbf{w}-\mathbf{w}_*\|_2^2 \geq 0, \tag{82}$$

and for all $t_2 \geq t_1 \geq 0$,

$$\left(\nabla\mathcal{N}(t_1\mathbf{w}) - \nabla\mathcal{N}(t_2\mathbf{w}_*)\right)^\top(t_1\mathbf{w} - t_2\mathbf{w}_*) \leq L(L-1)\mathcal{N}(\mathbf{w}_*)t_2^{L-2}\|t_1\mathbf{w} - t_2\mathbf{w}_*\|_2^2. \tag{83}$$

The first and second inequalities follow from Lemma 23 and Lemma 20. The third and fourth inequalities follow from Lemma 23 and Lemma 22. Define $f(\mathbf{w}) = \mathcal{J}(\mathbf{X};\mathbf{w})^\top(\ell'(\mathcal{H}(\mathbf{X};\mathbf{w}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y}))$, then, as shown in the proof of Lemma 30, there exists a $\beta > 0$ such that

$$\|f(\mathbf{w})\|_2 \leq \beta\|\mathbf{w}\|_2^{2L-1}, \text{ for all } \mathbf{w} \in \mathbb{R}^d. \tag{84}$$

Let $\mathbf{z}_1(t)$ denote the solution of

$$\dot{\mathbf{z}} = \nabla\mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_0. \tag{85}$$

Since $\mathbf{w}_0 \in \mathcal{S}(\mathbf{w}_*)$, $\mathbf{z}_1(t)$ converges to $\mathbf{w}_*$ in direction. Thus, we can assume there exists some time T_γ and a constant B_1 such that

$$\frac{\mathbf{w}_*^\top\mathbf{z}_1(T_\gamma)}{\|\mathbf{z}_1(T_\gamma)\|_2} = 1 - \gamma^2/8 \text{ and } \|\mathbf{z}_1(t)\|_2 \leq B_1, \text{ for all } t \in [0, T_\gamma].$$

Define $\mathbf{w}(t) = \boldsymbol{\psi}(t, \delta \mathbf{w}_0)$ and $\mathbf{s}_w(t) = \frac{1}{\delta}\mathbf{w}\left(\frac{t}{\delta^{L-2}}\right)$, then $\mathbf{s}_w(t)$ is the solution of

$$\dot{\mathbf{s}} = \nabla \mathcal{N}(\mathbf{s}) - \mathcal{J}(\mathbf{X}; \mathbf{s})^\top (\ell'(\mathcal{H}(\mathbf{X}; \delta \mathbf{s}), \mathbf{y}) - \ell'(\mathbf{0}, \mathbf{y})) = \nabla \mathcal{N}(\mathbf{s}) - f(\delta \mathbf{s})/\delta^{L-1}, \mathbf{s}(0) = \mathbf{w}_0.$$

Note that $\|\mathbf{s}_w(0) - \mathbf{z}_1(0)\|_2 = 0$. Define

$$T_1^* = \min_{t \geq 0} \{t : \|\mathbf{s}_w(t) - \mathbf{z}_1(t)\|_2 = \gamma^2/4\}.$$

Then, for all $t \in [0, T_1^*]$, $\|\mathbf{s}_w(t) - \mathbf{z}_1(t)\|_2 \leq \gamma^2/4$. Next, we show that for all sufficiently small $\delta > 0$, $T_1^* > T_\gamma$. For the sake of contradiction, suppose $T_1^* \leq T_\gamma$. Let $B_2 = B_1 + \gamma^2/4$, then $\|\mathbf{s}_w(t)\|_2 \leq B_2$, for all $t \in [0, T_1^*]$. Since $\nabla \mathcal{N}(\mathbf{s})$ is locally Lipschitz, there exists a $\mu_1 > 0$ such that

$$\|\nabla \mathcal{N}(\mathbf{s}_w(t)) - \nabla \mathcal{N}(\mathbf{z}_1(t))\|_2 \leq \mu_1 \|\mathbf{s}_w(t) - \mathbf{z}_1(t)\|_2, \text{ for all } t \in [0, T_1^*].$$

Hence, for all $t \in [0, T_1^*]$,

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\mathbf{s}_w(t) - \mathbf{z}_1(t)\|_2^2 &= (\mathbf{s}_w - \mathbf{z}_1)^\top (\dot{\mathbf{s}}_w - \dot{\mathbf{z}}_1) \\
&= (\mathbf{s}_w - \mathbf{z}_1)^\top (\nabla \mathcal{N}(\mathbf{s}_w) - \nabla \mathcal{N}(\mathbf{z}_1)) - (\mathbf{s}_w - \mathbf{z}_1)^\top f(\delta \mathbf{s}_w)/\delta^{L-1} \\
&\leq \mu_1 \|\mathbf{s}_w - \mathbf{z}_1\|_2^2 + \beta \delta^L \|\mathbf{s}_w - \mathbf{z}_1\|_2 \|\mathbf{s}_w\|_2^{2L-1} \\
&\leq \mu_1 \|\mathbf{s}_w - \mathbf{z}_1\|_2^2 + \beta \delta^L B_2^{2L-1} \|\mathbf{s}_w - \mathbf{z}_1\|_2,
\end{aligned}$$

which implies

$$\frac{d}{dt}\|\mathbf{s}_w(t) - \mathbf{z}_1(t)\|_2 \leq \mu_1 \|\mathbf{s}_w - \mathbf{z}_1\|_2 + \beta \delta^L B_2^{2L-1}.$$

Hence, using Lemma 29, we get

$$\|\mathbf{s}_w(t) - \mathbf{z}_1(t)\|_2 \leq \beta \delta^L B_2^{2L-1} e^{\mu_1 t} \int_0^t e^{-\mu_1 s} ds \leq B_3 \delta^L e^{\mu_1 t},$$

where B_3 is a sufficiently large constant. From the definition of T_1^*, we get

$$\gamma^2/4 = \|\mathbf{s}_w(T_1^*) - \mathbf{z}_1(T_1^*)\|_2 \leq B_3 \delta^L e^{\mu_1 T_1^*} \leq B_3 \delta^L e^{\mu_1 T_\gamma},$$

where the second inequality uses $T_1^* \leq T_\gamma$. Clearly, the above inequality is not possible if δ is sufficiently small, leading to a contradiction. Hence, $T_1^* > T_\gamma$. Let $A_1 = B_3 e^{\mu_1 T_\gamma}$, then

$$\|\mathbf{s}_w(T_\gamma) - \mathbf{z}_1(T_\gamma)\|_2 \leq A_1 \delta^L. \tag{86}$$

Next, let $\tau_1(t) = \mathbf{s}_w(t) - \mathbf{z}_1(t)$, then, for all sufficiently small $\delta > 0$, we have

$$\begin{aligned}
\frac{\mathbf{w}_*^\top \mathbf{s}_w(T_\gamma)}{\|\mathbf{s}_w(T_\gamma)\|_2} = \frac{\mathbf{w}_*^\top \mathbf{z}_1(T_\gamma) + \mathbf{w}_*^\top \tau_1(T_\gamma)}{\|\mathbf{s}_w(T_\gamma)\|_2} &\geq \frac{(1 - \gamma^2/8)\|\mathbf{z}_1(T_\gamma)\|_2 - A_1 \delta^L}{\|\mathbf{s}_w(T_\gamma)\|_2} \\
&\geq \frac{(1 - \gamma^2/8)\|\mathbf{z}_1(T_\gamma)\|_2 - A_1 \delta^L}{\|\mathbf{z}_1(T_\gamma)\|_2 + A_1 \delta^L} \geq 1 - \gamma^2/4.
\end{aligned}$$

Let $A_2 = \|\mathbf{z}_1(T_\gamma)\|_2$, where note that A_2 is a constant that does not depend on δ. Define $\mathbf{w}_1 = \mathbf{s}_w(T_\gamma)/\|\mathbf{s}_w(T_\gamma)\|_2$ and $\widetilde{\delta} = \delta\|\mathbf{s}_w(T_\gamma)\|_2$. From eq. (86), we have

$$\widetilde{\delta} \in (A_2\delta - A_1\delta^{L+1}, A_2\delta + A_1\delta^{L+1}),$$

thus, $\widetilde{\delta}$ can be made sufficiently small by choosing δ sufficiently small. Next, let $\mathbf{u}(t) = \boldsymbol{\psi}(t + T_\gamma/\delta^{L-2}, \delta\mathbf{w}_0)$. Then,

$$\mathbf{u}(t) = \boldsymbol{\psi}\left(t + \frac{T_\gamma}{\delta^{L-2}}, \delta\mathbf{w}_0\right) = \boldsymbol{\psi}\left(t, \boldsymbol{\psi}\left(\frac{T_\gamma}{\delta^{L-2}}, \delta\mathbf{w}_0\right)\right) = \boldsymbol{\psi}\left(t, \delta\|\mathbf{s}_w(T_\gamma)\|_2\mathbf{w}_1\right) = \boldsymbol{\psi}\left(t, \widetilde{\delta}\mathbf{w}_1\right).$$

Define $\mathbf{s}_u(t) = \frac{1}{\widetilde{\delta}}\mathbf{u}\left(\frac{t}{\widetilde{\delta}^{L-2}}\right)$, then $\mathbf{s}_u(t)$ is the solution of

$$\dot{\mathbf{s}} = \nabla\mathcal{N}(\mathbf{s}) - \mathcal{J}(\mathbf{X};\mathbf{s})^\top(\ell'(\mathcal{H}(\mathbf{X};\widetilde{\delta}\mathbf{s}),\mathbf{y}) - \ell'(\mathbf{0},\mathbf{y})) = \nabla\mathcal{N}(\mathbf{s}) - f(\widetilde{\delta}\mathbf{s})/\widetilde{\delta}^{L-1}, \mathbf{s}(0) = \mathbf{w}_1.$$

Define $\mathbf{z}_2(t)$ to be the solution of

$$\dot{\mathbf{z}} = \nabla\mathcal{N}(\mathbf{z}), \mathbf{z}(0) = \mathbf{w}_1. \tag{87}$$

Since $\mathbf{z}_2(0)^\top\mathbf{w}_* \geq 1 - \gamma^2/4 > 1 - \gamma$, we may assume without loss of generality that $\gamma > 0$ is sufficiently small such that Lemma 25 is applicable. Hence, there exists a $T > 0$ such that

$$\left(1 - \frac{\mathbf{w}_*^\top\mathbf{z}_2(t)}{\|\mathbf{z}_2(t)\|_2}\right) \leq \gamma\left(1 - \frac{t}{T}\right)^{\frac{\Delta}{L(L-2)\mathcal{N}(\mathbf{w}_*)}} \text{ and } \|\mathbf{z}_2(t)\|_2 = \frac{\|\mathbf{g}(t)\|_2}{(T-t)^{\frac{1}{L-2}}}, \tag{88}$$

for all $t \in [0, T)$, where $\|\mathbf{z}_2(t)\|$ is an increasing function and $\|\mathbf{g}(t)\|_2$ is a decreasing function in $[0, T)$, and $\|\mathbf{g}(0)\|_2^{L-2} = T, \|\mathbf{g}(T)\|_2^{L-2} = 1/(L(L-2)\mathcal{N}(\mathbf{w}_*))$. Also, from Lemma 25 and eq. (81), we have

$$T \leq \frac{1}{L(L-2)\mathcal{N}(\mathbf{w}_1)} \leq \frac{1}{L(L-2)(\mathcal{N}(\mathbf{w}_*) - \alpha)}. \tag{89}$$

Next, define $\eta = \frac{2(L-2)L^2\mathcal{N}(\mathbf{w}_*)}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}$ and $\overline{T} = T(1 - \widetilde{\delta}^\eta)$, then

$$\left(1 - \frac{\mathbf{w}_*^\top\mathbf{z}_2(\overline{T})}{\|\mathbf{z}_2(\overline{T})\|_2}\right) \leq \gamma\widetilde{\delta}^{\frac{2L\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}, \text{ and } \|\mathbf{z}_2(\overline{T})\|_2^{L-2} = \frac{\|\mathbf{g}(\overline{T})\|_2^{L-2}}{T\widetilde{\delta}^\eta}.$$

Therefore,

$$\left\|\frac{\mathbf{z}_2(\overline{T})}{\|\mathbf{z}_2(\overline{T})\|_2} - \mathbf{w}_*\right\|_2^2 = 2\left(1 - \frac{\mathbf{w}_*^\top\mathbf{z}_2(\overline{T})}{\|\mathbf{z}_2(\overline{T})\|_2}\right) \leq 2\gamma\widetilde{\delta}^{\frac{2L\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}. \tag{90}$$

Define $\mu_2 = (L(L-1)\mathcal{N}(\mathbf{w}_*) + \alpha)(1+\alpha)^{L-2}$, and let

$$C_1 = \frac{2\beta(1+\alpha)^{2L-1}}{\left(\frac{L+1}{L-2} - \mu_2\|\mathbf{g}(T)\|_2^{L-2}\right)}\frac{\|\mathbf{g}(0)\|_2^{\frac{2L-1}{L-2}}}{\|\mathbf{g}(T)\|_2^{\frac{L+1}{L-2}}} > 0,$$

where the inequality holds since, from eq. (80), $\frac{L+1}{L-2} - \mu_2 \|\mathbf{g}(T)\|_2^{L-2} > 0$. Now, since $\mathbf{w}_*^\top \mathbf{s}_u(0)/\|\mathbf{s}_u(0)\|_2 \geq 1 - \gamma^2/4$ and $0 = \|\mathbf{s}_u(0) - \mathbf{z}_2(0)\|_2 < C_1 \widetilde{\delta}^L \|\mathbf{z}_2(0)\|_2^{L+1}$, we define

$$T_2^* = \min_{t \geq 0} \left\{ t : \frac{\mathbf{w}_*^\top \mathbf{s}_u(t)}{\|\mathbf{s}_u(t)\|_2} = 1 - \gamma \right\}, T_3^* = \min_{t \geq 0} \{ t : \|\mathbf{s}(t) - \mathbf{z}_2(t)\|_2 = C_1 \widetilde{\delta}^L \|\mathbf{z}_2(t)\|_2^{L+1} \},$$

and $T_4^* = \min(T_2^*, T_3^*)$. Thus, for all $t \in [0, T_4^*]$, we have

$$\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2 \leq C_1 \widetilde{\delta}^L \|\mathbf{z}_2(t)\|_2^{L+1}, \text{ and } \frac{\mathbf{s}_u(t)^\top \mathbf{w}_*}{\|\mathbf{s}_u(t)\|_2} \geq 1 - \gamma.$$

We next show that $\overline{T} \leq T_4^*$, for all sufficiently small $\delta > 0$. For the sake of contradiction, let $\overline{T} > T_4^*$. Now, since $\|\mathbf{z}_2(t)\|_2$ increases with time, for all $t \in [0, T_4^*]$, we have

$$C_1 \widetilde{\delta}^L \|\mathbf{z}_2(t)\|_2^L \leq C_1 \widetilde{\delta}^L \|\mathbf{z}_2(\overline{T})\|_2^L \leq \frac{C_1 \|\mathbf{g}(\overline{T})\|_2^L}{T^{\frac{L}{L-2}}} \widetilde{\delta}^{L - \frac{L\eta}{(L-2)}} = \frac{C_1 \|\mathbf{g}(\overline{T})\|_2^L}{T^{\frac{L}{L-2}}} \widetilde{\delta}^{\frac{L\Delta}{2L^2 \mathcal{N}(\mathbf{w}_*) + \Delta}} \leq \alpha,$$

for all sufficiently small $\delta > 0$. Since $\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2 \leq C_1 \widetilde{\delta}^L \|\mathbf{z}_2(t)\|_2^{L+1}$, we get

$$\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2 \leq \alpha \|\mathbf{z}_2(t)\|_2, \text{ for all } t \in [0, T_4^*],$$

which implies

$$0 < (1 - \alpha) \|\mathbf{z}_2(t)\|_2 \leq \|\mathbf{s}_u(t)\|_2 \leq (1 + \alpha) \|\mathbf{z}_2(t)\|_2 . \tag{91}$$

Then, for all $t \in [0, T_4^*]$,

$$\begin{aligned}
\frac{d}{dt} \frac{\mathbf{w}_*^\top \mathbf{s}_u(t)}{\|\mathbf{s}_u(t)\|_2} &= \mathbf{w}_*^\top \left(\mathbf{I} - \frac{\mathbf{s}_u \mathbf{s}_u^\top}{\|\mathbf{s}_u\|_2^2} \right) \frac{\dot{\mathbf{s}}_u}{\|\mathbf{s}_u\|_2} \\
&= \mathbf{w}_*^\top \left(\mathbf{I} - \frac{\mathbf{s}_u \mathbf{s}_u^\top}{\|\mathbf{s}_u\|_2^2} \right) \frac{\nabla \mathcal{N}(\mathbf{s}_u)}{\|\mathbf{s}_u\|_2} - \mathbf{w}_*^\top \left(\mathbf{I} - \frac{\mathbf{s}_u \mathbf{s}_u^\top}{\|\mathbf{s}_u\|_2^2} \right) \frac{f(\widetilde{\delta} \mathbf{s}_u)}{\widetilde{\delta}^{L-1} \|\mathbf{s}_u\|_2} \\
&\geq \frac{\mathbf{w}_*^\top \nabla \mathcal{N}(\mathbf{s}_u)}{\|\mathbf{s}_u\|_2} - \frac{L \mathbf{w}_*^\top \mathbf{s}_u \mathcal{N}(\mathbf{s}_u)}{\|\mathbf{s}_u\|_2^3} - \beta \widetilde{\delta}^L \|\mathbf{s}_u\|_2^{2L-2} \\
&= \|\mathbf{s}_u\|_2^{L-2} \left(\mathbf{w}_*^\top \nabla \mathcal{N} \left(\frac{\mathbf{s}_u}{\|\mathbf{s}_u\|_2} \right) - L \mathbf{w}_*^\top \left(\frac{\mathbf{s}_u}{\|\mathbf{s}_u\|_2} \right) \mathcal{N} \left(\frac{\mathbf{s}_u}{\|\mathbf{s}_u\|_2} \right) \right) - \beta \widetilde{\delta}^L \|\mathbf{s}_u\|_2^{2L-2} \\
&\geq \|\mathbf{s}_u\|_2^{L-2} \frac{\Delta}{2} \left\| \frac{\mathbf{s}_u(t)}{\|\mathbf{s}_u(t)\|_2} - \mathbf{w}_* \right\|_2^2 - \beta \widetilde{\delta}^L \|\mathbf{s}_u\|_2^{2L-2} \\
&\geq -\beta \widetilde{\delta}^L \|\mathbf{s}_u\|_2^{2L-2} \geq -\beta (1 + \alpha)^{2L-2} \widetilde{\delta}^L \|\mathbf{z}_2(t)\|_2^{2L-2},
\end{aligned}$$

where the second inequality uses eq. (82), and the last inequality uses eq. (91). Hence,

$$\begin{aligned}
\frac{\mathbf{w}_*^\top \mathbf{s}_u(t)}{\|\mathbf{s}_u(t)\|_2} &\geq \frac{\mathbf{w}_*^\top \mathbf{s}_u(0)}{\|\mathbf{s}_u(0)\|_2} - \beta \widetilde{\delta}^L (1 + \alpha)^{2L-2} \int_0^t \frac{\|\mathbf{g}(s)\|_2^{2L-2} ds}{(T - s)^{\frac{2L-2}{L-2}}} \\
&\geq \frac{\mathbf{w}_*^\top \mathbf{s}_u(0)}{\|\mathbf{s}_u(0)\|_2} - \beta \|\mathbf{g}(0)\|_2^{2L-2} \widetilde{\delta}^L (1 + \alpha)^{2L-2} \int_0^t \frac{ds}{(T - s)^{\frac{2L-2}{L-2}}} \\
&\geq \frac{\mathbf{w}_*^\top \mathbf{s}_u(0)}{\|\mathbf{s}_u(0)\|_2} - \frac{\beta (L-2) \|\mathbf{g}(0)\|_2^{2L-2} \widetilde{\delta}^L (1 + \alpha)^{2L-2}}{L (T - t)^{\frac{L}{L-2}}},
\end{aligned}$$

where the second inequality holds since $\mathbf{g}(t)$ is a decreasing function. Since $\overline{T} > T_4^*$, we get

$$\frac{\mathbf{w}_*^\top \mathbf{s}_u(T_4^*)}{\|\mathbf{s}_u(T_4^*)\|_2} - \frac{\mathbf{w}_*^\top \mathbf{s}_u(0)}{\|\mathbf{s}_u(0)\|_2} \geq -\frac{(L-2)\beta\|\mathbf{g}(0)\|_2^{2L-2}\widetilde{\delta}^L(1+\alpha)^{2L-2}}{L(T-\overline{T})^{\frac{L}{L-2}}}$$
$$= -\frac{(L-2)\beta\|\mathbf{g}(0)\|_2^{2L-2}\widetilde{\delta}^L(1+\alpha)^{2L-2}}{LT^{\frac{L}{L-2}}\widetilde{\delta}^{\frac{\eta L}{L-2}}}. \tag{92}$$

Next, note that for $t \in [0, T_4^*]$, using the mean value theorem, we have

$$\|\nabla\mathcal{N}(\mathbf{s}_u(t)) - \nabla\mathcal{N}(\mathbf{z}_2(t))\|_2 \leq \|\nabla^2\mathcal{N}(\mathbf{r})\|_2 \|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2,$$

where $\mathbf{r} = (1-\lambda)\mathbf{s}_u(t) + \lambda\mathbf{z}_2(t)$, for some $\lambda \in (0,1)$. Since $\nabla^2\mathcal{N}(\cdot)$ is $(L-2)$-homogeneous,

$$\|\nabla^2\mathcal{N}(\mathbf{r})\|_2 = \|\nabla^2\mathcal{N}(\mathbf{r}/\|\mathbf{r}\|_2)\|_2\|\mathbf{r}\|_2^{L-2}$$
$$\leq \|\nabla^2\mathcal{N}(\mathbf{r}/\|\mathbf{r}\|_2)\|_2 \max(\|\mathbf{z}_2(t)\|_2^{L-2}, \|\mathbf{s}_u(t)\|_2^{L-2})$$
$$\leq \|\nabla^2\mathcal{N}(\mathbf{r}/\|\mathbf{r}\|_2)\|_2(1+\alpha)^{L-2}\|\mathbf{z}_2(t)\|_2^{L-2}.$$

From eq. (81), we know $\|\nabla^2\mathcal{N}(\mathbf{r}/\|\mathbf{r}\|_2)\|_2 \leq L(L-1)\mathcal{N}(\mathbf{w}_*) + \alpha$. Thus, for all $t \in [0, T_4^*]$,

$$\|\nabla\mathcal{N}(\mathbf{s}_u(t)) - \nabla\mathcal{N}(\mathbf{z}_2(t))\|_2 \leq (L(L-1)\mathcal{N}(\mathbf{w}_*) + \alpha)(1+\alpha)^{L-2}\|\mathbf{z}_2(t)\|_2^{L-2} \|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2.$$

Recall that $\mu_2 = (L(L-1)\mathcal{N}(\mathbf{w}_*) + \alpha)(1+\alpha)^{L-2}$, then

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2^2 = (\mathbf{s}_u - \mathbf{z}_2)^\top(\dot{\mathbf{s}}_u - \dot{\mathbf{z}}_2)$$
$$= (\mathbf{s}_u - \mathbf{z}_2)^\top(\nabla\mathcal{N}(\mathbf{s}_u) - \nabla\mathcal{N}(\mathbf{z}_2)) - (\mathbf{s}_u - \mathbf{z}_2)^\top f(\widetilde{\delta}\mathbf{s}_u)/\widetilde{\delta}^{L-1}$$
$$\leq \mu_2\|\mathbf{z}_2(t)\|_2^{L-2}\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2^2 + \beta\widetilde{\delta}^L\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2\|\mathbf{s}_u(t)\|_2^{2L-1}$$
$$\leq \frac{\mu_2\|\mathbf{g}(t)\|_2^{L-2}}{(T-t)}\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2^2 + \beta(1+\alpha)^{2L-1}\widetilde{\delta}^L\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2\|\mathbf{z}_2(t)\|_2^{2L-1}.$$

Define $\beta_1 = \beta(1+\alpha)^{2L-1}$, then, from the above equation, we have

$$\frac{d}{dt}\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2 \leq \frac{\mu_2\|\mathbf{g}(t)\|_2^{L-2}}{(T-t)}\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2 + \beta_1\widetilde{\delta}^L\|\mathbf{z}_2(t)\|_2^{2L-1}.$$

Using Lemma 29, we get

$$\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2 \leq \frac{\beta_1\widetilde{\delta}^L}{P(t)}\int_0^t P(s)\|\mathbf{z}_2(t)\|_2^{2L-1}ds,$$

where $P(t) = e^{-\int_0^t b(s)ds}$ and $b(t) = \frac{\mu_2\|\mathbf{g}(t)\|_2^{L-2}}{(T-t)}$. The above equation implies that

$$\frac{\|\mathbf{s}_u(t) - \mathbf{z}_2(t)\|_2}{\|\mathbf{z}_2(t)\|_2^{L+1}} \leq \frac{\beta_1\widetilde{\delta}^L}{P(t)\|\mathbf{z}(t)\|_2^{L+1}}\int_0^t P(s)\|\mathbf{z}(t)\|_2^{2L-1}ds$$
$$= \beta_1\widetilde{\delta}^L\frac{\int_0^t P(s)\|\mathbf{g}(s)\|_2^{\frac{2L-1}{L-2}}/(T-t)^{\frac{2L-1}{L-2}}}{P(t)\|\mathbf{g}(t)\|_2^{\frac{L+1}{L-2}}/(T-t)^{\frac{L+1}{L-2}}}$$
$$\leq \frac{\beta_1\|\mathbf{g}(0)\|_2^{\frac{2L-1}{L-2}}\widetilde{\delta}^L}{\|\mathbf{g}(T)\|_2^{\frac{L+1}{L-2}}}\frac{\int_0^t P(s)/(T-t)^{\frac{2L-1}{L-2}}}{P(t)/(T-t)^{\frac{L+1}{L-2}}},$$

where in the last inequality we used $\|\mathbf{g}(T)\|_2 \le \|\mathbf{g}(s)\|_2 \le \|\mathbf{g}(0)\|_2$, for all $s \in [0,T)$. Let

$$h(t) := \frac{\int_0^t P(s)/(T-s)^{\frac{2L-1}{L-2}}}{P(t)/(T-t)^{\frac{L+1}{L-2}}}, \text{ for all } t \in [0,T).$$

We will later show that $h(t)$ is an increasing function in $[0,T)$ and

$$\lim_{t\to T} h(t) = \frac{1}{\left(\frac{L+1}{L-2} - \mu_2 \|\mathbf{g}(T)\|_2^{L-2}\right)}.$$

Assuming the above statement is true and since $T > \overline{T} > T_4^*$, we have

$$\frac{\|\mathbf{s}_u(T_4^*) - \mathbf{z}_2(T_4^*)\|_2}{\beta_1 \widetilde{\delta}^L \|\mathbf{z}_2(T_4^*)\|_2^{L+1}} \le \frac{h(T_4^*)\|\mathbf{g}(0)\|_2^{\frac{2L-1}{L-2}}}{\|\mathbf{g}(T)\|_2^{\frac{L+1}{L-2}}} \le \frac{h(T)\|\mathbf{g}(0)\|_2^{\frac{2L-1}{L-2}}}{\|\mathbf{g}(T)\|_2^{\frac{L+1}{L-2}}} = \frac{C_1}{2\beta_1}. \tag{93}$$

Now, by definition of T_4^*, either

$$\frac{\mathbf{w}_*^\top \mathbf{s}_u(T_4^*)}{\|\mathbf{s}_u(T_4^*)\|_2} = 1-\gamma \text{ or } \|\mathbf{s}_u(T_4^*) - \mathbf{z}_2(T_4^*)\|_2 = C_1 \widetilde{\delta}^L \|\mathbf{z}_2(T_4^*)\|_2^{L+1}.$$

However, from eq. (92),

$$\begin{aligned}
\frac{\mathbf{w}_*^\top \mathbf{s}_u(T_4^*)}{\|\mathbf{s}_u(T_4^*)\|_2} &\ge \frac{\mathbf{w}_*^\top \mathbf{s}_u(0)}{\|\mathbf{s}_u(0)\|_2} - \frac{(L-2)\beta\|\mathbf{g}(0)\|_2^{2L-2}\widetilde{\delta}^L(1+\alpha)^{2L-2}}{LT^{\frac{L}{L-2}}\widetilde{\delta}^{\frac{\eta L}{L-2}}} \\
&\ge 1 - \gamma^2/4 - \frac{(L-2)\beta\|\mathbf{g}(0)\|_2^{2L-2}\widetilde{\delta}^{\frac{L\Delta}{L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}(1+\alpha)^{2L-2}}{LT^{\frac{L}{L-2}}} > 1 - \gamma/2,
\end{aligned}$$

for all sufficiently small $\delta > 0$. Also, from eq. (93), we have

$$C_1\widetilde{\delta}^L = \frac{\|\mathbf{s}_u(T_4^*) - \mathbf{z}_2(T_4^*)\|_2}{\|\mathbf{z}_2(T_4^*)\|_2^{L+1}} \le C_1\widetilde{\delta}^L/2.$$

Hence, there is a contradiction, which implies $T_4^* \ge \overline{T}$. Therefore,

$$\begin{aligned}
\|\mathbf{u}(\overline{T}/\widetilde{\delta}^{L-2}) - \widetilde{\delta}\|\mathbf{z}_2(\overline{T})\|_2\mathbf{w}_*\|_2 &= \|\widetilde{\delta}\mathbf{s}_u(\overline{T}) - \widetilde{\delta}\|\mathbf{z}_2(\overline{T})\|_2\mathbf{w}_*\|_2 \\
&\le \widetilde{\delta}\|\mathbf{s}_u(\overline{T}) - \mathbf{z}_2(\overline{T})\|_2 + \widetilde{\delta}\|\mathbf{z}_2(\overline{T}) - \|\mathbf{z}_2(\overline{T})\|_2\mathbf{w}_*\|_2 \\
&\le C_1\widetilde{\delta}^{L+1}\|\mathbf{z}_2(\overline{T})\|_2^{L+1} + \widetilde{\delta}\|\mathbf{z}_2(\overline{T})\|_2\|\mathbf{z}_2(\overline{T})/\|\mathbf{z}_2(\overline{T})\|_2 - \mathbf{w}_*\|_2 \\
&\le \frac{C_1\widetilde{\delta}^{L+1}\|\mathbf{g}(\overline{T})\|_2^{L+1}}{T^{\frac{L+1}{L-2}}\widetilde{\delta}^{\frac{\eta(L+1)}{L-2}}} + \frac{\sqrt{2\gamma}\|\mathbf{g}(\overline{T})\|_2\widetilde{\delta}\widetilde{\delta}^{\frac{L\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}}{T^{\frac{1}{L-2}}\widetilde{\delta}^{\frac{\eta}{L-2}}} \\
&\le C_2\widetilde{\delta}^{\frac{(L+1)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} + C_3\widetilde{\delta}^{\frac{(L+1)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}},
\end{aligned}$$

for some constants C_2, C_3 and for all sufficiently small $\delta > 0$, where the third inequality follows from eq. (88) and eq. (90). Also, note that

$$\widetilde{\delta}\|\mathbf{z}_2(\overline{T})\|_2 = \widetilde{\delta}\|\mathbf{g}(\overline{T})\|_2/(T^{\frac{1}{L-2}}\widetilde{\delta}^{\frac{\eta}{L-2}}) = \|\mathbf{g}(\overline{T})\|_2\widetilde{\delta}^{\frac{\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}/T^{\frac{1}{L-2}}.$$

Thus, using the above two equations and the definition of $\mathbf{u}(t)$, we get

$$\left\| \boldsymbol{\psi}(T_\gamma/\delta^{L-2} + \overline{T}/\widetilde{\delta}^{L-2}, \delta\mathbf{w}_0) - \|\mathbf{g}(\overline{T})\|_2 \widetilde{\delta}^{\frac{\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} / T^{\frac{1}{L-2}} \right\|_2 \leq a_1 \widetilde{\delta}^{\frac{(L+1)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}},$$

where a_1 is a sufficiently large constant. Note that, since $\overline{T}$ depends on δ, $\|\mathbf{g}(\overline{T})\|_2$ may also depend on δ, but $\|\mathbf{g}(\overline{T})\|_2^{L-2} \in [1/(L(L-2)\mathcal{N}(\mathbf{w}_*)), T]$.

We next prove $h(t)$ is an increasing function by showing $h'(t) \geq 0$ for $t \in [0, T)$. Note that

$$P'(t) = -\mu_2 \|\mathbf{g}(t)\|_2^{L-2} P(t)/(T-t).$$

From the quotient rule of differentiation, the denominator of $h'(t)$ is $P^2(t)/(T-t)^{\frac{2L+2}{L-2}}$, and the numerator can be written as

$$\frac{P(t)}{(T-t)^{\frac{2L-1}{L-2}}} \frac{P(t)}{(T-t)^{\frac{L+1}{L-2}}} - \left(\int_0^t \frac{P(s)}{(T-s)^{\frac{2L-1}{L-2}}} ds \right) \left(\frac{-\mu_2 \|\mathbf{g}(t)\|_2^{L-2} P(t)}{(T-t)^{\frac{2L-1}{L-2}}} + \frac{L+1}{L-2} \frac{P(t)}{(T-t)^{\frac{2L-1}{L-2}}} \right)$$
$$= \frac{P^2(t)}{(T-t)^{\frac{3L}{L-2}}} \left(1 - \left(\frac{L+1}{L-2} - \mu_2 \|\mathbf{g}(t)\|_2^{L-2} \right) h(t) \right).$$

Let $r(t) = \left(\frac{L+1}{L-2} - \mu_2 \|\mathbf{g}(t)\|_2^{L-2} \right)$, then $h'(t) = \frac{1 - r(t)h(t)}{(T-t)}$, and $r(t)$ is an increasing function since $\|\mathbf{g}(t)\|_2$ is a decreasing function. Next, note that

$$1 - r(t)h(t) = \frac{P(t)/(T-t)^{\frac{L+1}{L-2}} - r(t) \int_0^t P(s)/(T-s)^{\frac{2L-1}{L-2}} ds}{P(t)/(T-t)^{\frac{L+1}{L-2}}}.$$

Now, the denominator is an increasing function since its derivative is

$$\left(\frac{-\mu_2 \|\mathbf{g}(t)\|_2^{L-2} P(t)}{(T-t)^{\frac{2L-1}{L-2}}} + \frac{L+1}{L-2} \frac{P(t)}{(T-t)^{\frac{2L-1}{L-2}}} \right) = \frac{P(t) r(t)}{(T-t)^{\frac{2L-1}{L-2}}} \geq 0,$$

where the inequality is true since, for all $t \in (0, T)$, from eq. (80) and eq. (89), we have

$$\frac{L+1}{L-2} - \mu_2 \|\mathbf{g}(t)\|_2^{L-2} \geq \frac{L+1}{L-2} - \mu_2 \|\mathbf{g}(0)\|_2^{L-2} \geq \frac{L+1}{L-2} - \frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*) - \alpha)} \geq 0.$$

Also, the numerator is a decreasing function since its derivative is

$$\frac{P(t)r(t)}{(T-t)^{\frac{2L-1}{L-2}}} - \frac{P(t)r(t)}{(T-t)^{\frac{2L-1}{L-2}}} - r'(t) \int_0^t \frac{P(s)}{(T-s)^{\frac{2L-1}{L-2}}} ds = -r'(t) \int_0^t \frac{P(s)}{(T-s)^{\frac{2L-1}{L-2}}} ds \leq 0,$$

where the inequality is true since $r(t)$ is an increasing function. Hence, $1 - r(t)h(t)$ is a decreasing function in $(0, T)$. Now, for the sake of contradiction, let $h'(t_1) = -\epsilon < 0$, for some $t_1 \in (0, T)$. Then, we may assume $1 - r(t_1)h(t_1) = -\tilde{\epsilon} < 0$. Since $1 - r(t)h(t)$ is a decreasing function in $(0, T)$, $1 - r(t)h(t) \leq -\tilde{\epsilon}$, for all $t \geq t_1$. Hence,

$$h(t) - h(t_1) \leq \int_{t_1}^t \frac{-\tilde{\epsilon}}{(T-s)} ds = -\tilde{\epsilon} \ln \left(\frac{T-t_1}{T-t} \right), \text{ for all } t \in (t_1, T).$$

Now, if t is chosen sufficiently close to T, then $h(t)$ will become negative, leading to a contradiction. Hence, $h(t)$ is an increasing function in $(0, T)$.

We next compute $\lim_{t\to T} h(t)$. We first show that

$$\lim_{t\to T} \int_0^t P(s)/(T-s)^{\frac{2L-1}{L-2}} = \infty \text{ and } \lim_{t\to T} P(t)/(T-t)^{\frac{L+1}{L-2}} = \infty.$$

From eq. (89), we know $\|\mathbf{g}(t)\|_2^{L-2} \le \|\mathbf{g}(0)\|_2^{L-2} \le \frac{1}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)}$. Hence,

$$\frac{P(t)}{(T-t)^{\frac{L+1}{L-2}}} \ge \frac{e^{-\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)}\int_0^t \frac{ds}{(T-s)}}}{(T-t)^{\frac{L+1}{L-2}}} = \left(\frac{T-t}{T}\right)^{\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)}} (T-t)^{\frac{-L-1}{L-2}}.$$

Combining the above inequality with eq. (80), we get $\lim_{t\to T} P(t)/(T-t)^{\frac{L+1}{L-2}} = \infty$. Next,

$$\begin{aligned}
&\int_0^t P(s)/(T-s)^{\frac{2L-1}{L-2}} ds \ge \frac{1}{T^{\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)}}} \int_0^t (T-s)^{\left(\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)} - \frac{2L-1}{L-2}\right)} ds \\
&= \frac{\left((T-t)^{\left(\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)} - \frac{L+1}{L-2}\right)} - T^{\left(\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)} - \frac{L+1}{L-2}\right)}\right)}{T^{\frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)}} \left(\frac{L+1}{L-2} - \frac{\mu_2}{L(L-2)(\mathcal{N}(\mathbf{w}_*)-\alpha)}\right)}.
\end{aligned}$$

Combining the above inequality with eq. (80), we get $\lim_{t\to T} \int_0^t P(s)/(T-s)^{\frac{2L-1}{L-2}} = \infty$. Thus, using L'Hopital rule, we have

$$\lim_{t\to T} h(t) = \lim_{t\to T} \frac{P(t)/(T-t)^{\frac{2L-1}{L-2}}}{\frac{P(t)}{(T-t)^{\frac{2L-1}{L-2}}}\left(\frac{L+1}{L-2} - \mu_2\|\mathbf{g}(t)\|_2^{L-2}\right)} = \frac{1}{\left(\frac{L+1}{L-2} - \mu_2\|\mathbf{g}(T)\|_2^{L-2}\right)}.$$

■

Proof of Theorem 10: From Lemma 32, we have

$$\left\|\boldsymbol{\psi}\left(\frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}}\left(1 - \widetilde{\delta}^{\frac{2(L-2)L^2\mathcal{N}(\mathbf{w}_*)}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}\right), \delta\mathbf{w}_0\right) - b_\delta \widetilde{\delta}^{\frac{\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} \mathbf{w}_*\right\|_2 \le a_1 \widetilde{\delta}^{\frac{(L+1)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}},$$

for some $T_1, a_1 > 0$. Define $\bar{\delta} = b_\delta \widetilde{\delta}^{\frac{\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}$, then the above equation implies

$$\left\|\boldsymbol{\psi}\left(\frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}}\left(1 - \widetilde{\delta}^{\frac{2(L-2)L^2\mathcal{N}(\mathbf{w}_*)}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}\right), \delta\mathbf{w}_0\right) - \bar{\delta}\mathbf{w}_*\right\|_2 \le a_2 \bar{\delta}^{L+1},$$

where $a_2 \ge a_1/b_\delta^{L+1}$ is a large enough constant. Define $T_\delta := \frac{T_1}{\delta^{L-2}} + \frac{T\left(1-\widetilde{\delta}^{\frac{2(L-2)L^2\mathcal{N}(\mathbf{w}_*)}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}\right)}{\widetilde{\delta}^{L-2}}$, then, using Lemma 30, for any fixed $\widetilde{T} \in (-\infty, \infty)$, there exists a constant $\widetilde{C}_1 > 0$ such that for all sufficiently small $\delta > 0$, we have

$$\left\|\boldsymbol{\psi}\left(t + \frac{1/\bar{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \boldsymbol{\psi}\left(T_\delta, \delta\mathbf{w}_0\right)\right) - \boldsymbol{\psi}\left(t + \frac{1/\bar{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \bar{\delta}\mathbf{w}_*\right)\right\|_2 \le \widetilde{C}_1\bar{\delta},$$

for all $t \in [-\widetilde{T}, \widetilde{T}]$, and from Lemma 31, there exists a $\widetilde{C}_2 > 0$ such that

$$\left\| \mathbf{p}(t) - \boldsymbol{\psi}\left(t + \frac{1/\overline{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \overline{\delta}\mathbf{w}_* \right) \right\|_2 \leq \widetilde{C}_2 \overline{\delta}, \text{ for all } t \in [-\widetilde{T}, \widetilde{T}].$$

Since

$$\begin{aligned}\frac{1/\overline{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} + T_\delta &= \frac{T_1}{\delta^{L-2}} + \frac{1/(b_\delta^{L-2}\widetilde{\delta}^{\frac{(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}})}{L(L-2)\mathcal{N}(\mathbf{w}_*)} + \frac{T}{\widetilde{\delta}^{L-2}} - \frac{T}{\widetilde{\delta}^{\frac{(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}} \\ &= \frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}} + \left(\frac{1}{b_\delta^{L-2}L(L-2)\mathcal{N}(\mathbf{w}_*)} - T \right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}},\end{aligned}$$

it follows that

$$\begin{aligned}&\boldsymbol{\psi}\left(t + \frac{1/\overline{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \boldsymbol{\psi}\left(T_\delta, \delta\mathbf{w}_0\right) \right) \\ &= \boldsymbol{\psi}\left(t + \frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}} + \left(\frac{1}{b_\delta^{L-2}L(L-2)\mathcal{N}(\mathbf{w}_*)} - T \right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}, \delta\mathbf{w}_0 \right).\end{aligned}$$

Hence, for all $t \in [-\widetilde{T}, \widetilde{T}]$ and for all sufficiently small δ, we have

$$\begin{aligned}&\left\| \boldsymbol{\psi}\left(t + \frac{T_1}{\delta^{L-2}} + \frac{T}{\widetilde{\delta}^{L-2}} + \left(\frac{1}{b_\delta^{L-2}L(L-2)\mathcal{N}(\mathbf{w}_*)} - T \right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}, \delta\mathbf{w}_0 \right) - \mathbf{p}(t) \right\|_2 \\ &\leq \left\| \boldsymbol{\psi}\left(t + \frac{1/\overline{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \boldsymbol{\psi}\left(T_\delta, \delta\mathbf{w}_0\right) \right) - \boldsymbol{\psi}\left(t + \frac{1/\overline{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \overline{\delta}\mathbf{w}_* \right) \right\|_2 + \\ &\left\| \boldsymbol{\psi}\left(t + \frac{1/\overline{\delta}^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)}, \overline{\delta}\mathbf{w}_* \right) - \mathbf{p}(t) \right\|_2 \leq \widetilde{C}_1\overline{\delta} + \widetilde{C}_2\overline{\delta} \leq \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}},\end{aligned}$$

where $\widetilde{C}$ is a positive constant. The last inequality is true since $b_\delta \leq 1$, and $\widetilde{\delta} \leq A_2\delta + A_1\delta^{L+1} \leq 2A_2\delta$, for all sufficiently small δ. Thus, the proof is complete. ■

Proof of Corollary 11: Since $\mathbf{p}(t)$ is bounded for all $t \geq 0$, there exists a constant $B > 0$ such that $\|\mathbf{p}(t)\|_2 \leq B$, for all $t \geq 0$. Moreover, since $\mathcal{L}(\cdot)$ has locally Lipschitz gradient, there exists a constant $\widetilde{A} > 0$ such that, if $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \leq 2B$, then

$$\|\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)\|_2 \leq \widetilde{A}\|\mathbf{w}_1 - \mathbf{w}_2\|_2. \tag{94}$$

Since $\mathbf{p}^* = \lim_{t\to\infty}\mathbf{p}(t)$ and $\nabla\mathcal{L}(\mathbf{p}^*) = \mathbf{0}$, for any $\epsilon \in (0, B)$, we can choose a T_ϵ such that

$$\|\mathbf{p}(T_\epsilon) - \mathbf{p}_*\|_2 \leq \epsilon/2 \text{ and } \|\nabla\mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 \leq \epsilon/2. \tag{95}$$

Since T_ϵ does not depend on δ, from Theorem 10, for all sufficiently small $\delta > 0$,

$$\begin{aligned}&\left\| \boldsymbol{\psi}\left(t + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T \right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}, \delta\mathbf{w}_0 \right) - \mathbf{p}(t) \right\|_2 \\ &\leq \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon/2, \text{ for all } t \in [-T_\epsilon, T_\epsilon].\end{aligned} \tag{96}$$

Putting $t = T_\epsilon$ in the above equation, and using eq. (95), we get

$$\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right) - \mathbf{p}^*\|_2 \leq \epsilon,$$

where $T_\delta \coloneqq T_\epsilon + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T\right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}$. Next, since $\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2 \leq B + \epsilon/2 \leq 2B$, using eq. (94), eq. (95) and eq. (96), we get

$$\|\nabla \mathcal{L}\left(\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\right)\|_2 \leq \epsilon/2 + \widetilde{A}\widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon,$$

where the final inequality is true for all sufficiently small $\delta > 0$. This completes the proof. ■

Proof of Corollary 12: Since $\lim_{t\to\infty} \mathbf{p}(t)/\|\mathbf{p}(t)\|_2 = \mathbf{p}^*$, $\lim_{t\to\infty} \|\mathbf{p}(t)\|_2 = \infty$, and $\lim_{t\to\infty} \nabla\mathcal{L}(\mathbf{p}(t)) = \mathbf{0}$, we have that for any $\epsilon \in (0,1)$, we can choose a T_ϵ such that

$$\mathbf{p}(T_\epsilon)^\top \mathbf{p}^*/\|\mathbf{p}(T_\epsilon)\|_2 \geq 1 - \epsilon/2, \|\mathbf{p}(T_\epsilon)\|_2 \geq 1/\epsilon, \text{ and } \|\nabla\mathcal{L}(\mathbf{p}(T_\epsilon))\|_2 \leq \epsilon/2. \tag{97}$$

Let $B_\epsilon \coloneqq \max_{t\in[0,T_\epsilon]} \|\mathbf{p}(t)\|_2$. Then, since $\mathcal{L}(\cdot)$ has locally Lipschitz gradient, there exists a constant $\widetilde{A}_\epsilon > 0$ such that, if $\|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \leq B_\epsilon + \epsilon$, then

$$\|\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)\|_2 \leq \widetilde{A}_\epsilon \|\mathbf{w}_1 - \mathbf{w}_2\|_2. \tag{98}$$

Here, $\widetilde{A}_\epsilon$ depends on ϵ. Since T_ϵ does not depend on δ, from Theorem 10, for all sufficiently small $\delta > 0$ and for all $t \in [-T_\epsilon, T_\epsilon]$,

$$\begin{aligned}
&\left\|\boldsymbol{\psi}\left(t + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T\right) \frac{1}{\widetilde{\delta}^{\frac{(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}} + \frac{T}{\widetilde{\delta}^{L-2}}, \delta\mathbf{w}_0\right) - \mathbf{p}(t)\right\|_2 \\
&\leq \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon/2.
\end{aligned} \tag{99}$$

Putting $t = T_\epsilon$ in the above equation, and using eq. (97), we get

$$\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2 \geq \|\mathbf{p}(T_\epsilon)\|_2 - \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}} \geq \frac{1}{\epsilon} - \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}} \geq \frac{1}{2\epsilon},$$

where $T_\delta \coloneqq T_\epsilon + \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b_\delta^{L-2}}{L(L-2)\mathcal{N}(\mathbf{w}_*)} - T\right) \frac{1}{\widetilde{\delta}^{\frac{(L-2)\Delta}{2L^2\mathcal{N}(\mathbf{w}_*)+\Delta}}} + \frac{T}{\widetilde{\delta}^{L-2}}$, and the second inequality is true for all sufficiently small $\delta > 0$. We also have

$$\frac{\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)^\top \mathbf{p}^*}{\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2} \geq \frac{\mathbf{p}(T_\epsilon)^\top \mathbf{p}_* - \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}}}{\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2} \geq \frac{(1-\epsilon/2)\|\mathbf{p}(T_\epsilon)\|_2 - \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}}}{\|\mathbf{p}(T_\epsilon)\|_2 + \widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}}} \geq 1-\epsilon,$$

where the first inequality uses eq. (99). The second inequality uses eq. (99) and eq. (97). The final inequality is true for all sufficiently small $\delta > 0$. Next, since $\|\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\|_2 \leq B_\epsilon + \epsilon/2$, using eq. (98) and eq. (97), we get

$$\|\nabla \mathcal{L}\left(\boldsymbol{\psi}\left(T_\delta, \delta \mathbf{w}_0\right)\right)\|_2 \leq \epsilon/2 + \widetilde{A}_\epsilon\widetilde{C}\delta^{\frac{\Delta}{\Delta+2L^2\mathcal{N}(\mathbf{w}_*)}} \leq \epsilon,$$

where the final inequality is true for all sufficiently small $\delta > 0$. ■

Appendix D. Proof Omitted from Section 4

Proof of Lemma 16: We first compute the gradient of $\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L)$ with respect to $\mathbf{W}_{l+1}[:, j]$ and $\mathbf{W}_l[j, :]$. Let

$$\begin{aligned}&\phi_0(\mathbf{x}) = \mathbf{x} \text{ and } \phi_l(\mathbf{x}) = \sigma(\mathbf{W}_l \phi_{l-1}(\mathbf{x})), \text{ for all } l \geq 1, \text{ and}\\ &\psi_L(\mathbf{z}) = \mathbf{W}_L \mathbf{z}, \text{ and } \psi_l(\mathbf{z}) = \psi_{l+1}(\sigma(\mathbf{W}_l \mathbf{z})), \text{ for all } l \leq L-1.\end{aligned}$$

Then

$$\begin{aligned}\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L) &= \psi_{l+1}\left(\mathbf{W}_{l+1}\sigma(\mathbf{W}_l \phi_{l-1}(\mathbf{x}))\right)\\ &= \psi_{l+1}\left(\sum_{p=1}^{k_l} \mathbf{W}_{l+1}[:, p]\sigma(\mathbf{W}_l[p, :]^\top \phi_{l-1}(\mathbf{x}))\right),\end{aligned}$$

which implies

$$\frac{d\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L)}{d\mathbf{W}_{l+1}[:, j]} = \nabla \psi_{l+1}\left(\sum_{p=1}^{k_l} \mathbf{W}_{l+1}[:, p]\sigma(\mathbf{W}_l[p, :]^\top \phi_{l-1}(\mathbf{x}))\right)\sigma(\mathbf{W}_l[j, :]^\top \phi_{l-1}(\mathbf{x})),$$

and

$$\begin{aligned}&\frac{d\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L)}{d\mathbf{W}_l[j, :]}\\ &= \mathbf{W}_{l+1}[:, j]^\top \nabla \psi_{l+1}\left(\sum_{p=1}^{k_l} \mathbf{W}_{l+1}[:, p]\sigma(\mathbf{W}_l[p, :]^\top \phi_{l-1}(\mathbf{x}))\right)\sigma'(\mathbf{W}_l[j, :]^\top \phi_{l-1}(\mathbf{x}))\phi_{l-1}(\mathbf{x}).\end{aligned}$$

Now, note that, if the weights are bounded, then there exists a $K > 0$ such that

$$\left\|\frac{d\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L)}{d\mathbf{W}_{l+1}[:, j]}\right\|_2 \leq K\|\mathbf{W}_l[j, :]\|_2, \left\|\frac{d\mathcal{H}(\mathbf{x}; \mathbf{W}_1, \cdots, \mathbf{W}_L)}{d\mathbf{W}_l[j, :]}\right\|_2 \leq K\|\mathbf{W}_{l+1}[:, j]\|_2. \tag{100}$$

The first inequality follows from Cauchy-Schwartz inequality and since $|\sigma(\mathbf{W}_l[j, :]^\top \phi_{l-1}(\mathbf{x}))| \leq K_1|\mathbf{W}_l[j, :]^\top \phi_{l-1}(\mathbf{x})|$, for some $K_1 > 0$, which follows from $\sigma(\cdot)$ being locally Lipshitz and $\sigma(0) = 0$. The second inequality follows from Cauchy-Schwartz inequality.

Now, let $(\mathbf{W}_1(t), \cdots, \mathbf{W}_L(t))$ be the solution of

$$\dot{\mathbf{W}}_i = -\nabla_{\mathbf{W}_i} \mathcal{L}(\mathbf{W}_1, \cdots, \mathbf{W}_L), \text{ for all } i \in [L].$$

If $\mathbf{W}_{l+1}[:, j], \mathbf{W}_l[j, :] \in \mathbf{w}_z$, for some $l \in [L-1]$ and $j \in [k_l]$, then $\|\mathbf{W}_{l+1}[:, j](0)\|_2 = 0 = \|\mathbf{W}_l[j, :](0)\|_2$. Also, note that

$$\begin{aligned}\dot{\mathbf{W}}_{l+1}[:, j] &= -\sum_{i=1}^{n} \ell'\left(\mathcal{H}(\mathbf{x}_i; \mathbf{W}_1, \cdots, \mathbf{W}_L), y_i\right)\frac{d\mathcal{H}(\mathbf{x}_i; \mathbf{W}_1, \cdots, \mathbf{W}_L)}{d\mathbf{W}_{l+1}[:, j]}, \text{ and}\\ \dot{\mathbf{W}}_l[j, :] &= -\sum_{i=1}^{n} \ell'\left(\mathcal{H}(\mathbf{x}_i; \mathbf{W}_1, \cdots, \mathbf{W}_L), y_i\right)\frac{d\mathcal{H}(\mathbf{x}_i; \mathbf{W}_1, \cdots, \mathbf{W}_L)}{d\mathbf{W}_l[j, :]}.\end{aligned}$$

Let $T \in (0, \infty)$, then for all $t \in [-T, T]$, we can assume $(\mathbf{W}_1(t), \cdots, \mathbf{W}_L(t))$ are bounded. Since the output of the neural network and the loss function are bounded for bounded input, from eq. (100), there exists a $B > 0$ such that, for all $t \in [0, T]$, we have

$$\frac{1}{2}\frac{d\|\mathbf{W}_{l+1}[:, j]\|_2^2}{dt} \leq \|\mathbf{W}_{l+1}[:, j]\|_2 \|\dot{\mathbf{W}}_{l+1}[:, j]\|_2 \leq B\|\mathbf{W}_{l+1}[:, j]\|_2 \|\mathbf{W}_l[j, :]\|_2, \text{ and}$$
$$\frac{1}{2}\frac{d\|\mathbf{W}_l[j, :]\|_2^2}{dt} \leq \|\mathbf{W}_l[j, :]\|_2 \|\dot{\mathbf{W}}_l[j, :]\|_2 \leq B\|\mathbf{W}_l[j, :]\|_2 \|\mathbf{W}_{l+1}[:, j]\|_2.$$

Adding the above two equations, and since $2xy \leq x^2 + y^2$, we get

$$\frac{d\left(\|\mathbf{W}_{l+1}[:, j]\|_2^2 + \|\mathbf{W}_l[j, :]\|_2^2\right)}{dt} \leq 2B\left(\|\mathbf{W}_{l+1}[:, j]\|_2^2 + \|\mathbf{W}_l[j, :]\|_2^2\right). \tag{101}$$

Integrating the above equation from 0 to $t \in [0, T]$, we get

$$\|\mathbf{W}_{l+1}[:, j](t)\|_2^2 + \|\mathbf{W}_l[j, :](t)\|_2^2 \leq (\|\mathbf{W}_{l+1}[:, j](0)\|_2^2 + \|\mathbf{W}_l[j, :](0)\|_2^2)e^{2tB}.$$

Since $\|\mathbf{W}_{l+1}[:, j](0)\|_2 = 0 = \|\mathbf{W}_l[j, :](0)\|_2$, the above equation implies $\|\mathbf{W}_{l+1}[:, j](t)\|_2 = 0 = \|\mathbf{W}_l[j, :](t)\|_2$, for all $t \in [0, T]$. Similarly, there exists a $B > 0$ such that, for all $t \in [-T, 0]$, we have

$$\frac{1}{2}\frac{d\|\mathbf{W}_{l+1}[:, j]\|_2^2}{dt} \geq -\|\mathbf{W}_{l+1}[:, j]\|_2 \|\dot{\mathbf{W}}_{l+1}[:, j]\|_2 \geq -B\|\mathbf{W}_{l+1}[:, j]\|_2 \|\mathbf{W}_l[j, :]\|_2, \text{ and}$$
$$\frac{1}{2}\frac{d\|\mathbf{W}_l[j, :]\|_2^2}{dt} \geq -\|\mathbf{W}_l[j, :]\|_2 \|\dot{\mathbf{W}}_l[j, :]\|_2 \geq -B\|\mathbf{W}_l[j, :]\|_2 \|\mathbf{W}_{l+1}[:, j]\|_2.$$

Adding the above two equation gives us

$$\frac{d\left(\|\mathbf{W}_{l+1}[:, j]\|_2^2 + \|\mathbf{W}_l[j, :]\|_2^2\right)}{dt} \geq -2B\left(\|\mathbf{W}_{l+1}[:, j]\|_2^2 + \|\mathbf{W}_l[j, :]\|_2^2\right). \tag{102}$$

Integrating the above equation from $t \in [-T, 0]$ to 0, we get

$$\|\mathbf{W}_{l+1}[:, j](t)\|_2^2 + \|\mathbf{W}_l[j, :](t)\|_2^2 \leq (\|\mathbf{W}_{l+1}[:, j](0)\|_2^2 + \|\mathbf{W}_l[j, :](0)\|_2^2)e^{-2tB},$$

which implies $\|\mathbf{W}_{l+1}[:, j](t)\|_2 = 0 = \|\mathbf{W}_l[j, :](t)\|_2$, for all $t \in [-T, 0]$. ■

Proof of Theorem 18: Note that $\mathbf{w}_z$ is a zero-preserving subset. If $\mathbf{W}_l[j, :] \in \mathbf{w}_z$, for some $l \in [L-1]$ and $j \in [k_l]$, then $\|\overline{\mathbf{W}}_l[j, :]\|_2 = 0$, and from Lemma 17, $\|\overline{\mathbf{W}}_{l+1}[:, j]\|_2 = 0$, which implies $\mathbf{W}_{l+1}[:, j] \in \mathbf{w}_z$. Similarly, if $\mathbf{W}_{l+1}[:, j] \in \mathbf{w}_z$, then $\mathbf{W}_l[j, :] \in \mathbf{w}_z$. Hence, $\mathbf{w}_z$ is a zero-preserving subset.

Throughout this proof, let $\overline{\mathbf{w}}$ denote the vector containing the entries of $(\overline{\mathbf{W}}_1, \cdots, \overline{\mathbf{W}}_L)$, and $\overline{\mathbf{w}}_z$ be the vector containing the entries of $(\overline{\mathbf{W}}_1, \cdots, \overline{\mathbf{W}}_L)$ that belong to $\mathbf{w}_z$. Now, consider the case when $\mathcal{H}$ is two-homogeneous. From Lemma 8, we know

$$\left\| \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\overline{\mathcal{N}}}, \delta\mathbf{W}_{1:L}^0\right) - b_\delta \widetilde{\delta}^{\frac{\Delta}{\Delta + 8\overline{\mathcal{N}}}} \overline{\mathbf{w}} \right\|_2 \leq a_1 \widetilde{\delta}^{\frac{3\Delta}{\Delta + 8\overline{\mathcal{N}}}}, \tag{103}$$

where $T_1, a_1, b_\delta, \widetilde{\delta}$ are as defined in Lemma 8. Since $\|\overline{\mathbf{w}}\|_2 = 1$, $b_\delta \geq \kappa_1 > 0$, and $\widetilde{\delta} \geq A_2\delta/2$ for all sufficiently small δ, there exists a b_1 such that

$$\left\| \boldsymbol{\psi}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\overline{\mathcal{N}}}, \delta \mathbf{W}_{1:L}^0\right)\right\|_2 \geq b_1 \widetilde{\delta}^{\frac{\Delta}{\Delta+8\overline{\mathcal{N}}}}. \tag{104}$$

Now, since $\|\overline{\mathbf{w}}_z\|_2 = 0$, from eq. (103), we have

$$\left\| \boldsymbol{\psi}_{\mathbf{w}_z}\left(T_1 + \frac{4\ln(1/\widetilde{\delta})}{\Delta + 8\overline{\mathcal{N}}}, \delta \mathbf{W}_{1:L}^0\right)\right\|_2 \leq a_1 \widetilde{\delta}^{\frac{3\Delta}{\Delta+8\overline{\mathcal{N}}}}. \tag{105}$$

Dividing eq. (105) by eq. (104) and using $\widetilde{\delta} \leq 2A_2\delta$, for all sufficiently small δ, we get eq. (11). Next, from the definition of zero-preserving subset, we know

$$\|\boldsymbol{\psi}_{\mathbf{w}_z}(t, \delta\overline{\mathbf{W}}_{1:L})\|_2 = 0, \text{ for all } t \in (-\infty, \infty).$$

Recall that

$$\mathbf{p}(t) = \lim_{\delta\to 0} \boldsymbol{\psi}\left(t + \frac{\ln(1/\delta)}{2\overline{\mathcal{N}}}, \delta\overline{\mathbf{W}}_{1:L}\right), \text{ and let } \mathbf{p}_{\mathbf{w}_z}(t) = \lim_{\delta\to 0} \boldsymbol{\psi}_{\mathbf{w}_z}\left(t + \frac{\ln(1/\delta)}{2\overline{\mathcal{N}}}, \delta\overline{\mathbf{W}}_{1:L}\right).$$

From Theorem 4, for all sufficiently small δ and for all $t \in [-\widetilde{T}, \widetilde{T}]$, we have

$$\left\| \boldsymbol{\psi}\left(t + T_1 + \frac{\ln(1/b_\delta)}{2\overline{\mathcal{N}}} + \frac{\ln(1/\widetilde{\delta})}{2\overline{\mathcal{N}}}, \delta \mathbf{W}_{1:L}^0\right) - \mathbf{p}(t)\right\|_2 \leq \widetilde{C}\delta^{\frac{\Delta}{\Delta+8\overline{\mathcal{N}}}}. \tag{106}$$

Since, for any $t \in (-\infty, \infty)$ and $\delta > 0$ we have $\|\boldsymbol{\psi}_{\mathbf{w}_z}(t + \ln(1/\delta)/2\overline{\mathcal{N}}, \delta\overline{\mathbf{W}}_{1:L})\|_2 = 0$, it follows that $\|\mathbf{p}_{\mathbf{w}_z}(t)\|_2 = 0$, for all $t \in (-\infty, \infty)$. Combining this fact with eq. (106) gives us eq. (12).

The proof for L-homogeneous networks is similar, as shown next. From Lemma 32,

$$\left\| \boldsymbol{\psi}\left(T_{\widetilde{\delta}}^1, \delta\mathbf{W}_{1:L}^0\right) - b_\delta \widetilde{\delta}^{\frac{\Delta}{2L^2\overline{\mathcal{N}}+\Delta}} \overline{\mathbf{w}}\right\|_2 \leq a_1 \widetilde{\delta}^{\frac{(L+1)\Delta}{2L^2\overline{\mathcal{N}}+\Delta}}, \tag{107}$$

where $T_{\widetilde{\delta}}^1 = \frac{T_1}{\delta^{L-2}} + \frac{T\left(1-\widetilde{\delta}^{\frac{2(L-2)L^2\overline{\mathcal{N}}}{2L^2\overline{\mathcal{N}}+\Delta}}\right)}{\widetilde{\delta}^{L-2}}$, and $T_1, T, a_1, b_\delta, \widetilde{\delta}$ are as defined in Lemma 32. Since $\|\overline{\mathbf{w}}\|_2 = 1$, $b_\delta^{L-2} \geq 1/(TL(L-2)\overline{\mathcal{N}}) > 0$, and $\widetilde{\delta} \geq A_2\delta/2$, for all sufficiently small $\delta > 0$, there exists a b_1 such that

$$\left\| \boldsymbol{\psi}\left(T_{\widetilde{\delta}}^1, \delta\mathbf{W}_{1:L}^0\right)\right\|_2 \geq b_1 \widetilde{\delta}^{\frac{\Delta}{\Delta+2L^2\overline{\mathcal{N}}}}. \tag{108}$$

Now, since $\|\overline{\mathbf{w}}_z\|_2 = 0$, from eq. (107), we have

$$\left\| \boldsymbol{\psi}_{\mathbf{w}_z}\left(T_{\widetilde{\delta}}^1, \delta\mathbf{W}_{1:L}^0\right)\right\|_2 \leq a_1 \widetilde{\delta}^{\frac{(L+1)\Delta}{\Delta+2L^2\overline{\mathcal{N}}}}. \tag{109}$$

Dividing eq. (109) by eq. (108) and using $\widetilde{\delta} \leq 2A_2\delta$, for all sufficiently small δ, gives us eq. (13). Next, from the definition of zero-preserving subset, we know

$$\|\boldsymbol{\psi}_{\mathbf{w}_z}(t, \delta\overline{\mathbf{W}}_{1:L})\|_2 = 0, \text{ for all } t \in (-\infty, \infty).$$

Recall that

$$\mathbf{p}(t) = \lim_{\delta \to 0} \boldsymbol{\psi}\left(t + \frac{1/\delta^{L-2}}{L(L-2)\overline{\mathcal{N}}}, \delta \overline{\mathbf{W}}_{1:L}\right),$$

and let

$$\mathbf{p}_{\mathbf{w}_z}(t) = \lim_{\delta \to 0} \boldsymbol{\psi}_{\mathbf{w}_z}\left(t + \frac{1/\delta^{L-2}}{L(L-2)\overline{\mathcal{N}}}, \delta \overline{\mathbf{W}}_{1:L}\right).$$

From Theorem 10, for all sufficiently small δ and for all $t \in [-\widetilde{T}, \widetilde{T}]$, we have

$$\left\| \boldsymbol{\psi}\left(t + T^2_{\widetilde{\delta}}, \delta \mathbf{W}^0_{1:L}\right) - \mathbf{p}(t) \right\|_2 \leq \widetilde{C} \delta^{\frac{\Delta}{2L^2\overline{\mathcal{N}}+\Delta}}, \tag{110}$$

where

$$T^2_{\widetilde{\delta}} = \frac{T_1}{\delta^{L-2}} + \left(\frac{1/b^{L-2}_{\delta}}{L(L-2)\overline{\mathcal{N}}} - T\right) \widetilde{\delta}^{\frac{-(L-2)\Delta}{2L^2\overline{\mathcal{N}}\Delta}} + \frac{T}{\widetilde{\delta}^{L-2}}.$$

Since, for any $t \in (-\infty, \infty)$ and $\delta > 0$, $\|\boldsymbol{\psi}_{\mathbf{w}_z}(t + (1/\delta)^{L-2}/(L(L-2)\overline{\mathcal{N}}), \delta \overline{\mathbf{W}}_{1:L})\|_2 = 0$, we have $\|\mathbf{p}_{\mathbf{w}_z}(t)\|_2 = 0$, for all $t \in (-\infty, \infty)$. Combining this fact with the above inequality gives us eq. (14). ■

Appendix E. Additional Results

The following lemma describes an instance where $\boldsymbol{\psi}(t, \delta \mathbf{w}_*)$ does not escape from the origin, if $\mathbf{w}_*$ is zero KKT point of the constrained NCF in eq. (4).

Lemma 33 *Let $\mathcal{H}$ be a single-layer squared ReLU network with single neuron, i.e, $\mathcal{H}(\mathbf{x}; \mathbf{w}) = \max(0, \mathbf{w}^\top \mathbf{x})^2$. Let $\mathbf{w}_* \in \mathbb{S}^{d-1}$ be such that $\mathbf{w}_*^\top \mathbf{x}_i < 0$, for all $i \in [n]$. Then, $\mathbf{w}_*$ is a first-order KKT point of eq. (4) such that $\mathcal{N}(\mathbf{w}_*) = 0$. Also, $\boldsymbol{\psi}(t, \delta \mathbf{w}_*) = \delta \mathbf{w}_*$, for all $t \geq 0$.*

Proof Since $\mathbf{w}_*^\top \mathbf{x}_i < 0$, $\mathcal{H}(\mathbf{x}_i; \mathbf{w}_*) = 0$ and $\nabla \mathcal{H}(\mathbf{x}_i; \mathbf{w}_*) = \mathbf{0}$, for all $i \in [n]$, which implies $\mathcal{J}(\mathbf{X}; \mathbf{w}_*) = \mathbf{0}$. Therefore, $\mathcal{N}(\mathbf{w}_*) = 0$ and $\nabla \mathcal{N}(\mathbf{w}_*) = \mathbf{0}$, implying $\mathbf{w}_*$ is a first-order KKT point of eq. (4).

Next, note that $\nabla \mathcal{L}(\delta \mathbf{w}_*) = \mathcal{J}(\mathbf{X}; \mathbf{w}_*)^\top \ell'(\mathcal{H}(\mathbf{X}; \delta \mathbf{w}_*), \mathbf{y}) = \mathbf{0}$, for all $\delta > 0$. Thus, $\delta \mathbf{w}_*$ is a critical point of the training loss and therefore, $\boldsymbol{\psi}(t, \delta \mathbf{w}_*) = \delta \mathbf{w}_*$. ■

The above lemma essentially implies that if the weights are chosen such that the neuron is inactive for all the training data, then those weights are a first-order KKT point of eq. (4), and moreover, in such cases the gradient flow does not escape from the origin

Proof of Example 1: For square loss, $-\ell'(\mathbf{0}, \mathbf{y}) = 2\mathbf{y}$, hence, the NCF in this example is $\mathcal{N}(\mathbf{w}) = 8w_1^2 + 2w_2^2$. Now, $\boldsymbol{\phi}(t, \mathbf{w}_0) = (\sqrt{2}e^{16t}, \sqrt{2}e^{4t})$, which implies

$$\lim_{t \to \infty} \frac{\boldsymbol{\phi}(t, \mathbf{w}_0)}{\|\boldsymbol{\phi}(t, \mathbf{w}_0)\|_2} = (1, 0) = \mathbf{w}_*.$$

Let $(w_1(t), w_2(t)) = \boldsymbol{\psi}(t, \delta \mathbf{w}_0)$, then

$$\dot{w}_1 = -4w_1(w_1^2 - 4), w_1(0) = \delta/\sqrt{2}, \text{ and } \dot{w}_2 = -4w_2(w_2^2 - 1), w_2(0) = \delta/\sqrt{2}.$$

Therefore, $w_1(t) = \frac{2\delta}{\sqrt{\delta^2+(8-\delta^2)e^{-32t}}}$ and $w_2(t) = \frac{\delta}{\sqrt{\delta^2+(2-\delta^2)e^{-8t}}}$. Next, let $(u_1(t), u_2(t)) = \boldsymbol{\psi}(t, \delta \mathbf{w}_*)$, then

$$\dot{u}_1 = -4u_1(u_1^2 - 4), u_1(0) = \delta, \text{ and } \dot{u}_2 = -4u_2(u_2^2 - 1), u_2(0) = 0.$$

Therefore, $u_1(t) = \frac{2\delta}{\sqrt{\delta^2+(4-\delta^2)e^{-32t}}}, u_2(t) = 0$. Since $\mathcal{N}(\mathbf{w}_*) = 8$,

$$\mathbf{p}(0) = \lim_{\delta \to 0} \left(u_1 \left(\frac{\ln(1/\delta)}{2\mathcal{N}(\mathbf{w}_*)} \right), 0 \right) = \lim_{\delta \to 0} \left(\frac{2\delta}{\sqrt{\delta^2 + (4-\delta^2)\delta^2}}, 0 \right) = (2/\sqrt{5}, 0),$$

which implies $\mathbf{p}(t) = \left(\frac{2}{\sqrt{1+4e^{-32t}}}, 0 \right)$. ■

Difficulty in proving Theorem 4 using Lemma 3: In Remark 14, we briefly discussed why Theorem 4 can not be proved using Lemma 3. Here, we elaborate more on that topic.

Suppose $\mathbf{w}_0 = \mathbf{w}_* + \epsilon \mathbf{b}$, where $\mathbf{b} \in \mathbf{w}_*^{\perp}$, that is, approximate directional convergence holds at the initialization itself. We assume the training loss has a globally Lipschitz gradient with Lipschitz constant $\kappa > 0$ to simplify the argument, though in reality, the gradient is only locally Lipschitz, a weaker condition. Then,

$$\begin{aligned}
&\frac{1}{2} \frac{d\|\boldsymbol{\psi}(t, \delta(\mathbf{w}_* + \epsilon \mathbf{b})) - \boldsymbol{\psi}(t, \delta \mathbf{w}_*)\|_2^2}{dt} \\
&= -(\boldsymbol{\psi}(t, \delta(\mathbf{w}_* + \epsilon \mathbf{b})) - \boldsymbol{\psi}(t, \delta \mathbf{w}_*))^\top (\nabla \mathcal{L}(\boldsymbol{\psi}(t, \delta(\mathbf{w}_* + \epsilon \mathbf{b}))) - \nabla \mathcal{L}(\boldsymbol{\psi}(t, \delta \mathbf{w}_*))) \\
&\leq \kappa \|\boldsymbol{\psi}(t, \delta(\mathbf{w}_* + \epsilon \mathbf{b})) - \boldsymbol{\psi}(t, \delta \mathbf{w}_*)\|_2^2,
\end{aligned}$$

which implies

$$\|\boldsymbol{\psi}(t, \delta(\mathbf{w}_* + \epsilon \mathbf{b})) - \boldsymbol{\psi}(t, \delta \mathbf{w}_*)\|_2 \leq e^{\kappa t} \delta \epsilon. \tag{111}$$

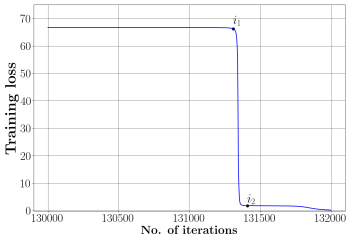
The above bound becomes vacuous after $\ln(1/(\delta\epsilon))/\kappa$ time has elapsed, where note that ϵ is small but *fixed*. However, from Theorem 4, we know that $\boldsymbol{\psi}(t, \delta \mathbf{w}_*)$ escapes from the origin after $\ln(1/(\delta))/(2\mathcal{N}(\mathbf{w}_*))$ time has elapsed. Therefore, the above inequality can be useful in describing the gradient flow dynamics beyond the origin if, for all sufficiently small $\delta > 0$, $\kappa \leq 2\mathcal{N}(\mathbf{w}_*)$, which we are not sure can be proven. The problem is further aggravated if we consider deep homogeneous neural networks. In that case, eq. (111) still holds, but note that $\boldsymbol{\psi}(t, \delta \mathbf{w}_*)$ escapes from the origin after $(1/\delta^{L-2})/(L(L - 2\mathcal{N}(\mathbf{w}_*)))$ time has elapsed, which is always greater than $\ln(1/(\delta\epsilon))/\kappa$, for all all sufficiently small $\delta > 0$ and $L > 2$.

Appendix F. Additional Experiments

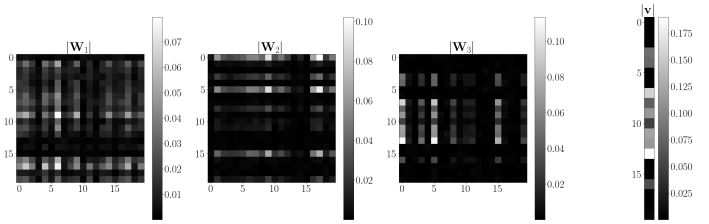
F.1 Four-layer ReLU Neural Network

To further demonstrate the preservation of sparsity structure after weights escape from the origin, we conduct experiments with four-layer ReLU neural networks. The setup closely follows that of the earlier experiments: the weights are trained using gradient descent with small initialization, and training is continued until the weights escape the origin and reach the next saddle point.

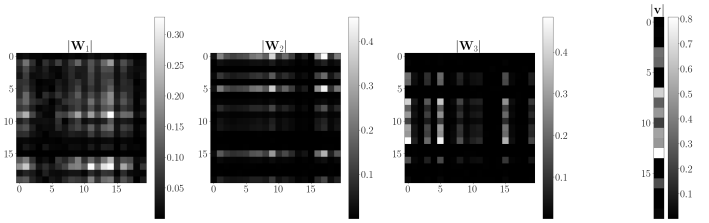
In Figure 7, similar to earlier experiments, the training set consists of 100 points sampled uniformly from the unit sphere in $\mathbb{R}^{20}$, with the corresponding labels generated by a smaller four-layer network. We observe that certain rows and columns of the weight matrices become relatively small before the weights escape the origin, and they remain small even until reaching the next saddle point. Moreover, these rows and columns follow the definition of zero-preserving subset: if the j-th row of $\mathbf{W}_l$ is small, then the j-th column of $\mathbf{W}_{l+1}$ is small, and vice-versa. For instance, the first row of $\mathbf{W}_1$ and first column of $\mathbf{W}_2$, rows 16-18 of $\mathbf{W}_2$ and columns 16-18 of $\mathbf{W}_3$, and rows 17-19 of $\mathbf{W}_3$ and columns 17-19 of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2.



(a) Evolution of training loss with iterations



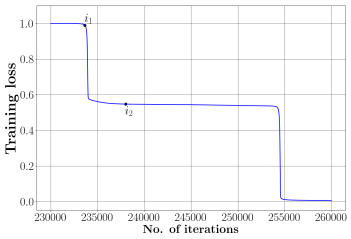
(b) Weights at iteration i_1



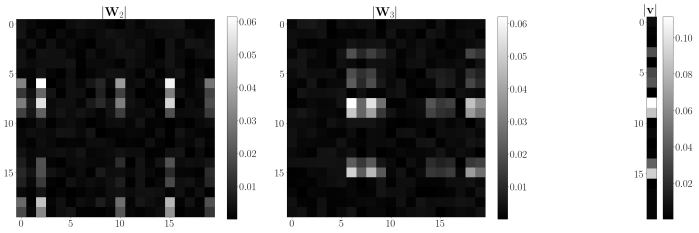
(c) Weights at iteration i_2

Figure 7: We train a four-layer neural network whose output is $\mathbf{v}^\top\sigma(\mathbf{W}_3\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})))$, where $\sigma(x) = \max(x, 0)$, and $\mathbf{v} \in \mathbb{R}^{20}, \mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1 \in \mathbb{R}^{20\times 20}$ are the trainable weights. We observe that rows and columns of the weight matrices that become small near the origin remain small until gradient descent reaches the next saddle point, demonstrating preservation of the sparsity structure.

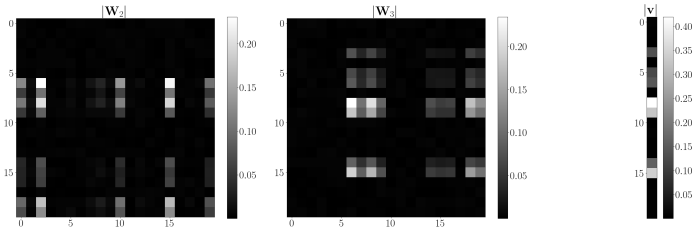
We next consider a more realistic setting in Figure 8, where the task is to classify digits 0 and 1 from the MNIST dataset (LeCun et al., 2010). Digits are labeled as -1 (for 0) and 1 (for 1). The original 28×28 images are down-sampled to size 14×14, and each image is flattened into a vector of length 196. Training is performed using a four-layer ReLU network with square loss. Once again, we observe that certain rows and columns of the weight matrices become relatively small before escaping the origin, and they remain small until reaching the next saddle point. Also, these rows and columns follow the definition of zero-preserving subset. For example, the second row of $\mathbf{W}_1$ and second column of $\mathbf{W}_2$, first five rows of $\mathbf{W}_2$ and first five columns of $\mathbf{W}_3$, and first three rows of $\mathbf{W}_3$ and first three columns of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2.



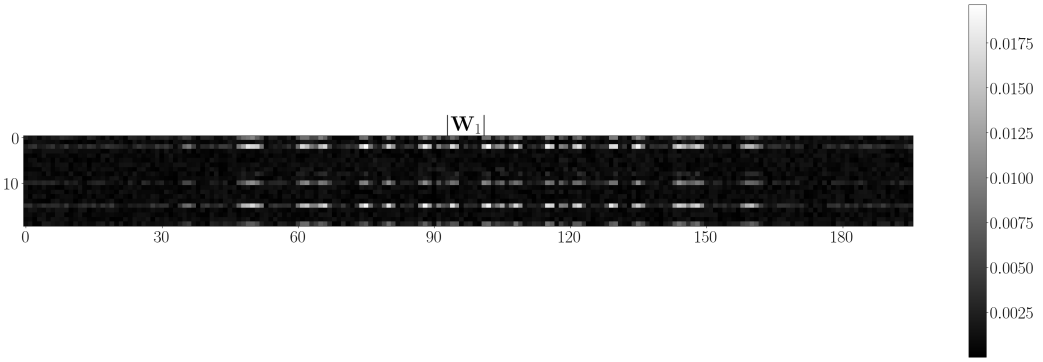
(a) Evolution of training loss with iterations



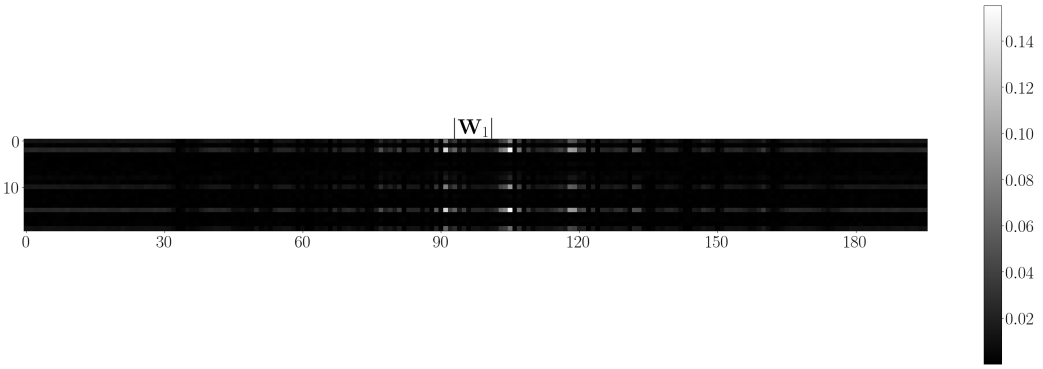
(b) Weights of last three layers at iteration i_1



(c) Weights of last three layers at iteration i_2



(d) Weights of first layer at iteration i_1



(e) Weights of first layer at iteration i_2

Figure 8: We train a four-layer neural network to classify digits 0 and 1 from the MNIST dataset. The images are down-sampled by a factor of 2 (from 28×28 to 14×14) and reshaped into vectors of length 196. The network output is $\mathbf{v}^\top \sigma(\mathbf{W}_3 \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x})))$, where $\sigma(x) = \max(x, 0)$, and $\mathbf{v} \in \mathbb{R}^{20}, \mathbf{W}_3, \mathbf{W}_2 \in \mathbb{R}^{20\times 20}, \mathbf{W}_1 \in \mathbb{R}^{20\times 196}$ are the trainable weights. The rows and columns of the weight matrices that become small near the origin remain small until the next saddle point.

F.2 Non-homogeneous Activation Function

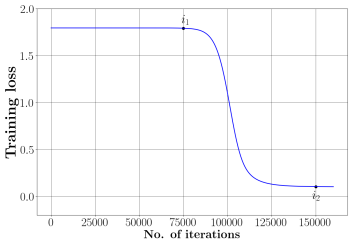
Although our theoretical results are stated for homogeneous activations, we evaluate whether the preservation of the sparsity structure also occurs with non-homogeneous activations, specifically tanh and Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel, 2016). In all experiments the weights are trained using gradient descent with small initialization, and training is continued until the weights escape the origin and reach the next saddle point.

Tanh activation function. We train two-, three-, and four-layer neural network with tanh activation function, where the results are depicted in Figure 9, Figure 10 and Figure 11, respectively. The training set consists of 100 points sampled uniformly from the unit sphere in $\mathbb{R}^{20}$, with the corresponding labels generated by a smaller network. In all three cases we observe the same qualitative behavior seen for homogeneous activations: certain rows and columns of the weight matrices become relatively small before the weights escape from the origin, and they remain small until reaching the next saddle point. Moreover, these rows and columns also follow the definition of zero-preserving subset. For instance, in the two-layer case, first row of $\mathbf{W}_1$ and first column of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2. In the three-layer case, last row of $\mathbf{W}_1$ and last column of $\mathbf{W}_2$, and row 27 of $\mathbf{W}_2$ and column 27 of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2. In the four-layer case, row 14 of $\mathbf{W}_1$ and column 14 of $\mathbf{W}_2$, the row 25 of $\mathbf{W}_2$ and column 25 of $\mathbf{W}_3$, and first row of $\mathbf{W}_3$ and first column of $\mathbf{v}^\top$ stay small at iteration i_1 and i_2.

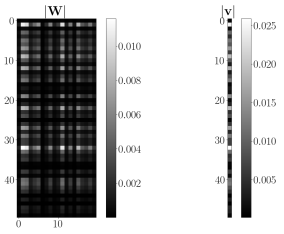
GELU activation function. We train two- and three-layer neural network with GELU activation function, where the results are depicted in Figure 12 and Figure 13, respectively. The training set consists of 100 points sampled uniformly from the unit sphere in $\mathbb{R}^{20}$, with the corresponding labels generated by a smaller network. In both cases, before the weights escape from the origin, certain rows and columns of the weight matrices become relatively small. However, as the weights escape from the origin and training further progresses, it appears that the sparsity among the weights increases, as even the rows and columns which were not small near the origin become small as the training progresses. For example, in the two-layer case, row 40-42 of $\mathbf{W}_1$ is not relatively small at iteration i_1, but it becomes quite small at iteration i_3 and beyond. In the three-layer case, first two rows of $\mathbf{W}_1$ are not relatively small at iteration i_1, but it becomes quite small at iteration i_4 and beyond.

This behavior contrasts with the above experiments using homogeneous activation and tanh activation. While the precise reason behind this is not entirely clear to us, one possible explanation lies in the motivation behind GELU. As discussed in Hendrycks and Gimpel (2016), GELU activation was derived by combining dropout with ReLU. Since dropout is known to encourage sparsity, perhaps this leads to increased sparsity in the weights after escaping from the origin.

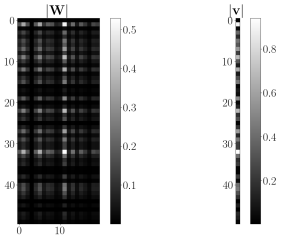
Overall, these set of experiments indicate that the phenomenon of preservation of the sparsity structure extends beyond homogeneous activation functions, though it may not hold in general. Exploring the conditions under which this phenomenon persists is an important direction for future research.



(a) Evolution of training loss with iterations

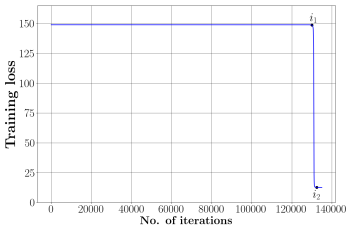


(b) Weights at iteration i_1

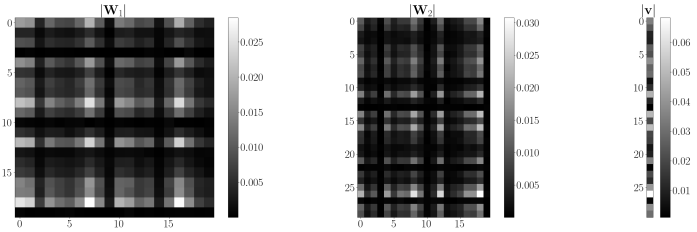


(c) Weights at iteration i_2

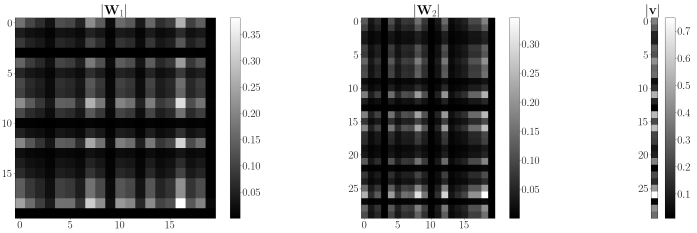
Figure 9: We train a two-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_1 \mathbf{x})$, where $\sigma(x) = \tanh(x)$, and $\mathbf{v} \in \mathbb{R}^{50}, \mathbf{W}_1 \in \mathbb{R}^{50\times 20}$ are the trainable weights. The sparsity structure that emerges among the weights before before escaping the origin is preserved post-escape and until reaching the next saddle point.



(a) Evolution of training loss with iterations

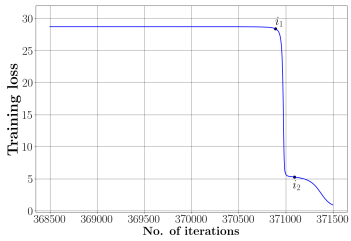


(b) Weights at iteration i_1

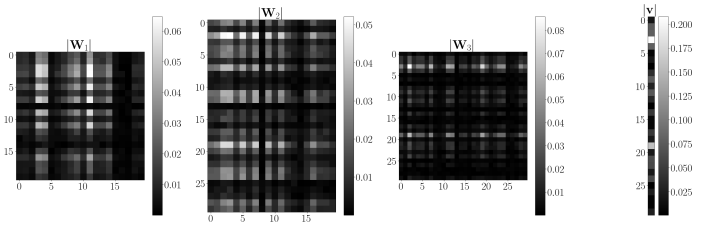


(c) Weights at iteration i_2

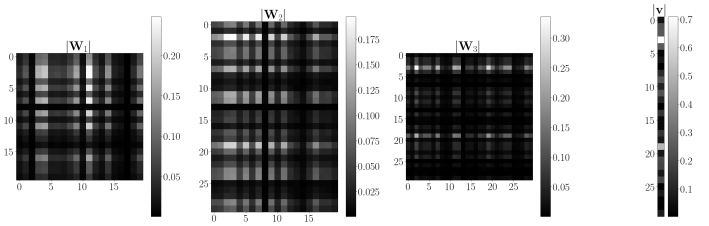
Figure 10: We train a three-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))$, where $\sigma(x) = \tanh(x)$, and $\mathbf{v} \in \mathbb{R}^{30}, \mathbf{W}_2 \in \mathbb{R}^{30\times 20}, \mathbf{W}_1 \in \mathbb{R}^{20\times 20}$ are the trainable weights. The sparsity structure is preserved upon escaping from the origin.



(a) Evolution of training loss with iterations



(b) Weights at iteration i_1



(c) Weights at iteration i_2

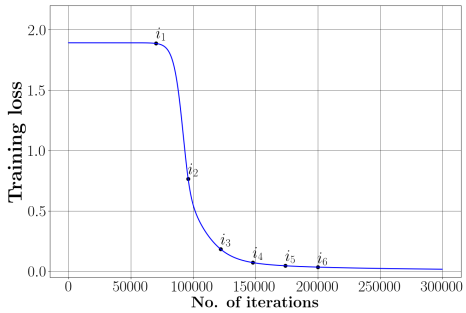
Figure 11: We train a four-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_3\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})))$, where $\sigma(x) = \tanh(x)$, and $\mathbf{v} \in \mathbb{R}^{30}, \mathbf{W}_3 \in \mathbb{R}^{30\times 30}, \mathbf{W}_2 \in \mathbb{R}^{30\times 20}, \mathbf{W}_1 \in \mathbb{R}^{20\times 20}$ are the trainable weights. The rows and columns of the weight matrices that become small near the origin remain small until gradient descent reaches the next saddle point, demonstrating preservation of the sparsity structure.

Appendix G. Training Dynamics Beyond the First Saddle Point

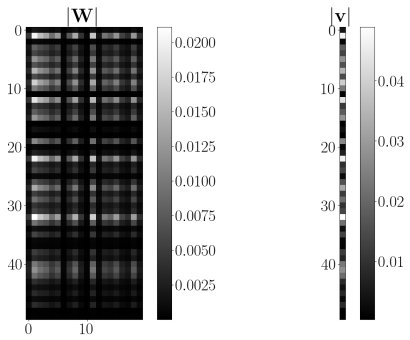
Our theoretical results and experiments so far have focused on the regime after escaping the origin and until reaching the first saddle point. As discussed in Section 1.1, it is widely believed that the training dynamics follows a saddle-to-saddle trajectory: after escaping from one saddle point, the weights reach another, and this process continues until convergence. This naturally raises the question of whether the phenomenon of sparsity structure preservation also holds after escaping the saddle points.

To investigate this, we re-consider some of our earlier experiments, running them for additional iterations when needed to ensure that the weights have escaped from the first saddle point. In Figure 14 and Figure 15, we revisit the experiment with two- and three-layer networks from Figure 1 and Figure 4, respectively, while Figure 16 and Figure 17 considers the experiment in Figure 8 with four-layer network and MNIST data. In all cases, we plot the weights after they escape from the first saddle point. We observe that the sparsity structure is preserved among the weights after it escapes from the first saddle point and until reaching the next saddle point.

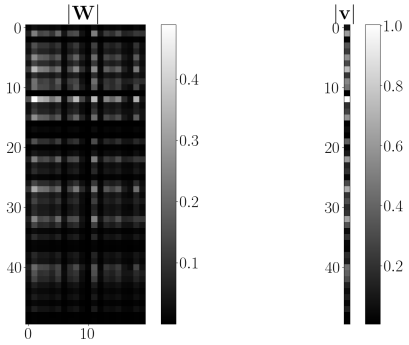
Overall, these experiments suggest that the sparsity structure is perhaps preserved even after escaping the saddle point and until reaching the next saddle point. Establishing a theoretical explanation for this phenomenon is an important direction for future research.



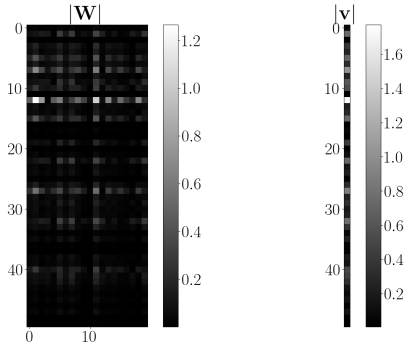
(a) Evolution of training loss



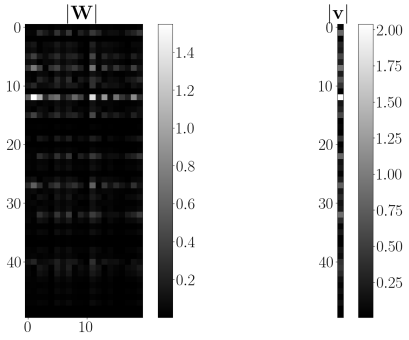
(b) Weights at iteration i_1



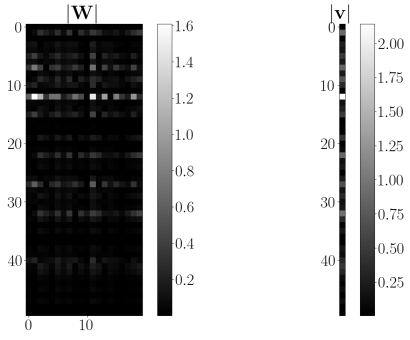
(c) Weights at iteration i_2



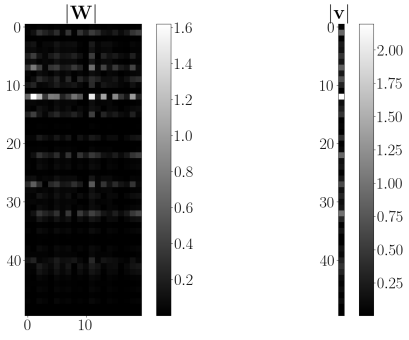
(d) Weights at iteration i_3



(e) Weights at iteration i_4

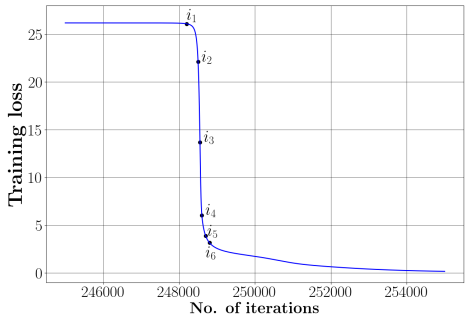


(f) Weights at iteration i_5

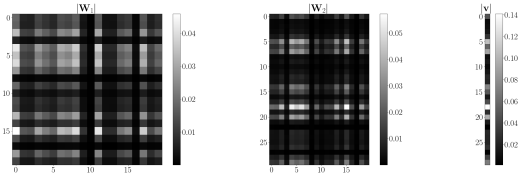


(g) Weights at iteration i_6

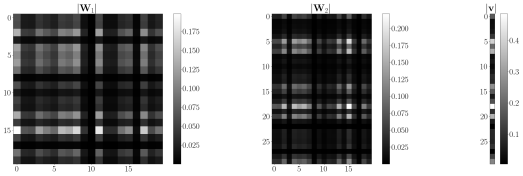
Figure 12: We train a two-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_1 \mathbf{x})$, where $\sigma(x) = \mathrm{GELU}(x)$, and $\mathbf{v} \in \mathbb{R}^{50}, \mathbf{W}_1 \in \mathbb{R}^{50 \times 20}$ are the trainable weights.



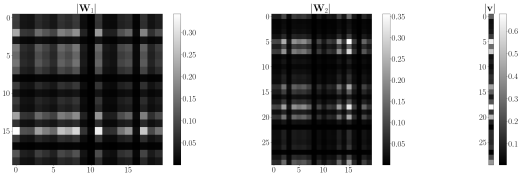
(a) Evolution of training loss



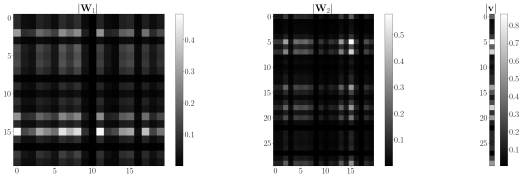
(b) Weights at iteration i_1



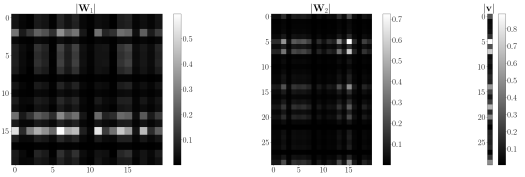
(c) Weights at iteration i_2



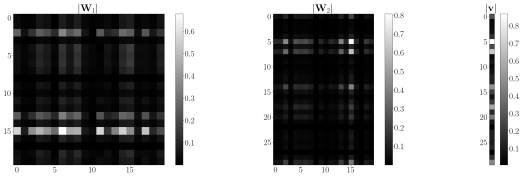
(d) Weights at iteration i_3



(e) Weights at iteration i_4

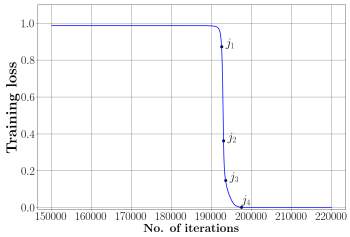


(f) Weights at iteration i_5

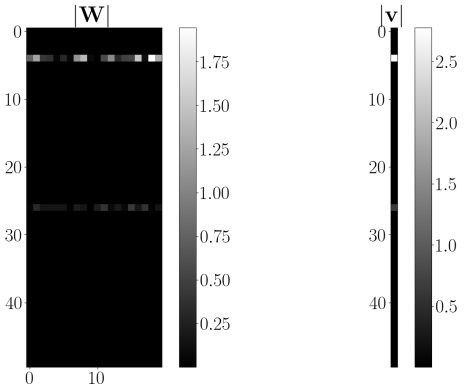


(g) Weights at iteration i_6

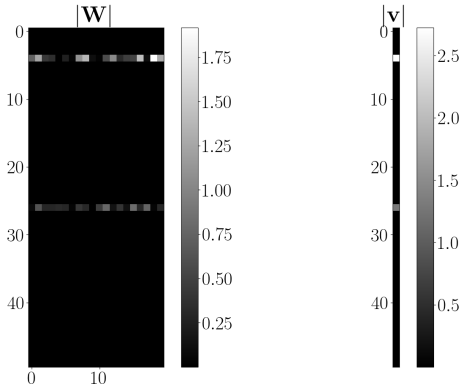
Figure 13: We train a three-layer neural network whose output is $\mathbf{v}^\top \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))$, where $\sigma(x) = \mathrm{GELU}(x)$, and $\mathbf{v} \in \mathbb{R}^{30}, \mathbf{W}_2 \in \mathbb{R}^{30\times 20}, \mathbf{W}_1 \in \mathbb{R}^{20\times 20}$ are the trainable weights.



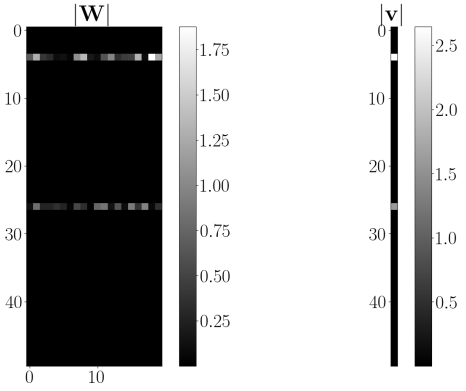
(a) Evolution of training loss



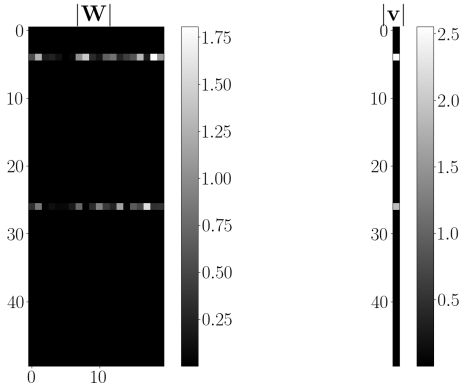
(b) Weights at iteration j_1



(c) Weights at iteration j_2

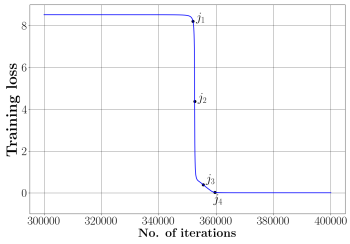


(d) Weights at iteration j_3

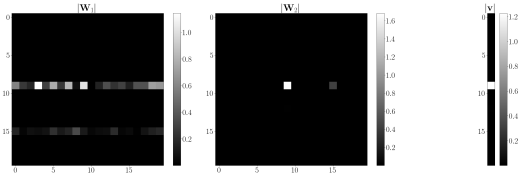


(e) Weights at iteration j_4

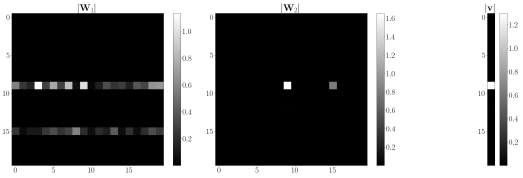
Figure 14: The experiment in Figure 1 is considered again. Panel (a) depicts the evolution of training loss near and after escaping from the first saddle point. Panels (b)-(e) depict the weights at different stages of training (marked in panel (a)).



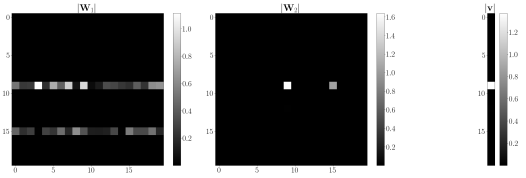
(a) Evolution of training loss



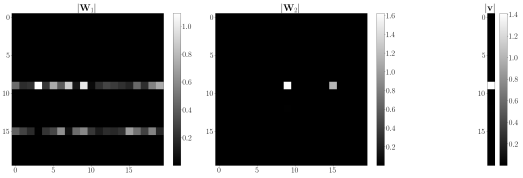
(b) Weights at iteration j_1



(c) Weights at iteration j_2

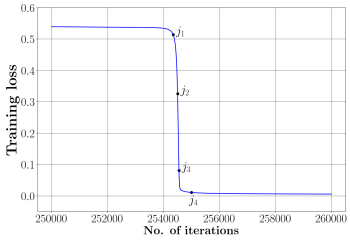


(d) Weights at iteration j_3

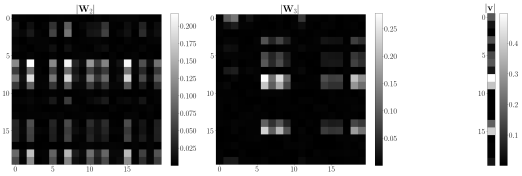


(e) Weights at iteration j_4

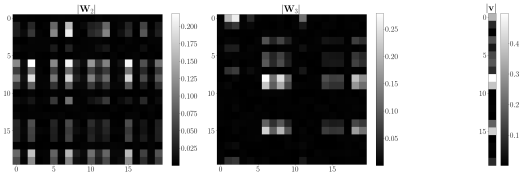
Figure 15: The experiment in Figure 4 is run for more iterations, specifically, until the weights escapes from the first saddle point. Panel (a) depicts the evolution of training loss near and after escaping from the first saddle point. Panels (b)-(e) depict the weights at different stages of training (marked in panel (a)).



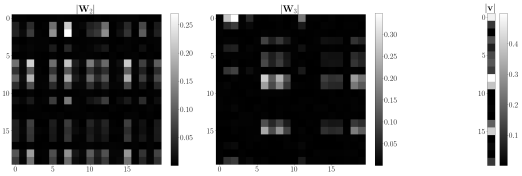
(a) Evolution of training loss



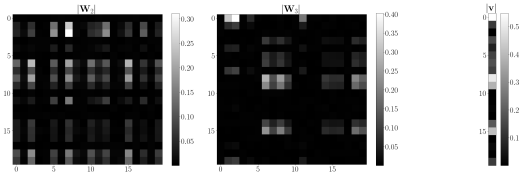
(b) Weights of last three layers at iteration j_1



(c) Weights of last three layers at iteration j_2

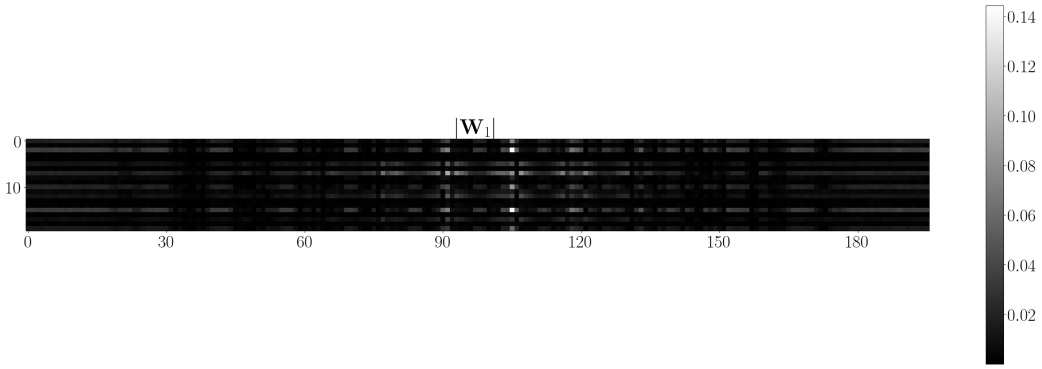


(d) Weights of last three layers at iteration j_3

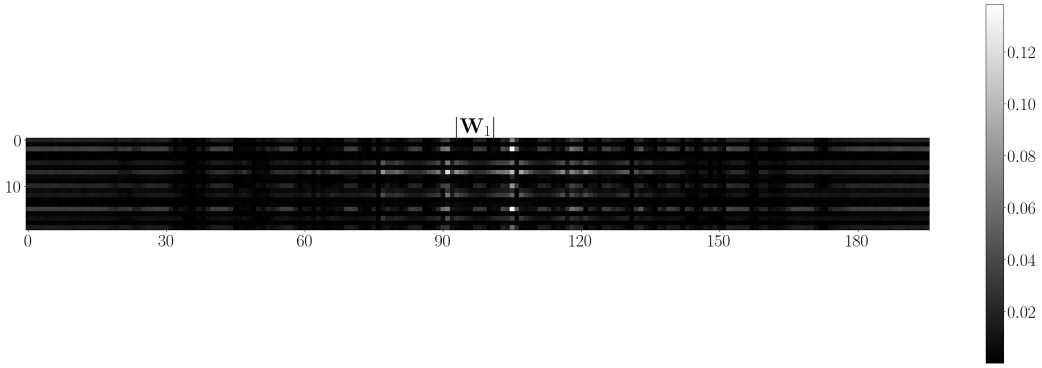


(e) Weights of last three layers at iteration j_4

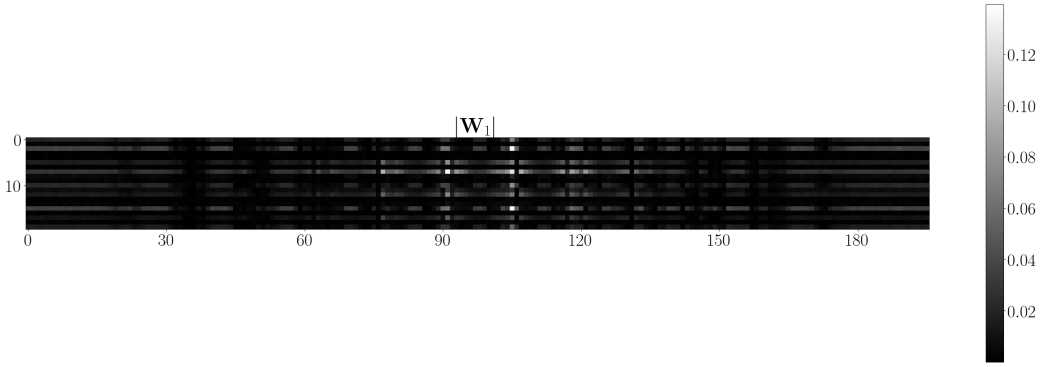
Figure 16: The experiment in Figure 8 is considered again. Panel (a) depicts the evolution of training loss near and after escaping from the first saddle point. Panels (b)-(e) depict the weights of last three layers at different stages of training (marked in panel (a)). The weights of first layer are shown in Figure 17 due to limited space.



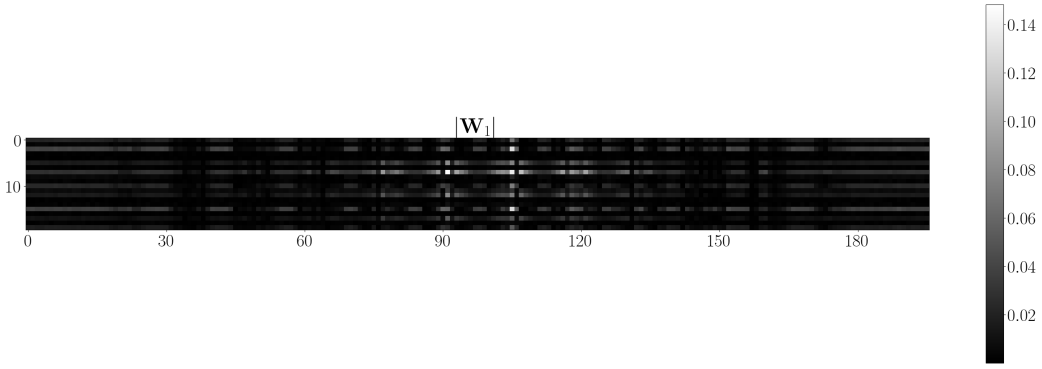
(a) Weights of first layer at iteration j_1



(b) Weights of first layer at iteration j_2



(c) Weights of first layer at iteration j_3



(d) Weights of first layer at iteration j_4

Figure 17: The weights of first layer from the experiment of Figure 16.

References

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: Leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, 2023.

Emmanuel Abbe, Samy Bengio, Enric Boix-Adsera, Etai Littwin, and Joshua Susskind. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 2024.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019b.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, 2024.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 2025.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, 2022.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis*, 68, 2024.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, 2022.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2023.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

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.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, 2020.

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 *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 2024.

Akshay Kumar and Jarvis Haupt. Early directional convergence in deep homogeneous neural networks for small initializations. *Transactions on Machine Learning Research*, 2025.

Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T Kiani. Implicit bias of linear equivariant networks. In *International Conference on Machine Learning*, 2022.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, 2016.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 2019.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer ReLU neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22 (71), 2021.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Advances in Neural Information Processing Systems*, 2021.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Conference on learning theory*, 2019.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*, 2022.

Yonatan Slutzky, Yotam Alexander, Noam Razin, and Nadav Cohen. The implicit bias of structured state space models can be poisoned with clean labels. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(1), 2018.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, 2021.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, 2023.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 2019.

Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of ReLU networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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 *Proceedings of Thirty Third Conference on Learning Theory*, pages 3635–3673, 2020.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Hanxu Zhou, Zhou Qixuan, Zhenyuan Jin, Tao Luo, Yaoyu Zhang, and Zhi-Qin Xu. Empirical phase diagram for three-layer neural networks with infinite width. In *Advances in Neural Information Processing Systems*, 2022.