Phase Diagram of Initial Condensation for Two-layer Neural Networks

Zhengan Chen[1,2], Yuqing Li[1,2]*, Tao Luo[1,2,3,4,6], Zhangchen Zhou[1,5], Zhi-Qin John Xu[1,3,4]†

[1] School of Mathematical Sciences, Shanghai Jiao Tong University
[2] CMA-Shanghai, Shanghai Jiao Tong University
[3] Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University
[4] Qing Yuan Research Institute, Shanghai Jiao Tong University
[5] Zhiyuan College, Shanghai Jiao Tong University
[6] Shanghai Artificial Intelligence Laboratory
{zhengan_chen, liyuqing_551, luotao41, zczhou1115, xuzhiqin}@sjtu.edu.cn.

April 7, 2023

Abstract

The phenomenon of distinct behaviors exhibited by neural networks under varying scales of initialization remains an enigma in deep learning research. In this paper, based on the earlier work by Luo et al. [16], we present a phase diagram of initial condensation for two-layer neural networks. Condensation is a phenomenon wherein the weight vectors of neural networks concentrate on isolated orientations

*Corresponding author: liyuqing_551@sjtu.edu.cn
†Corresponding author: xuzhiqin@sjtu.edu.cn.

during the training process, and it is a feature in non-linear learning process that enables neural networks to possess better generalization abilities. Our phase diagram serves to provide a comprehensive understanding of the dynamical regimes of neural networks and their dependence on the choice of hyperparameters related to initialization. Furthermore, we demonstrate in detail the underlying mechanisms by which small initialization leads to condensation at the initial training stage.

Keywords— two-layer neural network, phase diagram, dynamical regime, condensation

1 Introduction

In deep learning, one intriguing observation is the distinct behaviors exhibited by Neural Networks (NNs) depending on the scale of initialization. Specifically, in a particular regime, NNs trained with gradient descent can be viewed as a kernel regression predictor known as the Neural Tangent Kernel (NTK) [11, 5, 10, 15], and Chizat et al. [4] identify it as the lazy training regime in which the parameters of overparameterized NNs trained with gradient based methods hardly varies. However, under a different scaling, the Gradient Flow (GF) of NN shows highly nonlinear features and a mean-field analysis [19, 23, 3, 24] has been established for infinitely wide two-layer networks to analyze its behavior. Additionally, small initialization is proven to give rise to condensation [18, 16, 31, 32], a phenomenon where the weight vectors of NNs concentrate on isolated orientations during the training process. This is significant as NNs with condensed weight vectors are equivalent to "smaller" NNs with fewer parameters, as revealed by the embedding principle (the loss landscape of a DNN "contains" all the critical points of all the narrower DNNs [30, 29]), thus reducing the complexity of the output functions of NNs. As the generalization error can be bounded in terms of the complexity [1], NNs with condensed parameters tend to possess better generalization abilities. In addition, the study of the embedding principle also found the number of the descent directions in a condensed large network is no less than that of the equivalent small effective network, which may lead to easier training of a large network [30, 29].

Taken together, identifying the regime of condensation and understanding the mechanism of condensation are important to understand the non-linear training of neural networks. Our contributions can be categorized into two aspects.

Firstly, we established the phase diagram of initial condensation for two-layer neural networks (NNs) with a wide class of smooth activation functions, as illustrated Figure 1. Note that the phase diagram drawn in [16] is only for two-layer

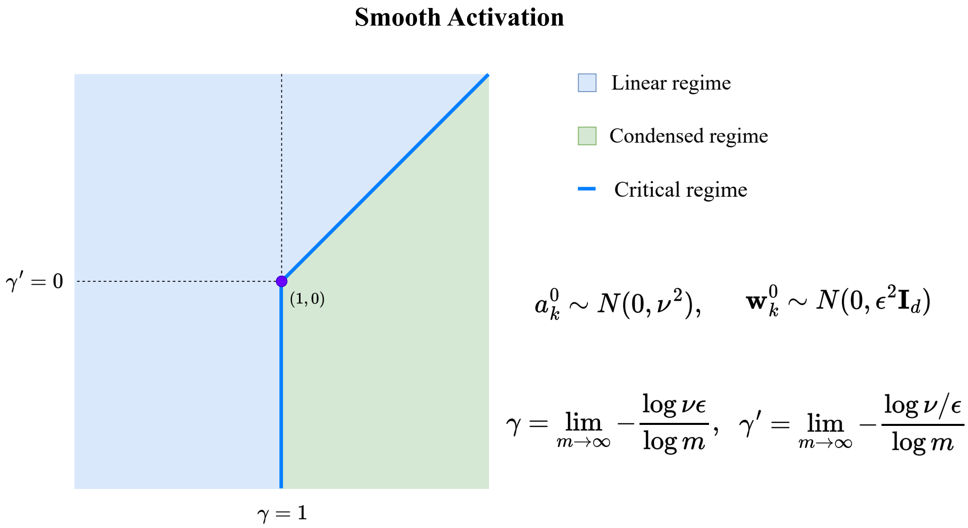


Figure 1: Phase diagram of two-layer NNs.

wide ReLU networks and the phase diagram in [32] is empirical for three-layer wide ReLU networks. The phase diagram of a two-layer neural network refers to a graphical representation of the dynamical behavior of the network as a function of its initialization scales. In this diagram, different regions correspond to different types of behaviors exhibited by NNs, such as the linear regime, where the network behaves like a linear model, and the condensed regime, where the network exhibits the initial condensation phenomenon.

Secondly, we reveal the mechanism of initial condensation for two-layer NNs and identify the directions towards which the weight parameters condense. There has been a flurry of recent papers endeavor to analyze the mechanism underlying the condensation of NNs at the initial training stage under small initialization [18, 21, 16, 17, 32]. For instance, Maennel et al. [18] uncovered that for two-layer ReLU NNs, the GF limits the weight vectors to a certain number of directions depending sorely on the input data. Zhou et al. [32] showed empirically that condensation is a common feature in non-linear training regime for three-layer ReLU NNs. Theoretically, Maennel et al. [18] argued that GF prefers "simple" functions over "complex" ones, and Zhou et. al. [31] demonstrated that the maximal number of condensed orientations at initial training stage is twice the multiplicity (Definition 1) of the activation function. However, these proofs are heuristic as they do not account for the dynamics of parameters. Pellegrini and Biroli [21] derived a mean-field model demonstrating that two-layer ReLU NNs, when trained with hinge loss and infinite data, lead to a linear classifier. Nonetheless, their analysis does not illustrate how the initial condensation depends on the scale of initialization and does not specify which directions NNs condense on.

The organization of the paper is listed as follows. In Section 2, we discuss some related works. In Section 3, we give some preliminary introduction to our problems. In Section 4, we state our main results and show some empirical evidence. In Section 5, we give out the outline of proofs for our main results, and conclusions are drawn in Section 6. All the details of the proof are deferred to the Appendix.

2 Related Works

There has been a rich literature on the choice of initialization schemes in order to facilitate neural network training [7, 9, 19, 24], and most of the work identified the width m as a hyperparameter, where the kernel regime is reached when the width grows towards infinity [11, 27, 6]. However, with the introduction of lazy training by Chizat et al. [4], instead of the width m, one shall take the initialization scale as the relevant hyperparameter. The lazy training refers to the phenomenon in which a heavily over-parameterized NN trained with gradient-based methods could converge exponentially fast to zero training loss with its parameters hardly varying, and such phenomenon can be observed in any non-convex model accompanied by the choice of an appropriate scaling factor of the initialization. Follow-up work by Woodworth et al. [26] focus on how the scale of initialization acts as a controlling quantity for the transition between two very different regimes, namely the kernel regime and the rich regime, for the matrix factorization problems. As for two-layer ReLU NNs, the phase digram in Luo et al. [16] identified three regimes, namely the linear regime, the critical regime and the condensed regime, based on the relative change of input weights as the width m approaches infinity. In summary, the selection of appropriate initialization scales plays a crucial role in the training of NNs.

Several theoretical works studying the dynamical behavior of NNs with small initialization can be connected to implicit regularization effect provided by the weight initialization schemes, and the condensation phenomenon has also been studied under different names. Ji and Telgarsky [12] analyzed the implicit regularization of GF on deep linear networks and observed the matrix alignment phenomena, i.e., weight matrices belonging to different layers share the same direction. The weight quantization effect [18] in training two-layer ReLU NNs with small initialization is really the condensation phenomenon in disguise, and so is the case for the weight cluster effect [2] in learning the MNIST task for a three-layer CNN. Luo et al. [16] focused on how the condensation phenomenon can be clearly detected by the choice of initialization schemes, but they did not show the reason behind it. Zhang et al. [28, 30] proposed a general Embedding Principle of loss landscape of DNNs, showing that a larger DNN can experience critical points with condensed parameter, and its output is the same as that of a much smaller DNN, but their analysis did not involve its dynamical behavior. Zhou et al. [31] presented a theory for the initial direction towards which the weight vector condenses, yet it is far from satisfactory.

3 Preliminaries

3.1 Notations

We begin this section by introducing some notations that will be used in the rest of this paper. We set n for the number of input samples and m for the width of the neural network. We set $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ as the normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ. We let $[n] = \{1, 2, \ldots, n\}$. We denote vector L^2 norm as $\|\cdot\|_2$, vector or function L_∞ norm as $\|\cdot\|_\infty$, matrix spectral (operator) norm as $\|\cdot\|_{2\to2}$, matrix infinity norm as $\|\cdot\|_{\infty\to\infty}$, and matrix Frobenius norm as $\|\cdot\|_{\mathrm{F}}$. For a matrix $\mathbf{A}$, we use $\mathbf{A}_{i,j}$ to denote its (i,j)-th entry. We will also use $\mathbf{A}_{i,:}$ to denote the i-th row vector of $\mathbf{A}$ and define $\mathbf{A}_{i,j:k} = [\mathbf{A}_{i,j}, \mathbf{A}_{i,j+1}, \cdots, \mathbf{A}_{i,k}]^\intercal$ as part of the vector. Similarly, $\mathbf{A}_{:,i}$ is the i-th column vector and $\mathbf{A}_{j:k,i}$ is a part of the i-th column vector. For a semi-positive-definite matrix $\boldsymbol{A}$, we denote its smallest eigenvalue by $\lambda_{\min}(\boldsymbol{A})$, and correspondingly, its largest eigenvalue by $\lambda_{\max}(\boldsymbol{A})$. We use $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ for the standard Big-O and Big-Omega notations. We finally denote the set of continuous functions $f(\cdot) : \mathbb{R} \to \mathbb{R}$ possessing continuous derivatives of order up to and including r by $\mathcal{C}^r(\mathbb{R})$, the set of analytic functions $f(\cdot) : \mathbb{R} \to \mathbb{R}$ by $\mathcal{C}^\omega(\mathbb{R})$, and $\langle\cdot,\cdot\rangle$ for standard inner product between two vectors.

3.2 Problem Setup

We use almost the same settings in Luo et al. [16] by starting with the original model

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}), \tag{3.1}$$

whose parameters $\boldsymbol{\theta}^0 := \mathrm{vec}(\boldsymbol{\theta}_a^0, \boldsymbol{\theta}_{\boldsymbol{w}}^0)$ are initialized by

$$a_k^0 \sim \mathcal{N}(0, \nu^2), \quad \boldsymbol{w}_k^0 \sim \mathcal{N}(\mathbf{0}, \varepsilon^2 \boldsymbol{I}_d), \tag{3.2}$$

and the empirical risk is

$$R_S(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2. \tag{3.3}$$

Then the training dynamics based on gradient descent (GD) at the continuous limit obeys the following gradient flow precisely reads: For $k \in [m]$,

$$\begin{aligned}
\frac{\mathrm{d}a_k}{\mathrm{d}t} &= -\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k'=1}^{m} a_{k'} \sigma(\boldsymbol{w}_{k'}^\intercal \boldsymbol{x}_i) - y_i \right) \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i), \\
\frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &= -\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k'=1}^{m} a_{k'} \sigma(\boldsymbol{w}_{k'}^\intercal \boldsymbol{x}_i) - y_i \right) a_k \sigma^{(1)}(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \boldsymbol{x}_i.
\end{aligned} \tag{3.4}$$

We identify the parameters $\boldsymbol{\theta}_a := \mathrm{vec}(\{a_k\}_{k=1}^m)$ and $\boldsymbol{\theta}_{\boldsymbol{w}} := \mathrm{vec}(\{\boldsymbol{w}_k\}_{k=1}^m)$ as variables of order one by setting

$$a_k = \nu \bar{a}_k, \quad \boldsymbol{w}_k = \varepsilon \bar{\boldsymbol{w}}_k,$$

then the rescaled dynamics can be written as

$$\begin{aligned}
\nu \frac{\mathrm{d}\bar{a}_k}{\mathrm{d}t} &= -\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k'=1}^{m} \nu \varepsilon \bar{a}_{k'} \frac{\sigma(\varepsilon \bar{\boldsymbol{w}}_{k'}^\intercal \boldsymbol{x}_i)}{\varepsilon} - y_i \right) \varepsilon \frac{\sigma(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}, \\
\varepsilon \frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}t} &= -\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k'=1}^{m} \nu \varepsilon \bar{a}_{k'} \frac{\sigma(\varepsilon \bar{\boldsymbol{w}}_{k'}^\intercal \boldsymbol{x}_i)}{\varepsilon} - y_i \right) \nu \bar{a}_k \sigma^{(1)}(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i) \boldsymbol{x}_i.
\end{aligned} \tag{3.5}$$

For the case where $\varepsilon \ll 1$ and $\varepsilon \gg 1$, the expressions $\frac{\sigma(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}$ and $\sigma^{(1)}(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)$ are hard to handle at first glance. However, in the case where $\varepsilon \ll 1$, under the condition (Assumption 1) that $\sigma(0) = 0$ and $\sigma^{(1)}(0) = 1$, we obtain that

$$\frac{\sigma(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} \approx \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i, \quad \sigma^{(1)}(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i) \approx 1,$$

hence $\frac{\sigma(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}$ and $\sigma^{(1)}(\varepsilon \bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)$ are of order one.

In the case where $\varepsilon \gg 1$, under the condition (Assumption 2) that

$$\lim_{x \to -\infty} \sigma^{(1)}(x) = a, \quad \lim_{x \to +\infty} \sigma^{(1)}(x) = b,$$

we obtain that

$$\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)}{\varepsilon}\approx\sigma^{(1)}(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i),$$

hence $\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)}{\varepsilon}$ and $\sigma^{(1)}(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)$ are also of order one. Under these two aforementioned conditions, $\sigma(\cdot)$ acts like a linear activation in the case where $\varepsilon\ll 1$, and acts like a leaky-ReLU activation activation in the case where $\varepsilon\gg 1$, both of which are homogeneous functions. Hence the above dynamics can be simplified into

$$\begin{aligned}\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^n\left(\sum_{k'=1}^m\nu\varepsilon\bar{a}_{k'}\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)}{\varepsilon}-y_i\right)\frac{\varepsilon}{\nu}\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)}{\varepsilon},\\ \frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^n\left(\sum_{k'=1}^m\nu\varepsilon\bar{a}_{k'}\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)}{\varepsilon}-y_i\right)\frac{\nu}{\varepsilon}\bar{a}_k\sigma^{(1)}(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i.\end{aligned}\tag{3.6}$$

We hereby introduce two scaling parameters

$$\kappa:=\nu\varepsilon,\quad \kappa':=\frac{\nu}{\varepsilon},\tag{3.7}$$

then the dynamics (3.6) can be written as a *normalized* flow

$$\begin{aligned}\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^n\left(\sum_{k'=1}^m\kappa\bar{a}_{k'}\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)}{\varepsilon}-y_i\right)\frac{1}{\kappa'}\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)}{\varepsilon},\\ \frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^n\left(\sum_{k'=1}^m\kappa\bar{a}_{k'}\frac{\sigma(\varepsilon\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)}{\varepsilon}-y_i\right)\kappa'\bar{a}_k\sigma^{(1)}(\varepsilon\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i.\end{aligned}\tag{3.8}$$

with the following initialization

$$\bar{a}_k^0\sim\mathcal{N}(0,1),\quad \bar{\boldsymbol{w}}_k^0\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d).\tag{3.9}$$

In the following discussion throughout this paper, we always refer to this rescaled model (3.8) and drop all the "bar"s of $\{a_k\}_{k=1}^m$ and $\{\boldsymbol{w}_k\}_{k=1}^m$ for notational simplicity.

As κ and κ' are always in specific power-law relations to the width m, we introduce two independent coordinates

$$\gamma:=\lim_{m\to\infty}-\frac{\log\kappa}{\log m},\quad \gamma':=\lim_{m\to\infty}-\frac{\log\kappa'}{\log m},\tag{3.10}$$

which meet all the guiding principles [16] for finding the coordinates of a phase diagram.

Before we end this section, we list out some commonly-used initialization methods with their scaling parameters as shown in Table 1.

Name	ν	ε	κ $(\nu\varepsilon)$	κ' (ν/ε)	γ	γ'
LeCun [14]	$\sqrt{\frac{1}{m}}$	$\sqrt{\frac{1}{d}}$	$\sqrt{\frac{1}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
He [8]	$\sqrt{\frac{2}{m}}$	$\sqrt{\frac{2}{d}}$	$\sqrt{\frac{4}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
Xavier [7]	$\sqrt{\frac{2}{m+1}}$	$\sqrt{\frac{2}{m+d}}$	$\sqrt{\frac{4}{(m+1)(d+1)}}$	$\sqrt{\frac{m+d}{m+1}}$	1	0
Huang [10]	1	$\sqrt{\frac{1}{m}}$	$\sqrt{\frac{1}{m}}$	$\sqrt{m}$	$\frac{1}{2}$	$-\frac{1}{2}$

Table 1: Initialization methods with their scaling parameters

4 Main Results

4.1 Activation function and input

In this part, we shall impose some technical conditions on the activation function and input samples. We start with a technical condition [31, Definition 1] on the activation function $\sigma(\cdot)$

Definition 1 (Multiplicity p). *$\sigma(\cdot):\mathbb{R}\to\mathbb{R}$ has multiplicity p if there exists an integer $p\geq 1$, such that for all $0\leq s\leq p-1$, the s-th order derivative satisfies $\sigma^{(s)}(0)=0$, and $\sigma^{(p)}(0)\neq 0$.*

We list out some examples of activation functions with different multiplicity.

Remark 1.

- *$tanh(x):=\frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$ is with multiplicity $p=1$;*
- *$SiLU(x):=\frac{x}{1+\exp(-x)}$ is with multiplicity $p=1$;*
- *$xtanh(x):=\frac{x\exp(x)-x\exp(-x)}{\exp(x)+\exp(-x)}$ is with multiplicity $p=2$.*

Assumption 1 (Multiplicity 1). *The activation function $\sigma\in\mathcal{C}^2(\mathbb{R})$, and there exists a universal constant $C_L>0$, such that its first and second derivatives satisfy*

$$\left\|\sigma^{(1)}(\cdot)\right\|_{\infty}\leq C_L,\quad \left\|\sigma^{(2)}(\cdot)\right\|_{\infty}\leq C_L. \tag{4.1}$$

Moreover,

$$\sigma(0)=0,\quad \sigma^{(1)}(0)=1. \tag{4.2}$$

Remark 2. *We remark that σ has multiplicity 1. $\sigma^{(1)}(0) = 1$ can be replaced by $\sigma^{(1)}(0) \neq 0$, and we set $\sigma^{(1)}(0) = 1$ for simplicity, and it can be easily satisfied by replacing the original activation $\sigma(\cdot)$ with $\frac{\sigma(\cdot)}{\sigma^{(1)}(0)}$.*

We note that Assumption 1 can be satisfied by using the tanh activation:

$$\sigma(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)},$$

and the scaled SiLU activation

$$\sigma(x) = \frac{2x}{1 + \exp(-x)}.$$

Assumption 2. *The activation function $\sigma \in \mathcal{C}^{\omega}(\mathbb{R})$ and is not a polynomial function, also its function value at 0 satisfy*

$$\sigma(0) = 0, \tag{4.3}$$

also there exists a universal constant $C_L > 0$, such that its first and second derivatives satisfy

$$\sigma^{(1)}(0) = 1, \quad \left\|\sigma^{(1)}(\cdot)\right\|_{\infty} \leq C_L, \quad \left\|\sigma^{(2)}(\cdot)\right\|_{\infty} \leq C_L. \tag{4.4}$$

Moreover,

$$\lim_{x \to -\infty} \sigma^{(1)}(x) = a, \quad \lim_{x \to +\infty} \sigma^{(1)}(x) = b, \tag{4.5}$$

and $a \neq b$.

Remark 3. *We note that Assumption 2 can be satisfied by by using the scaled SiLU activation:*

$$\sigma(x) = \frac{2x}{1 + \exp(-x)},$$

where $a = 0$ and $b = 2$.

Some other functions also satisfy this assumption, for instance, the modified scaled softplus activation:

$$\sigma(x) = 2\left(\log(1 + \exp(x)) - \log 2\right),$$

where $a = 0$ and $b = 2$.

Assumption 3 (Non-degenerate data)**.** *The training inputs and labels $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ satisfy that there exists a universal constant $c > 0$, such that for all $i \in [n]$,*

$$\frac{1}{c} \leq \|\boldsymbol{x}_i\|_2, \quad |y_i| \leq c,$$

and

$$\sum_{i=1}^{n} y_i \boldsymbol{x}_i \neq \mathbf{0}. \tag{4.6}$$

We denote by

$$\boldsymbol{z} := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i, \tag{4.7}$$

and assume further that for some universal constant $c > 0$, the following holds

$$\frac{1}{c} \leq \|\boldsymbol{z}\|_2 \leq c, \tag{4.8}$$

and its unit vector

$$\hat{\boldsymbol{z}} := \frac{\sum_{i=1}^{n} y_i \boldsymbol{x}_i}{\|\sum_{i=1}^{n} y_i \boldsymbol{x}_i\|_2}. \tag{4.9}$$

Assumption 4. *The training inputs and labels $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ satisfy that there exists a universal constant $c > 0$, such that for all $i \in [n]$,*

$$\frac{1}{c} \leq \|\boldsymbol{x}_i\|_2, \quad |y_i| \leq c,$$

and all training inputs are non-parallel with each other, i.e., for any $i \neq j$ and $i, j \in [n]$,

$$\boldsymbol{x}_i \nparallel \boldsymbol{x}_j.$$

We remark that the requirements in Assumption 3 are easier to meet compared with Assumption 4, and both assumptions impose the input sample $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ to be of order one.

Assumption 5. *The following limit exists*

$$\gamma_1 := \lim_{m\to\infty} -\frac{\log \nu}{\log m}, \quad \gamma_2 := \lim_{m\to\infty} -\frac{\log \varepsilon}{\log m}, \tag{4.10}$$

then by definition

$$\gamma = \lim_{m\to\infty} -\frac{\log \nu\varepsilon}{\log m} = \gamma_1 + \gamma_2, \quad \gamma' = \lim_{m\to\infty} -\frac{\log \frac{\nu}{\varepsilon}}{\log m} = \gamma_1 - \gamma_2.$$

4.2 Regime Characterization at Initial Stage

Before presenting our theory that establishes a consistent boundary to separate the diagram into two distinct areas, namely the linear regime area and the condensed regime area, we introduce a quantity that has proven to be valuable in the analysis of NNs.

It is known that the output of a two-layer NN is linear with respect to $\boldsymbol{\theta}_a$, hence if the set of parameter $\boldsymbol{\theta}_{\boldsymbol{w}}$ remain stuck to its initialization throughout the whole training process, then the training dynamics of a two-layer NN can be linearized around the initialization. In the phase diagram, the linear regime area precisely corresponds to the region where the output function of a two-layer NN can be well approximated by its linearized model, i.e., in the linear regime area, the following holds

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx f\left(\boldsymbol{x}, \boldsymbol{\theta}(0)\right) &+ \left\langle \nabla_{\boldsymbol{\theta}_a} f\left(\boldsymbol{x}, \boldsymbol{\theta}(0)\right), \boldsymbol{\theta}_a(t) - \boldsymbol{\theta}_a(0) \right\rangle \\ &+ \left\langle \nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}} f\left(\boldsymbol{x}, \boldsymbol{\theta}(0)\right), \boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0) \right\rangle. \end{aligned} \tag{4.11}$$

In general, this linear approximation holds only when $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ remains within a small neighbourhood of $\boldsymbol{\theta}_{\boldsymbol{w}}(0)$. Since the size of this neighbourhood scales with $\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2$, therefore we use the following quantity as an indicator of how far $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ deviates away from $\boldsymbol{\theta}_{\boldsymbol{w}}(0)$ throughout the training process

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}. \tag{4.12}$$

We demonstrate that as $m \to \infty$, under suitable choice of the initialization scales (the blue area in Figure 1), the NN training dynamics fall into the linear regime (Theorem 1), and for large enough m,

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to 0.$$

We also demonstrate that under some other choices of the initialization scales (the green area in Figure 1), the NN training dynamics fall into the condensed regime (Theorem 2), where

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to +\infty,$$

and the phenomenon of condensation can be observed, and $\boldsymbol{\theta}_{\boldsymbol{w}}$ condense toward the direction of $\boldsymbol{z}$. We observe that in both cases, as $\boldsymbol{\theta}_{\boldsymbol{w}}$ deviates far away from initialization, the approximation (4.11) fails, and NN training dynamics is essentially nonlinear with respect to $\boldsymbol{\theta}_{\boldsymbol{w}}$.

Moreover, under the remaining choice of the initialization scales (the solid blue line in Figure 1), the NN training dynamics fall into the critical regime area, and we conjecture that

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to \mathcal{O}(1),$$

whose study is beyond the scope of this paper.

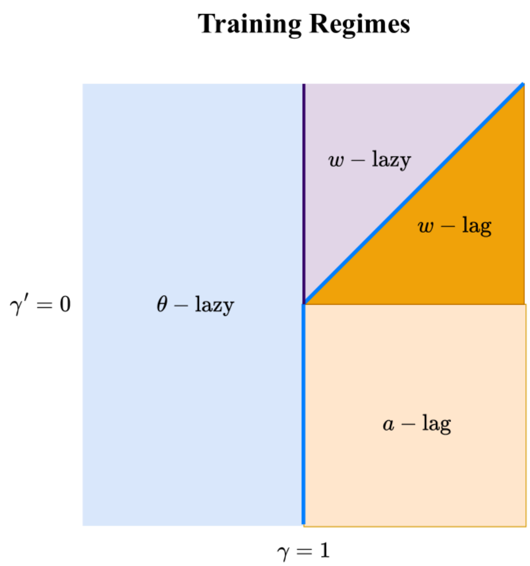


Figure 2: Different training regimes in the phase diagram.

Theorem 1 (Linear regime)**.** *Given any $\delta \in (0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if $\gamma < 1$ or $\gamma' > \gamma - 1$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,*

$$\lim_{m\to\infty} \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = 0. \tag{4.13}$$

Remark 4. *The linear regime area is split into two parts, one is termed the $\boldsymbol{\theta}$-lazy area (blue area in Figure 2), where $\gamma < 1$, the other is termed the $\boldsymbol{w}$-lazy area (pink area in Figure 2), where $\gamma \geq 1$ and $\gamma' > \gamma - 1 > 0$.*

In the $\boldsymbol{\theta}$-lazy area, the following relation holds

$$\lim_{m\to\infty} \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2} = 0, \tag{4.14}$$

whose detailed reasoning can be found in Appendix B.2, and in the $\boldsymbol{w}$-lazy area, relation (4.14) does not hold.

Theorem 2 (Condensed regime)**.** *Given any $\delta \in (0,1)$, under Assumption 1, Assumption 3 and Assumption 5, if $\gamma > 1$ and $\gamma' < \gamma - 1$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, there exists $T > 0$, such that*

$$\lim_{m\to\infty} \sup_{t\in[0,T]} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = +\infty, \tag{4.15}$$

and

$$\lim_{m\to\infty} \sup_{t\in[0,T]} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2} = 1, \tag{4.16}$$

where $\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t) := [\langle \boldsymbol{w}_1, \hat{\boldsymbol{z}}\rangle, \langle \boldsymbol{w}_2, \hat{\boldsymbol{z}}\rangle, \cdots, \langle \boldsymbol{w}_m, \hat{\boldsymbol{z}}\rangle]^\intercal$.

Remark 5. *The condensed regime area is split into two parts, one is termed the $\boldsymbol{w}$-lag area (orange area in Figure 2), where $\gamma > 1$ and $0 \leq \gamma' < \gamma - 1$, the other is termed the a-lag area (yellow area in Figure 2), where $\gamma > 1$ and $\gamma' < 0$.*

In the $\boldsymbol{w}$-lag regime area, as illustrated in (5.14)*, $\boldsymbol{\theta}_{\boldsymbol{w}}$ waits for a period of time of order one until $\boldsymbol{\theta}_a$ attains a magnitude that is commensurate with that of $\boldsymbol{\theta}_{\boldsymbol{w}}$, and the time T in Theorem 2 satisfies that*

$$T \geq \log\left(\frac{1}{4}\right) + \frac{\gamma - \gamma' - 1}{8}\log(m), \tag{4.17}$$

and as $m \to \infty$, $T \to \infty$.

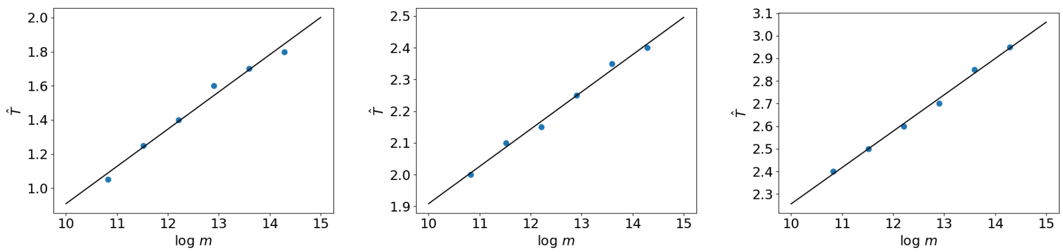
In the a-lag regime area, as illustrated in (5.14)*, $\boldsymbol{\theta}_a$ waits for a period of time of order one until $\boldsymbol{\theta}_{\boldsymbol{w}}$ attains a magnitude that is commensurate with that of $\boldsymbol{\theta}_a$, and it is exactly during this interval of time that the phenomenon of initial condensation can be observed. Hence for some $\alpha > 0$, the time T in Theorem 2 can be chosen as*

$$m^{-\alpha} \leq T \leq 2m^{-\alpha}, \tag{4.18}$$

which is obviously of order one, see Appendix C.4 for more details.

4.3 Experimental Demonstration

In order to distinguish between the $\boldsymbol{w}$-lag regime and a-lag regime, it is necessary to estimate the time T in Remark 5, which is also a reasonale way to empirically validate our theoretical analysis. An empirical approximation of T can be obtained by determining the time interval $\hat{T}$, starting from the initial stage at $t = 0$, up to the point at which the quantity $\frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2}$ reaches its climax for sufficiently large values of m ($m = 50000, 100000, 200000, 400000, 800000, 1600000$), as we are unable to run experiments at $m \to \infty$.

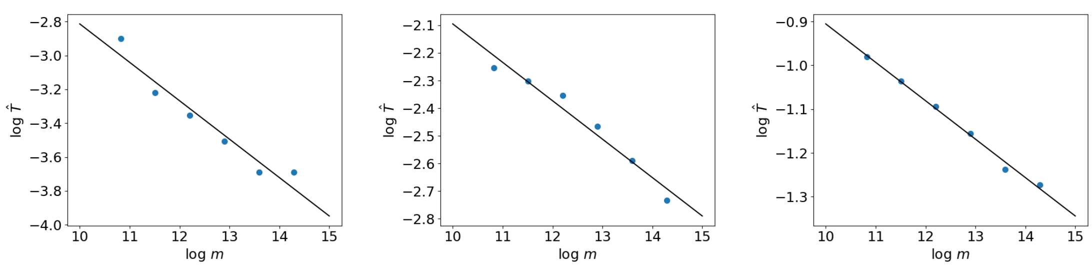


(a) $\gamma = 1.4, \quad R^2 = 0.983$ (b) $\gamma = 1.6, \quad R^2 = 0.991$ (c) $\gamma = 1.8, \quad R^2 = 0.995$

Figure 3: $\hat{T}$ (ordinate) vs $\log m$ (abscissa) with different values of γ but fixed $\gamma' = 0$ for two-layer NNs with tanh activation indicated by blue dots. The black line is a linear fit, and R^2 is *the coefficient of determination* that provides information about the goodness of fit of a linear regression. The closer R^2 is to 1, the better the model fits the data.

4.3.1 $\boldsymbol{w}$-lag regime

We validate the effectiveness of our estimates by performing a simple linear regression to visualize the relation (4.17), where $\hat{T}$ is set as the response variable and $\log m$ as the single independent variable. Figure 3 shows that NNs with different values of γ but fixed γ' satisfy the relation (4.17), thereby demonstrating the accuracy and reliability of our estimates.



(a) $\gamma = 1.2, \quad R^2 = 0.936$ (b) $\gamma = 1.4, \quad R^2 = 0.977$ (c) $\gamma = 1.6, \quad R^2 = 0.993$

Figure 4: $\log \hat{T}$ (ordinate) vs $\log m$ (abscissa) with different values of γ but fixed $\gamma' = -0.4$ for two-layer NNs with tanh activation indicated by blue dots. The black line is a linear fit, and R^2 is the coefficient of determination.

4.3.2 a-lag regime

We repeat the strategy in Section 4.3.1 except that in one hand we are hereby to visualize the relation (4.18) in Figure 4, and in the other, $\log \hat{T}$ is set as the response variable and γ' is no longer 0. We can still see a good agreement between the experimental data and its linear fitting, thus, validating the relation (4.18).

5 Technique Overview

In this part, we describe some technical tools and present the sketch of proofs for the above two theorems. Before we proceed, a rigorous description of the updated notations and definitions is required.

We start by a two-layer normalized NN model

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{m} \nu\varepsilon a_k \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x})}{\varepsilon}, \tag{5.1}$$

with the normalized parameters $\boldsymbol{\theta}^0 := \mathrm{vec}(\boldsymbol{\theta}_a^0, \boldsymbol{\theta}_{\boldsymbol{w}}^0)$ initialized by

$$\begin{aligned} a_k^0 &:= a_k(0) \sim \mathcal{N}(0,1), \\ \boldsymbol{w}_k^0 &:= \boldsymbol{w}_k(0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d). \end{aligned}$$

For all $i \in [n]$, we denote hereafter that

$$e_i := e_i(\boldsymbol{\theta}) := f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i,$$

and

$$\boldsymbol{e} := \boldsymbol{e}(\boldsymbol{\theta}) := [e_1(\boldsymbol{\theta}), e_2(\boldsymbol{\theta}), \ldots, e_n(\boldsymbol{\theta})]^\intercal.$$

Then the normalized flow reads: For all $k \in [m]$,

$$\begin{aligned} \frac{\mathrm{d}a_k}{\mathrm{d}t} &= -\frac{\varepsilon}{\nu}\frac{1}{n}\sum_{i=1}^{n} e_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}, \\ \frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &= -\frac{\nu}{\varepsilon}\frac{1}{n}\sum_{i=1}^{n} e_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i. \end{aligned} \tag{5.2}$$

5.1 Linear Regime

We define the normalized kernels as follows

$$\begin{aligned} k^{[a]}(\boldsymbol{x}, \boldsymbol{x}') &:= \frac{1}{\varepsilon^2}\mathbb{E}_{\boldsymbol{w}}\sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x})\sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}'), \\ k^{[\boldsymbol{w}]}(\boldsymbol{x}, \boldsymbol{x}') &:= \mathbb{E}_{(a,\boldsymbol{w})} a^2 \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x})\sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}') \left\langle \boldsymbol{x}, \boldsymbol{x}' \right\rangle, \end{aligned} \tag{5.3}$$

thus, the components of the Gram matrices $\boldsymbol{K}^{[a]}$ and $\boldsymbol{K}^{[\boldsymbol{w}]}$ of at infinite width respectively reads: For any $i, j \in [n]$,

$$\begin{aligned} \boldsymbol{K}^{[a]} &:= \left[K_{ij}^{[a]}\right]_{n\times n}, \\ K_{ij}^{[a]} &:= k^{[a]}(\boldsymbol{x}_i, \boldsymbol{x}_j), \\ \boldsymbol{K}^{[\boldsymbol{w}]} &:= \left[K_{ij}^{[\boldsymbol{w}]}\right]_{n\times n}, \\ K_{ij}^{[\boldsymbol{w}]} &:= k^{[\boldsymbol{w}]}(\boldsymbol{x}_i, \boldsymbol{x}_j), \end{aligned} \tag{5.4}$$

we conclude that under Assumption 2 and Assumption 4, $\boldsymbol{K}^{[a]}$ and $\boldsymbol{K}^{[\boldsymbol{w}]}$ are strictly positive definite, and both of their least eigenvalues are of order one (Theorem 4).

We define the normalized Gram matrices $\boldsymbol{G}^{[a]}(\boldsymbol{\theta})$, $\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta})$, and $\boldsymbol{G}(\boldsymbol{\theta})$ for a finite width two-layer network as follows: For any $i, j \in [n]$,

$$\begin{aligned} \boldsymbol{G}^{[a]}(\boldsymbol{\theta}) &:= \left[G_{ij}^{[a]}(\boldsymbol{\theta})\right]_{n\times n}, \\ G_{ij}^{[a]}(\boldsymbol{\theta}) &:= \frac{1}{m}\sum_{k=1}^m \left\langle \nabla_{a_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \frac{\varepsilon}{\nu}\nabla_{a_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\right\rangle \\ &= \frac{\nu^2}{m}\frac{\varepsilon}{\nu}\sum_{k=1}^m \sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_j) \\ &= \frac{\nu\varepsilon^3}{m}\sum_{k=1}^m \frac{1}{\varepsilon^2}\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_j), \\ \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}) &:= \left[G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta})\right]_{n\times n}, \\ G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}) &:= \frac{1}{m}\sum_{k=1}^m \left\langle \nabla_{\boldsymbol{w}_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \frac{\nu}{\varepsilon}\nabla_{\boldsymbol{w}_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\right\rangle \\ &= \frac{\nu^3\varepsilon}{m}\sum_{k=1}^m a_k^2\sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_j)\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j\right\rangle, \end{aligned} \tag{5.5}$$

and

$$\boldsymbol{G}(\boldsymbol{\theta}) := \boldsymbol{G}^{[a]}(\boldsymbol{\theta}) + \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}). \tag{5.6}$$

Remark 6. *We conclude that*

$$\lambda_{\min}\left(\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)\right) \sim \Omega(\nu\varepsilon^3), \quad \lambda_{\min}\left(\boldsymbol{G}^{[\boldsymbol{w}]}\left(\boldsymbol{\theta}^0\right)\right) \sim \Omega(\nu^3\varepsilon), \tag{5.7}$$

and it has been rigorously achieved by Proposition 2 located in Appendix B.1.

Finally, we obtain that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R_S(\boldsymbol{\theta}) &= -\left(\sum_{k=1}^{m}\frac{\varepsilon}{\nu}\left\langle \nabla_{a_k}R_S(\boldsymbol{\theta}), \nabla_{a_k}R_S(\boldsymbol{\theta})\right\rangle + \sum_{k=1}^{m}\frac{\nu}{\varepsilon}\left\langle \nabla_{\boldsymbol{w}_k}R_S(\boldsymbol{\theta}), \nabla_{\boldsymbol{w}_k}R_S(\boldsymbol{\theta})\right\rangle\right)\\ &= -\frac{m}{n^2}\boldsymbol{e}^{\intercal}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}) + \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta})\right)\boldsymbol{e}.\end{aligned}$$

In the case where $\gamma < 1$ ($\boldsymbol{\theta}$-lazy regime), the following holds for all $t > 0$:

$$\lambda_{\min}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))\right) \geq \frac{1}{2}\nu\varepsilon^3\lambda, \quad \lambda_{\min}\left(\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))\right) \geq \frac{1}{2}\nu^3\varepsilon\lambda,$$

for some universal constant $\lambda > 0$. Hence, we obtain that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R_S(\boldsymbol{\theta}(t)) &= -\frac{m}{n^2}\boldsymbol{e}^{\intercal}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t)) + \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))\right)\boldsymbol{e}\\ &\leq -\frac{2m}{n}\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta}(t))\right)R_S(\boldsymbol{\theta}(t))\\ &\leq -\frac{m}{n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)R_S(\boldsymbol{\theta}(t)),\end{aligned}$$

then

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m}{n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)t\right)R_S(\boldsymbol{\theta}(0)). \tag{5.8}$$

The following relation

$$\lim_{m\to\infty}\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2} = 0, \tag{5.9}$$

is illustrated through an intuitive scaling analysis. Since

$$\frac{\mathrm{d}}{\mathrm{d}t}R_S(\boldsymbol{\theta}(t)) = -\frac{\varepsilon}{\nu}\left\|\nabla_{\boldsymbol{\theta}_a}R_S(\boldsymbol{\theta}(t))\right\|_2^2 - \frac{\nu}{\varepsilon}\left\|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}}R_S(\boldsymbol{\theta}(t))\right\|_2^2 \sim -\frac{m}{n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)R_S(\boldsymbol{\theta}(t)),$$

then we have that

$$R_S(\boldsymbol{\theta}(t)) \sim \exp\left(-\frac{m}{n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)t\right)R_S(\boldsymbol{\theta}(0)),$$

and

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}_a}R_S(\boldsymbol{\theta}(t))\right\|_2 &\sim \sqrt{\frac{m}{n}\nu^3\varepsilon\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)}\sqrt{R_S(\boldsymbol{\theta}(t))}\\ &\sim \sqrt{\frac{m}{n}\nu^3\varepsilon\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)t\right)\sqrt{R_S(\boldsymbol{\theta}(0))},\\ \left\|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}}R_S(\boldsymbol{\theta}(t))\right\|_2 &\sim \sqrt{\frac{m}{n}\nu\varepsilon^3\left(\frac{\varepsilon}{\nu}\lambda + \frac{\nu}{\varepsilon}\lambda\right)}\sqrt{R_S(\boldsymbol{\theta}(t))}\end{aligned}$$

$$\sim \sqrt{\frac{m}{n}\nu\varepsilon^3\left(\frac{\varepsilon}{\nu}\lambda+\frac{\nu}{\varepsilon}\lambda\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda+\frac{\nu}{\varepsilon}\lambda\right)t\right)\sqrt{R_S(\boldsymbol{\theta}(0))},$$

both holds, hence

$$\begin{aligned}
\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2 &\leq \|\boldsymbol{\theta}_a(t)-\boldsymbol{\theta}_a(0)\|_2+\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2\\
&\leq \frac{\varepsilon}{\nu}\int_0^t \|\nabla_{\boldsymbol{\theta}_a}R_S(\boldsymbol{\theta}(s))\|_2\,\mathrm{d}s+\frac{\nu}{\varepsilon}\int_0^t \|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}}R_S(\boldsymbol{\theta}(s))\|_2\,\mathrm{d}s\\
&\leq \frac{\varepsilon}{\nu}\int_0^\infty \|\nabla_{\boldsymbol{\theta}_a}R_S(\boldsymbol{\theta}(s))\|_2\,\mathrm{d}s+\frac{\nu}{\varepsilon}\int_0^\infty \|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}}R_S(\boldsymbol{\theta}(s))\|_2\,\mathrm{d}s\\
&\lesssim \left(\sqrt{\frac{\varepsilon}{\nu}}+\sqrt{\frac{\nu}{\varepsilon}}\right)\sqrt{\frac{n}{m\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda+\frac{\nu}{\varepsilon}\lambda\right)}}\sqrt{R_S(\boldsymbol{\theta}(0))},
\end{aligned}$$

and

$$\|\boldsymbol{\theta}(0)\|_2\sim\sqrt{m},$$

hence

$$\begin{aligned}
\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2} &\lesssim \left(\sqrt{\frac{\varepsilon}{\nu}}+\sqrt{\frac{\nu}{\varepsilon}}\right)\sqrt{\frac{n}{m^2\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda+\frac{\nu}{\varepsilon}\lambda\right)}}\sqrt{R_S(\boldsymbol{\theta}(0))}\\
&\lesssim \sqrt{\frac{n}{m^2\nu^2\varepsilon^2}}\sqrt{R_S(\boldsymbol{\theta}(0))}.
\end{aligned}\tag{5.10}$$

The rigorous statements of relations (5.8) and (5.9) are given in Theorem 5.

In the case where $\gamma\geq 1$ and $\gamma'>\gamma-1$ ($\boldsymbol{w}$-lazy regime), the following holds for all $t>0$:

$$\lambda_{\min}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))\right)\geq\frac{1}{2}\nu\varepsilon^3\lambda,$$

for some universal constant $\lambda>0$. Hence, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}R_S(\boldsymbol{\theta}(t)) &= -\frac{m}{n^2}\boldsymbol{e}^\intercal\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))+\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))\right)\boldsymbol{e}\\
&\leq -\frac{2m}{n}\lambda_{\min}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))\right)R_S(\boldsymbol{\theta}(t))\\
&\leq -\frac{m}{n}\nu^2\varepsilon^2\frac{\varepsilon}{\nu}\lambda R_S(\boldsymbol{\theta}(t))\\
&= -\frac{m}{n}\nu\varepsilon^3\lambda R_S(\boldsymbol{\theta}(t)),
\end{aligned}$$

thus the following holds

$$R_S(\boldsymbol{\theta}(t))\leq\exp\left(-\frac{m\nu\varepsilon^3\lambda t}{n}\right)R_S(\boldsymbol{\theta}(0)),$$

and (5.9) does not hold anymore. However, we still have

$$\lim_{m\to\infty} \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = 0, \tag{5.11}$$

and it can also be illustrated through an intuitive scaling analysis. Since

$$\frac{\mathrm{d}}{\mathrm{d}t} R_S(\boldsymbol{\theta}(t)) = -\frac{\varepsilon}{\nu}\left\|\nabla_{\boldsymbol{\theta}_a} R_S(\boldsymbol{\theta}(t))\right\|_2^2 - \frac{\nu}{\varepsilon}\left\|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}} R_S(\boldsymbol{\theta}(t))\right\|_2^2 \sim -\frac{m}{n}\nu\varepsilon^3\lambda R_S(\boldsymbol{\theta}(t)),$$

then

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}} R_S(\boldsymbol{\theta}(t))\right\|_2 &\sim \sqrt{\frac{m}{n}\varepsilon^4\lambda}\sqrt{R_S(\boldsymbol{\theta}(t))}\\ &\sim \sqrt{\frac{m}{n}\varepsilon^4\lambda}\exp\left(-\frac{m\nu\varepsilon^3\lambda t}{2n}\right)\sqrt{R_S(\boldsymbol{\theta}(0))},\end{aligned}$$

hence

$$\begin{aligned}\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2 &\leq \frac{\nu}{\varepsilon}\int_0^t \left\|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}} R_S(\boldsymbol{\theta}(s))\right\|_2 \mathrm{d}s\\ &\leq \frac{\nu}{\varepsilon}\int_0^\infty \left\|\nabla_{\boldsymbol{\theta}_{\boldsymbol{w}}} R_S(\boldsymbol{\theta}(s))\right\|_2 \mathrm{d}s\\ &\lesssim \sqrt{\frac{n}{m\varepsilon^4\lambda}}\sqrt{R_S(\boldsymbol{\theta}(0))},\end{aligned}$$

and as

$$\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2 \sim \sqrt{m},$$

then

$$\begin{aligned}\frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} &\lesssim \sqrt{\frac{n}{m^2\varepsilon^4\lambda}}\sqrt{R_S(\boldsymbol{\theta}(0))}\\ &\lesssim \sqrt{\frac{n}{m^2\varepsilon^4}}\sqrt{R_S(\boldsymbol{\theta}(0))}.\end{aligned} \tag{5.12}$$

The rigorous statements of relation (5.11) are given in Theorem 6. To end this part, we provide a sketch of the proofs for Theorem 1, see Figure 5.

5.2 Condensed Regime

We remark that the $\{a_k, \boldsymbol{w}_k\}_{k=1}^m$ dynamics

$$\begin{aligned}\frac{\mathrm{d}a_k}{\mathrm{d}t} &= -\frac{\varepsilon}{\nu}\frac{1}{n}\sum_{i=1}^n e_i \frac{\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)}{\varepsilon},\\ \frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &= -\frac{\nu}{\varepsilon}\frac{1}{n}\sum_{i=1}^n e_i a_k \sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i,\end{aligned} \tag{5.13}$$

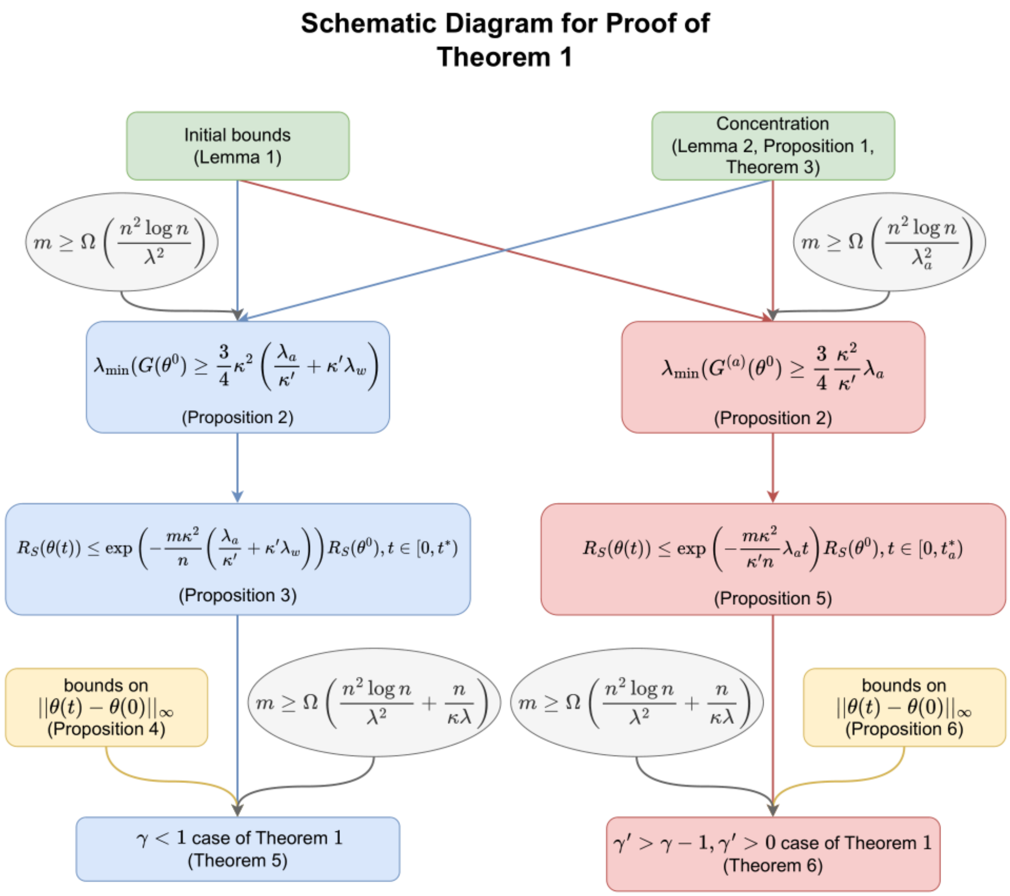


Figure 5: Sketch of proof for Theorem 1.

is coupled in the sense that the solution of at least one of the equations in the system depends on knowing one of the other solutions in the system, and a coupled system is usually hard to solve.

However, in the condense regime, as $\varepsilon \ll 1$ and $\varepsilon\nu \ll \frac{1}{m}$, the evolution of $\{e_i\}_{i=1}^n$ is slow enough so that it remains close to $\{-y_i\}_{i=1}^n$ over a period of time $T > 0$ at the initial stage, hence (5.13) approximately reads

$$\begin{aligned}
\frac{\mathrm{d}a_k}{\mathrm{d}t} &\approx \frac{\varepsilon}{\nu}\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i = \frac{\varepsilon}{\nu}\boldsymbol{w}_k^\intercal \boldsymbol{z}, \\
\frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &\approx \frac{\nu}{\varepsilon}\frac{1}{n}\sum_{i=1}^n y_i a_k \sigma^{(1)}(0)\boldsymbol{x}_i = \frac{\nu}{\varepsilon}a_k \boldsymbol{z},
\end{aligned} \tag{5.14}$$

and the coupled dynamics is reduced to linear dynamics.

We are able to solve out the linear dynamics (5.14) (Proposition 7), whose solutions read: For each $k \in [m]$, under the initial condition $\left[\nu a_k(0), \varepsilon \boldsymbol{w}_k^\intercal(0)\right]^\intercal = \left[\nu a_k^0, (\varepsilon \boldsymbol{w}_k^0)^\intercal\right]^\intercal$, we obtain that

$$\begin{aligned}
\nu a_k(t) &= \nu\left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) + \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right) a_k^0 \\
&\quad + \varepsilon\left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) - \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right)\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle, \\
\varepsilon \boldsymbol{w}_k(t) &= \nu\left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) - \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right) a_k^0 \hat{\boldsymbol{z}} \\
&\quad + \varepsilon\left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) + \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right)\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}} \\
&\quad - \varepsilon\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}} + \varepsilon \boldsymbol{w}_k^0.
\end{aligned} \tag{5.15}$$

We remark that $\{a_k, \boldsymbol{w}_k\}_{k=1}^m$ are the normalized parameters, then $\{\nu a_k, \varepsilon \boldsymbol{w}_k\}_{k=1}^m$ corresponds to the original parameters in (3.4).

In the case where $\gamma > 1$ and $0 \leq \gamma' < \gamma - 1$ ($\boldsymbol{w}$-lag regime), as $\varepsilon \gg \nu$, then the magnitude of $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$ is much larger than that of $\{\nu a_k\}_{k=1}^m$ at $t = 0$. Based on (5.15), it can be observed that $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$ remains dormant until $\{\nu a_k\}_{k=1}^m$ attain a magnitude that is commensurate with that of $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$, and only then do the magnitudes of $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$ and $\{\nu a_k\}_{k=1}^m$ experience exponential growth simultaneously. In order for the initial condensation of $\boldsymbol{\theta}_{\boldsymbol{w}}$ to be observed, one has to wait for some growth in the magnitude of $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$, hence $T \sim \Omega(\log(m))$. More importantly, compared with the $\boldsymbol{w}$-lazy regime, the condition $\gamma' < \gamma - 1$ enforces $\varepsilon \ll \frac{1}{\sqrt{m}}$, thus providing enough room for $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$ to grow in the $\boldsymbol{z}$-direction before $\{e_i\}_{i=1}^n$ deviates away from $\{-y_i\}_{i=1}^m$.

In the case where $\gamma > 1$ and $\gamma' < 0$ (a-lag regime), as $\varepsilon \ll \nu$, then the initial magnitudes of $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$ is much smaller than that of $\{\nu a_k\}_{k=1}^m$. Based on (5.15), it takes only $T \sim \Omega(1)$ for $\{\varepsilon \boldsymbol{w}_k\}_{k=1}^m$ to attain a magnitude comparable to $\{\nu a_k\}_{k=1}^m$, and this rapid growth leads to the observation of initial condensation towards the $\boldsymbol{z}$-direction, where $\gamma > 1$ impose $\{e_i\}_{i=1}^n$ to a small neighbourhood of $\{-y_i\}_{i=1}^m$ for a period of time which is at least of order one. To end this part, we provide a sketch of the proofs for Theorem 2, see Figure 6.

6 Conclusions

In this paper, we present the phase diagram of initial condensation for two-layer NNs with a wide class of smooth activation functions. We demonstrate the distinct features exhibited by NNs in the linear regime area and condensed regime

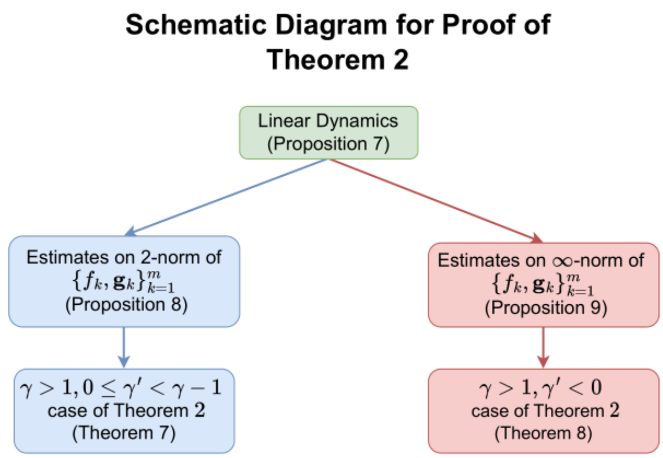


Figure 6: Sketch of proof for Theorem 2.

area, and we provide a complete and detailed analysis for the transition across the boundary (critical regime) in the phase diagram. Moreover, in comparison with the work of Luo et al. [16], we identify the direction towards which the weight parameters condense, and we draw estimates on the time required for initial condensation to occur in contrast to the work of Zhou et al. [31]. The phase diagram at initial stage is crucial in that it is a valuable tool for understanding the implicit regularization effect provided by weight initialization schemes, and it serves as a cornerstone upon which future works can be done to provide thorough characterization of the dynamical behavior of general NNs at each of the identified regime.

In future, we endeavor to establish a framework for the analysis of initial condensation by a series of papers. In our upcoming publication, we plan to extend this formalism to Convolutional Neural Network (CNN) [20], and apply it to investigate the phenomenon of condensation for a wide range of NN architectures, including fully-connected deep network (DNN) and Residual Network (ResNet) [9].

Acknowledgments

This work is sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200 (Z. X., T. L.), the Shanghai Sailing Program, the Natural Science Foundation of Shanghai Grant No. 20ZR1429000 (Z. X.), the National Natural

Science Foundation of China Grant No. 62002221 (Z. X.), the National Natural Science Foundation of China Grant No. 12101401 (T. L.), Shanghai Municipal Science and Technology Key Project No. 22JC1401500 (T. L.), Shanghai Municipal of Science and Technology Major Project No. 2021SHZDZX0102, and the HPC of School of Mathematical Sciences and the Student Innovation Center, and the Siyuan-1 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

References

[1] P. L. Bartlett and S. Mendelson, *Rademacher and gaussian complexities: Risk bounds and structural results*, Journal of Machine Learning Research, 3 (2002), pp. 463–482.

[2] A. Brutzkus and A. Globerson, *Why do larger models generalize better? a theoretical perspective via the xor problem*, in International Conference on Machine Learning, PMLR, 2019, pp. 822–830.

[3] L. Chizat and F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in neural information processing systems, 31 (2018).

[4] L. Chizat, E. Oyallon, and F. Bach, *On lazy training in differentiable programming*, Advances in Neural Information Processing Systems, 32 (2019).

[5] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, *Gradient descent finds global minima of deep neural networks*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., vol. 97 of Proceedings of Machine Learning Research, PMLR, 09–15 Jun 2019, pp. 1675–1685.

[6] S. S. Du, X. Zhai, B. Póczos, and A. Singh, *Gradient descent provably optimizes over-parameterized neural networks*, in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.

[7] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[9] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[10] J. Huang and H.-T. Yau, *Dynamics of deep neural networks and neural tangent hierarchy*, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 4542–4551.

[11] A. Jacot, F. Gabriel, and C. Hongler, *Neural Tangent Kernel: Convergence and Generalization in Neural Networks*, in Advances in neural information processing systems, 2018, pp. 8571–8580.

[12] Z. Ji and M. Telgarsky, *Gradient descent aligns the layers of deep linear networks*, arXiv preprint arXiv:1810.02032, (2018).

[13] B. Laurent and P. Massart, *Adaptive Estimation of a Quadratic Functional by Model Selection*, Annals of Statistics, (2000), pp. 1302–1338.

[14] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient backprop*, Neural networks: Tricks of the trade: Second edition, (2012), pp. 9–48.

[15] Y. Li, T. Luo, and N. K. Yip, *Towards an understanding of residual networks using neural tangent hierarchy (nth)*, CSIAM Transactions on Applied Mathematics, 3 (2022), pp. 692–760.

[16] T. Luo, Z.-Q. J. Xu, Z. Ma, and Y. Zhang, *Phase diagram for two-layer relu neural networks at infinite-width limit*, Journal of Machine Learning Research, 22 (2021), pp. 1–47.

[17] K. Lyu, Z. Li, R. Wang, and S. Arora, *Gradient descent on two-layer nets: Margin maximization and simplicity bias*, Advances in Neural Information Processing Systems, 34 (2021), pp. 12978–12991.

[18] H. Maennel, O. Bousquet, and S. Gelly, *Gradient descent quantizes relu network features*, arXiv preprint arXiv:1803.08367, (2018).

[19] S. Mei, A. Montanari, and P.-M. Nguyen, *A Mean Field View of the Landscape of Two-layer Neural Networks*, Proceedings of the National Academy of Sciences, 115 (2018), pp. E7665–E7671.

[20] K. O'Shea and R. Nash, *An introduction to convolutional neural networks*, arXiv preprint arXiv:1511.08458, (2015).

[21] F. Pellegrini and G. Biroli, *An analytic theory of shallow networks dynamics for hinge loss classification*, Journal of Statistical Mechanics: Theory and Experiment, 2021 (2021), p. 124005.

[22] A. Radhakrishnan, *Lecture 4: Nngp, dual activations, and overparameterization*, 2022.

[23] G. Rotskoff and E. Vanden-Eijnden, *Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks*, Advances in neural information processing systems, 31 (2018).

[24] J. Sirignano and K. Spiliopoulos, *Mean field analysis of neural networks: A central limit theorem*, Stochastic Processes and their Applications, 130 (2020), pp. 1820–1852.

[25] R. Vershynin, *Introduction to the Non-asymptotic Analysis of Random Matrices*, arXiv preprint arXiv:1011.3027, (2010).

[26] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, *Kernel and rich regimes in overparametrized models*, in Conference on Learning Theory, PMLR, 2020, pp. 3635–3673.

[27] G. Yang, *Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation*, arXiv preprint arXiv:1902.04760, (2019).

[28] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning (still) requires rethinking generalization*, Communications of the ACM, 64 (2021), pp. 107–115.

[29] Y. Zhang, Y. Li, Z. Zhang, T. Luo, and Z.-Q. J. Xu, *Embedding principle: a hierarchical structure of loss landscape of deep neural networks*, Journal of Machine Learning vol, 1 (2022), pp. 1–45.

[30] Y. Zhang, Z. Zhang, T. Luo, and Z. J. Xu, *Embedding principle of loss landscape of deep neural networks*, Advances in Neural Information Processing Systems, 34 (2021), pp. 14848–14859.

[31] H. Zhou, T. Luo, Y. Zhang, and Z.-Q. J. Xu, *Towards understanding the condensation of two-layer neural networks at initial training*, arXiv preprint arXiv:2105.11686, (2021).

[32] H. Zhou, Q. Zhou, Z. Jin, T. Luo, Y. Zhang, and Z.-Q. J. Xu, *Empirical phase diagram for three-layer neural networks with infinite width*, arXiv preprint arXiv:2205.12101, (2022).

A Several Estimates on the Initialization

We begin this part by an estimate on standard Gaussian vectors.

Lemma 1 (Bounds on initial parameters)**.** *Given any* $\delta \in (0,1)$*, we have with probability at least* $1-\delta$ *over the choice of* $\boldsymbol{\theta}^0$*,*

$$\max_{k\in[m]} \left\{|a_k^0|,\ \|\boldsymbol{w}_k^0\|_\infty\right\} \leq \sqrt{2\log\frac{2m(d+1)}{\delta}}, \tag{A.1}$$

Proof. If $\mathrm{X} \sim \mathcal{N}(0,1)$, then for any $\eta > 0$,

$$\mathbb{P}(|\mathrm{X}| > \eta) \leq 2\exp\left(-\frac{1}{2}\eta^2\right).$$

Since for $k \in [m]$,

$$a_k^0 \sim \mathcal{N}(0,1), \quad \boldsymbol{w}_k^0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d),$$

where

$$\boldsymbol{w}_k^0 := \left[\left(w_k^0\right)_1, \left(w_k^0\right)_2, \cdots, \left(w_k^0\right)_d\right]^\intercal,$$

then for any $j \in [d]$,

$$\left(w_k^0\right)_j \sim \mathcal{N}(0,1),$$

and they are all independent with each other. As we set

$$\eta = \sqrt{2\log\frac{2m(d+1)}{\delta}},$$

we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\max_{k\in[m]} \left\{|a_k^0|,\ \|\boldsymbol{w}_k^0\|_\infty\right\} > \eta\right) \\
&= \mathbb{P}\left(\max_{k\in[m], j\in[d]} \left\{|a_k^0|,\ |(w_k^0)_j|\right\} > \eta\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^m \left[\left(|a_k^0| > \eta\right) \bigcup \left(\bigcup_{j=1}^d \left(|(w_k^0)_j| > \eta\right)\right)\right]\right) \\
&\leq \sum_{k=1}^m \mathbb{P}\left(|a_k^0| > \eta\right) + \sum_{k=1}^m\sum_{j=1}^d \mathbb{P}\left(|(w_k^0)_j| > \eta\right) \\
&\leq 2m\exp\left(-\frac{1}{2}\eta^2\right) + 2md\exp\left(-\frac{1}{2}\eta^2\right) \\
&= 2m(d+1)\exp\left(-\frac{1}{2}\eta^2\right) = \delta.
\end{aligned}$$

□

Next we would like to introduce the sub-exponential norm [25] of a random variable and Bernstein's Inequality.

Definition 2 (Sub-exponential norm)**.** *The sub-exponential norm of a random variable* X *is defined as*

$$\|\mathrm{X}\|_{\psi_1} := \inf\left\{ s > 0 \mid \mathbb{E}_{\mathrm{X}}\left[\exp\left(\frac{|\mathrm{X}|}{s}\right)\right] \le 2 \right\}. \tag{A.2}$$

In particular, we denote $\mathrm{X} := \chi^2(d)$ *as a chi-square distribution with* d *degrees of freedom [13], and its sub-exponential norm by*

$$C_{\psi,d} := \|\mathrm{X}\|_{\psi_1}.$$

Remark 7. *As the probability density function of* $\mathrm{X} = \chi^2(d)$ *reads*

$$f_{\mathrm{X}}(z) := \frac{1}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})} z^{\frac{d}{2}-1} \exp\left(-\frac{z}{2}\right), \quad z \ge 0,$$

we note that

$$\mathbb{E}_{\mathrm{X}\sim\chi^2(1)} \exp\left(\frac{|\mathrm{X}|}{s}\right) = \int_0^{+\infty} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} z^{-\frac{1}{2}} \exp\left(-\left(\frac{1}{2} - \frac{1}{s}\right) z\right) \mathrm{d}z = \frac{1}{\sqrt{1-\frac{2}{s}}},$$

Then we obtain that

$$\frac{8}{3} \le C_{\psi,1} < 3.$$

Moreover, we notice that

$$\mathbb{E}_{\mathrm{X}\sim\chi^2(d)} \exp\left(\frac{|\mathrm{X}|}{s}\right) = \left(\mathbb{E}_{\mathrm{Y}\sim\chi^2(1)} \exp\left(\frac{|\mathrm{Y}|}{s}\right)\right)^d,$$

as we set

$$\frac{1}{\sqrt{1-\frac{2}{s}}} = 2^{\frac{1}{d}},$$

then

$$s = \frac{2}{1-2^{-\frac{2}{d}}},$$

hence

$$\frac{2}{1-2^{-\frac{2}{d}}} \le C_{\psi,d} < 3,$$

and

$$C_{\psi,d} \ge C_{\psi,1},$$

for $d \ge 1$*.*

Lemma 2. *Given* $\|\boldsymbol{x}\|_2 \le 1$ *and* $\|\boldsymbol{y}\|_2 \le 1$ *equipped with* $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$, *with* $\sigma \in \mathcal{C}^1(\mathbb{R})$, *satisfying*

$$\sigma(0) = 0,$$

and

$$\left\|\sigma^{(1)}(\cdot)\right\|_\infty \le 1.$$

Under the condition that

$$a \sim \mathcal{N}(0,1), \quad \boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d),$$

then for any $\varepsilon > 0$,

- *if*

$$\mathrm{X}_1 := \frac{1}{\varepsilon^2}\sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x})\sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{y}),$$

  then

$$\|\mathrm{X}_1\|_{\psi_1} \le C_{\psi,d},$$

- *and if*

$$\mathrm{X}_2 := a^2\sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x})\sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{y}) \left\langle \boldsymbol{x}, \boldsymbol{y} \right\rangle,$$

  then

$$\|\mathrm{X}_2\|_{\psi_1} \le C_{\psi,1}.$$

Proof. Let

$$\mathrm{Z} := \|\boldsymbol{w}\|_2^2 \sim \chi^2(d).$$

Since

$$|\mathrm{X}_1| \le |\boldsymbol{w}^\intercal \boldsymbol{x}| \, |\boldsymbol{w}^\intercal \boldsymbol{y}| = \|\boldsymbol{w}\|_2^2,$$

then by definition,

$$\begin{aligned}
\|\mathrm{X}_1\|_{\psi_1} &= \inf\left\{s > 0 \mid \mathbb{E}_{\mathrm{X}_1} \exp\left(\frac{|\mathrm{X}_1|}{s}\right) \le 2\right\} \\
&= \inf\left\{s > 0 \mid \mathbb{E}_{\boldsymbol{w}} \exp\left(\frac{|\sigma(\boldsymbol{w}^\intercal \boldsymbol{x})\sigma(\boldsymbol{w}^\intercal \boldsymbol{y})|}{s}\right) \le 2\right\} \\
&\le \inf\left\{s > 0 \mid \mathbb{E}_{\boldsymbol{w}} \exp\left(\frac{\|\boldsymbol{w}\|_2^2}{s}\right) \le 2\right\},
\end{aligned}$$

hence

$$\|\mathrm{X}_1\|_{\psi_1} \le C_{\psi,d}. \tag{A.3}$$

By similar reasoning, as

$$|\mathrm{X}_2| \le a^2,$$

hence

$$\|\mathrm{X}_2\|_{\psi_1} \le C_{\psi,1}. \tag{A.4}$$

□

We state an important theorem without proof, details of which can be found in [25].

Theorem 3 (Bernstein's inequality)**.** *Let $\{\mathrm{X}_k\}_{k=1}^m$ be i.i.d. sub-exponential random variables satisfying*

$$\mathbb{E}\mathrm{X}_1 = \mu,$$

then for any $\eta \ge 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^m \mathrm{X}_k - \mu\right| \ge \eta\right) \le 2\exp\left(-C_0 m \min\left(\frac{\eta^2}{\|\mathrm{X}_1\|_{\psi_1}^2}, \frac{\eta}{\|\mathrm{X}_1\|_{\psi_1}}\right)\right),$$

for some absolute constant C_0.

Proposition 1 (Upper and lower bounds of initial parameters)**.** *Given any $\delta \in (0,1)$, if*

$$m = \Omega\left(\log\frac{4}{\delta}\right),$$

then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\sqrt{\frac{m}{2}} \le \|\boldsymbol{\theta}_a^0\|_2 \le \sqrt{\frac{3m}{2}}, \tag{A.5}$$

$$\sqrt{\frac{md}{2}} \le \|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2 \le \sqrt{\frac{3md}{2}}, \tag{A.6}$$

and

$$\sqrt{\frac{m(d+1)}{2}} \le \|\boldsymbol{\theta}^0\|_2 \le \sqrt{\frac{3m(d+1)}{2}}. \tag{A.7}$$

Proof. Since

$$\left(a_1^0\right)^2, \left(a_2^0\right)^2, \cdots, \left(a_m^0\right)^2 \sim \chi^2(1)$$

are i.i.d. sub-exponential random variables with

$$\mathbb{E}\left(a_1^0\right)^2 = 1.$$

By application of Theorem 3, we have

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^m \left(a_k^0\right)^2 - 1\right| \ge \eta\right) \le 2\exp\left(-C_0 m \min\left(\frac{\eta^2}{C_{\psi,1}^2}, \frac{\eta}{C_{\psi,1}}\right)\right),$$

since $C_{\psi,1} \geq \frac{8}{3} > 2$, then for any $0 \leq \eta \leq 2$, it is obvious that

$$\min\left(\frac{\eta^2}{C_{\psi,1}^2}, \frac{\eta}{C_{\psi,1}}\right) = \frac{\eta^2}{C_{\psi,1}^2}.$$

We set

$$2\exp\left(-C_0 m \frac{\eta^2}{C_{\psi,1}^2}\right) = \frac{\delta}{2},$$

and consequently,

$$\eta = \sqrt{\frac{C_{\psi,1}^2}{mC_0}\log\frac{4}{\delta}},$$

then with probability at least $1 - \frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$m\left(1 - \sqrt{\frac{C_{\psi,1}^2}{mC_0}\log\frac{4}{\delta}}\right) \leq \|\boldsymbol{\theta}_a^0\|_2^2 \leq m\left(1 + \sqrt{\frac{C_{\psi,1}^2}{mC_0}\log\frac{4}{\delta}}\right), \tag{A.8}$$

by choosing

$$m \geq \frac{4C_{\psi,1}^2}{C_0}\log\frac{4}{\delta},$$

we obtain that

$$\sqrt{\frac{m}{2}} \leq \|\boldsymbol{\theta}_a^0\|_2 \leq \sqrt{\frac{3m}{2}}.$$

As for $\boldsymbol{\theta}_{\boldsymbol{w}}^0$, by application of Theorem 3,

$$\mathbb{P}\left(\left|\frac{1}{md}\sum_{k=1}^{md}\left(w_k^0\right)^2 - 1\right| \geq \eta\right) \leq 2\exp\left(-C_0 md \min\left(\frac{\eta^2}{C_{\psi,1}^2}, \frac{\eta}{C_{\psi,1}}\right)\right).$$

We set

$$2\exp\left(-C_0 md \frac{\eta^2}{C_{\psi,1}^2}\right) = \frac{\delta}{2},$$

and consequently,

$$\eta = \sqrt{\frac{C_{\psi,1}^2}{mdC_0}\log\frac{4}{\delta}},$$

then with probability at least $1 - \frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$md\left(1 - \sqrt{\frac{C_{\psi,1}^2}{mdC_0}\log\frac{4}{\delta}}\right) \leq \|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2^2 \leq md\left(1 + \sqrt{\frac{C_{\psi,1}^2}{mdC_0}\log\frac{4}{\delta}}\right), \tag{A.9}$$

by choosing

$$m \geq \frac{4C_{\psi,1}^2}{dC_0} \log \frac{4}{\delta},$$

we obtain that

$$\sqrt{\frac{md}{2}} \leq \|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2 \leq \sqrt{\frac{3md}{2}}.$$

Finally, as

$$\|\boldsymbol{\theta}^0\|_2^2 = \|\boldsymbol{\theta}_a^0\|_2^2 + \|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2^2,$$

then with probability at least $1-\delta$,

$$\sqrt{\frac{m(d+1)}{2}} \leq \|\boldsymbol{\theta}^0\|_2 \leq \sqrt{\frac{3m(d+1)}{2}}.$$

□

B Linear Regime

B.1 Full Rankness of the Gram matrices

We shall state two lemmas concerning full rankness of the Gram matrices, which have been stated as Lemma F.1. and Lemma F.2. in Du et al. [5].

Lemma 3. *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider input data set as $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$, and non-parallel with each other, i.e., for any $j \neq k$,*

$$\boldsymbol{z}_j \notin \mathrm{span}\left(\boldsymbol{z}_k\right),$$

we define

$$\begin{aligned} \boldsymbol{G}(\mathcal{Z}) &:= \left[\boldsymbol{G}(\mathcal{Z})_{ij}\right], \\ \boldsymbol{G}(\mathcal{Z})_{ij} &:= \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sigma(\boldsymbol{w}^\intercal \boldsymbol{z}_i)\sigma(\boldsymbol{w}^\intercal \boldsymbol{z}_j)\right], \end{aligned} \tag{B.1}$$

then $\lambda_{\min}\left(\boldsymbol{G}(\mathcal{Z})\right) > 0$.

Similar to Lemma 3, we have another Lemma.

Lemma 4. *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider input data set as $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$, and non-parallel with each other, we define*

$$\begin{aligned} \boldsymbol{G}(\mathcal{Z}) &:= \left[\boldsymbol{G}(\mathcal{Z})_{ij}\right], \\ \boldsymbol{G}(\mathcal{Z})_{ij} &:= \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sigma^{(1)}(\boldsymbol{w}^\intercal \boldsymbol{z}_i)\sigma^{(1)}(\boldsymbol{w}^\intercal \boldsymbol{z}_j)\left\langle \boldsymbol{z}_i, \boldsymbol{z}_j\right\rangle\right], \end{aligned} \tag{B.2}$$

then $\lambda_{\min}\left(\boldsymbol{G}(\mathcal{Z})\right) > 0$.

We state an important theorem concerning the least eigenvalue of the normalized Gram matrices $\boldsymbol{K}^{[a]}$ and $\boldsymbol{K}^{[\boldsymbol{w}]}$ at infinite width limit. Recall that

$$\begin{aligned} K_{ij}^{[a]} &= \frac{1}{\varepsilon^2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\sigma(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_j), \\ K_{ij}^{[\boldsymbol{w}]} &= \mathbb{E}_{(a,\boldsymbol{w})\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_{d+1})}a^2\sigma^{(1)}(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma^{(1)}(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_j)\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle. \end{aligned} \tag{B.3}$$

Theorem 4 (Least eigenvalue of Gram matrices at infinite width)**.** *Under Assumption 2 and Assumption 4, the normalized Gram matrices $\boldsymbol{K}^{[a]}$ and $\boldsymbol{K}^{[\boldsymbol{w}]}$ are strictly positive definite, and both of their least eigenvalues are of order* 1*.*

In other words, if we denote

$$\lambda := \min\{\lambda_a, \lambda_{\boldsymbol{w}}\} > 0,$$

where

$$\lambda_a := \lambda_{\min}\left(\boldsymbol{K}^{[a]}\right), \quad \lambda_{\boldsymbol{w}} := \lambda_{\min}\left(\boldsymbol{K}^{[\boldsymbol{w}]}\right), \tag{B.4}$$

then

$$\lambda > 0,$$

and

$$\lambda \sim \Omega(1).$$

Proof. Since the following limit exists

$$\gamma_2 = \lim_{m\to\infty} -\frac{\log\varepsilon}{\log m},$$

then we conclude that

$$\lim_{m\to\infty}\varepsilon = 0,$$

or

$$\lim_{m\to\infty}\varepsilon = 1,$$

or

$$\lim_{m\to\infty}\varepsilon = +\infty,$$

and the normalized matrices $\boldsymbol{K}^{[a]} = \left[K_{ij}^{[a]}\right]_{n\times n}$ and $\boldsymbol{K}^{[\boldsymbol{w}]} = \left[K_{ij}^{[\boldsymbol{w}]}\right]_{n\times n}$ read

$$\begin{aligned} K_{ij}^{[a]} &= \frac{1}{\varepsilon^2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\sigma(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_j), \\ K_{ij}^{[\boldsymbol{w}]} &= \mathbb{E}_{(a,\boldsymbol{w})\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_{d+1})}a^2\sigma^{(1)}(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma^{(1)}(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_j)\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle \\ &= \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\sigma^{(1)}(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma^{(1)}(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_j)\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle. \end{aligned}$$

For the case where $\lim_{m\to\infty} \varepsilon = 1$, by direct application of Lemma 3 and Lemma 4, $\lambda \sim \Omega(1)$ can be easily achieved.

For the case where $\lim_{m\to\infty} \varepsilon = 0$, the normalized matrices read

$$\begin{aligned}
K_{ij}^{[a]} &= \lim_{m\to\infty} \frac{1}{\varepsilon^2} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i) \sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j) \\
&= \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} (\boldsymbol{w}^\intercal \boldsymbol{x}_i)(\boldsymbol{w}^\intercal \boldsymbol{x}_j), \\
K_{ij}^{[\boldsymbol{w}]} &= \lim_{m\to\infty} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i) \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j) \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle \\
&= \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \left[\sigma^{(1)}(0)^2 \right] \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle \\
&= \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle .
\end{aligned}$$

hence both $\boldsymbol{K}^{[a]}$ and $\boldsymbol{K}^{[\boldsymbol{w}]}$ are independent of $\boldsymbol{\theta}_{\boldsymbol{w}}$ and $\boldsymbol{\theta}_{\boldsymbol{w}}$, and depend merely on the input sample $\mathcal{S}$. Consequently, $\lambda \sim \Omega(1)$.

For the case where $\lim_{m\to\infty} \varepsilon = +\infty$, the normalized matrices read

$$\begin{aligned}
K_{ij}^{[a]} &= \lim_{m\to\infty} \frac{1}{\varepsilon^2} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i) \sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j) \\
&= \lim_{m\to\infty} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i) \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j), \\
K_{ij}^{[\boldsymbol{w}]} &= \lim_{m\to\infty} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i) \sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j) \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle ,
\end{aligned}$$

where the entries of the matrix $\boldsymbol{K}^{[a]}$ exhibit leaky-ReLU-like behaviors. It has been proven in Du et al. [6] that under the unit data and nonparallel assumption, $\boldsymbol{K}^{[a]}$ is positive definite. Moreover, its expression can be explicitly computed, see [22]. As for the matrix $\boldsymbol{K}^{[\boldsymbol{w}]}$, we define a function $\boldsymbol{G} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, such that

$$\boldsymbol{G}(\boldsymbol{K})_{ij} := \boldsymbol{K}_{ij} \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{K}_{ii} & \boldsymbol{K}_{ij} \\ \boldsymbol{K}_{ji} & \boldsymbol{K}_{jj} \end{pmatrix}\right)} \sigma^{(1)}(u) \sigma^{(1)}(v).$$

We denote that $\boldsymbol{A} \succeq \boldsymbol{B}$ if and only if $\boldsymbol{A} - \boldsymbol{B}$ is a semi-positive definite matrix, and $\boldsymbol{A} \succ \boldsymbol{B}$ if and only if $\boldsymbol{A} - \boldsymbol{B}$ is a strictly positive definite matrix. Consequently, a scalar function $g(t)$ is defined as

$$g(t) := \min_{\boldsymbol{K}:\boldsymbol{K}\succ 0,\ \frac{1}{c} \leq \boldsymbol{K}_{ii} \leq c,\ \lambda_{\min}(\boldsymbol{K}) \geq t} \lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{K})\right).$$

Then Lemma 4 guarantees that

$$g(\lambda_0) > 0,$$

and $\lambda \sim \Omega(1)$. □

Finally, we are hereby to show that the Gram matrix $\boldsymbol{G}$ at $t=0$ is also positive definite. Recall that

$$\begin{aligned}\boldsymbol{G}^{[a]}(\boldsymbol{\theta}) &= \frac{\nu\varepsilon^3}{m}\sum_{k=1}^{m}\frac{1}{\varepsilon^2}\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_j),\\ \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}) &= \frac{\nu^3\varepsilon}{m}\sum_{k=1}^{m}a_k^2\sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_j)\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle,\end{aligned} \tag{B.5}$$

Proposition 2 (Least eigenvalue of initial Gram matrices)**.** *Given any $\delta\in(0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if*

$$m=\Omega\left(\frac{n^2}{\lambda^2}\log\frac{4n^2}{\delta}\right),$$

where $\lambda=\min\{\lambda_a,\lambda_{\boldsymbol{w}}\}$, whose definition can be found in Theorem 4. Then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have

$$\lambda_{\min}\left(\boldsymbol{G}\left(\boldsymbol{\theta}^0\right)\right)\geq\frac{3}{4}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right). \tag{B.6}$$

Proof. For any $\eta>0$ and all $i,j\in[n]$, we define the events

$$\begin{aligned}\Omega_{ij}^{[a]} &:= \left\{\boldsymbol{\theta}^0 \mid \left|\frac{1}{\nu\varepsilon^3}G_{ij}^{[a]}\left(\boldsymbol{\theta}^0\right)-K_{ij}^{[a]}\right|\leq\frac{\eta}{n}\right\},\\ \Omega_{ij}^{[\boldsymbol{w}]} &:= \left\{\boldsymbol{\theta}^0 \mid \left|\frac{1}{\nu^3\varepsilon}G_{ij}^{[\boldsymbol{w}]}\left(\boldsymbol{\theta}^0\right)-K_{ij}^{[\boldsymbol{w}]}\right|\leq\frac{\eta}{n}\right\}.\end{aligned}$$

By application of Lemma 2 and Theorem 3, we obtain that for sufficiently small $\eta>0$, the following holds

$$\begin{aligned}\mathbb{P}(\Omega_{ij}^{[a]}) &\geq 1-2\exp\left(-\frac{mC_0\eta^2}{n^2C_{\psi,d}^2}\right)\\ \mathbb{P}(\Omega_{ij}^{[\boldsymbol{w}]}) &\geq 1-2\exp\left(-\frac{mC_0\eta^2}{n^2C_{\psi,1}^2}\right)\geq 1-2\exp\left(-\frac{mC_0\eta^2}{n^2C_{\psi,d}^2}\right),\end{aligned}$$

hence with probability at least $1-4n^2\exp\left(-\frac{mC_0\eta^2}{n^2C_{\psi,d}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have

$$\begin{aligned}\left\|\frac{1}{\nu\varepsilon^3}\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)-\boldsymbol{K}^{[a]}\right\|_{\mathrm{F}} &\leq n\left\|\frac{1}{\nu\varepsilon^3}\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)-\boldsymbol{K}^{[a]}\right\|_{\infty}\leq\eta,\\ \left\|\frac{1}{\nu^3\varepsilon}\boldsymbol{G}^{[\boldsymbol{w}]}\left(\boldsymbol{\theta}^0\right)-\boldsymbol{K}^{[\boldsymbol{w}]}\right\|_{\mathrm{F}} &\leq n\left\|\frac{1}{\nu^3\varepsilon}\boldsymbol{G}^{[\boldsymbol{w}]}\left(\boldsymbol{\theta}^0\right)-\boldsymbol{K}^{[\boldsymbol{w}]}\right\|_{\infty}\leq\eta.\end{aligned}$$

By taking $\eta = \frac{\lambda}{4}$,

$$\delta = 4n^2 \exp\left(-\frac{mC_0\lambda^2}{16n^2 C_{\psi,d}^2}\right),$$

and we conclude that

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{G}\left(\boldsymbol{\theta}^0\right)\right) &\geq \lambda_{\min}\left(\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)\right) + \lambda_{\min}\left(\boldsymbol{G}^{[\boldsymbol{w}]}\left(\boldsymbol{\theta}^0\right)\right) \\
&\geq \nu\varepsilon^3\lambda_a - \nu\varepsilon^3\left\|\frac{1}{\nu\varepsilon^3}\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right) - \boldsymbol{K}^{[a]}\right\|_{\mathrm{F}} \\
&\quad + \nu^3\varepsilon\lambda_{\boldsymbol{w}} - \nu^3\varepsilon\left\|\frac{1}{\nu^3\varepsilon}\boldsymbol{G}^{[\boldsymbol{w}]}\left(\boldsymbol{\theta}^0\right) - \boldsymbol{K}^{[\boldsymbol{w}]}\right\|_{\mathrm{F}} \\
&\geq \nu\varepsilon^3(\lambda_a - \eta) + \nu^3\varepsilon(\lambda_{\boldsymbol{w}} - \eta) \\
&\geq \frac{3}{4}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right).
\end{aligned}$$

□

B.2 $\boldsymbol{\theta}$-lazy Regime

In the rest of the paper, we define two quantities

$$\alpha(t) := \max_{k\in[m], s\in[0,t]} |a_k(s)|, \quad \omega(t) := \max_{k\in[m], s\in[0,t]} \|\boldsymbol{w}_k(s)\|_\infty, \tag{B.7}$$

and we denote

$$t^* = \inf\{t \mid \boldsymbol{\theta}(t) \notin \mathcal{N}\left(\boldsymbol{\theta}^0\right)\}, \tag{B.8}$$

where the event is defined as

$$\mathcal{N}\left(\boldsymbol{\theta}^0\right) := \left\{\boldsymbol{\theta} \mid \|\boldsymbol{G}(\boldsymbol{\theta}) - \boldsymbol{G}\left(\boldsymbol{\theta}^0\right)\|_{\mathrm{F}} \leq \frac{1}{4}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)\right\}. \tag{B.9}$$

We observe immediately that the event $\mathcal{N}\left(\boldsymbol{\theta}^0\right) \neq \varnothing$, since $\boldsymbol{\theta}^0 \in \mathcal{N}\left(\boldsymbol{\theta}^0\right)$. Recall that $\lambda = \min\{\lambda_a, \lambda_{\boldsymbol{w}}\}$, where

$$\lambda_a = \lambda_{\min}\left(\boldsymbol{K}^{[a]}\right), \quad \lambda_{\boldsymbol{w}} = \lambda_{\min}\left(\boldsymbol{K}^{[\boldsymbol{w}]}\right),$$

whose definition can be found in Theorem 4. Then we have the following lemma.

Proposition 3. *Given any $\delta \in (0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if*

$$m = \Omega\left(\frac{n^2}{\lambda^2}\log\frac{4n^2}{\delta}\right),$$

then with probability at least $1-\delta$ *over the choice of* $\boldsymbol{\theta}^0$*, we have for any time* $t\in[0,t^*)$*,*

$$R_S(\boldsymbol{\theta}(t)) \le \exp\left(-\frac{m}{n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)R_S\left(\boldsymbol{\theta}^0\right). \tag{B.10}$$

Proof. Proposition 2 implies that for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and for any $\boldsymbol{\theta}\in\mathcal{N}\left(\boldsymbol{\theta}^0\right)$, we have

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta})\right) &\ge \lambda_{\min}\left(\boldsymbol{G}\left(\boldsymbol{\theta}^0\right)\right)-\|\boldsymbol{G}(\boldsymbol{\theta})-\boldsymbol{G}\left(\boldsymbol{\theta}^0\right)\|_{\mathrm{F}}\\
&\ge \frac{3}{4}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)-\frac{1}{4}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)\\
&= \frac{1}{2}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
G_{ij}(\boldsymbol{\theta}) &= G^{[a]}_{ij}(\boldsymbol{\theta})+G^{[\boldsymbol{w}]}_{ij}(\boldsymbol{\theta})\\
&= \frac{\nu\varepsilon^3}{m}\sum_{k=1}^m\frac{1}{\varepsilon^2}\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_j)+\frac{\nu^3\varepsilon}{m}\sum_{k=1}^m a_k^2\sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_j)\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle,
\end{aligned}$$

and the normalized flow

$$\begin{aligned}
\nabla_{a_k}R_S(\boldsymbol{\theta}) &= \nu\varepsilon\frac{1}{n}\sum_{i=1}^n e_i\frac{\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)}{\varepsilon},\\
\nabla_{\boldsymbol{w}_k}R_S(\boldsymbol{\theta}) &= \nu\varepsilon\frac{1}{n}\sum_{i=1}^n e_i a_k\sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i.
\end{aligned}$$

Finally, we obtain that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}R_S(\boldsymbol{\theta}) &= -\left(\sum_{k=1}^m\frac{\varepsilon}{\nu}\left\langle\nabla_{a_k}R_S(\boldsymbol{\theta}),\nabla_{a_k}R_S(\boldsymbol{\theta})\right\rangle+\sum_{k=1}^m\frac{\nu}{\varepsilon}\left\langle\nabla_{\boldsymbol{w}_k}R_S(\boldsymbol{\theta}),\nabla_{\boldsymbol{w}_k}R_S(\boldsymbol{\theta})\right\rangle\right)\\
&= -\frac{m}{n^2}\boldsymbol{e}^\intercal\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{e}\\
&\le -\frac{2m}{n}\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta})\right)R_S(\boldsymbol{\theta})\\
&\le -\frac{m}{n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)R_S(\boldsymbol{\theta}),
\end{aligned}$$

and immediate integration yields the result. □

Proposition 4. *Given any $\delta \in (0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if $\gamma < 1$, and*

$$m = \max\left(\Omega\left(\frac{n^2}{\lambda^2}\log\frac{8n^2}{\delta}\right), \Omega\left(\left(\frac{n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda}\right)^{\frac{1}{1-\gamma}}\right)\right),$$

then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, for any time $t \in [0, t^*)$ and for any $k \in [m]$, both

$$\begin{aligned}
\max_{k\in[m]}|a_k(t) - a_k(0)| &\leq 2\max\left\{\frac{\varepsilon}{\nu}, 1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}p, \\
\max_{k\in[m]}\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty &\leq 2\max\left\{\frac{\nu}{\varepsilon}, 1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}p,
\end{aligned} \tag{B.11}$$

and

$$\max_{k\in[m]}\{|a_k(0)|,\ \|\boldsymbol{w}_k(0)\|_\infty\} \leq \sqrt{2\log\frac{4m(d+1)}{\delta}}, \tag{B.12}$$

hold, where

$$p := \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}.$$

Proof. Since

$$\alpha(t) = \max_{k\in[m],s\in[0,t]}|a_k(s)|, \quad \omega(t) = \max_{k\in[m],s\in[0,t]}\|\boldsymbol{w}_k(s)\|_\infty,$$

we obtain

$$\begin{aligned}
|\nabla_{a_k}R_S|^2 &= \left|\frac{1}{n}\sum_{i=1}^n e_i\nu\sigma(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\right|^2 \leq 2\|\varepsilon\boldsymbol{w}_k\|_1^2\nu^2R_S(\boldsymbol{\theta}) \leq 2d^2(\omega(t))^2\nu^2\varepsilon^2R_S(\boldsymbol{\theta}), \\
\|\nabla_{\boldsymbol{w}_k}R_S\|^2 &= \left\|\frac{1}{n}\sum_{i=1}^n e_i\nu\varepsilon a_k\sigma^{(1)}(\varepsilon\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i\right\|_\infty^2 \leq 2|a_k|^2\nu^2\varepsilon^2R_S(\boldsymbol{\theta}) \leq 2(\alpha(t))^2\nu^2\varepsilon^2R_S(\boldsymbol{\theta}).
\end{aligned}$$

By Proposition 2 and Proposition 3, we have if

$$m \geq \frac{16n^2C_{\psi,d}^2}{\lambda^2C_0}\log\frac{8n^2}{\delta},$$

then with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, we have that

$$
\begin{aligned}
|a_k(t)-a_k(0)| &\leq \frac{\varepsilon}{\nu}\int_0^t |\nabla_{a_k} R_S(\boldsymbol{\theta}(s))|\,\mathrm{d}s \\
&\leq \sqrt{2}d\varepsilon^2 \int_0^t \omega(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\leq \sqrt{2}d\varepsilon^2\omega(t)\int_0^t \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s \\
&\leq \sqrt{2}d\varepsilon^2\omega(t)\int_0^\infty \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s \\
&= \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\omega(t) = \frac{\varepsilon}{\nu}p\omega(t),
\end{aligned}
$$

and

$$
\begin{aligned}
\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty &\leq \frac{\nu}{\varepsilon}\int_0^t \|\nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}(s))\|_\infty\,\mathrm{d}s \\
&\leq \sqrt{2}\nu^2 \int_0^t \alpha(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\leq \sqrt{2}\nu^2\alpha(t)\int_0^t \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s \\
&\leq \sqrt{2}\nu^2\alpha(t)\int_0^\infty \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s \\
&= \frac{2\sqrt{2}n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\alpha(t) \leq \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\alpha(t) = \frac{\nu}{\varepsilon}p\alpha(t).
\end{aligned}
$$

Thus

$$
\begin{aligned}
\alpha(t) &\leq \alpha(0)+\frac{\varepsilon}{\nu}p\omega(t), \\
\omega(t) &\leq \omega(0)+\frac{\nu}{\varepsilon}p\alpha(t),
\end{aligned}
$$

moreover, by Lemma 1, we have with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, as

$$
\max_{k\in[m]}\{|a_k(0)|,\ \|\boldsymbol{w}_k(0)\|_\infty\} \leq \sqrt{2\log\frac{4m(d+1)}{\delta}},
$$

then

$$
\alpha(0) \leq \sqrt{2\log\frac{4m(d+1)}{\delta}},\quad \omega(0) \leq \sqrt{2\log\frac{4m(d+1)}{\delta}}
$$

Hence if

$$m \geq \left(\frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda}\right)^{\frac{1}{1-\gamma}},$$

we have

$$p = \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)} \leq \frac{\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\lambda} \leq \frac{1}{2}.$$

Thus

$$\alpha(t) \leq \alpha(0) + \frac{\varepsilon}{\nu}p\omega(0) + p^2\alpha(t),$$

hence

$$\alpha(t) \leq \frac{4}{3}\alpha(0) + \frac{2}{3}\frac{\varepsilon}{\nu}\omega(0).$$

Therefore

$$\alpha(t) \leq 2\max\left\{\frac{\varepsilon}{\nu}, 1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}.$$

Similarly, one can obtain the estimate of $\omega(t)$ as

$$\omega(t) \leq 2\max\left\{\frac{\nu}{\varepsilon}, 1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}.$$

To sum up, for any $t \in [0, t^*)$, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{k\in[m]}|a_k(t) - a_k(0)| \leq 2\max\left\{\frac{\varepsilon}{\nu}, 1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}p,$$
$$\max_{k\in[m]}\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2\max\left\{\frac{\nu}{\varepsilon}, 1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}p.$$

□

Theorem 5. *Given any $\delta \in (0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if $\gamma < 1$ and*

$$m = \max\left(\Omega\left(\frac{n^2}{\lambda^2}\log\frac{16n^2}{\delta}\right), \Omega\left(\left(\frac{n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda}\right)^{\frac{1}{1-\gamma}}\right),\right.$$
$$\left.\Omega\left(\left(\frac{n^2\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda^2}\right)^{\frac{1}{1-\gamma}}\right), \Omega\left(\log\frac{8}{\delta}\right)\right), \tag{B.13}$$

then with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, we have for all time $t > 0$,

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{2m\nu^2\varepsilon^2\lambda t}{n}\right)R_S\left(\boldsymbol{\theta}^0\right), \tag{B.14}$$

and with probability at least $1-\delta$ *over the choice of* $\boldsymbol{\theta}^0$,

$$\lim_{m\to\infty}\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}=0. \tag{B.15}$$

Remark 8. *In this scenario, we also obtain that with probability at least* $1-\delta$ *over the choice of* $\boldsymbol{\theta}^0$,

$$\lim_{m\to\infty}\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2}=0. \tag{B.16}$$

Proof. According to Proposition 4, it suffices to show that $t^*=\infty$.
(i). Firstly, from Proposition 3, we have with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, for any $t\in[0,t^*)$, the following holds

$$\begin{aligned}|a_k(t)-a_k(0)| &\leq \frac{\varepsilon}{\nu}p\omega(t)\leq 2\max\left\{\frac{\varepsilon}{\nu},1\right\}\xi p,\\ \|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty &\leq \frac{\nu}{\varepsilon}p\alpha(t)\leq 2\max\left\{\frac{\nu}{\varepsilon},1\right\}\xi p,\end{aligned}$$

where

$$p=\frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)},$$

and

$$\xi:=\sqrt{2\log\frac{8m(d+1)}{\delta}}.$$

For m large enough, i.e.,

$$m\geq\left(\frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda}\right)^{\frac{1}{1-\gamma}},$$

then we have

$$p=\frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\leq\frac{\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\lambda}\leq\frac{1}{2}.$$

It is noteworthy to emphasize that the demonstration of (B.14) requires solely the utilization of the aforementioned relations.
(ii). Let

$$g_{ij}^{[a]}(\boldsymbol{w}):=\frac{1}{\varepsilon^2}\sigma(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma(\varepsilon\boldsymbol{w}^\intercal\boldsymbol{x}_j),$$

then

$$\left|G_{ij}^{[a]}(\boldsymbol{\theta}(t))-G_{ij}^{[a]}(\boldsymbol{\theta}(0))\right|\leq\frac{\nu\varepsilon^3}{m}\sum_{k=1}^m\left|g_{ij}^{[a]}(\boldsymbol{w}_k(t))-g_{ij}^{[a]}(\boldsymbol{w}_k(0))\right|.$$

By mean value theorem, for some $c \in (0,1)$,

$$\left|g_{ij}^{[a]}(\boldsymbol{w}_k(t)) - g_{ij}^{[a]}(\boldsymbol{w}_k(0))\right| \le \left\|\nabla_{\boldsymbol{w}} g_{ij}^{[a]}\left(c\boldsymbol{w}_k(t) + (1-c)\boldsymbol{w}_k(0)\right)\right\|_\infty \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_1,$$

where

$$\nabla_{\boldsymbol{w}} g_{ij}^{[a]}(\boldsymbol{w}) = \frac{1}{\varepsilon}\sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j)\boldsymbol{x}_i + \frac{1}{\varepsilon}\sigma(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma^{(1)}(\varepsilon \boldsymbol{w}^\intercal \boldsymbol{x}_j)\boldsymbol{x}_j,$$

and

$$\left\|\nabla_{\boldsymbol{w}} g_{ij}^{[a]}(\boldsymbol{w})\right\|_\infty \le 2\|\boldsymbol{w}\|_1.$$

then

$$\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_1 \le 2d \max\left\{\frac{\nu}{\varepsilon}, 1\right\} \xi p,$$

thus, we have

$$\begin{aligned}\|c\boldsymbol{w}_k(t) + (1-c)\boldsymbol{w}_k(0)\|_1 &\le d\left(\|\boldsymbol{w}_k(0)\|_\infty + \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty\right)\\ &\le d\left(\xi + 2\max\left\{\frac{\nu}{\varepsilon}, 1\right\}\xi p\right)\\ &\le d\left(\xi + \max\left\{\frac{\nu}{\varepsilon}, 1\right\}\xi\right)\\ &\le 2\max\left\{\frac{\nu}{\varepsilon}, 1\right\} d\xi,\end{aligned}$$

hence

$$\begin{aligned}\left|G_{ij}^{[a]}(\boldsymbol{\theta}(t)) - G_{ij}^{[a]}(\boldsymbol{\theta}(0))\right| &\le \frac{\nu\varepsilon^3}{m}\sum_{k=1}^m \left|g_{ij}^{[a]}(\boldsymbol{w}_k(t)) - g_{ij}^{[a]}(\boldsymbol{w}_k(0))\right|\\ &\le 8d^2\nu^2\varepsilon^2\xi^2 \max\left\{\frac{\nu}{\varepsilon}, \frac{\varepsilon}{\nu}\right\} p,\end{aligned}$$

then by using the same technique in Proposition 2,

$$\begin{aligned}&\left\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t)) - \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0))\right\|_{\mathrm{F}}\\ &\quad\le 8d^2 n\nu^2\varepsilon^2 \max\left\{\frac{\nu}{\varepsilon}, \frac{\varepsilon}{\nu}\right\}\left(2\log\frac{8m(d+1)}{\delta}\right)\frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\\ &\le \nu\varepsilon\max\left\{\frac{\nu}{\varepsilon}, \frac{\varepsilon}{\nu}\right\}\frac{32\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}.\end{aligned}$$

We shall notice that as

$$\frac{1}{\lambda^2} \ge 16\left(\frac{1}{27}\right)^{\frac{1}{2}} \frac{1}{\left(\left(\frac{\varepsilon}{\nu}\right)^{\frac{3}{2}}\lambda_a + \left(\frac{\nu}{\varepsilon}\right)^{\frac{1}{2}}\lambda_{\boldsymbol{w}}\right)^2}$$

$$= 16\left(\frac{1}{27}\right)^{\frac{1}{2}}\frac{\nu}{\varepsilon}\frac{1}{\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2},$$

and

$$\begin{aligned}\frac{1}{\lambda^2} &\geq 16\left(\frac{1}{27}\right)^{\frac{1}{2}}\frac{1}{\left(\left(\frac{\varepsilon}{\nu}\right)^{\frac{1}{2}}\lambda_a+\left(\frac{\nu}{\varepsilon}\right)^{\frac{3}{2}}\lambda_{\boldsymbol{w}}\right)^2}\\ &= 16\left(\frac{1}{27}\right)^{\frac{1}{2}}\frac{\varepsilon}{\nu}\frac{1}{\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2}.\end{aligned}$$

As $\gamma < 1$, then we may choose m large enough, such that

$$\begin{aligned}m\nu\varepsilon &\geq \frac{96\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda^2}\\ &\geq \frac{256\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{16\left(\frac{1}{27}\right)^{\frac{1}{2}}\lambda^2}\\ &\geq \max\left\{\frac{\nu}{\varepsilon},\frac{\varepsilon}{\nu}\right\}\frac{256\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2}.\end{aligned}$$

Then for any $t \in [0, t^*)$,

$$\left\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))-\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0))\right\|_{\mathrm{F}} \leq \frac{1}{8}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right). \tag{B.17}$$

(iii). As for $\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))$, we observe that directly from Proposition 3, for any $t \in [0, t^*)$, the following holds

$$\begin{aligned}&\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))-G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right|\\ &\leq \frac{\nu^3\varepsilon\left|\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle\right|}{m}\left(\sum_{k=1}^m\left|a_k^2(t)\sigma^{(1)}\left(\langle\boldsymbol{w}_k(t),\boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle\boldsymbol{w}_k(t),\boldsymbol{x}_j\rangle\right)\right.\right.\\ &\qquad\left.\left.-a_k^2(0)\sigma^{(1)}\left(\langle\boldsymbol{w}_k(0),\boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle\boldsymbol{w}_k(0),\boldsymbol{x}_j\rangle\right)\right|\right),\end{aligned}$$

we define a new quantity

$$D_{k,i,j}(t) := \sigma^{(1)}\left(\langle\boldsymbol{w}_k(t),\boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle\boldsymbol{w}_k(t),\boldsymbol{x}_j\rangle\right)-\sigma^{(1)}\left(\langle\boldsymbol{w}_k(0),\boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle\boldsymbol{w}_k(0),\boldsymbol{x}_j\rangle\right),$$

then we obtain that

$$\begin{aligned}
&\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right| \\
&\leq \frac{\nu^3\varepsilon}{m}\sum_{k=1}^{m}\left(a_k^2(t)D_{k,i,j}(t) + \left|a_k^2(t) - a_k^2(0)\right|\right).
\end{aligned}$$

We shall make an estimate on the quantity $D_{k,i,j}(t)$:

$$\begin{aligned}
D_{k,i,j}(t) &= \sigma^{(1)}\left(\langle \boldsymbol{w}_k(t), \boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle \boldsymbol{w}_k(t), \boldsymbol{x}_j\rangle\right) - \sigma^{(1)}\left(\langle \boldsymbol{w}_k(t), \boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle \boldsymbol{w}_k(0), \boldsymbol{x}_j\rangle\right) \\
&\quad + \sigma^{(1)}\left(\langle \boldsymbol{w}_k(t), \boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle \boldsymbol{w}_k(0), \boldsymbol{x}_j\rangle\right) - \sigma^{(1)}\left(\langle \boldsymbol{w}_k(0), \boldsymbol{x}_i\rangle\right)\sigma^{(1)}\left(\langle \boldsymbol{w}_k(0), \boldsymbol{x}_j\rangle\right) \\
&\leq \left|\langle \boldsymbol{w}_k(t), \boldsymbol{x}_i\rangle - \langle \boldsymbol{w}_k(0), \boldsymbol{x}_i\rangle\right| + \left|\langle \boldsymbol{w}_k(t), \boldsymbol{x}_j\rangle - \langle \boldsymbol{w}_k(0), \boldsymbol{x}_j\rangle\right| \\
&\leq 2\left\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\right\|_1 \leq 2d\left\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\right\|_\infty \\
&\leq 4d\max\left\{\frac{\nu}{\varepsilon}, 1\right\}\xi p.
\end{aligned}$$

Hence, we observe that

$$\begin{aligned}
\left|a_k^2(t) - a_k^2(0)\right| &\leq \left|a_k(t) - a_k(0)\right|^2 + 2\left|a_k(0)\right|\left|a_k(t) - a_k(0)\right| \\
&\leq \left(2\max\left\{\frac{\varepsilon}{\nu}, 1\right\}\xi p\right)^2 + 2\xi\left(2\max\left\{\frac{\varepsilon}{\nu}, 1\right\}\xi p\right) \\
&\leq 6\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2, 1\right\}\xi^2 p,
\end{aligned}$$

and consequently

$$\begin{aligned}
a_k(t)^2 &\leq \left|a_k^2(t) - a_k^2(0)\right| + a_k^2(0) \\
&\leq 6\xi^2\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2, 1\right\}p + \xi^2 \\
&\leq 3\xi^2\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2, 1\right\} + \xi^2 \\
&\leq 4\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2, 1\right\}\xi^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right| \\
&\leq \frac{\nu^3\varepsilon}{m}\sum_{k=1}^{m}\left(a_k^2(t)D_{k,i,j}(t) + \left|a_k^2(t) - a_k^2(0)\right|\right) \\
&\leq \nu^3\varepsilon\left(16d\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2, \frac{\nu}{\varepsilon}\right\}\xi^3 p + 6\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2, 1\right\}\xi^2 p\right)
\end{aligned}$$

$$\leq \nu^3\varepsilon\left(16d\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2,\frac{\nu}{\varepsilon}\right\}\xi^3p+6d\max\left\{\left(\frac{\varepsilon}{\nu}\right)^2,1\right\}\xi^2p\right)$$

$$\leq 24\nu^2\varepsilon^2 d\left(2\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\max\left\{\left(\frac{\nu}{\varepsilon}\right)^2,\frac{\varepsilon}{\nu}\right\}p,$$

then by using the same technique in Proposition 2,

$$\begin{aligned}
&\left\|G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))-G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right\|_{\mathrm{F}}\\
&\leq n\left\|G^{[\boldsymbol{w}]}_{ij}(\boldsymbol{\theta}(t))-G^{[\boldsymbol{w}]}_{ij}(\boldsymbol{\theta}(0))\right\|_{\infty}\\
&\leq 24\nu^2\varepsilon^2 d\left(2\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\max\left\{\left(\frac{\nu}{\varepsilon}\right)^2,\frac{\varepsilon}{\nu}\right\}np\\
&=\nu\varepsilon\max\left\{\left(\frac{\nu}{\varepsilon}\right)^2,\frac{\varepsilon}{\nu}\right\}\frac{96d^2n^2\left(\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)},
\end{aligned}$$

We notice that as

$$\frac{1}{\lambda^2}\geq\frac{\left(\frac{\nu}{\varepsilon}\right)^2}{\left(\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2}=\left(\frac{\nu}{\varepsilon}\right)^2\frac{1}{\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2},$$

and

$$\begin{aligned}
\frac{1}{\lambda^2}&\geq 16\left(\frac{1}{27}\right)^{\frac{1}{2}}\frac{1}{\left(\left(\frac{\varepsilon}{\nu}\right)^{\frac{1}{2}}\lambda_a+\left(\frac{\nu}{\varepsilon}\right)^{\frac{3}{2}}\lambda_{\boldsymbol{w}}\right)^2}\\
&=16\left(\frac{1}{27}\right)^{\frac{1}{2}}\frac{\varepsilon}{\nu}\frac{1}{\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2}.
\end{aligned}$$

As $\gamma<1$, then we may choose m large enough, such that

$$\begin{aligned}
m\nu\varepsilon&\geq\frac{384d^2n^2\left(\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda^2}\\
&\geq\max\left\{\frac{384d^2n^2\left(\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda^2},\frac{384d^2n^2\left(\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{16\left(\frac{1}{27}\right)^{\frac{1}{2}}\lambda^2}\right\}\\
&\geq\max\left\{\left(\frac{\nu}{\varepsilon}\right)^2,\frac{\varepsilon}{\nu}\right\}\frac{d^2n^2\left(\log\frac{8m(d+1)}{\delta}\right)^{\frac{3}{2}}\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)^2}.
\end{aligned}$$

Then for any $t \in [0, t^*)$,

$$\left\| \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0)) \right\|_{\mathrm{F}} \leq \frac{1}{8}\nu^2\varepsilon^2 \left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}} \right). \tag{B.18}$$

To sum up, for $t \in [0, t^*)$, the following holds

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m}{n}\nu^2\varepsilon^2 \left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}} \right) t \right) R_S\left(\boldsymbol{\theta}^0\right).$$

Suppose we have $t^* < +\infty$, then one can take the limit $t \to t^*$ in (B.17) and (B.18), which leads to a contradiction with the definition of t^*. Therefore $t^* = +\infty$.

Directly from Proposition 1, we have with probability at least $1 - \frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$\left\| \boldsymbol{\theta}^0_{\boldsymbol{w}} \right\|_2 \geq \sqrt{\frac{md}{2}},$$

thus we have

$$\sup_{t \in [0, +\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \leq \sqrt{\frac{2}{md}} \sup_{t \in [0, +\infty)} \|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2,$$

via Proposition 4, with probability at least $1 - \frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{k \in [m]} \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2\max\left\{ \frac{\nu}{\varepsilon}, 1 \right\} \xi p,$$

then

$$\begin{aligned}
\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2 &\leq \left[\sum_{k=1}^{m} \left(\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty^2 \right) \right]^{\frac{1}{2}} \\
&\leq 2\max\left\{ \frac{\nu}{\varepsilon}, 1 \right\} \sqrt{md}\,\xi p \\
&\leq 2\max\left\{ \frac{\nu}{\varepsilon}, 1 \right\} \sqrt{md} \sqrt{2\log\frac{8m(d+1)}{\delta}} \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}} \right)} \\
&\leq 8\max\left\{ \frac{\nu}{\varepsilon}, 1 \right\} \sqrt{\log\frac{8m(d+1)}{\delta}} \frac{d^{\frac{3}{2}}n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\sqrt{m}\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}} \right)},
\end{aligned}$$

hence

$$\begin{aligned}
\sup_{t \in [0, +\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} &\leq \sqrt{\frac{2}{md}} \sup_{t \in [0, +\infty)} \|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2 \\
&\leq 8\sqrt{2}\max\left\{ \frac{\nu}{\varepsilon}, 1 \right\} \sqrt{\log\frac{8m(d+1)}{\delta}} \frac{dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a + \frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}} \right)}
\end{aligned}$$

$$\leq 4\sqrt{2}\sqrt{\log\frac{8m(d+1)}{\delta}}\frac{dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\lambda},$$

as $\gamma < 1$, then we obtain that

$$\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}\lesssim\frac{\sqrt{\log\frac{8m(d+1)}{\delta}}}{m^{1-\gamma}}, \tag{B.19}$$

hence

$$\lim_{m\to\infty}\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}=0. \tag{B.20}$$

Similarly, directly from Proposition 1, we have with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$\left\|\boldsymbol{\theta}^0\right\|_2\geq\sqrt{\frac{m(d+1)}{2}},$$

thus we have

$$\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2}\leq\sqrt{\frac{2}{m(d+1)}}\sup_{t\in[0,+\infty)}\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2,$$

via Proposition 4, with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{k\in[m]}|a_k(t)-a_k(0)|\leq 2\max\left\{\frac{\varepsilon}{\nu},1\right\}\xi p,$$

$$\max_{k\in[m]}\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty\leq 2\max\left\{\frac{\nu}{\varepsilon},1\right\}\xi p,$$

then

$$\begin{aligned}
\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2&\leq\left[\sum_{k=1}^m\left(|a_k(t)-a_k(0)|^2+\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty^2\right)\right]^{\frac{1}{2}}\\
&\leq 2\max\left\{\frac{\varepsilon}{\nu},\frac{\nu}{\varepsilon}\right\}\sqrt{m(d+1)}\xi p\\
&\leq 2\max\left\{\frac{\varepsilon}{\nu},\frac{\nu}{\varepsilon}\right\}\sqrt{m(d+1)}\sqrt{2\log\frac{8m(d+1)}{\delta}}\frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\\
&\leq 8\max\left\{\frac{\varepsilon}{\nu},\frac{\nu}{\varepsilon}\right\}\sqrt{\log\frac{8m(d+1)}{\delta}}\frac{\sqrt{d+1}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\sqrt{m}\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)},
\end{aligned}$$

hence

$$\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2}\leq\sqrt{\frac{2}{m(d+1)}}\sup_{t\in[0,+\infty)}\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2$$

$$\leq 8\sqrt{2}\max\left\{\frac{\varepsilon}{\nu},\frac{\nu}{\varepsilon}\right\}\sqrt{\log\frac{8m(d+1)}{\delta}}\frac{dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}$$
$$\leq 8\sqrt{2}\sqrt{\log\frac{8m(d+1)}{\delta}}\frac{dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\lambda},$$

as $\gamma < 1$, then we obtain that

$$\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2}\lesssim\frac{\sqrt{\log\frac{8m(d+1)}{\delta}}}{m^{1-\gamma}}, \tag{B.21}$$

moreover,

$$\lim_{m\to\infty}\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}(0)\|_2}{\|\boldsymbol{\theta}(0)\|_2}=0. \tag{B.22}$$

□

B.3 $\boldsymbol{w}$-lazy Regime

We denote

$$t_a^*=\inf\{t\mid\boldsymbol{\theta}(t)\notin\mathcal{N}\left(\boldsymbol{\theta}^0\right)\}, \tag{B.23}$$

where the event is defined as

$$\mathcal{N}_a\left(\boldsymbol{\theta}^0\right):=\left\{\boldsymbol{\theta}\mid\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta})-\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)\|_{\mathrm{F}}\leq\frac{1}{4}\nu\varepsilon^3\lambda_a\right\}. \tag{B.24}$$

We observe immediately that the event $\mathcal{N}_a\left(\boldsymbol{\theta}^0\right)\neq\varnothing$, since $\boldsymbol{\theta}^0\in\mathcal{N}_a\left(\boldsymbol{\theta}^0\right)$. Recall that

$$\lambda_a=\lambda_{\min}\left(\boldsymbol{K}^{[a]}\right),$$

whose definition can be found in Theorem 4.

Proposition 5. *Given any $\delta\in(0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if*

$$m=\Omega\left(\frac{n^2}{\lambda_a^2}\log\frac{4n^2}{\delta}\right),$$

then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have for any time $t\in[0,t_a^*)$,

$$R_S(\boldsymbol{\theta}(t))\leq\exp\left(-\frac{m}{n}\nu\varepsilon^3\lambda_a t\right)R_S\left(\boldsymbol{\theta}^0\right). \tag{B.25}$$

Proof. Similar to the proof in Proposition 3, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and for any $\boldsymbol{\theta} \in \mathcal{N}\left(\boldsymbol{\theta}^0\right)$, we have

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta})\right) &\geq \lambda_{\min}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta})\right) \\
&\geq \lambda_{\min}\left(\boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)\right) - \|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}) - \boldsymbol{G}^{[a]}\left(\boldsymbol{\theta}^0\right)\|_{\mathrm{F}} \\
&\geq \frac{1}{2}\nu^3 \varepsilon \lambda_a.
\end{aligned}$$

Finally, we obtain that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} R_S(\boldsymbol{\theta}(t)) &= -\frac{m}{n^2}\boldsymbol{e}^\intercal \boldsymbol{G}(\boldsymbol{\theta}(t))\boldsymbol{e} \\
&\leq -\frac{2m}{n}\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta}(t))\right) R_S(\boldsymbol{\theta}(t)) \\
&\leq -\frac{m}{n}\nu^3 \varepsilon \lambda_a R_S(\boldsymbol{\theta}(t)),
\end{aligned}$$

and immediate integration yields the result. □

Proposition 6. *Given any $\delta \in (0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if $\gamma > 1$, $\gamma' > \gamma - 1$, and*

$$m = \max\left(\Omega\left(\frac{n^2}{\lambda_a^2}\log\frac{8n^2}{\delta}\right), \Omega\left(\frac{n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda_a}\right)^{\frac{1}{1-\gamma+\gamma'}}\right),$$

then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, for any time $t \in [0, t_a^*)$ and for any $k \in [m]$, both

$$\begin{aligned}
\max_{k\in[m]}|a_k(t) - a_k(0)| &\leq 2\frac{\varepsilon}{\nu}\sqrt{2\log\frac{4m(d+1)}{\delta}}\, p_a, \\
\max_{k\in[m]}\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty &\leq 2\sqrt{2\log\frac{4m(d+1)}{\delta}}\, p_a,
\end{aligned} \tag{B.26}$$

and

$$\max_{k\in[m]}\{|a_k(0)|,\ \|\boldsymbol{w}_k(0)\|_\infty\} \leq \sqrt{2\log\frac{4m(d+1)}{\delta}}, \tag{B.27}$$

hold, where

$$p_a := \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a}.$$

Proof. Since

$$\alpha(t) = \max_{k\in[m],s\in[0,t]} |a_k(s)|\,, \quad \omega(t) = \max_{k\in[m],s\in[0,t]} \|\boldsymbol{w}_k(s)\|_\infty\,,$$

we obtain

$$\begin{aligned}
|\nabla_{a_k} R_S|^2 &= \left|\frac{1}{n}\sum_{i=1}^n e_i \nu \sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right|^2 \leq 2\|\varepsilon \boldsymbol{w}_k\|_1^2 \nu^2 R_S(\boldsymbol{\theta}) \leq 2d^2(\omega(t))^2\nu^2\varepsilon^2 R_S(\boldsymbol{\theta}),\\
\|\nabla_{\boldsymbol{w}_k} R_S\|^2 &= \left\|\frac{1}{n}\sum_{i=1}^n e_i \nu\varepsilon a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i\right\|_\infty^2 \leq 2|a_k|^2\nu^2\varepsilon^2 R_S(\boldsymbol{\theta}) \leq 2(\alpha(t))^2\nu^2\varepsilon^2 R_S(\boldsymbol{\theta}).
\end{aligned}$$

By Proposition 5, we have if

$$m \geq \frac{16n^2d^2C_{\psi,d}^2}{\lambda^2 C_0}\log\frac{8n^2}{\delta},$$

then with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, we have that

$$\begin{aligned}
|a_k(t)-a_k(0)| &\leq \frac{\varepsilon}{\nu}\int_0^t |\nabla_{a_k} R_S(\boldsymbol{\theta}(s))|\,\mathrm{d}s\\
&\leq \sqrt{2}d\varepsilon^2\int_0^t \omega(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s\\
&\leq \sqrt{2}d\varepsilon^2\omega(t)\int_0^t \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s\\
&\leq \sqrt{2}d\varepsilon^2\omega(t)\int_0^\infty \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s\\
&= \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\omega(t)\\
&\leq \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\nu\varepsilon\lambda_a}\omega(t)\\
&\leq \frac{\varepsilon}{\nu}p_a\omega(t),
\end{aligned}$$

and similarly

$$
\begin{aligned}
\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty &\le \frac{\nu}{\varepsilon}\int_0^t \|\nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}(s))\|_\infty \,\mathrm{d}s \\
&\le \sqrt{2}\nu^2 \int_0^t \alpha(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\le \sqrt{2}\nu^2\alpha(t)\int_0^t \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s \\
&\le \sqrt{2}\nu^2\alpha(t)\int_0^\infty \sqrt{R_S\left(\boldsymbol{\theta}^0\right)}\exp\left(-\frac{m}{2n}\nu^2\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)t\right)\mathrm{d}s \\
&= \frac{2\sqrt{2}n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\left(\frac{\varepsilon}{\nu}\lambda_a+\frac{\nu}{\varepsilon}\lambda_{\boldsymbol{w}}\right)}\alpha(t) \\
&\le \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\frac{\varepsilon^3}{\nu}\lambda_a}\alpha(t) \\
&\le \frac{\nu}{\varepsilon}p_a\alpha(t).
\end{aligned}
$$

Thus

$$
\begin{aligned}
\alpha(t) &\le \alpha(0)+\frac{\varepsilon}{\nu}p_a\omega(t), \\
\omega(t) &\le \omega(0)+\frac{\nu}{\varepsilon}p_a\alpha(t),
\end{aligned}
$$

moreover, by Lemma 1, with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$
\max_{k\in[m]}\{|a_k(0)|,\ \|\boldsymbol{w}_k(0)\|_\infty\} \le \sqrt{2\log\frac{4m(d+1)}{\delta}}.
$$

Then, if

$$
m \ge \left(\frac{4\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda_a}\right)^{\frac{1}{1-\gamma+\gamma'}},
$$

we have

$$
p_a = \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a} \le \frac{1}{2}.
$$

Therefore

$$
\alpha(t) \le 2\max\left\{\frac{\varepsilon}{\nu},1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}} = 2\frac{\varepsilon}{\nu}\sqrt{2\log\frac{4m(d+1)}{\delta}}.
$$

Similarly, one can obtain the estimate of $\omega(t)$ as

$$
\omega(t) \le 2\max\left\{\frac{\nu}{\varepsilon},1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}} = 2\sqrt{2\log\frac{4m(d+1)}{\delta}}.
$$

Hence, for any $t \in [0, t_a^*)$, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{k\in[m]} |a_k(t) - a_k(0)| \leq 2\frac{\varepsilon}{\nu}\sqrt{2\log\frac{4m(d+1)}{\delta}}p_a,$$
$$\max_{k\in[m]} \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2\sqrt{2\log\frac{4m(d+1)}{\delta}}p_a.$$

□

Theorem 6. *Given any $\delta \in (0,1)$, under Assumption 2, Assumption 4 and Assumption 5, if $\gamma \geq 1$, $\gamma' > \gamma - 1$, and*

$$m = \max\left(\Omega\left(\frac{n^2}{\lambda_a^2}\log\frac{16n^2}{\delta}\right), \Omega\left(\left(\frac{n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda_a}\right)^{\frac{1}{1-\gamma+\gamma'}}\right),\right.$$
$$\left.\Omega\left(\left(\frac{n^2\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda_a^2}\right)^{\frac{1}{1-\gamma+\gamma'}}\right), \Omega\left(\log\frac{8}{\delta}\right)\right), \tag{B.28}$$

then with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, we have for all time $t > 0$,

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m\nu\varepsilon^3\lambda_a t}{n}\right) R_S\left(\boldsymbol{\theta}^0\right), \tag{B.29}$$

and with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\lim_{m\to\infty} \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = 0. \tag{B.30}$$

Proof. According to Proposition 6, it suffices to show that $t_a^* = \infty$.

(i). Firstly, from Proposition 5, we have with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$, for any $t \in [0, t_a^*)$, the following holds

$$|a_k(t) - a_k(0)| \leq 2\frac{\varepsilon}{\nu}\xi p_a,$$
$$\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2\xi p_a,$$

where

$$p_a = \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a},$$

and

$$\xi = \sqrt{2\log\frac{8m(d+1)}{\delta}}.$$

For m large enough, i.e.,

$$m \geq \left(\frac{4\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda_a} \right)^{\frac{1}{1-\gamma+\gamma'}},$$

we have

$$p_a = \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a} \leq \frac{1}{2}.$$

We inherit the proof in Theorem 5 and obtain that

$$\begin{aligned}
\left| G_{ij}^{[a]}(\boldsymbol{\theta}(t)) - G_{ij}^{[a]}(\boldsymbol{\theta}(0)) \right| &\leq \frac{\nu\varepsilon^3}{m} \sum_{k=1}^{m} \left| g_{ij}^{[a]}(\boldsymbol{w}_k(t)) - g_{ij}^{[a]}(\boldsymbol{w}_k(0)) \right| \\
&\leq 8d^2\nu^2\varepsilon^2\xi^2 \max\left\{ \frac{\nu}{\varepsilon}, \frac{\varepsilon}{\nu} \right\} p_a \\
&\leq 8d^2\nu\varepsilon^3\xi^2 p_a,
\end{aligned}$$

by using the same technique in Proposition 2,

$$\begin{aligned}
&\left\| \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t)) - \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0)) \right\|_{\mathrm{F}} \\
&\quad \leq 8d^2 n\nu\varepsilon^3 \left(2\log\frac{8m(d+1)}{\delta} \right) \frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a} \\
&\leq \nu\varepsilon^3 \frac{32\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a}.
\end{aligned}$$

As $\gamma \geq 1$, and $\gamma' > \gamma - 1$, then we may choose m large enough, such that

$$m\varepsilon^2 \geq \frac{128\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\lambda_a^2},$$

then for any $t \in [0, t_a^*)$,

$$\left\| \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t)) - \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0)) \right\|_{\mathrm{F}} \leq \frac{1}{4}\nu\varepsilon^3\lambda_a. \tag{B.31}$$

Hence, for $t \in [0, t_a^*)$, the following holds

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m}{n}\nu\varepsilon^3\lambda_a t \right) R_S\left(\boldsymbol{\theta}^0\right).$$

Suppose we have $t_a^* < +\infty$, then one can take the limit $t \to t_a^*$ in (B.31), which leads to a contradiction with the definition of t_a^*. Therefore $t_a^* = +\infty$.

Directly from Proposition 1, we have with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$\left\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\right\|_2 \geq \sqrt{\frac{md}{2}},$$

thus we have

$$\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \leq \sqrt{\frac{2}{md}} \sup_{t\in[0,+\infty)} \left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2,$$

via Proposition 6, with probability at least $1-\frac{\delta}{2}$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{k\in[m]} \left\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\right\|_\infty \leq 2\xi p_a,$$

then

$$\begin{aligned}
\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2 &\leq \left[\sum_{k=1}^m \left(\left\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\right\|_\infty^2\right)\right]^{\frac{1}{2}} \\
&\leq 2\sqrt{md}\,\xi p_a \\
&\leq 2\sqrt{md}\sqrt{2\log\frac{8m(d+1)}{\delta}}\,\frac{2\sqrt{2}dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a} \\
&\leq 8\sqrt{\log\frac{8m(d+1)}{\delta}}\,\frac{d^{\frac{3}{2}}n\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{\sqrt{m}\varepsilon^2\lambda_a},
\end{aligned}$$

hence

$$\begin{aligned}
\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} &\leq \sqrt{\frac{2}{md}} \sup_{t\in[0,+\infty)} \left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2 \\
&\leq 8\sqrt{2}\sqrt{\log\frac{8m(d+1)}{\delta}}\,\frac{dn\sqrt{R_S\left(\boldsymbol{\theta}^0\right)}}{m\varepsilon^2\lambda_a},
\end{aligned}$$

as $\gamma \geq 1$, and $\gamma' > \gamma - 1$, then we obtain that

$$\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \lesssim \frac{\sqrt{\log\frac{8m(d+1)}{\delta}}}{m^{1-\gamma+\gamma'}}, \tag{B.32}$$

hence

$$\lim_{m\to\infty} \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = 0. \tag{B.33}$$

□

C Condensed Regime

C.1 Effective Linear Dynamics

As the normalized flow reads

$$\begin{aligned}\frac{\mathrm{d}a_k}{\mathrm{d}t} &= \frac{\varepsilon}{\nu}\left(-\frac{1}{n}\sum_{i=1}^{n} e_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}\right),\\ \frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &= \frac{\nu}{\varepsilon}\left(-\frac{1}{n}\sum_{i=1}^{n} e_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i\right),\end{aligned} \tag{C.1}$$

since $e_i \approx -y_i$, and by means of perturbation expansion with respect to ε and keep the order 1 term, we obtain that

$$\begin{aligned}\frac{\mathrm{d}a_k}{\mathrm{d}t} &\approx \frac{\varepsilon}{\nu}\frac{1}{n}\sum_{i=1}^{n} y_i \sigma^{(1)}(0)\boldsymbol{w}_k^\intercal \boldsymbol{x}_i = \frac{\varepsilon}{\nu}\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i,\\ \frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &\approx \frac{\nu}{\varepsilon}\frac{1}{n}\sum_{i=1}^{n} y_i a_k \sigma^{(1)}(0)\boldsymbol{x}_i = \frac{\nu}{\varepsilon}\frac{1}{n}\sum_{i=1}^{n} y_i a_k \boldsymbol{x}_i,\end{aligned}$$

so the normalized flow approximately reads

$$\begin{aligned}\frac{\mathrm{d}a_k}{\mathrm{d}t} &= \frac{\varepsilon}{\nu}\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i = \frac{\varepsilon}{\nu}\boldsymbol{w}_k^\intercal \boldsymbol{z},\\ \frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &= \frac{\nu}{\varepsilon}\frac{1}{n}\sum_{i=1}^{n} y_i a_k \boldsymbol{x}_i = \frac{\nu}{\varepsilon}a_k \boldsymbol{z}.\end{aligned} \tag{C.2}$$

We observe that (C.2) reveals that the training dynamics of two-layer NNs at initial stage has a close relationship to power iteration of a matrix that only depends on the input sample $\mathcal{S}$. We denote by

$$\boldsymbol{A} := \begin{bmatrix} 0 & \boldsymbol{z}^\intercal \\ \boldsymbol{z} & \boldsymbol{0}_{d\times d}\end{bmatrix}, \tag{C.3}$$

where

$$\boldsymbol{z} = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i,$$

whose definition can be traced back to Assumption 3, and (C.2) can be written into

$$\frac{\mathrm{d}\boldsymbol{q}_k}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{q}_k, \tag{C.4}$$

where

$$\boldsymbol{q}_k := \left[\frac{\nu}{\varepsilon} a_k, \boldsymbol{w}_k\right]^\mathsf{T}.$$

Moreover, simple linear algebra shows that $\boldsymbol{A}$ has two nonzero eigenvalues $\lambda_1 = \|\boldsymbol{z}\|_2$ and $\lambda_2 = -\|\boldsymbol{z}\|_2$. Moreover, the unit eigenvector for $\lambda_1 = \|\boldsymbol{z}\|_2$ is

$$\boldsymbol{u}_1 := \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \hat{\boldsymbol{z}}^\mathsf{T}\right]^\mathsf{T},$$

where

$$\hat{\boldsymbol{z}} = \frac{\sum_{i=1}^n y_i \boldsymbol{x}_i}{\left\|\sum_{i=1}^n y_i \boldsymbol{x}_i\right\|_2}.$$

whose definition can be traced back to Assumption 3. We obtain further that the unit eigenvector for $\lambda_2 = -\|\boldsymbol{z}\|_2$ is

$$\boldsymbol{u}_2 := \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \hat{\boldsymbol{z}}^\mathsf{T}\right]^\mathsf{T},$$

and $\langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle = 0$.

We also observe that the rest of the eigenvalues for $\boldsymbol{A}$ is all zero, i.e.,

$$\lambda_3 = \lambda_4 = \cdots = \lambda_{d+1} = 0,$$

and their eigenvectors $\{\boldsymbol{u}_k\}_{k=3}^{d+1}$ read,

$$\boldsymbol{u}_k := \left[0, \boldsymbol{b}_k^\mathsf{T}\right]^\mathsf{T}, \quad \boldsymbol{b}_k \in \boldsymbol{z}^\perp,$$

and

$$\operatorname{span}\{\boldsymbol{b}_k\}_{k=3}^{d+1} = \boldsymbol{z}^\perp,$$

whose first component is zero, and the rest of the components spans the orthogonal complement of $\boldsymbol{z}$, i.e.,

$$\boldsymbol{z}^\perp = \{\boldsymbol{x} \in \mathbb{R}^d \mid \langle \boldsymbol{z}, \boldsymbol{x} \rangle = 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^d\},$$

since

$$\dim\left(\operatorname{span}\{\boldsymbol{b}_k\}_{k=3}^{d+1}\right) = \dim\left(\boldsymbol{z}^\perp\right) = d - 1.$$

We hereafter denote that: For $t \geq 0$,

$$r(t) := \exp\left(\frac{1}{2}\|\boldsymbol{z}\|_2 t\right),$$

and since

$$\frac{1}{c} \leq \|\boldsymbol{z}\|_2 \leq c,$$

for some universal constant $c > 0$, then we obtain that for some universal constants $c_1 > 0$ and $c_2 > 0$,

$$\exp(c_1 t) \leq r(t) \leq \exp(c_2 t).$$

Proposition 7. *The solution to the linear differential equation*

$$\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{q}, \quad \boldsymbol{q}(0) = \boldsymbol{q}^0, \tag{C.5}$$

where

$$\boldsymbol{q}(t) := [a(t), (\boldsymbol{w}(t))^\intercal]^\intercal, \quad \boldsymbol{q}^0 := \left[a^0, \left(\boldsymbol{w}^0\right)^\intercal\right]^\intercal,$$

reads

$$\begin{aligned} a(t) &= \left(\frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t)\right) a^0 + \left(\frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t)\right) \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle, \\ \boldsymbol{w}(t) &= \left(\frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t)\right) a^0 \hat{\boldsymbol{z}} + \left(\frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t)\right) \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle \hat{\boldsymbol{z}} \\ &\quad - \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle \hat{\boldsymbol{z}} + \boldsymbol{w}^0, \end{aligned} \tag{C.6}$$

where $r(t) = \exp\left(\frac{1}{2}\|\boldsymbol{z}\|_2 t\right)$.

Proof. We only need to solve out the matrix exponential for $\boldsymbol{A}$, as $\boldsymbol{A}$ is symmetric, then it can be diagonalized by an orthogonal matrix $\boldsymbol{P}$, where

$$\boldsymbol{P} := [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{d+1}],$$

and

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^\intercal,$$

where

$$\boldsymbol{J} := \begin{bmatrix} \|\boldsymbol{z}\|_2 & & & \\ & -\|\boldsymbol{z}\|_2 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix},$$

since

$$\exp(t\boldsymbol{A}) = \boldsymbol{P}\exp(t\boldsymbol{J})\boldsymbol{P}^\intercal,$$

then

$$\begin{pmatrix} a(t) \\ \boldsymbol{w}(t) \end{pmatrix} = \exp(t\boldsymbol{A}) \begin{pmatrix} a^0 \\ \boldsymbol{w}^0 \end{pmatrix},$$

thus, we obtain that

$$\begin{aligned} a(t) = &\left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) + \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right) a^0 \\ &+ \left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) - \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right) \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle, \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{w}(t) = &\left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) - \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right) a^0 \hat{\boldsymbol{z}} \\
&+ \left(\frac{1}{2}\exp(\|\boldsymbol{z}\|_2 t) + \frac{1}{2}\exp(-\|\boldsymbol{z}\|_2 t)\right) \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle \hat{\boldsymbol{z}} \\
&- \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle \hat{\boldsymbol{z}} + \boldsymbol{w}^0.
\end{aligned}
$$

□

Remark 9. *It is noteworthy that* $\boldsymbol{w}$ *has two components, one is the projection of* $\boldsymbol{w}^0$ *into the direction of* $\boldsymbol{z}$*:*

$$
\left(\frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t)\right) a^0 \hat{\boldsymbol{z}} + \left(\frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t)\right) \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle \hat{\boldsymbol{z}},
$$

which evolves with respect to t*. The other is the projection of* $\boldsymbol{w}^0$ *onto* $\boldsymbol{z}^\perp$*:*

$$
\boldsymbol{w}^0 - \left\langle \boldsymbol{w}^0, \hat{\boldsymbol{z}} \right\rangle \hat{\boldsymbol{z}},
$$

which remains frozen as t *evolves.*

C.2 Difference between Real and Linear Dynamics

We observe that the real dynamics can be written into

$$
\begin{aligned}
\frac{\nu}{\varepsilon}\frac{\mathrm{d}a_k}{\mathrm{d}t} &= \left(-\frac{1}{n}\sum_{i=1}^n e_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} - \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i + \boldsymbol{w}_k^\intercal \boldsymbol{z}\right), \\
\frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t} &= \frac{\nu}{\varepsilon}\left(-\frac{1}{n}\sum_{i=1}^n e_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i - \frac{1}{n}\sum_{i=1}^n y_i a_k \boldsymbol{x}_i + a_k \boldsymbol{z}\right),
\end{aligned}
\tag{C.7}
$$

hence the difference between the real and linear dynamics is characterized by $\{f_k, \boldsymbol{g}_k\}_{k=1}^m$, where

$$
\begin{aligned}
f_k &:= \frac{1}{n}\sum_{i=1}^n \left(e_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} + y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i\right), \\
\boldsymbol{g}_k &:= \frac{1}{n}\sum_{i=1}^n \left(e_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i + y_i a_k \boldsymbol{x}_i\right).
\end{aligned}
$$

In other words, the real dynamics can be written into

$$
\frac{\mathrm{d}\begin{pmatrix}\frac{\nu}{\varepsilon}a_k \\ \boldsymbol{w}_k\end{pmatrix}}{\mathrm{d}t} = \boldsymbol{A}\begin{pmatrix}\frac{\nu}{\varepsilon}a_k \\ \boldsymbol{w}_k\end{pmatrix} + \begin{pmatrix}f_k \\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix}, \quad \begin{pmatrix}\frac{\nu}{\varepsilon}a_k(0) \\ \boldsymbol{w}_k(0)\end{pmatrix} = \begin{pmatrix}\frac{\nu}{\varepsilon}a_k^0 \\ \boldsymbol{w}_k^0\end{pmatrix},
\tag{C.8}
$$

and its solution reads

$$\begin{pmatrix} \frac{\nu}{\varepsilon} a_k \\ \boldsymbol{w}_k \end{pmatrix} = \exp(t\boldsymbol{A}) \begin{pmatrix} \frac{\nu}{\varepsilon} a_k^0 \\ \boldsymbol{w}_k^0 \end{pmatrix} + \int_0^t \exp((t-s)\boldsymbol{A}) \begin{pmatrix} f_k(s) \\ \frac{\nu}{\varepsilon} \boldsymbol{g}_k(s) \end{pmatrix} \mathrm{d}s. \tag{C.9}$$

C.3 $\boldsymbol{w}$-lag regime

Definition 3 (Neuron 2-energy, $\boldsymbol{w}$-lag regime)**.** *In real dynamics, we define the 2-energy at time t for each single neuron, i.e., for each $k \in [m]$,*

$$q_k(t) := \left(\frac{\nu^2}{\varepsilon^2} |a_k(t)|^2 + \|\boldsymbol{w}_k(t)\|_2^2 \right)^{\frac{1}{2}}. \tag{C.10}$$

We denote

$$q_{\max}(t) := \max_{k \in [m]} q_k(t). \tag{C.11}$$

For simplicity, we hereafter drop the (t)s for all $q_k(t)$ and $q_{\max}(t)$. Then the estimates on $\{f_k, \boldsymbol{g}_k\}_{k=1}^m$ read

Proposition 8. *For any time $t > 0$,*

$$\begin{aligned} |f_k| &\leq m\varepsilon^2 q_{\max}^2 \|\boldsymbol{w}_k\|_2 + \varepsilon \|\boldsymbol{w}_k\|_2^2, \\ \|\boldsymbol{g}_k\|_2 &\leq m\varepsilon^2 q_{\max}^2 |a_k| + \varepsilon \|\boldsymbol{w}_k\|_2 |a_k|. \end{aligned} \tag{C.12}$$

Proof. We obtain that

$$\begin{aligned} |f_k| &= \left| \frac{1}{n} \sum_{i=1}^n \left((e_i + y_i) \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} + y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i - y_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} \right) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \left(\left(\sum_{k=1}^m \nu a_k \sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \right) \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} + y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i - y_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} \right) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left(\left(\sum_{k=1}^m \nu\varepsilon |a_k| \|\boldsymbol{w}_k\|_2 \right) \|\boldsymbol{w}_k\|_2 + \varepsilon (\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)^2 \right) \\ &\leq \frac{1}{2n} \sum_{i=1}^n \left(\nu\varepsilon \left(\sum_{k=1}^m \left(\frac{\nu}{\varepsilon} |a_k|^2 + \frac{\varepsilon}{\nu} \|\boldsymbol{w}_k\|_2^2 \right) \right) \|\boldsymbol{w}_k\|_2 + \varepsilon (\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)^2 \right) \\ &\leq \frac{1}{2n} \sum_{i=1}^n \left(\left(\sum_{k=1}^m \varepsilon^2 q_k^2 \right) \|\boldsymbol{w}_k\|_2 + \varepsilon (\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)^2 \right) \\ &\leq \frac{1}{2n} \sum_{i=1}^n \left(\left(\sum_{k=1}^m \varepsilon^2 q_k^2 \right) \|\boldsymbol{w}_k\|_2 + \varepsilon \|\boldsymbol{w}_k\|_2^2 \right) \end{aligned}$$

$$\leq m\varepsilon^2 q_{\max}^2 \left\| \boldsymbol{w}_k \right\|_2 + \varepsilon \left\| \boldsymbol{w}_k \right\|_2^2,$$

and

$$\begin{aligned}
\left\| \boldsymbol{g}_k \right\|_2 &= \left\| \frac{1}{n} \sum_{i=1}^{n} \left((e_i + y_i) a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i) \boldsymbol{x}_i + y_i a_k \boldsymbol{x}_i - y_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i) \boldsymbol{x}_i \right) \right\|_2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\left(\sum_{k=1}^{m} \nu a_k \sigma(\varepsilon \boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i) \right) a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i) \boldsymbol{x}_i + y_i a_k \boldsymbol{x}_i - y_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i) \boldsymbol{x}_i \right) \right\|_2 \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left(\left(\sum_{k=1}^{m} \nu \varepsilon \left| a_k \right| \left\| \boldsymbol{w}_k \right\|_2 \right) \left| a_k \right| + \varepsilon \left\| \boldsymbol{w}_k \right\|_2 \left| a_k \right| \right) \\
&\leq \frac{1}{2n} \sum_{i=1}^{n} \left(\left(\sum_{k=1}^{m} \varepsilon^2 q_k^2 \right) \left| a_k \right| + \varepsilon \left\| \boldsymbol{w}_k \right\|_2 \left| a_k \right| \right) \\
&\leq m\varepsilon^2 q_{\max}^2 \left| a_k \right| + \varepsilon \left\| \boldsymbol{w}_k \right\|_2 \left| a_k \right|.
\end{aligned}$$

□

We denote a useful quantity

$$\phi(t) := \sup_{0 \leq s \leq t} q_{\max}(s). \tag{C.13}$$

Then directly from Lemma 1, we have with probability at least $1 - \delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{k \in [m]} \left\{ |a_k^0|, \ \|\boldsymbol{w}_k^0\|_\infty \right\} \leq \sqrt{2 \log \frac{2m(d+1)}{\delta}}, \tag{C.14}$$

hence

$$\phi(0) \leq \sqrt{2(d+1) \log \frac{2m(d+1)}{\delta}}. \tag{C.15}$$

We define

$$T_{\text{eff}} := \inf \left\{ t > 0 \mid m\varepsilon^2 \phi^3(t) > m^{-\tau}, \quad \tau = \frac{\gamma - \gamma' - 1}{4} \right\}, \tag{C.16}$$

then for m large enough, as $\gamma - \gamma' > 1$, based on (C.15),

$$m\varepsilon^2 \phi^3(0) \leq m\varepsilon^2 \left(2(d+1) \log \frac{2m(d+1)}{\delta} \right)^{\frac{3}{2}} \leq m^{-\frac{\gamma - \gamma' - 1}{2}}, \tag{C.17}$$

hence $T_{\text{eff}} \geq 0$.

We observe further that by taking the 2-norm on both sides of (C.9), the following holds

$$q_k(t) \leq \exp\left(t \left\|\boldsymbol{A}\right\|_{2\to 2}\right) q_k(0) + \int_0^t \exp\left((t-s)\left\|\boldsymbol{A}\right\|_{2\to 2}\right)\left(m\varepsilon^2 q_{\max}^2(s) + \varepsilon q_{\max}(s)\right) q_k(s)\mathrm{d}s,$$

and by taking supreme over the index k and time $0 \leq t \leq T_{\mathrm{eff}}$ on both sides, and for large enough m, the following holds

$$\begin{aligned}\phi(t) &\leq \phi(0)\exp(t) + 2m^{-\min\left\{\frac{1}{2},\tau\right\}} \int_0^t \exp(t-s)\mathrm{d}s \\ &\leq \phi(0)\exp(t) + 2m^{-\min\left\{\frac{1}{2},\tau\right\}} \left(\exp(t) - 1\right) \\ &\leq \phi(0)\exp(t) + 2m^{-\min\left\{\frac{1}{2},\tau\right\}} \exp(t),\end{aligned} \tag{C.18}$$

then based on (C.15), we have with probability $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, for sufficiently large m,

$$\begin{aligned}\phi(t) &\leq \phi(0)\exp(t) + 2m^{-\min\left\{\frac{1}{2},\tau\right\}} \exp(t) \\ &\leq 2\phi(0)\exp(t) \leq 2\sqrt{2(d+1)\log\frac{2m(d+1)}{\delta}}\exp(t),\end{aligned} \tag{C.19}$$

we set t_0 as the time satisfying

$$2\sqrt{2(d+1)\log\frac{2m(d+1)}{\delta}}\exp(t_0) = \frac{1}{2}m^{\frac{\gamma-\gamma'-1}{4}}, \tag{C.20}$$

then we obtain that, for any $\eta_0 > \frac{\gamma-\gamma'-1}{100} > 0$,

$$T_{\mathrm{eff}} \geq t_0 > \log\left(\frac{1}{4}\right) + \left(\frac{\gamma-\gamma'-1}{4} - \eta_0\right)\log(m). \tag{C.21}$$

Theorem 7 (Condensed regime, $\boldsymbol{w}$-lag regime)**.** *Given any $\delta \in (0,1)$, under Assumption 1, Assumption 3 and Assumption 5, if $\gamma > 1$ and $0 \leq \gamma' < \gamma - 1$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{m\to+\infty} \sup_{t\in[0,T_{\mathrm{eff}}]} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = +\infty, \tag{C.22}$$

and

$$\lim_{m\to+\infty} \sup_{t\in[0,T_{\mathrm{eff}}]} \frac{\left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\right\|_2} = 1. \tag{C.23}$$

Proof. Since we have

$$\frac{\mathrm{d}\begin{pmatrix}\frac{\nu}{\varepsilon}a_k\\ \boldsymbol{w}_k\end{pmatrix}}{\mathrm{d}t}=\boldsymbol{A}\begin{pmatrix}\frac{\nu}{\varepsilon}a_k\\ \boldsymbol{w}_k\end{pmatrix}+\begin{pmatrix}f_k\\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix},\quad \begin{pmatrix}\frac{\nu}{\varepsilon}a_k(0)\\ \boldsymbol{w}_k(0)\end{pmatrix}=\begin{pmatrix}\frac{\nu}{\varepsilon}a_k^0\\ \boldsymbol{w}_k^0\end{pmatrix},$$

and its solution reads

$$\begin{pmatrix}\frac{\nu}{\varepsilon}a_k\\ \boldsymbol{w}_k\end{pmatrix}=\exp\left(t\boldsymbol{A}\right)\begin{pmatrix}\frac{\nu}{\varepsilon}a_k^0\\ \boldsymbol{w}_k^0\end{pmatrix}+\int_0^t\exp\left((t-s)\boldsymbol{A}\right)\begin{pmatrix}f_k\\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix}\mathrm{d}s.$$

As we notice that for any $k\in[m]$, $\begin{pmatrix}\frac{\nu}{\varepsilon}a_k\\ \boldsymbol{w}_k\end{pmatrix}$ can be written into two parts, the first one is the linear part, the second one is the residual part. For simplicity of proof, we need to introduce some further notations.

As we already identify the parameters $\boldsymbol{\theta}_a=\mathrm{vec}(\{a_k\}_{k=1}^m)$ and $\boldsymbol{\theta}_{\boldsymbol{w}}=\mathrm{vec}(\{\boldsymbol{w}_k\}_{k=1}^m)$, with some slight misuse of notations, we denote $\boldsymbol{\theta}_a:=\mathrm{vec}(\{\frac{\nu}{\varepsilon}a_k\}_{k=1}^m)$. From the observations above,

$$\boldsymbol{\theta}:=\mathrm{vec}\left(\boldsymbol{\theta}_a,\boldsymbol{\theta}_{\boldsymbol{w}}\right)=\mathrm{vec}\left(\left\{\frac{\nu}{\varepsilon}a_k\right\}_{k=1}^m,\{\boldsymbol{w}_k\}_{k=1}^m\right),$$

and $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_{\boldsymbol{w}}$ can be written into

$$\begin{aligned}\boldsymbol{\theta}_a(t)&=\bar{\boldsymbol{\theta}}_a(t)+\widetilde{\boldsymbol{\theta}}_a(t),\\ \boldsymbol{\theta}_{\boldsymbol{w}}(t)&=\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t),\end{aligned}$$

where the k-th component of $\bar{\boldsymbol{\theta}}_a$ and $\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}$ respectively reads

$$\begin{aligned}\left(\bar{\boldsymbol{\theta}}_a\right)_k&:=\frac{\nu}{\varepsilon}\left(\frac{1}{2}r^2(t)+\frac{1}{2}r^{-2}(t)\right)a_k^0+\left(\frac{1}{2}r^2(t)-\frac{1}{2}r^{-2}(t)\right)\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle,\\ \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}\right)_k&:=\frac{\nu}{\varepsilon}\left(\frac{1}{2}r^2(t)-\frac{1}{2}r^{-2}(t)\right)a_k^0\hat{\boldsymbol{z}}+\left(\frac{1}{2}r^2(t)+\frac{1}{2}r^{-2}(t)\right)\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\hat{\boldsymbol{z}}\\ &\quad-\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\hat{\boldsymbol{z}}+\boldsymbol{w}_k^0,\end{aligned}$$

and the k-th component of $\widetilde{\boldsymbol{\theta}}_a$ and $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}$ altogether reads

$$\begin{pmatrix}\left(\widetilde{\boldsymbol{\theta}}_a(t)\right)_k\\ \left(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right)_k\end{pmatrix}:=\int_0^t\exp\left((t-s)\boldsymbol{A}\right)\begin{pmatrix}f_k\\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix}\mathrm{d}s.$$

Moreover, we observe that $\boldsymbol{\theta}_{\boldsymbol{w}}$ can be decomposed into two parts, one is the projection of $\boldsymbol{w}^0$ into the direction of $\boldsymbol{z}$, i.e., $\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}$, and the other is the projection of

$\boldsymbol{w}^0$ onto $\boldsymbol{z}^\perp$, i.e., $\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}$. As $\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}$ inherits the same structure as $\boldsymbol{\theta}_{\boldsymbol{w}}$, we may apply the same decomposition to $\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}$. Hence, we obtain that

$$\begin{aligned}\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t) &= \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t),\\ \boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) &= \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t),\end{aligned}$$

and these relations concerning 2-norm hold simultaneously for any $t \geq 0$,

$$\begin{aligned}\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2^2 &= \|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2^2 + \left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2,\\ \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2^2 &= \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2 + \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2,\\ \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2^2 &= \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2 + \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2,\end{aligned}$$

and the k-th component of $\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}$ and $\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}$ altogether reads

$$\begin{aligned}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}\right)_k &:= \frac{\nu}{\varepsilon}\left(\frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t)\right)a_k^0\hat{\boldsymbol{z}} + \left(\frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t)\right)\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}},\\ \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}\right)_k &:= \boldsymbol{w}_k^0 - \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}},\end{aligned}$$

and finally, based on the relations concerning 2-norm above, for any $t \geq 0$,

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2 &\leq \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2,\\ \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2 &\leq \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2.\end{aligned}$$

We are hereby to prove (C.22). Firstly, we observe that

$$\boldsymbol{\theta}_{\boldsymbol{w}}(0) = \bar{\boldsymbol{\theta}}_{\boldsymbol{w}}(0),$$

hence

$$\begin{aligned}&\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2^2\\ &= \left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w}}(0)\right\|_2^2\\ &= \left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0) + \boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(0)\right\|_2\\ &= \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2 + \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(0) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2,\end{aligned}$$

by choosing $\eta_0 = \frac{\gamma-\gamma'-1}{8}$, then for time $0 \leq t \leq \bar{t}_0 := \left(\frac{\gamma-\gamma'-1}{8}\right)\log(m) - \log(2)$,

$$\left\|\int_0^t \exp\left((t-s)\boldsymbol{A}\right)\begin{pmatrix} f_k \\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix}\mathrm{d}s\right\|_2$$

$$
\begin{aligned}
&\leq \left(m\varepsilon^2\phi^3(t)+\varepsilon\phi^2(t)\right)\int_0^t \exp((t-s)\left\|\boldsymbol{z}\right\|_2)\mathrm{d}s\\
&\leq 2m^{-\min\left\{\tau,\frac{1}{2}\right\}}\int_0^t \exp((t-s))\mathrm{d}s\\
&\leq 2m^{-\frac{\gamma-\gamma'-1}{4}}\exp(t)\leq 2m^{-\frac{\gamma-\gamma'-1}{4}}\exp(\bar{t}_0)=m^{-\frac{\gamma-\gamma'-1}{8}}.
\end{aligned}
$$

We conclude that for $t\leq \bar{t}_0$, the following holds

$$
\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2\leq\sqrt{m}\left\|\int_0^t\exp\left((t-s)\boldsymbol{A}\right)\begin{pmatrix} f_k\\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix}\mathrm{d}s\right\|_2\leq\sqrt{m}m^{-\frac{\gamma-\gamma'-1}{8}},
$$

since the k-th component of $\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}$ reads

$$
\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)=\boldsymbol{w}_k^0-\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\hat{\boldsymbol{z}},
$$

we observe that since the RHS is independent of time t, hence we have

$$
\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)=\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(0),
$$

so we obtain that

$$
\begin{aligned}
&\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2\\
=&\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2+\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)\right\|_2^2\\
=&\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2+\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)\right\|_2^2.
\end{aligned}
$$

thus the ratio reads

$$
\begin{aligned}
\left(\frac{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}\right)^2&=\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2+\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)\right\|_2^2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2}\\
&=\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2}+\frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)\right\|_2^2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2}.
\end{aligned}
$$

As

$$
\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2=\left\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\right\|_2^2=\sum_{k=1}^m\left(\boldsymbol{w}_k^0\right)^2,
$$

and we observe that since $\boldsymbol{w}_k^0\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$, then $\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\sim\mathcal{N}(0,1)$. Moreover, $\left\{a_k^0,\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\right\}_{k=1}^m\sim\mathcal{N}(0,1)$ are i.i.d. Gaussian variables. Hence, by application of Theorem 3, with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for m large enough,

$$
\frac{d}{2}\leq\frac{1}{m}\left\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\right\|_2^2\leq\frac{3d}{2},
$$

and with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for m large enough,

$$\frac{1}{2} \leq \frac{1}{m}\sum_{k=1}^{m}\left(a_k^0\right)^2 \leq \frac{3}{2},$$

and with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for m large enough,

$$\frac{1}{2} \leq \frac{1}{m}\sum_{k=1}^{m}\left(\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle\right)^2 \leq \frac{3}{2},$$

and with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for m large enough,

$$-\frac{1}{4} \leq \frac{1}{m} a_k^0 \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \leq \frac{1}{4}.$$

To sum up, the ratio

$$\left(\frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2}\right)^2 = \frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2^2} + \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^{\perp}}(t)\right\|_2^2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2^2},$$

for the first part of the RHS:

$$\begin{aligned}
&\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \\
&\geq \frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)\right\|_2-\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \\
&= \frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} - \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2},
\end{aligned}$$

and

$$
\begin{aligned}
&\frac{1}{m}\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)\right\|_2^2\\
&=\sum_{k=1}^{m}\left[\frac{1}{2}\left(r^2(t)-r^{-2}(t)\right)\frac{\nu}{\varepsilon}a_k^0+\frac{1}{2}\left(r^2(t)+r^{-2}(t)-2\right)\left\langle \boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\right]^2\\
&=\frac{1}{4}\left(r(t)-r^{-1}(t)\right)^2\sum_{k=1}^{m}\left[\left(r(t)+r^{-1}(t)\right)\frac{\nu}{\varepsilon}a_k^0+\left(r(t)-r^{-1}(t)\right)\left\langle \boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\right]^2\\
&=\frac{1}{4}\left(r(t)-r^{-1}(t)\right)^2\sum_{k=1}^{m}\Bigg\{\left[\left(r(t)+r^{-1}(t)\right)^2\left(\frac{\nu}{\varepsilon}a_k^0\right)^2+\left(r(t)-r^{-1}(t)\right)^2\left(\left\langle \boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\right)^2\right]\\
&\qquad\qquad+2\left(r^2(t)-r^{-2}(t)\right)\frac{\nu}{\varepsilon}a_k^0\left\langle \boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\Bigg\}\\
&\geq\frac{1}{4}\left(r(t)-r^{-1}(t)\right)^2\Bigg[\left(r(t)+r^{-1}(t)\right)^2\left(\frac{\nu}{\varepsilon}\right)^2\frac{1}{2}+\left(r(t)-r^{-1}(t)\right)^2\frac{1}{2}\\
&\qquad\qquad-\left(r^2(t)-r^{-2}(t)\right)\frac{\nu}{\varepsilon}\frac{1}{2}\Bigg]\\
&\geq\frac{3}{32}\left(r(t)-r^{-1}(t)\right)^2\left(r(t)-r^{-1}(t)\right)^2=\frac{3}{32}\left(r(t)-r^{-1}(t)\right)^4.
\end{aligned}
$$

Then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough m, for any $0\leq t\leq \bar{t}_0=\left(\frac{\gamma-\gamma'-1}{8}\right)\log(m)-\log(2)$, the following holds:

$$
\begin{aligned}
&\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}\\
&\geq\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}-\frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}\\
&\geq\sqrt{\frac{2}{3md}}\sqrt{\frac{3m}{32}}\left(r(t)-r^{-1}(t)\right)^2-\sqrt{\frac{2}{d}}m^{-\frac{\gamma-\gamma'-1}{8}}.
\end{aligned}
$$

Specifically, if we choose $t=\bar{t}_0$,

$$
\begin{aligned}
&\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t_0)-\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)+\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t_0)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}\\
&\geq\sqrt{\frac{1}{16d}}\left(r^2(t_0)+r^{-2}(t_0)-2\right)-\sqrt{\frac{2}{d}}m^{-\frac{\gamma-\gamma'-1}{8}}
\end{aligned}
$$

$$\gtrsim m^{\frac{\gamma-\gamma'-1}{8}} - m^{-\frac{\gamma-\gamma'-1}{8}}.$$

By taking limit, we obtain that

$$\lim_{m\to\infty} \frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t_0) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t_0)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = +\infty.$$

For the second part of the RHS:

$$\frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \leq \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2},$$

then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough m, for any $0 \leq t \leq \bar{t}_0 = \left(\frac{\gamma-\gamma'-1}{8}\right)\log(m) - \log(2)$, the following holds:

$$\frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \leq \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} \leq \sqrt{\frac{2}{d}} m^{-\frac{\gamma-\gamma'-1}{8}}.$$

By taking limit, we obtain that

$$\lim_{m\to\infty} \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2} = 0.$$

To sum up, since $t_0 \leq T_{\text{eff}}$, we have that

$$\lim_{m\to+\infty} \sup_{t\in[0,T_{\text{eff}}]} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = +\infty + 0 = +\infty, \tag{C.24}$$

which finishes the proof of (C.22).

In order to prove (C.23), firstly we have

$$\frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2} \leq 1,$$

moreover, we observe that

$$\begin{aligned}\left(\frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2}\right)^2 &= \frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2^2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2^2} = \frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2^2}{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2^2 + \left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2} \\ &= \frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2}{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2 + \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2}.\end{aligned}$$

Then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough m, for any $0 \le t \le \bar{t}_0 = \left(\frac{\gamma-\gamma'-1}{8}\right)\log(m) - \log(2)$, the following holds:

$$
\begin{aligned}
\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2 &\ge \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2 - \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(0)\right\|_2 \\
&\ge \sqrt{\frac{3m}{32}}\left(r(t) - r^{-1}(t)\right)^2 - \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2 - \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}(0)\right\|_2 \\
&\ge \sqrt{\frac{3m}{32}}\left(r(t) - r^{-1}(t)\right)^2 - \sqrt{m}m^{-\frac{\gamma-\gamma'-1}{8}} - \sqrt{\frac{3md}{2}}, \\
\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2 &\le \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(0) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2 + \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(0)\right\|_2 \\
&\le \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w}}(t)\right\|_2 + \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w}}(0)\right\|_2 \\
&\le \sqrt{m}m^{-\frac{\gamma-\gamma'-1}{8}} + \sqrt{\frac{3md}{2}}.
\end{aligned}
$$

By taking $t = \bar{t}_0$, we observe that $\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2$ is of order at least $\sqrt{m}m^{\frac{\gamma-\gamma'-1}{8}}$, while $\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2$ is of order at most $\sqrt{m}$, which finishes the proof of (C.23). □

C.4 a-lag regime

Recall the real dynamics,

$$
\frac{\mathrm{d}\begin{pmatrix}\frac{\nu}{\varepsilon}a_k \\ \boldsymbol{w}_k\end{pmatrix}}{\mathrm{d}t} = \boldsymbol{A}\begin{pmatrix}\frac{\nu}{\varepsilon}a_k \\ \boldsymbol{w}_k\end{pmatrix} + \begin{pmatrix}f_k \\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k\end{pmatrix}, \quad \begin{pmatrix}\frac{\nu}{\varepsilon}a_k(0) \\ \boldsymbol{w}_k(0)\end{pmatrix} = \begin{pmatrix}\frac{\nu}{\varepsilon}a_k^0 \\ \boldsymbol{w}_k^0\end{pmatrix},
$$

and its solution (C.8) reads

$$
\begin{pmatrix}\frac{\nu}{\varepsilon}a_k \\ \boldsymbol{w}_k\end{pmatrix} = \exp\left(t\boldsymbol{A}\right)\begin{pmatrix}\frac{\nu}{\varepsilon}a_k^0 \\ \boldsymbol{w}_k^0\end{pmatrix} + \int_0^t \exp\left((t-s)\boldsymbol{A}\right)\begin{pmatrix}f_k(s) \\ \frac{\nu}{\varepsilon}\boldsymbol{g}_k(s)\end{pmatrix}\mathrm{d}s.
$$

Then, for any $\boldsymbol{u} \in \boldsymbol{z}^\perp$, and $\|\boldsymbol{u}\|_2 = 1$, we obtain that

$$
\begin{aligned}
\frac{\mathrm{d}\left\langle \boldsymbol{w}_k, \boldsymbol{u}\right\rangle}{\mathrm{d}t} &= \frac{\nu}{\varepsilon}\left\langle \boldsymbol{z}, \boldsymbol{u}\right\rangle + \frac{\nu}{\varepsilon}\left\langle \boldsymbol{g}_k, \boldsymbol{u}\right\rangle \\
&= \frac{\nu}{\varepsilon}\left\langle \boldsymbol{g}_k, \boldsymbol{u}\right\rangle,
\end{aligned}
$$

and as

$$
\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp} = \bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp} + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}
$$

hence the k-th component of $\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}$, $\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}$ and $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}$ altogether reads

$$
\begin{aligned}
\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}\right)_k &= \boldsymbol{w}_k^0 - \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}},\\
\left(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{w},\boldsymbol{z}^\perp}\right)_k &= \frac{\nu}{\varepsilon}\int_0^t \left[\boldsymbol{g}_k(s) - \left\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}}\right] \mathrm{d}s,\\
\left(\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}\right)_k &= \boldsymbol{w}_k^0 - \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}} + \frac{\nu}{\varepsilon}\int_0^t \left[\boldsymbol{g}_k(s) - \left\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}}\right] \mathrm{d}s.
\end{aligned}
$$

Moreover, the real dynamics can also be written as

$$
\frac{\mathrm{d}\begin{pmatrix} \frac{\nu}{\varepsilon}a_k \\ \langle \boldsymbol{w}_k, \hat{\boldsymbol{z}}\rangle \end{pmatrix}}{\mathrm{d}t} = \boldsymbol{B}\begin{pmatrix} \frac{\nu}{\varepsilon}a_k \\ \langle \boldsymbol{w}_k, \hat{\boldsymbol{z}}\rangle \end{pmatrix} + \begin{pmatrix} f_k \\ \frac{\nu}{\varepsilon}\langle \boldsymbol{g}_k, \hat{\boldsymbol{z}}\rangle \end{pmatrix}, \quad \begin{pmatrix} \frac{\nu}{\varepsilon}a_k(0) \\ \langle \boldsymbol{w}_k, \hat{\boldsymbol{z}}\rangle (0) \end{pmatrix} = \begin{pmatrix} \frac{\nu}{\varepsilon}a_k^0 \\ \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \end{pmatrix},
$$

where

$$
\boldsymbol{B} := \begin{pmatrix} 0 & \|\boldsymbol{z}\|_2 \\ \|\boldsymbol{z}\|_2 & 0 \end{pmatrix} = \|\boldsymbol{z}\|_2 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},
$$

since $\boldsymbol{B}$ is a full rank matrix, the solution to the above dyanmics can be explicitly written out as

$$
\begin{aligned}
\begin{pmatrix} \frac{\nu}{\varepsilon}a_k \\ \langle \boldsymbol{w}_k, \hat{\boldsymbol{z}}\rangle \end{pmatrix} &= \exp(t\boldsymbol{B})\begin{pmatrix} \frac{\nu}{\varepsilon}a_k^0 \\ \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \end{pmatrix} + \int_0^t \exp((t-s)\boldsymbol{B})\begin{pmatrix} f_k(s) \\ \frac{\nu}{\varepsilon}\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\rangle \end{pmatrix}\mathrm{d}s\\
&= \begin{pmatrix} \frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t) & \frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t) \\ \frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t) & \frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t) \end{pmatrix}\begin{pmatrix} \frac{\nu}{\varepsilon}a_k^0 \\ \left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \end{pmatrix}\\
&\quad + \int_0^t \begin{pmatrix} \frac{1}{2}r^2(t-s) + \frac{1}{2}r^{-2}(t-s) & \frac{1}{2}r^2(t-s) - \frac{1}{2}r^{-2}(t-s) \\ \frac{1}{2}r^2(t-s) - \frac{1}{2}r^{-2}(t-s) & \frac{1}{2}r^2(t-s) + \frac{1}{2}r^{-2}(t-s) \end{pmatrix}\begin{pmatrix} f_k(s) \\ \frac{\nu}{\varepsilon}\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\rangle \end{pmatrix}\mathrm{d}s,
\end{aligned}
$$

hence we obtain that

$$
\begin{aligned}
(\boldsymbol{\theta}_a)_k = \boldsymbol{a}_k &= \left(\frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t)\right)a_k^0 + \frac{\varepsilon}{\nu}\left(\frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t)\right)\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle\\
&\quad + \frac{\varepsilon}{\nu}\int_0^t \left(\frac{1}{2}r^2(t-s) + \frac{1}{2}r^{-2}(t-s)\right) f_k(s)\mathrm{d}s\\
&\quad + \int_0^t \left(\frac{1}{2}r^2(t-s) - \frac{1}{2}r^{-2}(t-s)\right)\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\rangle \,\mathrm{d}s,\\
(\boldsymbol{\theta}_{\boldsymbol{w}})_k = \boldsymbol{w}_k &= \frac{\nu}{\varepsilon}\left(\frac{1}{2}r^2(t) - \frac{1}{2}r^{-2}(t)\right)a_k^0\hat{\boldsymbol{z}} + \left(\frac{1}{2}r^2(t) + \frac{1}{2}r^{-2}(t)\right)\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}}\\
&\quad + \int_0^t \left(\frac{1}{2}r^2(t-s) - \frac{1}{2}r^{-2}(t-s)\right) f_k(s)\mathrm{d}s\hat{\boldsymbol{z}}\\
&\quad + \frac{\nu}{\varepsilon}\int_0^t \left(\frac{1}{2}r^2(t-s) + \frac{1}{2}r^{-2}(t-s)\right)\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\rangle \,\mathrm{d}s\hat{\boldsymbol{z}}
\end{aligned}
$$

$$-\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}} + \boldsymbol{w}_k^0 + \frac{\nu}{\varepsilon}\int_0^t \left[\boldsymbol{g}_k(s) - \left\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}}\right] \mathrm{d}s.$$

Definition 4 (Neuron ∞-energy, a-lag regime)**.** *In real dynamics, we define the ∞-energy at time t for each single neuron, i.e., for each $k \in [m]$,*

$$p_k(t) := \max_{k\in[m]} \left\{\left|a_k(t)\right|, \left\|\boldsymbol{w}_k(t)\right\|_\infty\right\}. \tag{C.25}$$

We denote

$$p_{\max}(t) := \max_{k\in[m]} p_k(t). \tag{C.26}$$

For simplicity, we hereafter drop the (t)s for all $p_k(t)$ and $p_{\max}(t)$. Then the estimates on $\{f_k, \boldsymbol{g}_k\}_{k=1}^m$ read

Proposition 9. *For any time $t > 0$,*

$$\begin{aligned} |f_k| &\leq dm\nu\varepsilon p_{\max}^2 \left\|\boldsymbol{w}_k\right\|_\infty + \varepsilon d \left\|\boldsymbol{w}_k\right\|_\infty^2, \\ \left\|\boldsymbol{g}_k\right\|_\infty &\leq \sqrt{d} m\nu\varepsilon p_{\max}^2 |a_k| + \varepsilon\sqrt{d}\left\|\boldsymbol{w}_k\right\|_\infty |a_k|. \end{aligned} \tag{C.27}$$

Proof. We obtain that

$$\begin{aligned} |f_k| &= \left|\frac{1}{n}\sum_{i=1}^n \left((e_i + y_i)\frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} + y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i - y_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}\right)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^n \left(\left(\sum_{k=1}^m \nu a_k \sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right)\frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon} + y_i \boldsymbol{w}_k^\intercal \boldsymbol{x}_i - y_i \frac{\sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)}{\varepsilon}\right)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \left(\left(\sum_{k=1}^m \nu\varepsilon |a_k| \left\|\boldsymbol{w}_k\right\|_2\right)\left\|\boldsymbol{w}_k\right\|_2 + \varepsilon(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)^2\right) \\ &\leq \frac{1}{n}\sum_{i=1}^n \left(\left(\sum_{k=1}^m \nu\varepsilon |a_k| \sqrt{d}\left\|\boldsymbol{w}_k\right\|_\infty\right)\sqrt{d}\left\|\boldsymbol{w}_k\right\|_\infty + \varepsilon d\left\|\boldsymbol{w}_k\right\|_\infty^2\right) \\ &\leq dm\nu\varepsilon r_{\max}^2 \left\|\boldsymbol{w}_k\right\|_\infty + \varepsilon d\left\|\boldsymbol{w}_k\right\|_\infty^2, \end{aligned}$$

and

$$\begin{aligned} \left\|\boldsymbol{g}_k\right\|_\infty &= \left\|\frac{1}{n}\sum_{i=1}^n \left((e_i + y_i)a_k\sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i + y_i a_k \boldsymbol{x}_i - y_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i\right)\right\|_\infty \\ &= \left\|\frac{1}{n}\sum_{i=1}^n \left(\left(\sum_{k=1}^m \nu a_k \sigma(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right) a_k\sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i + y_i a_k \boldsymbol{x}_i - y_i a_k \sigma^{(1)}(\varepsilon \boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i\right)\right\|_\infty \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\left(\sum_{k=1}^{m}\nu\varepsilon\left|a_k\right|\left\|\boldsymbol{w}_k\right\|_2\right)\left|a_k\right|+\varepsilon\left\|\boldsymbol{w}_k\right\|_2\left|a_k\right|\right)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\left(\sum_{k=1}^{m}\nu\varepsilon\left|a_k\right|\sqrt{d}\left\|\boldsymbol{w}_k\right\|_\infty\right)\left|a_k\right|+\varepsilon\sqrt{d}\left\|\boldsymbol{w}_k\right\|_\infty\left|a_k\right|\right)\\
&\leq \sqrt{d}m\nu\varepsilon r_{\max}^2\left|a_k\right|+\varepsilon\sqrt{d}\left\|\boldsymbol{w}_k\right\|_\infty\left|a_k\right|.
\end{aligned}
$$

□

We denote a useful quantity

$$
\psi(t):=\sup_{0\leq s\leq t}p_{\max}(s). \tag{C.28}
$$

Then directly from Lemma 1, we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$
\max_{k\in[m]}\left\{|a_k^0|,\ \|\boldsymbol{w}_k^0\|_\infty\right\}\leq\sqrt{2\log\frac{2m(d+1)}{\delta}}, \tag{C.29}
$$

hence

$$
\psi(0)\leq\sqrt{2(d+1)\log\frac{2m(d+1)}{\delta}}, \tag{C.30}
$$

and for all $k\in[m]$ and any time $t>0$,

$$
\begin{aligned}
|f_k|&\leq dm\nu\varepsilon\psi^3(t)+\varepsilon d\psi^2(t),\\
\left\|\boldsymbol{g}_k\right\|_\infty&\leq\sqrt{d}m\nu\varepsilon\psi^3(t)+\varepsilon\sqrt{d}\psi^2(t).
\end{aligned}
$$

We are hereby to conduct some simple calculations. Let $\alpha>0$, and there exists $t_\alpha>0$, such that

$$
r^2(t_\alpha)-r^{-2}(t_\alpha)=2m^{-\alpha},
$$

then we obtain immediately that

$$
r^2(t_\alpha)=\exp(\left\|\boldsymbol{z}\right\|_2 t_\alpha)=\sqrt{1+m^{-\alpha}}+m^{-\alpha},
$$

and

$$
t_\alpha\lesssim\log(\sqrt{1+m^{-\alpha}}+m^{-\alpha})\sim\frac{3}{2}m^{-\alpha}\sim\mathcal{O}(1),
$$

with

$$
t_\alpha\gtrsim\log(\sqrt{1+m^{-\alpha}}+m^{-\alpha})\sim\frac{3}{2}m^{-\alpha}\sim\Omega(1).
$$

Moreover,

$$
\frac{1}{2}r^2(t_\alpha)+\frac{1}{2}r^{-2}(t_\alpha)=\sqrt{1+m^{-\alpha}},
$$

$$\frac{1}{2}r^2(t_\alpha) - \frac{1}{2}r^{-2}(t_\alpha) = m^{-\alpha},$$
$$\int_0^{t_\alpha} \frac{1}{2}r^2(t_\alpha - s) + \frac{1}{2}r^{-2}(t_\alpha - s)\mathrm{d}s = \frac{m^{-\alpha}}{\|\boldsymbol{z}\|_2},$$
$$\int_0^{t_\alpha} \frac{1}{2}r^2(t_\alpha - s) - \frac{1}{2}r^{-2}(t_\alpha - s)\mathrm{d}s = \frac{\sqrt{1+m^{-\alpha}}-1}{\|\boldsymbol{z}\|_2},$$

hence

$$\begin{aligned} |a_k| &\leq \psi(0) + m^{-\alpha}\left(m\nu\varepsilon\psi^3(t_\alpha) + \varepsilon\psi^2(t_\alpha)\right), \\ \|\boldsymbol{w}_k\|_\infty &\leq \frac{\nu}{\varepsilon}m^{-\alpha}\psi(0) + \frac{\nu}{\varepsilon}m^{-\alpha}\left(m\nu\varepsilon\psi^3(t_\alpha) + \varepsilon\psi^2(t_\alpha)\right), \end{aligned} \tag{C.31}$$

For some $\beta > 1$, we define

$$\widetilde{T}_{\mathrm{eff},\beta} := \inf\left\{t > 0 \mid m\nu\varepsilon\psi^3(t) + \varepsilon\psi^2(t) > m^{-\widetilde{\tau}}, \quad \widetilde{\tau} = -\frac{\gamma'}{\beta}\right\}, \tag{C.32}$$

then for m large enough, as $\gamma > 1$, based on (C.30), we choose β satisfying

$$\min\left\{\gamma - 1, \frac{1}{2}\right\} > -\frac{\gamma'}{\beta},$$

then, we obtain that

$$m\nu\varepsilon\psi^3(0)+\varepsilon\psi^2(0) \leq m\nu\varepsilon\left(2(d+1)\log\frac{2m(d+1)}{\delta}\right)^{\frac{3}{2}} + m^{-\frac{1}{2}}\left(2(d+1)\log\frac{2m(d+1)}{\delta}\right) \leq m^{-\frac{\gamma-1}{2\beta}},$$

hence $\widetilde{T}_{\mathrm{eff},\beta} \geq 0$.

We observe further that by taking the ∞-norm on both sides of (C.31), and by taking supreme over the index k and time $0 \leq t \leq \min\left\{t_\alpha, \widetilde{T}_{\mathrm{eff},\beta}\right\}$ on both sides, and for large enough m, the following holds

$$\psi(t) \leq \frac{\nu}{\varepsilon}m^{-\alpha}\psi(0) + \frac{\nu}{\varepsilon}m^{-\alpha}m^{-\frac{\gamma'}{\beta}} \leq m^{-\gamma'-\alpha}\sqrt{2(d+1)\log\frac{2m(d+1)}{\delta}},$$

now we shall choose $\alpha > 0$ and $\beta > 1$, such that

$$t_\alpha \leq \widetilde{T}_{\mathrm{eff},\beta},$$

which is equivalent to solve out the relation

$$m^{1-\gamma}m^{-3\gamma'-3\alpha} + m^{-\frac{1}{2}}m^{-2\gamma'-2\alpha} \leq m^{\frac{\gamma'}{\beta}}.$$

Then, we choose α satisfying

$$0 < \gamma' - \alpha \leq \min\left\{\frac{1}{3}\left(\gamma - 1 + \frac{\gamma'}{\beta}\right), \frac{1}{2}\left(\frac{1}{2} + \frac{\gamma'}{\beta}\right)\right\},$$

the existence of α can be guaranteed since $\gamma' > 0$, and

$$\min\left\{\frac{1}{3}\left(\gamma - 1 + \frac{\gamma'}{\beta}\right), \frac{1}{2}\left(\frac{1}{2} + \frac{\gamma'}{\beta}\right)\right\} \geq \frac{\gamma'}{6\beta} > 0.$$

We observe that

$$\widetilde{T}_{\mathrm{eff},\beta} \geq t_\alpha, \quad \widetilde{T}_{\mathrm{eff},\beta} \sim \Omega(1). \tag{C.33}$$

WLOG, we choose $\bar{\beta} := \max\left\{-1024\frac{\gamma'}{\gamma-1}, -512\gamma'\right\}$, and $\bar{\alpha}$ accordingly, and finally we have

Theorem 8 (Condensed regime, a-lag regime)**.** *Given any $\delta \in (0,1)$, under Assumption 1, Assumption 3 and Assumption 5, if $\gamma > 1$ and $\gamma' < 0$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{m\to+\infty} \sup_{t\in[0,\widetilde{T}_{\mathrm{eff},\bar{\beta}}]} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = +\infty, \tag{C.34}$$

and

$$\lim_{m\to+\infty} \sup_{t\in[0,\widetilde{T}_{\mathrm{eff},\bar{\beta}}]} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)\|_2} = 1. \tag{C.35}$$

As the details are almost the same as the ones in Theorem 7, the proof of Theorem 8 is written in a slightly sketchy way.

Proof. We observe that

$$\begin{aligned}
&(\boldsymbol{\theta}_{\boldsymbol{w}}(t_{\bar{\alpha}}) - \boldsymbol{\theta}_{\boldsymbol{w}}(0))_k \\
&= \frac{\nu}{\varepsilon}\left(\frac{1}{2}r^2(t_{\bar{\alpha}}) - \frac{1}{2}r^{-2}(t_{\bar{\alpha}})\right)a_k^0\hat{\boldsymbol{z}} + \left(\frac{1}{2}r^2(t_{\bar{\alpha}}) + \frac{1}{2}r^{-2}(t_{\bar{\alpha}}) - 1\right)\left\langle \boldsymbol{w}_k^0, \hat{\boldsymbol{z}}\right\rangle \hat{\boldsymbol{z}} \\
&\quad + \int_0^{t_{\bar{\alpha}}} \left(\frac{1}{2}r^2(t_{\bar{\alpha}} - s) - \frac{1}{2}r^{-2}(t_{\bar{\alpha}} - s)\right) f_k(s)\mathrm{d}s\hat{\boldsymbol{z}} \\
&\quad + \frac{\nu}{\varepsilon}\int_0^{t_{\bar{\alpha}}} \left(\frac{1}{2}r^2(t_{\bar{\alpha}} - s) + \frac{1}{2}r^{-2}(t_{\bar{\alpha}} - s)\right)\langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\rangle\,\mathrm{d}s\hat{\boldsymbol{z}} \\
&\quad + \frac{\nu}{\varepsilon}\int_0^{t_{\bar{\alpha}}} \left[\boldsymbol{g}_k(s) - \langle \boldsymbol{g}_k(s), \hat{\boldsymbol{z}}\rangle\,\hat{\boldsymbol{z}}\right]\mathrm{d}s,
\end{aligned}$$

so we obtain that with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough m, for any $0 \leq t \leq t_{\bar{\alpha}}$, the following holds:

$$
\begin{aligned}
&\frac{1}{m}\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2 \\
&\geq m^{-\gamma'-\bar{\alpha}}\left(\frac{1}{2}-\sqrt{d}\max_{k\in[m]}\left\|\boldsymbol{g}_k\right\|_\infty\right)-\left(\sqrt{1+m^{-\bar{\alpha}}}-1\right)\left(\frac{3d}{2}+\sqrt{d}\max_{k\in[m]}\left|f_k\right|\right) \\
&\geq m^{-\gamma'-\bar{\alpha}}\left(\frac{1}{2}-\sqrt{d}m^{-\widetilde{\tau}}\right)-\frac{1}{2}m^{-\bar{\alpha}}\left(\frac{3d}{2}+\sqrt{d}m^{-\widetilde{\tau}}\right),
\end{aligned}
$$

hence the ratio

$$
\begin{aligned}
\left(\frac{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}\right)^2 &= \frac{\frac{1}{m}\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2}{\frac{1}{m}\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2^2} \\
\geq \frac{2}{3d}&\left[m^{-\gamma'-\bar{\alpha}}\left(\frac{1}{2}-\sqrt{d}m^{-\widetilde{\tau}}\right)-\frac{1}{2}m^{-\bar{\alpha}}\left(\frac{3d}{2}+\sqrt{d}m^{-\widetilde{\tau}}\right)\right],
\end{aligned}
$$

by taking limit, we obtain that for any $0 \leq t \leq t_{\bar{\alpha}}$

$$
\lim_{m\to\infty}\frac{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\right\|_2}=+\infty.
$$

Moreover, since

$$
\begin{aligned}
\left(\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t_{\bar{\alpha}})\right)_k &= \frac{\nu}{\varepsilon}\left(\frac{1}{2}r^2(t_{\bar{\alpha}})-\frac{1}{2}r^{-2}(t_{\bar{\alpha}})\right)a_k^0+\left(\frac{1}{2}r^2(t_{\bar{\alpha}})+\frac{1}{2}r^{-2}(t_{\bar{\alpha}})\right)\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle \\
&\quad+\int_0^{t_{\bar{\alpha}}}\left(\frac{1}{2}r^2(t_{\bar{\alpha}}-s)-\frac{1}{2}r^{-2}(t_{\bar{\alpha}}-s)\right)f_k(s)\mathrm{d}s \\
&\quad+\frac{\nu}{\varepsilon}\int_0^{t_{\bar{\alpha}}}\left(\frac{1}{2}r^2(t_{\bar{\alpha}}-s)+\frac{1}{2}r^{-2}(t_{\bar{\alpha}}-s)\right)\left\langle\boldsymbol{g}_k(s),\hat{\boldsymbol{z}}\right\rangle\mathrm{d}s, \\
\left(\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t_{\bar{\alpha}})\right)_k &= \boldsymbol{w}_k^0-\left\langle\boldsymbol{w}_k^0,\hat{\boldsymbol{z}}\right\rangle\hat{\boldsymbol{z}}+\frac{\nu}{\varepsilon}\int_0^t\left[\boldsymbol{g}_k(s)-\left\langle\boldsymbol{g}_k(s),\hat{\boldsymbol{z}}\right\rangle\hat{\boldsymbol{z}}\right]\mathrm{d}s,
\end{aligned}
$$

so we obtain that with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough m, for any $0 \leq t \leq t_{\bar{\alpha}}$, the following holds:

$$
\begin{aligned}
\frac{1}{m}\left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\right\|_2^2 &\geq m^{-\gamma'-\bar{\alpha}}\left(\frac{1}{2}-\sqrt{d}\max_{k\in[m]}\left\|\boldsymbol{g}_k\right\|_\infty\right)-\sqrt{1+m^{-\bar{\alpha}}}\left(\frac{3d}{2}+\sqrt{d}\max_{k\in[m]}\left|f_k\right|\right) \\
&\geq m^{-\gamma'-\bar{\alpha}}\left(\frac{1}{2}-\sqrt{d}m^{-\widetilde{\tau}}\right)-2\left(\frac{3d}{2}+\sqrt{d}m^{-\widetilde{\tau}}\right), \\
\frac{1}{m}\left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2 &\leq \frac{3d}{2}+m^{-\gamma'-\bar{\alpha}}\sqrt{d}\max_{k\in[m]}\left\|\boldsymbol{g}_k\right\|_\infty \leq \frac{3d}{2}+m^{-\gamma'-\bar{\alpha}}\sqrt{d}m^{-\widetilde{\tau}}.
\end{aligned}
$$

By taking $t = t_{\bar{\alpha}}$, we observe that $\frac{1}{m} \|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}}(t)\|_2^2$ is of order at least $m^{-\gamma'-\bar{\alpha}}$, while $\frac{1}{m} \left\|\boldsymbol{\theta}_{\boldsymbol{w},\boldsymbol{z}^\perp}(t)\right\|_2^2$ is of order at most one, which finishes the proof. $\square$