Journal of Machine Learning Research 22 (2021) 1-47 Submitted 10/20; Revised 1/21; Published 2/21

Phase Diagram for Two-layer ReLU Neural Networks at Infinite-width Limit

Tao Luo[a,*] LUOTAO41@SJTU.EDU.CN
Zhi-Qin John Xu[a,*] XUZHIQIN@SJTU.EDU.CN
Zheng Ma[a] ZHENGMA@SJTU.EDU.CN
Yaoyu Zhang[a,b,†] ZHYY.SJTU@SJTU.EDU.CN
[a] *School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, China*
[b] *Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, 200031, China*

Editor: Ohad Shamir

Keywords: two-layer ReLU neural network, infinite-width limit, phase diagram, dynamical regime, condensation

Abstract

How neural network behaves during the training over different choices of hyperparameters is an important question in the study of neural networks. In this work, inspired by the phase diagram in statistical mechanics, we draw the phase diagram for the two-layer ReLU neural network at the infinite-width limit for a complete characterization of its dynamical regimes and their dependence on hyperparameters related to initialization. Through both experimental and theoretical approaches, we identify three regimes in the phase diagram, i.e., *linear* regime, *critical* regime and *condensed* regime, based on the relative change of input weights as the width approaches infinity, which tends to 0, $O(1)$ and $+\infty$, respectively. In the linear regime, NN training dynamics is approximately linear similar to a random feature model with an exponential loss decay. In the condensed regime, we demonstrate through experiments that active neurons are condensed at several discrete orientations. The critical regime serves as the boundary between above two regimes, which exhibits an intermediate nonlinear behavior with the mean-field model as a typical example. Overall, our phase diagram for the two-layer ReLU NN serves as a map for the future studies and is a first step towards a more systematical investigation of the training behavior and the implicit regularization of NNs of different structures.

1. Introduction

It has been widely observed that, given training data, neural networks (NNs) may exhibit distinctive dynamical behaviors during the training, depending on the choices of hyperparameters. As an example, we consider a two-layer NN with m hidden neurons

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

∗. The first two authors contributed equally.
†. Corresponding author.

where $\boldsymbol{x} \in \mathbb{R}^d$, α is the scaling factor, $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\theta}_a, \boldsymbol{\theta}_{\boldsymbol{w}})$ with $\boldsymbol{\theta}_a = \mathrm{vec}(\{a_k\}_{k=1}^m)$, $\boldsymbol{\theta}_{\boldsymbol{w}} = \mathrm{vec}(\{\boldsymbol{w}_k\}_{k=1}^m)$ is the set of parameters initialized by $a_k^0 \sim N(0, \beta_1^2)$, $\boldsymbol{w}_k^0 \sim N(0, \beta_2^2 \boldsymbol{I}_d)$. The bias term b_k can be incorporated by expanding $\boldsymbol{x}$ and $\boldsymbol{w}_k$ to $(\boldsymbol{x}^\intercal, 1)^\intercal$ and $(\boldsymbol{w}_k^\intercal, b_k)^\intercal$. At the infinite-width limit $m \to \infty$, given $\beta_1, \beta_2 \sim O(1)$, for $\alpha \sim \sqrt{m}$, the gradient flow of NN can be approximated by a linear dynamics of neural tangent kernel (NTK) (Jacot et al., 2018; Arora et al., 2019; Zhang et al., 2020), whereas for $\alpha \sim m$, gradient flow of NN exhibits highly nonlinear mean-field dynamics (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020). The current situation of NN study is similar to an early era of statistical mechanics, when we observe different states of a matter at several discrete conditions without the guidance of a unified phase diagram.

In this work, we present the first phase diagram for the two-layer neural networks with rectified linear units (ReLU NN). To this end, two difficulties need to be overcome. The first difficulty is that one can not identify sharply distinctive regimes/states required for a phase diagram with finite neurons. This situation is similar to the analysis in statistical mechanics, e.g., Ising model, where phase transition can not happen with finite particles. Therefore, in analogy to the thermodynamic limit, we take the infinite-width limit $m \to \infty$ as our starting point and successfully identify three dynamical regimes of NNs, i.e., *linear* regime, *critical* regime, and *condensed* regime. In the linear regime, $\boldsymbol{\theta}_{\boldsymbol{w}}$ almost does not change and NN training dynamics can be linearized around the initialization similar to an NTK or a random feature model. In the condensed regime, the relative change of $\boldsymbol{\theta}_{\boldsymbol{w}}$ tends to infinity and is condensed at several discrete directions in the feature space. In the critical regime, which serves as the boundary between above two regimes, relative change of $\boldsymbol{\theta}_{\boldsymbol{w}}$ is $O(1)$ with the mean-field model as an example. The second difficulty is the identification of phase diagram coordinates. For the vanilla gradient flow training dynamics of NN in Eq. (1), there are three hyperparameters α, β_1 and β_2, which in general are functions of m. However, through appropriate rescaling and normalization of the gradient flow dynamics, which accounts for the dynamical similarity up to a time scaling, we arrive at two independent coordinates

$$\gamma = \lim_{m\to\infty} -\frac{\log \beta_1\beta_2/\alpha}{\log m}, \quad \gamma' = \lim_{m\to\infty} -\frac{\log \beta_1/\beta_2}{\log m}. \tag{2}$$

The resulting phase diagram is shown in Fig. 1. Examples studied in previous literature are also marked, for example, Ref. E et al. (2020) studied NNs with settings represented by the red dashed line.

This phase diagram is obtained through experimental and theoretical approaches. We first present an intuitive scaling analysis to provide a rationale for the boundary that separates the linear regime and the condensed regime. Then, we experimentally demonstrate the transition across this boundary in the phase diagram for an 1-d data set. Finally, we establish a rigorous theory for general data sets.

Our work is a first step towards a systematical effort in drawing the phase diagrams for NNs of different structures. With the guidance of these phase diagrams, detailed experimental * and theoretical works can be done to further characterize the dynamical behavior and the corresponding implicit regularization effect at each of the identified regime.

∗. Code can be found in `https://github.com/xuzhiqin1990/phasediagram_twolayerNN`

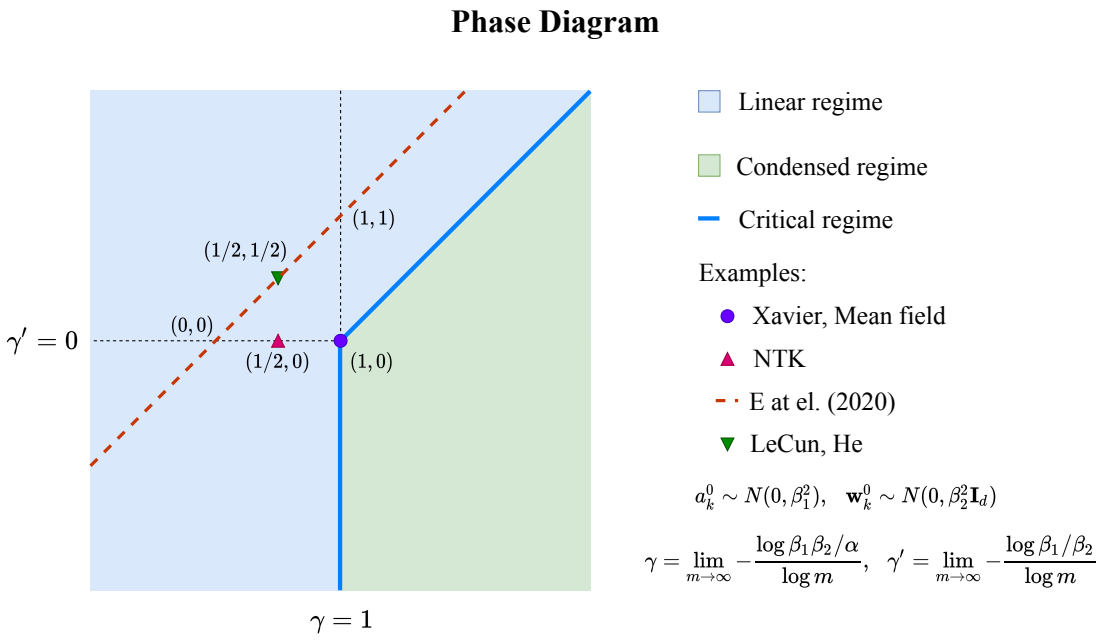


Figure 1: Phase diagram of two-layer ReLU NNs at infinite-width limit. The marked examples are studied in existing literature (see Table 1 for details.)

2. Related Works

The study of regimes in the literature usually revolves around the choice of scaling factor α in specific power-law relations to the width m. For example, the NTK scaling $\alpha \sim \sqrt{m}$ (Jacot et al., 2018; Arora et al., 2019; Zhang et al., 2020) and the mean-field scaling $\alpha \sim m$ (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020) has been studied extensively. In Chizat et al. (2019), the authors identify the lazy training behavior for $\lim_{m\to\infty} m/\alpha = \infty$, by which NN parameters stay close to initialization during the training. In Williams et al. (2019), for two-layer ReLU network, lazy and active regimes and their corresponding regularization effect are studied for 1-d problems. Their analysis uses different quantities for regime separation, which cannot serve as coordinates for a phase diagram. In Geiger et al. (2020), the authors show empirically that the mean-field scaling of $\alpha \sim m$ for $\beta_1, \beta_2 \sim O(1)$ fixed (i.e., $\gamma = 1$ for $\gamma' = 0$ fixed in our phase diagram Fig. 1) is critical for regime separation. All above works do not account for the effect of specific power-law scaling of initialization over different layers used in practice. Note that, in Woodworth et al. (2020), for the matrix factorization problem, a similar critical scaling of $1/m$ is identified for regime separation.

In E et al. (2020), for two-layer NNs with $\alpha = 1$, $\beta_2 \sim O(1)$, the authors study the effect of β ($\sim \beta_1$) in relation to m. Specifically, they prove that NN training dynamics can be linearized for $\beta = o(m^{-1/6})$ as $m \to \infty$, which constitutes a line in Fig. 1. In Ma et al. (2020), they further study such cases in the under-parameterized and mildly over-parameterized settings and experimentally identified the quenching-activation behavior for

finite m, which phenomenologically is closely related to the condensed regime we identified at $m \to \infty$.

Another work related to the condensed regime is Maennel et al. (2018). The authors study the two-layer ReLU NNs and prove that, as the initialization of parameters goes to zero, a quantization effect emerges, that is, the weight vectors tend to concentrate at a small number of orientations determined by the input data at an early stage of training. However, the limit of $m \to \infty$ is not considered in their work.

Our linear regime is closely related to NTK, kernel and lazy regimes, which, despite defined under different settings, exhibit the same characteristic linear training dynamics; our critical regime and condensed regime are related to mean-field, active, adaptive, deep, rich or feature-learning regimes, which characterizes nonlinear training dynamics of NNs from different perspectives (Jacot et al., 2018; Chizat et al., 2019; Mei et al., 2019; Williams et al., 2019; Woodworth et al., 2020; Moroshko et al., 2020; Geiger et al., 2020; Chen et al., 2020).

3. Rescaling and the Normalized Model

Identification of the coordinates is important for drawing the phase diagram. Unlike in some thermodynamic systems where temperature and pressure are natural choices, for NNs, it is not obvious which quantities of hyperparameters are keys to the regime separation. However, there are some guiding principles for finding the coordinates of a phase diagram at $m \to \infty$:

(i) They should be effectively independent.

(ii) Given a specific coordinate in the phase diagram, the learning dynamics of all the corresponding NNs statistically should be similar up to a time scaling.

(iii) They should well differentiate dynamical differences except for the time scaling.

Guided by above principles, in this section, we perform the following rescaling procedure for a fair comparison between different choices of hyperparameters and obtain a normalized model with two independent quantities irrespective of the time scaling of the gradient flow dynamics. We start with the original model (1)

$$f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x}) = \frac{1}{\alpha} \sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^{\intercal} \boldsymbol{x}), \tag{3}$$

defined on a given sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in [n]$, network width m and a scaling parameter $1/\alpha$ and $\sigma = \mathrm{ReLU}$. The parameters are initialized by

$$a_k^0 \sim N(0, \beta_1^2), \quad \boldsymbol{w}_k^0 \sim N(0, \beta_2^2 \boldsymbol{I}_d), \tag{4}$$

where a_k and $\boldsymbol{w}_k$ are separated into different scales β_1 and β_2. The empirical risk is

$$R_S(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x}_i) - y_i)^2. \tag{5}$$

Then the training dynamics based on gradient descent (GD) at the continuous limit obeys the following gradient flow of $\boldsymbol{\theta}$,

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}). \tag{6}$$

More precisely, $\boldsymbol{\theta} = \mathrm{vec}(\{\boldsymbol{q}_k\}_{k=1}^m)$ with $\boldsymbol{q}_k = (a_k, \boldsymbol{w}_k^\intercal)^\intercal$, $k \in [m]$ solves

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

Let

$$\bar{a}_k = \beta_1^{-1} a_k, \quad \bar{\boldsymbol{w}}_k = \beta_2^{-1}\boldsymbol{w}_k, \quad \bar{t} = \frac{1}{\beta_1\beta_2}t, \tag{7}$$

then

$$\begin{aligned}
\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}\bar{t}} &= -\frac{\beta_2}{\beta_1}\frac{1}{n}\sum_{i=1}^n \frac{\beta_1\beta_2}{\alpha}\sigma(\bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)\left(\frac{\beta_1\beta_2}{\alpha}\sum_{k'=1}^m \bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^\intercal \boldsymbol{x}_i) - y_i\right), \\
\frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}\bar{t}} &= -\frac{\beta_1}{\beta_2}\frac{1}{n}\sum_{i=1}^n \frac{\beta_1\beta_2}{\alpha}\bar{a}_k\sigma'(\bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i\left(\frac{\beta_1\beta_2}{\alpha}\sum_{k'=1}^m \bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^\intercal \boldsymbol{x}_i) - y_i\right).
\end{aligned}$$

We introduce two scaling parameters

$$\kappa := \frac{\beta_1\beta_2}{\alpha}, \quad \kappa' := \frac{\beta_1}{\beta_2}, \tag{8}$$

where κ and κ' are called the energetic scaling parameter and the dynamical scaling parameter, respectively. Then the above dynamics can be written as

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

The above recaled dynamics can be treated as a weighted gradient flow of NN scaled by κ equipped with the empirical risk

$$f_{\boldsymbol{\theta}}^\kappa(\boldsymbol{x}) = \kappa\sum_{k=1}^m \bar{a}_k\sigma(\bar{\boldsymbol{w}}_k^\intercal \boldsymbol{x}), \tag{9}$$

$$R_{S,\kappa}(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^n (f_{\boldsymbol{\theta}}^\kappa(\boldsymbol{x}_i) - y_i)^2, \tag{10}$$

with the following initialization

$$\bar{a}_k^0 \sim N(0,1), \quad \bar{\boldsymbol{w}}_k^0 \sim N(0, \boldsymbol{I}_d), \tag{11}$$

where we can see they are of standard normal distributions. The weighted GD dynamics then can be written simply as

$$\frac{\mathrm{d}\bar{\boldsymbol{q}}_k}{\mathrm{d}\bar{t}} = -\boldsymbol{M}_{\kappa'} \nabla_{\boldsymbol{q}_k} R_{S,\kappa}(\bar{\boldsymbol{\theta}}), \tag{12}$$

where the mobility matrix

$$\boldsymbol{M}_{\kappa'} = \begin{pmatrix} 1/\kappa' & \\ & \kappa' \boldsymbol{I}_d \end{pmatrix}. \tag{13}$$

In the following discussion throughout this paper, we will refer to this rescaled model (9) as *normalized* model and drop superscript κ and all the "bar"s of a_k, $\boldsymbol{w}_k$, t for simplicity. Note that κ and κ' do not follow principle (ii) and (iii) above at infinite-width limit. They are in general functions of m, which attains 0, $O(1)$, $+\infty$ at $m \to \infty$. For example, $\kappa = 0$ and $\kappa' = 1$ for both the NTK and mean-field model, however, they are known to have distinctive training behaviors. To account for such dynamical difference under different widely considered power-law scalings of α, β_1 and β_2 shown in Table. 1, we arrive at

$$\gamma = \lim_{m\to\infty} -\frac{\log \kappa}{\log m}, \quad \gamma' = \lim_{m\to\infty} -\frac{\log \kappa'}{\log m}, \tag{14}$$

which meets all above principles as demonstrated later by theory and experiments.

Remark 1. *We remark that the above rescaling technique can be viewed in analogy to the nondimensionalization in physics, which is the partial or full removal of physical dimensions from an equation involving physical quantities by a suitable substitution of variables. In more general point of view, nondimensionalization can also recover characteristic properties of a system, which in our case recovers the different behaviors of training dynamics for different regimes.*

More specifically, we can view, in the original model (1)*,* $\boldsymbol{q}_k = (a_k, \boldsymbol{w}_k^\intercal)^\intercal$*,* $k \in [m]$ *as the generalized coordinates which have the unit of "length" denoted as* [L]*. Then in the two-layer NN* (1)*,* α *should have the unit of "volume" as a normalization factor depending on* m *to avoid blowing up of the model. Particularly, if* σ *is* ReLU *then we can think* α*'s unit is* $[\mathrm{L}]^2$ *(unit of area on a plane).*

Finally, following above analysis, $\kappa = \frac{\beta_1\beta_2}{\alpha}$ *and* $\kappa' = \frac{\beta_1}{\beta_2}$ *are two* nondimensional *parameters (without unit) so as for* γ *and* γ'*, which are suitable to serve as the coordinations of our phase diagram.*

Remark 2. *Here we list some commonly-used initialization methods and/or related works with their scaling parameters as shown in Table 1.*

Name (related works)	α	β_1	β_2	κ $(\frac{\beta_1\beta_2}{\alpha})$	κ' $(\frac{\beta_1}{\beta_2})$	γ $(\lim_{m\to\infty}\frac{\log 1/\kappa}{\log m})$	γ' $(\lim_{m\to\infty}\frac{\log 1/\kappa'}{\log m})$
LeCun (LeCun et al., 2012)	1	$\sqrt{\frac{1}{m}}$	$\sqrt{\frac{1}{d}}$	$\sqrt{\frac{1}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
He (He et al., 2015)	1	$\sqrt{\frac{2}{m}}$	$\sqrt{\frac{2}{d}}$	$\sqrt{\frac{4}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
Xavier (Glorot and Bengio, 2010)	1	$\sqrt{\frac{2}{m+1}}$	$\sqrt{\frac{2}{m+d}}$	$\sqrt{\frac{4}{(m+1)(m+d)}}$	$\sqrt{\frac{m+d}{m+1}}$	1	0
NTK (Jacot et al., 2018)	$\sqrt{m}$	1	1	$\sqrt{\frac{1}{m}}$	1	$\frac{1}{2}$	0
Mean-field (Mei et al., 2018) (Sirignano and Spiliopoulos, 2020) (Rotskoff and Vanden-Eijnden, 2018)	m	1	1	$\frac{1}{m}$	1	1	0
E et al. (E et al., 2020)	1	β	1	β	β	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$

Table 1: Initialization methods with their scaling parameters

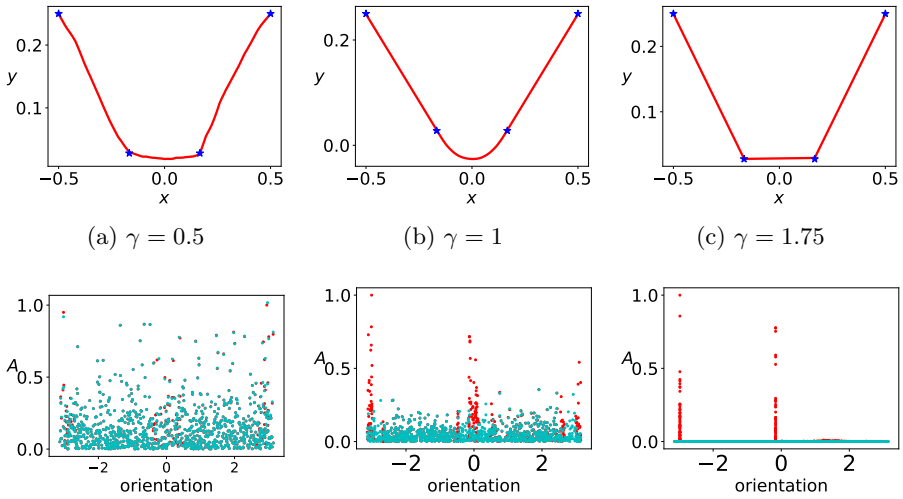
3.1 Typical Cases over the Phase Diagram

With γ and γ' as coordinates, in this subsection, we illustrate through experiments the behavior of a diversity of typical cases over the phase diagram using a simple 1-d problem of 4 training points, which allows easy visualization.

The first row in Fig. 2 shows typical learning results over different γ's, from a relatively jagged interpolation (NTK scaling) to a smooth cubic-spline-like interpolation (mean-field scaling) and further to a linear spline interpolation. To probe into details of their parameter space representation, we notice for the ReLU activation that the parameter pair $(a_k, \boldsymbol{w}_k)$ of each neuron can be separated into a unit orientation feature $\hat{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|_2$ and an amplitude $A = |a|\|\boldsymbol{w}\|_2$ indicating its contribution to the output, that is, $(A, \hat{\boldsymbol{w}})$. For the one-dimensional input, $\boldsymbol{w}$ is two dimensional due to the incorporation of bias. Therefore, we use the angle to the x-axis in $[-\pi, \pi)$ to indicate the orientation of each $\hat{\boldsymbol{w}}$. The scatter plot of $\{(A_k, \hat{\boldsymbol{w}}_k)\}_{k=1}^m$ is shown in the second row in Fig. 2. Clearly, the evolution of the parameters of the examples in the first row of Fig. 2 are different. For $\gamma = 0.5$, the initial scatter plot is very close to the one after training. However, for $\gamma = 1.75$, active neurons (i.e., neurons with significant amplitude A) are condensed at a few orientations, which strongly deviates from the initial scatter plot.

4. Phase Diagram

In this section, with γ and γ' as coordinates, we characterize at $m \to \infty$ the dynamical regimes of NNs and identify their boundaries in the phase diagram through experimental and theoretical approaches. How to characterize and classify different types of training behaviors of NNs is an important open question. Currently, a behavior of NN dynamics, by which gradient flow of the NN can be effectively linearized around initialization during the training, has been extensively studied both empirically and theoretically (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019; E et al., 2020). We refer to the regime with this behavior as the linear regime. As shown in Fig. 1, many works have proved that a



(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 1.75$

(d) $\gamma = 0.5$ (e) $\gamma = 1$ (f) $\gamma = 1.75$

Figure 2: Learning four data points by two-layer ReLU NNs with different γ's shown in the first row. The corresponding scatter plots of initial (cyan) and final (red) $\{(A_k, \hat{\boldsymbol{w}}_k)\}_{k=1}^m$ are shown in the second row. $\gamma' = 0$ ($\beta_1 = \beta_2 = 1$), hidden neuron number $m = 1000$.

specific point or line in the phase diagram belong to the linear regime. However, its exact range in the phase diagram remains unclear. On the other hand, NN training dynamics can also be highly nonlinear at $m \to \infty$ as widely studied for the mean-field model as a point shown in the phase diagram (Mei et al., 2018; Sirignano and Spiliopoulos, 2020; Rotskoff and Vanden-Eijnden, 2018). However, whether there are other points in the phase diagram that has similar training behavior is not well understood. In addition, it is not clear if there are other regimes in the phase diagram that are nonlinear but behaves distinctively comparing to the mean-field model. In the following, we will address these problems and draw the phase diagram.

4.1 Regime Identification and Separation

The linear regime refers to the set of coordinates with which the gradient flow of $f_{\boldsymbol{\theta}}$ at any t is well approximated by gradient flow of its linearized model, i.e.,

$$f_{\boldsymbol{\theta}}^{\text{lin}} = \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(0)} \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)). \tag{15}$$

Note that, the zeroth order term $f_{\boldsymbol{\theta}(0)}$ does not appear because, without loss of generality, it is always offset to 0 by the ASI trick to eliminate the extra generalization error induced by a random initial function as studied in Zhang et al. (2020). In general, this linear behavior only happens when $\boldsymbol{\theta}(t)$ always stays within a small neighbourhood of $\boldsymbol{\theta}(0)$ such that the first order Taylor expansion is a good approximation. For a two-layer NN, because its output layer is always linear w.r.t. output weights, this requirement of small neighbourhood is reduced to the one for the input weights, that is, $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ always stays within a neighbourhood

of $\boldsymbol{\theta}_{\boldsymbol{w}}(0)$. Since the size of this neighbourhood of good linear approximation scales with $\|\boldsymbol{\theta}_{\boldsymbol{w}}(0)\|_2$, therefore we use the following relative distance as an indicator of how far $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ deviates from $\boldsymbol{\theta}_{\boldsymbol{w}}(0)$ during the training

$$\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \frac{\|\theta_{\boldsymbol{w}}(t) - \theta_{\boldsymbol{w}}(0)\|_2}{\|\theta_{\boldsymbol{w}}(0)\|_2}. \tag{16}$$

Specifically, we focus on quantity $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t))$, which is the maximum deviation of $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ from initialization during the training. As $m \to \infty$, if $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to 0$, then the NN training dynamics falls into the linear regime. Otherwise, if it approaches $O(1)$ or $+\infty$, then NN training dynamics is nonlinear. Note that, for the latter case, in which $\boldsymbol{\theta}_{\boldsymbol{w}}$ deviates infinitely far away from initialization, a very strong nonlinear dynamical behavior of condensation in feature space can be observed as illustrated in Fig. 2f. We refer to the regime of $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to +\infty$ the condensed regime, which is justified latter in Sec. 4.2 by detailed experiments. For $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to O(1)$, NNs exhibit an intermediate level of nonlinear behavior. We refer to this regime as the critical regime.

In the following, we will separate exactly the linear regime and the condensed regime in the phase diagram through experimental and theoretical approaches. We first present an intuitive scaling analysis to provide a rationale for the boundary that separates these two regimes in the phase diagram. Then, we experimentally demonstrate the validity of this boundary in regime separation in the phase diagram for an 1-d data set. Finally, we establish a rigorous theory which proves the transition across this boundary for two-layer ReLU NNs at $m \to \infty$ for general data sets.

4.1.1 Intuitive Scaling Analysis

Before we jump into a detailed analysis, through an intuitive scaling analysis, we first illustrate the separation between the linear regime and the condensed regime. The capability, i.e., the magnitude of target function that can be fitted, of the two-layer ReLU NN around initialization can be roughly estimated as

$$C = m\beta_1\beta_2/\alpha = m\kappa.$$

Without loss of generality, the target function is always $O(1)$. Therefore, a necessary condition for the linear regime is that NN has the capability of fitting the target in the vicinity of initialization, i.e., $C \gtrsim O(1)$. Therefore,

$$\kappa \gtrsim 1/m,$$

yielding $\gamma \leq 1$ at $m \to \infty$. We further notice that, the output layer is always linear. Therefore, even when the output weight $\boldsymbol{\theta}_a$ changes significantly, the dynamics can still be linearized if the input layer weight $\boldsymbol{\theta}_{\boldsymbol{w}}$ stays in the vicinity of its initialization. As indicated by the dynamics Eq. (12), this is possible when (i) $\kappa' \ll 1$ at initialization and (ii) the scale of a, say quantified by expectation $\mathbb{E}(|a|)$, satisfies $\mathbb{E}(|a|) \ll \beta_2$ throughout the training. In this case, at the end of the training,

$$C = m\beta_2\mathbb{E}(|a|)/\alpha \ll m\beta_2^2/\alpha = m\kappa/\kappa'. \tag{17}$$

Because $C \gtrsim O(1)$, we got

$$1/\kappa' \gg 1/m\kappa, \tag{18}$$

which yields the condition $\gamma' > \gamma - 1$ for $\gamma' > 0$ at $m \to \infty$.

In contrary, if $\gamma' < \gamma - 1$ and $\gamma > 1$, i.e., $m\kappa \ll 1$ and $m\kappa/\kappa' \ll 1$ as $m \to \infty$, then the NN has no capability in fitting a $O(1)$ target when $\boldsymbol{\theta_w}$ stays at the vicinity of its initialization. The capability of NN must undergo a magnificent increase to be able to fit the data, which is a feature of the condensed regime.

Above scaling analysis provides an intuitive argument about the separation of linear and condensed regimes by the boundary $\gamma = 1$ for $\gamma' \leq 0$ and $\gamma' = \gamma - 1$ for $\gamma' > 0$ in the phase diagram. To further demonstrate the criticality of this boundary, we sort to the following experimental studies for a specific case.

4.1.2 Experimental Demonstration

To experimentally distinguish the linear and nonlinear regimes, we need to estimate

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta_w}(t)),$$

which empirically can be approximated by $\mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}})$ ($\boldsymbol{\theta}^*_{\boldsymbol{w}} := \boldsymbol{\theta_w}(\infty)$) without loss of generality. Next, because we can never run experiments at $m \to \infty$, we alternatively quantify the growth of $\mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}})$ as $m \to \infty$. By Fig. 3 (a-c), they approximately have a power-law relation. Therefore we define

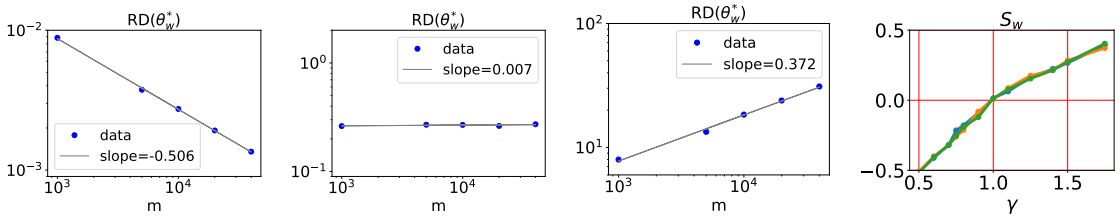
$$S_{\boldsymbol{w}} = \lim_{m\to\infty} \frac{\log \mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}})}{\log m}, \tag{19}$$

which is empirically obtained by estimating the slope in the log-log plot like in Fig. 3. As shown in Fig. 3 (d), NNs with the same pair of γ and γ', but different α, β_1, and β_2, have very similar S_w, which validates the effectiveness of the normalized model. In the following experiments, we only show result of one combination of α, β_1, and β_2 for a pair of γ and γ'.

Then, we visualize the phase diagram by experimentally scanning S_w over the phase space. The result for the same 1-d problem as in Fig. 2 is presented in Fig. 4. In the red zone, where $S_{\boldsymbol{w}}$ is less than zero, $\mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}}) \to 0$ as $m \to \infty$, indicating a linear regime. In contrast, in the blue zone, where $S_{\boldsymbol{w}}$ is greater than zero, $\mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}}) \to \infty$ as $m \to \infty$, indicating a highly nonlinear behavior. Their boundary are experimentally identified through interpolation indicated by stars in Fig. 4, where $\mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}}) \sim O(1)$. They are close to the boundary identified through the scaling analysis indicated by the auxiliary lines, justifying its criticality. Similarly, we use two-layer ReLU NNs to fit MNIST data set with mean squared loss. In our experiments, the input is a 784 dimensional vector and the output is the one-dimensional label ($0 \sim 9$) of the input image. As shown in Fig. 5, the phase diagram obtained by the synthetic data also applies for such real high-dimensional data set.

4.1.3 Theoretical Results for General Two-layer ReLU NNs

The intuitive scaling analysis and the experimental demonstration result in a consistent boundary to separate the linear and condensed regimes. A question naturally arises—is



(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 1.75$ (d) S_w vs. γ

Figure 3: Growth of $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)$ w.r.t. $m \to \infty$ with $\gamma' = 0$. For (a-c), $\beta_1 = 1$, $\beta_2 = 1$, and the plot is $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)$ vs. m of NNs with $1000, 5000, 10000, 20000, 40000$ hidden neurons indicated by five blue dots, respectively. The gray line is a linear fit with slope indicated. For (d), the plot is $S_{\boldsymbol{w}}$ vs. γ for $\gamma' = 0$. Each line is for a pair of β_1 and β_2: Blue: $\beta_1 = 1$, $\beta_2 = 1$; Orange: $\beta_1 = m^{-1/2}$, $\beta_2 = m^{-1/2}$; Green: $\beta_1 = m^{-1}$, $\beta_2 = m^{-1}$. Note that α is determined by $\alpha = \beta_1\beta_2 m^{\gamma}$.

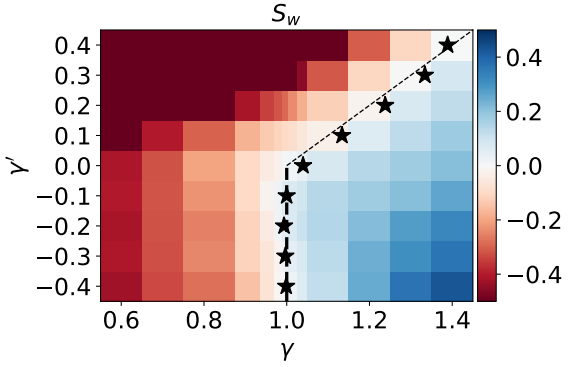


Figure 4: For synthetic data, $S_{\boldsymbol{w}}$ estimated on two-layer ReLU NNs of 1000, 5000, 10000, 20000, 40000 hidden neurons over γ (ordinate) and γ' (abscissa). The stars are zero points obtained by the linear interpolation over different γ for each fixed γ'. Dashed lines are auxiliary lines indicating the theoretically obtained boundary.

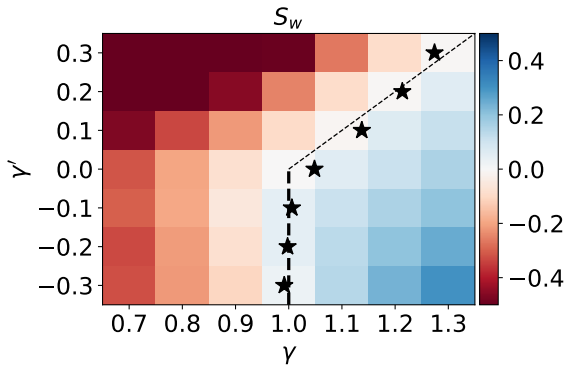


Figure 5: For MNIST data, $S_{\boldsymbol{w}}$ estimated on two-layer ReLU NNs of 1000, 10000, 50000, 250000, 40000 hidden neurons over γ (ordinate) and γ' (abscissa). The stars are zero points obtained by the linear interpolation over different γ for each fixed γ'. Dashed lines are auxiliary lines indicating the theoretically obtained boundary.

there a theory that makes the intuitive scaling analysis rigorous and generalizes above empirical phase diagram for an 1-d example to general high-dimensional data for two-layer ReLU NNs. In the following, we address this question by providing two theorems in informal statements, which proves the criticality of $\lim_{m\to+\infty} \sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t))$ at above identified boundary in the phase diagram. Their rigorous statements can be found in Section 5.

Theorem 1*. *(Informal statement of Theorem 6) If $\gamma < 1$ or $\gamma' > \gamma - 1$, then with a high probability over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{m\to+\infty} \sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = 0. \tag{20}$$

Theorem 2*. *(Informal statement of Theorem 8) If $\gamma > 1$ and $\gamma' < \gamma - 1$, then with a high probability over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{m\to+\infty} \sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = +\infty. \tag{21}$$

Remark 3. $\lim_{m\to+\infty} \sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t))$ *is like an order parameter in the analysis of phase transition in statistical mechanics, which is key to the regime separation and exhibits discontinuity at the boundary.*

In Theorem 1*, focusing on the linear regime, the negligible relative change of $\boldsymbol{w}$ is essentially proved by showing the kernel of the training dynamics undergoes no significant change during the whole dynamics. However, the kernel of the training dynamics might be out of control for the condensed regime. This difficulty makes the result of Theorem 2* nontrivial. Instead of studying the kernel, more detailed information of the dynamics should be used. Indeed, we establish a neural-wise estimate, $|a_k(t)| \leq \frac{1}{\kappa'}\|\boldsymbol{w}_k(t)\|_2 + |a_k^0|$,

which holds for any κ, κ' and any $t \geq 0$. We believe that this estimate can be extended to other network structures and general activation functions for the regimes of nonlinear dynamics.

Above two theorems complete the phase diagram of two-layer ReLU NN with distinctive dynamical regimes separated based on $\lim_{m\to+\infty} \sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t))$. The behavior of NN in the linear regime, e.g., exponential decay of loss, implicit regularization in terms of a RKHS norm, and etc., is very well studied. However, the critical and condense regimes is largely not understood. In the following, we make a further step to unravel a signature nonlinear behavior—condensation as $m \to \infty$ through experiments, which sheds light on future theoretical study.

4.2 Critical and Condensed Regimes

By Fig. 2 (d-f) in previous section, it can be observed that the condensation of NN representation in feature space $\{(A_k, \hat{\boldsymbol{w}}_k)\}_{k=1}^m$ comparing to initialization is a distinctive feature for the nonlinear training dynamics of NNs. Specifically, we care about this condensation at the $m \to \infty$ limit when relative change of $\boldsymbol{\theta}_w$ approaches $+\infty$. Therefore, using the same 1-d data as in Fig. 2, we scan the learned distribution of $(A_k, \hat{\boldsymbol{w}}_k)$ pair for $m = 10^3, 10^4, 10^6$ over the phase diagram to experimentally find out the limiting behavior. The result is shown in Fig. 6 with the corresponding $S_{\boldsymbol{w}}$ shown in Fig. 4. It is easy to observe that, right to the boundary indicated by blue boxes, the condensation becomes stronger for larger $S_{\boldsymbol{w}}$ and as $m \to \infty$, implying a delta-function-like condensation behavior in the limit. This conforms with our intuition that the farther away $\boldsymbol{\theta}_{\boldsymbol{w}}$ deviates from initialization, the stronger nonlinearity of NNs exhibited here in the form of condensation. Therefore, as introduced before, we refer to this regime as the *condensed* regime. In the critical regime as the boundary between the linear and the condensed regimes, the level of condensation is almost fixed as $m \to \infty$, which resembles a mean-field behavior. Indeed, the well-studied mean-field model is one point in the critical regime shown in the phase diagram Fig. 1. In general, the mechanism of condensation as well as its implicit regularization effect is not well understood, which remain as important open questions for the future research.

We also examine the condensation of NNs for MNIST data set. For such high-dimensional data, it is impossible to directly visualize the distribution in the high-dimensional feature space like above 1-d case. Therefore, we consider a projection approach, by which we project each $\hat{\boldsymbol{w}}$ to a reference direction $\boldsymbol{p}$ and plot A_k vs. $I_{\hat{\boldsymbol{w}}} = \hat{\boldsymbol{w}} \cdot \boldsymbol{p}$. Note that the reference direction can be arbitrary selected and does not affect our conclusion. Without loss of generality, we pick $\boldsymbol{p} = \mathbf{1}/\sqrt{n}$. Clearly, if neurons indeed condensed at several directions in the high-dimensional feature space, then their 1-d projection should also condense at several points. As shown in Fig. 7 (corresponding $S_{\boldsymbol{w}}$ is shown in Fig. 5), similar to the 1-d case, condensation behavior can be observed in the condensed regime identified above. As the parameters move further away from the boundary in the condensed regime, condensation becomes more salient.

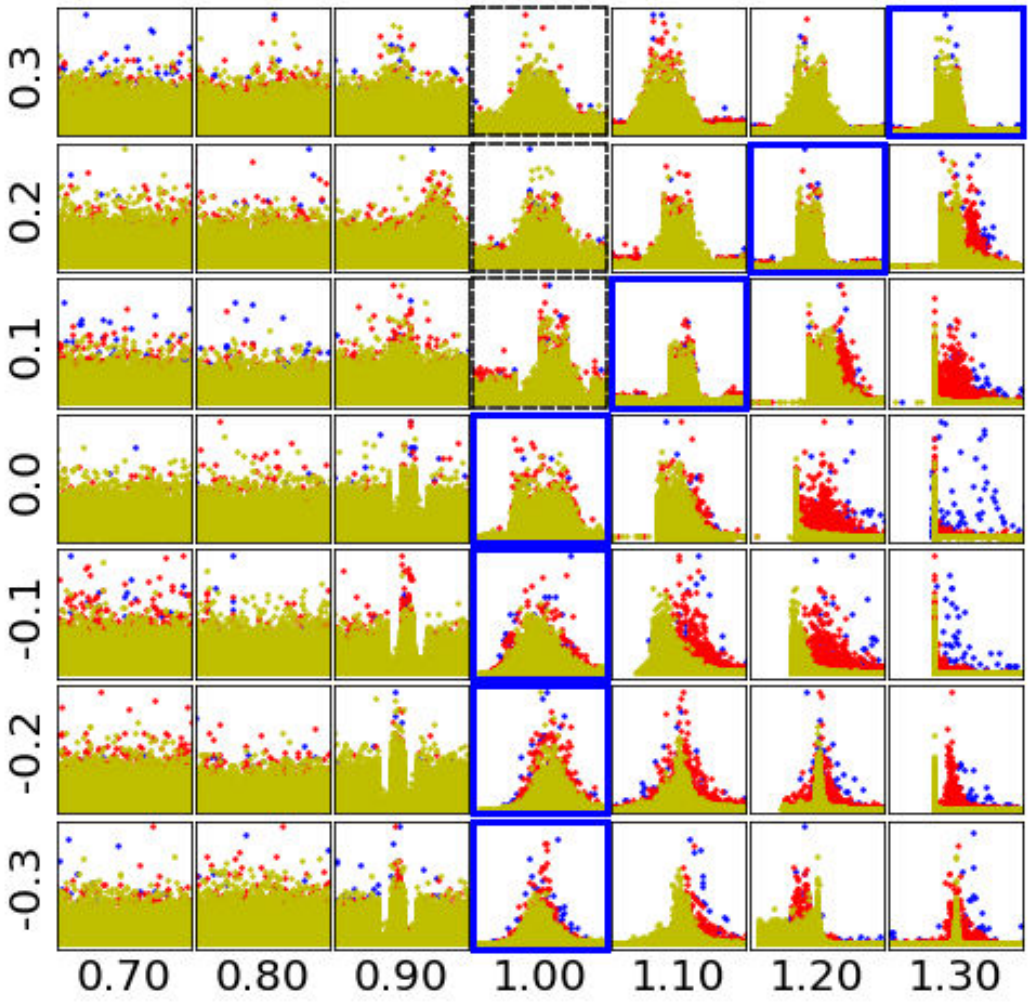


Figure 6: Condensation map for 1-d synthetic data. Each color in each box is a scatter of $\{(A_k, \hat{\boldsymbol{w}}_k)\}_{k=1}^m$ for a NN with corresponding γ and γ'. The hidden neuron numbers are: $m = 10^3$ (blue), 10^4 (red), 10^6 (yellow). The abscissa coordinate is γ and the ordinate one is γ'. Note that the corresponding $S_{\boldsymbol{w}}$ is shown in Fig. 4.

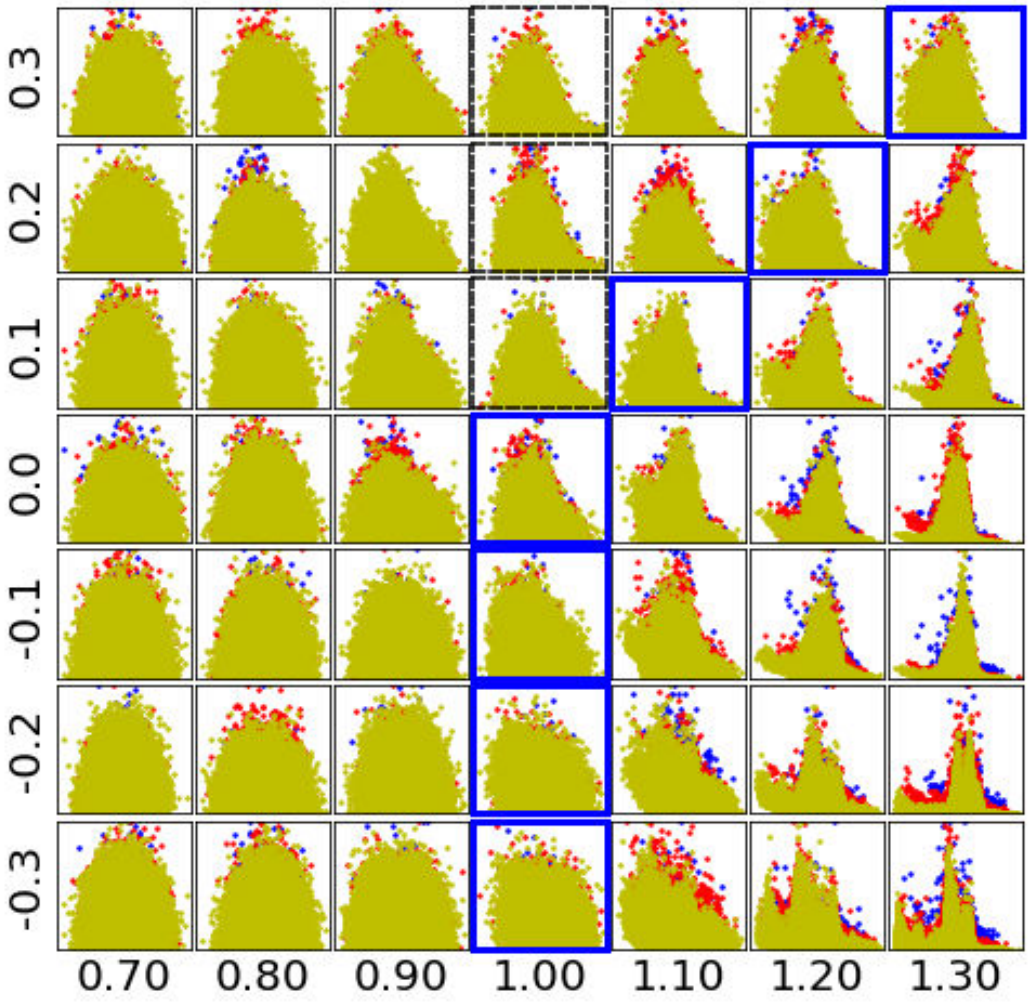


Figure 7: Condensation map for MNIST data set. Each color in each box is a scatter of $\{(A_k, I_{\hat{\boldsymbol{w}}}\}_{k=1}^{m}$ for a NN with corresponding γ and γ'. The hidden neuron numbers are: $m = 10^3$ (blue), 10^4 (red), 2.5×10^5 (yellow). The abscissa coordinate is γ and the ordinate one is γ'. Note that the corresponding $S_{\boldsymbol{w}}$ is shown in Fig. 5.

5. Theoretical Regime Characterization

We illustrate above how our phase diagram Fig. 1 is obtained through experimental and theoretical approaches. To obtain a more detailed understanding of general properties of these regimes, we present our theoretical results in detail in this section, which follows a rigorous description of our notations and definitions in the beginning. The proofs can be found in the appendix.

To start with, let us consider a two layer neural network

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{1}{\kappa} f_{\boldsymbol{\theta}}^{\kappa}(\boldsymbol{x}) = \sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^{\intercal} \boldsymbol{x}), \tag{22}$$

with the activation function $\sigma(z) = \mathrm{ReLU}(z) = \max(z, 0)$. Denote the data set

$$S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}, \tag{23}$$

where $\boldsymbol{x}_i$'s are i.i.d. sampled from the (unknown) distribution $\mathcal{D}$ over $\Omega = [0,1]^d$ with $(\boldsymbol{x}_i)_d = 1$ and $y_i = f(\boldsymbol{x}_i) \in [0,1]$ for all $i \in [n]$.

We denote $e_i = \kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i = \kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)$, $i \in [n]$ and $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^{\intercal}$. Then the empirical risk can be written as

$$R_S(\boldsymbol{\theta}) := R_{S,\kappa}(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (\kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2 = \frac{1}{2n} \boldsymbol{e}^{\intercal} \boldsymbol{e}. \tag{24}$$

Its gradient flow is

$$\dot{\boldsymbol{\theta}} = -M_{\kappa'} \nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}), \tag{25}$$

with a more explicit form for a_k and $\boldsymbol{w}_k$ respectively

$$\begin{cases} \dot{a}_k = -\dfrac{1}{\kappa'} \nabla_{a_k} R_S(\boldsymbol{\theta}) = -\dfrac{\kappa}{\kappa' n} \displaystyle\sum_{i=1}^{n} e_i \sigma(\boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i), \\ \dot{\boldsymbol{w}}_k = -\kappa' \nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}) = -\dfrac{\kappa \kappa'}{n} \displaystyle\sum_{i=1}^{n} e_i a_i \sigma'(\boldsymbol{w}_k^{\intercal} \boldsymbol{x}_i) \boldsymbol{x}_i. \end{cases} \tag{26}$$

Here κ, κ' are scaling parameters proposed in Section 3. The parameters are initialized as

$$a_k^0 := a_k(0) \sim N(0,1), \tag{27}$$

$$\boldsymbol{w}_k^0 := \boldsymbol{w}_k(0) \sim N(0, \boldsymbol{I}_d), \tag{28}$$

$$\boldsymbol{\theta}^0 := \boldsymbol{\theta}(0) = \mathrm{vec}(\{a_k^0, \boldsymbol{w}_k^0\}_{k=1}^m). \tag{29}$$

The kernels $k^{[a]}$ and $k^{[\boldsymbol{w}]}$ of the GD dynamics are

$$\begin{aligned} k^{[a]}(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}_{\boldsymbol{w}} \sigma(\boldsymbol{w}^{\intercal} \boldsymbol{x}) \sigma(\boldsymbol{w}^{\intercal} \boldsymbol{x}'), \\ k^{[\boldsymbol{w}]}(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}_{(a,\boldsymbol{w})} a^2 \sigma'(\boldsymbol{w}^{\intercal} \boldsymbol{x}) \sigma'(\boldsymbol{w}^{\intercal} \boldsymbol{x}') \boldsymbol{x} \cdot \boldsymbol{x}'. \end{aligned} \tag{30}$$

The Gram matrices $\boldsymbol{K}^{[a]}$ and $\boldsymbol{K}^{[\boldsymbol{w}]}$ of an infinite width two-layer network are

$$\begin{aligned} K_{ij}^{[a]} &= k^{[a]}(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad \boldsymbol{K}^{[a]} = (K_{ij}^{[a]})_{n\times n}, \\ K_{ij}^{[\boldsymbol{w}]} &= k^{[\boldsymbol{w}]}(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad \boldsymbol{K}^{[\boldsymbol{w}]} = (K_{ij}^{[\boldsymbol{w}]})_{n\times n}. \end{aligned} \tag{31}$$

For the simplicity of proof, we now define the normalized Gram matrices $\boldsymbol{G}^{[a]}$, $\boldsymbol{G}^{[\boldsymbol{w}]}$, and $\boldsymbol{G}$ for a finite width two-layer network as follows:

$$\begin{aligned} G_{ij}^{[a]}(\boldsymbol{\theta}) &= \frac{1}{\kappa' m}\sum_{k=1}^{m} \nabla_{a_k}\kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\cdot\nabla_{a_k}\kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) = \frac{\kappa^2}{\kappa' m}\sum_{k=1}^{m}\sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_j), \\ G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}) &= \frac{\kappa'}{m}\sum_{k=1}^{m} \nabla_{\boldsymbol{w}_k}\kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\cdot\nabla_{\boldsymbol{w}_k}\kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) = \frac{\kappa^2\kappa'}{m}\sum_{k=1}^{m} a_k^2\sigma'(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\sigma'(\boldsymbol{w}_k^\intercal \boldsymbol{x}_j)\boldsymbol{x}_i\cdot\boldsymbol{x}_j, \\ \boldsymbol{G} &= \boldsymbol{G}^{[a]} + \boldsymbol{G}^{[\boldsymbol{w}]}. \end{aligned} \tag{32}$$

Assumption 1. *Suppose that the Gram matrices are strictly positive definite. In other words,*

$$\lambda := \min\{\lambda_a, \lambda_{\boldsymbol{w}}\} > 0, \tag{33}$$

where

$$\lambda_a := \lambda_{\min}\left(\boldsymbol{K}^{[a]}\right), \quad \lambda_{\boldsymbol{w}} := \lambda_{\min}\left(\boldsymbol{K}^{[\boldsymbol{w}]}\right). \tag{34}$$

We remark that if for any $i \neq j$, $\boldsymbol{x}_i \nparallel \boldsymbol{x}_j$, then Assumption 1 holds. This follows the proof of (Du et al., 2019, Theorem 3.1). Indeed, we do have the fact that, for any $i \neq j$, if $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ then $\boldsymbol{x}_i \nparallel \boldsymbol{x}_j$, because in our paper the notation $\boldsymbol{x}_i$ means the augmented vector $(\boldsymbol{x}_i^\intercal, 1)^\intercal$ (See the first paragraph in Section 1).

Assumption 2. *Suppose that the following limits exist*

$$\gamma := \lim_{m\to\infty} -\frac{\log\kappa}{\log m}, \quad \gamma' := \lim_{m\to\infty} -\frac{\log\kappa'}{\log m}. \tag{35}$$

Remark 4. *We expect that*

$$\boldsymbol{G}^{[a]}(\boldsymbol{\theta}^0) \approx \frac{\kappa^2}{\kappa'}\boldsymbol{K}^{[a]}, \quad \boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}^0) \approx \kappa^2\kappa'\boldsymbol{K}^{[\boldsymbol{w}]}, \tag{36}$$

and these will be rigorously achieved in the following proofs. We also remark that $\lambda \leq d$, which will be used in the following proofs.

Remark 5. *When $\gamma \leq \frac{1}{2}$, we consider NNs with non-zero initial parameters and zero initial output, which can be achieved in NNs by applying the AntiSymmetrical Initialization (ASI) trick (Zhang et al., 2020).*

Our main results are as follows.

Theorem 6 (linear regime)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 and Assumption 2 hold. ASI is used when $\gamma \leq \frac{1}{2}$. Suppose that $\gamma < 1$ or $\gamma' > \gamma - 1$ and the dynamics (26)–(29) is considered. Then for sufficiently large m, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

(a) (changes of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\boldsymbol{w}}$)

$$\sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2 \le \sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2 \lesssim \frac{1}{\sqrt{m\kappa}}\log m. \tag{37}$$

(b) (linear convergence rate)

$$R_S(\boldsymbol{\theta}(t)) \le \exp\left(-\frac{2m\kappa^2\lambda t}{n}\right) R_S(\boldsymbol{\theta}^0). \tag{38}$$

Moreover, for sufficiently large m, with probability at least $1-\delta-2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have

(c) (relative change of $\boldsymbol{\theta}$)

$$\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} \lesssim \frac{1}{m\kappa}\log m. \tag{39}$$

In particular, if $\gamma<1$, $\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} \ll 1$.

(d) (relative change of $\boldsymbol{\theta}_{\boldsymbol{w}}$)

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2} \lesssim \begin{cases} \frac{1}{m\kappa}\log m, & \gamma<1, \\ \frac{\kappa^{\gamma'}}{m\kappa}\log m, & \gamma'>\gamma-1. \end{cases} \tag{40}$$

In particular, if either $\gamma<1$ or $\gamma'>\gamma-1$, $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2} \ll 1$.

Remark 7. *In the regions $\gamma<1$ or $\gamma'>\gamma-1$, Theorem 6 shows that with a high probability over the initialization, the relative changes of $\boldsymbol{w}_k$'s are negligible. This implies that the features change only slightly during the whole gradient flow dynamics. Therefore, in this regime and with large width m, one can expect the training result to be close to that of some proper linear regression model. Note that the relative change of $\boldsymbol{\theta}$ is negligible only in the sub-region $\gamma<1$. For $\gamma\ge 1$ and $\gamma'>\gamma-1$, the relative changes of a_k's can be fairly large, which may lead to unbounded relative change of $\boldsymbol{\theta}$. The relative changes of $\boldsymbol{\theta}$ and a_k's are also empirically validated in Appendix D.*

In order to obtain the theorem that characterize the condensed regime, we need further assumption as follows,

Assumption 3. *We assume that, without loss of generality,*

$$\max_{i\in[n]} y_i \ge \frac{1}{2}, \tag{41}$$

and that the neural network can be well-trained to the empirical risk less than $O(\frac{1}{n})$. More quantitatively, we require that there exists a $T^*>0$ such that

$$R_S(\boldsymbol{\theta}(T^*)) \le \frac{1}{32n}. \tag{42}$$

Then we can get the following theorem

Theorem 8 (condensed regime). *The sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 2 and Assumption 3 hold. Suppose that $\gamma > 1$ and $\gamma' < \gamma - 1$ and the dynamics (26)–(29) is considered. Then for sufficiently large m, with probability at least $1 - 2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have*

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2} \gg 1. \tag{43}$$

To end this section, we provide a sketch of the proofs for the main theorems. In particular, two schematic diagrams 8 and 9 are provided for the proofs of Theorem 6 (Theorem 1*) and Theorem 8 (Theorem 2*), respectively, since they are proved in totally different ways. For Theorem 6, we first establish bounds and concentration inequalities for initial parameters. Then the lower bound for the minimal eigenvalue of initial Gram matrix is obtained, which leads to a local in time linear convergence result for the empirical risk. Finally, for sufficiently wide neural networks, we show that the previous estimate is essentially global in time. We remark that for different (γ, γ')'s, the details are quite different in the proofs of Theorem 6, which causes the two branches shown in Figure 8. For Theorem 8, as shown in Figure 9, the schematic diagram of the proof is short and straightforward, thanks to a key observation of the neuron-wise estimate, i.e., Proposition 27.

6. Conclusions and Discussion

In this paper, we characterized the linear, critical, and condensed regimes with distinctive features and draw the phase diagram for the two-layer ReLU NN at the infinite-width limit. We experimentally demonstrate and theoretically prove the transition across the boundary (critical regime) in the phase diagram. Through experiments, we further identify the condensation as the signature behavior in the condensed regime of very strong nonlinearity. Note that, the phenomenon of condensation in a broad sense is observed in several early studies of nonlinear training dynamics of NNs in different settings (Maennel et al., 2018; Chizat and Bach, 2018; Ma et al., 2020).

A phase diagram serves as a map that guides the future research. In our phase diagram for two-layer ReLU NNs, the linear regimes is very well understood both theoretically and experimentally. However, the critical and condensed regimes are still largely not understood from both experimental and theoretical perspectives. The following problems for these regimes requires further studies: (i) whether the dynamics always converges to a global minimizer; (ii) what is the convergence rate; (iii) what is the mechanism of condensation; (iv) how to characterize the implicit regularization of condensation.

Our phase diagram is obtained specifically for the ReLU activation, however, our methodology and thus obtained regime characterization can be naturally extended to more general activations, which is an immediate next step of this work. In addition, how to characterize the effect of other hyperparameters, e.g., choice of optimization method, learning rate, regularization techniques, and etc., to the NN training dynamics requires future studies.

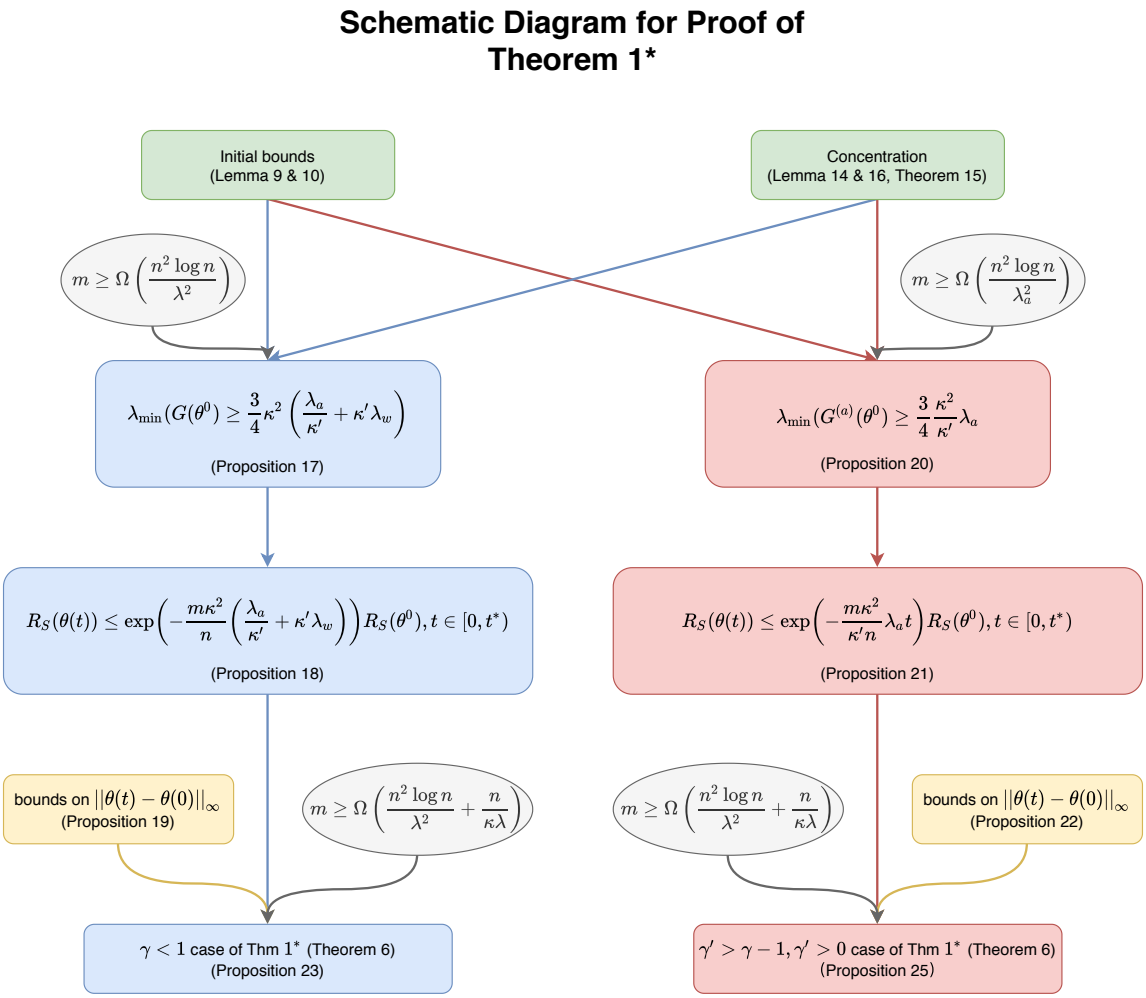


Figure 8: Sketch of proof for Theorem 1*.

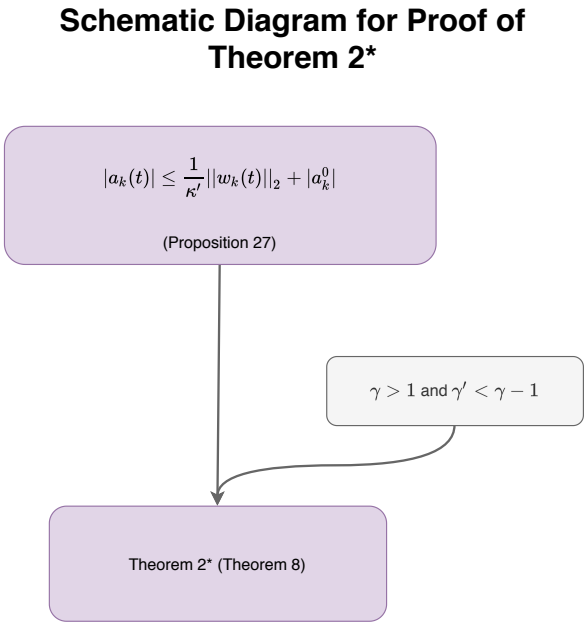


Figure 9: Sketch of proof for Theorem 2*.

Other important future problems include drawing the phase diagram for NNs of three or more layers or for convolutional networks.

In analogy to statistical mechanics, a clean regime separation may be only possible at the infinite width limit, which is not realistic in practice. Nevertheless, rich insight about a finite size system often can be derived from the analysis at the limit, which is usually much easier. Therefore, we believe it is an important task to systematically draw such phase diagrams for NNs of different structures through a combination of theoretical and experimental approaches as demonstrated in this work. These phase diagrams can be continuously refined and provides a clear pathway to open the black box of deep learning.

Acknowledgments

This work is sponsored by the National Key R&D Program of China Grant No. 2019YFA0709503 (Z. X.), 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.), Shanghai Municipal of Science and Technology Project Grant No. 20JC1419500 (Y.Z.), and the HPC of School of Mathematical Sciences and the Student Innovation Center at Shanghai Jiao Tong University.

Appendix A. Technical Lemmas

This section collects some technical lemmas and propositions. For convenience, we define the two quantities

$$\alpha(t) := \max_{k\in[m],s\in[0,t]} |a_k(s)|, \quad \omega(t) := \max_{k\in[m],s\in[0,t]} \|\boldsymbol{w}_k(s)\|_\infty. \tag{44}$$

Lemma 9 (bounds of initial parameters)**.** *Given $\delta \in (0,1)$, we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$*

$$\max_{k\in[m]} \left\{|a_k^0|,\ \|\boldsymbol{w}_k^0\|_\infty\right\} \le \sqrt{2\log\frac{2m(d+1)}{\delta}}, \tag{45}$$

Proof If $\mathrm{X} \sim N(0,1)$, then $\mathbb{P}(|\mathrm{X}| > \varepsilon) \le 2\mathrm{e}^{-\frac{1}{2}\varepsilon^2}$ for all $\varepsilon > 0$. Since $a_k^0 \sim N(0,1)$, $(w_k^0)_\alpha \sim N(0,1)$ for $k = 1,2,\dots,m,\ \alpha = 1,\dots,d$ and they are all independent, by setting

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

one can obtain

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

■

Lemma 10 (bound of initial empirical risk)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. We have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$*

$$R_S(\boldsymbol{\theta}^0) \le \frac{1}{2}\left[1 + 2d\left(\log\frac{4m(d+1)}{\delta}\right)\left(2 + 3\sqrt{2\log(8/\delta)}\right)\kappa\sqrt{m}\right]^2. \tag{46}$$

Proof Let

$$\mathcal{G} = \{g_{\boldsymbol{x}}(a,\boldsymbol{w}) \mid g_{\boldsymbol{x}}(a,\boldsymbol{w}) := a\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}), \boldsymbol{x} \in \Omega\}. \tag{47}$$

Lemma 9 implies that with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$, we have

$$|g_{\boldsymbol{x}}(a_k^0, \boldsymbol{w}_k^0)| \le d|a_k^0|\|\boldsymbol{w}_k^0\| \le 2d\log\frac{4m(d+1)}{\delta}$$

Then

$$\frac{1}{m}\sup_{\boldsymbol{x}\in\Omega}|f_{\boldsymbol{\theta}^0}(\boldsymbol{x})| = \sup_{\boldsymbol{x}\in\Omega}\left|\frac{1}{m}\sum_{k=1}^{m}a_k^0\sigma(\boldsymbol{w}_k^0\cdot\boldsymbol{x}) - \mathbb{E}_{(a,\boldsymbol{w})}a\sigma(\boldsymbol{w}^\intercal\boldsymbol{x})\right| \\ \leq 2\mathrm{Rad}_{\boldsymbol{\theta}^0}(\mathcal{G}) + 6d\left(\log\frac{4m(d+1)}{\delta}\right)\sqrt{\frac{2\log(8/\delta)}{m}}.$$

The Rademacher complexity can be estimated by

$$\begin{aligned}\mathrm{Rad}_{\boldsymbol{\theta}^0}(\mathcal{G}) &= \frac{1}{m}\mathbb{E}_\tau\left[\sup_{\boldsymbol{x}\in\Omega}\sum_{k=1}^{m}\tau_k a_k^0\sigma(\boldsymbol{w}_k^0\cdot\boldsymbol{x})\right] \\ &\leq \frac{1}{m}\sqrt{2\log\frac{4m(d+1)}{\delta}}\mathbb{E}_\tau\left[\sup_{\boldsymbol{x}\in\Omega}\sum_{k=1}^{m}\tau_k\boldsymbol{w}_k^0\cdot\boldsymbol{x}\right] \\ &\leq \sqrt{2\log\frac{4m(d+1)}{\delta}}\sqrt{2d\log\frac{4m(d+1)}{\delta}}\frac{\sqrt{d}}{\sqrt{m}} \\ &= \frac{2d\log\frac{4m(d+1)}{\delta}}{\sqrt{m}}.\end{aligned}$$

Therefore

$$\sup_{\boldsymbol{x}\in\Omega}|f_{\boldsymbol{\theta}^0}(\boldsymbol{x})| \leq 2d\left(\log\frac{4m(d+1)}{\delta}\right)(2+3\sqrt{2\log(8/\delta)}\sqrt{m}),$$

and

$$\begin{aligned}R_S(\boldsymbol{\theta}^0) &\leq \frac{1}{2n}\sum_{i=1}^{n}(1+\kappa|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)|)^2 \\ &\leq \frac{1}{2}\left[1+2d\left(\log\frac{4m(d+1)}{\delta}\right)\left(2+3\sqrt{2\log(8/\delta)}\kappa\sqrt{m}\right)\right]^2.\end{aligned}$$

■

Remark 11. *If $\gamma > \frac{1}{2}$, then $\kappa = o(\frac{1}{\sqrt{m}\log m})$ and $R_S(\boldsymbol{\theta}^0) = O(1)$. One can use ASI trick (Zhang et al., 2020) to guarantee $R_S(\boldsymbol{\theta}^0) \leq \frac{1}{2}$ for any κ.*

Next we introduce the sub-exponential norm of a random variable and the sub-exponential Bernstein's inequality.

Definition 12 (sub-exponential norm (Vershynin, 2018))**.** *The sub-exponential norm of a random variable* X *is defined as*

$$\|\mathrm{X}\|_{\psi_1} := \inf\{s>0 \mid \mathbb{E}_{\mathrm{X}}[\mathrm{e}^{|\mathrm{X}|/s}] \leq 2\}. \tag{48}$$

In particular, we denote the sub-exponential norm of a $\chi^2(d)$ random variable X *by $C_{\psi,d} := \|\mathrm{X}\|_{\psi_1}$. Here the χ^2 distribution with d degrees of freedom has the probability density function*

$$f_{\mathrm{X}}(z) = \frac{1}{2^{d/2}\Gamma(d/2)}z^{d/2-1}\mathrm{e}^{-z/2}.$$

Remark 13. *Note that*

$$\mathbb{E}_{\mathrm{X}\sim\chi^2(d)}\mathrm{e}^{|\mathrm{X}|/2} = \int_0^{+\infty} \frac{1}{2^{d/2}\Gamma(d/2)} z^{d/2-1}\,\mathrm{d}z = +\infty,$$
$$\lim_{s\to+\infty} \mathbb{E}_{\mathrm{X}\sim\chi^2(d)}\mathrm{e}^{|\mathrm{X}|/s} = \lim_{s\to+\infty}\int_0^{+\infty} \frac{1}{2^{d/2}\Gamma(d/2)} z^{d/2-1}\mathrm{e}^{-z/2+z/s}\,\mathrm{d}z = 1 < 2.$$

These imply that $2 \le C_{\psi,d} < +\infty$.

Lemma 14. *Suppose that* $\boldsymbol{w} \sim N(0, \boldsymbol{I}_d)$, $a \sim N(0,1)$ *and given* $\boldsymbol{x}_i, \boldsymbol{x}_j \in \Omega$. *Then we have*

(i) if $\mathrm{X} := \sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma(\boldsymbol{x}\cdot\boldsymbol{x}_j)$, *then* $\|\mathrm{X}\|_{\psi_1} \le dC_{\psi,d}$.

(ii) if $\mathrm{X} := a^2\sigma'(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma'(\boldsymbol{w}^\intercal \boldsymbol{x}_j)\boldsymbol{x}_i\cdot\boldsymbol{x}_j$, *then* $\|\mathrm{X}\|_{\psi_1} \le dC_{\psi,d}$.

Proof Let $\mathrm{Z} := \|\boldsymbol{w}\|_2^2 = \chi^2(d)$.
(i) $|\mathrm{X}| \le d\|\boldsymbol{w}\|_2^2 = d\mathrm{Z}$ and

$$\begin{aligned}
\|\mathrm{X}\|_{\psi_1} &= \inf\{s>0 \mid \mathbb{E}_{\mathrm{X}}\exp(|\mathrm{X}|/s) \le 2\} \\
&= \inf\{s>0 \mid \mathbb{E}_{\boldsymbol{w}}\exp\left(|\sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_j)|/s\right) \le 2\} \\
&\le \inf\{s>0 \mid \mathbb{E}_{\boldsymbol{w}}\exp(d\|\boldsymbol{w}\|_2^2/s) \le 2\} \\
&= \inf\{s>0 \mid \mathbb{E}_{\mathrm{Z}}\exp(d|\mathrm{Z}|/s) \le 2\} \\
&= d\inf\{s>0 \mid \mathbb{E}_{\mathrm{Z}}\exp(|\mathrm{Z}|/s) \le 2\} \\
&= d\|\chi^2(d)\|_{\psi_1} \\
&\le dC_{\psi,d}.
\end{aligned}$$

(ii) $|\mathrm{X}| \le d|a|^2 \le d\mathrm{Z}$ and $\|\mathrm{X}\|_{\psi_1} \le dC_{\psi,d}$. ■

Theorem 15 (sub-exponential Bernstein's inequality (Vershynin, 2018))**.** *Suppose that* $\mathrm{X}_1, \ldots, \mathrm{X}_m$ *are i.i.d. sub-exponential random variables with* $\mathbb{E}\mathrm{X}_1 = \mu$, *then for any* $s \ge 0$ *we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^m \mathrm{X}_k - \mu\right| \ge s\right) \le 2\exp\left(-C_0 m \min\left(\frac{s^2}{\|\mathrm{X}_1\|_{\psi_1}^2}, \frac{s}{\|\mathrm{X}_1\|_{\psi_1}}\right)\right), \tag{49}$$

where C_0 *is an absolute constant.*

Proposition 16 (norm of initial parameters)**.** *Given* $\delta \in (0,1)$, *we have with probability at least* $1 - 2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$ *over the choice of* $\boldsymbol{\theta}^0$

$$\sqrt{\frac{m(d+1)}{2}} \le \|\boldsymbol{\theta}^0\|_2 \le \sqrt{\frac{3m(d+1)}{2}}, \tag{50}$$
$$\sqrt{\frac{md}{2}} \le \|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2 \le \sqrt{\frac{3md}{2}}. \tag{51}$$
$$\sqrt{\frac{m}{2}} \le \|\boldsymbol{\theta}_a^0\|_2 \le \sqrt{\frac{3m}{2}}. \tag{52}$$

Proof Let $\mathrm{X}_1, \ldots, \mathrm{X}_{m(d+1)}$ be the squares of the entries of $\boldsymbol{\theta}^0$, which are drawn i.i.d. from $\chi^2(1)$. The latter is sub-exponential and $\mathbb{E}\mathrm{X}_k = 1$. Then by Theorem 15

$$\mathbb{P}\left(\left|\frac{1}{m(d+1)}\sum_{k=1}^{m(d+1)}\mathrm{X}_k - 1\right| \geq s\right) \leq 2\exp\left(-C_0 m(d+1)\min\left(\frac{s^2}{C_{\psi,1}^2}, \frac{s}{C_{\psi,1}}\right)\right).$$

Setting $s = \frac{1}{2}$, we have $\frac{s}{C_{\psi,1}} \leq \frac{1/2}{2} < 1$ and

$$\begin{aligned}\mathbb{P}\left(\frac{1}{2} \leq \frac{1}{m(d+1)}\sum_{k=1}^{m(d+1)}\mathrm{X}_k \leq \frac{3}{2}\right) &\leq 2\exp\left(-C_0 m(d+1)\min\left(\frac{1}{4C_{\psi,1}^2}, \frac{1}{2C_{\psi,1}}\right)\right) \\ &= 2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right).\end{aligned}$$

Therefore, with probability at least $1 - 2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$,

$$\frac{1}{2} \leq \frac{1}{m(d+1)}\sum_{k=1}^{m(d+1)}\mathrm{X}_k = \frac{1}{m(d+1)}\|\boldsymbol{\theta}^0\|_2^2 \leq \frac{3}{2}.$$

In other words, (50) holds. The proofs of (51) and (52) are similar. ■

Proposition 17 (minimal eigenvalue of Gram matrix $\boldsymbol{G}$ at initial)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. If $m \geq \frac{16n^2d^2C_{\psi,d}}{C_0\lambda^2}\log\frac{4n^2}{\delta}$ then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta}^0)\right) \geq \frac{3}{4}\kappa^2\left(\frac{1}{\kappa'}\lambda_a + \kappa'\lambda_{\boldsymbol{w}}\right). \tag{53}$$

Proof For any $\varepsilon > 0$, we define

$$\begin{aligned}\Omega_{ij}^{[a]} &:= \left\{\boldsymbol{\theta}^0 \mid \left|\frac{\kappa'}{\kappa^2}G_{ij}^{[a]}(\boldsymbol{\theta}^0) - K_{ij}^{[a]}\right| \leq \frac{\varepsilon}{n}\right\}, \\ \Omega_{ij}^{[\boldsymbol{w}]} &:= \left\{\boldsymbol{\theta}^0 \mid \left|\frac{1}{\kappa^2\kappa'}G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}^0) - K_{ij}^{[\boldsymbol{w}]}\right| \leq \frac{\varepsilon}{n}\right\}.\end{aligned}$$

By Theorem 15 and Lemma 14, if $\frac{\varepsilon}{ndC_{\psi,d}} \leq 1$, then

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

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

$$\left\|\frac{\kappa'}{\kappa^2}\boldsymbol{G}^{[a]}(\boldsymbol{\theta}^0)-\boldsymbol{K}^{[a]}\right\|_{\mathrm{F}}\leq\varepsilon,$$
$$\left\|\frac{1}{\kappa^2\kappa'}\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}^0)-\boldsymbol{K}^{[\boldsymbol{w}]}\right\|_{\mathrm{F}}\leq\varepsilon,$$

Hence by taking $\varepsilon=\lambda/4$, that is, $\delta=4n^2\exp\left(-\frac{mC_0\lambda^2}{16n^2d^2C_{\psi,d}^2}\right)$

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

We remark that for $\varepsilon=\lambda/4$, we have $\frac{\varepsilon}{ndC_{\psi,d}}=\frac{\lambda}{8nd}\leq\frac{1}{8}<1$. ■

In the following we denote

$$t^*=\inf\{t\mid\boldsymbol{\theta}(t)\notin\mathcal{N}(\boldsymbol{\theta}^0)\},\tag{54}$$

where

$$\mathcal{N}(\boldsymbol{\theta}^0):=\left\{\theta\mid\|\boldsymbol{G}(\boldsymbol{\theta})-\boldsymbol{G}(\boldsymbol{\theta}^0)\|_{\mathrm{F}}\leq\frac{1}{4}\kappa^2\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right)\right\}.\tag{55}$$

Then we have the following lemma.

Proposition 18 (local in time exponential decay of R_S, $\boldsymbol{\theta}$-lazy training)**.** *Given $\delta\in(0,1)$ and the sample set $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n\subset\Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. If $m\geq\frac{16n^2d^2C_{\psi,d}^2}{\lambda^2C_0}\log\frac{4n^2}{\delta}$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have for any $t\in[0,t^*)$*

$$R_S(\boldsymbol{\theta}(t))\leq\exp\left(-\frac{m\kappa^2}{n}\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right)\right)R_S(\boldsymbol{\theta}^0).\tag{56}$$

Proof Prop. 17 implies that for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and for any $\boldsymbol{\theta}\in\mathcal{N}(\boldsymbol{\theta}^0)$, we have

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

Note that

$$G_{ij} = G_{ij}^{[a]} + G_{ij}^{[\boldsymbol{w}]} = \frac{\kappa^2}{\kappa' m} \sum_{k=1}^{m} \nabla_{a_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \cdot \nabla_{a_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) + \frac{\kappa^2 \kappa'}{m} \sum_{k=1}^{m} \nabla_{\boldsymbol{w}_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \cdot \nabla_{\boldsymbol{w}_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j),$$

and

$$\nabla_{a_k} R_S(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} e_i \kappa \nabla_{a_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i),$$
$$\nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} e_i \kappa \nabla_{\boldsymbol{w}_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i).$$

Thus

$$\frac{m}{n^2} \boldsymbol{e}^\intercal \boldsymbol{G} \boldsymbol{e} = \frac{1}{\kappa'} \sum_{k=1}^{m} \nabla_{a_k} R_S(\boldsymbol{\theta}) \cdot \nabla_{a_k} R_S(\boldsymbol{\theta}) + \kappa' \sum_{k=1}^{m} \nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}).$$

Then finally we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} R_S(\boldsymbol{\theta}(t)) &= -\left(\frac{1}{\kappa'} \sum_{k=1}^{m} \nabla_{a_k} R_S(\boldsymbol{\theta}) \cdot \nabla_{a_k} R_S(\boldsymbol{\theta}) + \kappa' \sum_{k=1}^{m} \nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}) \right), \\
&= -\frac{m}{n^2} \boldsymbol{e}^\intercal \boldsymbol{G} \boldsymbol{e}, \\
&\leq -\frac{2m}{n} \lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{\theta}(t))\right) R_S(\boldsymbol{\theta}(t)) \\
&\leq -\frac{m\kappa^2}{n} \left(\frac{1}{\kappa'} \lambda_a + \kappa' \lambda_{\boldsymbol{w}} \right) R_S(\boldsymbol{\theta}(t)),
\end{aligned}$$

and an integration yields the result. ∎

Proposition 19 (bounds on the change of parameters, $\boldsymbol{\theta}$-lazy training)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. If $m \geq \max\left\{ \frac{16 n^2 d^2 C_{\psi,d}^2}{\lambda^2 C_0} \log \frac{8n^2}{\delta}, \frac{4\sqrt{2d}n\sqrt{R_S(\boldsymbol{\theta}^0)}}{\kappa(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}})} \right\}$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, for any $t \in [0, t^*)$ and any $k \in [m]$,*

$$\max_{k\in[m]} |a_k(t) - a_k(0)| \leq 2 \max\left\{ \frac{1}{\kappa'}, 1 \right\} \sqrt{2 \log \frac{4m(d+1)}{\delta}} p, \tag{57}$$

$$\max_{k\in[m]} \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2 \max\{\kappa', 1\} \sqrt{2 \log \frac{4m(d+1)}{\delta}} p, \tag{58}$$

and

$$\max_{k\in[m]} \{ |a_k(0)|,\ \|\boldsymbol{w}_k(0)\|_\infty \} \leq \sqrt{2 \log \frac{4m(d+1)}{\delta}}, \tag{59}$$

where $p := \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}})}$.

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

$$|\nabla_{a_k} R_S|^2 = \left|\frac{1}{n}\sum_{i=1}^n e_i \kappa \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right|^2 \le 2\|\boldsymbol{w}_k\|_1^2 \kappa^2 R_S(\boldsymbol{\theta}) \le 2d^2(\omega(t))^2\kappa^2 R_S(\boldsymbol{\theta}),$$
$$\|\nabla_{\boldsymbol{w}_k} R_S\|^2 = \left\|\frac{1}{n}\sum_{i=1}^n e_i \kappa a_k \sigma'(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i\right\|_\infty^2 \le 2|a_k|^2\kappa^2 R_S(\boldsymbol{\theta}) \le 2(\alpha(t))^2\kappa^2 R_S(\boldsymbol{\theta}).$$

By Prop. 18, we have if $m \ge \frac{16n^2d^2C_{\psi,d}^2}{\lambda^2 C_0}\log\frac{8n^2}{\delta}$, then with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

$$\begin{aligned}
|a_k(t) - a_k(0)| &\le \frac{1}{\kappa'}\int_0^t |\nabla_{a_k} R_S(\boldsymbol{\theta}(s))|\,\mathrm{d}s \\
&\le \frac{\sqrt{2}d\kappa}{\kappa'}\int_0^t \omega(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\le \frac{\sqrt{2}d\kappa}{\kappa'}\omega(t)\int_0^t \sqrt{R_S(\boldsymbol{\theta}^0)}\exp\left(-\frac{m\kappa^2}{2n}\left(\frac{1}{\kappa'}\lambda_a + \kappa'\lambda_{\boldsymbol{w}}\right)s\right)\mathrm{d}s \\
&\le \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\kappa'\left(\lambda_S^{[a]}/\kappa' + \kappa'\lambda_{\boldsymbol{w}}\right)}\omega(t) \\
&= \frac{p}{\kappa'}\omega(t).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty &\le \kappa'\int_0^t \|\nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}(s))\|_\infty\,\mathrm{d}s \\
&\le \sqrt{2}\kappa\kappa'\int_0^t \alpha(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\le \sqrt{2}\kappa\kappa'\alpha(t)\int_0^t \sqrt{R_S(\boldsymbol{\theta}^0)}\exp\left(-\frac{m\kappa^2}{2n}\left(\frac{1}{\kappa'}\lambda_a + \kappa'\lambda_{\boldsymbol{w}}\right)s\right)\mathrm{d}s \\
&\le \frac{2\sqrt{2}n\sqrt{R_S(\boldsymbol{\theta}^0)}\kappa'}{m\kappa\left(\lambda_S^{[a]}/\kappa' + \kappa'\lambda_{\boldsymbol{w}}\right)}\alpha(t) \\
&\le p\kappa'\alpha(t).
\end{aligned}$$

Thus

$$\alpha(t) \le \alpha(0) + p\omega(t)\frac{1}{\kappa'},$$
$$\omega(t) \le \omega(0) + p\alpha(t)\kappa'.$$

By Lemma 9, we have with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

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

If

$$m \ge \frac{4\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)},$$

then we have

$$p = \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)} \le \frac{1}{2}.$$

Thus

$$\alpha(t) \le \alpha(0) + \frac{p}{\kappa'}\omega(0) + p^2\alpha(t),$$
$$\alpha(t) \le \frac{4}{3}\alpha(0) + \frac{2}{3}\frac{1}{\kappa'}\omega(0).$$

Therefore

$$\alpha(t) \le 2\max\left\{1, \frac{1}{\kappa'}\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}.$$

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

$$\omega(t) \le 2\max\{1,\kappa'\}\sqrt{2\log\frac{4m(d+1)}{\delta}}.$$

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

$$\max_{k\in[m]}|a_k(t)-a_k(0)| \le 2\max\left\{\frac{1}{\kappa'},1\right\}\sqrt{2\log\frac{4m(d+1)}{\delta}}p,$$
$$\max_{k\in[m]}\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty \le 2\max\{\kappa',1\}\sqrt{2\log\frac{4m(d+1)}{\delta}}p,$$

which completes the proof. ■

To show our main results with $\gamma' > \gamma - 1$, we further define

$$t_a^* = \inf\{t \mid \boldsymbol{\theta}(t)\in\mathcal{N}_a(\boldsymbol{\theta}^0)\},\quad t_{\boldsymbol{w}}^* = \inf\{t\mid\boldsymbol{\theta}(0)\in\mathcal{N}_{\boldsymbol{w}}(\boldsymbol{\theta}^0)\}, \tag{60}$$

where

$$\mathcal{N}_a(\boldsymbol{\theta}^0) := \left\{\boldsymbol{\theta}\mid\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta})-\boldsymbol{G}^{[a]}(\boldsymbol{\theta}^0)\|_{\mathrm{F}} \le \frac{1}{4}\frac{\kappa^2}{\kappa'}\lambda_a\right\}, \tag{61}$$

$$\mathcal{N}_{\boldsymbol{w}}(\boldsymbol{\theta}^0) := \left\{\boldsymbol{\theta}\mid\|\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta})-\boldsymbol{G}^{[\boldsymbol{w}]}(\boldsymbol{\theta}^0)\|_{\mathrm{F}} \le \frac{1}{4}\kappa^2\kappa'\lambda_{\boldsymbol{w}}\right\}. \tag{62}$$

Proposition 20 (minimal eigenvalue of Gram matrix $\boldsymbol{G}^{[a]}$ at initial)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. If $m \geq \frac{16n^2d^2C_{\psi,d}^2}{C_0\lambda_a^2}\log\frac{2n^2}{\delta}$, then we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,*

$$\lambda_{\min}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}^0)\right) \geq \frac{3}{4}\frac{\kappa^2}{\kappa'}\lambda_a. \tag{63}$$

Proof For any $\varepsilon > 0$ define

$$\Omega_{ij}^{[a]} := \left\{\boldsymbol{\theta}^0 \mid \left|\frac{\kappa'}{\kappa^2}G_{ij}^{[a]}(\boldsymbol{\theta}^0) - K_{ij}^{[a]}\right| \leq \frac{\varepsilon}{n}\right\}. \tag{64}$$

By Theorem 15 and Lemma 14, if $\frac{\varepsilon}{ndC_{\psi,d}} \leq 1$ then

$$\mathbb{P}(\Omega_{ij}^{[a]}) \geq 1 - 2\exp\left(-\frac{mC_0\varepsilon^2}{n^2d^2C_{\psi,d}^2}\right),$$

with probability at least

$$\left[1 - 2\exp\left(-\frac{mC_0\varepsilon^2}{n^2d^2C_{\psi,d}^2}\right)\right]^{n^2} \geq 1 - 2n^2\exp\left(-\frac{mC_0\varepsilon^2}{n^2d^2C_{\psi,d}^2}\right)$$

over the choice of $\boldsymbol{\theta}^0$, we have

$$\left\|\frac{\kappa'}{\kappa^2}\boldsymbol{G}^{[a]}(\boldsymbol{\theta}^0) - \boldsymbol{K}^{[a]}\right\|_{\mathrm{F}} \leq \varepsilon.$$

Taking $\varepsilon = \lambda_a/4$, i.e., $\delta = 2n^2\exp\left(-\frac{mC_0\lambda_a^2}{16n^2d^2C_{\psi,d}^2}\right)$, we obtain the estimate

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

■

Proposition 21 (local in time exponential decay of R_S, $\boldsymbol{w}$-lazy training)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. If $m \geq \frac{16n^2d^2C_{\psi,d}^2}{C_0\lambda_a^2}\log\frac{2n^2}{\delta}$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, for $t \in [0, t_a^*)$,*

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m\kappa^2}{\kappa' n}\lambda_a t\right)R_S(\boldsymbol{\theta}^0). \tag{65}$$

Proof By Prop. 20, for any $\delta \in (0,1)$ with probability $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and for any $\boldsymbol{\theta} \in \mathcal{N}_a(\boldsymbol{\theta}^0)$,

$$\begin{aligned}\lambda_{\min}(\boldsymbol{G}^{[a]}(\boldsymbol{\theta})) &\geq \lambda_{\min}(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}^0)) - \|\boldsymbol{G}(\boldsymbol{\theta}) - \boldsymbol{G}(\boldsymbol{\theta}^0)\|_{\mathrm{F}} \\ &\geq \frac{3}{4}\frac{\kappa^2}{\kappa'}\lambda_a - \frac{1}{4}\frac{\kappa^2}{\kappa'}\lambda_a \\ &= \frac{1}{2}\frac{\kappa^2}{\kappa'}\lambda_a.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R_S(\boldsymbol{\theta}(t)) &= -\frac{m}{n^2}\boldsymbol{e}^\intercal \boldsymbol{G}\boldsymbol{e} \\ &\leq -\frac{m}{n^2}\boldsymbol{e}^\intercal \boldsymbol{G}^{[a]}\boldsymbol{e} \\ &\leq -\frac{2m}{n}\lambda_{\min}\left(\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))\right) R_S(\boldsymbol{\theta}(t)) \\ &\leq -\frac{m\kappa^2}{\kappa' n}\lambda_a R_S(\boldsymbol{\theta}(t)).\end{aligned}$$

This leads to the linear convergence rate. ■

Proposition 22 (bounds on the change of parameters, $\boldsymbol{w}$-lazy training)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 holds. If $m \geq \frac{16n^2d^2C_{\psi,d}^2}{C_0\lambda_a^2}\log\frac{4n^2}{\delta}$ and $\frac{m\kappa}{\kappa'} \geq \frac{4\sqrt{2d}n\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda_a}$ and $\kappa' \leq 1$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and for any $t \in [0, t_a^*)$, $k \in [m]$,*

$$\max_{k\in[m]}|a_k(t) - a_k(0)| \leq 2\frac{1}{\kappa'}\sqrt{2\log\frac{4m(d+1)}{\delta}}p_a, \tag{66}$$

$$\max_{k\in[m]}\|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2\sqrt{2\log\frac{4m(d+1)}{\delta}}p_a, \tag{67}$$

and

$$\max_{k\in[m]}\{|a_k(0)|,\ \|\boldsymbol{w}_k(0)\|_\infty\} \leq \sqrt{2\log\frac{4m(d+1)}{\delta}}, \tag{68}$$

where $p_a = \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\lambda_a/\kappa'}$.

Proof Since

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

then

$$|\nabla_{a_k}R_S|^2 = \left|\frac{1}{n}\sum_{i=1}^n e_i\kappa\sigma(\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\right|^2 \leq 2\|\boldsymbol{w}_k\|_1^2\kappa^2 R_S(\boldsymbol{\theta}) \leq 2d^2(\omega(t))^2\kappa^2 R_S(\boldsymbol{\theta}),$$

$$\|\nabla_{\boldsymbol{w}_k}R_S\|^2 = \left\|\frac{1}{n}\sum_{i=1}^n e_i\kappa a_k\sigma'(\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i\right\|_\infty^2 \leq 2|a_k|^2\kappa^2 R_S(\boldsymbol{\theta}) \leq 2(\alpha(t))^2\kappa^2 R_S(\boldsymbol{\theta}).$$

By Prop. 18, 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 $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

$$\begin{aligned}
|a_k(t)-a_k(0)| &\leq \frac{1}{\kappa'}\int_0^t |\nabla_{a_k} R_S(\boldsymbol{\theta}(s))|\,\mathrm{d}s \\
&\leq \frac{\sqrt{2}d\kappa}{\kappa'}\int_0^t \omega(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\leq \frac{\sqrt{2}d\kappa}{\kappa'}\omega(t)\int_0^t \sqrt{R_S(\boldsymbol{\theta}^0)}\exp\left(-\frac{m\kappa^2}{2n}\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right)s\right)\mathrm{d}s \\
&\leq \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\kappa'\left(\lambda_S^{[a]}/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}\omega(t) \\
&= \frac{p_a}{\kappa'}\omega(t).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty &\leq \kappa'\int_0^t \|\nabla_{\boldsymbol{w}_k} R_S(\boldsymbol{\theta}(s))\|_\infty\,\mathrm{d}s \\
&\leq \sqrt{2}\kappa\kappa'\int_0^t \alpha(s)\sqrt{R_S(\boldsymbol{\theta}(s))}\,\mathrm{d}s \\
&\leq \sqrt{2}\kappa\kappa'\alpha(t)\int_0^t \sqrt{R_S(\boldsymbol{\theta}^0)}\exp\left(-\frac{m\kappa^2}{2n}\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right)s\right)\mathrm{d}s \\
&\leq \frac{2\sqrt{2}n\sqrt{R_S(\boldsymbol{\theta}^0)}\kappa'}{m\kappa\left(\lambda_S^{[a]}/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}\alpha(t) \\
&\leq p_a\kappa'\alpha(t).
\end{aligned}$$

Thus

$$\begin{aligned}
\alpha(t) &\leq \alpha(0)+p_a\omega(t)\frac{1}{\kappa'}, \\
\omega(t) &\leq \omega(0)+p_a\alpha(t)\kappa'.
\end{aligned}$$

By Lemma 9, we have with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

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

If

$$m \geq \frac{4\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)},$$

then

$$p_a = \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\lambda_a/\kappa'} \leq \frac{1}{2}.$$

Thus

$$\alpha(t) \leq \alpha(0) + \frac{p_a}{\kappa'}\omega(0) + p_a^2\alpha(t),$$
$$\alpha(t) \leq \frac{4}{3}\alpha(0) + \frac{2}{3}\frac{1}{\kappa'}\omega(0),$$

Therefore

$$\alpha(t) \leq 2\frac{1}{\kappa'}\sqrt{2\log\frac{4m(d+1)}{\delta}}.$$

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

$$\omega(t) \leq 2\sqrt{2\log\frac{4m(d+1)}{\delta}}.$$

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

$$\max_{k\in[m]} |a_k(t) - a_k(0)| \leq 2\frac{1}{\kappa'}\sqrt{2\log\frac{4m(d+1)}{\delta}}p_a,$$
$$\max_{k\in[m]} \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \leq 2\sqrt{2\log\frac{4m(d+1)}{\delta}}p_a,$$

which completes the proof. ■

Appendix B. Proof of Theorem 6

We further divide the linear regime into two part: $\gamma < 1$ where the training dynamics is $\boldsymbol{\theta}$-lazy and $\gamma' > \gamma - 1$ where the training dynamics is $\boldsymbol{w}$-lazy. Theorem 6 is hence covered by Proposition 23 and Proposition 25 whose proofs are given in this section.

Proposition 23 ($\boldsymbol{\theta}$-lazy training)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 and Assumption 2 hold. ASI is used if $\gamma \leq \frac{1}{2}$. Suppose that $\gamma < 1$ and the dynamics (26)–(29) is considered. Then for sufficiently large m, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

(a) $\sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2 \lesssim \frac{1}{\sqrt{m}\kappa}\log m.$

(b) $R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{2m\kappa^2\lambda t}{n}\right) R_S(\boldsymbol{\theta}^0).$

Moreover, we have with probability at least $1-\delta-2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$.

(c) $\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} \lesssim \frac{1}{m\kappa}\log m.$

Proof Let $t \in [0, t^*)$, $p = \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}})}$ and $\xi = \sqrt{2\log\frac{8m(d+1)}{\delta}}$.

(a) From Proposition 19 we have with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

$$\begin{aligned}
\sup_{t\in[0,t^*]}\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2 &\le \left[m(d+1)\left(2\sqrt{2\log\frac{8m(d+1)}{\delta}}\frac{\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}\right)^2\right]^{\frac{1}{2}}\\
&= \max\left\{\kappa',\frac{1}{\kappa'}\right\}\sqrt{m(d+1)}2\sqrt{2\log\frac{8m(d+1)}{\delta}}\frac{\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}\\
&\le \max\left\{\kappa',\frac{1}{\kappa'}\right\}\frac{4\sqrt{d+1}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\sqrt{m}\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}\\
&\le \frac{4\sqrt{d+1}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\sqrt{m}\kappa\lambda}\\
&\lesssim \frac{1}{\sqrt{m}\kappa}\log m,
\end{aligned}$$

where we use the fact

$$\frac{\max\left\{\kappa',\frac{1}{\kappa'}\right\}}{\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}}\le\max\left\{\frac{\kappa'}{\kappa'\lambda_{\boldsymbol{w}}},\frac{1/\kappa'}{\lambda_a/\kappa'}\right\}\le\frac{1}{\lambda}.$$

(b) The linear convergence rate is essentially proved by Prop. 18 with $t^*=+\infty$. We divide the proof into the following three steps. In particular, $t^*=+\infty$ is proved in the step (iii).

(i) Let

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

then

$$\left|G_{ij}^{[a]}(\boldsymbol{\theta}(t))-G_{ij}^{[a]}(\boldsymbol{\theta}(0))\right|\le\frac{\kappa^2}{m\kappa'}\sum_{k=1}^m\left|g_{ij}^{[a]}(\boldsymbol{w}_k(t))-g_{ij}^{[a]}(\boldsymbol{w}_k(0))\right|.$$

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

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

where

$$\nabla g_{ij}^{[a]}(\boldsymbol{w})=\sigma'(\boldsymbol{w}\cdot\xi_i)\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}_j)\boldsymbol{x}_i+\sigma(\boldsymbol{w}^\intercal\boldsymbol{x}_i)\sigma'(\boldsymbol{w}^\intercal\boldsymbol{x}_j)\boldsymbol{x}_j,$$

and

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

From Proposition 19 we have with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

$$\begin{aligned}
\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty&\le p\alpha(t)\kappa'\le 2\max\{\kappa',1\}\xi p,\\
\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_1&\le 2d\max\{\kappa',1\}\xi p.
\end{aligned}$$

Thus

$$\begin{aligned}\|c\boldsymbol{w}_k(t)+(1-c)\boldsymbol{w}_k(0)\|_1 &\le d\left(\|\boldsymbol{w}_k(0)\|_\infty+\|\boldsymbol{w}_k(t)-\boldsymbol{w}_k(0)\|_\infty\right)\\ &\le d\left(\xi+2\max\{\kappa',1\}\xi p\right)\\ &\le 2d\xi\max\{\kappa',1\}.\end{aligned}$$

Then

$$|G_{ij}^{[a]}(\boldsymbol{\theta}(t))-G_{ij}^{[a]}(\boldsymbol{\theta}(0))|\le 8d^2\kappa^2\xi^2\max\left\{\kappa',\frac{1}{\kappa'}\right\}p,$$

and

$$\begin{aligned}\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))-\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}} &\le 8d^2n\kappa^2\max\left\{\kappa',\frac{1}{\kappa'}\right\}\left(2\log\frac{8m(d+1)}{\delta}\right)\frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}\\ &\le \kappa\max\left\{\kappa',\frac{1}{\kappa'}\right\}\frac{32\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)}.\end{aligned}$$

If we choose

$$m\kappa\ge\frac{128\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda^2},\tag{69}$$

then noticing that

$$\begin{aligned}\frac{1}{\lambda^2} &\ge \frac{1}{\left(4\left(\frac{1}{27}\lambda_a(\lambda_{\boldsymbol{w}})^3\right)^{1/4}\right)^2}\\ &\ge \frac{1}{\left(\lambda_a/(\kappa')^{3/2}+\sqrt{\kappa'}\lambda_{\boldsymbol{w}}\right)^2}\\ &= \frac{\kappa'}{\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)^2}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\lambda^2} &\ge \frac{1}{\left(4\left(\frac{1}{27}(\lambda_a)^3\lambda_{\boldsymbol{w}}\right)^{1/4}\right)^2}\\ &\ge \frac{1}{\left(\lambda_a/\sqrt{\kappa'}+(\kappa')^{3/2}\lambda_{\boldsymbol{w}}\right)^2}\\ &= \frac{1}{\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)^2\kappa'},\end{aligned}$$

we have

$$m\kappa\ge\max\left\{\kappa',\frac{1}{\kappa'}\right\}\frac{256\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\right)\sqrt{R_S(\boldsymbol{\theta}^0)}}{\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)^2}.$$

Therefore

$$\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t))-\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}}\le\frac{1}{8}\kappa^2\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right).\tag{70}$$

(ii) Define

$$D_{i,k} = \{\omega_k(0) \mid \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \le 2\xi \max\{\kappa', 1\}p, \\ \sigma'(\boldsymbol{w}_k(t^*) \cdot \boldsymbol{x}_i) \ne \sigma'(\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i)\}.$$

If $|\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i| > 4d \max\{\kappa', 1\}\xi p$, then

$$|\boldsymbol{w}_k(t) \cdot \boldsymbol{x}_i - \boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i| \le \|\boldsymbol{x}_i\|_1 \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty \le 2d\sqrt{2\log\frac{8m(d+1)}{\delta}}p,$$

thus $\boldsymbol{w}_k(t) \cdot \boldsymbol{x}_i$ and $\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i$ have the same sign which means $D_{i,k}$ is empty. Recall that $\boldsymbol{x}_i \in [0,1]^d$ with $(\boldsymbol{x}_i)_d = 1$, then $\|\boldsymbol{x}_i\|_2 \ge 1$. Let $\hat{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2}$ then $|\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i| \ge |\boldsymbol{x}_k(0) \cdot \hat{\boldsymbol{x}}_i|$ and

$$\begin{aligned}
\mathbb{P}(D_{i,k}) &\le \mathbb{P}(|\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i| \le 4d \max\{\kappa', 1\}\xi p) \\
&\le \mathbb{P}(|\boldsymbol{w}_k(0) \cdot \hat{\boldsymbol{x}}_i| \le 4d \max\{\kappa', 1\}\xi p) \\
&= \mathbb{P}(|\boldsymbol{w}_k(0) \cdot (1,0,0,\ldots,0)^\intercal| \le 4d \max\{\kappa', 1\}\xi p) \\
&= \mathbb{P}(|(\boldsymbol{w}_k(0))_1| \le 4d \max\{\kappa', 1\}\xi p) \\
&= 2\int_0^{4d\max\{\kappa',1\}\xi p} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \,\mathrm{d}x \\
&\le \frac{8}{\sqrt{2\pi}} d \max\{\kappa', 1\}\xi p \\
&\le 4d \max\{\kappa', 1\}\xi p.
\end{aligned}$$

Then

$$\begin{aligned}
|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))| &\le \frac{\kappa^2\kappa'|\boldsymbol{x}_i \cdot \boldsymbol{x}_j|}{m} \sum_{k=1}^m \Big| a_k^2(t^*)\sigma'(\boldsymbol{w}_k(t) \cdot \boldsymbol{x}_i)\sigma'(\boldsymbol{w}_k(t) \cdot \boldsymbol{x}_j) \\
&\qquad - a_k^2(0)\sigma'(\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i)\sigma'(\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_j)\Big| \\
&\le \frac{\kappa^2\kappa' d}{m} \sum_{k=1}^m \left[a_k^2(t)|D_{k,i,j}| + |a_k^2(t) - a_k^2(0|\right],
\end{aligned}$$

where

$$D_{k,i,j} := \sigma'(\boldsymbol{w}_k(t) \cdot \boldsymbol{x}_i)\sigma'(\boldsymbol{w}_k(t) \cdot \boldsymbol{x}_j) - \sigma'(\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_i)\sigma'(\boldsymbol{w}_k(0) \cdot \boldsymbol{x}_j).$$

Thus

$$\mathbb{E}|D_{k,i,j}| \le \mathbb{P}(D_{k,i} \cup D_{k,j}) \le 8d \max\{\kappa', 1\}\xi p.$$

At the same time

$$\begin{aligned}
|a_k^2(t) - a_k^2(0)| &\le |a_k(t) - a_k(0)|^2 + 2|a_k(0)||a_k(t) - a_k(0)| \\
&\le \left(2\max\left\{\frac{1}{\kappa'}, 1\right\}\xi p\right)^2 + 2\xi\left(2\max\left\{\frac{1}{\kappa'}, 1\right\}\xi p\right) \\
&\le 6\xi^2 \max\left\{\frac{1}{\kappa'^2}, 1\right\} p,
\end{aligned}$$

so

$$a_k^2(t) \le |a_k^2(t) - a_k^2(0)| + a_k^2(0) \le \left(2\max\left\{\frac{1}{\kappa'}, 1\right\}\xi p\right)^2 + 2\xi\left(2\max\left\{\frac{1}{\kappa'}, 1\right\}\xi p\right) + \xi^2$$
$$\le 4\max\left\{\frac{1}{\kappa'^2}, 1\right\}\xi^2.$$

Then

$$\begin{aligned}
&\mathbb{E}\sum_{i,j=1}^{n}\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right| \\
&\quad\le \sum_{i,j=1}^{n}\frac{\kappa^2\kappa' d}{m}\sum_{k=1}^{m}\left(4\max\left\{\frac{1}{\kappa'^2}, 1\right\}\xi^2\mathbb{E}|D_{k,i,j}| + 6\max\left\{\frac{1}{\kappa'^2}, 1\right\}\xi^2 p\right) \\
&\quad\le \sum_{i,j=1}^{n}\frac{\kappa^2\kappa' d}{m}\sum_{k=1}^{m}\left(4\max\left\{\frac{1}{\kappa'^2}, 1\right\}\xi^2 8d\max\{\kappa', 1\}\xi p + 6\max\left\{\frac{1}{\kappa'^2}, 1\right\}\xi^2 p\right) \\
&\quad\le \kappa^2\kappa' d n^2\left(32d\xi\max\{\kappa', \frac{1}{\kappa'^2}\} + 6\max\left\{\frac{1}{\kappa'^2}, 1\right\}\right)\xi^2 p \\
&\quad\le 40\kappa^2 d^2 n^2\left(2\log\frac{8m(d+1)}{\delta}\right)^{3/2}\max\{\kappa'^2, \frac{1}{\kappa'}\}p.
\end{aligned}$$

By Markov's inequality, with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$, we have

$$\begin{aligned}
&\|G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}} \\
&\le \sum_{i,j=1}^{n}\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right| \\
&\le \max\left\{\kappa'^2, \frac{1}{\kappa'}\right\}\frac{40\kappa^2 d^2 n^2\left(2\log\frac{8m(d+1)}{\delta}\right)^{3/2} p}{\delta/2} \\
&\le \max\left\{\kappa'^2, \frac{1}{\kappa'}\right\}\frac{80\kappa^2 d^2 n^2 2\sqrt{2}}{\delta}\left(\log\frac{8m(d+1)}{\delta}\right)^{3/2}\frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\left(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}}\right)} \\
&\le \kappa\max\left\{\kappa'^2, \frac{1}{\kappa'}\right\}\frac{640 d^3 n^3\left(\log\frac{8m(d+1)}{\delta}\right)^{3/2}\sqrt{R_S(\boldsymbol{\theta}^0)}\delta^{-1}}{m\left(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}}\right)}.
\end{aligned}$$

If

$$m \ge \frac{5120\delta^{-1}d^3n^3\left(\log\frac{8m(d+1)}{\delta}\right)^{3/2}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda^2},$$

then noticing that

$$\frac{1}{\lambda^2} \ge \frac{\kappa'^2}{\left(\kappa'\lambda_{\boldsymbol{w}}\right)^2} \ge \frac{\kappa'^2}{\left(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}}\right)^2}$$

and

$$\begin{aligned}\frac{1}{\lambda^2} &\geq \frac{1}{\left(4\left(\frac{1}{27}(\lambda_a)^3\lambda_{\boldsymbol{w}}\right)^{1/4}\right)^2} \\ &\geq \frac{1}{\left(\lambda_a/\sqrt{\kappa'}+(\kappa')^{3/2}\lambda_{\boldsymbol{w}}\right)^2} \\ &= \frac{1}{\left(\lambda_a/\kappa'+\kappa'\lambda_{\boldsymbol{w}}\right)^2\kappa'},\end{aligned}$$

we have

$$\|G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t))-G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}} \leq \frac{1}{8}\kappa^2\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right). \tag{71}$$

(iii) For $t\in[0,t^*)$,

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m\kappa^2}{n}\left(\frac{1}{\kappa'}\lambda_a+\kappa'\lambda_{\boldsymbol{w}}\right)t\right)R_S(\boldsymbol{\theta}^0) \leq \exp\left(-\frac{2m\kappa^2\lambda}{n}\right)R_S(\boldsymbol{\theta}^0).$$

Suppose that $t^*<+\infty$ then one can take the limit $t\to t^*$ in (70) and (71). This will lead to a contradiction with the definition of t^*. Therefore $t^*=+\infty$.

(c) By Proposition 16, we have with probability at least $1-2\exp\left(-\frac{C_0m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$,

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

Therefore, with probability at least $1-\delta-2\exp\left(-\frac{C_0m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have

$$\begin{aligned}\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} &\leq \sqrt{\frac{2}{m(d+1)}}\sup_{t\in[0,+\infty)}\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}^0\|_2 \\ &\leq \sqrt{\frac{2}{m(d+1)}}\frac{4\sqrt{d+1}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\sqrt{m}\kappa\lambda} \\ &\leq \frac{1}{m\kappa}\frac{4\sqrt{2}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda} \\ &\lesssim \frac{1}{m\kappa}\log m.\end{aligned}$$

■

Remark 24. *The proof indicates more quantitative conditions on m and κ for Proposition 23 to hold:*

$$m \geq \frac{16n^2d^2C_{\psi,d}^2}{\lambda^2C_0}\log\frac{16n^2}{\delta}, \tag{72}$$

and

$$m\kappa \geq \max\left\{\frac{2\sqrt{2d}n\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda}, \frac{128\sqrt{2}d^3n^2\left(\log\frac{8m(d+1)}{\delta}\sqrt{R_S(\boldsymbol{\theta}^0)}\right)}{\lambda^2}, \frac{5120\delta^{-1}d^3n^3\left(\log\frac{8m(d+1)}{\delta}\right)^{3/2}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda^2}\right\}. \tag{73}$$

Proposition 25 ($\boldsymbol{w}$-lazy training)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 and Assumption 2 hold. Suppose that $\gamma' > \gamma - 1$, $\gamma' > 0$, and the dynamics (26)–(29) is considered. Then for sufficiently large m, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

(a)

$$\sup_{t\in[0,+\infty)}\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}^0_{\boldsymbol{w}}\|_2 \leq \sup_{t\in[0,+\infty)}\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2 \lesssim \frac{1}{\sqrt{m}\kappa}\log m.$$

(b) $R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m\kappa^2\lambda_a t}{\kappa' n}\right) R_S(\boldsymbol{\theta}^0)$.

Moreover we have with probability at least $1-\delta-2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$.

(c)

$$\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} \lesssim \frac{1}{m\kappa}\log m, \quad (not\ \ll 1),$$
$$\sup_{t\in[0,+\infty)}\frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}^0_{\boldsymbol{w}}\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}\|_2} \lesssim \frac{\kappa'}{m\kappa}\log m, \quad (\ll 1).$$

Proof Let $t \in [0, t_a^*)$, $p_a = \frac{2\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\lambda_a/\kappa'}$, and $\xi = \sqrt{2\log\frac{8m(d+1)}{\delta}}$.

(a) From Proposition 19 we have with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$

$$\begin{aligned}
\sup_{t\in[0,t_a^*)}\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}^0_{\boldsymbol{w}}\|_2 &\leq \sup_{t\in[0,t_a^*)}\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2 \\
&\leq \left[\left(\frac{m}{\kappa'^2} + md\right)\left(2\sqrt{2\log\frac{8m(d+1)}{\delta}}p_a\right)^2\right]^{\frac{1}{2}} \\
&= \sqrt{m(d+1)}2\frac{1}{\kappa'}\sqrt{2\log\frac{8m(d+1)}{\delta}}\frac{\sqrt{2}dn\sqrt{R_S(\boldsymbol{\theta}^0)}}{m\kappa\lambda_a/\kappa'} \\
&\leq \max\left\{\kappa', \frac{1}{\kappa'}\right\}\frac{4\sqrt{d+1}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\sqrt{m}\kappa\left(\lambda_a/\kappa' + \kappa'\lambda_{\boldsymbol{w}}\right)} \\
&\leq \frac{8\sqrt{d+1}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\sqrt{m}\kappa\lambda_a}.
\end{aligned}$$

(b) We divide this proof into the following two steps.

(i) Let

$$G_{ij}^{[a]}(\boldsymbol{w}) := \sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_j),$$

then

$$|G_{ij}^{[a]}(\boldsymbol{\theta}(t)) - G_{ij}^{[a]}(\boldsymbol{\theta}(0))| \le \frac{\kappa^2}{m\kappa'} \sum_{k=1}^{m} |g_{ij}^{[a]}(\boldsymbol{w}_k(t)) - g_{ij}^{[a]}(\boldsymbol{w}_k(0))|.$$

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

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

where

$$\nabla g_{ij}^{[a]}(\boldsymbol{w}) = \sigma'(\boldsymbol{w}\cdot\xi_i)\sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_j)\boldsymbol{x}_i + \sigma(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\sigma'(\boldsymbol{w}^\intercal \boldsymbol{x}_j)\boldsymbol{x}_j,$$

and

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

From Proposition 19 we have with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$,

$$\begin{aligned} \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty &\le p_a \alpha(t)\kappa' \le 2\xi p_a, \\ \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_1 &\le 2d\xi p_a. \end{aligned}$$

Thus

$$\begin{aligned} \|c\boldsymbol{w}_k(t) + (1-c)\boldsymbol{w}_k(0)\|_1 &\le d\left(\|\boldsymbol{w}_k(0)\|_\infty + \|\boldsymbol{w}_k(t) - \boldsymbol{w}_k(0)\|_\infty\right) \\ &\le d\left(\xi + 2\xi p_a\right) \\ &\le 2d\xi. \end{aligned}$$

Then

$$|G_{ij}^{[a]}(\boldsymbol{\theta}(t)) - G_{ij}^{[a]}(\boldsymbol{\theta}(0))| \le 8d^2 \frac{\kappa^2}{\kappa'}\xi^2 p_a,$$

and

$$\begin{aligned} \|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t)) - \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}} &\le 16d^2 n \left(\log \frac{8m(d+1)}{\delta}\right) \frac{\kappa^2}{\kappa'} p_a \\ &\le \frac{32\sqrt{2}d^3 n^2 \left(\log \frac{8m(d+1)}{\delta}\right) \sqrt{R_S(\boldsymbol{\theta}^0)}\kappa}{m\lambda_a}. \end{aligned}$$

If

$$\frac{m\kappa}{\kappa'} \ge \frac{256\sqrt{2}d^3 n^2 \left(\log \frac{8m(d+1)}{\delta})\right) \sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda_a^2},$$

then we have

$$\|\boldsymbol{G}^{[a]}(\boldsymbol{\theta}(t)) - \boldsymbol{G}^{[a]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}} \le \frac{1}{8}\frac{\kappa^2}{\kappa'}\lambda_a. \tag{74}$$

(ii) For $t \in [0, t_a^*)$ by Prop. 18,

$$R_S(\boldsymbol{\theta}(t)) \leq \exp\left(-\frac{m\kappa^2\lambda_a t}{n\kappa'}\right) R_S(\boldsymbol{\theta}^0).$$

Suppose that $t_a^* < +\infty$ then one can take the limit $t \to t_a^*$ in (74). This will lead to a contradiction with the definition of t_a^*. Therefore $t_a^* = +\infty$.

(c) By Proposition 16, we have with probability at least $1 - 2\exp(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2})$ over the choice of $\boldsymbol{\theta}^0$,

$$\|\boldsymbol{\theta}^0\|_2^2 \geq \frac{d+1}{2} m.$$

So with probability at least $1 - \delta - 2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have

$$\begin{aligned}
\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} &\leq \sqrt{\frac{2}{m(d+1)}} \sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2 \\
&\leq \sqrt{\frac{2}{m(d+1)}} \frac{8\sqrt{d+1}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\sqrt{m}\kappa\lambda} \\
&\leq \frac{1}{m\kappa} \frac{8\sqrt{2}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda_a} \\
&\lesssim \frac{1}{m\kappa}\log m.
\end{aligned}$$

Similarly, by Proposition 16, we have with probability at least $1 - 2\exp(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2})$ over the choice of $\boldsymbol{\theta}^0$,

$$\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2^2 \geq \frac{d}{2} m.$$

So with probability at least $1 - \delta - 2\exp\left(-\frac{C_0 md}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have

$$\begin{aligned}
\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2} &\leq \sqrt{\frac{2}{dm}} \sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}_{\boldsymbol{w}}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2 \\
&\leq \sqrt{\frac{2}{dm}} \frac{\kappa'}{\sqrt{m}\kappa} \frac{8\sqrt{d}dn\sqrt{\log\frac{8m(d+1)}{\delta}}\sqrt{R_S(\boldsymbol{\theta}^0)}}{\lambda} \\
&\lesssim \frac{\kappa'}{m\kappa}\log m.
\end{aligned}$$

■

We remark that in fact we can prove a similar result about the change of parameter a_k's. We state this result as follows without proof.

Proposition 26 (a-lazy training)**.** *Given $\delta \in (0,1)$ and the sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \Omega$ with $\boldsymbol{x}_i$'s drawn i.i.d. from some unknown distribution $\mathcal{D}$. Suppose that Assumption 1 and Assumption 2 hold. Suppose that $\gamma' < \gamma - 1$, $\gamma' < 0$, and the dynamics (26)–(29) is considered. Then for sufficiently large m, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

(a)

$$\sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}_a(t) - \boldsymbol{\theta}_a^0\|_2 \le \sup_{t\in[0,+\infty)} \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2 \lesssim \frac{1}{\sqrt{m}\kappa} \log m.$$

(b) $R_S(\boldsymbol{\theta}(t)) \le \exp\left(-\frac{m\kappa^2\kappa'\lambda_{\boldsymbol{w}} t}{n}\right) R_S(\boldsymbol{\theta}^0)$.

Moreover we have with probability at least $1-\delta-2\exp\left(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2}\right)$ over the choice of $\boldsymbol{\theta}^0$, we have

(c)

$$\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} \lesssim \frac{1}{m\kappa} \log m, \quad (\textit{not } \ll 1),$$
$$\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_a(t) - \boldsymbol{\theta}_a^0\|_2}{\|\boldsymbol{\theta}_a\|_2} \lesssim \frac{1}{m\kappa\kappa'} \log m, \quad (\ll 1).$$

Appendix C. Proof of Theorem 8

In order to characterize the condensed regime, we need a crucial proposition that ravels a natrual relation between $a_k(t)$ and $\boldsymbol{w}_k(t)$ during the GD training dynamics.

Proposition 27. *Consider the GD training dynamics (26)–(29), then we have*

$$|a_k(t)| \le \frac{1}{\kappa'}\|\boldsymbol{w}_k(t)\|_2 + |a_k^0|, \tag{75}$$

which holds for any $t \ge 0$ and $k \in [m]$.

Proof Multiplying equations in (26) by $\kappa' a_k$ and $\frac{\boldsymbol{w}_k}{\kappa'}$ respectively, we obtain

$$\begin{cases} \kappa' \dot{a}_k a_k = -\dfrac{\kappa}{n} \displaystyle\sum_{i=1}^n e_i a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i), \\ \dfrac{1}{\kappa'} \dot{\boldsymbol{w}}_k \boldsymbol{w}_k = -\dfrac{\kappa}{n} \displaystyle\sum_{i=1}^n e_i a_k \sigma'(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \boldsymbol{w}_k^\intercal \boldsymbol{x}_i. \end{cases} \tag{76}$$

Notice that for ReLU activation $\sigma(z) = z\sigma'(z)$, $z \in \mathbb{R}$. by comparing the right hand side of (76), one can obtain

$$\kappa'^2 \frac{\mathrm{d}}{\mathrm{d}t}|a_k|^2 = \frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{w}_k\|_2^2. \tag{77}$$

Integrating this from 0 to t leads to

$$\kappa'^2 \left(|a_k(t)|^2 - |a_k^0|^2\right) = \|\boldsymbol{w}_k(t)\|_2^2 - \|\boldsymbol{w}_k^0\|_2^2,$$

which then can be written as

$$|a_k(t)|^2 = \frac{1}{\kappa'^2}\left(\|\boldsymbol{w}_k(t)\|_2^2 - \|\boldsymbol{w}_k^0\|_2^2\right) + |a_k^0|^2. \tag{78}$$

Finally we have

$$|a_k(t)| \le \sqrt{\frac{1}{\kappa'^2}\|\boldsymbol{w}_k(t)\|_2^2 + |a_k^0|^2} \le \frac{1}{\kappa'}\|\boldsymbol{w}_k(t)\|_2 + |a_k^0|.$$

■

We remark that the key identity (77) in the proof of Proposition 27 has been previously proved in (Du et al., 2018, Theorem 2.1). A special case of it is (Williams et al., 2019, Lemma 3).

Proof [Proof of Theorem 8] By Assumption 3, there exits a $T^* > 0$ such that

$$R_S(\boldsymbol{\theta}(T^*)) \le \frac{1}{32n}.$$

Without loss of generality, we assume $f(\boldsymbol{x}_1) \ge \frac{1}{2}$. Therefore

$$\frac{1}{2n}e_1(T^*)^2 \le \frac{1}{2n}\boldsymbol{e}(T^*)^\intercal \boldsymbol{e}(T^*) = R_S(\boldsymbol{\theta}(T^*)) \le \frac{1}{32n},$$

which means

$$|e_1(T^*)| \le \frac{1}{4}.$$

Recalling the definition that $e_1 = \kappa f_{\boldsymbol{\theta}}(\boldsymbol{x}_1) - f(\boldsymbol{x}_1)$, we have

$$\kappa\sum_{k=1}^m a_k(T^*)\sigma(\boldsymbol{w}_k(T^*)^\intercal \boldsymbol{x}_1) \ge f(\boldsymbol{x}_1) - \frac{1}{4} \ge \frac{1}{4}.$$

So

$$\begin{aligned}
\frac{1}{4\kappa} &\le \sum_{k=1}^m a_k(T^*)\sigma(\boldsymbol{w}_k(T^*)^\intercal \boldsymbol{x}_1) \\
&\le \sqrt{d}\sum_{k=1}^m |a_k(T^*)|\|\boldsymbol{w}_k(T^*)\|_2 \\
&\le \sqrt{d}\sum_{k=1}^m \left(\frac{1}{\kappa'}\|\boldsymbol{w}_k(T^*)\|_2 + |a_k^0|\right)\|\boldsymbol{w}_k(T^*)\|_2 \\
&= \sqrt{d}\left(\frac{1}{\kappa'}\sum_{k=1}^m \|\boldsymbol{w}_k(T^*)\|_2^2 + \sum_{k=1}^m |a_k^0|\|\boldsymbol{w}_k(T^*)\|_2\right) \\
&\le \sqrt{d}\left(\frac{1}{\kappa'}\|\boldsymbol{\theta}_{\boldsymbol{w}}(T^*)\|_2^2 + \frac{1}{4}\|\boldsymbol{\theta}_a^0\|_2^2 + \|\boldsymbol{\theta}_{\boldsymbol{w}}(T^*)\|_2^2\right) \\
&\le 2\sqrt{d}\max\left\{\frac{1}{\kappa'}, 1\right\}\|\boldsymbol{\theta}_{\boldsymbol{w}}(T^*)\|_2^2 + \frac{\sqrt{d}}{4}\|\boldsymbol{\theta}_a^0\|_2^2,
\end{aligned}$$

where we have used Proposition 27. By Proposition 16, we have with probability at least $1-2\exp(-\frac{C_0 m(d+1)}{4C_{\psi,1}^2})$ over the choice of $\boldsymbol{\theta}^0$,

$$\|\boldsymbol{\theta}_a^0\| \le \sqrt{\frac{3}{2}m},$$
$$\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\| \le \sqrt{\frac{3}{2}dm}.$$

If $m\kappa \le \frac{1}{3\sqrt{d}}$, then $\frac{3\sqrt{d}m}{8} \le \frac{1}{8\kappa}$ and

$$\begin{aligned}\frac{1}{8\kappa} \le \frac{1}{4\kappa} - \frac{3\sqrt{d}m}{8} &\le \frac{1}{4\kappa} - \frac{\sqrt{d}}{4}\|\boldsymbol{\theta}_a^0\|_2^2 \\ &\le 2\sqrt{d}\max\left\{\frac{1}{\kappa'}, 1\right\}\|\boldsymbol{\theta}_{\boldsymbol{w}}(T^*)\|_2^2.\end{aligned}$$

Thus

$$\frac{\min\{1,\kappa'\}}{16\sqrt{d}\kappa} \le \|\boldsymbol{\theta}_{\boldsymbol{w}}(T^*)\|_2^2.$$

Therefore

$$\begin{aligned}\sup_{t\in[0,+\infty)} \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(t)-\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2} &\ge \frac{\|\boldsymbol{\theta}_{\boldsymbol{w}}(T^*)-\boldsymbol{\theta}_{\boldsymbol{w}}^0\|_2}{\|\boldsymbol{\theta}^0\|_2} \\ &\ge \sqrt{\frac{\frac{\min\{1,\kappa'\}}{16\sqrt{d}\kappa}}{\frac{3}{2}dm} - 1} \\ &\gtrsim \sqrt{\frac{\min\{1,\kappa'\}}{\kappa m}}.\end{aligned}$$

If $\gamma' < \gamma - 1$, then

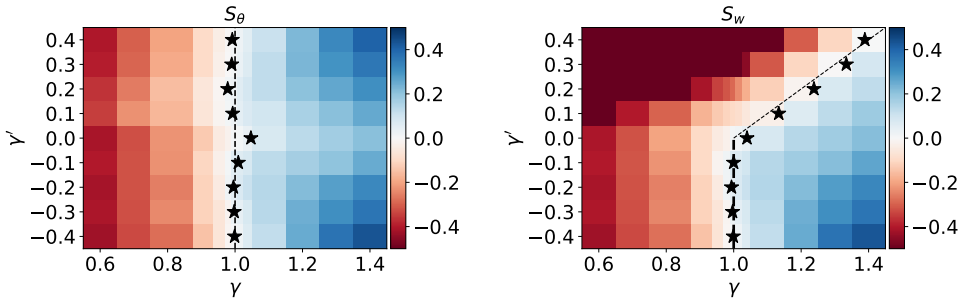
$$\frac{\min\{1,\kappa'\}}{\kappa m} \gg 1,$$

which completes the proof. ■

Remark 28. *Suppose that Assumption 1 and 2 hold. If $\gamma > 1$ and $\gamma' < 1-\gamma$, then Theorem 8 can hold without taking Assumption 3. Actually, for any $\delta \in (0,1)$, Proposition 26 guarantees the Assumption 3 with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, when m is sufficiently large. Therefore, under Assumptions 1 and 2, if m is sufficiently large, then we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, the relative change* $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \gg 1$.

Appendix D. Relative Deviation of Parameters

For completion, we can also similarly define the slope of the relative deviation for $\boldsymbol{\theta}$ and a denoted by $S_{\boldsymbol{\theta}}$ and S_a, respectively. As shown in Fig. 10 (a), the boundary for the $\boldsymbol{\theta}$ is



(a) (b)

Figure 10: $S_{\boldsymbol{\theta}}$ in (a) and $S_{\boldsymbol{w}}$ in (b) estimated on NNs of $1000, 5000, 10000, 20000, 40000$ neurons over γ (ordinate) and γ' (abscissa). The stars are zero points obtained by the linear interpolation over different γ for each fixed γ'. Dashed lines are auxiliary lines.

$\gamma = 1$, regardless of γ', that is, all parameters are close to their initialization after training. For output weight a, as shown in Fig. 10(b), the boundary consists of two rays, one is $\gamma = 1$ and $\gamma' \geq 0$, the other is $\gamma + \gamma' = 1$ and $\gamma' \leq 0$. This verifies that, in the area between $\gamma = 1$ and $\gamma - \gamma' = 1$ of $\gamma' \leq 0$, the change of scatter plot from a Gaussian initialization is induced by the change of a.

References

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *NeurIPS 2019*, 2019.

Zhengdao Chen, Grant M. Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. In *NeurIPS 2020*, 2020.

Lénaïc Chizat and Francis R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS 2018*, 2018.

Lénaïc Chizat, Edouard Oyallon, and Francis R. Bach. On lazy training in differentiable programming. In *NeurIPS 2019*, 2019.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *NeurIPS 2018*, 2018.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR 2019*, 2019.

Weinan E, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Sci. China Math.*, 63, 2020.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV 2015*, 2015.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS 2018*, 2018.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS 2019*, 2019.

Chao Ma, Lei Wu, and Weinan E. The quenching-activation behavior of the gradient descent dynamics for two-layer neural network models. *arXiv preprint arXiv:2006.14450*, 2020.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *COLT 2019*, 2019.

Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *NeurIPS 2020*, 2020.

Grant M. Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *NeurIPS 2018*, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Francis Williams, Matthew Trager, Daniele Panozzo, Cláudio T. Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. In *NeurIPS 2019*, 2019.

Blake E. Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in over-parametrized models. In *COLT 2020*, 2020.

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A type of generalization error induced by initialization in deep neural networks. In *MSML 2020*, 2020.