Over-Parameterization Exponentially Slows Down Gradient Descent for Learning a Single Neuron

Weihang Xu WEIHANG_XU@OUTLOOK.COM
Tsinghua University

Simon S. Du SSDU@CS.WASHINGTON.EDU
University of Washington

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We revisit the canonical problem of learning a single neuron with ReLU activation under Gaussian input with square loss. We particularly focus on the over-parameterization setting where the student network has $n \geq 2$ neurons. We prove the global convergence of randomly initialized gradient descent with a $O\left(T^{-3}\right)$ rate. This is the first global convergence result for this problem beyond the exact-parameterization setting ($n = 1$) in which the gradient descent enjoys an $\exp(-\Omega(T))$ rate. Perhaps surprisingly, we further present an $\Omega\left(T^{-3}\right)$ lower bound for randomly initialized gradient flow in the over-parameterization setting. These two bounds jointly give an exact characterization of the convergence rate and imply, for the first time, that *over-parameterization can exponentially slow down the convergence rate*. To prove the global convergence, we need to tackle the interactions among student neurons in the gradient descent dynamics, which are not present in the exact-parameterization case. We use a three-phase structure to analyze GD's dynamics. Along the way, we prove gradient descent automatically balances student neurons, and use this property to deal with the non-smoothness of the objective function. To prove the convergence rate lower bound, we construct a novel potential function that characterizes the pairwise distances between the student neurons (which cannot be done in the exact-parameterization case). We show this potential function converges slowly, which implies the slow convergence rate of the loss function.

Keywords: over-parameterization, global convergence, non-convex optimization

1. Introduction

In recent years, theoretical explanations of the success of gradient descent (GD) on training deep neural networks emerge as an important problem. A prominent line of work Allen-Zhu et al. (2018); Du et al. (2018c); Jacot et al. (2018); Safran and Shamir (2018); Chizat et al. (2019) suggests that over-parameterization plays a key role in the successful training of neural networks.

However, the drawback of over-parameterization is under-explored. In this paper, we consider training two-layer ReLU networks, with a particular focus on learning a single neuron in the over-parameterization setting. We give a rigorous proof for the following surprising phenomenon:

Over-parameterization exponentially slows down the convergence of gradient descent.

Specifically, we consider two-layer ReLU networks with n neurons and input dimension d:

$$\boldsymbol{x} \to \sum_{i=1}^{n} [\boldsymbol{w}_i^\top \boldsymbol{x}]_+, \tag{1}$$

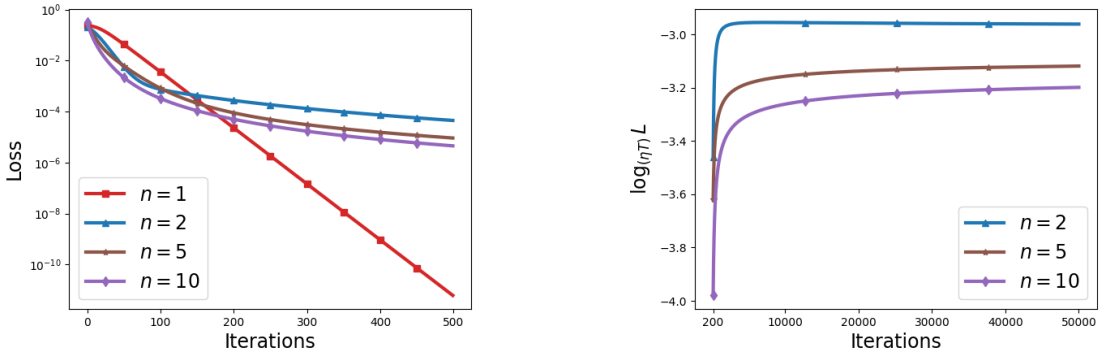

Figure 1: Setting: $\sigma = 0.1, \eta = 0.05, \|\boldsymbol{v}\| = 1$. Left: The loss converges much slower when $n > 1$, compared to the case of $n = 1$. Right: $\log_{(\eta T)} L(\boldsymbol{w}(T))$ converges to -3, with a small perturbation that converges extremely slow (note if we want $\log_{(\eta T)} \frac{C}{(\eta T)^3} \in (-3-\epsilon, -3+\epsilon)$, then $T \geq \frac{1}{\eta} C^{1/\epsilon}$ is needed.)

where $[x]_+ = \max\{0, x\}$ denotes the ReLU function, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n \in \mathbf{R}^d$ are n neurons. The input $\boldsymbol{x} \sim \mathcal{N}(0, I)$ follows a standard Gaussian distribution.

We consider the canonical teacher-student setting, where a student network is trained to learn a ground truth teacher network. Following the architecture (1), the student network $f : \mathbf{R}^d \to \mathbf{R}$ is given by $f(\boldsymbol{x}) = \sum_{i=1}^n [\boldsymbol{w}_i^\top \boldsymbol{x}]_+$, where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n \in \mathbf{R}^d$ are n student neurons. Similarly, the teacher network is given by $f^*(\boldsymbol{x}) = \sum_{i=1}^m [\boldsymbol{v}_i^\top \boldsymbol{x}]_+$, where $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbf{R}^d$ are m teacher neurons. It is natural to study the square loss:

$$L(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I)} \left[\frac{1}{2} \left(\sum_{i=1}^n [\boldsymbol{w}_i^\top \boldsymbol{x}]_+ - \sum_{i=1}^m [\boldsymbol{v}_i^\top \boldsymbol{x}]_+ \right)^2 \right], \tag{2}$$

where $\boldsymbol{w} = (\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \ldots, \boldsymbol{w}_n^\top)^\top \in \mathbf{R}^{n \times d}$ denotes the parameter vector formed by student neurons.

In this paper, we focus on the special case where the teacher network consists of one single neuron $\boldsymbol{v}_1$, i.e., $m = 1$. For simplicity, we omit the subscript and denote $\boldsymbol{v}_1$ with $\boldsymbol{v}$. Then the loss becomes

$$L(\boldsymbol{w}) = E_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I)} \left[\frac{1}{2} \left(\sum_{i=1}^n [\boldsymbol{w}_i^\top \boldsymbol{x}]_+ - [\boldsymbol{v}^\top \boldsymbol{x}]_+ \right)^2 \right]. \tag{3}$$

The student network is initialized with a Gaussian distribution: $\forall 1 \leq i \leq n, \boldsymbol{w}_i(0) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, ($\sigma \in \mathbf{R}^+$ denotes the initialization scale), then trained by gradient descent with step size η.

In this widely-studied setting, we discover a new phenomenon: compared to the exact-parameterized case ($n = 1$), the loss $L(\boldsymbol{w}(t))$ converges much slower in the over-parameterized case. Empirically (see Figure 1), the slow-down effect happens universally for all $n \geq 2$. Moreover, $\log_{(\eta T)} L(\boldsymbol{w}(T))$ has a tendency of converging towards -3, which seems to suggest that the convergence rate should be $L(\boldsymbol{w}(T)) = \Theta(T^{-3})$. In this paper, we prove rigorously that this is indeed true.

For the exact-parameterized case ($n = 1$), Yehudai and Ohad (2020) proved that $L(\boldsymbol{w}(t))$ convergences with a linear rate: $L(\boldsymbol{w}(t)) \leq \exp(-\Omega(t))$, which is also validated in Figure 1. For the over-parameterized case, an exact characterization of the convergence rate is given in this paper as

$L(\boldsymbol{w}(t)) = \Theta(t^{-3})$. As a result, we show that (even very mild) over-parameterization exponentially slows down the convergence rate. Specifically, our main results are the following two theorems.

Theorem 1 (Global Convergence, Informal) *For $\forall \delta > 0$, suppose the dimension $d = \Omega(\log(n/\delta))$, the initialization scale [1] $\sigma\sqrt{d} = poly(n^{-1})\|\boldsymbol{v}\|$, the learning rate $\eta = poly(\sigma\sqrt{d}, n^{-1}, \|\boldsymbol{v}\|^{-1})$. Then with probability at least $1-\delta$, gradient descent converges to a global minimum with rate $L(\boldsymbol{w}(t)) \leq poly(n, \|\boldsymbol{v}\|, \eta^{-1})t^{-3}$.*

Theorem 2 (Convergence Rate Lower Bound, Informal) *Suppose the student network is over-parameterized, i.e., $n \geq 2$. Consider gradient flow: $\frac{\partial \boldsymbol{w}(t)}{\partial t} = -\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}$. If the requirements on d and σ in Theorem 1 hold, then with high probability, there exist constants Γ_1, Γ_2 which do not depend on time t, such that $\forall t \geq 0, L(\boldsymbol{w}(t)) \geq (\Gamma_1 t + \Gamma_2)^{-3}$.*

Theorem 1 shows the global convergence of GD, while Theorem 2 provides a convergence rate lower bound. These two bounds together imply an exact characterization of the convergence rate for GD. We further highlight the significance of our contributions below:

• To our knowledge, Theorem 1 is the first global convergence result of gradient descent for the square loss beyond the special exact-parameterization cases of $m = n = 1$ (Tian, 2017; Brutzkus and Globerson, 2017; Yehudai and Ohad, 2020; Du et al., 2017) and $m = n = 2$ (Wu et al., 2018).

• While over-parameterization is well-known for its benefit in establishing global convergence in the finite-data regime, this is the first work proving it can slow down gradient-based methods.

1.1. Related Works

The problem of learning a single neuron is actually well-understood and can be solved with minimal assumptions by classical single index models algorithms (Kakade et al., 2011). For learning a single-neuron, Brutzkus and Globerson (2017); Tian (2017); Soltanolkotabi (2017) proved convergence for GD assuming Gaussian input distribution, which was later improved by Yehudai and Ohad (2020) who proved linear convergence of GD for learning one single neuron properly. These results are also generalized to learning a convolutional filter (Goel et al., 2018; Du et al., 2017, 2018a; Zhou et al., 2019; Liu et al., 2019). These works only focus on the exact-parameterization setting, while we focus on the over-parameterization setting.

Another direction focuses on the optimization landscape. Safran and Shamir (2018) showed spurious local minima exists for large m in the exact-parameterization setting. Safran et al. (2020) studied problem (2) with orthogonal teacher neurons. They showed that neither one-point strong convexity nor Polyak-Łojasiewicz (PL) condition hold locally near the global minimum. Wu et al. (2018) showed that problem (2) has no spurious local minima for $m = n = 2$. Zhong et al. (2017); Zhang et al. (2019) studied the exact-parameterization setting and showed the local strong convexity of loss and therefore with tensor initialization, GD can converge to a global minimum. Arjevani and Field (2022) proved that over-parameterization annihilates certain types of spurious local minima.

A popular line of works, known as neural tangent kernel (NTK) (Jacot et al., 2018; Chizat et al., 2019; Du et al., 2018c, 2019; Cao and Gu, 2019; Allen-Zhu et al., 2019; Arora et al., 2019; Oymak and Soltanolkotabi, 2020; Zou et al., 2020; Li and Liang, 2018) connects the training of ultra-wide neural networks with kernel methods. Another line of works uses the mean-field analysis

1. Note that $\|\boldsymbol{w}_i(0)\|$ scales with $\sigma\sqrt{d}$ rather than σ.

to study the training of infinite-width neural networks (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Wei et al., 2019; Nguyen and Pham, 2020; Fang et al., 2021; Lu et al., 2020). All of these works considered the finite-data regime and require the neural network to be ultra-wide, sometimes infinitely wide. Their techniques cannot explain the learnability of a single neuron, as pointed out by Yehudai and Shamir (2019).

More related to our works are results on the dynamics of gradient descent in the teacher-student setting. Li and Yuan (2017) studied the exact-parameterized setting and proved convergence for SGD with initialization in a region near identity. Li et al. (2020) showed that GD can learn two-layer networks better than any kernel methods, but their final upper bound of loss is constantly large and no convergence is proven. Zhou et al. (2021) proved *local* convergence for mildly over-parameterized two-layer networks. While our global convergence analysis uses their idea of establishing a gradient lower bound, we also propose new techniques to get rid of their architectural modifications, and improved their gradient lower bound to yield a tight convergence rate upper bound (see Section 2 for details). Also, Zhou et al. (2021) only provided a local convergence theory, while we prove convergence globally. On the other hand, their results hold for general $m \geq 1$ whereas we only study $m = 1$.

The first phase of our analysis is similar to the initial alignment phenomenon in Boursier et al. (2022). Their analysis also relies on the finite-data regime and the orthogonality of inputs, hence does not apply to our setting.

Similar slow-down effects of over-parameterization on the convergence rate have been observed in other scenarios. Richert et al. (2022) considered error function activation and empirically observed an $O(T^{-2})$ convergence rate. Going beyond neural network training, Dwivedi et al. (2018); Wu and Zhou (2019) showed such a phenomenon for Expectation-Maximization (EM) algorithm on Gaussian mixture models. Zhang et al. (2022) exhibited similar empirical behaviors of GD on Burer–Monteiro factorization, but no rigorous proof was given.

Paper Organization. In Section 2 we describe the main technical challenges in our analysis, and our ideas for addressing them. In section 3 we define some notations and preliminary notions. In Section 4 we formalize the global convergence result (Theorem 1) and provide a proof sketch. In Section 5 we formalize the convergence rate lower bound (Theorem 2) and provide a proof sketch.

2. Technical Overview

Three-Phase Convergence Analysis. Our global convergence analysis is divided into three phases. We define θ_i as the angle between $\boldsymbol{w}_i$ and $\boldsymbol{v}$, and $H := \|\boldsymbol{v}\| - \sum_i \langle \boldsymbol{w}_i, \overline{\boldsymbol{v}} \rangle$. Intuitively, θ_i represents the radial difference between teacher and students, while H represents the tangential difference between teacher and students.

When the initialization $\sigma\sqrt{d}$ is small enough, in phase 1, for every $i \in [n]$, θ_i decreases to a small value while $\|\boldsymbol{w}_i\|$ remains small. In phase 2, $\forall i \in [n], \theta_i$ remains bounded by a small value while H decreases with an exponential rate. Both θ_i and H being small at the end of phase 2 implies that GD enters a local region near a global minimum. In phase 3, we establish the local convergence by proving two properties: a lower bound of gradient, and a regularity condition of student neurons.

Non-Benign Optimization Landscape. Compared to the exact-parameterization setting, the optimization landscape becomes significantly different and much harder to analyze when the network is over-parameterized. Zhou et al. (2021) provided an intuitive illustration for this in their Section 4. For the general problem (2), Safran et al. (2020) showed that nice geometric properties that hold

when $m = n$, including one-point strong convexity and PL condition, do not hold when $m < n$. In this paper, we go further and show that the difference in geometric landscape leads to totally different convergence rates.

Non-smoothness and Implicit Regularization. The loss function is not smooth when student neurons are close to $\mathbf{0}$, which brings a major technical challenge for a local convergence analysis. Zhou et al. (2021) reparameterized the student neural network architecture to make the loss L smooth. We show this artificial change is not necessary. Our observation is that GD implicitly regularizes the student neurons and keeps them away from the non-smooth regions near $\mathbf{0}$. To prove this, we show that $\boldsymbol{w}_i$ cannot move too far in phase 3, by applying an algebraic trick to upper-bound $\sum_{t=T}^{\infty} \eta \|\nabla_{\boldsymbol{w}_i} L(\boldsymbol{w}(t))\|$ with $L(\boldsymbol{w}(T))$ (Lemma 24). A similar regularization property for GD was given in Du et al. (2018b), but it applies layer-wise rather than neuron-wise as in our paper.

Improving the Gradient Lower Bound. In our local convergence phase, we establish a local gradient lower bound similar to Theorem 3 in Zhou et al. (2021). Moreover, we improve their bound from $\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\| \geq \Omega(L(\boldsymbol{w}))$ to $\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\| \geq \Omega(L^{2/3}(\boldsymbol{w}))$ (Theorem 7). The idea in Zhou et al. (2021) is to pick an arbitrary global minimum $\{\boldsymbol{w}_i^*\}_{i=1}^n$ and show $\sum_i \langle \nabla_{\boldsymbol{w}_i} L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^* \rangle \geq L(\boldsymbol{w})$. We improve their proof technique by carefully choosing a specific $\{\boldsymbol{w}_i^*\}_{i=1}^n$ such that $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\|$ is small, then applying Cauchy inequality to get a tighter bound. This improvement is crucial since it improves the final bound of convergence rate from $L = O(T^{-1})$ in Zhou et al. (2021) to $L = O(T^{-3})$, which matches the lower bound in Theorem 2. This also indicates the optimality of the improved dependency $L^{2/3}$.

Non-degeneracy Condition. While the lower bound for the convergence rate is straightforward to prove in the worst-case (i.e., from a bad initialization), the average-case (i.e., with random initialization) lower bound is highly-nontrivial due to the existence of several counter-examples in the benign cases (see Appendix A). To distinguish these counter-examples from general cases, we establish a new non-degeneracy condition and build our lower bound upon it. We define a potential function $Z(t) = \sum_{i<j} \|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\|$, where $\boldsymbol{z}_i := \boldsymbol{w}_i - \langle \boldsymbol{w}_i, \overline{\boldsymbol{v}} \rangle \overline{\boldsymbol{v}}$. As long as the initialization is non-degenerate (See Definition 13), then $Z(t) = \Omega(t^{-1})$ and $L(\boldsymbol{w}(t)) \geq \Omega(Z^3(t) n^{-5} / \|\boldsymbol{v}\|)$, which imply $L(\boldsymbol{w}(t)) \geq \Omega(t^{-3})$. Intuitively, the slow convergence rate of L when $n \geq 2$ is due to the slow convergence of term $\boldsymbol{z}_i - \boldsymbol{z}_j, (i \neq j)$, and we define $Z(t)$ to formalize this idea.

3. Preliminaries

Notations. In this paper, bold-faced letters denote vectors. We use $[n]$ to denote $\{1, 2, \ldots, n\}$. For any nonzero vector $\boldsymbol{v} \in \mathbf{R}^d$, the corresponding normalized vector is denoted with $\overline{\boldsymbol{v}} := \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$. For two nonzero vectors $\boldsymbol{w}, \boldsymbol{v} \in \mathbf{R}^d$, $\theta(\boldsymbol{w}, \boldsymbol{v}) \coloneqq \arccos\left(\langle \overline{\boldsymbol{w}}, \overline{\boldsymbol{v}} \rangle\right)$ denotes the angle between them.

For simplicity, we also adopt some notational conventions. Denote the gradient of the i^{th} student neuron with $\nabla_i \coloneqq \frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}_i}$. For any variable $\boldsymbol{w}$ that changes during the training process, $\boldsymbol{w}(t)$ denotes its value at the t^{th} iteration, *e.g.*, $\boldsymbol{w}_i(t)$ indicates the value of $\boldsymbol{w}_i$ at the t^{th} iteration. Sometimes we omit the iteration index t when this causes no ambiguity. We abbreviate the expectation taken w.r.t the standard Gaussian as $\mathbb{E}_{\boldsymbol{x}}[\cdot] \coloneqq \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I)}[\cdot]$.

Special Notations for Important Terms. There are several important terms in our analysis and we give each of them a special notation. $\theta_i \coloneqq \theta(\boldsymbol{w}_i, \boldsymbol{v})$ denotes the angle between $\boldsymbol{w}_i$ and $\boldsymbol{v}$.

$\theta_{ij} \coloneqq \theta(\boldsymbol{w}_i, \boldsymbol{w}_j)$ denotes the angle between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$. Define

$$\boldsymbol{r} \coloneqq \sum_{i=1}^{n} \boldsymbol{w}_i - \boldsymbol{v}, \quad \text{and} \quad R : \mathbf{R}^d \to \mathbf{R}, R(\boldsymbol{x}) \coloneqq \sum_{j=1}^{n} [\boldsymbol{w}_j^\top \boldsymbol{x}]_+ - [\boldsymbol{v}^\top \boldsymbol{x}]_+.$$

Then $L(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}}[\frac{1}{2}R^2(\boldsymbol{x})]$. Define the length of the projection of $\boldsymbol{w}_i$ onto $\boldsymbol{v}$ as

$$h_i = \langle \boldsymbol{w}_i, \overline{\boldsymbol{v}} \rangle.$$

Lastly, define

$$H \coloneqq \|\boldsymbol{v}\| - \sum_{i \in [n]} h_i = \langle \overline{\boldsymbol{v}}, -\boldsymbol{r} \rangle.$$

Closed Form Expressions of Loss and Gradient. When the input distribution is standard Gaussian, closed form expressions of $L(\boldsymbol{w})$ and $\nabla L(\boldsymbol{w})$ can be obtained (Safran and Shamir, 2018). The complete form is deferred to Appendix B. Here we only present the closed form of gradient as it is used extensively in our analysis: Safran and Shamir (2018) showed that when $\boldsymbol{w}_i \neq \mathbf{0}, \forall i \in [n]$, the loss function is differentiable with gradient given by:

$$\nabla_i = \frac{1}{2}\left(\sum_j \boldsymbol{w}_j - \boldsymbol{v}\right) + \frac{1}{2\pi}\left[\left(\sum_{j \neq i} \|\boldsymbol{w}_j\| \sin\theta_{ij} - \|\boldsymbol{v}\| \sin\theta_i\right)\overline{\boldsymbol{w}}_i - \sum_{j \neq i} \theta_{ij}\boldsymbol{w}_j + \theta_i \boldsymbol{v}\right]. \quad (4)$$

Big-O notation. In this paper, we slightly abuse the use of big-O notation. We say for $\forall \epsilon = O(p)$, proposition A holds, if there *exists* an absolute constant $C \in \mathbf{R}^+$ such that for $\forall \epsilon \leq Cp$, proposition A holds. (See Theorem 4, 5, 6, 9, 14 for details.)

4. Proof Overview: Global Convergence

In this section we provide a proof sketch for Theorem 1. Full proofs for all theorems and lemmas can be found in the Appendix. We start with the initialization.

4.1. Initialization

We need the following conditions, which hold with high probability by random initialization.

Lemma 3 *Let $s_1 \coloneqq \frac{1}{2}\sigma\sqrt{d}, s_2 \coloneqq 2\sigma\sqrt{d}$. When $d = \Omega(\log(n/\delta))$, with probability at least $1-\delta$, the following properties hold at the initialization:*

$$\forall i \in [n], s_1 \leq \|\boldsymbol{w}_i(0)\| \leq s_2, \quad \textit{and} \quad \frac{\pi}{3} \leq \theta_i(0) \leq \frac{2\pi}{3}. \quad (5)$$

Condition (5) gives upper bound s_2 and lower bound s_1 for the norms of $\boldsymbol{w}_i(0)$, and states θ_i will fall in the interval $[\frac{\pi}{3}, \frac{2\pi}{3}]$ initially. These are standard facts in high-dimensional probability. See Appendix F.1 for proof details. The rest of our analysis will proceed *deterministically*.

4.2. Phase 1

We present the main theorem of Phase 1, which starts at time 0 and ends at time T_1.

Theorem 4 (Phase 1) *Suppose the initial condition in Lemma 3 holds. For any $\epsilon_1 = O(1), (\epsilon_1 > 0)$, there exists $C = O\left(\frac{\epsilon_1^2}{n}\right)$ such that for any $\sigma = O\left(C\epsilon_1^{48} d^{-1/2}\|\boldsymbol{v}\|\right)$ and $\eta = O\left(\frac{nC\sigma\sqrt{d}}{\|\boldsymbol{v}\|}\right)$, by setting $T_1 := \frac{C}{\eta}$, the following holds for $\forall 1 \le i \le n, 0 \le t \le T_1$:*

$$s_1 \le \|\boldsymbol{w}_i(t)\| \le s_2 + 2\eta\|\boldsymbol{v}\|t, \tag{6}$$

$$\textit{and} \quad \sin^2\left(\frac{\theta_i(t)}{2}\right) - \epsilon_1^2 \le \left(1 + \frac{\eta t}{s_2/\|\boldsymbol{v}\|}\right)^{-1/24}\left(\sin^2\left(\frac{\theta_i(0)}{2}\right) - \epsilon_1^2\right). \tag{7}$$

Consequently, at the end of Phase 1, we have

$$\forall i \in [n], \theta_i(T_1) \le 4\epsilon_1, \tag{8}$$

$$\textit{and} \quad h_i(T_1) \le 2h_j(T_1), \forall i, j \in [n]. \tag{9}$$

(6) gives upper and lower bounds for $\|\boldsymbol{w}_i\|$. (7) is used to bound the dynamics of θ_i. (8) shows that θ_i is small at the end of Phase 1, so the student neurons are approximately aligned with the teacher neuron. (9) states that the student neurons' projections on the teacher neuron are balanced.

Now we briefly describe our proof ideas.

Proof of (6). Proving the upper bound of $\|\boldsymbol{w}_i\|$ is straightforward, since the triangle inequality implies an upper bound of gradient norm $\|\nabla_i\| = O(\|\boldsymbol{v}\| + \sum_i \|\boldsymbol{w}_i\|)$, and the increasing rate of $\|\boldsymbol{w}_i\|$ is bounded by $\eta\|\nabla_i\|$. Note that we use $\|\boldsymbol{w}_i\|$ to upper bound $\|\nabla_i\|$, and use $\|\nabla_i\|$ to upper bound $\|\boldsymbol{w}_i\|$, so the argument can proceed inductively.

Given with the upper bound, we know that $\|\boldsymbol{w}_i\| = O(\eta\|\boldsymbol{v}\|t) = O(\epsilon_1^2\|\boldsymbol{v}\|/n)$ is a small term. Then the gradient (4) can be rewritten as:

$$\nabla_i = -\frac{1}{2\pi}(\|\boldsymbol{v}\|\sin\theta_i\overline{\boldsymbol{w}}_i + (\pi - \theta_i)\boldsymbol{v}) + O(\epsilon_1^2\|\boldsymbol{v}\|^2). \tag{10}$$

With (10), we prove the lower bound $\|\boldsymbol{w}_i\| \ge s_1$ by showing that $\|\boldsymbol{w}_i\|$ monotonically increases.

Proof of (7). The condition (7) aims to show that θ_i would decrease. Our intuition is clear: Since in each GD iteration, the update of $\boldsymbol{w}_i$ (the inverse of gradient (10)) is approximately a linear combination of $\overline{\boldsymbol{w}}_i$ and $\boldsymbol{v}$, the angle between $\boldsymbol{w}_i$ and $\boldsymbol{v}$ is going to decrease.

However, there is a technical difficulty when converting the above intuition into a rigorous proof, which is caused by the small perturbation term $O(\epsilon_1^2\|\boldsymbol{v}\|^2)$ in (10). When θ_i is large, showing θ_i would decrease is easy since this term is negligible. But when θ_i is too small, the effects of this perturbation term on the dynamics of θ_i is no longer negligible. As a result, we cannot directly show that θ_i decreases *monotonically*. Instead, we prove a weaker condition on the dynamics of θ_i and perform an algebraic trick (See (33) (34) in Appendix C):

$$\begin{aligned}\chi_i(t) - \chi_i(t+1) &\ge \frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t+1)\|}\left(\chi_i(t) - \epsilon_1^2\right)\\ \Rightarrow \chi_i(t+1) - \epsilon_1^2 &\le \left(1 - \frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t+1)\|}\right)\left(\chi_i(t) - \epsilon_1^2\right).\end{aligned} \tag{11}$$

Here $\chi_i(t) := \sin^2\left(\frac{\theta_i(t)}{2}\right)$. Note that (11) holds regardless of the sign of $\chi_i(t) - \epsilon_1^2$, hence both cases of θ_i being large and θ_i being small are gracefully handled. Therefore we can apply (11) iteratively to get $\chi_i(t+1) - \epsilon_1^2 \le \Pi_{t'=1}^{t+1}\left(1 - \frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t')\|}\right)\left(\chi_i(0) - \epsilon_1^2\right)$ (even if $\chi_i(t) - \epsilon_1^2$ might be negative for some t). This bound, combined with algebraic calculations, yields (7).

Proof of (8). Applying (7) with $t = T_1$ and some basic algebraic calculations yields (8).

Proof of (9). To prove (9), we divide Phase 1 into two intervals: $[0, T_1/50]$ and $[T_1/50, T_1]$. We first show that θ_i remains small in the second interval: $[T_1/50, T_1]$. Given with θ_i being small, nice properties of the gradient implies that h_i monotonically increases, and its increasing rate approximately equals $\frac{\eta}{2}H$ (see (36)), which is identical for all i. Therefore, the increases of h_i in the second interval: $h_i(T_1) - h_i(T_1/50)$ are balanced. Then we show that $h_i(T_1/50)$ is small compared to $h_i(T_1) - h_i(T_1/50)$. These two properties together shows that $h_i(T_1)$ are balanced.

4.3. Phase 2

Our second phase starts at time $T_1 + 1$ and ends at time T_2. The main theorem is as follows.

Theorem 5 (Phase 2) *Suppose the initial condition in Lemma 3 holds. For* $\forall \epsilon_2 = O(1)$, *set* $\epsilon_1 = O\left(\epsilon_2^6 n^{-1/2}\right)$ *in Theorem 4,* $\eta = O\left(\frac{\epsilon_1^2\sigma^2 d}{\|\boldsymbol{v}\|^2}\right)$ *and* $T_2 = T_1 + \left\lceil \frac{1}{n\eta}\ln\left(\frac{1}{36\epsilon_2}\right)\right\rceil$, *then* $\forall T_1 \le t \le T_2$,

$$h_i(t) \le 2h_j(t), \forall i, j, \tag{12}$$

$$\left(1 - \frac{n\eta}{2}\right)^{t-T_1}\|\boldsymbol{v}\| + 6\epsilon_2\|\boldsymbol{v}\| \ge H(t) \ge \frac{2}{3}\left(1 - \frac{n\eta}{2}\right)^{t-T_1}\|\boldsymbol{v}\| - 6\epsilon_2\|\boldsymbol{v}\| \ge 18\epsilon_2\|\boldsymbol{v}\|, \tag{13}$$

$$\frac{2\|\boldsymbol{v}\|}{n} \ge h_i(t) \ge \frac{s_1}{2}, \forall i. \tag{14}$$

$$\theta_i(t) \le \epsilon_2, \forall i. \tag{15}$$

(12) is the continuation of (6), which shows that the projections h_i remain balanced in Phase 2. (13) bounds the dynamics of $H(t)$. It shows that $H(t)$ exponentially decreases and gives upper and lower bounds. (14) gives upper and lower bounds for h_i. (15) shows that θ_i remains upper bounded by a small term ϵ_2 in Phase 2. Below we prove (12) (13) (14) (15) together inductively.

Proof of (12). Similar to (9), note that (15) guarantees that θ_i is small, so we still have that, for $\forall i$, h_i monotonically increases with rate approximately $\frac{\eta}{2}H$. Therefore, h_i will remain balanced.

Proof of (13). To understand why we need the bound (13), note that the gradient (4) has the following property:

$$\nabla_i = \frac{1}{2}\boldsymbol{r} + O((n\max_i\|\boldsymbol{w}_i\| + \|\boldsymbol{v}\|)\max_i\theta_i). \tag{16}$$

By (14) and (15), $\max_i\theta_i \le \epsilon_2$, and $\max_i\|\boldsymbol{w}_i\| = O(\|\boldsymbol{v}\|/n)$. So the second term in (16) can be bounded as $O((n\max_i\|\boldsymbol{w}_i\| + \|\boldsymbol{v}\|)\max_i\theta_i) \le O(\epsilon_2\|\boldsymbol{v}\|)$. When $O(\epsilon_2\|\boldsymbol{v}\|)$ is much smaller than the first term $\boldsymbol{r}/2$ in (16), we have $\nabla_i \approx \boldsymbol{r}/2$. Consequently, $\boldsymbol{r}$ and $H = \langle\overline{\boldsymbol{v}}, -\boldsymbol{r}\rangle$ will decrease with an exponential rate. But this will end when $\boldsymbol{r}$ becomes no larger than $O(\epsilon_2\|\boldsymbol{v}\|)$ and the approximation $\nabla_i \approx \boldsymbol{r}/2$ no longer holds, and that is the end of Phase 2.

So H should decrease (with exponential rate) to a small value, and it also should not be too small to ensure that $\|\boldsymbol{r}/2\| \gg O(\epsilon_2\|\boldsymbol{v}\|)$ (since $H = \langle\overline{\boldsymbol{v}}, -\boldsymbol{r}\rangle$). So we need to use (13) to simultaneously

upper and lower bound H. With the above intuition, proving (13) is straightforward as: $\nabla_i \approx \boldsymbol{r}/2 \Rightarrow H(t+1) \approx (1-n\eta/2)H(t) \approx \cdots \approx (1-n\eta/2)^{t-T_1}H(T_1)$.

It is worth noting that, we also need to handle a perturbation term when bounding the dynamics of $H(t)$, and we used the same trick as in proving (7).

Proof of (14). The left inequality can be derived from (12) and (13). The right inequality can be derived from the monotonicity of h_i, (15) and (6).

Proof of (15). This is the most difficult part in Theorem 5. Recall that in Phase 1 we used the gradient approximation (10) to bound θ_i, but (10) relies on $\|\boldsymbol{w}_i\|$ being a small term, which only holds in phase 1. So this time we use a totally different method to bound θ_i.

First we calculate the dynamics of $\cos\theta_i$ and get (see the proof in Appendix D for details): $\cos(\theta_i(t+1))-\cos(\theta_i(t)) = I_1+I_2$, where term $I_1 \geq -\frac{\eta}{2}\sum_{j\neq i}\sin\theta_i(t)\sin(\theta_i(t)+\theta_j(t))\frac{\|\boldsymbol{w}_j(t)\|}{\|\boldsymbol{w}_i(t+1)\|}$, and term I_2 is a small perturbation term. The next step is to establish the condition (12), then use it to bound the term $\frac{\|\boldsymbol{w}_j(t)\|}{\|\boldsymbol{w}_i(t+1)\|}$ in I_1. Consequently, we have

$$\cos(\theta_i(t+1)) - \cos(\theta_i(t)) = I_1 + I_2 \geq -2\eta\sum_{j\neq i}\sin\theta_i(t)(\sin\theta_i(t)+\sin\theta_j(t)) + I_2. \tag{17}$$

However, this is still not enough to prove the bound. The lower bound of the dynamics of $\cos\theta_i$ in (17) depends on θ_j where $j\neq i$. Since θ_j might be much larger then θ_i, the increasing rate of θ_i still cannot be upper-bounded.

To solve this problem, our key idea is to consider all θ_i's together. Define a potential function $V(t) := \sum_i \sin^2(\theta_i(t)/2)$, then we can sum the bound in (17) over all i's to get an upper bound for the increasing rate of V. Although the bound for θ_i depends on other θ_j's, the bound for V only depends on V itself. Consequently, the dynamics of the potential function V can be upper bounded, which yields the final upper bound (15).

4.4. Phase 3

Theorem 6 (Phase 3) *Suppose the initial condition in Lemma 3 holds. If we set $\epsilon_2 = O(n^{-14})$ in Theorem 5, $\eta = O\left(\frac{1}{n^2}\right)$, then $\forall T \in \mathbf{N}$ we have*

$$\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i(T+T_2)\| \geq \frac{\|\boldsymbol{v}\|}{4n} \quad \textit{and} \quad L(T+T_2) \leq O\left(\frac{n^4\|\boldsymbol{v}\|^2}{(\eta T)^3}\right). \tag{18}$$

This is the desired $1/T^3$ convergence rate. Our analysis consists of two steps:
1. Prove a gradient lower bound $\|\nabla L(\boldsymbol{w})\| \geq \text{poly}(n^{-1}, \|\boldsymbol{v}\|^{-1})L^{2/3}(\boldsymbol{w})$.
2. Prove that the loss function is smooth and Lipschitz on the gradient trajectory.
Given these two properties, the convergence can be established via the standard analysis for GD.

4.4.1. STEP 1: GRADIENT LOWER BOUND.

Theorem 7 (Gradient Lower Bound) *If for every student neuron we have $\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i\| \geq \frac{\|\boldsymbol{v}\|}{4n}$, and $L(\boldsymbol{w}) = O\left(\frac{\|\boldsymbol{v}\|^2}{n^{14}}\right)$, then $\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\| \geq \Omega\left(\frac{L^{2/3}(\boldsymbol{w})}{n^{2/3}\|\boldsymbol{v}\|^{1/3}}\right)$.*

As stated in Section 2, this theorem is an improved version of Theorem 3 in Zhou et al. (2021), improving the dependency of L from $\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\| \geq \Omega(L(\boldsymbol{w}))$ to $\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\| \geq \Omega(L^{2/3}(\boldsymbol{w}))$. Below we introduce our idea of improving the bound.

Lemma 8 (Gradient Projection Bound) *Suppose $\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, \ldots, \boldsymbol{w}_n^*$ is a global minimum of loss function L. Define $\theta_{\max} := \max_{i\in[n]} \theta_i$, then*

$$\sum_{i=1}^{n} \left\langle \frac{\partial}{\partial \boldsymbol{w}_i} L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^* \right\rangle \geq 2L(\boldsymbol{w}) - O\left(\theta_{\max}^2 \|\boldsymbol{r}\| \cdot \|\boldsymbol{v}\|\right). \tag{19}$$

Lemma 8 uses the idea of "descent direction" from Lemma C.1 in Zhou et al. (2021). The idea is to pick a global minimum $\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, \ldots, \boldsymbol{w}_n^*$ and lower bound the projection of gradient on the direction $\boldsymbol{w}_i - \boldsymbol{w}_i^*$. Recall that Zhou et al. (2021) made artificial modifications of the network architecture for technical reasons, e.g., they used the absolute value activation $x \to |x|$ instead of ReLU. Therefore, their proof cannot be directly applied to our lemma. However, we show that their idea still works in our setting, and modified their proof to prove Lemma 8 in Appendix E.2.

With Lemma 8 and several technical lemmas (Lemma 20, 21 in Appendix E.2), it is easy show that the last term $O\left(\theta_{\max}^2 \|\boldsymbol{r}\| \cdot \|\boldsymbol{v}\|\right)$ in (19) is small, so $\sum_{i=1}^{n} \left\langle \frac{\partial}{\partial \boldsymbol{w}_i} L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^* \right\rangle \geq L(\boldsymbol{w})$. Then we need to upper bound $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\|$, and that is the step where we make the improvement. In Zhou et al. (2021), they picked an *arbitrary* global minimum $\{\boldsymbol{w}_i^*\}_{i=1}^n$ and treated the term $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\|$ as constantly large. Consequently, their gradient lower bound scale with L^{-1}, yielding a final convergence rate of $L(\boldsymbol{w}(T)) \leq O(T^{-1})$. In contrast, our key observation is that we can pick a *specific* global minimum $\{\boldsymbol{w}_i^*\}_{i=1}^n$ that depends on $\{\boldsymbol{w}_i\}_{i=1}^n$. Specifically, we define

$$\forall i \in [n], \boldsymbol{w}_i^* := \frac{h_i}{\sum_j h_j} \boldsymbol{v}.$$

Then Lemma 22 shows that $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\| \leq O(L^{1/3}(\boldsymbol{w}))$ is a small term rather than a constant term. Finally, direct application of Cauchy inequality yields the improved bound Theorem 7.

4.4.2. STEP 2: SMOOTHNESS AND LIPSCHITZNESS

The aim of step 2 is to show the smoothness and Lipschitzness of L. However, one can see from (4) that L is neither Lipschitz nor smooth. The problem of non-Lipschitzness is easy to address, since (4) implies that $\|\nabla L\|$ is upper bounded by $\|\boldsymbol{w}_i\|$, and $\|\boldsymbol{w}_i\|$ is upper bounded by $L(\boldsymbol{w})$. However, the non-smoothness property of L is hard to handle. By the closed form expression of $\nabla^2 L$ (see (51)), one can see that $\|\nabla^2 L\|$ scales with $\frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i\|}$. Then $\|\nabla^2 L\| \to \infty$ as $\|\boldsymbol{w}_i\| \to 0$.

As stated in Section 2, our idea of solving this problem is to show that GD implicitly regularizes $\boldsymbol{w}_i$ such that $\|\boldsymbol{w}_i\|$ is always lower and upper bounded, namely (18) in Theorem 6. This property ensures the smoothness of L on GD trajectory (see Lemma 23 for details).

Implicit Regularization of Student Neurons. Next we describe our idea of proving the implicit regularization condition (18). It is not hard to give $\|\boldsymbol{w}_i(T_2)\|$ lower and upper bounds (see Lemma 19). Therefore, we only need to show that the student neurons do not move very far in phase 3. In other words, we wish to bound $\sum_{t=T_2}^{T} \eta \|\nabla L(\boldsymbol{w}(t))\|$ for $\forall T > T_2$. The intuition is very clear: in phase 3, the loss being small implies that the decrease of loss is small. Since the move of student neurons results in the decrease of loss, the change of $\|\boldsymbol{w}_i\|$ should also be small. However, the following subtlety emerges when constructing a rigorous proof.

The Importance of the Improved Gradient Lower Bound. We want to emphasize that our improved gradient lower bound (Theorem 7) is crucial for bounding the movement of student neurons $\sum_{t=T_2}^{T} \eta \|\nabla L(\boldsymbol{w}(t))\|$. There is an intuitive explanation for this: The weaker bound $\|\nabla L(\boldsymbol{w}(t))\| \sim$

$L(\boldsymbol{w}(t))$ implies the rate $L(\boldsymbol{w}(T)) \sim \frac{1}{T}$ (i.e., the rate in Zhou et al. (2021)). Then $\|\nabla L(\boldsymbol{w}(t))\| \sim L(\boldsymbol{w}(t)) \sim \frac{1}{T}$ and $\sum_{t=T_2}^{T} \eta \|\nabla L(\boldsymbol{w}(t))\| \sim \sum_{t=T_2}^{T} \frac{1}{t}$. But the infinite sum $\sum_{t=T_2}^{\infty} \frac{1}{t}$ diverges, so we cannot derive any meaningful bound.

On the other hand, the improved gradient lower bound $\|\nabla L(\boldsymbol{w}(t))\| \sim L^{2/3}(\boldsymbol{w}(t))$ implies the convergence rate $L(\boldsymbol{w}(T)) \sim \frac{1}{T^3} \Rightarrow \|\nabla L(\boldsymbol{w}(t))\| \sim L^{2/3}(\boldsymbol{w}(t)) \sim \frac{1}{T^2} \Rightarrow \sum_{t=T_2}^{T} \eta \|\nabla L(\boldsymbol{w}(t))\| \sim \sum_{t=T_2}^{T} \frac{1}{t^2}$, which is finite. See Lemma 24 for the rigorous argument.

4.5. Main Theorem

Now we are ready to state and prove the formal version of Theorem 1.

Theorem 9 (Global Convergence) *For $\forall \delta > 0$, if $d = \Omega(\log(n/\delta))$, $\sigma = O\left(n^{-4226} d^{-1/2} \|\boldsymbol{v}\|\right)$, $\eta = O\left(\frac{\sigma^2 d}{n^{169}\|\boldsymbol{v}\|^2}\right)$, then there exists $T_2 = O\left(\frac{\log n}{n\eta}\right)$ such that with probability at least $1-\delta$ over the initialization, for any $T \in \mathbf{N}$, $L(\boldsymbol{w}(T+T_2)) \le O\left(\frac{n^4\|\boldsymbol{v}\|^2}{(\eta T)^3}\right)$.*

To combine three phases of our analysis together, the last step is to assign values to the parameters in Theorem 4, 5, 6 $(\epsilon_2, \epsilon_1, C, \sigma, \eta, T_1, T_2)$ such that the previous phase satisfies the requirements of the next phase. For a complete list of the values, we refer the readers to Appendix F.2. With the parameter valuations in Appendix F.2, combining the initialization condition (Lemma 3) and three phases of our analysis (Theorem 4, 5, 6) together proves Theorem 9 immediately.

Remark 10 *Careful readers might notice that, if there exists i such that $\boldsymbol{w}_i = \mathbf{0}$, then L is not differentiable and gradient descent is not well-defined. However, such a corner case has been naturally excluded in our previous analysis. (See Appendix F.3 for a detailed discussion.)*

Remark 11 *Some readers might think that the polynomial dependencies of $\sigma = O\left(n^{-4226} d^{-1/2} \|\boldsymbol{v}\|\right)$, $\eta = O\left(\frac{\sigma^2 d}{n^{169}\|\boldsymbol{v}\|^2}\right)$ in Theorem 9 is too large. Here we would like to stress that these dependencies are not optimized, and we leave fine-grained optimization of them as a future direction.*

Taking $\eta \to 0$ in Theorem 9 (more rigorously, replacing our discrete analysis for GD with its continuous counterpart), we can obtain a corresponding global convergence theorem for gradient flow.

Corollary 12 (Global Convergence, Gradient Flow Version) *For $\forall \delta > 0$, if $d = \Omega(\log(n/\delta))$, $\sigma = O(n^{-4226} d^{-1/2} \|v\|)$, then there exists $T_2 = O(\log n/n)$ such that with probability at least $1-\delta$ over the initialization, for any $T > 0$, randomly initialized gradient flow has convergence rate upper bound $L(w(T+T_2)) \le O(n^4\|v\|^2/T^3)$.*

5. Proof Overview: Convergence Rate Lower Bound

In this section, we provide a general overview for the convergence rate lower bound. Full proofs of all theorems can be found in Appendix G. We consider the gradient flow (gradient descent with infinitesimal step size):

$$\frac{\partial \boldsymbol{w}(t)}{\partial t} = -\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}, \forall t \ge 0,$$

while keeping other settings (network architecture, initialization scheme, etc.) unchanged.

Our goal is to prove an $\Omega\left(1/T^3\right)$ bound. We note there exist fast-converging initialization points that break this lower bound. In Appendix A, we list several examples. Therefore, we need to utilize the property of random initialization to show our lower bound of $\Omega\left(1/T^3\right)$.

We first define a few important terms. For $\forall i \in [n]$, define $\boldsymbol{z}_i := \boldsymbol{w}_i - \langle \boldsymbol{w}_i, \overline{\boldsymbol{v}} \rangle \overline{\boldsymbol{v}}$ as the projection of $\boldsymbol{w}_i$ onto the orthogonal complement of $\boldsymbol{v}$. Define $Z(t) = \sum_{1 \le i < j \le n} \|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\|$. Define $Q^+(t) := \{i \in [n] | \boldsymbol{z}_i(t) \neq \boldsymbol{0}\}$ as the index set containing all i with $\boldsymbol{z}_i$ nonzero at time t. For $i, j \in Q^+$, define $\kappa_{ij} := \theta(\boldsymbol{z}_i, \boldsymbol{z}_j)$ as the angle between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. Define $\kappa_{\max}(t) := \max_{i,j \in Q^+(t)} \kappa_{ij}(t)$ as the maximum angle between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$.

Our idea is to show the lower bound holds as long as the initialization is "non-degenerate", formalized by the following definition.

Definition 13 (Non-degeneracy) *When $n \ge 2$, we say the initialization is non-degenerate if the following two conditions are satisfied. (1) All $\boldsymbol{z}_i$'s are nonzero: $\forall i \in [n]$, $\boldsymbol{z}_i(0) \neq 0$. (2) $\boldsymbol{z}_i$'s are not parallel: $\kappa_{\max}(0) > 0$.*

Since $\boldsymbol{z}_i$'s are initialized with a Gaussian distribution, the initialization is only degenerate on a set with Lebesgue measure zero, so the probability of the initialization being non-degenerate is 1. Now we are ready to state the formal version of Theorem 2 whose proof is in Appendix G.3.

Theorem 14 (Convergence Rate Lower Bound) *Suppose the network is over-parameterized, i.e., $n \ge 2$. Consider gradient flow: $\frac{\partial \boldsymbol{w}(t)}{\partial t} = -\frac{\partial L(\boldsymbol{w}(t))}{\partial \boldsymbol{w}}$. For $\forall \delta > 0$, if the initialization is non-degenerate, $d = \Omega(\log(n/\delta))$, $\sigma = O\left(n^{-4226} d^{-1/2} \|\boldsymbol{v}\|\right)$, then there exists $T_2 = O\left(\frac{\log n}{n}\right)$ such that with probability at least $1 - \delta$, for $\forall t \ge T_2$ we have*

$$L(\boldsymbol{w}(t))^{-1/3} \le O\left(\frac{n^{17/3}}{\kappa_{\max}^2(0) \|\boldsymbol{v}\|^{2/3}}\right)(t - T_2) + \gamma,$$

where $\gamma \in \mathbf{R}^+$ is a constant that does not depend on t.

Remark 15 *The bound in Theorem 14 depends on $1/\kappa_{\max}^{-2}(0)$. Such a dependence is reasonable since we have shown that there would be counter-examples if $\kappa_{\max}(0) = 0$ (See toy case 3 in Appendix A).*

5.1. Proof Sketch

Our key idea of proving Theorem 14 is to consider the potential function

$$Z(t) = \sum_{i<j} \|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\| \quad \text{where} \quad \boldsymbol{z}_i(t) := \boldsymbol{w}_i(t) - \langle \boldsymbol{w}_i(t), \overline{\boldsymbol{v}} \rangle \overline{\boldsymbol{v}}.$$

With $Z(t)$, our proof consists of three steps:
1. Show that with the non-degeneracy condition, $\kappa_{\max}(t)$ is lower bounded. (Lemma 30)
2. Show that when $\kappa_{\max}$ is lower bounded by a positive constant, $\frac{\partial}{\partial t} Z(t)$ can be lower bounded by $Z^2(t)$ (See (66) in Appendix G.3), so the convergence rate of $Z(t)$ is at most $Z(t) \sim t^{-1}$.
3. Use $Z(t)$ to lower bound $L(\boldsymbol{w}(t))$: $L(\boldsymbol{w}(t)) \ge \Omega\left(\frac{Z^3(t)}{n^5 \|\boldsymbol{v}\|}\right)$. ((68) in Appendix G.3).

Remark 16 *The potential function $Z(t)$ provides two implications.*
• *It explains why the convergence rate is different for $n = 1$ and $n \geq 2$. For $n \geq 2$, our analysis implies that the slow convergence rate of Z ($Z(t) \sim t^{-1}$) induces the slow convergence rate of $L(t) \sim t^{-3}$. When $n = 1$, Z is always zero, so L converges with linear rate. Intuitively, this is because optimizing the difference between student neurons is hard, which is a phenomenon that only exists in the over-parameterized case.*
• *It explains why the convergence rates in the two counter-examples (toy case 2 and 3 in Appendix A) are linear. In these two cases, the potential function Z degenerates to* 0.

We need several technical properties of the gradient flow trajectory. The first one is the implicit regularization condition: (18) in Theorem 6, and we use its gradient flow version (see Theorem 27 for details). We also need Corollary 28 and Lemma 29 to exclude the corner cases when $\boldsymbol{w}_i = \mathbf{0}$ and $\boldsymbol{z}_i(t) = \mathbf{0}$, where $\kappa_{\max}$ is not well-defined. The proofs are deferred to Appendix G.1.

Lower bounding $\kappa_{\max}$ is the most non-trivial step. We need to use the following lemma.

Lemma 17 (Automatic Separation of $\boldsymbol{z}_i$) *If there exists i, j such that $\kappa_{ij}(t) = \kappa_{\max}(t) < \frac{\pi}{2}$, then $\cos \kappa_{ij}(t)$ is well-defined in an open neighborhood of t, differentiable at t, and*

$$\frac{\partial}{\partial t} \cos \kappa_{ij}(t) \leq -\frac{\pi - \theta_{ij}(t)}{\pi}(1 - \cos \kappa_{ij}^2(t)). \tag{20}$$

Lemma 17 states that, when the vectors $\boldsymbol{z}_i$ are too close in direction, gradient flow will automatically separate them, which immediately implies a lower bound of $\kappa_{\max}$ (See Theorem 30). Its proof idea is also interesting: we can easily compute the dynamics of $\cos \kappa_{ij}$, which splits into two terms I_1 and I_2 (see (61) for them). I_1 is a simple term that can be handled easily, but the second term I_2 is very complicated and seems intractable. Our key observation is that, although I_2 is hard to bound for general i, j, it is always non-positive if we pick the pair of i, j such that $\kappa_{ij} = \kappa_{\max}$, and that property implies Lemma 17 via some routine computations.

Remark 18 *We note that in toy case 3 in Appendix A, all $\boldsymbol{z}_i$'s remain parallel and will not be separated. This is because the bound (20) in Lemma 17 implies that the initial condition $\kappa_{ij} = 0, \forall i, j$ is unstable. To see this, consider the ordinary differential equation $\dot{x} = -\tilde{C}(1 - x^2)$ where $\tilde{C} > 0$ is a constant. The initial condition $x(0) = 1$ induces the solution $x(t) \equiv 1$, which corresponds to toy case 3. But this initial condition is unstable since any perturbation of $x(0)$ results in solution $x(t) = \frac{1-\exp(2\tilde{C}t+c_0)}{1+\exp(2\tilde{C}t+c_0)}$, which implies an exponential increase of the perturbation, hence the separation of $\boldsymbol{z}_i$.*

Given with a lower bound of $\kappa_{\max}(t)$, and the implicit regularization property in Theorem 27, step 2 and step 3 can be proved with some geometric lemmas See Lemma 33, Lemma 31 and the proof of Theorem 14 in Appendix G. Combining three steps together finishes our proof.

Acknowledgments

SSD acknowledges the support of NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844, NSF IIS 2229881.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers, 2018. URL https://arxiv.org/abs/1811.04918.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Yossi Arjevani and Michael Field. Annihilation of spurious minima in two-layer relu networks. *arXiv preprint arXiv:2210.06088*, 2022.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *arXiv preprint arXiv:2206.00939*, 2022.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *International conference on machine learning*, pages 605–614. PMLR, 2017.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Sanjoy Dasgupta and Leonard Schulman. A two-round variant of em for gaussian mixtures, 2013. URL https://arxiv.org/abs/1301.3850.

Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1339–1348. PMLR, 2018a.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018b.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2018c. URL https://arxiv.org/abs/1810.02054.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Michael I. Jordan, Martin J. Wainwright, and Bin Yu. Singularity, misspecification, and the convergence rate of em, 2018. URL https://arxiv.org/abs/1810.00828.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021.

Surbhi Goel, Adam Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In *International Conference on Machine Learning*, pages 1783–1791. PMLR, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, 30, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer relu neural networks beyond ntk. *arXiv preprint arXiv:2007.04596*, 2020.

Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. Towards understanding the importance of shortcut connections in residual networks. *Advances in neural information processing systems*, 32, 2019.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pages 6426–6436. PMLR, 2020.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Frederieke Richert, Roman Worschech, and Bernd Rosenow. Soft mode in the dynamics of overrealizable online learning for soft committee machines. *Physical Review E*, 105(5):L052302, 2022.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4433–4441. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/safran18a.html.

Itay Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. *CoRR*, abs/2006.01005, 2020. URL https://arxiv.org/abs/2006.01005.

Mahdi Soltanolkotabi. Learning relus via gradient descent. *Advances in neural information processing systems*, 30, 2017.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *International conference on machine learning*, pages 3404–3413. PMLR, 2017.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

Chenwei Wu, Jiajun Luo, and Jason D Lee. No spurious local minima in a two hidden unit relu network. 2018.

Yihong Wu and Harrison H Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv preprint arXiv:1908.10935*, 2019.

Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3756–3786. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/yehudai20a.html.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Gavin Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification. *arXiv preprint arXiv:2206.03345*, 2022.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pages 1524–1534. PMLR, 2019.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pages 4140–4149. PMLR, 2017.

Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In *International Conference on Machine Learning*, pages 7594–7602. PMLR, 2019.

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Appendix A. Case study

It is a known result that under the case of exact-parameterization where $n = 1$, the loss converges exponentially fast [Yehudai and Ohad (2020)]. To understand why the convergence rate becomes much slower when $n \geq 2$, we investigate several toy cases.

Toy Case 1. Set $n = 2$, $\boldsymbol{w}_1(0) = \lambda_1(0)\boldsymbol{v} + \lambda_2(0)\boldsymbol{v}^\perp, \boldsymbol{w}_2(0) = \lambda_1(0)\boldsymbol{v} - \lambda_2(0)\boldsymbol{v}^\perp$, where $\lambda_1(0), \lambda_2(0) > 0$, $\boldsymbol{v}^\perp$ is a vector orthogonal with $\boldsymbol{v}$ such that $\|\boldsymbol{v}^\perp\| = \|\boldsymbol{v}\|$. Then $\boldsymbol{w}_1(0)$ and $\boldsymbol{w}_2(0)$ are reflection symmetric with respect to $\boldsymbol{v}$ (See Figure 2). Consider gradient descent with step size η initialized from $(\boldsymbol{w}_1(0), \boldsymbol{w}_2(0))$. It is easy to see that the symmetry of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is preserved in GD update, so for $t = 0, 1, 2, \ldots$ there exists $\lambda_1(t), \lambda_2(t)$ such that $\boldsymbol{w}_1(t) = \lambda_1(t)\boldsymbol{v} + \lambda_2(t)\boldsymbol{v}^\perp, \boldsymbol{w}_2(t) = \lambda_1(t)\boldsymbol{v} - \lambda_2(t)\boldsymbol{v}^\perp$. Since $\theta_1(t) = \theta_2(t), \forall t \in \mathbf{N}$, we denote $\theta := \theta_1 = \theta_2$. Then gradient (4) has the form

$$\begin{aligned}
\nabla_1 &= \left(\lambda_1 - \frac{1}{2}\right)\boldsymbol{v} + \frac{1}{2\pi}\left[\left(\|\boldsymbol{w}_2\|\sin(2\theta) - \|\boldsymbol{v}\|\sin\theta\right)\overline{\boldsymbol{w}}_1 - 2\theta\boldsymbol{w}_2 + \theta\boldsymbol{v}\right] \\
&= \left(\lambda_1 - \frac{1}{2} + \frac{1}{2\pi}\left(\left(\sin(2\theta) - \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_1\|}\sin\theta\right)\lambda_1 - \theta(2\lambda_1 - 1)\right)\right)\boldsymbol{v} \\
&\quad + \frac{1}{2\pi}\left(2\theta + \sin(2\theta) - \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_1\|}\sin\theta\right)\lambda_2\boldsymbol{v}^\perp \\
&= \left(\lambda_1 - \frac{1}{2}\right)\left(1 - \frac{\theta}{\pi} + \frac{\sin(2\theta)}{\lambda_1}\right)\boldsymbol{v} + \frac{1}{2\pi}\left(2\theta + \frac{\lambda_1 - 1/2}{\lambda_1}\sin(2\theta)\right)\lambda_2\boldsymbol{v}^\perp,
\end{aligned}$$

where the last equality is because $\sin(2\theta) - \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_1\|}\sin\theta = \sin(2\theta) - \frac{\|\boldsymbol{v}\|}{\lambda_1\|\boldsymbol{v}\|/\cos\theta}\sin\theta = \frac{\sin(2\theta)}{\lambda_1}\left(\lambda_1 - \frac{1}{2}\right)$. A similar expression can be computed for ∇_2.

Then we can write out the dynamics of λ_1 and λ_2 as

$$\lambda_1(t+1) - \frac{1}{2} = \left(\lambda_1(t) - \frac{1}{2}\right)\left(1 - \eta\left(1 - \frac{\theta(t)}{\pi} + \frac{\sin(2\theta(t))}{\lambda_1(t)}\right)\right), \tag{21}$$

$$\lambda_2(t+1) = \lambda_2(t)\left(1 - \frac{\eta}{2\pi}\left(2\theta + \frac{\lambda_1 - 1/2}{\lambda_1}\sin(2\theta)\right)\right). \tag{22}$$

Since $\theta = o(1)$, λ_1 is a constant term, $1 - \frac{\theta(t)}{\pi} + \frac{\sin(2\theta(t))}{\lambda_1(t)} \approx 1$, then (21) implies [2] $\lambda_1(t+1) - \frac{1}{2} \approx \left(\lambda_1(t) - \frac{1}{2}\right)(1 - \eta)$. This indicates that λ_1 converges to $\frac{1}{2}$ exponentially fast. So $\lambda_1 - 1/2 = o(1) \Rightarrow 2\theta + \frac{\lambda_1 - 1/2}{\lambda_1}\sin(2\theta) \approx 2\theta \approx 2\tan\theta = 2\frac{\lambda_2}{\lambda_1} \approx 4\lambda_2$. Then (22) can be rewritten as $\lambda_2(t+1) \approx \lambda_2(t)\left(1 - \frac{2\eta}{\pi}\lambda_2(t)\right)$. This indicates that λ_2 converges to 0 with rate $\lambda_2(t) \sim t^{-1}$.

Finally, we can compute the loss with (23) as $L(\boldsymbol{w}) = \Theta\left((2\lambda_1 - 1)^2 + (\sin\theta - \theta\cos\theta)\right)\|\boldsymbol{v}\|^2$. Since $(\sin\theta - \theta\cos\theta) \sim \theta^3 \sim \lambda_2^3 \sim t^{-3}$, we know that the convergence rate is $L(\boldsymbol{w}(t)) \sim t^{-3}$.

From the above toy case, we already know that the convergence rate given by Theorem 9 is *worst case optimal*. However, our ultimate goal is to prove an *average case* lower bound for the convergence rate: Theorem 14. We would like to point out that there is a huge gap between the worst case optimality and the average case optimality: proving the latter is much more difficult. To see this, we present two more toy cases.

2. Here we use the $\approx$ sign to omit higher order terms.

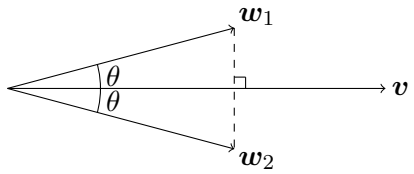

Figure 2: Toy Case 1

Toy Case 2. Let $n \geq 2$. We consider the case where all student neurons are parallel with the teacher neuron: $\boldsymbol{w}_1 = \lambda_1 \boldsymbol{v}, \ldots, \boldsymbol{w}_n = \lambda_n \boldsymbol{v}$, where $\lambda_1, \ldots, \lambda_n \in \mathbf{R}^+$. Then the gradient (4) becomes $\nabla_i = \frac{1}{2}(\sum_j \boldsymbol{w}_j - \boldsymbol{v})$. One can easily see that $\sum_{i\in[n]} \lambda_i$ converges exponentially fast to 0, which means that the convergence rate in this toy case is actually linear.

Toy Case 3. Let $n \geq 2$. We consider the case where all student neurons are equal: $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \ldots = \boldsymbol{w}_n$. Then the gradient (4) becomes $\nabla_i = \frac{1}{2}(n\boldsymbol{w}_i - \boldsymbol{v}) + \frac{1}{2\pi}\left[-\|\boldsymbol{v}\| \sin\theta_i \overline{\boldsymbol{w}}_i + \theta_i \boldsymbol{v}\right]$. One can see that the gradient in this case is just n times the gradient in the single student neuron case where the student neuron is $\boldsymbol{w}_i$ and the teacher neuron is $\boldsymbol{v}/n$. So the training process is actually equivalent with learning one teacher neuron $\|\boldsymbol{v}\|/n$ with one student neuron, with the step size η being multiplied by a factor of n. So in this toy case, the loss also have linear convergence.

Appendix B. Closed Form Expressions for L and ∇L

In this section, we present closed forms of L and ∇L, as computed in Safran and Shamir (2018).

Closed Form of $L(\boldsymbol{w})$.

$$L(\boldsymbol{w}) = \frac{1}{2}\sum_{i,j=1}^{n} \Upsilon(\boldsymbol{w}_i, \boldsymbol{w}_j) - \sum_{i=1}^{n} \Upsilon(\boldsymbol{w}_i, \boldsymbol{v}) + \frac{1}{2}\Upsilon(\boldsymbol{v}, \boldsymbol{v}),$$

where

$$\begin{aligned}\Upsilon(\boldsymbol{w}, \boldsymbol{v}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\left[\boldsymbol{w}^\top \boldsymbol{x}\right]_+ \left[\boldsymbol{v}^\top \boldsymbol{x}\right]_+\right] \\ &= \frac{1}{2\pi}\|\boldsymbol{w}\|\,\|\boldsymbol{v}\|(\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}) + (\pi - \theta_{\boldsymbol{w},\boldsymbol{v}})\cos(\theta_{\boldsymbol{w},\boldsymbol{v}})).\end{aligned}$$

Rearranging terms yields

$$L(\boldsymbol{w}) = \frac{1}{4}\|\sum_i \boldsymbol{w}_i - \boldsymbol{v}\|^2 + \frac{1}{2\pi}\left[\sum_{i<j}(\sin\theta_{ij} - \theta_{ij}\cos\theta_{ij})\|\boldsymbol{w}_i\|\,\|\boldsymbol{w}_j\| - \sum_i (\sin\theta_i - \theta_i\cos\theta_i)\|\boldsymbol{w}_i\|\,\|\boldsymbol{v}\|\right]. \tag{23}$$

Closed Form of $\nabla L(\boldsymbol{w})$. When $\boldsymbol{w}_i \neq \mathbf{0}, \forall i \in [n]$, Safran and Shamir (2018) showed that the loss function is differentiable and the gradient is given by

$$\begin{aligned}\nabla_i &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\left(\sum_{j=1}^{n}\left[\boldsymbol{w}_j^\top\boldsymbol{x}\right]_+ - \left[\boldsymbol{v}^\top\boldsymbol{x}\right]_+\right)\mathbb{1}\left\{\boldsymbol{w}_i^\top\boldsymbol{x}\geq 0\right\}\boldsymbol{x}\right] \\ &= \frac{1}{2}\left(\sum_j \boldsymbol{w}_j - \boldsymbol{v}\right) + \frac{1}{2\pi}\left[\left(\sum_{j\neq i}\|\boldsymbol{w}_j\|\sin\theta_{ij} - \|\boldsymbol{v}\|\sin\theta_i\right)\overline{\boldsymbol{w}}_i - \sum_{j\neq i}\theta_{ij}\boldsymbol{w}_j + \theta_i\boldsymbol{v}\right].\end{aligned}$$

Appendix C. Global Convergence: phase 1

Theorem 4 *Suppose the initial condition in Lemma 3 holds. For any $\epsilon_1 = O(1), (\epsilon_1 > 0)$, there exists $C = O\left(\frac{\epsilon_1^2}{n}\right)$ such that for any $\sigma = O\left(C\epsilon_1^{48}d^{-1/2}\|\boldsymbol{v}\|\right)$ and $\eta = O\left(\frac{nC\sigma\sqrt{d}}{\|\boldsymbol{v}\|}\right)$, by setting $T_1 := \frac{C}{\eta}$,[3] the following holds for $\forall 1 \leq i \leq n, 0 \leq t \leq T_1$:*

$$s_1 \leq \|\boldsymbol{w}_i(t)\| \leq s_2 + 2\eta\|\boldsymbol{v}\|t, \tag{24}$$

$$\sin^2\left(\frac{\theta_i(t)}{2}\right) - \epsilon_1^2 \leq \left(1 + \frac{\eta t}{s_2/\|\boldsymbol{v}\|}\right)^{-1/24}\left(\sin^2\left(\frac{\theta_i(0)}{2}\right) - \epsilon_1^2\right). \tag{25}$$

Consequently, at the end of Phase 1, we have

$$\forall i \in [n], \theta_i(T_1) \leq 4\epsilon_1, \tag{26}$$

and

$$h_i(T_1) \leq 2h_j(T_1), \forall i, j \in [n]. \tag{27}$$

Proof By Lemma 3, (24) and (25) holds for $t = 0$, and we have $s_1 \leq \|\boldsymbol{w}_i(0)\| \leq s_2, \forall i$.

Now we show with induction that (24) and (25) holds for $\forall t \leq T_1$.

For $t < T_1$, assume (24) and (25) holds for $0, 1, \ldots, t$, we prove the case of $t+1$.

First note that (25) holds for $0, 1, \ldots, t$ implies that $\forall t' \leq t, \sin^2(\theta_i(t')/2) \leq \max\{\sin^2(\theta_i(0)/2), \epsilon_1^2\} \leq \sin^2(\pi/3) \Rightarrow \theta_i(t') \leq 2\pi/3$.

Proof of the right inequality of (24).

Consider $\forall 0 \leq t' \leq t$, note that $\|\boldsymbol{w}_i(t')\| \geq s_1 > 0, \forall i$ implies that, for any i, j, the gradient $\nabla_i(t')$ and the angles $\theta_i(t'), \theta_{ij}(t')$ are well-defined.

Note that $s_2 = 2\sigma\sqrt{d} = O\left((\eta T_1)\epsilon_1^{48}\|\boldsymbol{v}\|\right) \leq \eta T_1\|\boldsymbol{v}\|$, so $\forall 0 \leq t' \leq t$ we have

$$\|\boldsymbol{w}_i(t')\| \leq s_2 + 2\eta\|\boldsymbol{v}\|t' \leq s_2 + 2\eta\|\boldsymbol{v}\|T_1 \leq 3C\|\boldsymbol{v}\| = O\left(\epsilon_1^2/n\right)\|\boldsymbol{v}\| \leq \frac{\|\boldsymbol{v}\|}{3n}. \tag{28}$$

By triangle inequality, for $\forall i \in [n], 0 \leq t' \leq t$,

$$\begin{aligned}\|\nabla_i(t')\| &\leq \frac{1}{2}\left(\sum_j\left\|\boldsymbol{w}_j(t')\right\| + \|\boldsymbol{v}\|\right) + \frac{1}{2\pi}\left[\sum_{j\neq i}\|\boldsymbol{w}_j(t')\| + \|\boldsymbol{v}\| + \sum_{j\neq i}\pi\|\boldsymbol{w}_j(t')\| + \pi\|\boldsymbol{v}\|\right] \\ &\leq \frac{1}{2}\left(\frac{1}{3}+1\right)\|\boldsymbol{v}\| + \frac{1}{2\pi}\left(\frac{1}{3}+1+\frac{\pi}{3}+\pi\right)\|\boldsymbol{v}\| \leq 2\|\boldsymbol{v}\|.\end{aligned} \tag{29}$$

3. Here we set η such that $T_1 = C/\eta \in \mathbf{N}$.

Then $\|\boldsymbol{w}_i(t+1)\|$ can be upper-bounded as

$$\|\boldsymbol{w}_i(t+1)\| = \|\boldsymbol{w}_i(0) - \sum_{t'=0}^{t} \eta \nabla_i(t')\| \leq \|\boldsymbol{w}_i(0)\| + \sum_{t'=0}^{t} \eta \|\nabla_i(t')\| \leq s_2 + 2\eta(t+1)\|\boldsymbol{v}\|.$$

Proof of the left inequality of (24).

Next we show that $\|\boldsymbol{w}_i(t+1)\| \geq \|\boldsymbol{w}_i(t)\| \geq s_1$.

Note that

$$\|\boldsymbol{w}_i(t+1)\|^2 - \|\boldsymbol{w}_i(t)\|^2 = \|\boldsymbol{w}_i(t) - \eta\nabla_i(t)\|^2 - \|\boldsymbol{w}_i(t)\|^2 = -2\eta\langle \boldsymbol{w}_i(t), \nabla_i(t)\rangle + \eta^2\|\nabla_i(t)\|^2,$$

so to show $\|\boldsymbol{w}_i(t+1)\| \geq \|\boldsymbol{w}_i(t)\|$, we only need to prove that $\langle \overline{\boldsymbol{w}}_i(t), \nabla_i(t)\rangle < 0$ (note that by the induction hypothesis we have $\|\boldsymbol{w}_i(t)\| > 0$, therefore $\overline{\boldsymbol{w}}_i(t)$ is well-defined):

$$\begin{aligned}
&\langle \overline{\boldsymbol{w}}_i(t), \nabla_i(t)\rangle \\
=& -\frac{(\pi - \theta_i(t))\,\langle \overline{\boldsymbol{w}}_i(t), \boldsymbol{v}\rangle + \langle \overline{\boldsymbol{w}}_i(t), \|\boldsymbol{v}\|\sin\theta_i(t)\overline{\boldsymbol{w}}_i(t)\rangle}{2\pi} \\
&+ \sum_j \frac{\langle \overline{\boldsymbol{w}}_i(t), (\pi - \theta_{ij}(t))\boldsymbol{w}_j(t)\rangle}{2\pi} + \frac{\left\langle \overline{\boldsymbol{w}}_i(t), \left(\sum_{j\neq i}\|\boldsymbol{w}_j(t)\|\sin\theta_{ij}(t)\right)\overline{\boldsymbol{w}}_i(t)\right\rangle}{2\pi} \\
=& -\frac{(\pi - \theta_i(t))\cos\theta_i(t) + \sin\theta_i(t)}{2\pi}\|\boldsymbol{v}\| + \sum_j \frac{(\pi - \theta_{ij}(t))\cos\theta_{ij}(t) + \sin\theta_{ij}(t)}{2\pi}\|\boldsymbol{w}_j(t)\| \\
\overset{(28)}{\leq}& -\frac{(\pi - \theta_i(t))\cos\theta_i(t) + \sin\theta_i(t)}{2\pi}\|\boldsymbol{v}\| + O\left(\epsilon_1^2\right)\|\boldsymbol{v}\| \\
\leq& -\frac{1/12}{2\pi}\|\boldsymbol{v}\| + O\left(\epsilon_1^2\right)\|\boldsymbol{v}\| \\
<& 0.
\end{aligned}$$

The reason for the second to last inequality is that, it is easy to verify by taking derivatives that the expression $(\pi - \theta)\cos\theta + \sin\theta$ monotonically decreases on the interval $[0, \pi]$, and the induction hypothesis implies $\theta_i(t) \leq 2\pi/3$, therefore $(\pi - \theta_i(t))\cos\theta_i(t) + \sin\theta_i(t) \geq (\pi - 2\pi/3)\cos(2\pi/3) + \sin(2\pi/3) > 1/12$.

Then we have $\|\boldsymbol{w}_i(t+1)\| \geq \|\boldsymbol{w}_i(t)\| \geq s_1$.

Proof of (25).

First we calculate the dynamics of $\cos\theta_i$.

$$
\begin{aligned}
&\cos(\theta_i(t+1)) - \cos(\theta_i(t)) \\
=&\langle \overline{\boldsymbol{w}}_i(t+1), \overline{\boldsymbol{v}} \rangle - \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \\
=&\frac{\|\boldsymbol{w}_i(t)\| \langle \boldsymbol{w}_i(t+1), \overline{\boldsymbol{v}} \rangle - \|\boldsymbol{w}_i(t+1)\| \langle \boldsymbol{w}_i(t), \overline{\boldsymbol{v}} \rangle}{\|\boldsymbol{w}_i(t+1)\| \cdot \|\boldsymbol{w}_i(t)\|} \\
=&\frac{\langle \boldsymbol{w}_i(t), \overline{\boldsymbol{v}} \rangle (\|\boldsymbol{w}_i(t)\| - \|\boldsymbol{w}_i(t+1)\|) - \eta \|\boldsymbol{w}_i(t)\| \langle \nabla_i(t), \overline{\boldsymbol{v}} \rangle}{\|\boldsymbol{w}_i(t+1)\| \cdot \|\boldsymbol{w}_i(t)\|} \\
=&\frac{\langle \boldsymbol{w}_i(t), \overline{\boldsymbol{v}} \rangle \frac{\|\boldsymbol{w}_i(t)\|^2 - \|\boldsymbol{w}_i(t) - \eta \nabla_i(t)\|^2}{\|\boldsymbol{w}_i(t)\| + \|\boldsymbol{w}_i(t+1)\|} - \eta \|\boldsymbol{w}_i(t)\| \langle \nabla_i(t), \overline{\boldsymbol{v}} \rangle}{\|\boldsymbol{w}_i(t+1)\| \cdot \|\boldsymbol{w}_i(t)\|} \\
=&\frac{1}{\|\boldsymbol{w}_i(t+1)\|} \left[\langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \frac{2\eta \langle \boldsymbol{w}_i(t), \nabla_i(t) \rangle - \eta^2 \|\nabla_i(t)\|^2}{\|\boldsymbol{w}_i(t)\| + \|\boldsymbol{w}_i(t+1)\|} - \eta \langle \nabla_i(t), \overline{\boldsymbol{v}} \rangle \right] \\
=&\frac{1}{\|\boldsymbol{w}_i(t+1)\|} \Bigg[\eta \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \langle \overline{\boldsymbol{w}}_i(t), \nabla_i(t) \rangle + \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \left(\frac{2\eta \langle \boldsymbol{w}_i(t), \nabla_i(t) \rangle}{\|\boldsymbol{w}_i(t)\| + \|\boldsymbol{w}_i(t+1)\|} - \frac{2\eta \langle \boldsymbol{w}_i(t), \nabla_i(t) \rangle}{2\|\boldsymbol{w}_i(t)\|} \right) \\
&- \eta^2 \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \frac{\|\nabla_i(t)\|^2}{\|\boldsymbol{w}_i(t)\| + \|\boldsymbol{w}_i(t+1)\|} - \eta \langle \nabla_i(t), \overline{\boldsymbol{v}} \rangle \Bigg] \\
=&\underbrace{\frac{\eta}{\|\boldsymbol{w}_i(t+1)\|} \langle \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \overline{\boldsymbol{w}}_i(t) - \overline{\boldsymbol{v}}, \nabla_i(t) \rangle}_{I_1} \\
&+ \underbrace{\frac{\eta \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle}{\|\boldsymbol{w}_i(t+1)\|} \left[\frac{\langle \overline{\boldsymbol{w}}_i(t), \nabla_i(t) \rangle (\|\boldsymbol{w}_i(t)\| - \|\boldsymbol{w}_i(t+1)\|)}{\|\boldsymbol{w}_i(t)\| + \|\boldsymbol{w}_i(t+1)\|} - \eta \frac{\|\nabla_i(t)\|^2}{\|\boldsymbol{w}_i(t)\| + \|\boldsymbol{w}_i(t+1)\|} \right]}_{I_2}.
\end{aligned}
\tag{30}
$$

For the first term I_1, note that the vector $\langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \overline{\boldsymbol{w}}_i(t) - \overline{\boldsymbol{v}}$ is orthogonal with $\boldsymbol{w}_i$, therefore,

$$
\begin{aligned}
I_1 =& \frac{\eta}{\|\boldsymbol{w}_i(t+1)\|} \left\langle \langle \overline{\boldsymbol{w}}_i(t), \overline{\boldsymbol{v}} \rangle \overline{\boldsymbol{w}}_i(t) - \overline{\boldsymbol{v}}, \frac{1}{2\pi} \left[\sum_{j \neq i} (\pi - \theta_{ij}(t)) \boldsymbol{w}_j(t) - (\pi - \theta_i(t)) \boldsymbol{v} \right] \right\rangle \\
=& \frac{\eta}{2\pi \|\boldsymbol{w}_i(t+1)\|} \left[(\pi - \theta_i(t)) \sin^2 \theta_i(t) \|\boldsymbol{v}\| - \sum_{j \neq i} (\pi - \theta_{ij}(t)) \left(\cos\theta_j(t) - \cos\theta_i(t) \cos\theta_{ij}(t) \right) \|\boldsymbol{w}_j(t)\| \right] \\
\geq& \frac{\eta}{2\pi \|\boldsymbol{w}_i(t+1)\|} \left[(\pi - \theta_i(t)) \sin^2 \theta_i(t) \|\boldsymbol{v}\| - n\pi \cdot 2 \|\boldsymbol{w}_j(t)\| \right] \\
\overset{(28)}{\geq}& \frac{\eta}{2\pi \|\boldsymbol{w}_i(t+1)\|} \left[(\pi - \theta_i(t)) \sin^2 \theta_i(t) \|\boldsymbol{v}\| - n\pi \cdot 2O(C) \|\boldsymbol{v}\| \right] \\
\geq& \frac{\eta \|\boldsymbol{v}\|}{2\pi \|\boldsymbol{w}_i(t+1)\|} \left[\frac{\pi}{3} \sin^2 \theta_i(t) - O\left(nC\right) \right],
\end{aligned}
\tag{31}
$$

where the last inequality is because $\theta_i(t) \leq \pi/3$.

The second term I_2 is a small perturbation term, which can be lower bounded as:

$$\begin{aligned} I_2 &\geq -\frac{\eta}{\|\boldsymbol{w}_i(t+1)\|}\left[\frac{\|\nabla_i(t)\|\cdot\|\eta\nabla_i(t)\|}{2s_1}+\eta\frac{\|\nabla_i(t)\|^2}{2s_1}\right] \\ &= -\frac{\eta^2}{s_1\|\boldsymbol{w}_i(t+1)\|}\|\nabla_i(t)\|^2 \overset{(29)}{\geq} -\frac{4\eta^2}{s_1\|\boldsymbol{w}_i(t+1)\|}\|\boldsymbol{v}\|^2. \end{aligned} \tag{32}$$

Combining both terms together, we get

$$\begin{aligned} \cos(\theta_i(t+1))-\cos(\theta_i(t)) = I_1+I_2 &\geq \frac{\eta\|\boldsymbol{v}\|}{2\pi\|\boldsymbol{w}_i(t+1)\|}\left[\frac{\pi}{3}\sin^2\theta_i(t)-O\left(nC\right)-8\pi\eta\frac{\|\boldsymbol{v}\|}{s_1}\right] \\ &\geq \frac{\eta\|\boldsymbol{v}\|}{6\|\boldsymbol{w}_i(t+1)\|}\left[\sin^2\theta_i(t)-O\left(nC\right)\right], \end{aligned}$$

where the last inequality is because $\eta = O\left(\frac{nC\sigma\sqrt{d}}{\|\boldsymbol{v}\|}\right) \Rightarrow 8\pi\eta\frac{\|\boldsymbol{v}\|}{s_1} = O\left(nC\right)$.

Therefore,

$$\begin{aligned} &\sin^2\left(\frac{\theta_i(t)}{2}\right)-\sin^2\left(\frac{\theta_i(t+1)}{2}\right) = \frac{\cos(\theta_i(t+1))-\cos(\theta_i(t))}{2} \\ &\geq \frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t+1)\|}\left[\sin^2\theta_i(t)-O\left(nC\right)\right] \geq \frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t+1)\|}\left[\sin^2\left(\frac{\theta_i(t)}{2}\right)-\epsilon_1^2\right], \end{aligned} \tag{33}$$

where the last inequality is because $\cos(\theta_i(t)/2) \geq \cos(\pi/3) = 1/2 \Rightarrow \sin\theta_i(t) = 2\sin(\theta_i(t)/2)\cos(\theta_i(t)/2) \geq \sin(\theta_i(t)/2)$, and $C = O(\epsilon_1^2/n) \Rightarrow O(nC) \leq \epsilon_1^2$.

Then we have

$$\begin{aligned} \sin^2\left(\frac{\theta_i(t+1)}{2}\right)-\epsilon_1^2 &\leq \sin^2\left(\frac{\theta_i(t)}{2}\right)-\frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t+1)\|}\left[\sin^2\left(\frac{\theta_i(t)}{2}\right)-\epsilon_1^2\right]-\epsilon_1^2 \\ &= \left(1-\frac{\eta\|\boldsymbol{v}\|}{12\|\boldsymbol{w}_i(t+1)\|}\right)\left(\sin^2\left(\frac{\theta_i(t)}{2}\right)-\epsilon_1^2\right) \\ &\leq \left(1-\frac{\eta}{12\left(s_2/\|\boldsymbol{v}\|+2\eta(t+1)\right)}\right)\left(\sin^2\left(\frac{\theta_i(t)}{2}\right)-\epsilon_1^2\right). \end{aligned}$$

For the same reason, for any $t' \in \{0,1,\ldots,t\}$ we have

$$\sin^2\left(\frac{\theta_i(t'+1)}{2}\right)-\epsilon_1^2 \leq \left(1-\frac{\eta}{12\left(s_2/\|\boldsymbol{v}\|+2\eta(t'+1)\right)}\right)\left(\sin^2\left(\frac{\theta_i(t')}{2}\right)-\epsilon_1^2\right). \tag{34}$$

$\sin^2\left(\frac{\theta_i(t')}{2}\right)-\epsilon_1^2$ can both be positive or negative, but (34) always holds regardless of its sign. Since $1-\frac{\eta}{12(s_2/\|\boldsymbol{v}\|+2\eta(t'+1))}$ is always positive, and multiplying both sides of an inequality by a positive number does not change the direction of the inequality, we can iteratively apply (34) and

get

$$\begin{aligned}
\sin^2\left(\frac{\theta_i(t+1)}{2}\right)-\epsilon_1^2 &\le \prod_{u=1}^{t+1}\left(1-\frac{\eta}{12\left(s_2/\|\boldsymbol{v}\|+2\eta u\right)}\right)\left(\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2\right)\\
&\le \prod_{u=1}^{t+1}\exp\left(-\frac{\eta}{12\left(s_2/\|\boldsymbol{v}\|+2\eta u\right)}\right)\left(\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2\right)\\
&\le \exp\left(\int_{u=1}^{t+2}-\frac{\eta}{12\left(s_2/\|\boldsymbol{v}\|+2\eta u\right)}du\right)\left(\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2\right)\\
&= \exp\left(-\frac{1}{24}\ln\left(\frac{s_2+(t+2)2\eta\|\boldsymbol{v}\|}{s_2+2\eta\|\boldsymbol{v}\|}\right)\right)\left(\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2\right)\\
&\le \left(1+\frac{\eta(t+1)}{s_2/\|\boldsymbol{v}\|}\right)^{-1/24}\left(\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2\right),
\end{aligned}$$

where the last inequality is because $2\eta\|\boldsymbol{v}\| \le s_1 \le s_2$. (Note that, by Lemma 3, $\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2$ is always positive.)

Proof of (26). W.L.O.G., suppose $T_1/50 \in \mathbb{N}$. By (25), for $\forall t \in [T_1/50, T_1]$ we have that

$$\sin^2\left(\frac{\theta_i(t)}{2}\right)-\epsilon_1^2 \le \left(1+\frac{\eta t}{s_2/\|\boldsymbol{v}\|}\right)^{-1/24}\left(\sin^2\left(\frac{\theta_i(0)}{2}\right)-\epsilon_1^2\right) \le \left(\frac{\eta T_1/50}{O\left((\eta T_1)\epsilon_1^{48}\right)}\right)^{-1/24} \le \epsilon_1^2. \tag{35}$$

Since $\epsilon_1 = O(1)$ is a sufficiently small constant, we have

$$\forall t \in [T_1/50, T_1], \sin\left(\frac{\theta_i(t)}{2}\right) \le \sqrt{2}\epsilon_1 \Rightarrow \forall t \in [T_1/50, T_1], \theta_i(t) \le 4\epsilon_1.$$

This implies (26) immediately.

Proof of (27). Consider $\forall t \in [T_1/50, T_1]$. The dynamics of h_i is given by

$$\begin{aligned}
&h_i(t+1)-h_i(t) = -\eta\langle\nabla_i(t), \overline{\boldsymbol{v}}\rangle\\
&= \underbrace{\frac{\eta}{2}\left(\|\boldsymbol{v}\|-\sum_j h_j(t)\right)}_{\frac{\eta}{2}H(t)}\\
&\quad \underbrace{-\frac{\eta}{2\pi}\left[\left(\sum_{j\neq i}\|\boldsymbol{w}_j(t)\|\sin\theta_{ij}(t)-\|\boldsymbol{v}\|\sin\theta_i(t)\right)\cos\theta_i(t)-\sum_{j\neq i}\theta_{ij}(t)h_j(t)+\theta_i(t)\|\boldsymbol{v}\|\right]}_{Q_i(t)}.
\end{aligned} \tag{36}$$

The first term is just $\frac{\eta}{2}H(t)$. Denote the second term with $Q_i(t)$. Then $h_i(t+1) = h_i(t)+\frac{\eta}{2}H(t)-Q_i(t)$.

$H(t)$ can be lower bounded as

$$H(t) = \|\boldsymbol{v}\|-\sum_j h_j(t) \ge \|\boldsymbol{v}\|-\sum_j\|\boldsymbol{w}_j(t)\| \overset{(28)}{\ge} \|\boldsymbol{v}\|-n\cdot\frac{\|\boldsymbol{v}\|}{3n} = \frac{2\|\boldsymbol{v}\|}{3}. \tag{37}$$

On the other hand, the second term $Q_i(t)$ is a small perturbation term, whose norm can be upper bounded by

$$\begin{aligned}|Q_i(t)| \leq & \frac{\eta}{2\pi}\left[n\cdot\frac{\|\boldsymbol{v}\|}{3n}\theta_{ij}(t)+\|\boldsymbol{v}\|\theta_i(t)+n\cdot\frac{\|\boldsymbol{v}\|}{3n}\theta_{ij}(t)+\theta_i(t)\|\boldsymbol{v}\|\right]\\ \leq & \frac{\eta}{2\pi}\left[\frac{\|\boldsymbol{v}\|}{3}8\epsilon_1+\|\boldsymbol{v}\|4\epsilon_1+\frac{\|\boldsymbol{v}\|}{3}8\epsilon_1+4\epsilon_1\|\boldsymbol{v}\|\right]\\ \leq & 3\eta\|\boldsymbol{v}\|\epsilon_1\leq 9\epsilon_1\frac{\eta}{2}H(t)\leq 0.1\frac{\eta}{2}H(t),\end{aligned} \tag{38}$$

where the second inequality is because $\theta_{ij}(t)\leq\theta_i(t)+\theta_j(t)\leq 8\epsilon_1$.

Therefore

$$0.3\eta\|\boldsymbol{v}\|\leq 0.9\frac{\eta}{2}H(t)\leq h_i(t+1)-h_i(t)=\frac{\eta}{2}H(t)-Q_i(t)\leq 1.1\frac{\eta}{2}H(t),\forall t\in[T_1/50,T_1].$$

Then we have the following bound, which shows that, $\forall i$, $h_i(T_1)-h_i(T_1/50)$ approximately equals to $\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)$:

$$0.9\left(\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)\right)\leq h_i(T_1)-h_i\left(T_1/50\right)=\sum_{t=T_1/50}^{T_1-1}\left(h_i(t+1)-h_i(t)\right)\leq 1.1\left(\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)\right). \tag{39}$$

The next bounds shows that, $\forall i$, $|h_i(T_1/50)|$ is small comparing to $\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)$:

$$|h_i(T_1/50)|\leq\|\boldsymbol{w}_i(T_1/50)\|\overset{(24)}{\leq}s_2+2\eta\|v\|\frac{T_1}{50}\leq\frac{1}{20}\eta\|\boldsymbol{v}\|T_1\overset{(37)}{\leq}0.2\left(\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)\right). \tag{40}$$

(39) and (40) jointly yields

$$0<0.7\left(\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)\right)\leq h_i(T_1)\leq 1.3\left(\sum_{t=T_1/50}^{T_1-1}\frac{\eta}{2}H(t)\right),\forall i,$$

which implies (27) immediately.

■

Appendix D. Global Convergence: phase 2

Theorem 5 *Suppose the initial condition in Lemma 3 holds. For $\forall\epsilon_2=O(1)$, set $\epsilon_1=O\left(\epsilon_2^6 n^{-1/2}\right)$ in Theorem 4, $\eta=O\left(\frac{\epsilon_1^2\sigma^2 d}{\|\boldsymbol{v}\|^2}\right)$ and $T_2=T_1+\left\lceil\frac{1}{n\eta}\ln\left(\frac{1}{36\epsilon_2}\right)\right\rceil$, then $\forall T_1\leq t\leq T_2$,*

$$h_i(t)\leq 2h_j(t),\forall i,j, \tag{41}$$

$$\left(1-\frac{n\eta}{2}\right)^{t-T_1}\|\boldsymbol{v}\|+6\epsilon_2\|\boldsymbol{v}\|\geq H(t)\geq\frac{2}{3}\left(1-\frac{n\eta}{2}\right)^{t-T_1}\|\boldsymbol{v}\|-6\epsilon_2\|\boldsymbol{v}\|\geq 18\epsilon_2\|\boldsymbol{v}\|, \tag{42}$$

$$\frac{2\|\boldsymbol{v}\|}{n}\geq h_i(t)\geq\frac{s_1}{2},\forall i. \tag{43}$$

$$\theta_i(t)\leq\epsilon_2,\forall i. \tag{44}$$

Proof

We prove (41), (42), (43) and (44) together inductively. First we show the induction base holds.

Note that Theorem 4 directly implies (41) and (44) for $t = T_1$. For (43), by (28) we have $h_i(T_1) \le \|\boldsymbol{w}_i(T_1)\| \le \frac{2\|\boldsymbol{v}\|}{n}$, and by (24) we have $h_i(T_1) = \|\boldsymbol{w}_i(T_1)\| \cos\theta_i(T_1) \ge \|\boldsymbol{w}_i(T_1)\|/2 \ge s_1/2$. For (42), note that $0 \le h_i(T_1) \le \|\boldsymbol{w}_i\| \overset{(28)}{\le} \|\boldsymbol{v}\|/(3n) \Rightarrow \|\boldsymbol{v}\| \ge H(T_1) \ge 2\|\boldsymbol{v}\|/3$.

Now suppose (41), (42) (43) and (44) holds for $T_1, T_1+1, \ldots, t$, next we show the case of $t+1$.

Proof of (41).

First note that due to $\theta_i(t) \le \epsilon_2$ and $\frac{2\|\boldsymbol{v}\|}{n} \ge h_i(t) \ge \frac{s_1}{2}$ we have

$$\frac{3\|\boldsymbol{v}\|}{n} \ge \frac{h_i(t)}{\cos\epsilon_2} \ge \frac{h_i(t)}{\cos\theta_i(t)} = \|\boldsymbol{w}_i(t)\| \ge h_i \ge \frac{s_1}{2}, \forall i. \tag{45}$$

As computed in (36), $h_i(t+1) = h_i(t) + \frac{\eta}{2}H(t) - Q_i(t), \forall i$.

Note that $\theta_{ij}(t) \le \theta_i(t) + \theta_j(t) \overset{(44)}{\le} 2\epsilon_2, \forall i, j$. Similar to (38), we have

$$|Q_i(t)| \le \frac{\eta}{2\pi} 14\epsilon_2\|\boldsymbol{v}\| \le 3\epsilon_2\eta\|\boldsymbol{v}\| \overset{(42)}{\le} \frac{\eta}{2}\cdot\frac{1}{3}H(t), \tag{46}$$

which implies

$$0 < \frac{\eta}{2}\cdot\frac{2}{3}H(t) \le \frac{\eta}{2}H(t) - Q_i(t) \le \frac{\eta}{2}\cdot\frac{4}{3}H(t), \forall i. \tag{47}$$

Then $\frac{\eta}{2}H(t) - Q_i(t) \le 2\left(\frac{\eta}{2}H(t) - Q_j(t)\right), \forall i, j$.

Finally, $\forall i, j$ we have $h_i(t+1) = h_i(t) + \frac{\eta}{2}H(t) - Q_i(t) \le 2h_j(t) + 2\left(\frac{\eta}{2}H(t) - Q_j(t)\right) = 2h_j(t+1)$.

Proof of (42).

The dynamics of $H(t)$ is given by $H(t+1) = H(t) - \sum_i (h_i(t+1) - h_i(t)) = H(t) - \sum_i \frac{\eta}{2}H(t) + \sum_i Q_i(t) = (1 - n\eta/2)H(t) + \sum_i Q_i(t)$.

Note that (46) implies $|\sum_i Q_i(t)| \le 3n\epsilon_2\|\boldsymbol{v}\|\eta$, therefore

$$H(t+1) - 6\epsilon_2\|\boldsymbol{v}\| \le \left(1 - \frac{n\eta}{2}\right)H(t) + 3n\epsilon_2\|\boldsymbol{v}\|\eta - 6\epsilon_2\|\boldsymbol{v}\| = \left(1 - \frac{n\eta}{2}\right)\left(H(t) - 6\epsilon_2\|\boldsymbol{v}\|\right).$$

Iterative application of the above bound yields $H(t+1) - 6\epsilon_2\|\boldsymbol{v}\| \le \left(1 - \frac{n\eta}{2}\right)^{t+1-T_1}\left(H(T_1) - 6\epsilon_2\|\boldsymbol{v}\|\right)$.

For the same reason, we also have $H(t+1) + 6\epsilon_2\|\boldsymbol{v}\| \ge \left(1 - \frac{n\eta}{2}\right)^{t+1-T_1}\left(H(T_1) + 6\epsilon_2\|\boldsymbol{v}\|\right)$.

On the other hand, $\forall i, \|\boldsymbol{v}\|/(3n) \overset{(28)}{\ge} \|\boldsymbol{w}_i(T_1)\| \ge h_i(T_1) = \|\boldsymbol{w}_i(T_1)\|\cos\theta_i(T_1) \ge 0$ implies $\|\boldsymbol{v}\| \ge H(T_1) \ge 2\|\boldsymbol{v}\|/3$. Combining three aforementioned bounds yields the first and second inequality in (42).

Now we prove the rightmost inequality in (42). Note that $n\eta = o(1) \Rightarrow 1 - n\eta/2 \geq \exp(-2n\eta/3)$. Then

$$\begin{aligned}
&\frac{2}{3}\left(1-\frac{n\eta}{2}\right)^{t-T_1}\|\boldsymbol{v}\| - 6\epsilon_2\|\boldsymbol{v}\| \\
\geq &\frac{2}{3}\exp(-2n\eta(T_2-T_1)/3)\|\boldsymbol{v}\| - 6\epsilon_2\|\boldsymbol{v}\| \\
\geq &\frac{2}{3}\exp\left(-\frac{2n\eta}{3}\cdot\frac{3}{2}\frac{1}{n\eta}\ln\left(\frac{1}{36\epsilon_2}\right)\right)\|\boldsymbol{v}\| - 6\epsilon_2\|\boldsymbol{v}\| \\
= &\frac{2}{3}\cdot 36\epsilon_2\|\boldsymbol{v}\| - 6\epsilon_2\|\boldsymbol{v}\| \\
= &18\epsilon_2\|\boldsymbol{v}\|,
\end{aligned}$$

where the second inequality is because $T_2 - T_1 = \left\lceil \frac{1}{n\eta}\ln\left(\frac{1}{36\epsilon_2}\right)\right\rceil \leq \frac{3}{2}\frac{1}{n\eta}\ln\left(\frac{1}{36\epsilon_2}\right)$.

Proof of (43).

Since we have already shown (41) and (42) for $t+1$, $H(t+1) \geq 18\epsilon_2\|\boldsymbol{v}\| > 0$ implies

$$\frac{n}{2}h_i(t+1) \overset{(41)}{\leq} \sum_j h_j(t+1) \leq \|\boldsymbol{v}\|, \forall i \;\Rightarrow\; h_i(t+1) \leq \frac{2}{n}\|\boldsymbol{v}\|, \forall i.$$

For the lower bound, by (36) and (47) we have

$$h_i(t+1) = h_i(t) + \frac{\eta}{2}H(t) - Q_i(t) \geq h_i(t) \geq \frac{s_1}{2}. \tag{48}$$

Proof of (44).

Recall that the dynamics of $\cos(\theta_i)$ is given by (30) as $\cos(\theta_i(t+1)) - \cos(\theta_i(t)) = I_1 + I_2$. Then we have

$$\begin{aligned}
I_1 =&\frac{\eta}{2\pi\|\boldsymbol{w}_i(t+1)\|}\left[(\pi-\theta_i(t))\sin^2\theta_i(t)\|\boldsymbol{v}\| - \sum_{j\neq i}(\pi-\theta_{ij}(t))\left(\cos\theta_j(t) - \cos\theta_i(t)\cos\theta_{ij}(t)\right)\|\boldsymbol{w}_j(t)\|\right] \\
\geq &-\frac{\eta}{2}\sum_{j\neq i}\left(\cos\theta_j(t) - \cos\theta_i(t)\cos\theta_{ij}(t)\right)\frac{\|\boldsymbol{w}_j(t)\|}{\|\boldsymbol{w}_i(t+1)\|} \\
\geq &-\frac{\eta}{2}\sum_{j\neq i}\sin\theta_i(t)\sin(\theta_i(t)+\theta_j(t))\frac{\|\boldsymbol{w}_j(t)\|}{\|\boldsymbol{w}_i(t+1)\|},
\end{aligned}$$

where the last inequality is because $\theta_{ij}(t) \leq \theta_i(t)+\theta_j(t) \leq 2\epsilon_2 < \pi \Rightarrow \cos\theta_j(t) - \cos\theta_i(t)\cos\theta_{ij}(t) \leq \cos\theta_j(t) - \cos\theta_i(t)\cos(\theta_i(t)+\theta_j(t)) = \sin\theta_i(t)\sin(\theta_i(t)+\theta_j(t))$.

Since we have already shown (43) for $t+1$, $\|\boldsymbol{w}_i(t+1)\| \geq h_i(t+1) \overset{(48)}{\geq} h_i(t)$ holds. Also we have $\|\boldsymbol{w}_j(t)\| = h_j(t)/\cos\theta_j(t) \leq 2h_j(t)$. Then

$$I_1 \geq -\frac{\eta}{2}\sum_{j\neq i}\sin\theta_i(t)(\sin\theta_i(t)+\sin\theta_j(t))\frac{2h_j(t)}{h_i(t)} \overset{(41)}{\geq} -2\eta\sum_{j\neq i}\sin\theta_i(t)(\sin\theta_i(t)+\sin\theta_j(t)).$$

To bound I_2, first note that $\|\boldsymbol{w}_i(t)\| \leq 3\|\boldsymbol{v}\|/n, \forall i$. Then by applying elementary triangle inequality in a similar manner as (29), we have $\|\nabla_i(t)\| \leq 5\|\boldsymbol{v}\|, \forall i$. Since $\|\boldsymbol{w}_i(t+1)\| \geq h_i(t+1) \geq s_1/2$, for similar reasons as (32), I_2 could be lower bounded as

$$I_2 \geq -\frac{\eta}{s_1/2} \cdot \frac{\|\nabla_i(t)\| \cdot \eta\|\nabla_i(t)\| + \eta\|\nabla_i(t)\|^2}{s_1} \geq -100\frac{\eta^2\|\boldsymbol{v}\|^2}{s_1^2}.$$

So we have

$$\cos(\theta_i(t+1)) - \cos\theta_i(t) \geq -2\eta\sum_{j\neq i}\sin\theta_i(t)(\sin\theta_i(t) + \sin\theta_j(t)) - 100\frac{\eta^2\|\boldsymbol{v}\|^2}{s_1^2}. \tag{49}$$

Define a potential function $V(t) \coloneqq \sum_i \sin^2(\theta_i(t)/2)$, we consider the dynamics of $V(t)$:

$$\begin{aligned} V(t+1) - V(t) =& \frac{1}{2}\sum_i (\cos\theta_i(t) - \cos\theta_i(t+1)) \\ \leq& \frac{1}{2}\sum_i \left(2\eta\sum_{j\neq i}\sin\theta_i(t)(\sin\theta_i(t) + \sin\theta_j(t)) + 100\frac{\eta^2\|\boldsymbol{v}\|^2}{s_1^2}\right) \\ \leq& \eta\sum_i \left(\frac{3}{2}n\sin^2\theta_i(t) + \sum_j \sin^2\theta_j(t)\right) + 50\frac{n\eta^2\|\boldsymbol{v}\|^2}{s_1^2} \\ \leq& 10n\eta\sum_i \sin^2\left(\frac{\theta_i(t)}{2}\right) + 50\frac{n\eta^2\|\boldsymbol{v}\|^2}{s_1^2}, \end{aligned} \tag{50}$$

where the last inequality is because $\sin^2\theta_i(t) = 4\sin^2(\theta_i(t)/2)\cos^2(\theta_i(t)/2) \leq 4\sin^2(\theta_i(t)/2)$.

Then we have

$$V(t+1) + \frac{5\eta\|\boldsymbol{v}\|^2}{s_1^2} \leq (1+10n\eta)\left(V(t) + \frac{5\eta\|\boldsymbol{v}\|^2}{s_1^2}\right) \leq \cdots \leq (1+10n\eta)^{t+1-T_1}\left(V(T_1) + \frac{5\eta\|\boldsymbol{v}\|^2}{s_1^2}\right).$$

Note that by (35) we have $V(T_1) \leq 2n\epsilon_1^2$, and by setting $\eta = O\left(\frac{\epsilon_1^2\sigma^2 d}{\|\boldsymbol{v}\|^2}\right) = O\left(\frac{\epsilon_1^2 s_1^2}{\|\boldsymbol{v}\|^2}\right)$ we have $\frac{5\eta\|\boldsymbol{v}\|^2}{s_1^2} \leq n\epsilon_1^2$. Then $V(t+1) \leq (1+10n\eta)^{t+1-T_1}\, 3n\epsilon_1^2 \leq \exp(10n\eta(T_2 - T_1))3n\epsilon_1^2 \leq \epsilon_2^2/16$. ■

Appendix E. Global Convergence: phase 3

E.1. Initial Condition of Phase 3

First we prove some initial conditions that are satisfied at time T_2, *i.e.*, the start of Phase 3.

Lemma 19 *Suppose the conditions* (41) (42) (43) (44) *in Theorem 5 holds, then at the start of Phase 3 we have*

$$\forall i \in [n], \|\boldsymbol{v}\|/(3n) \leq \|\boldsymbol{w}_i(T_2)\| \leq 3\|\boldsymbol{v}\|/n,$$

and

$$L(\boldsymbol{w}(T_2)) \leq 20\epsilon_2\|\boldsymbol{v}\|^2.$$

Proof

Proof of the First Condition.

By Theorem 5 we have $\theta_i(T_2) \leq \epsilon_2, \forall i$, and $H(T_2) \leq \left(1-\frac{n\eta}{2}\right)^{T_2-T_1}\|\boldsymbol{v}\| + 6\epsilon_2\|\boldsymbol{v}\| \leq \exp(-\frac{n\eta}{2}(T_2-T_1))\|\boldsymbol{v}\| + 6\epsilon_2\|\boldsymbol{v}\| = (36\epsilon_2)^{1/2}\|\boldsymbol{v}\| + 6\epsilon_2\|\boldsymbol{v}\| \leq 7\epsilon_2^{1/2}\|\boldsymbol{v}\|$.

Then for $\forall i \in [n]$, $\frac{2}{3}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\| - H(T_2) = \sum_j h_j(T_2) \leq 2nh_i(T_2) \Rightarrow \|\boldsymbol{w}_i(T_2)\| \geq h_i(T_2) \geq \|\boldsymbol{v}\|/(3n)$.

Similarly, for $\forall i \in [n]$, $H(T_2) \geq 0 \Rightarrow \|\boldsymbol{v}\| \geq \sum_j h_j(T_2) \geq nh_i(T_2)/2 \Rightarrow h_i(T_2) \leq 2\|\boldsymbol{v}\|/n \Rightarrow \|\boldsymbol{w}_i(T_2)\| = h_i(T_2)/\cos(\theta_i(T_2)) \leq \frac{3}{2}h_i(T_2) \leq 3\|\boldsymbol{v}\|/n$.

Proof of the Second Condition.

Since $\|\boldsymbol{v}\|/(3n) \leq h_i(T_2) \leq 2\|\boldsymbol{v}\|/n, \forall i$, we have

$$\begin{aligned}\left\|\sum_i \boldsymbol{w}_i(T_2) - \boldsymbol{v}\right\| &\leq \sum_i \|\boldsymbol{w}_i(T_2) - h_i(T_2)\overline{\boldsymbol{v}}\| + \left\|\sum_i h_i(T_2)\overline{\boldsymbol{v}} - \boldsymbol{v}\right\| \\ &= \sum_i h_i(T_2)\tan\theta_i(T_2) + H(T_2) \leq 8\epsilon_2^{1/2}\|\boldsymbol{v}\|,\end{aligned}$$

where the last inequality is because $\forall i, \theta_i(T_2) \leq \epsilon_2 = o(1) \Rightarrow \tan\theta_i(T_2) \leq 2\epsilon_2 \leq o(\epsilon_2^{1/2})$.

So according to (23) we have

$$L(\boldsymbol{w}(T_2)) \leq \frac{1}{4}\left(8\epsilon_2^{1/2}\|\boldsymbol{v}\|\right)^2 + \frac{1}{2\pi}\left(n^2\cdot 2\epsilon_2\cdot\left(\frac{2}{n}\|\boldsymbol{v}\|\right)^2 + n\epsilon_2\frac{2}{n}\|\boldsymbol{v}\|^2\right) \leq 20\epsilon_2\|\boldsymbol{v}\|^2.$$

■

E.2. Proofs for Gradient Lower Bound

Before proving Theorem 7, we need some auxiliary lemmas.

E.2.1. AUXILIARY LEMMAS

Lemma 8 *Recall the global minimum $\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, \ldots, \boldsymbol{w}_n^*$ defined as $\boldsymbol{w}_i^* = \frac{h_i}{\sum_{j\in[n]} h_j}\boldsymbol{v}$. Define $\theta_{\max} := \max_{i\in[n]}\theta_i$, then*

$$\sum_{i=1}^n \left\langle \frac{\partial}{\partial \boldsymbol{w}_i}L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^*\right\rangle \geq 2L(\boldsymbol{w}) - O\left(\theta_{\max}^2\|\boldsymbol{r}\|\cdot\|\boldsymbol{v}\|\right).$$

Proof First we introduce the idea of residual decomposition in Zhou et al. (2021), which decomposes the residual function $R(\boldsymbol{x})$ in two terms :

$$R(\boldsymbol{x}) = \sum_{j=1}^n \left[\boldsymbol{w}_j^\top\boldsymbol{x}\right]_+ - \left[\boldsymbol{v}^\top\boldsymbol{x}\right]_+ = \boldsymbol{r}^\top\boldsymbol{x}\cdot\mathbb{1}\{\boldsymbol{v}^\top\boldsymbol{x}\geq 0\} + \sum_{j=1}^n \boldsymbol{w}_j^\top\boldsymbol{x}\left(\mathbb{1}\{\boldsymbol{w}_j^\top\boldsymbol{x}\geq 0\} - \mathbb{1}\{\boldsymbol{v}^\top\boldsymbol{x}\geq 0\}\right).$$

Define $R_1(\boldsymbol{x}) = \boldsymbol{r}^\top\boldsymbol{x}\cdot\mathbb{1}\{\boldsymbol{v}^\top\boldsymbol{x}\geq 0\}$ and $R_2(\boldsymbol{x}) = \sum_{j=1}^n \boldsymbol{w}_j^\top\boldsymbol{x}\left(\mathbb{1}\{\boldsymbol{w}_j^\top\boldsymbol{x}\geq 0\} - \mathbb{1}\{\boldsymbol{v}^\top\boldsymbol{x}\geq 0\}\right)$, then $R(\boldsymbol{x}) = R_1(\boldsymbol{x}) + R_2(\boldsymbol{x})$.

Back to the lemma, first we have the following algebraic calculations:

$$\begin{aligned}
&\sum_{i=1}^{n}\left\langle \frac{\partial}{\partial \boldsymbol{w}_i} L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^* \right\rangle \\
&= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{x}}\left[R(\boldsymbol{x}) \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\} \boldsymbol{x}^\top (\boldsymbol{w}_i - \boldsymbol{w}_i^*)\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[R(\boldsymbol{x}) \sum_{i=1}^{n}\left(\left[\boldsymbol{w}_i^\top \boldsymbol{x}\right]_+ - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\} \boldsymbol{x}^\top \boldsymbol{w}_i^*\right)\right] \\
&= 2L(\boldsymbol{w}) + \mathbb{E}_{\boldsymbol{x}}\left[R(\boldsymbol{x}) \sum_{i=1}^{n}\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^*\right]
\end{aligned}$$

With the residual decomposition, the last term above can be decomposed into two terms I_1, I_2 as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[R(\boldsymbol{x}) \sum_{i=1}^{n}\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^*\right] \\
=& \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[R_1(\boldsymbol{x}) \sum_{i=1}^{n}\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^*\right]}_{I_1} \\
&+ \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[R_2(\boldsymbol{x}) \sum_{i=1}^{n}\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^*\right]}_{I_2}.
\end{aligned}$$

For the second term I_2, note that

$$\forall j, \boldsymbol{w}_j^\top \boldsymbol{x}\left(\mathbb{1}\{\boldsymbol{w}_j^\top \boldsymbol{x} \geq 0\} - \mathbb{1}\{\boldsymbol{v}^\top \boldsymbol{x} \geq 0\}\right) \geq 0 \Rightarrow R_2(\boldsymbol{x}) \geq 0$$

and $\forall \boldsymbol{w}_i^*, \boldsymbol{w}_i$,

$$\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^* \geq 0,$$

so I_2 is always non-negative.

For the first term $I_1 = \sum_{i\in[n]} \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{r}^\top \boldsymbol{x} \mathbb{1}(\boldsymbol{v}^\top \boldsymbol{x} \geq 0)\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^*\right]$, we bound each term in the summation as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{r}^\top \boldsymbol{x} \mathbb{1}(\boldsymbol{v}^\top \boldsymbol{x} \geq 0)\left(\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right) \boldsymbol{x}^\top \boldsymbol{w}_i^*\right] \\
&\geq -\mathbb{E}_{\boldsymbol{x}}\left[\left|\boldsymbol{r}^\top \boldsymbol{x}\right| \cdot \left|\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\right\}\right| \cdot \|\tilde{\boldsymbol{x}}\| \cdot \|\boldsymbol{w}_i^*\| \theta_i\right] \\
&= -\|\boldsymbol{r}\| \theta_i \|\boldsymbol{w}_i^*\| \cdot \mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\|\tilde{\boldsymbol{x}}\|^2 \left|\mathbb{1}\left\{\boldsymbol{w}_i^{*\top} \tilde{\boldsymbol{x}} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^\top \tilde{\boldsymbol{x}} \geq 0\right\}\right|\right] \\
&\geq -O\left(\|\boldsymbol{r}\| \theta_i^2 \|\boldsymbol{w}_i^*\|\right)
\end{aligned}$$

where $\tilde{\boldsymbol{x}}$ is the projection of $\boldsymbol{x}$ onto $span(\boldsymbol{w}_i^*, \boldsymbol{w}_i, \boldsymbol{r})$ and follows a three-dimensional Gaussian. Here the first inequality is because $\mathbb{1}\left\{\boldsymbol{w}_i^{*\top}\boldsymbol{x} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^{\top}\boldsymbol{x} \geq 0\right\} \neq 0 \Rightarrow \theta(\boldsymbol{w}_i^*, \boldsymbol{x}) \in [\pi/2 - \theta_i, \pi/2 + \theta_i] \Rightarrow |\boldsymbol{x}^\top \boldsymbol{w}_i^*| = |\tilde{\boldsymbol{x}}^\top \boldsymbol{w}_i^*| \leq \|\tilde{\boldsymbol{x}}\| \cdot \|\boldsymbol{w}_i^*\|\theta_i$, and the last inequality is because

$$\mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\|\tilde{\boldsymbol{x}}\|^2 \left|\mathbb{1}\left\{\boldsymbol{w}_i^{*\top}\tilde{\boldsymbol{x}} \geq 0\right\} - \mathbb{1}\left\{\boldsymbol{w}_i^{\top}\tilde{\boldsymbol{x}} \geq 0\right\}\right|\right] = O(\theta_i).$$

(See Lemma C.5 in Zhou et al. (2021) for detailed calculations.)

Note that $\sum_i \|\boldsymbol{w}_i^*\| = \|\boldsymbol{v}\|$, so we have

$$\sum_{i=1}^n \left\langle \frac{\partial}{\partial \boldsymbol{w}_i} L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^* \right\rangle = 2L(\boldsymbol{w}) + I_1 + I_2 \geq 2L(\boldsymbol{w}) - \sum_{i\in[n]} O\left(\|\boldsymbol{r}\|\theta_i^2\|\boldsymbol{w}_i^*\|\right) \geq 2L(\boldsymbol{w}) - O\left(\theta_{\max}^2\|\boldsymbol{r}\| \cdot \|\boldsymbol{v}\|\right).$$

■

Lemma 20 (Bound of θ_i)

$$\|\boldsymbol{w}_i\|^2\theta_i^3 \leq 30\pi L(\boldsymbol{w}), \forall i.$$

Proof W.L.O.G., suppose $\boldsymbol{v} = (\|\boldsymbol{v}\|, 0, \ldots, 0)^\top$ and $\boldsymbol{w}_i = (\cos\theta_i, \sin\theta_i, 0, \ldots, 0)^\top\|\boldsymbol{w}_i\|$.

Define $S_i \coloneqq \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{x}^\top\boldsymbol{w}_i \geq 0 \wedge \boldsymbol{x}^\top\boldsymbol{v} < 0\}$. One can see that $S_i = \{\boldsymbol{x} : \theta(\boldsymbol{x}, \boldsymbol{w}_i) \leq \pi/2, \theta(\boldsymbol{x}, \boldsymbol{v}) \geq \pi/2\}$. On the other hand, $\forall \boldsymbol{x} \in S_i$, $R(\boldsymbol{x}) = \sum_{j=1}^n \left[\boldsymbol{w}_j^\top\boldsymbol{x}\right]_+ - \left[\boldsymbol{v}^\top\boldsymbol{x}\right]_+ \geq \left[\boldsymbol{w}_i^\top\boldsymbol{x}\right]_+ \geq 0$. Therefore,

$$\begin{aligned}
L(\boldsymbol{w}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},I)}\left[\frac{1}{2}\left(\sum_{i=1}^n \left[\boldsymbol{w}_i^\top\boldsymbol{x}\right]_+ - \left[\boldsymbol{v}^\top\boldsymbol{x}\right]_+\right)^2\right] \\
&\geq \int_{\boldsymbol{x}\in S_i} \frac{1}{2}(\boldsymbol{w}_i^\top\boldsymbol{x})^2 \frac{e^{-\frac{\|\boldsymbol{x}\|^2}{2}}}{(2\pi)^{d/2}} dx \\
&= \int_{\rho=0}^{+\infty}\int_{\omega=\pi/2}^{\theta_i+\pi/2} \frac{1}{4\pi}(\|\boldsymbol{w}_i\|\rho\cos(\omega-\theta_i))^2\rho e^{-\rho^2/2} d\omega d\rho \\
&= \frac{\|\boldsymbol{w}_i\|^2}{2\pi}\int_{\omega=\pi/2}^{\theta_i+\pi/2} \cos^2(\omega-\theta_i) d\omega \\
&= \frac{\|\boldsymbol{w}_i\|^2}{8\pi}\left(2\theta_i - \sin(2\theta_i)\right) \\
&\geq \frac{\|\boldsymbol{w}_i\|^2}{8\pi} \cdot \frac{(2\theta_i)^3}{30}.
\end{aligned}$$

Rearranging terms yields the result. ■

Lemma 21 (Bound of $\|\boldsymbol{r}\|$) *Given that* $\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i\| \geq \frac{\|\boldsymbol{v}\|}{4n}$ *for all* $i \in [n]$ *and* $L(\boldsymbol{w}) = O(n^{-2}\|\boldsymbol{v}\|^2)$, *then*

$$\|\boldsymbol{r}\| = O\left(nL^{1/2}(\boldsymbol{w})\right).$$

Proof Lemma 20 and $\|\boldsymbol{w}_i\| = \Theta(\|\boldsymbol{v}\|/n)$ implies $\theta_i = O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right) = o(1)$. For all $i \in [n]$, by Taylor expansion we have $|(\sin\theta_i - \theta_i\cos\theta_i)| = O(\theta_i^3)$, then $|(\sin\theta_i - \theta_i\cos\theta_i)\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{v}\|| = O(\theta_i^3)\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{v}\| = O\left(nL(\boldsymbol{w})\right)$. Similarly, $\forall i, j \in [n]$ we have $\theta_{ij} \leq \theta_i + \theta_j \leq O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right) \Rightarrow |(\sin\theta_{ij} - \theta_{ij}\cos\theta_{ij})\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{w}_j\|| = O(\theta_{ij}^3)\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{w}_j\| = O\left(L(\boldsymbol{w})\right)$.

Then by (23),

$$\begin{aligned}\|\boldsymbol{r}\|^2 &= 4L(\boldsymbol{w}) - \frac{2}{\pi}\left[\sum_{i<j}(\sin\theta_{ij} - \theta_{ij}\cos\theta_{ij})\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{w}_j\| - \sum_i(\sin\theta_i - \theta_i\cos\theta_i)\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{v}\|\right] \\ &\leq 4L(\boldsymbol{w}) + n^2 O(L(\boldsymbol{w})) + nO(nL(\boldsymbol{w})) \leq O(n^2 L(\boldsymbol{w})),\end{aligned}$$

which implies $\|\boldsymbol{r}\| = O\left(nL^{1/2}(\boldsymbol{w})\right)$. ■

Lemma 22 (Bound of $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\|$) *Suppose* $\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i\| \geq \frac{\|\boldsymbol{v}\|}{4n}, \forall i \in [n]$ *and* $L(\boldsymbol{w}) = O(\|\boldsymbol{v}\|^2/n^2)$. *Then* $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\| \leq O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right)\|\boldsymbol{w}_i\|$.

Proof Lemma 20 and $\|\boldsymbol{w}_i\| = \Theta(\|\boldsymbol{v}\|/n)$ implies $\theta_i = O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right)$. Lemma 21 implies $|H| = |\langle \boldsymbol{r}, \overline{\boldsymbol{v}}\rangle| = O(nL^{1/2}(\boldsymbol{w}))$.

We first decompose $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\|$ into two parts as $\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\| \leq \|\boldsymbol{w}_i - h_i\overline{\boldsymbol{v}}\| + \|h_i\overline{\boldsymbol{v}} - \boldsymbol{w}_i^*\|$.

The first part can be bounded as $\|\boldsymbol{w}_i - h_i\overline{\boldsymbol{v}}\| = \|\boldsymbol{w}_i\|\sin\theta_i \leq O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right)\|\boldsymbol{w}_i\|$.

The second part can be bounded as $\|h_i\overline{\boldsymbol{v}} - \boldsymbol{w}_i^*\| = \left|h_i\left(1 - \frac{\|\boldsymbol{v}\|}{\sum_j h_j}\right)\right| = h_i\frac{|H|}{\|\boldsymbol{v}\| - H} \leq \|\boldsymbol{w}_i\|\frac{|H|}{\|\boldsymbol{v}\| - H}$. Note that $|H| = |\langle \boldsymbol{r}, \overline{\boldsymbol{v}}\rangle| \leq \|\boldsymbol{r}\| = O(nL^{1/2}(\boldsymbol{w})) \leq O(\|\boldsymbol{v}\|) \Rightarrow \|\boldsymbol{v}\| - |H| \geq \|\boldsymbol{v}\|/2$. So we have $\|h_i\overline{\boldsymbol{v}} - \boldsymbol{w}_i^*\| \leq \|\boldsymbol{w}_i\|\frac{|H|}{\|\boldsymbol{v}\| - H} \leq \|\boldsymbol{w}_i\| \cdot \frac{1}{\|\boldsymbol{v}\|/2}O(nL^{1/2}(\boldsymbol{w})) \leq O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right)\|\boldsymbol{w}_i\|$.

Combining two parts together yields the bound. ■

E.2.2. PROOF OF THEOREM 7

Now we are ready to prove Theorem 7.

Theorem 7 *If for every student neuron we have* $\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i\| \geq \frac{\|\boldsymbol{v}\|}{4n}$, *and*

$$L(\boldsymbol{w}) = O\left(\frac{\|\boldsymbol{v}\|^2}{n^{14}}\right),$$

then $\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\| \geq \Omega\left(\frac{L^{2/3}(\boldsymbol{w})}{n^{2/3}\|\boldsymbol{v}\|^{1/3}}\right)$.

Proof Lemma 20 and $\|\boldsymbol{w}_i\| = \Theta(\|\boldsymbol{v}\|/n)$ implies $\theta_i = O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right)$. Lemma 21 implies $\|\boldsymbol{r}\| = O(nL^{1/2}(\boldsymbol{w}))$.

Combined with lemma 8 and $L(\boldsymbol{w}) = O\left(\frac{\|\boldsymbol{v}\|^2}{n^{14}}\right)$ we have

$$\sum_{i=1}^{n}\left\langle \frac{\partial}{\partial \boldsymbol{w}_i}L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^*\right\rangle \geq 2L(\boldsymbol{w}) - O\left(\theta_{\max}^2\|\boldsymbol{r}\|\cdot\|\boldsymbol{v}\|\right) \geq 2L(\boldsymbol{w}) - O(n^{7/3}L^{7/6}(\boldsymbol{w})\|\boldsymbol{v}\|^{-1/3}) \geq L(\boldsymbol{w}).$$

Then

$$\begin{aligned} L(\boldsymbol{w}) &\leq \sum_{i=1}^{n}\left\langle \frac{\partial}{\partial \boldsymbol{w}_i}L(\boldsymbol{w}), \boldsymbol{w}_i - \boldsymbol{w}_i^*\right\rangle \leq \sum_{i=1}^{n}\left\|\frac{\partial}{\partial \boldsymbol{w}_i}L(\boldsymbol{w})\right\|\cdot\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\| \leq \left\|\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}\right\|\sum_{i\in[n]}\|\boldsymbol{w}_i - \boldsymbol{w}_i^*\| \\ &\overset{\text{Lemma 22}}{\leq} \left\|\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}\right\| O\left(n^{2/3}\left(\frac{L(\boldsymbol{w})}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right)\sum_{i\in[n]}\|\boldsymbol{w}_i^*\| = O\left(n^{2/3}L(\boldsymbol{w})^{1/3}\|\boldsymbol{v}\|^{1/3}\right)\left\|\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}\right\|. \end{aligned}$$

So $\left\|\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}\right\| \geq \Omega\left(\frac{L^{2/3}(\boldsymbol{w})}{n^{2/3}\|\boldsymbol{v}\|^{1/3}}\right)$. ■

E.3. Handling Non-smoothness

In this section, we establish two lemmas needed for handling the non-smoothness of L.

Define the Hessian matrix of L as $\Lambda := \frac{\partial^2 L(\boldsymbol{w})}{\partial \boldsymbol{w}^2}$. The next lemma ensures the smoothness of L when the student neurons are regularized, $i.e$, their norms are upper and lower bounded.

Lemma 23 (Conditional Smoothness of L) *If for every student neuron we have $\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i\| \geq \frac{\|\boldsymbol{v}\|}{4n}$, then $\|\Lambda\|_2 \leq O(n^2)$.*

Proof When $\boldsymbol{w}_i \neq \boldsymbol{0}, \forall i$, Safran et al. (2020) has shown that $L(\boldsymbol{w})$ is twice differentiable and computed the closed form expression of Hessian $\Lambda = \frac{\partial^2 L}{\partial \boldsymbol{w}^2} \in \mathbf{R}^{nd\times nd}$:

$$\Lambda = \begin{pmatrix} \Lambda_{1,1} & \cdots & \Lambda_{1,n} \\ \vdots & & \vdots \\ \Lambda_{n,1} & \cdots & \Lambda_{n,n} \end{pmatrix}, \tag{51}$$

where $\Lambda_{i,j} \in \mathbb{R}^{d\times d}, i, j \in [n]$ are $d \times d$ matrices with the following forms:

The i^{th} diagonal block matrix of Λ is

$$\Lambda_{i,i} = \frac{1}{2}I + \sum_{j\neq i}\zeta(\boldsymbol{w}_i, \boldsymbol{w}_j) - \zeta(\boldsymbol{w}_i, \boldsymbol{v}),$$

where

$$\zeta(\boldsymbol{w}, \boldsymbol{v}) = \frac{\sin\theta(\boldsymbol{w}, \boldsymbol{v})\|\boldsymbol{v}\|}{2\pi\|\boldsymbol{w}\|}(I - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top + \overline{\boldsymbol{n}}_{\boldsymbol{v},\boldsymbol{w}}\overline{\boldsymbol{n}}_{\boldsymbol{v},\boldsymbol{w}}^\top),$$

and $\boldsymbol{n}_{\boldsymbol{v},\boldsymbol{w}} = \overline{\boldsymbol{v}} - \cos\theta(\boldsymbol{w}, \boldsymbol{v})\overline{\boldsymbol{w}}$.

For $i \neq j$, the off-diagonal entry is

$$\Lambda_{i,j} = \frac{1}{2\pi}\left[(\pi - \theta(\boldsymbol{w}_i, \boldsymbol{w}_j))I + \overline{\boldsymbol{n}}_{\boldsymbol{w_i},\boldsymbol{w_j}}\overline{\boldsymbol{w}}_j^\top + \overline{\boldsymbol{n}}_{\boldsymbol{w_j},\boldsymbol{w_i}}\overline{\boldsymbol{w}}_i^\top\right].$$

Note that $\|\boldsymbol{w}_i\| = \Theta(\frac{\|\boldsymbol{v}\|}{n}), \forall i$ implies $\frac{\|\boldsymbol{w}_j\|}{\|\boldsymbol{w}_i\|} = O(1), \forall i, j$ and $\frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i\|} = O(n), \forall i$. Then $\|\zeta(\boldsymbol{w}_i, \boldsymbol{w}_j)\| \leq \frac{\|w_j\|}{2\pi\|\boldsymbol{w}_i\|} = O(1)$ and $\|\zeta(\boldsymbol{w}_i, \boldsymbol{v})\| \leq O(1)\frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i\|} \leq O(n)$, so $\|\Lambda_{i,i}\| \leq \|\frac{1}{2}I\| + \sum_{j\neq i}\|\zeta(\boldsymbol{w}_i, \boldsymbol{w}_j)\| + \|\zeta(\boldsymbol{w}_i, \boldsymbol{v})\| \leq O(n)$.

Also note that $\|\Lambda_{i,j}\| \leq \frac{1}{2\pi}(\pi + 1 + 1) \leq 1$ for all $i \neq j$.

Then $\|\Lambda(\boldsymbol{w})\| \leq \sum_{i,j}\|\Lambda_{i,j}\| \leq nO(n) + (n^2 - n) \leq O(n^2)$. ■

The following lemma shows that each student neuron $\boldsymbol{w}_i$ will not move too far in the third phase.

Lemma 24 (Bound of the Change of Neurons) *If the initial loss at phase 3 is upper bounded by $L(\boldsymbol{w}(T_2)) \leq C_l$, and there exists constant $C_s > 0$ such that $L(\boldsymbol{w}(t+1)) \leq L(\boldsymbol{w}(t)) - \frac{\eta}{2}\|\nabla_W(t)\|^2 \leq L(\boldsymbol{w}(t)) - C_s\eta L^{4/3}(\boldsymbol{w}(t)), \forall T + T_2 - 1 \geq t \geq T_2$, then*

$$L(T + T_2) \leq \frac{1}{(L(\boldsymbol{w}(T_2))^{-1/3} + C_s\eta T/3)^3},$$

and

$$\sum_{t=0}^{T-1}\eta\|\nabla_W(t + T_2)\| \leq 8C_s^{-1/2}C_l^{1/3}.$$

Proof We bound the loss as

$$\begin{aligned}
&\frac{1}{L^{1/3}(\boldsymbol{w}(t+1))} \\
&\geq \frac{1}{\left(L(\boldsymbol{w}(t)) - C_s\eta L^{4/3}(\boldsymbol{w}(t))\right)^{1/3}} \\
&= \frac{1}{L^{1/3}(\boldsymbol{w}(t))}\left(1 + \frac{1 - \left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{1/3}}{\left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{1/3}}\right) \\
&= \frac{1}{L^{1/3}(\boldsymbol{w}(t))}\left(1 + \frac{C_s\eta L^{1/3}(\boldsymbol{w}(t))}{\left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{1/3}\left(1 + \left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{1/3} + \left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{2/3}\right)}\right) \\
&= \frac{1}{L^{1/3}(\boldsymbol{w}(t))} + C_s\eta\frac{1}{\left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{1/3} + \left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)^{2/3} + \left(1 - C_s\eta L^{1/3}(\boldsymbol{w}(t))\right)} \\
&\geq \frac{1}{L^{1/3}(\boldsymbol{w}(t))} + \frac{C_s\eta}{3}.
\end{aligned}$$

Therefore, $L^{-1/3}(\boldsymbol{w}(t + T_2)) \geq L^{-1/3}(\boldsymbol{w}(T_2)) + \frac{C_s\eta}{3}t, \forall T \geq t \geq 0$. Let $l_1 \coloneqq L^{-1/3}(\boldsymbol{w}(T_2))$, then $1/l_1^3 \leq C_l$ and

$$L(\boldsymbol{w}(T_2 + t)) \leq \frac{1}{(l_1 + C_s\eta t/3)^3}, \forall t \leq T,$$

this proves the first inequality.

For the second inequality, note that

$$L(\boldsymbol{w}(t+1)) \leq L(\boldsymbol{w}(t)) - \frac{\eta}{2}\|\nabla_W(t)\|^2 \Rightarrow \|\nabla_W(t)\|^2 \leq \frac{2}{\eta}\left(L(\boldsymbol{w}(t)) - L(\boldsymbol{w}(t+1))\right).$$

By Cauchy inequality, $\forall T > 0$,

$$
\begin{aligned}
&\left(\sum_{t=0}^{T-1} \|\nabla_W(T_2+t)\|\right)^2 \\
\leq &\left(\sum_{t=0}^{T-1} \frac{1}{(l_1+C_s\eta t/3)^2}\right)\left(\sum_{t=0}^{T-1}(l_1+C_s\eta t/3)^2\|\nabla_W(T_2+t)\|^2\right) \\
\leq &\left(\sum_{t=0}^{T-1} \frac{1}{(l_1+C_s\eta t/3)^2}\right)\left(\sum_{t=0}^{T-1}(l_1+C_s\eta t/3)^2\frac{2}{\eta}\left(L(\boldsymbol{w}(T_2+t))-L(\boldsymbol{w}(T_2+t+1))\right)\right) \\
\leq &\frac{2}{\eta}\left(\sum_{t=0}^{T-1} \frac{1}{(l_1+C_s\eta t/3)^2}\right)\left(l_1^2 L(\boldsymbol{w}(T_2+T))+\sum_{t=1}^{T-1}\left((l_1+C_s\eta t/3)^2-(l_1+C_s\eta(t-1)/3)^2\right)L(\boldsymbol{w}(T_2+t))\right) \\
\leq &\frac{2}{\eta}\left(\sum_{t=0}^{T-1} \frac{1}{(l_1+C_s\eta t/3)^2}\right)\left(l_1^2 L(\boldsymbol{w}(T_2+T))+\frac{2C_s\eta}{3}\sum_{t=1}^{T-1}(l_1+C_s\eta t/3)\,L(\boldsymbol{w}(T_2+t))\right) \\
\leq &\frac{2}{\eta}\left(\sum_{t=0}^{T-1} \frac{1}{(l_1+C_s\eta t/3)^2}\right)\left(\frac{1}{l_1}+\frac{2C_s\eta}{3}\sum_{t=1}^{T-1}\frac{1}{(l_1+C_s\eta t/3)^2}\right)
\end{aligned}
$$

Note that

$$
\sum_{t=0}^{T-1}\frac{1}{(l_1+C_s\eta t/3)^2} \leq \sum_{t=0}^{+\infty}\frac{1}{(l_1+C_s\eta t/3)^2} \leq \frac{3}{C_s\eta}\sum_{t=0}^{+\infty}\left(\frac{1}{l_1+C_s\eta(t-1)/3}-\frac{1}{l_1+C_s\eta t/3}\right) \leq \frac{6}{C_s\eta l_1}.
$$

Therefore,

$$
\left(\sum_{t=0}^{T-1}\|\nabla_W(T_2+t)\|\right)^2 \leq \frac{2}{\eta}\left(\frac{6}{C_s\eta l_1}\right)\left(\frac{1}{l_1}+\frac{2C_s\eta}{3}\frac{6}{C_s\eta l_1}\right) \leq \frac{2}{\eta}\frac{6}{C_s\eta l_1}\frac{5}{l_1} = \frac{60}{C_s\eta^2 l_1^2}.
$$

So we get

$$
\eta\sum_{t=0}^{T}\|\nabla(T_2+t)\| \leq \eta\sqrt{\frac{60}{C_s\eta^2 l_1^2}} \leq 8C_s^{-1/2}C_l^{1/3}.
$$

■

E.4. Proof of Theorem 6

Theorem 6 *Suppose the initial condition in Lemma 3 holds. If we set $\epsilon_2 = O(n^{-14})$ in Theorem 5, $\eta = O\left(\frac{1}{n^2}\right)$, then $\forall T \in \mathbf{N}$ we have*

$$
\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i(T+T_2)\| \geq \frac{\|\boldsymbol{v}\|}{4n}, \tag{52}
$$

and

$$
L(T+T_2) \leq O\left(\frac{n^4\|\boldsymbol{v}\|^2}{(\eta T)^3}\right). \tag{53}
$$

Proof Since we have set $\epsilon_2 = O(n^{-14})$, by Lemma 19 we have $\frac{3\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i(T_2)\| \geq \frac{\|\boldsymbol{v}\|}{3n}, \forall i$, and $L(\boldsymbol{w}(T_2)) \leq 20\epsilon_2\|\boldsymbol{v}\|^2 = O\left(\frac{\|\boldsymbol{v}\|^2}{n^{14}}\right)$.

To prove the theorem, we just need to prove (52) and a stronger version of (53):

$$L(T+T_2) \leq \frac{1}{\left(L(T_2)^{-1/3} + \Omega\left(\frac{1}{n^{4/3}\|\boldsymbol{v}\|^{2/3}}\right)\eta T\right)^3}. \tag{54}$$

We prove (52) and (54) together inductively.

The induction base holds for $T = 0$ by Lemma 19.

Now suppose (52) (54) hold for $0, 1, \ldots, T-1$, we show the case of T.

First note that (4) (52) and routine computation implies $\|\nabla_i(t+T_2)\| = O(\|\boldsymbol{v}\|), \forall 0 \leq t \leq T-1$.

For $\forall T_2 \leq t \leq T+T_2-1$, since the induction condition together with lemma 23 guarantee the smoothness of L, the classical analysis of gradient descent can be applied ([Nesterov et al. (2018)], lemma 1.2.3) to bound the decrease of loss at time t as

$$\begin{aligned} L(\boldsymbol{w}(t+1)) =& L(\boldsymbol{w}(t)) + \langle \nabla_W(t), -\eta\nabla_W(t)\rangle + \\ & \int_{\tau=0}^{1} (1-\tau)(-\eta\nabla_W(t))^\top \frac{\partial^2 L}{\partial \boldsymbol{w}^2}(\boldsymbol{w}(t) - \tau\eta\nabla_W(t))(-\eta\nabla_W(t))\mathrm{d}\tau. \end{aligned} \tag{55}$$

For $\forall \tau \in [0,1]$, $\|\boldsymbol{w}_i(t) - \tau\eta\nabla_i(t)\| \geq \|\boldsymbol{w}_i(t)\| - \eta\|\nabla_i(t)\| \geq \frac{\|\boldsymbol{v}\|}{4n} - \eta O(\|\boldsymbol{v}\|) \geq \frac{\|\boldsymbol{v}\|}{5n}$, similarly we have $\|\boldsymbol{w}_i(t) - \tau\eta\nabla_i(t)\| \leq \|\boldsymbol{w}_i(t)\| + \eta\|\nabla_i(t)\| \leq \frac{5\|\boldsymbol{v}\|}{n}$. Then lemma 23 implies the smoothness of L at $\boldsymbol{w}(t) - \tau\eta\nabla_W(t)$: $\left\|\frac{\partial^2 L}{\partial \boldsymbol{w}^2}(\boldsymbol{w}(t) - \tau\eta\nabla_W(t))\right\| \leq O\left(n^2\right)$. Combined with gradient lower bound Theorem 7 (note that $L(\boldsymbol{w}(t)) = O(\|\boldsymbol{v}\|^2/n^{14})$), the dynamic of loss can be bounded as

$$\begin{aligned} L(\boldsymbol{w}(t)) - L(\boldsymbol{w}(t+1)) &\geq \eta\|\nabla_W(t)\|^2 - \int_{\tau=0}^{1} (1-\tau)O\left(n^2\right)\| - \eta\nabla_W(t)\|^2\mathrm{d}\tau \\ &\geq \frac{\eta}{2}\|\nabla_W(t)\|^2 \\ &\overset{\text{Theorem 7}}{\geq} \Omega\left(\frac{1}{n^{4/3}\|\boldsymbol{v}\|^{2/3}}\right)\eta L^{4/3}(\boldsymbol{w}(t)). \end{aligned}$$

Set C_s in Lemma 24 as $C_s = \Omega\left(\frac{1}{n^{4/3}\|\boldsymbol{v}\|^{2/3}}\right)$. For $\forall T+T_2-1 \geq t \geq T_2$, the above inequality implies that $L(\boldsymbol{w}(t+1)) \leq L(\boldsymbol{w}(t)) - \frac{\eta}{2}\|\nabla_W(t)\|^2 \leq L(\boldsymbol{w}(t)) - C_s\eta L^{4/3}(\boldsymbol{w}(t))$. So we can apply Lemma 24 here, which immediately implies (54).

For (52), Lemma 24 yields

$$\begin{aligned} \|\boldsymbol{w}_i(T+T_2)\| &\geq \|\boldsymbol{w}_i(T_2)\| - \sum_{t=0}^{T-1}\eta\|\nabla_W(t+T_2)\| \geq \frac{\|\boldsymbol{v}\|}{3n} - 8C_s^{-1/2}L(\boldsymbol{w}(T_2))^{1/3} \\ &= \frac{\|\boldsymbol{v}\|}{3n} - O\left(n^{2/3}\|\boldsymbol{v}\|^{1/3}\cdot\left(\frac{\|\boldsymbol{v}\|^2}{n^{14}}\right)^{1/3}\right) \geq \frac{\|\boldsymbol{v}\|}{4n}, \end{aligned}$$

similarly

$$\|\boldsymbol{w}_i(T+T_2)\| \leq \|\boldsymbol{w}_i(T_2)\| + \sum_{t=0}^{T-1}\eta\|\nabla_W(t+T_2)\| \leq \frac{3\|\boldsymbol{v}\|}{n} + O\left(n^{2/3}\|\boldsymbol{v}\|^{1/3}\cdot\left(\frac{\|\boldsymbol{v}\|^2}{n^{14}}\right)^{1/3}\right) \leq \frac{4\|\boldsymbol{v}\|}{n}.$$

■

Appendix F. Supplementary Materials for Section 4

F.1. Proof of Lemma 3

Lemma 3 *Let $s_1 \coloneqq \frac{1}{2}\sigma\sqrt{d}, s_2 \coloneqq 2\sigma\sqrt{d}$. When $d = \Omega(\log(n/\delta))$, with probability at least $1-\delta$, the following properties holds:*

$$\forall i \in [n], s_1 \le \|\boldsymbol{w}_i(0)\| \le s_2, \tag{56}$$

$$\forall i \in [n], \frac{\pi}{3} \le \theta_i(0) \le \frac{2\pi}{3}. \tag{57}$$

Proof By concentration inequality of Gaussian (See Section 2 in Dasgupta and Schulman (2013)), For $\forall i$ we have $\Pr\left[\|\boldsymbol{w}_i(0)\| < \frac{1}{2}\sigma\sqrt{d} \vee \|\boldsymbol{w}_i(0)\| > 2\sigma\sqrt{d}\right] \le \Pr\left[|\|\boldsymbol{w}_i(0)\|^2 - \sigma^2 d| > \frac{3}{4}\sigma^2 d\right] \le \exp\left(-\left(\frac{3}{4}\right)^2 d/24\right) \le \exp(-\Omega(\log(n/\delta))) \le \frac{\delta}{3n}$. By union bound, (56) holds with probability at least $1-\delta/3$.

For (57), note that for $\forall i \in [n]$,

$$|\langle \boldsymbol{w}_i(0), \overline{\boldsymbol{v}}\rangle| \le \frac{1}{4}\sigma\sqrt{d} \wedge \|\boldsymbol{w}_i(0)\| \ge \frac{1}{2}\sigma\sqrt{d} \Rightarrow \frac{|\langle \boldsymbol{w}_i(0), \overline{\boldsymbol{v}}\rangle|}{\|\boldsymbol{w}_i(0)\|} \le \frac{1}{2} \Rightarrow \frac{\pi}{3} \le \theta_i(0) \le \frac{2\pi}{3}.$$

By concentration inequality of Gaussian, $\Pr\left[|\langle \boldsymbol{w}_i(0), \overline{\boldsymbol{v}}\rangle| > \frac{1}{4}\sigma\sqrt{d}\right] \le 2\exp\left(-\frac{(\frac{1}{4}\sigma\sqrt{d})^2}{2\sigma^2}\right) \le \frac{\delta}{3n}$. Then $\Pr\left[\theta_i(0) < \frac{\pi}{3} \vee \theta_i(0) > \frac{2\pi}{3}\right] \le \Pr\left[|\langle \boldsymbol{w}_i(0), \overline{\boldsymbol{v}}\rangle| > \frac{1}{4}\sigma\sqrt{d}\right] + \Pr\left[\|\boldsymbol{w}_i(0)\| < \frac{1}{2}\sigma\sqrt{d}\right] \le \frac{2\delta}{3n}$. By union bound, (57) holds with probability at least $1-2\delta/3$. Applying union bound again finishes the proof. ■

F.2. Parameter Valuation

In this section, we assign values to all intermediate parameters appeared in Theorem 4, Theorem 5 and Theorem 6, according to the requirements of these theorems.

- First we set $\epsilon_2 = O(n^{-14})$ in Theorem 5 as required by Theorem 6.
- Set $\epsilon_1 = O(\epsilon_2^6 n^{-1/2}) = O(n^{-84.5})$ in Theorem 4 as required by Theorem 5.
- Set $C = O\left(\frac{\epsilon_1^2}{n}\right) = O(n^{-170})$ in Theorem 4.
- Set $\sigma = O\left(C\epsilon_1^{48} d^{-1/2}\|\boldsymbol{v}\|\right) = O\left(\epsilon_1^{50} d^{-1/2}\|\boldsymbol{v}\|/n\right) = O\left(n^{-4226} d^{-1/2}\|\boldsymbol{v}\|\right)$ in Theorem 4.
- Set $\eta = O\left(\frac{\epsilon_1^2 \sigma^2 d}{\|\boldsymbol{v}\|^2}\right) = O\left(\frac{\sigma^2 d}{n^{169}\|\boldsymbol{v}\|^2}\right)$ as required by Theorem 5. (Note that in Phase 1 and 3, Theorem 4 and Theorem 6 also have requirements for η, but the bound in Theorem 5 is the tightest one.)

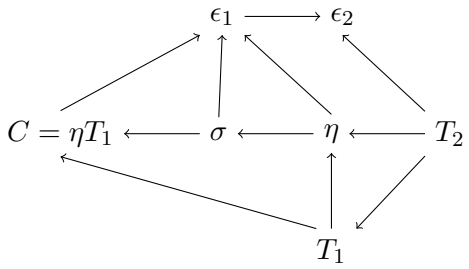

Figure 3: Parameter Dependency Graph

- Set $T_1 = \frac{C}{\eta} = O\left(\frac{\epsilon_1^2}{n\eta}\right)$ in Theorem 4.
- Finally, set $T_2 = T_1 + \left\lceil \frac{1}{n\eta} \ln\left(\frac{1}{36\epsilon_2}\right) \right\rceil = O\left(\frac{\epsilon_1^2}{n\eta}\right) + O\left(\frac{\log(1/\epsilon_2)}{n\eta}\right) = O\left(\frac{\log(1/\epsilon_2)}{n\eta}\right) = O\left(\frac{\log n}{n\eta}\right)$ in Theorem 5.

The dependence between these parameters is shown in Figure 3. (We use arrows to indicate dependency, $e.g.$, the arrow from ϵ_1 to ϵ_2 indicates that ϵ_1 depends on ϵ_2.)

F.3. Non-degeneracy of Student Neurons

There is a technical issue in the convergence analysis: if one of the student neuron $\boldsymbol{w}_i$ is degenerate and $\boldsymbol{w}_i = \mathbf{0}$, the loss function $L(\boldsymbol{w})$ is not differentiable, hence gradient descent is not well-defined.

However, our proof shows that such case would not happen and the student neurons are always non-degenerate. Note that the student neuron's norm $\|\boldsymbol{w}_i\|$ is always lower-bounded in all three phases of our analysis (Phase 1: (24) in Theorem 4, Phase 2: (43) in Theorem 5, Phase 3: (52) in Theorem 6). By these bounds we have the following corollary describing the non-degeneracy of student neurons.

Corollary 25 *If the initialization conditions in Lemma 3 hold, then for* $\forall i \in [n], t \in \mathbf{N}$, $\|\boldsymbol{w}_i(t)\| > 0$.

Remark 26 *Note that an assumption on the initialization condition such as the one in Corollary 25 is necessary, otherwise there would be counter-examples where all student neurons are degenerate. For example,* $\forall c > 0$, *if we set* $\boldsymbol{w}_i(0) = -c\boldsymbol{v}, \forall i$ *and* $\eta = \frac{2c}{1+nc}$, *then straightforward calculation shows that* $\forall i, \nabla_i(0) = -\frac{1+nc}{2}\boldsymbol{v} \Rightarrow \forall i, \boldsymbol{w}_i(1) = \mathbf{0}$..

Appendix G. Lower Bound of the Convergence Rate

G.1. Preliminaries

In this section, we do some technical preparations for proving Theorem 14.

Taking $\eta \to 0$, we get the gradient flow version of Theorem 6.

Theorem 27 *Suppose gradient flow is initialized from a point $\boldsymbol{w}(0)$ where the conditions in Lemma 3 hold. If $\sigma = O\left(n^{-4226}d^{-1/2}\|\boldsymbol{v}\|\right)$, then there exists $T_2 = O\left(\frac{\log n}{n}\right)$ such that $\forall T \in \mathbf{R}_{\geq 0}$ we have*

$$\frac{4\|\boldsymbol{v}\|}{n} \geq \|\boldsymbol{w}_i(T+T_2)\| \geq \frac{\|\boldsymbol{v}\|}{4n}, \tag{58}$$

and

$$L(T+T_2) \leq O\left(\frac{n^4\|\boldsymbol{v}\|^2}{T^3}\right). \tag{59}$$

Similarly, we also have the gradient flow version of Corollary 25.

Corollary 28 *Given that the initial condition in Lemma 3 holds, then for $\forall i \in [n], t \in \mathbf{N}$, $\|\boldsymbol{w}_i(t)\| > 0$.*

Lemma 29 *Given that the initial condition in Lemma 3 holds, and the initialization is non-degenerate, then $\forall t, \exists i \in [n],\ s.t.\ \boldsymbol{z}_i(t) \neq \mathbf{0}$.*

Proof Assume for contradiction that $\exists t \in \mathbb{R}^+$ such that $\boldsymbol{z}_1(t) = \cdots = \boldsymbol{z}_n(t) = \mathbf{0}$. Define

$$\bar{t} := \inf\{t | \boldsymbol{z}_1(t) = \cdots = \boldsymbol{z}_n(t) = \mathbf{0}\}.$$

Then the continuity of $\boldsymbol{z}_i$ implies $\boldsymbol{z}_1(\bar{t}) = \cdots = \boldsymbol{z}_n(\bar{t}) = \mathbf{0}$, so $\bar{t} > 0$. On the other hand, Corollary 25 indicates that $\boldsymbol{w}_i(\bar{t}) \neq \mathbf{0}, \forall i \in [n]$. Since $\boldsymbol{w}_i$ is continuous, there exists a neighborhood of $\bar{t}$ such that for $\forall i, j \in [n]$, $\|\boldsymbol{w}_i(t)\|/\|\boldsymbol{w}_j(t)\|$ and $\|\boldsymbol{v}\|/\|\boldsymbol{w}_i(t)\|$ are bounded by a fixed constant when t is in this neighborhood. Furthermore, since $\bar{t} > 0$, $\exists \epsilon > 0$ and constant $C > 1$ such that for $\forall t \in [\bar{t} - \epsilon, \bar{t}], \forall i \in [n]$ we have

$$\left|\pi + \sum_{j\neq i} \frac{\|\boldsymbol{w}_j(t)\|}{\|\boldsymbol{w}_i(t)\|} \sin\theta_{ij}(t) - \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i(t)\|}\sin\theta_i(t)\right| \leq \pi + \sum_{j\neq i} \frac{\|\boldsymbol{w}_j(t)\|}{\|\boldsymbol{w}_i(t)\|}| + \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i(t)\|} \leq C.$$

Then in the interval $[\bar{t} - \epsilon, \bar{t}]$ we have

$$\left\|\frac{\partial \boldsymbol{z}_i}{\partial t}\right\| = \left\|-\frac{1}{2\pi}\left(\pi + \sum_{j\neq i}\frac{\|\boldsymbol{w}_j\|}{\|\boldsymbol{w}_i\|}\sin\theta_{ij} - \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i\|}\sin\theta_i\right)\boldsymbol{z}_i - \sum_{j\neq i}\frac{\pi-\theta_{ij}}{2\pi}\boldsymbol{z}_j\right\| \leq C\|\boldsymbol{z}_i\| + \sum_{j\neq i}\|\boldsymbol{z}_j\|,$$

$$\Rightarrow \frac{\partial\|\boldsymbol{z}_i\|^2}{\partial t} = 2\left\langle\frac{\partial \boldsymbol{z}_i}{\partial t}, \boldsymbol{z}_i\right\rangle \geq -2C\|\boldsymbol{z}_i\|\left(\sum_{j\in[n]}\|\boldsymbol{z}_j\|\right)$$

$$\Rightarrow \frac{\partial}{\partial t}\sum_{j\in[n]}\|\boldsymbol{z}_j\|^2 \geq -2C\left(\sum_{j\in[n]}\|\boldsymbol{z}_j\|\right)^2 \geq -2nC\left(\sum_{j\in[n]}\|\boldsymbol{z}_j\|^2\right)$$

$$\Rightarrow \frac{\partial}{\partial t}\left[e^{2nCt}\left(\sum_{j\in[n]}\|\boldsymbol{z}_j\|^2\right)\right] = e^{2nCt}\left[2nC\left(\sum_{j\in[n]}\|\boldsymbol{z}_j\|^2\right) + \frac{\partial}{\partial t}\left(\sum_{j\in[n]}\|\boldsymbol{z}_j\|^2\right)\right] \geq 0$$

$$\Rightarrow \sum_{j\in[n]}\|\boldsymbol{z}_j(\bar{t})\|^2 \geq e^{-2nC\epsilon}\left(\sum_{j\in[n]}\|\boldsymbol{z}_j(\bar{t}-\epsilon)\|^2\right).$$

Here the (t) indicator is omitted for simplicity. Note that we bound $\frac{\partial \|\boldsymbol{z}_i\|^2}{\partial t}$ instead of $\frac{\partial \|\boldsymbol{z}_i\|}{\partial t}$ here, since $\|\boldsymbol{z}_i\|$ might not be differentiable if $\boldsymbol{z}_i = \boldsymbol{0}$, while $\|\boldsymbol{z}_i\|^2$ is always differentiable.

Finally, due to the definition of $\bar{t}$, there exists $i \in [n]$ such that $\boldsymbol{z}_i(\bar{t}-\epsilon) \neq \boldsymbol{0}$. Then $\sum_{j\in[n]} \|\boldsymbol{z}_j(\bar{t})\|^2 \geq e^{-2nC\epsilon}\left(\sum_{j\in[n]} \|\boldsymbol{z}_j(\bar{t}-\epsilon)\|^2\right) > 0$, a contradiction. ■

G.2. Proofs for Section 5

By the closed form formula of gradient (4), the dynamics of $\boldsymbol{z}_i$ is given by

$$\frac{\partial \boldsymbol{z}_i}{\partial t} = -\frac{1}{2\pi}\left(\pi + \sum_{j\neq i}\frac{\|\boldsymbol{w}_j\|}{\|\boldsymbol{w}_i\|}\sin\theta_{ij} - \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{w}_i\|}\sin\theta_i\right)\boldsymbol{z}_i - \sum_{j\neq i}\frac{\pi-\theta_{ij}}{2\pi}\boldsymbol{z}_j. \tag{60}$$

Lemma 17 *If there exists i, j such that $\kappa_{ij}(t) = \kappa_{\max}(t) < \frac{\pi}{2}$, then $\cos\kappa_{ij}(t)$ is well-defined in an open neighborhood of t, differentiable at t, and*

$$\frac{\partial}{\partial t}\cos\kappa_{ij}(t) \leq -\frac{\pi-\theta_{ij}(t)}{\pi}(1-\cos\kappa_{ij}^2(t)).$$

Proof First note that for $\forall i \in Q^+(t)$, $\|\boldsymbol{z}_i(t)\| > 0 \Rightarrow \overline{\boldsymbol{z}_i(t)} = \frac{\boldsymbol{z}_i(t)}{\|\boldsymbol{z}_i(t)\|}$ is differentiable at t. Therefore, for $\forall i, j \in Q^+(t)$, $\cos\kappa_{ij} = \langle\overline{\boldsymbol{z}_i(t)}, \overline{\boldsymbol{z}_j(t)}\rangle$ is well-defined in an open neighborhood of t and differentiable at t. According to the dynamics of $\boldsymbol{z}_i$ (60), $\forall i \in Q^+(t)$ we have

$$\frac{\partial}{\partial t}\overline{\boldsymbol{z}_i} = \frac{\|\boldsymbol{z}_i\|\frac{\partial \boldsymbol{z}_i}{\partial t} - \frac{\partial\|\boldsymbol{z}_i\|}{\partial t}\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|^2} = \frac{\frac{\partial \boldsymbol{z}_i}{\partial t} - \langle\frac{\partial \boldsymbol{z}_i}{\partial t}, \overline{\boldsymbol{z}_i}\rangle\overline{\boldsymbol{z}_i}}{\|\boldsymbol{z}_i\|} = -\sum_{k\neq i}\frac{\pi-\theta_{ik}}{2\pi}\frac{\boldsymbol{z}_k - \langle\boldsymbol{z}_k, \overline{\boldsymbol{z}_i}\rangle\overline{\boldsymbol{z}_i}}{\|\boldsymbol{z}_i\|}.$$

Then

$$\begin{aligned}
&\frac{\partial\cos\kappa_{ij}}{\partial t} = \frac{\partial\langle\overline{\boldsymbol{z}_i}, \overline{\boldsymbol{z}_j}\rangle}{\partial t}\\
=&\left\langle\frac{\partial}{\partial t}\overline{\boldsymbol{z}_i}, \overline{\boldsymbol{z}_j}\right\rangle + \left\langle\overline{\boldsymbol{z}_i}, \frac{\partial}{\partial t}\overline{\boldsymbol{z}_j}\right\rangle\\
=&-\sum_{k\neq i, k\in Q^+}\frac{\pi-\theta_{ik}}{2\pi}\frac{\|\boldsymbol{z}_k\|}{\|\boldsymbol{z}_i\|}(\langle\overline{\boldsymbol{z}_k}, \overline{\boldsymbol{z}_j}\rangle - \langle\overline{\boldsymbol{z}_k}, \overline{\boldsymbol{z}_i}\rangle\langle\overline{\boldsymbol{z}_i}, \overline{\boldsymbol{z}_j}\rangle) - \sum_{k\neq j, k\in Q^+}\frac{\pi-\theta_{jk}}{2\pi}\frac{\|\boldsymbol{z}_k\|}{\|\boldsymbol{z}_j\|}(\langle\overline{\boldsymbol{z}_k}, \overline{\boldsymbol{z}_i}\rangle - \langle\overline{\boldsymbol{z}_k}, \overline{\boldsymbol{z}_j}\rangle\langle\overline{\boldsymbol{z}_i}, \overline{\boldsymbol{z}_j}\rangle)\\
=&\underbrace{-\frac{\pi-\theta_{ij}}{2\pi}\left(\frac{\|\boldsymbol{z}_i\|}{\|\boldsymbol{z}_j\|} + \frac{\|\boldsymbol{z}_j\|}{\|\boldsymbol{z}_i\|}\right)(1-\cos^2\kappa_{ij})}_{I_1}\\
&\underbrace{-\sum_{k\neq i, j\wedge k\in Q^+}\left[\frac{\pi-\theta_{ik}}{2\pi}\frac{\|\boldsymbol{z}_k\|}{\|\boldsymbol{z}_i\|}(\cos\kappa_{kj} - \cos\kappa_{ki}\cos\kappa_{ij}) + \frac{\pi-\theta_{jk}}{2\pi}\frac{\|\boldsymbol{z}_k\|}{\|\boldsymbol{z}_j\|}(\cos\kappa_{ik} - \cos\kappa_{kj}\cos\kappa_{ij})\right]}_{I_2}.
\end{aligned} \tag{61}$$

The expression above splits into two terms I_1 and I_2. The most important observation is that, by setting $\overline{\boldsymbol{z}_i}, \overline{\boldsymbol{z}_j}$ to be the pair of maximally separated vectors, i.e., $\kappa_{ij} = \kappa_{\max}$, the second term I_2

is guaranteed to be nonpositive. This is because for $\forall k$,

$$\kappa_{ki} \le \kappa_{\max} < \pi/2, \kappa_{kj} \le \kappa_{\max} = \kappa_{ij} \Rightarrow \cos\kappa_{kj} \ge \cos\kappa_{ij} \ge \cos\kappa_{ki}\cos\kappa_{ij} \Rightarrow \cos\kappa_{kj} - \cos\kappa_{ki}\cos\kappa_{ij} \ge 0,$$

similarly we have $\cos\kappa_{ik} - \cos\kappa_{kj}\cos\kappa_{ij} \ge 0, \forall k$. So $I_2 \le 0$ when $\kappa_{ij} = \kappa_{\max}$. This implies that when $\kappa_{ij} = \kappa_{\max}$,

$$\frac{\partial \cos\kappa_{ij}}{\partial t} \le I_1 = -\frac{\pi - \theta_{ij}}{2\pi}\left(\frac{\|\boldsymbol{z}_i\|}{\|\boldsymbol{z}_j\|} + \frac{\|\boldsymbol{z}_j\|}{\|\boldsymbol{z}_i\|}\right)(1 - \cos^2\kappa_{ij}) \le -\frac{\pi - \theta_{ij}}{\pi}(1 - \cos^2\kappa_{ij}).$$

■

Lemma 30 *Given that the initial condition in Lemma 3 holds, suppose the network is over-parameterized, i.e., $n \ge 2$, and the initialization is non-degenerate, then for $\forall t \in \mathbb{R}_{\ge 0}$, at least one of the following two conditions must hold:*

$$\exists i \in [n] \ s.t. \ \boldsymbol{z}_i(t) = \boldsymbol{0}, \tag{62}$$

$$\kappa_{\max}(t) \ge \frac{\kappa_{\max}(0)}{3}. \tag{63}$$

Proof Assume for contradiction that $\exists t$ such that $\boldsymbol{z}_i(t) \ne \boldsymbol{0}, \forall i$ and $\kappa_{\max}(t) < \frac{\kappa_{\max}(0)}{2}$. Then we can define

$$t^* = \inf\left\{t \in \mathbb{R} | \forall i, \boldsymbol{z}_i(t) \ne \boldsymbol{0} \wedge \kappa_{\max}(t) < \frac{\kappa_{\max}(0)}{3}\right\}.$$

Note that by lemma 29 we have $Q^+(t^*) \ge 1$. For $\forall i, j \in Q^+(t^*)$, if $\kappa_{ij}(t^*) > \kappa_{\max}(0)/3$, due to the continuity of κ_{ij}, $\kappa_{ij} > \kappa_{\max}(0)/3$ holds in an open neighborhood of t^*, which contradicts the definition of t^*. So $\forall i, j \in Q^+(t^*)$ we have $\kappa_{ij}(t^*) \le \kappa_{\max}(0)/3$.

Step 1: First we prove $\forall i, \boldsymbol{z}_i(t^*) \ne \boldsymbol{0}$.

If $\exists i$ such that $\boldsymbol{z}_i(t^*) = \boldsymbol{0}$, then for such i we have $\frac{\partial \boldsymbol{z}_i(t^*)}{\partial t} = -\sum_{j\ne i}\frac{\pi - \theta_{ij}(t^*)}{2\pi}\boldsymbol{z}_j(t^*)$. Since $Q^+(t^*) \ne \varnothing$, pick $k \in Q^+(t^*)$ and we have

$$\left\langle \frac{\partial \boldsymbol{z}_i(t^*)}{\partial t}, \boldsymbol{z}_k(t^*)\right\rangle = -\sum_{j\ne i \wedge j \in Q^+(t^*)}\frac{\pi - \theta_{ij}(t^*)}{2\pi}\langle \boldsymbol{z}_j(t^*), \boldsymbol{z}_k(t^*)\rangle < 0,$$

where the last inequality is because $\kappa_{jk}(t^*) \le \kappa_{\max}(0)/3 < \frac{\pi}{2}$.

On the other hand, the definition of $\frac{\partial \boldsymbol{z}_i(t^*)}{\partial t}$ implies that $\exists \epsilon > 0$, $\forall t' \in [t^*, t^* + \epsilon)$, $\boldsymbol{z}_i(t') = \boldsymbol{z}_i(t^*) + (t' - t^*)\frac{\partial \boldsymbol{z}_i(t^*)}{\partial t} + \boldsymbol{o}(t' - t^*) = (t' - t^*)\frac{\partial \boldsymbol{z}_i(t^*)}{\partial t} + \boldsymbol{o}(t' - t^*)$. Similarly, $\boldsymbol{z}_k(t') = \boldsymbol{z}_k(t^*) + (t' - t^*)\frac{\partial \boldsymbol{z}_k(t^*)}{\partial t} + \boldsymbol{o}(t' - t^*)$. Then

$$\langle \boldsymbol{z}_i(t'), \boldsymbol{z}_k(t')\rangle = (t' - t^*)\left\langle \frac{\partial \boldsymbol{z}_i(t^*)}{\partial t}, \boldsymbol{z}_k(t^*)\right\rangle + o(t' - t^*). \tag{64}$$

Since $\left\langle \frac{\partial \boldsymbol{z}_i(t^*)}{\partial t}, \boldsymbol{z}_k(t^*)\right\rangle < 0$ is a negative constant, there exists $\epsilon' > 0$ such that for $\forall t' \in [t^*, t^* + \epsilon')$, (64) is negative, consequently $\kappa_{ik}(t') = \arccos\left(\left\langle \overline{\boldsymbol{z}_i(t')}, \overline{\boldsymbol{z}_k(t')}\right\rangle\right) > \frac{\pi}{2}$. So $\forall t' \in [t^*, t^* + \epsilon')$, $\kappa_{\max}(t') \ge \kappa_{ik}(t') > \pi/2 \ge \kappa_{\max}(0)/3$, this contradicts the definition of t^*.

Step 2: After proving $\boldsymbol{z}_i(t^*) \neq \mathbf{0}, \forall i$ and $\kappa_{ij}(t^*) \leq \kappa_{\max}(0)/3 < \kappa_{\max}(0), \forall i, j \in [n]$, we aim to derive a contradiction.

Note that $t^* \neq 0$ due to the definition of $\kappa_{\max}$. By the continuity of $\boldsymbol{z}_i$, $\exists \epsilon_1 > 0$ such that for $\forall t \in (t^* - \epsilon_1, t^* + \epsilon_1), i \in [n]$, $\boldsymbol{z}_i(t) \neq \mathbf{0}$. Then the definition of t^* implies that $\forall t \in (t^* - \epsilon_1, t^*)$, $\kappa_{\max}(t) \geq \kappa_{\max}(0)/3$. Since on the interval $(t^* - \epsilon_1, t^* + \epsilon_1)$, $\kappa_{\max} = \max_{i,j\in[n]} \kappa_{ij}$ is continuous [4], we have that $\kappa_{\max}(t^*) \geq \kappa_{\max}(0)/3$. Then $\kappa_{\max}(t^*) = \kappa_{\max}(0)/3$. Pick i, j such that $\kappa_{ij}(t^*) = \kappa_{\max}(t^*)$. Note that $\theta_{ij}(t^*) < \pi$, otherwise $\kappa_{ij}(t^*) = \pi$, a contradiction. Then by lemma 17 we have

$$\frac{\partial}{\partial t} \cos \kappa_{ij}(t^*) \leq -\frac{\pi - \theta_{ij}(t^*)}{\pi}(1 - \cos \kappa_{ij}^2(t^*)) < 0.$$

So $\frac{\partial}{\partial t}\kappa_{ij}(t^*) > 0 \Rightarrow \exists \epsilon_2 > 0 \; s.t. \; \forall t \in (t^*, t^* + \epsilon_2), \kappa_{\max}(t) \geq \kappa_{ij}(t) > \kappa_{ij}(t^*) = \kappa_{\max}(t^*) = \kappa_{\max}(0)/3$, this contradicts the definition of t^*. ■

Lemma 31 *Given that the initial condition in Lemma 3 holds, suppose the network is over-parameterized, i.e., $n \geq 2$, and the initialization is non-degenerate, then for $\forall t \in \mathbb{R}_{\geq 0}$ we have*

$$Z(t) \geq \Omega(\kappa_{\max}(0) \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|).$$

Proof We show that for $\forall t \in \mathbb{R}_{\geq 0}$,

$$\max_{i,j\in[n]} \|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\| \geq \Omega(\kappa_{\max}(0) \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|). \tag{65}$$

W.L.O.G., suppose $\boldsymbol{z}_1(t) = \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|$. By lemma 30, for $\forall t$ one of the following two cases must happen:

- $\exists k \; s.t. \; \boldsymbol{z}_k(t) = \mathbf{0}$.

  By lemma 29, $k \neq 1$. Then $\max_{i,j\in[n]} \|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\| \geq \|\boldsymbol{z}_1(t) - \boldsymbol{z}_k(t)\| = \|\boldsymbol{z}_1(t)\| \geq O(\kappa_{\max}(0) \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|)$.

- $\kappa_{\max}(t) \geq \kappa_{\max}(0)/3$.

  Pick a pair i, j such that $\kappa_{ij}(t) = \kappa_{\max}(t)$. Then $\kappa_{1i}(t) + \kappa_{1j}(t) \geq \kappa_{ij}(t) \geq \kappa_{\max}(0)/3 \Rightarrow \max\{\kappa_{1i}(t), \kappa_{1j}(t)\} \geq \kappa_{\max}(0)/6$. W.L.O.G., suppose $\kappa_{1i}(t) \geq \kappa_{\max}(0)/6$. If $\kappa_{1i}(t) \leq \pi/2$, then $\|\boldsymbol{z}_1(t) - \boldsymbol{z}_i(t)\| \geq \|\boldsymbol{z}_1(t)\| \sin \kappa_{1i}(t) \geq \Omega(\kappa_{\max}(0) \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|)$. If $\kappa_{1i}(t) > \pi/2$, then $\|\boldsymbol{z}_1(t) - \boldsymbol{z}_i(t)\| \geq \|\boldsymbol{z}_1(t)\| \geq \Omega(\kappa_{\max}(0) \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|)$. So no matter which case happens, (65) always holds.

In conclusion, we have $Z(t) \geq \max_{i,j\in[n]} \|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\| \geq \Omega(\kappa_{\max}(0) \max_{i\in[n]} \|\boldsymbol{z}_i(t)\|)$. ■

Combined with Lemma 29, Lemma 31 immediately implies the following corollary.

Corollary 32 *Given that the initial condition in Lemma 3 holds, suppose $\boldsymbol{z}_i(0) \neq 0, \forall i \in [n]$, $\kappa_{\max}(0) > 0$ and $n \geq 2$, then for $\forall t \in \mathbb{R}_{\geq 0}$ we have $Z(t) > 0$.*

4. Generally $\kappa_{\max}$ may not be continuous since the range of taking the max ($i, j \in Q^+$) might change, but it is continuous on $(t^* - \epsilon_1, t^* + \epsilon_1)$ since all $\boldsymbol{z}_i$'s are nonzero on this interval.

Lemma 33 *Suppose the conditions (58) (59) in Theorem 27 holds. Suppose the network is over-parameterized, i.e., $n \geq 2$, and the initialization is non-degenerate. Then $\forall t \geq T_2$ we have*

$$\frac{\partial}{\partial t} Z(t) \geq -O(n^2 \|\boldsymbol{v}\| \theta_{\max}^2(t)).$$

Proof

Recall the closed form expression of gradient (4), which can be decomposed into two terms,

$$\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}_i} = \underbrace{\frac{1}{2}\left(\sum_j \boldsymbol{w}_j - \boldsymbol{v}\right)}_{I_1} + \underbrace{\frac{1}{2\pi}\left[\left(\sum_{j\neq i} \|\boldsymbol{w}_j\| \sin\theta_{ij} - \|\boldsymbol{v}\| \sin\theta_i\right)\overline{\boldsymbol{w}}_i - \sum_{j\neq i}\theta_{ij}\boldsymbol{w}_j + \theta_i \boldsymbol{v}\right]}_{I_2}.$$

The second term I_2 can be rewritten as

$$I_2 = \frac{1}{2\pi}\left[\sum_{j\neq i}\|\boldsymbol{w}_j\|(\sin\theta_{ij} - \theta_{ij})\overline{\boldsymbol{w}}_i + \sum_{j\neq i}\|\boldsymbol{w}_j\|\theta_{ij}(\overline{\boldsymbol{w}}_i - \overline{\boldsymbol{w}}_j) - \|\boldsymbol{v}\|(\sin\theta_i - \theta_i)\overline{\boldsymbol{w}}_i - \|\boldsymbol{v}\|\theta_i(\overline{\boldsymbol{w}}_i - \overline{\boldsymbol{v}})\right]$$

By Theorem 27, for $\forall t > T_2$, we have $\|\boldsymbol{w}_i(t)\| = \Theta(\|\boldsymbol{v}\|/n)$. Note that $\sin\theta_{ij}(t) - \theta_{ij}(t) = O(\theta_{ij}^3(t)) = O(\theta_{\max}^3(t))$, similarly $\sin\theta_i(t) - \theta_i(t) = O(\theta_{\max}^3(t))$. Combined with $\|\overline{\boldsymbol{w}}_i - \overline{\boldsymbol{w}}_j\| = 2\sin(\theta_{ij}/2) = O(\theta_{\max})$, we get

$$\begin{aligned}\|I_2\| \leq &\sum_{j\neq i}\Theta(\|\boldsymbol{v}\|/n)O(\theta_{\max}^3) + \sum_{j\neq i}\Theta(\|\boldsymbol{v}\|/n)\theta_{\max}O(\theta_{\max}) \\ &+\|\boldsymbol{v}\|O(\theta_{\max}^3) + \|\boldsymbol{v}\|\theta_{\max}O(\theta_{\max}) = O(\|\boldsymbol{v}\|\theta_{\max}^2).\end{aligned}$$

Then $\forall i$, $\frac{\partial \boldsymbol{w}_i}{\partial t} = -\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}_i} = -\frac{1}{2}\left(\sum_j \boldsymbol{w}_j - \boldsymbol{v}\right) + \boldsymbol{O}(\|\boldsymbol{v}\|\theta_{\max}^2)$. Note that the first term $-\frac{1}{2}\left(\sum_j \boldsymbol{w}_j - \boldsymbol{v}\right)$ is the same for all $\boldsymbol{w}_i$, so for $\forall i, j \in [n]$ we have $\left\|\frac{\partial \boldsymbol{w}_i - \boldsymbol{w}_j}{\partial t}\right\| = \left\|\frac{\partial \boldsymbol{w}_i}{\partial t} - \frac{\partial \boldsymbol{w}_j}{\partial t}\right\| = O(\|\boldsymbol{v}\|\theta_{\max}^2)$.

For $\forall i, j \in [n]$, if $\boldsymbol{z}_i(t) = \boldsymbol{z}_j(t)$ then $\frac{\partial \boldsymbol{z}_i(t)}{\partial t} = \frac{\partial \boldsymbol{z}_j(t)}{\partial t} \Rightarrow \frac{\partial}{\partial t}\|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\| = 0$. Otherwise $\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t) \neq \boldsymbol{0}$ and

$$\begin{aligned}\frac{\partial}{\partial t}\left(\|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\|\right) &= \left\langle \frac{\partial}{\partial t}\left(\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\right), \overline{\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)}\right\rangle \geq -\left\|\frac{\partial}{\partial t}(\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t))\right\| \\ &\geq -\left\|\frac{\partial}{\partial t}(\boldsymbol{w}_i(t) - \boldsymbol{w}_j(t))\right\| \geq -O(\|\boldsymbol{v}\|\theta_{\max}^2(t)).\end{aligned}$$

So for both cases we have $\frac{\partial}{\partial t}\left(\|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\|\right) \geq -O(\|\boldsymbol{v}\|\theta_{\max}^2(t))$.
Then $\frac{\partial}{\partial t} Z(t) = \sum_{1\leq i<j\leq n}\frac{\partial}{\partial t}\left(\|\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)\|\right) \geq -O(n^2\|\boldsymbol{v}\|\theta_{\max}^2(t))$. ■

G.3. Proof of Main Theorem

Theorem 14 *Suppose the network is over-parameterized, i.e., $n \geq 2$. For $\forall \delta > 0$, if the initialization is non-degenerate, $d = \Omega(\log(n/\delta))$, $\sigma = O\left(n^{-4226} d^{-1/2} \|\boldsymbol{v}\|\right)$, then there exists $T_2 = O\left(\frac{\log n}{n}\right)$ such that with probability at least $1-\delta$, for $\forall t \geq T_2$ we have*

$$L(\boldsymbol{w}(t))^{-1/3} \leq O\left(\frac{n^{17/3}}{\kappa_{\max}^2(0)\|\boldsymbol{v}\|^{2/3}}\right)(t - T_2) + \gamma,$$

where $\gamma \in \mathbf{R}^+$ is a constant that does not depend on t.

Proof For $\forall t \geq T_2$ we have $\max_{i\in[n]} \|\boldsymbol{z}_i(t)\| \geq \theta_{\max}(t)\Theta(\|\boldsymbol{v}\|/n)$. Then by lemma 31, for $\forall t \geq T_2$, $Z(t) \geq \Omega(\kappa_{\max}(0)\theta_{\max}(t)\|\boldsymbol{v}\|/n) \Rightarrow \theta_{\max}(t) = O\left(\frac{nZ(t)}{\kappa_{\max}(0)\|\boldsymbol{v}\|}\right)$. Combined with lemma 33 we have

$$\frac{\partial}{\partial t} Z(t) \geq -O(n^2\|\boldsymbol{v}\|\theta_{\max}^2(t)) \geq -O\left(\frac{n^4 Z^2(t)}{\kappa_{\max}^2(0)\|\boldsymbol{v}\|}\right). \tag{66}$$

By Corollary 32, $Z(t)$ is always strictly positive. We can therefore calculate the dynamics of $1/Z(t)$ as: $\forall t \geq T_2$,

$$\frac{\partial}{\partial t}\frac{1}{Z(t)} = -\frac{1}{Z^2(t)}\frac{\partial}{\partial t} Z(t) \leq O\left(\frac{n^4}{\kappa_{\max}^2(0)\|\boldsymbol{v}\|}\right) \Rightarrow \frac{1}{Z(t)} = O\left(\frac{n^4}{\kappa_{\max}^2(0)\|\boldsymbol{v}\|}(t - T_2)\right) + \frac{1}{Z(T_2)}. \tag{67}$$

On the other hand, by Theorem 27 we have

$$Z(t) \leq \sum_{1\leq i<j\leq n} (\|\boldsymbol{z}_i(t)\| + \|\boldsymbol{z}_j(t)\|) \leq \sum_{1\leq i<j\leq n} (\theta_i(t)\|\boldsymbol{w}_i(t)\| + \theta_j(t)\|\boldsymbol{w}_j(t)\|) \leq O(n\|\boldsymbol{v}\|\theta_{\max}).$$

By lemma 20 we have

$$\forall t \geq T_2, \theta_{\max}(t) = O\left(\left(\frac{L(\boldsymbol{w}(t))n^2}{\|\boldsymbol{v}\|^2}\right)^{1/3}\right) \Rightarrow L(\boldsymbol{w}(t)) \geq \Omega\left(\frac{Z^3(t)}{n^5\|\boldsymbol{v}\|}\right). \tag{68}$$

Combined with (67), we have

$$L(\boldsymbol{w}(t))^{-1/3} \leq O\left(\frac{n^{17/3}}{\kappa_{\max}^2(0)\|\boldsymbol{v}\|^{2/3}}\right)(t - T_2) + \frac{n^{5/3}\|\boldsymbol{v}\|^{1/3}}{Z(T_2)}.$$

Finally, since Corollary 32 implies $Z(T_2) > 0$, setting $\gamma = \frac{n^{5/3}\|\boldsymbol{v}\|^{1/3}}{Z(T_2)}$ finishes the proof. ■