
Disentangle Sample Size and Initialization Effect on Perfect Generalization for Single-Neuron Target

Jiajie Zhao¹, Zhiwei Bai¹, Yaoyu Zhang^{1,2*}

¹ School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.

² Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 200240, P.R. China {zjj0216, bai299, zhyy.sjtu}@sjtu.edu.cn.

Abstract

Overparameterized models like deep neural networks have the intriguing ability to recover target functions with fewer sampled data points than parameters (Zhang et al., 2023a). To gain insights into this phenomenon, we concentrate on a single-neuron target recovery scenario, offering a systematic examination of how initialization and sample size influence the performance of two-layer neural networks. Our experiments reveal that a smaller initialization scale is associated with improved generalization, and we identify a critical quantity called the "initial imbalance ratio" that governs training dynamics and generalization under small initialization, supported by theoretical proofs. Additionally, we empirically delineate two critical thresholds in sample size—termed the "optimistic sample size" and the "separation sample size"—that align with the theoretical frameworks established by Zhang et al. (2023a,b). Our results indicate a transition in the model's ability to recover the target function: below the optimistic sample size, recovery is unattainable; at the optimistic sample size, recovery becomes attainable albeit with a set of initialization of zero measure. Upon reaching the separation sample size, the set of initialization that can successfully recover the target function shifts from zero to positive measure. These insights, derived from a simplified context, provide a perspective on the intricate yet decipherable complexities of perfect generalization in overparameterized neural networks.

1 Introduction

In machine learning, a fundamental problem is to learn a function from data sampled from a target function f^* with the goal of minimizing the generalization error. Traditional learning theory suggests that overparameterized models, where the number of parameters exceeds the number of sample points, are prone to overfitting and poor generalization (Vapnik, 1998; Bartlett and Mendelson, 2002). However, in practice, overparameterized deep neural networks often exhibit good generalization performance (Breiman, 2018; Zhang et al., 2021). To demystify this generalization phenomenon, researchers have sought to devise theoretical complexity measures to determine an upper bound on the generalization gap. Many proposed complexity measures are predicated on worst-case analyses, assessing the most unfavorable generalization scenarios within a given hypothesis space. Nonetheless, empirical investigations often show a weak or nonexistent relationship between these theoretical predictors and the observed generalization performance of actual models (Jiang et al., 2019).

Recent research has explored a novel concept contrary to the worst-case scenario, referred to as the "optimistic estimate". This investigates the minimum number of samples that models need to

*Corresponding author: zhyy.sjtu@sjtu.edu.cn.

exactly reconstruct the target function in the recoverable setting (Zhang et al., 2022, 2023a). Their experimental findings indicate that with appropriate hyperparameter tuning, the number of data points required to recover the target function can approach, or even match, the proposed "optimistic sample size". Furthermore, Zhang et al. (2023b) characterized the structure of the loss landscape of two-layer neural networks near global minima. They discovered that as the sample size reaches a certain threshold, called the "separation sample size", the set of parameters with zero generalization error Q^* , referred to as the target set, separates out (see Definition 1 for the formal definition of separation). However, the target set Q^* generally consists of different branches, and it remains unclear to which branch the actual training process will converge for different sample sizes and initialization.

In our study, we conduct a systematic exploration of the impact that initialization and sample size have on the dynamics and convergence results of a model. The challenge in studying neural network dynamics stems from its dependency on multiple factors, including the specific architecture, dataset, optimization technique, and initialization method. To dissect the global dynamics and generalization capabilities within overparameterized neural networks, we concentrate on a simplified scenario: the recovery of a single-neuron target. In our context, term "recovery" and "perfect generalization" are both identical to zero generalization error. Despite its simplicity, this scenario still represents an overparameterized system, and an in-depth examination can provide insights on more intricate situations. Our principal conclusions are encapsulated as follows:

Effect of Initialization Scale: Within the context of single-neuron target recovery, we experimentally demonstrate that smaller initialization scales are conducive to enhanced generalization.

Effect of Randomness: Randomness retains its significance even as the initialization scale nears zero; we pinpoint a critical variable, termed the "initial imbalance ratio", which serves as a determinant of the training dynamics and generalization error.

Effect of Sample Size: Our empirical results highlight two critical thresholds in sample size—the optimistic sample size and the separation sample size—that align with theoretical forecasts by Zhang et al. (2023a,b). Specifically, we empirically establish that:

- (i) Below optimistic sample size, the model cannot recover the target function.
- (ii) At optimistic sample size, a zero-measure subset of initialization can recover the target function.
- (iii) Once the sample size reaches the separation sample size, there exists a non-zero probability that certain combinations of initialization and sampling will successfully recover the target function.
- (iv) When sample size equals the number of parameters, all small-scale initialization can recover the target function.

2 Related works

The single-neuron fitting problem has been extensively studied, with various works investigating the convergence properties of networks in both exactly parameterized and overparameterized settings (Yehudai and Ohad, 2020; Vardi et al., 2021; Xu and Du, 2023; Vempala and Wilmes, 2018). These works have established results on the convergence rates and conditions for neural networks, laying a theoretical groundwork for discussions on generalization. When it comes to generalization, several studies have derived polynomial generalization bounds (Wu, 2022; Frei et al., 2020), while others have presented theoretical results of implicit regularization (Chistikov et al., 2024; Oymak and Soltanolkotabi, 2019; Safran et al., 2022). However, these analyses are typically restricted to the exactly parameterized setting or are specific to the ReLU activation function. In this paper, we empirically examine generalization enigmas in overparameterized networks with analytic activation functions. We demonstrate that perfect generalization is attainable for a certain sample size within the single-neuron target framework, offering a more nuanced characterization of generalization than the polynomial generalization bounds previously reported.

Recent theoretical advancements by Zhang et al. (2023a) and Zhang et al. (2022) introduced an optimistic estimate framework for general nonlinear models, suggesting that above a certain "optimistic sample size," some global minima become locally linearly stable, thereby allowing initializations close to these points to converge to stable solutions. Furthermore, Zhang et al. (2023b) delved into the branch structure of global minima in two-layer neural networks, defining a "separation sample size." Despite the theoretical importance of these findings, empirical validation has been limited. Our

study aims to bridge this gap by providing a systematic empirical investigation of how initialization and sample size influence the actual dynamics and convergence outcomes in neural network models.

3 Preliminaries

3.1 Notations

In this paper, we investigate a two-layer fully connected neural network represented by $f_{\theta}(\mathbf{x}) = \sum_{i=1}^m a_i \sigma(\mathbf{w}_i^{\top} \mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^d$ and $\theta = (a_1, \mathbf{w}_1, a_2, \mathbf{w}_2, \dots, a_m, \mathbf{w}_m) \in \mathbb{R}^{(d+1)m}$. The function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$ denotes the activation function, and m represents the width of the network. The target function we aim to approximate is a single-neuron function $f^*(\mathbf{x}) = a_0 \sigma(\mathbf{w}_0^{\top} \mathbf{x})$. The dataset $(\mathbf{x}_i, y_i)_{i=1}^n$ is generated by sampling from the target function f^* , that is, $y_i = f^*(\mathbf{x}_i)$ for $i = 1, 2, \dots, n$. We define the loss function as $\ell(\theta) = \frac{1}{2} \sum_{i=1}^n (f_{\theta}(\mathbf{x}_i) - y_i)^2$.

3.2 Optimistic sample size and separation sample size

The *target set*, denoted by Q^* , is defined as the set of parameters that achieve perfect generalization:

$$Q^* := \{\theta \mid f_{\theta}(\mathbf{x}) = f^*(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^d\}.$$

Zhang et al. (2023b) classified Q^* into several affine subspaces for a two-layer neural network without a bias term. We illustrate this with Example 1, where Q^* is the union of two affine spaces.

Example 1. Consider a neural network model $f_{\theta}(x) = a_1 \tanh(w_1 x + b_1) + a_2 \tanh(w_2 x + b_2)$, where $x \in \mathbb{R}$. Let the target function be $f^*(x) = a_0 \tanh(w_0 x + b_0)$. We define:

$$\begin{aligned} Q^1 &:= \{\theta \mid (w_1, b_1) = (w_2, b_2) = (w_0, b_0), a_1 + a_2 = a_0\} \\ Q^2 &:= \{\theta \mid (w_1, b_1) = (w_0, b_0), (w_2, b_2) \neq (w_0, b_0), a_1 = a_0, a_2 = 0\} \end{aligned}$$

Treating parameters symmetric about the origin as identical and the interchange of the two neurons as identical, we have $Q^1 \cup Q^2 = Q^*$ (See Figure 3(a) for geometric structure of Q^1 and Q^2).

Definition 1 (Separation of Q^k). A set Q^k is said to be separated if there exists an open neighborhood M around Q^k such that $M \cap \ell^{-1}(0) = Q^k \cap \ell^{-1}(0)$, where $\ell^{-1}(0)$ is the set of global minima. The minimum number of samples required for Q^k to be separated is referred to as the "separation sample size".

Definition 1 introduces the concept of separation and separation sample size. Zhang et al. (2023b) proves that (i) the separation sample sizes of Q^1 and Q^2 are 4 and 5, respectively, and (ii) when $n = 6$, $Q^* = \ell^{-1}(0)$.

Zhang et al. (2023a) proposed the concept of the "optimistic sample size," which determines the minimum number of samples required to achieve zero generalization error of a target function. In Example 1, the optimistic sample size is 3. Table 1 summarizes the various sample sizes of Example 1.

| sample size n | Name |
|-----------------|---------------------------------|
| $n = 3$ | optimistic sample size |
| $n = 4$ | separation sample size of Q^2 |
| $n = 5$ | separation sample size of Q^1 |
| $n = 6$ | $Q^* = \ell^{-1}(0)$ |

Table 1: Different sample sizes in Example 1 (Zhang et al., 2023a,b)

3.3 Experimental setup

Our methodology involves sampling a set of data points from the target function f^* and train the network f_{θ} until the parameters converge to θ_{∞} . To evaluate the training effectiveness, we measure the L_2 distance between f^* and the learned function $f_{\theta_{\infty}}$. The ideal outcome is that $f^* = f_{\theta_{\infty}}$, a condition we term "recovery" or "perfect generalization".

For training, we employ gradient descent with the update rule $\theta_{n+1} = \theta_n - \eta \nabla \ell(\theta_n)$, using a fixed learning rate η . Parameters are initialized according to a Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix σI , where I denotes the identity matrix. The standard deviation σ is referred to as the "initialization scale". We utilize a random seed to generate the Gaussian distribution, with each seed uniquely identified by a corresponding number.

4 Effect of initialization scale

Our experimental findings suggest a relationship between a small initialization scale of network's parameters and a lower generalization error. Figures 1(a) to 1(e) depict the generalization error across various initialization. The scale of initialization is represented by σ on the x-axis. For each initialization scale, we generate initialization using 100 distinct random seeds. The sample points are fixed, evenly spaced over the interval $[-2, 2]$. We observe that, regardless of the value of n , the generalization error tends to increase with the sample size, with this trend being particularly noticeable for $n = 3$ and $n = 4$. In Figure 1(f), we use samples drawn independently from an standard Gaussian distribution. Both samples and initialization are generated using 50 random seeds, and the generalization error is calculated as the average over these 50 trials. Results of Figure 1(f) indicate that, statistically, the generalization error decreases as the sample size grows, indicating that larger datasets are conducive to better generalization. Moreover, this reduction in error is more significant at smaller scales of initialization, highlighting the benefits of smaller initializations for achieving improved generalization.

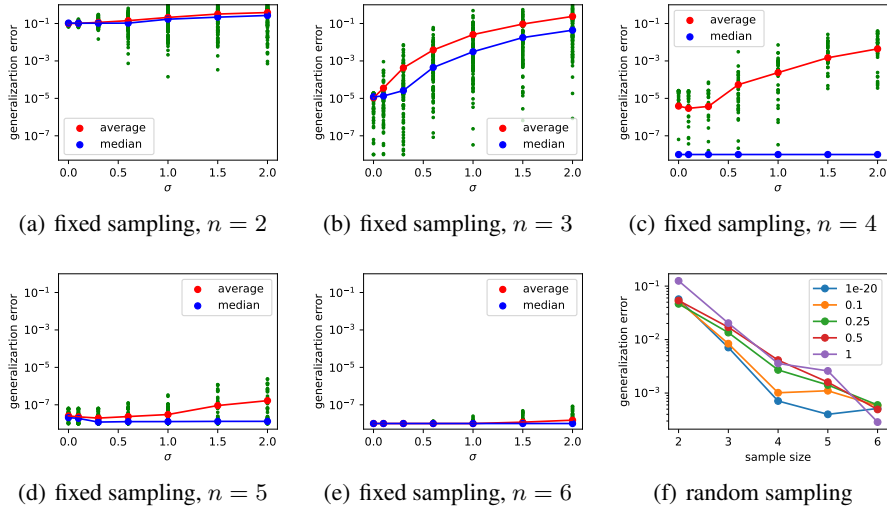


Figure 1: The network and target function correspond to Example 1. Here, n represents the sample size. For Figures 1(a) through 1(e), samples were evenly spaced on the interval $[-2, 2]$. In Figure 1(f), the dataset $\{(x_i, y_i)\}_{i=1}^n$ is such that $y_i = f^*(x_i)$, with the $\{x_i\}_{i=1}^n$ being independently and identically distributed according to a standard Gaussian distribution. For each combination of initialization scale and sample size, we conducted 50 trials with different seeds to generate data points and parameter initializations. The reported generalization error is the average over these trials. Curve legends indicate the initialization scale.

5 Effect of randomness of initialization

5.1 Initialization and trajectory of parameters

Previously, we empirically demonstrated that small-scale initialization enhances the generalization. This section delves into the dynamics of gradient flow under small initialization.

Theorem 1. Consider the gradient flow governed by the differential equation

$$\frac{d\theta}{dt} = -\nabla \ell(\theta(t)), \theta(0) = \theta_0, \quad (1)$$

where $\ell(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n (f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$ for $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, with the model $f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^\top \mathbf{x})$, and the parameter vector $\boldsymbol{\theta} = (a_1, \mathbf{w}_1, \dots, a_m, \mathbf{w}_m) \in \mathbb{R}^{m(d+1)}$. The solution to (1) is denoted by $\phi(\boldsymbol{\theta}_0, t)$. Define $\boldsymbol{\gamma} := \sum_{i=1}^n y_i \mathbf{x}_i$, $C_i(\boldsymbol{\theta}) := a_i \|\boldsymbol{\gamma}\|_2 + \mathbf{w}_i^\top \boldsymbol{\gamma}$, and $\mathbf{C}(\boldsymbol{\theta}) := (C_1(\boldsymbol{\theta}), \dots, C_m(\boldsymbol{\theta}))$. Assume the following conditions:

- (i) $\sigma(x)$ is twice continuously differentiable on \mathbb{R} , $\sigma(0) = 0$, and $\sigma'(0) \neq 0$.
- (ii) $\boldsymbol{\gamma} \neq \mathbf{0}$.

Under these assumptions, the following statements hold:

- (i) For any $t \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^{m(d+1)}$, the limit $h(\boldsymbol{\theta}, t) := \lim_{\alpha \rightarrow 0} \phi(\alpha \boldsymbol{\theta}, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \frac{1}{\alpha})$ exists.
- (ii) The function $h(\boldsymbol{\theta}_0, t)$ is determined by $\mathbf{C}(\boldsymbol{\theta}_0)$. That is, if $\mathbf{C}(\boldsymbol{\theta}_1) = \mathbf{C}(\boldsymbol{\theta}_2)$, then $h(\boldsymbol{\theta}_1, t) = h(\boldsymbol{\theta}_2, t)$ for all t .
- (iii) If $\mathbf{C}(\boldsymbol{\theta}_0) \neq \mathbf{0}$, then the trajectory $T_{\boldsymbol{\theta}_0} := \{h(\boldsymbol{\theta}_0, t) : t \in \mathbb{R}\}$ is determined by $\frac{\mathbf{C}(\boldsymbol{\theta}_0)}{\|\mathbf{C}(\boldsymbol{\theta}_0)\|_2}$. That is, if $\mathbf{C}(\boldsymbol{\theta}_1) = \mathbf{C}(\boldsymbol{\theta}_2)$, then $T_{\boldsymbol{\theta}_1} = T_{\boldsymbol{\theta}_2}$.

The proof of Theorem 1 is in the Appendix A.1. A more general result for dynamic systems is stated in Theorem 3 in Appendix A.1. To intuitively understand Theorem 1, we consider the linearization of Equation (1) at the origin. Assumption (i) of Theorem 1 ensures that $\nabla \ell(\mathbf{0}) = \mathbf{0}$, allowing us to approximate $-\nabla \ell(\boldsymbol{\theta}(t))$ by $-\text{Hess}(\ell(\mathbf{0}))\boldsymbol{\theta}$ when $\|\boldsymbol{\theta}\|_2$ is small. Under this linear approximation, the solution to Equation (1) can be expressed as $\boldsymbol{\theta}(t) \approx e^{-\text{Hess}(\ell(\mathbf{0}))t} \boldsymbol{\theta}_0$. When the norm $\|\boldsymbol{\theta}_0\|_2$ is sufficiently small, a large t is required for $\boldsymbol{\theta}(t)$ to move significantly away from the origin. In such cases, the largest eigenvalue of $-\text{Hess}(\ell(\mathbf{0}))$, denoted μ_1 , becomes dominant, leading to

$$\boldsymbol{\theta}(t) \approx e^{\mu_1 t} \sum_{i=1}^k (\boldsymbol{\theta}_0^\top \mathbf{v}_i) \mathbf{v}_i,$$

where $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is the orthonormal basis of the eigenspace corresponding to μ_1 . The evolution of $\boldsymbol{\theta}(t)$ is thus determined by the coefficients $\{\boldsymbol{\theta}_0^\top \mathbf{v}_i\}_{i=1}^k$, which are encapsulated in the vector $\mathbf{C}(\boldsymbol{\theta}_0)$ defined in Theorem 1.

In the context of two-layer neural networks, $\mathbf{C}(\boldsymbol{\theta}_0) = (C_1(\boldsymbol{\theta}_0), C_2(\boldsymbol{\theta}_0), \dots, C_m(\boldsymbol{\theta}_0))$ provides an insightful interpretation. The expression $\boldsymbol{\theta}(t) \approx e^{\mu_1 t} \sum_{i=1}^k (\boldsymbol{\theta}_0^\top \mathbf{v}_i) \mathbf{v}_i$ suggests that for $i = 1, 2, \dots, m$, the following approximations hold:

$$a_i(t) \approx \frac{C_i(\boldsymbol{\theta}_0)}{2\|\boldsymbol{\gamma}\|_2} e^{\|\boldsymbol{\gamma}\|_2 t}, \quad \mathbf{w}_i(t) \approx \frac{C_i(\boldsymbol{\theta}_0) \boldsymbol{\gamma}}{2\|\boldsymbol{\gamma}\|_2^2} e^{\|\boldsymbol{\gamma}\|_2 t}, \quad (2)$$

where $\boldsymbol{\gamma}$ is a vector determined by the data. Equation (2) indicates that the direction of the vector (a_i, \mathbf{w}_i) is consistent across all neurons, characterized by $\boldsymbol{\gamma}$. This observation is in line with the findings of Zhou et al. (2022), which show that neural networks with small initial weights tend to have input weights of hidden neurons aligning along certain data-determined directions. Moreover, Equation (2) indicates that the magnitude of (a_i, \mathbf{w}_i) is determined by $C_i(\boldsymbol{\theta}_0)$. Thus, the vector $\mathbf{C}(\boldsymbol{\theta}_0)$, representing the initial magnitudes of all neurons, determines early evolution of the parameters. The normalized vector $\frac{\mathbf{C}(\boldsymbol{\theta}_0)}{\|\mathbf{C}(\boldsymbol{\theta}_0)\|_2}$, representing the initial relative magnitudes of all neurons, determines trajectory of the parameters. Due to this, we refer $\frac{\mathbf{C}(\boldsymbol{\theta}_0)}{\|\mathbf{C}(\boldsymbol{\theta}_0)\|_2}$ as "initial imbalance ratio".

To corroborate the theoretical insights posited by Theorem 1, we conducted a series of experiments. We chose a model of the form $f_{\boldsymbol{\theta}}(x) = a_1 \tanh(w_1 x + b_1) + a_2 \tanh(w_2 x + b_2)$, with the target function defined as $f^*(x) = \tanh(x + 1)$. According to Theorem 1, under small initialization, the parameter trajectory is determined by the normalized vector $\left(\frac{C_1(\boldsymbol{\theta}_0)}{\sqrt{C_1^2(\boldsymbol{\theta}_0) + C_2^2(\boldsymbol{\theta}_0)}}, \frac{C_2(\boldsymbol{\theta}_0)}{\sqrt{C_1^2(\boldsymbol{\theta}_0) + C_2^2(\boldsymbol{\theta}_0)}} \right)$. Given that $\sigma(x) = \tanh(x)$ is an odd function, the sign inversion of both $C_1(\boldsymbol{\theta}_0)$ and $C_2(\boldsymbol{\theta}_0)$ leads to a symmetric trajectory about the origin, which we consider equivalent. Therefore, the trajectory is effectively characterized by the ratio $\frac{C_1(\boldsymbol{\theta}_0)}{C_2(\boldsymbol{\theta}_0)}$, denoted by $c(\boldsymbol{\theta}_0) := \frac{C_1(\boldsymbol{\theta}_0)}{C_2(\boldsymbol{\theta}_0)}$. For simplicity, we will henceforth denote $c(\boldsymbol{\theta}_0)$ by c .

Figure 2 shows the training results across five trials. Each trial used a different initialization scale and random seed to generate Gaussian distribution but kept the ratio $c = 0.5$ across five trials by scaling the initialization of second neuron. The results demonstrate that both the loss and the parameter trajectories were consistent across all trials, lending strong empirical support to the theoretical prediction that the trajectory is governed by the ratio c under small initialization.

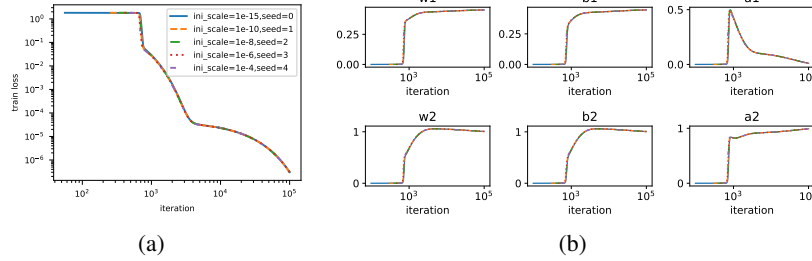


Figure 2: The network and target function correspond to Example 1. We trained the network across five trials, each utilizing an evenly spaced 6 data points within the interval $[-2, 2]$. Distinct initialization seeds and scales were used for each trial, but by scaling the initial parameters of the second neuron, we keep $c = 0.5$ across all trials. To align the curves, we applied translations based on distances calculated by Theorem 1.

5.2 Initialization and convergence point

As the scale of initialization approaches zero, the parameters' trajectory has a limit. A natural question arises: does the convergence point of the parameters also tend towards a limit as the initialization scale becomes infinitesimally small? We affirmatively address this question in Theorem 2.

Theorem 2. *Under the notations and assumptions of Theorem 1, and assuming that $\sigma(x)$ is analytic, then:*

- (i) *The limit $h(\boldsymbol{\theta}, t) := \lim_{\alpha \rightarrow 0} \phi(\alpha \boldsymbol{\theta}, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \frac{1}{\alpha})$ exists.*
- (ii) *For any $\boldsymbol{\theta}$, if the set $\{h(\boldsymbol{\theta}, t) : t \geq 0\}$ is bounded, then the limit $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t)$ exists and is determined by the normalized vector $\frac{\mathbf{C}(\boldsymbol{\theta})}{\|\mathbf{C}(\boldsymbol{\theta})\|_2}$. Specifically, if $\frac{\mathbf{C}(\boldsymbol{\theta}_1)}{\|\mathbf{C}(\boldsymbol{\theta}_1)\|_2} = \frac{\mathbf{C}(\boldsymbol{\theta}_2)}{\|\mathbf{C}(\boldsymbol{\theta}_2)\|_2}$, then $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_1, t) = \lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_2, t)$.*
- (iii) *If $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_0, t)$ exists and is not a saddle point of $\ell(\boldsymbol{\theta})$, then*

$$\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_0, t) = \lim_{\alpha \rightarrow 0} \lim_{t \rightarrow \infty} \phi(\alpha \boldsymbol{\theta}_0, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \frac{1}{\alpha}).$$

Additionally, the limit $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t)$ is continuous at $\boldsymbol{\theta}_0$.

The proof of Theorem 2 is presented in the Appendix A.3. We conduct experiments when $f_{\boldsymbol{\theta}}(x) = a_1 \tanh(w_1 x + b_1) + a_2 \tanh(w_2 x + b_2)$ and $f^*(x) = \tanh(x + 1)$. Literature suggests that convergence to a saddle point is rare (Panageas et al., 2019) for gradient flow. Moreover, in neural network experiments, divergence of $\boldsymbol{\theta}(t)$ to infinity is seldom observed. Thus, the conditions of (ii) and (iii) are typically met. The conclusion (iii) of Theorem 2 affirms that for a small initialization scale, the convergence point of the parameters can be brought arbitrarily close to $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_0, t)$. The conclusion of (ii) implies that $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_0, t)$ is determined by $\frac{\mathbf{C}}{\|\mathbf{C}\|_2}$. Therefore, under sufficiently small initialization, the vector $\frac{\mathbf{C}}{\|\mathbf{C}\|_2}$ almost determines the final convergence point of the parameters. Define $c := C_1/C_2$ and $\tilde{c} := \min\{|c|, |\frac{1}{c}|\}$. Networks initialized with c and $\frac{1}{c}$ are identical upon permuting the neurons and network initialized with c and $-c$ yields trajectories symmetric about the origin. Hence, \tilde{c} effectively encompasses all cases of initialization by accounting for symmetry.

In Figure 3, we investigate the impact of initialization on the convergence point under small initialization. Figure 3(a) visualizes the target set Q^* alongside the convergence points for various \tilde{c} . The line is Q^1 , while the surface are Q^2 . These simulations, run over 10^6 iterations with a learning rate of 0.05 and a sample size of 6, reveal two significant observations:

- (i) The parameters consistently converge to Q^* across different initialization, corroborating the theoretical results in Table 1 that for $n \geq 6$, the global minimum coincides with Q^* .
- (ii) The convergence points lie on a one-dimensional manifold parameterized by \tilde{c} . Notably, as \tilde{c} nears 1, the convergence point moves closer to Q^1 .

Figure 3(b) further illustrates the convergence behavior of the network’s parameters for diverse initialization, focusing on convergence to points Q^1 or Q^2 . The results demonstrate that convergence occurs at Q^1 when c approximates 1. For c values significantly divergent from 1, convergence at Q^2 becomes more likely, with the demarcation at $c = 1.35$ and $c = 0.74$.

In the small initialization dynamics of a neural network, ratio c governs relative magnitude of two neurons at initial stage. A c close to 1 suggests similar magnitudes for both neurons, leading to convergence at Q^1 , where each neuron contributes to the output function. Conversely, a substantially larger c implies that the first neuron’s magnitude predominates, resulting in convergence at Q^2 , where the second neuron’s output contribution is zero. The two extremes, $c = 1$ and $c = +\infty$, demonstrate this: for $c = 1$, the ratios $\frac{a_1}{a_2}$, $\frac{w_1}{w_2}$, and $\frac{b_1}{b_2}$ remain constant at 1 for all t , converging to Q^1 . For $c = +\infty$, we have $a_2 = w_2 = b_2 = 0$ for all t , resulting in convergence at Q^2 .

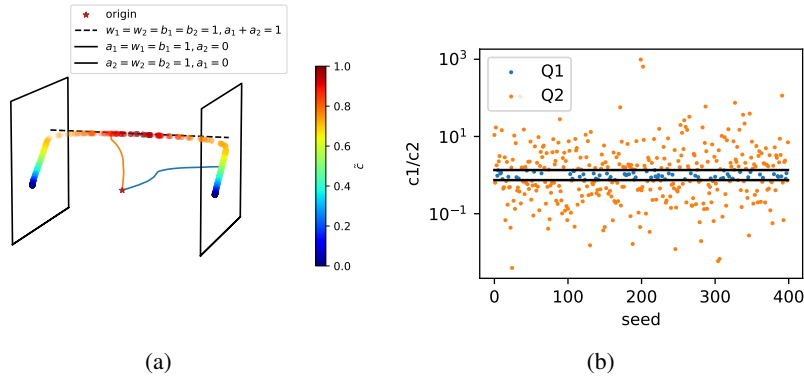


Figure 3: The network and target function correspond to Example 1 with a sample size of 6 and an initialization scale of 10^{-8} . We utilized 400 random seeds to initialize the parameters. Figure 3(a) shows the convergence points and the structures of Q^1 and Q^2 , along with the origin and two exemplary training trajectories. The dashed line is Q^1 and the affine surface is Q^2 . Figure 3(b) presents the convergence results using seeds 0-400, where blue and orange represent convergence to Q^1 and Q^2 , respectively. The x-axis denotes the seed index, and the y-axis measures the absolute value of the ratio C_1/C_2 . The two black horizontal lines mark the ratios at $y = 1.35$ and $y = 0.74$.

6 Effect of sample size

This section delves into the impact of training sample size on the network’s ability to achieve recovery. In Figure 4, we present the generalization error under various combinations of initialization and sample sizes. The network in question is defined as $f_\theta(x) = a_1 \tanh(w_1x + b_1) + a_2 \tanh(w_2x + b_2)$, with the target function being $f^*(x) = \tanh(x + 1)$. For small initialization scale, the ratio $\tilde{c} = \min\{|c|, |\frac{1}{c}|\}$ adequately represents all initialization. When $n = 2$, the network fails to recover the target function for any initialization, in line with the optimistic sample size theory (Zhang et al., 2023a). To elucidate, a single-neuron target encompasses 3 parameters. With only two samples, an infinite number of single-neuron targets could fit, preventing the network from identifying the desired target.

For $n = 3$ (see Figure 4(a)), recovery is feasible solely when $\tilde{c} = 0$ or $\tilde{c} = 1$. Initially, with $n = 3$, neither Q^1 nor Q^2 is separated (see Table 1). The target sets Q^1 and Q^2 , enveloped by global minima, lead the network to likely converge to these global minima rather than the target set. This accounts for the lack of recovery when $\tilde{c} \in (0, 1)$. Nonetheless, fortuitous scenarios occur. When $\tilde{c} = 1$, the ratios $\frac{a_1}{a_2} = \frac{w_1}{w_2} = \frac{b_1}{b_2} = 1$ hold during training, simplifying the two-neuron network to a single-neuron model, reducing six parameters to three. In such instances, three samples implies that the global

minimum equals the target set (Zhang et al., 2023b), facilitating recovery when $\tilde{c} = 1$. The case of $\tilde{c} = 0$ is similar.

For $n = 4$ (see Figure 4(c) and Figure 4(f)), recovery is attainable with a positive probability of sampling and initialization. Meanwhile, with some samples, recovery remains unattainable for all initialization. Table 1 indicates that for $n = 4$, Q^2 is separated, whereas Q^1 is not. Our findings corroborate that once Q^2 is separated, convergence to it is plausible with a positive probability. Additionally, we demonstrate the existence of samples for which no small-scale initialization leads to convergence to the target set.

For $n = 5$, some samples enable the network to recover for all small initialization, while others do not. Table 1 shows that for $n = 5$, both Q^1 and Q^2 are separated. Our experiments suggest that once Q^1 and Q^2 are separated, under some samples, recovery is achievable across all initialization. Moreover, at $n = 5$, certain global minima with non-zero generalization error may be encountered during training with specific samples. For $n = 6$, the network recovers for all small initialization, as all global minima are associated with zero generalization error (see Table 1).

Using an analogy, we can liken the process of recovery to archery. The size of the training sample dictates the structural configuration of the target. Once the sample size surpasses the threshold known as the separation sample size, certain sections of the target become exposed and unobscured by any other global minima, rendering them accessible with a non-zero probability. Concurrently, there are certain shortcuts that facilitate reaching the target more directly. Specifically, when $\tilde{c} = 0$ or $\tilde{c} = 1$, the network is capable of hitting the target at the so-called optimistic sample size, even if Q^1 and Q^2 remain concealed by global minima. In essence, the sample size molds the target's architecture, while the initialization steers the direction of the "shot".

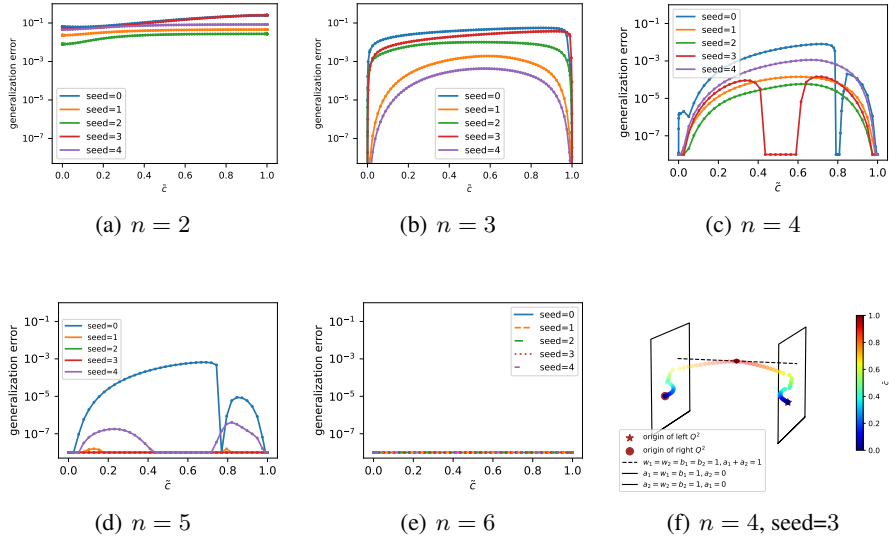


Figure 4: The network $f_\theta(x)$ and target function f^* correspond to Example 1. The samples $\{(x_i, y_i)\}_{i=1}^n$, where $y_i = f^*(x_i)$, is obtained by drawing $\{x_i\}_{i=1}^n$ from a standard Gaussian distribution. Five random seeds were used to generate the samples. Generalization errors below 10^{-8} are considered as successful recovery and identified with 10^{-8} . Figure 4(f) depicts the convergence point for $n = 4$ with samples generated by seed 3. The dashed line is Q^1 and the affine surface is Q^2 . All experiments were initialized with a scale of 10^{-20} .

7 Extension to multi-neuron networks

The insights from the two-neuron, two-layer neural network analysis can be generalized to networks with multiple neurons. Our experiments on a neural network with a width of 1000 and the activation function $\sigma(x) = \frac{x}{1+x^2}$ support this generalization. As depicted in Figure 5, the experiments reveal

that under small initialization conditions, the parameter trajectories conform to the set $\frac{C}{\|C\|_2}$, as postulated in Theorem 1.

Figure 6(a) shows that in a two-layer neural network approximating a single-neuron target function, only a subset of neurons develop substantial weights and become key contributors to output function. These neurons are distinguished by having the largest C_k values, aligning with earlier experimental observations of two-neuron network that neurons with higher C_k values tend to have greater magnitudes. In Figure 6(b), we note that the generalization error is significantly low in an overparameterized network. This is partly attributed to the phenomenon in Figure 6(a), where most neurons possess minimal weights compared to the neuron with the largest weight magnitude. Consequently, the neural network operates as if it has fewer active neurons. This decrease in active neuron count effectively reduces the network’s complexity and improves its generalization capability. Besides these experiments, we also conducted experiments with higher dimensional input (see Appendix B).

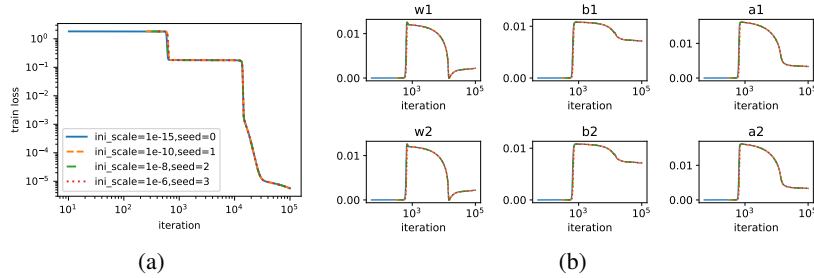


Figure 5: A two-layer neural network with a width of 1000 and activation function $\sigma(x) = \frac{x}{1+x^2}$ is trained on 6 evenly spaced data points in the interval $[-2, 2]$ with labels given by $y = \tanh(x + 1)$. Four trials with varying initialization seeds and scales were conducted. The ratio of initial parameters C_i/C_1 is set to $1.5 + 0.0015(i - 1)$ for each neuron $i = 1, 2, \dots, 1000$ in all trials. For visualization, curves in Figure 5(a) are translated based on distances derived from Theorem 1. Figure 5(b) shows the parameter trajectories for the first two neurons.

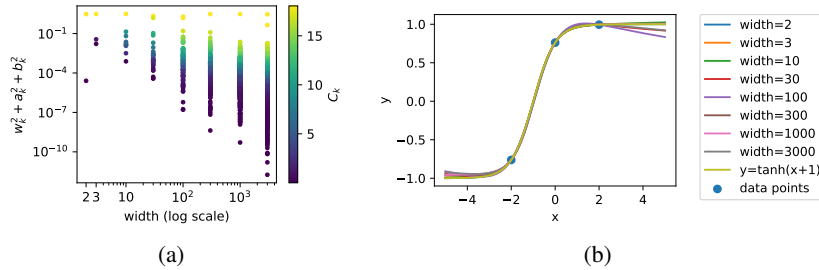


Figure 6: Visualization of training dynamics and outcomes for a two-layer neural network with variable widths. The networks use the $\tanh(x)$ activation function and are trained to approximate the target function $y = \tanh(x + 1)$ using a dataset of 3 points equally spaced within the interval $[-2, 2]$. Figure 6(a) shows the magnitude of the weights for individual, with each dot representing a neuron and the dot color indicating the absolute value of the scaling factor C_k for the k -th neuron. Figure 6(b) illustrates the final output functions of the networks with different widths after training.

8 Conclusion

Our investigation into the learning of single-neuron target functions within two-layer neural networks has elucidated the pivotal influence of initialization scale, randomness, and sample size on achieving perfect generalization. We found that smaller initialization scales and larger sample sizes tend to enhance generalization performance, while the element of randomness plays a significant role in shaping the learning outcomes. By honing in on the concept of perfect generalization, we have simplified the complexity inherent in the generalization puzzle and have empirically validated the

existence of both optimistic and separation sample size thresholds. These observations underscore the intricate interplay among initialization, stochastic elements, and sample size in the learning process of neural networks.

We must recognize the limitations imposed by the simplicity of our experimental framework, which was confined to the recovery of a single neuron. Additionally, in Theorem 2, we assume convergence to a local minimum instead of proving it. Further research into the learning behaviors of networks with more complex target functions, as well as the effects of critical points on learning dynamics, represents a compelling direction for future inquiry.

Acknowledgments and Disclosure of Funding

This work is sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200, the National Natural Science Foundation of China Grant No. 12101402, the Lingang Laboratory Grant No. LG-QS-202202-08, Shanghai Municipal of Science and Technology Major Project No. 2021SHZDZX0102.

References

- Y. Zhang, Z. Zhang, L. Zhang, Z. Bai, T. Luo, Z.-Q. J. Xu, Optimistic estimate uncovers the potential of nonlinear models, arXiv preprint arXiv:2307.08921 (2023a).
- L. Zhang, Y. Zhang, T. Luo, Structure and gradient dynamics near global minima of two-layer neural networks, 2023b. arXiv:2309.00508.
- V. N. Vapnik, Adaptive and learning systems for signal processing communications, and control, Statistical learning theory (1998).
- P. L. Bartlett, S. Mendelson, Rademacher and gaussian complexities: Risk bounds and structural results, Journal of Machine Learning Research 3 (2002) 463–482.
- L. Breiman, Reflections after refereeing papers for nips, in: The Mathematics of Generalization, CRC Press, 2018, pp. 11–15.
- C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM 64 (2021) 107–115.
- Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, S. Bengio, Fantastic generalization measures and where to find them, arXiv preprint arXiv:1912.02178 (2019).
- Y. Zhang, Z. Zhang, L. Zhang, Z. Bai, T. Luo, Z.-Q. J. Xu, Linear stability hypothesis and rank stratification for nonlinear models, arXiv preprint arXiv:2211.11623 (2022).
- G. Yehudai, S. Ohad, Learning a single neuron with gradient methods, in: Conference on Learning Theory, PMLR, 2020, pp. 3756–3786.
- G. Vardi, G. Yehudai, O. Shamir, Learning a single neuron with bias using gradient descent, Advances in Neural Information Processing Systems 34 (2021) 28690–28700.
- W. Xu, S. Du, Over-parameterization exponentially slows down gradient descent for learning a single neuron, in: The Thirty Sixth Annual Conference on Learning Theory, PMLR, 2023, pp. 1155–1198.
- S. Vempala, J. Wilmes, Polynomial convergence of gradient descent for training one-hidden-layer neural networks, arXiv preprint arXiv:1805.02677 (2018).
- L. Wu, Learning a single neuron for non-monotonic activation functions, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 4178–4197.
- S. Frei, Y. Cao, Q. Gu, Agnostic learning of a single neuron with gradient descent, Advances in Neural Information Processing Systems 33 (2020) 5417–5428.

- D. Chistikov, M. Englert, R. Lazic, Learning a neuron by a shallow relu network: Dynamics and implicit bias for correlated inputs, *Advances in Neural Information Processing Systems* 36 (2024).
- S. Oymak, M. Soltanolkotabi, Overparameterized nonlinear learning: Gradient descent takes the shortest path?, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 4951–4960.
- I. Safran, G. Vardi, J. D. Lee, On the effective number of linear regions in shallow univariate relu networks: Convergence guarantees and implicit bias, *Advances in Neural Information Processing Systems* 35 (2022) 32667–32679.
- H. Zhou, Z. Qixuan, T. Luo, Y. Zhang, Z.-Q. Xu, Towards understanding the condensation of neural networks at initial training, *Advances in Neural Information Processing Systems* 35 (2022) 2184–2196.
- I. Panageas, G. Piliouras, X. Wang, First-order methods almost always avoid saddle points: The case of vanishing step-sizes, *Advances in Neural Information Processing Systems* 32 (2019).
- Z. Li, Y. Luo, K. Lyu, Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning, *arXiv preprint arXiv:2012.09839* (2020).
- S. Lojasiewicz, *Ensembles semi-analytiques*, *Lectures Notes IHES (Bures-sur-Yvette)* (1965).

A Proof of Theorems

A.1 Proof of Theorem 1

We begin by establishing Theorem 3, and then leverage it to validate Theorem 1.

Let $\phi(\boldsymbol{\theta}_0, t)$ denote the solution to the following differential equation (3):

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= g(\boldsymbol{\theta}), \\ \boldsymbol{\theta}(0) &= \boldsymbol{\theta}_0, \end{aligned} \tag{3}$$

Theorem 3. *Assume the conditions:*

- (i) $g(\boldsymbol{\theta})$ is twice continuously differentiable.
- (ii) $g(\mathbf{0}) = \mathbf{0}$.
- (iii) $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$ is diagonalizable.
- (iv) The largest eigenvalue of $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$ is positive.

Denote the solution 3 as $\phi(\boldsymbol{\theta}_0, t)$. Then, the limit $\lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha}))$ exists, and the rate of convergence is $\alpha^{\frac{\mu_1 - \mu_2}{2\mu_1 - \mu_2}}$, where μ_1 and μ_2 are the largest and second-largest eigenvalues of $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$, respectively. \mathbf{v}_1 is a vector in the eigenspace corresponding to μ_1 . The vector \mathbf{u}_α is arbitrary, subject to the conditions:

- (i) \mathbf{u}_α is orthogonal to the eigenspace of μ_1 .
- (ii) $\exists c > 0$, such that $\forall \alpha > 0$, $\|\mathbf{u}_\alpha\|_2 \leq c\alpha$

Intuition of the Theorem:

Let us denote $\nabla g(\boldsymbol{\theta})$ by $\mathbf{J}(t)$. Linearizing the dynamics around the origin, we have:

$$\frac{d\boldsymbol{\theta}}{dt} = \mathbf{J}(0)\boldsymbol{\theta},$$

which yields the linearized solution:

$$\boldsymbol{\theta}(t) = e^{\mathbf{J}(0)t} \boldsymbol{\theta}_0.$$

When the initial condition $\boldsymbol{\theta}_0$ is very small, a large t is required to move away from zero. Consequently, the top eigenvalue of $\mathbf{J}(0)$ will dominate when computing $e^{\mathbf{J}(0)t}$. Hence, only the projection of $\boldsymbol{\theta}_0$ onto the eigenspace corresponding to μ_1 will significantly affect the trajectory of the dynamics.

Sketch of the Proof:

We consider an ϵ -ball centered at the origin and determine when the trajectory of $\boldsymbol{\theta}(t)$ intersects with this ϵ -ball. On one hand, ϵ cannot be too small; otherwise, the time t would be small, and the exponential term $e^{\mu_1 t}$ would not be dominant. On the other hand, ϵ cannot be too large; otherwise, the linear approximation of $g(\boldsymbol{\theta})$ would not be valid. Therefore, we must choose an appropriate value for ϵ and analyze the error caused by the two reasons mentioned above.

Formal Proof of Theorem 3: Because $\mathbf{J}(0)$ is diagonalizable, we can transform the coordinate system such that $\mathbf{J}(0)$ becomes a diagonal matrix. Without loss of generality, we assume that $\mathbf{J}(0)$ is the diagonal matrix $\text{diag}\{\mu_1, \mu_2, \dots, \mu_d\}$. We reference Lemma E.3, Lemma E.4, and Lemma E.5 from Li et al. (2020), which prove a special case of Theorem 3 where the eigenspace corresponding to μ_1 is one-dimensional.

We restate their lemmas, denoting $F(x) = \log(x) - \log(1 + \kappa x)$. Let $T_\alpha(r) = \frac{1}{\mu_1}(F(r) - F(\alpha))$. Let $R > 0$. Since $g(\boldsymbol{\theta})$ is \mathcal{C}^2 -smooth, there exists $\beta > 0$ such that $\|\mathbf{J}(\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta} + \mathbf{h})\|_2 \leq \beta \|\mathbf{h}\|_2$, for all $\|\boldsymbol{\theta}\|_2, \|\boldsymbol{\theta} + \mathbf{h}\|_2 \leq R$. Then we have:

Lemma E.3. For $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t)$ with $\|\boldsymbol{\theta}_0\|_2 \leq \alpha$ and $t \leq T_\alpha(r)$, it holds that

$$\|\boldsymbol{\theta}(t)\|_2 \leq \frac{1 + \kappa r}{1 + \kappa \alpha} \alpha \cdot e^{\mu_1 t} \leq r.$$

Lemma E.4. For $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t)$ with $\|\boldsymbol{\theta}_0\|_2 \leq \alpha$ and $t \leq T_\alpha(r)$, we have

$$\boldsymbol{\theta}(t) = e^{t\mathbf{J}(0)} \boldsymbol{\theta}_0 + O(r^2).$$

Lemma E.5. Let $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t)$ and $\hat{\boldsymbol{\theta}}(t) = \phi(\hat{\boldsymbol{\theta}}_0, t)$. If $\max\{\|\boldsymbol{\theta}_0\|_2, \|\hat{\boldsymbol{\theta}}_0\|_2\} \leq \alpha$, then for $t \leq T_\alpha(r)$,

$$\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \leq e^{\mu_1 t + \kappa r} \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0\|_2.$$

Where $\kappa = \frac{\beta}{\mu_1}$.

Given α , we compare $\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha}))$ and $\phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}))$, where α' is an arbitrary number smaller than α . Let ϵ be an indeterminate number. Define $t_1 = T_\alpha(\epsilon)$, $t_2 = T_{\alpha'}(\epsilon)$, and $t_0 = \frac{1}{\mu_1} \log(\frac{\alpha}{\alpha'})$. Then

$$t_2 - t_1 = \frac{1}{\mu_1} \log\left(\frac{\alpha}{\alpha'}\right) - \frac{1}{\mu_1} \log\left(\frac{1 + \kappa \alpha}{1 + \kappa \alpha'}\right) < t_0,$$

implying that $t_2 - t_0 < t_1$.

Applying Lemma E.3 with $r = \epsilon$, $t = t_2 - t_0$, $\alpha = \alpha$, we obtain

$$\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_2 - t_0) = e^{(t_2 - t_0)\mathbf{J}(0)}(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha) + O(\epsilon^2).$$

Similarly, applying Lemma E.3 with $r = \epsilon$, $t = t_2$, $\alpha = \alpha'$, we get

$$\phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_2) = e^{t_2 \mathbf{J}(0)}(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}) + O(\epsilon^2).$$

Since

$$e^{(t_2 - t_0)\mathbf{J}(0)} \alpha \mathbf{v}_1 = e^{(t_2 - t_0)\mu_1} \alpha \mathbf{v}_1 = e^{t_2 \mu_1} \alpha' \mathbf{v}_1,$$

we have

$$\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_2 - t_0) - \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_2)\|_2 = \|e^{t_2 \mathbf{J}(0)} \mathbf{u}_{\alpha'} + e^{(t_2 - t_0)\mathbf{J}(0)} \mathbf{u}_\alpha\|_2 + O(\epsilon^2).$$

Furthermore, we have

$$\|e^{t_2 \mathbf{J}(0)} \mathbf{u}_{\alpha'}\|_2 \leq e^{\mu_2 t_2} c \alpha' = \left(\frac{\epsilon(1 + \kappa \alpha')}{\alpha'(1 + \kappa \epsilon)}\right)^{\frac{\mu_2}{\mu_1}} c \alpha',$$

$$\|e^{(t_2 - t_0)\mathbf{J}(0)} \mathbf{u}_\alpha\|_2 \leq e^{\mu_2(t_2 - t_0)} c \alpha = \left(\frac{\epsilon(1 + \kappa \alpha')}{\alpha(1 + \kappa \epsilon)}\right)^{\frac{\mu_2}{\mu_1}} c \alpha.$$

Thus, we obtain

$$\begin{aligned} \|\phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_2) - \phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_2 - t_0)\|_2 &\leq O(\epsilon^2) + \left(\frac{\epsilon(1 + \kappa\alpha')}{\alpha(1 + \kappa\epsilon)}\right)^{\frac{\mu_2}{\mu_1}} c\alpha + \left(\frac{\epsilon(1 + \kappa\alpha')}{\alpha'(1 + \kappa\epsilon)}\right)^{\frac{\mu_2}{\mu_1}} c\alpha' \\ &= O(\epsilon^2) + O\left(\left(\frac{\epsilon}{\alpha}\right)^{\frac{\mu_2}{\mu_1}}\right) \alpha. \end{aligned}$$

Applying Lemma E.3, we conclude that $\|\phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_2)\|_2 \leq \epsilon$ and $\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_2 - t_0)\|_2 \leq \epsilon$. Define $\Delta t = t + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}) - t_2 = t + \frac{1}{\mu_1} \log(\frac{1}{\epsilon}) + \frac{1}{\mu_1} \log(\frac{1 + \kappa\epsilon}{1 + \kappa\alpha'})$.

Because $\Delta t \leq T_\epsilon(r) \iff \mu_1 t - \log(1 + \kappa\alpha') \leq \log(\frac{r}{1 + \kappa r})$, so we only need $\mu_1 t \leq \log(\frac{r}{1 + \kappa r})$ to satisfy $\Delta t \leq T_\epsilon(r)$.

Case1: Assume $\mu_1 t \leq \log(\frac{R}{1 + \kappa R})$.

let $r = \frac{1}{e^{-\mu_1 t - \kappa}}$, then $r \leq R$ and $\Delta t \leq T_\epsilon(r)$. Applying Lemma E.5 with $\theta_0 = \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_2)$, $\hat{\theta}_0 = \phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_2 - t_0)$, $t = \Delta t$, $\alpha = \epsilon$, and $r = \frac{1}{e^{-\mu_1 t - \kappa}}$, we get

$$\begin{aligned} &\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha})) - \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}))\|_2 \\ &\leq e^{\mu_1 \Delta t + k r} \times \|\phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_2) - \phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_2 - t_0)\|_2 \\ &\leq e^{\mu_1 t + k r} \times O\left(\frac{1}{\epsilon}\right) \times \left(O(\epsilon^2) + O\left(\left(\frac{\epsilon}{\alpha}\right)^{\frac{\mu_2}{\mu_1}}\right) \alpha\right) \\ &= e^{\mu_1 t + k r} \times \left(O(\epsilon) + O\left(\left(\frac{\alpha}{\epsilon}\right)^{1 - \frac{\mu_2}{\mu_1}}\right)\right). \end{aligned}$$

Setting $\epsilon = \alpha^s$ where $0 < s < 1$, we find

$$e^{\mu_1 t + k r} \times \left(O(\epsilon) + O\left(\left(\frac{\alpha}{\epsilon}\right)^{1 - \frac{\mu_2}{\mu_1}}\right)\right) = e^{\mu_1 t + k r} \times O\left(\alpha^{\min\{s, (1-s)(1 - \frac{\mu_2}{\mu_1})\}}\right).$$

Choosing $s = \frac{\mu_1 - \mu_2}{2\mu_1 - \mu_2}$, we obtain the tightest bound:

$$\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha})) - \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}))\|_2 \leq e^{\mu_1 t + k r} \times O\left(\alpha^{\frac{\mu_1 - \mu_2}{2\mu_1 - \mu_2}}\right).$$

Case2: Assume $\mu_1 t > \log(\frac{R}{1 + \kappa R})$.

Denote $t_s = \frac{1}{\mu_1} \log(\frac{R}{1 + \kappa R})$ and $\tau := t - t_s$. Then $\mu_1 t_s \leq \log(\frac{R}{1 + \kappa R})$. Applying results of **Case1**, we get:

$$\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_s + \frac{1}{\mu_1} \log(\frac{1}{\alpha})) - \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_s + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}))\|_2 \leq e^{\mu_1 t_s + k R} \times O\left(\alpha^{\frac{\mu_1 - \mu_2}{2\mu_1 - \mu_2}}\right)$$

Because $\phi(\theta, t)$ is locally Lipschitz over θ , so

$$\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha})) - \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}))\|_2 \quad (4)$$

$$= \|\phi\left(\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_s + \frac{1}{\mu_1} \log(\frac{1}{\alpha})), \tau\right) - \phi\left(\phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_s + \frac{1}{\mu_1} \log(\frac{1}{\alpha'})), \tau\right)\|_2 \quad (5)$$

$$= O(\|\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t_s + \frac{1}{\mu_1} \log(\frac{1}{\alpha})) - \phi(\alpha' \mathbf{v}_1 + \mathbf{u}_{\alpha'}, t_s + \frac{1}{\mu_1} \log(\frac{1}{\alpha'}))\|_2) \quad (6)$$

$$= O\left(\alpha^{\frac{\mu_1 - \mu_2}{2\mu_1 - \mu_2}}\right) \quad (7)$$

In both case, $\phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha}))$ as a sequence of α satisfies the Cauchy criterion. Therefore, $\lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}_1 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log(\frac{1}{\alpha})) := h(\mathbf{v}_1, t)$ exists. Moreover, since $e^{\mu_1 t + k r}$ is bounded in a neighborhood of \mathbf{v}_1 and t , the convergence is uniform, ensuring that $h(\mathbf{v}_1, t)$ is continuous with respect to \mathbf{v}_1 and t . \square

A.2 Corollary of Theorem 1

Corollary 1. *Assume the assumptions of Theorem 3 hold. Let \mathbf{v} be a vector in the parameter space, and let \mathbf{v}_1 be the projection of \mathbf{v} into the eigenspace corresponding to the largest eigenvalue of $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$. Then the limit $h(\mathbf{v}, t) := \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$ exists, and $h(\mathbf{v}, t) = h(\mathbf{v}_1, t)$.*

Proof. We have $\phi(\alpha \mathbf{v}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) = \phi(\alpha \mathbf{v}_1 + \alpha(\mathbf{v} - \mathbf{v}_1), t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$.

Let $\mathbf{u}_\alpha = \alpha(\mathbf{v} - \mathbf{v}_1)$. Then \mathbf{u}_α satisfies:

- (i) \mathbf{u}_α is orthogonal to the eigenspace of μ_1 .
- (ii) There exists $c > 0$ such that for all $\alpha > 0$, $\|\mathbf{u}_\alpha\|_2 \leq c\alpha$.

Applying Theorem 3, we get $\lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) = \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$, so $h(\mathbf{v}, t) = h(\mathbf{v}_1, t)$. \square

Corollary 2. *Assume the assumptions of Theorem 3 hold. Let $h(\mathbf{v}, t) := \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$. Then for all $s > 0$, $h(s\mathbf{v}, t) = h(\mathbf{v}, t + \frac{1}{\mu_1} \log(s))$.*

Proof. According to the definition, we have $h(s\mathbf{v}, t) = \lim_{\alpha \rightarrow 0} \phi(\alpha s\mathbf{v}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$. Let $\alpha' = \alpha s$, then $h(s\mathbf{v}, t) = \lim_{\alpha' \rightarrow 0} \phi(\alpha' \mathbf{v}, t + \frac{1}{\mu_1} \log(s) + \frac{1}{\mu_1} \log \frac{1}{\alpha'}) = h(\mathbf{v}, t + \frac{1}{\mu_1} \log(s))$. \square

Corollary 3. *Assume the assumptions of Theorem 3 hold. Let $h(\mathbf{v}, t) := \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$. Let $T_{\mathbf{v}} := \{h(\mathbf{v}, t) : t \in \mathbb{R}\}$. Let \mathbf{v}_1 be the projection of \mathbf{v} into the eigenspace of the largest eigenvalue of $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$. Then $T_{\mathbf{v}} = T_{\mathbf{v}_1}$, and for all $c > 0$, $T_{c\mathbf{v}} = T_{\mathbf{v}}$. If $\mathbf{v} \neq \mathbf{0}$, then $T_{\mathbf{v}} = T_{\frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}}$.*

Proof. From Corollary 1, we have $h(\mathbf{v}, t) = h(\mathbf{v}_1, t)$, so $T_{\mathbf{v}} = T_{\mathbf{v}_1}$. From Corollary 2, we have for all $c > 0$, $h(c\mathbf{v}, t) = h(\mathbf{v}, t + \frac{1}{\mu_1} \log(c))$, so $T_{c\mathbf{v}} = T_{\mathbf{v}}$ for all $c > 0$. \square

Remark. *Corollary 3 implies that the trajectory of parameters is determined by $\frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}$.*

Proof of Theorem 1: Denote $f_i(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\mathbf{x}_i)$. Then $\ell(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n (f_i(\boldsymbol{\theta}) - y_i)^2$. The gradient of ℓ at $\boldsymbol{\theta}$ is given by $-\nabla \ell(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - f_i(\boldsymbol{\theta})) \nabla f_i(\boldsymbol{\theta})$. Because $\nabla f_i(\mathbf{0}) = \mathbf{0}$, we have $-\nabla \ell(\mathbf{0}) = \mathbf{0}$.

The Hessian of $-\ell$ at $\mathbf{0}$ is $-\nabla^2 \ell(\mathbf{0}) = \sum_{i=1}^n y_i \nabla^2 f_i(\mathbf{0})$, which is a block diagonal matrix with blocks D_i for $i = 1, \dots, m$.

$$-\nabla^2 \ell(\mathbf{0}) = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_m \end{pmatrix}$$

where D_i is given by

$$D_i = \begin{pmatrix} 0 & \boldsymbol{\gamma}^\top \\ \boldsymbol{\gamma} & \mathbf{0}_{d \times d} \end{pmatrix}, i = 1, 2, \dots, m$$

The maximum eigenvalue of D_i is $\|\boldsymbol{\gamma}\|_2$, and the corresponding eigenvector is $(\|\boldsymbol{\gamma}\|_2, \boldsymbol{\gamma})$. Thus, the maximum eigenvalue of $-\nabla^2 \ell(\mathbf{0})$ is $\|\boldsymbol{\gamma}\|_2$. Denote the eigenspace of $\|\boldsymbol{\gamma}\|_2$ by \mathbf{V} , then the projection of $\boldsymbol{\theta}^*$ onto \mathbf{V} is determined by \mathbf{C} . It is easy to verify that the assumptions of Theorem 3 hold, by defining $g(\boldsymbol{\theta}) := -\nabla \ell(\boldsymbol{\theta})$. Then, by Corollary 1, $h(\boldsymbol{\theta}, t) = \lim_{\alpha \rightarrow 0} \phi(\alpha \boldsymbol{\theta}, t + \frac{1}{\mu} \log \frac{1}{\alpha})$ exists, and $h(\boldsymbol{\theta}, t)$ as a function of $\boldsymbol{\theta}$ is determined by \mathbf{C} . By Corollary 3, $T_{\boldsymbol{\theta}}$ is determined by $\frac{\mathbf{C}}{\|\mathbf{C}\|_2}$. \square

Proposition 1. *Assume the assumptions of Theorem 3 hold. Let $h(\mathbf{v}_0, t) := \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v}_0, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$. Then $h(\mathbf{v}_0, t)$ as a function of t is differentiable, and $\frac{d}{dt} h(\mathbf{v}_0, t) = g(h(\mathbf{v}_0, t))$.*

Proof. Since $\frac{d}{dt}\phi(\alpha\mathbf{v}_0, t + \frac{1}{\mu}\log\frac{1}{\alpha}) = g\left(\phi(\alpha\mathbf{v}_0, t + \frac{1}{\mu}\log\frac{1}{\alpha})\right)$, we have

$$\phi(\alpha\mathbf{v}_0, t + \frac{1}{\mu}\log\frac{1}{\alpha}) = h(\mathbf{v}_0, 0) + \int_0^t g\left(\phi(\alpha\mathbf{v}_0, s + \frac{1}{\mu}\log\frac{1}{\alpha})\right) ds. \quad (8)$$

Consider $\phi(\alpha\mathbf{v}_0, t + \frac{1}{\mu}\log\frac{1}{\alpha})$ as a function of t and α , and extend the value of $\phi(\alpha\mathbf{v}_0, t + \frac{1}{\mu}\log\frac{1}{\alpha})$ at $\alpha = 0$ by $\lim_{\alpha \rightarrow 0}\phi(\alpha\mathbf{v}_0, t + \frac{1}{\mu}\log\frac{1}{\alpha})$. Then ϕ is continuous for $(t, \alpha) \in D = [0, t_0] \times [0, 1]$. Thus, there exists $\delta > 0$ such that $\phi(D) \subset B_\delta(\mathbf{0})$. Because $g(\boldsymbol{\theta})$ is continuously differentiable, it is Lipschitz continuous in $B_\delta(\mathbf{0})$.

In Equation 8, let $\alpha \rightarrow 0$. Since $\phi(\alpha\mathbf{v}_0, s + \frac{1}{\mu}\log\frac{1}{\alpha})$ uniformly converges to $h(\mathbf{v}_0, s)$ and g is Lipschitz continuous, we obtain

$$h(\mathbf{v}_0, t) = h(\mathbf{v}_0, 0) + \int_0^t g(h(\mathbf{v}_0, s)) ds.$$

Therefore, $h(\mathbf{v}_0, t)$ is differentiable over t , and $\frac{d}{dt}h(\mathbf{v}_0, t) = g(h(\mathbf{v}_0, t))$. \square

A.3 Proof of Theorem 2

Next we begin by establishing Theorem 4, and then leverage it to validate Theorem 2.

Consider the gradient flow of the loss function ℓ :

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= -\nabla\ell(\boldsymbol{\theta}(t)), \\ \boldsymbol{\theta}(0) &= \boldsymbol{\theta}_0, \end{aligned} \quad (9)$$

and denote the solution of 9 as $\phi(\boldsymbol{\theta}_0, t)$.

Theorem 4. *Assume the following conditions:*

- (i) $\ell(\boldsymbol{\theta})$ is an analytic and nonnegative function.
- (ii) $\mathbf{0}$ is a strict saddle point of $\ell(\boldsymbol{\theta})$.

Denote \mathbf{V}_1 to be the eigenspace corresponding to the largest eigenvalue of $-\nabla^2\ell(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$. Let \mathbf{v}_1 be a vector in \mathbf{V}_1 . The vector \mathbf{u}_α is arbitrary, subject to the conditions:

- (i) \mathbf{u}_α is orthogonal to \mathbf{V}_1 .
- (ii) There exists $c > 0$ such that for all $\alpha > 0$, $\|\mathbf{u}_\alpha\|_2 \leq c\alpha$.

Denote the solution of 9 as $\phi(\boldsymbol{\theta}_0, t)$. Then $h(\mathbf{v}, t) = \lim_{\alpha \rightarrow 0}\phi(\alpha\mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1}\log\frac{1}{\alpha})$ exists. Given \mathbf{v}_0 , if $h(\mathbf{v}_0, t)$ is bounded for all $t \geq 0$, then the limit $\lim_{t \rightarrow \infty} h(\mathbf{v}_0, t)$ exists. Furthermore, if this limit is not a saddle point of $\ell(\boldsymbol{\theta})$, then there exists a neighborhood $B_\delta(\mathbf{v}_0)$ of \mathbf{v}_0 , and an $\alpha_0 > 0$, such that the limit $\lim_{t \rightarrow \infty}\phi(\alpha\mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1}\log\frac{1}{\alpha})$ exists for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$ and $0 < \alpha < \alpha_0$, and

$$\lim_{t \rightarrow \infty} h(\mathbf{v}_0, t) = \lim_{\alpha \rightarrow 0} \lim_{t \rightarrow \infty} \phi(\alpha\mathbf{v}_0 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1}\log\frac{1}{\alpha}).$$

Moreover, the limit $\lim_{t \rightarrow \infty} h(\mathbf{v}, t)$ is a continuous function of \mathbf{v} at \mathbf{v}_0 .

Proof. We attempt to apply Theorem 3 by setting $g(\boldsymbol{\theta}) = -\nabla\ell(\boldsymbol{\theta})$ and checking the conditions for $g(\boldsymbol{\theta})$:

- Since $\ell(\boldsymbol{\theta})$ is analytic, $g(\boldsymbol{\theta})$ is twice differentiable.
- As $\mathbf{0}$ is a strict saddle point of $\ell(\boldsymbol{\theta})$, we have $g(\mathbf{0}) = \mathbf{0}$.
- The matrix $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}} = -\nabla^2\ell(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$ is symmetric and thus diagonalizable.
- Because $\mathbf{0}$ is a strict saddle point, the largest eigenvalue of $\nabla g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}}$ is positive.

Therefore, $g(\boldsymbol{\theta})$ satisfies the conditions of Theorem 3. Applying Theorem 3, we conclude that $h(\mathbf{v}, t) = \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$ exists.

We then proceed in two steps to prove Theorem 4.

(i) First, we prove the existence of $\lim_{t \rightarrow \infty} h(\mathbf{v}_0, t)$ under the condition that $h(\mathbf{v}_0, t)$ is bounded for $t \geq 0$.

Let $\boldsymbol{\theta}(t) = h(\mathbf{v}_0, t)$. By Proposition 1, $\boldsymbol{\theta}'(t) = -\nabla \ell(\boldsymbol{\theta}(t))$. Because $\boldsymbol{\theta}(t)$ is bounded for all $t \geq 0$, there exists a sequence $\{t_n\}$ such that $\lim_{n \rightarrow \infty} t_n = +\infty$, and $\hat{\boldsymbol{\theta}} = \lim_{n \rightarrow \infty} \boldsymbol{\theta}(t_n)$ exists. Since

$$\frac{d}{dt} \ell(\boldsymbol{\theta}(t)) = \langle -\nabla \ell(\boldsymbol{\theta}(t)), \boldsymbol{\theta}'(t) \rangle = -\|\nabla \ell(\boldsymbol{\theta}(t))\|_2^2,$$

$\ell(\boldsymbol{\theta}(t))$ is monotonically decreasing over t , and $\ell(\boldsymbol{\theta}(t)) \geq \ell(\hat{\boldsymbol{\theta}})$ for all $t \geq 0$. Furthermore, $\int_0^{+\infty} \|\nabla \ell(\boldsymbol{\theta}(t))\|_2^2 dt \leq \ell(\boldsymbol{\theta}(0))$. Therefore, $\lim_{t \rightarrow \infty} \|\nabla \ell(\boldsymbol{\theta}(t))\|_2 = 0$, and $\nabla \ell(\hat{\boldsymbol{\theta}}) = \lim_{t \rightarrow \infty} \nabla \ell(\boldsymbol{\theta}(t_n)) = \mathbf{0}$.

By Łojasiewicz's inequality Łojasiewicz (1965), there exists $C > 0$ and $0 < \mu < 1$, and a neighborhood $B_\delta(\hat{\boldsymbol{\theta}})$ such that

$$\|\nabla \ell(\boldsymbol{\theta})\|_2 \geq C|\ell(\boldsymbol{\theta}) - \ell(\hat{\boldsymbol{\theta}})|^\mu, \quad \forall \boldsymbol{\theta} \in B_\delta(\hat{\boldsymbol{\theta}}).$$

Since $\ell(\boldsymbol{\theta}(t)) \geq \ell(\hat{\boldsymbol{\theta}})$, it follows that

$$\|\nabla \ell(\boldsymbol{\theta}(t))\|_2 \geq C(\ell(\boldsymbol{\theta}(t)) - \ell(\hat{\boldsymbol{\theta}}))^\mu, \quad \forall \boldsymbol{\theta} \in B_\delta(\hat{\boldsymbol{\theta}}).$$

Given that $\lim_{t \rightarrow \infty} \boldsymbol{\theta}(t_n) = \hat{\boldsymbol{\theta}}$, there exists an n such that $\|\boldsymbol{\theta}(t_n) - \hat{\boldsymbol{\theta}}\|_2 < \frac{\delta}{2}$ and $|\ell(\boldsymbol{\theta}(t_n)) - \ell(\hat{\boldsymbol{\theta}})| < \frac{1}{2}C(1 - \mu)\delta^{\frac{1}{1-\mu}}$. Because

$$\frac{d}{dt} \ell(\boldsymbol{\theta}(t)) = -\|\nabla \ell(\boldsymbol{\theta}(t))\|_2^2 = -\|\nabla \ell(\boldsymbol{\theta}(t))\|_2 \times \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2 \leq -C(\ell(\boldsymbol{\theta}(t)) - \ell(\hat{\boldsymbol{\theta}}))^\mu \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2,$$

we have

$$\frac{d}{dt} (\ell(\boldsymbol{\theta}(t)) - \ell(\hat{\boldsymbol{\theta}}))^{1-\mu} \leq -(1 - \mu)C \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2.$$

Thus,

$$\int_{t_n}^t \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2 dt \leq \frac{1}{C(1 - \mu)} (\ell(\boldsymbol{\theta}(t_n)) - \ell(\hat{\boldsymbol{\theta}}))^{1-\mu}.$$

Since $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t_n)\|_2 \leq \int_{t_n}^t \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2 dt$, it follows that

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t_n)\|_2 \leq \frac{1}{C(1 - \mu)} (\ell(\boldsymbol{\theta}(t_n)) - \ell(\hat{\boldsymbol{\theta}}))^{1-\mu} < \frac{\delta}{2}.$$

Therefore,

$$\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}\|_2 \leq \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t_n)\|_2 + \|\boldsymbol{\theta}(t_n) - \hat{\boldsymbol{\theta}}\|_2 < \delta,$$

so for all $t \geq t_n$, $\boldsymbol{\theta}(t) \in B_\delta(\hat{\boldsymbol{\theta}})$. Thus, we can apply Łojasiewicz's inequality for all $t \geq t_n$. Consequently,

$$\int_{t_n}^t \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2 dt \leq \frac{1}{C(1 - \mu)} (\ell(\boldsymbol{\theta}(t_n)) - \ell(\hat{\boldsymbol{\theta}}))^{1-\mu}, \quad \forall t \geq t_n.$$

Thus, the length of the trajectory of $\boldsymbol{\theta}(t)$ is finite, which implies that $\lim_{t \rightarrow \infty} \boldsymbol{\theta}(t)$ exists.

Lemma 1. Denote the solution of Equation 9 as $\phi(\boldsymbol{\theta}_0, t)$. Assume the following:

- (i) $\ell(\boldsymbol{\theta})$ is an analytic and nonnegative function.
- (ii) $\bar{\boldsymbol{\theta}} = \lim_{t \rightarrow \infty} \phi(\mathbf{v}_0, t)$ exists.
- (iii) $\bar{\boldsymbol{\theta}}$ is a local minimum of $\ell(\boldsymbol{\theta})$.

Then there exists a neighborhood of \mathbf{v}_0 , denoted $B_\delta(\mathbf{v}_0)$, such that for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$, the limit $\lim_{t \rightarrow \infty} \phi(\mathbf{v}, t)$ exists and is continuous at \mathbf{v}_0 .

Proof. By Łojasiewicz's inequality, there exists $C > 0$ and $0 < \mu < 1$, and a neighborhood $B_{\epsilon_0}(\bar{\boldsymbol{\theta}})$ such that

$$\|\nabla \ell(\boldsymbol{\theta})\|_2 \geq C|\ell(\boldsymbol{\theta}) - \ell(\bar{\boldsymbol{\theta}})|^\mu, \quad \forall \boldsymbol{\theta} \in B_{\epsilon_0}(\bar{\boldsymbol{\theta}}).$$

Since $\bar{\boldsymbol{\theta}} = \lim_{t \rightarrow \infty} \phi(\mathbf{v}_0, t)$, for all $\epsilon \in (0, \epsilon_0)$, there exists $t_0 > 0$ such that $\|\phi(\mathbf{v}_0, t_0) - \bar{\boldsymbol{\theta}}\|_2 < \frac{\epsilon}{4}$ and $|\ell(\phi(\mathbf{v}_0, t_0)) - \ell(\bar{\boldsymbol{\theta}})| < \frac{1}{4}C(1 - \mu)\epsilon^{\frac{1}{1-\mu}}$.

Because $\phi(\boldsymbol{\theta}, t_0)$ is locally Lipschitz continuous over $\boldsymbol{\theta}$, there exists $L > 0$ and $\delta > 0$ such that $\|\phi(\mathbf{v}, t_0) - \phi(\mathbf{v}_0, t_0)\|_2 < \frac{\epsilon}{4}$ for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$ and $|\ell(\phi(\mathbf{v}, t_0)) - \ell(\phi(\mathbf{v}_0, t_0))| < \frac{1}{4}C(1 - \mu)\epsilon^{\frac{1}{1-\mu}}$ for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$.

Thus, for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$, we have $\|\phi(\mathbf{v}, t_0) - \bar{\boldsymbol{\theta}}\|_2 < \|\phi(\mathbf{v}, t_0) - \phi(\mathbf{v}_0, t_0)\|_2 + \|\phi(\mathbf{v}_0, t_0) - \bar{\boldsymbol{\theta}}\|_2 < \frac{\epsilon}{2}$ and $|\ell(\phi(\mathbf{v}, t_0)) - \ell(\bar{\boldsymbol{\theta}})| \leq |\ell(\phi(\mathbf{v}, t_0)) - \ell(\phi(\mathbf{v}_0, t_0))| + |\ell(\phi(\mathbf{v}_0, t_0)) - \ell(\bar{\boldsymbol{\theta}})| < \frac{1}{2}C(1 - \mu)\epsilon^{\frac{1}{1-\mu}}$ for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$.

Applying Łojasiewicz's inequality as in Theorem 3, we get

$$\int_{t_0}^t \left\| \frac{\partial \phi(\mathbf{v}, \tau)}{\partial \tau} \right\|_2 d\tau \leq \frac{1}{C(1 - \mu)} (\ell(\phi(\mathbf{v}, t_0)) - \ell(\bar{\boldsymbol{\theta}}))^{1-\mu}. \quad (10)$$

and

$$\|\phi(\mathbf{v}, t) - \phi(\mathbf{v}, t_0)\|_2 \leq \frac{1}{C(1 - \mu)} (\ell(\phi(\mathbf{v}, t_0)) - \ell(\bar{\boldsymbol{\theta}}))^{1-\mu} < \frac{\epsilon}{2}. \quad (11)$$

Therefore, $\|\phi(\mathbf{v}, t) - \bar{\boldsymbol{\theta}}\|_2 \leq \|\phi(\mathbf{v}, t) - \phi(\mathbf{v}, t_0)\|_2 + \|\phi(\mathbf{v}, t_0) - \phi(\mathbf{v}_0, t_0)\|_2 + \|\phi(\mathbf{v}_0, t_0) - \bar{\boldsymbol{\theta}}\|_2 < \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon$.

From Equation 10, we know that for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$, the length of the trajectory of $\phi(\mathbf{v}, t)$ is finite, so $\lim_{t \rightarrow \infty} \phi(\mathbf{v}, t)$ exists.

From Equation 11, we know that for all $\epsilon \in (0, \epsilon_0)$, there exists $\delta > 0$ and $t_0 > 0$ such that for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$ and for all $t > t_0$, we have $\|\phi(\mathbf{v}, t) - \bar{\boldsymbol{\theta}}\|_2 < \epsilon$. Letting $t \rightarrow \infty$, we conclude: for all $\epsilon \in (0, \epsilon_0)$, there exists $\delta > 0$ such that for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$, $\|\lim_{t \rightarrow \infty} \phi(\mathbf{v}, t) - \bar{\boldsymbol{\theta}}\|_2 < \epsilon$. Therefore, $\lim_{t \rightarrow \infty} \phi(\mathbf{v}, t)$ is continuous at \mathbf{v}_0 . \square

Now we are back to proof of second part of Theorem 4. We assume that $\lim_{t \rightarrow \infty} h(\mathbf{v}_0, t)$ exists and is a local minimum of $\ell(\boldsymbol{\theta})$. We want to prove the following:

1. There exists a neighborhood $B_\delta(\mathbf{v}_0)$ of \mathbf{v}_0 , and an $\alpha_0 > 0$, such that the limit $\lim_{t \rightarrow \infty} \phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$ exists for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$ and $0 < \alpha < \alpha_0$.
2. $\lim_{t \rightarrow \infty} h(\mathbf{v}_0, t) = \lim_{\alpha \rightarrow 0} \lim_{t \rightarrow \infty} \phi(\alpha \mathbf{v}_0 + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$.
3. The limit $\lim_{t \rightarrow \infty} h(\mathbf{v}, t)$ is a continuous function of \mathbf{v} at \mathbf{v}_0 .

Denote $q(\mathbf{v}, \alpha) := \phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, \frac{1}{\mu_1} \log \frac{1}{\alpha})$. By Theorem 3, $\lim_{\alpha \rightarrow 0} q(\mathbf{v}, \alpha)$ exists and is continuous over \mathbf{v} .

We have

$$\phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) = \phi(\phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, \frac{1}{\mu_1} \log \frac{1}{\alpha}), t) = \phi(q(\mathbf{v}, \alpha), t).$$

And

$$h(\mathbf{v}, t) = \lim_{\alpha \rightarrow 0} \phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) = \phi(\lim_{\alpha \rightarrow 0} q(\mathbf{v}, \alpha), t).$$

Denote $q(\mathbf{v}, 0) := \lim_{\alpha \rightarrow 0} q(\mathbf{v}, \alpha)$. By setting \mathbf{v}_0 in Lemma 1 equal to $q(\mathbf{v}_0, 0)$, we get: there exists a neighborhood of $q(\mathbf{v}_0, 0)$, denoted $B_\delta(q(\mathbf{v}_0, 0))$, such that for all $\mathbf{v} \in B_\delta(q(\mathbf{v}_0, 0))$, the limit $\lim_{t \rightarrow \infty} \phi(\mathbf{v}, t)$ exists and is continuous at $q(\mathbf{v}_0, 0)$.

Because $q(\mathbf{v}, \alpha)$ is continuous at $(\mathbf{v}_0, 0)$, the pre-image $q^{-1}(B_\delta(q(\mathbf{v}_0, 0)))$ is open. Thus, there exists a neighborhood $B_\delta(\mathbf{v}_0)$ of \mathbf{v}_0 , and an $\alpha_0 > 0$, such that for all $\mathbf{v} \in B_\delta(\mathbf{v}_0)$ and $0 < \alpha < \alpha_0$, the limit $\lim_{t \rightarrow \infty} \phi(\alpha \mathbf{v} + \mathbf{u}_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$ exists.

Since $q(\mathbf{v}, \alpha)$ is continuous at $(\mathbf{v}_0, 0)$, and $\lim_{t \rightarrow \infty} \phi(\mathbf{v}, t)$ is continuous at $q(\mathbf{v}_0, 0)$, it follows that $\lim_{t \rightarrow \infty} \phi(q(\mathbf{v}, \alpha), t)$ is continuous at $(\mathbf{v}_0, 0)$. Therefore, we have

$$\lim_{\alpha \rightarrow 0} \lim_{t \rightarrow \infty} \phi(q(\mathbf{v}_0, \alpha), t) = \lim_{t \rightarrow \infty} \phi(q(\mathbf{v}_0, 0), t) = \lim_{t \rightarrow \infty} h(\mathbf{v}_0, t),$$

and $\lim_{t \rightarrow \infty} h(\mathbf{v}, t)$ is continuous at \mathbf{v}_0 . \square

Proof of Theorem 2. Under the notations and assumptions of Theorem 2, we have

$$-\nabla^2 \ell(\mathbf{0}) = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_m \end{pmatrix},$$

where each D_i is defined as

$$D_i = \begin{pmatrix} 0 & \boldsymbol{\gamma}^\top \\ \boldsymbol{\gamma} & \mathbf{0}_{d \times d} \end{pmatrix},$$

for $i = 1, 2, \dots, m$.

The maximum eigenvalue of D_i is $\|\boldsymbol{\gamma}\|_2$. Therefore, the maximum eigenvalue of $-\nabla^2 \ell(\mathbf{0})$ is $\|\boldsymbol{\gamma}\|_2 > 0$, indicating that $\mathbf{0}$ is a strict saddle point of $\ell(\boldsymbol{\theta})$. Given that $\ell(\boldsymbol{\theta})$ is an analytic and nonnegative function, and $\mathbf{0}$ is a strict saddle point of $\ell(\boldsymbol{\theta})$, the conditions of Theorem 4 are satisfied. Let us denote the eigenspace corresponding to $\|\boldsymbol{\gamma}\|_2$ by \mathbf{V} . Then the projection of $\boldsymbol{\theta}$ onto \mathbf{V} is determined by \mathbf{C} . We now proceed to verify the assertions of Theorem 2 point by point.

(i) By applying Theorem 4, we confirm that the limit

$$h(\boldsymbol{\theta}, t) = \lim_{\alpha \rightarrow 0} \phi(\alpha \boldsymbol{\theta}, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \frac{1}{\alpha})$$

exists.

(ii) When the set $\{h(\boldsymbol{\theta}, t) : t \geq 0\}$ is bounded, Theorem 4 ensures that the limit $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t)$ exists. Let $\hat{\boldsymbol{\theta}}$ be the projection of $\boldsymbol{\theta}$ onto \mathbf{V} . Since $h(\boldsymbol{\theta}, t) = h(\hat{\boldsymbol{\theta}}, t)$ and for any $\alpha > 0$, $h(\alpha \boldsymbol{\theta}, t) = h(\boldsymbol{\theta}, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \alpha)$, it follows that $h(\boldsymbol{\theta}, t) = h\left(\frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|_2}, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \frac{1}{\|\hat{\boldsymbol{\theta}}\|_2}\right)$. Given the existence of $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t)$, we have

$$\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t) = \lim_{t \rightarrow \infty} h\left(\frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|_2}, t\right).$$

As $\frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|_2}$ is determined by $\frac{\mathbf{C}}{\|\mathbf{C}\|_2}$, the limit $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t)$ is determined by $\frac{\mathbf{C}}{\|\mathbf{C}\|_2}$.

(iii) Applying Theorem 4 again, we deduce that if $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_0, t)$ exists and is not a saddle point of $\ell(\boldsymbol{\theta})$, then

$$\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}_0, t) = \lim_{\alpha \rightarrow 0} \lim_{t \rightarrow \infty} \phi\left(\alpha \boldsymbol{\theta}_0, t + \frac{1}{\|\boldsymbol{\gamma}\|_2} \log \frac{1}{\alpha}\right).$$

Furthermore, the limit $\lim_{t \rightarrow \infty} h(\boldsymbol{\theta}, t)$ is continuous at $\boldsymbol{\theta}_0$. \square

B Experiment of higher dimensional input

Consider a neuron network with high dimensional input $f_{\theta}(\mathbf{x}) = \sum_{i=1}^m a_i \sigma(\mathbf{w}_i^{\top} \mathbf{x})$. Theorem 1 shows that under small initialization, the trajectory is determined by $\frac{C(\theta)}{\|C(\theta)\|_2}$. Because $C(\theta)$ is an m dimensional vector, so the trajectory of θ under all initialization is restricted to a m dimensional manifold. Note the dimension of manifold of θ is independent of dimension of input. So even when input is very high dimensional, the trace of parameters stays in a low dimensional.

In Figure 7, we train a neural network for $m = 2$ and $d = 500$. The sample size is 500. The initialization of parameters is a Gaussian distribution centered at $\mathbf{0}$, with its standard in figure’s legend. By scaling the second neuron, we manage to make C_1/C_2 to be a constant in each subfigure. In this figure, the loss curve and parameters curve matches perfectly among different trials. The results show that, there existing a limiting trace of parameters as initialization scale approaches 0. Moreover, the trace of parameters is determined by C_1/C_2 .

In Figure 8, we conduct experiments for $m = 2$ and $d = 3$. $f_{\theta}(\mathbf{x}) = a_1 \tanh(\mathbf{w}_1^{\top} \mathbf{x} + b_1) + a_2 \tanh(\mathbf{w}_2^{\top} \mathbf{x} + b_2)$, with $x \in \mathbb{R}^3$. The target function is $f^*(\mathbf{x}) = \tanh(\mathbf{1}^T \mathbf{x} + 1)$. The training data $(\mathbf{x}_i, y_i)_{i=1}^n$ is obtained by $y_i = f^*(\mathbf{x}_i)$ and $\{\mathbf{x}_i\}_{i=1}^n$ draw independently from uniform distribution on interval $[-2, 2]^d$. The sample size $n = 10$. It is seen that in whatever initialization, the parameters all converge to Q^* , thus the generalization error is zero. Besides recovery, there is a clear sign that network initialized with \tilde{c} closer to 1 will converge to Q^1 and network initialized with \tilde{c} closer to 0 will converge to Q^2 . This phenomenon is in accordance to what we observed in one-dimensional experiment.

In Figure 9, we plot generalization error in high dimensional case. The network model is $f_{\theta}(\mathbf{x}) = a_1 \tanh(\mathbf{w}_1^{\top} \mathbf{x} + b_1) + a_2 \tanh(\mathbf{w}_2^{\top} \mathbf{x} + b_2)$, with $x \in \mathbb{R}^3$. The target function is $f^*(\mathbf{x}) = \tanh(\mathbf{1}^T \mathbf{x} + 1)$. The setting of experiment is $m = 2$ and $d = 3$. In this case, optimistic sample size is 5, the separation sample size of Q^2 is 6, the separation sample size of Q^1 is 9. It is seen that when the sample size is larger than optimistic sample size, the network can recover at $\tilde{c} = 1$ or $\tilde{c} = 0$. When sample size is above separation sample size of Q^1 , all initialization can recover target function. At separation sample size of Q^2 , which is $n = 6$, there is a small probability of initialization cannot recover target function. When sample size is above 6, 7, 8, the probability of recovery increases.

In Figure 10, we show that why when $n = 6, 7, 8$, recovery fails to happen at $\tilde{c} = 0$. Note that when sample size reaches the separation size of Q^2 , it only guarantees that separation of Q^2 is almost every. That is to say, it allows that a subset with zero-measure with respect to Q^2 is not separated. Unfortunately, the origin of Q^2 belongs to the set. There exists other global minimum rather than Q^2 around origin of Q^2 . Besides, $\tilde{c} = 0$ will lead to converge to origin of Q^2 . So it is possible for \tilde{c} near 0 to converge to these imperfect global minimum, thus failing in recovery. Overall, our understanding is that, both separation sample size and optimistic sample size is a lower bound for recovery. The recovery need not to happen at these sample size, but under these sample size, the recovery could not happen.

C Experimental Details

In Figure 1(a) to 1(e), the learning rate was set to 0.5. Points for calculating generalization error were 1000 points evenly from the interval $[-2, 2]$. For Figures 1(d) and 1(e), due to nonlinear convergence of seeds, it is difficult to get a extremely low training loss. So we train the network until loss is 10^{-8} . For For Figures 1(a) to 1(e), we train the network until training loss is 10^{-15} . In Figure 1(f), the generalization error is computed by 1000 points $(x_i, y_i)_{i=1}^{1000}$ with $y_i = f^*(x_i)$ and $\{x_i\}_{i=1}^{1000}$ following i.i.d standard Gaussian distribution. For $n = 2, 3$, the iterations is 10^6 . For $n = 4, 5$, the iterations is 4×10^5 . The training was halted once the loss reached 10^{-15} .

In Figure 2, training was performed using gradient descent with a learning rate of 0.01. Suppose θ_1 and θ_2 to be the initialization of the first and the second neuron, respectively. We transform θ_2 into $a\theta_2$, and choose appropriate value of a to keep $c = 0.5$ across all trials.

In Figure 3, Training was conducted using gradient descent with a learning rate of 0.05 and iterations of 10^6 . The dataset $(x_i, y_i)_{i=1}^n$ consists of 6 points, with $y_i = f^*(x_i)$ and $\{x_i\}_{i=1}^6$ equally spaced points on the interval $[-2, 2]$. We use seeds 0 to 400 to generate initialization of parameters.

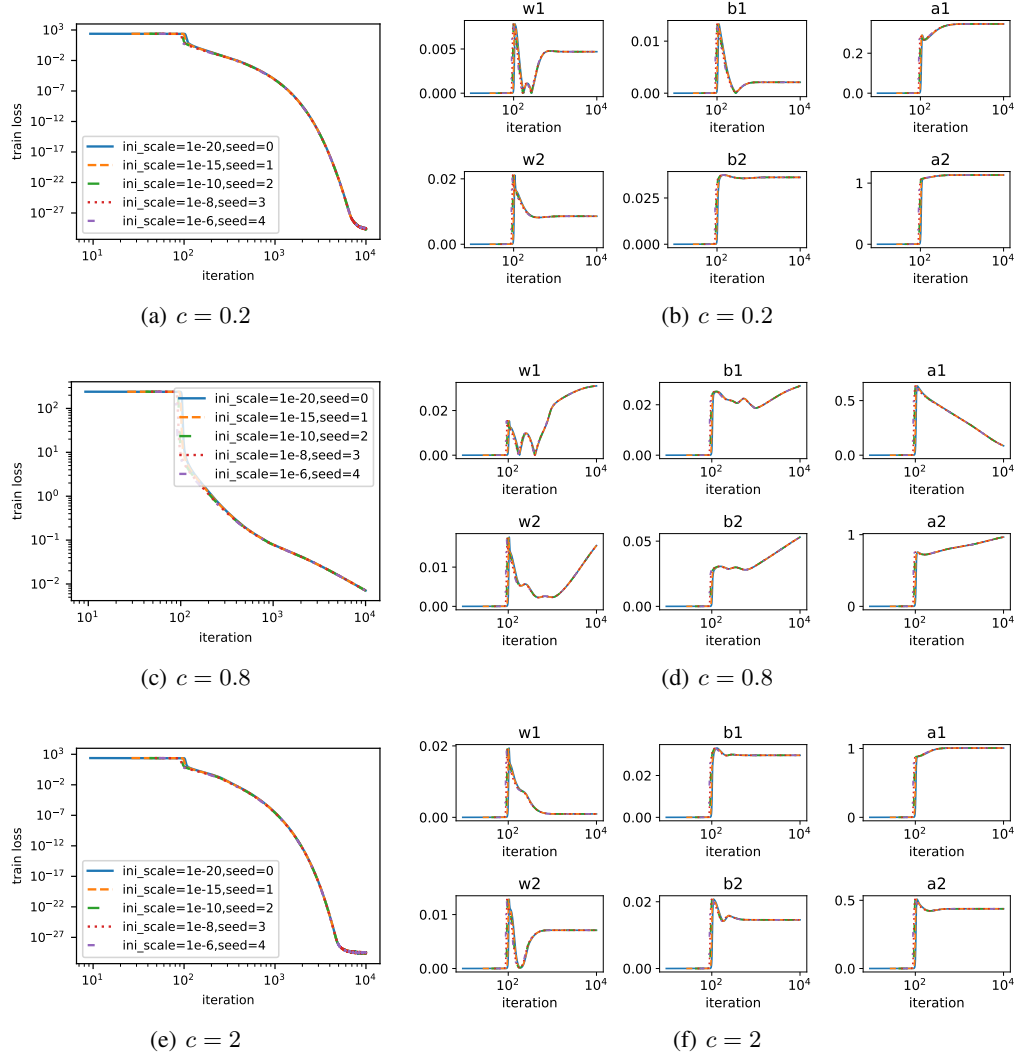


Figure 7: Training a two-layer neural network with varying widths. The network model is $f_{\theta}(\mathbf{x}) = a_1 \tanh(\mathbf{w}_1^T \mathbf{x} + b_1) + a_2 \tanh(\mathbf{w}_2^T \mathbf{x} + b_2)$, with $\mathbf{x} \in \mathbb{R}^3$. The target function is $f^*(\mathbf{x}) = \tanh(\mathbf{1}^T \mathbf{x} + 1)$. The training data $(\mathbf{x}_i, y_i)_{i=1}^n$ is obtained by $y_i = f^*(\mathbf{x}_i)$ and $\{\mathbf{x}_i\}_{i=1}^n$ draw independently from uniform distribution on interval $[-2, 2]^d$ with random seed 0. In this experiments, $m = 2, d = 500, n = 500$. The iterations is 10^4 , the learning rate is 0.001. Different initialization use different seeds to generate random number for Gaussian distribution of parameters. But we keep $c := C_1/C_2$ to be a constant for all initialization in a subfigure. The value of c is in the caption of subfigure. In the right part, the parameters w_1 and w_2 is the first coordinate of \mathbf{w}_1 and \mathbf{w}_2 .

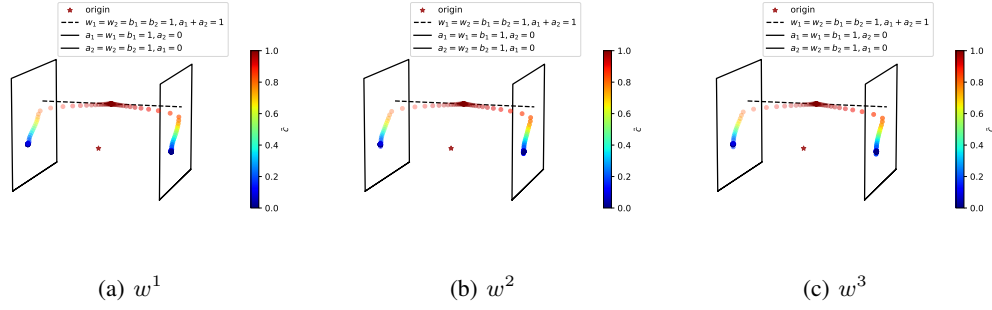


Figure 8: Training a two-layer neural network. The network model is $f_{\theta}(\mathbf{x}) = a_1 \tanh(\mathbf{w}_1^{\top} \mathbf{x} + b_1) + a_2 \tanh(\mathbf{w}_2^{\top} \mathbf{x} + b_2)$, with $x \in \mathbb{R}^3$. The target function is $f^*(\mathbf{x}) = \tanh(\mathbf{1}^T \mathbf{x} + 1)$. The training data $(\mathbf{x}_i, y_i)_{i=1}^n$ is obtained by $y_i = f^*(\mathbf{x}_i)$ and $\{\mathbf{x}_i\}_{i=1}^n$ draw independently from uniform distribution on interval $[-2, 2]^d$. In this experiments, $m = 2, d = 3, n = 10$. The iterations is 10^6 , the learning rate is 0.5. The initialization scale is 10^{-20} . Caption w^i represents the i -th dimension of \mathbf{w} .

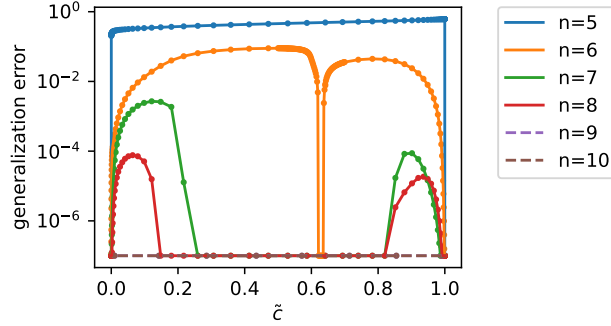


Figure 9: Training a two-layer neural network. The network model is $f_{\theta}(\mathbf{x}) = a_1 \tanh(\mathbf{w}_1^{\top} \mathbf{x} + b_1) + a_2 \tanh(\mathbf{w}_2^{\top} \mathbf{x} + b_2)$, with $x \in \mathbb{R}^3$. The target function is $f^*(\mathbf{x}) = \tanh(\mathbf{1}^T \mathbf{x} + 1)$. The training data $(\mathbf{x}_i, y_i)_{i=1}^n$ is obtained by $y_i = f^*(\mathbf{x}_i)$ and $\{\mathbf{x}_i\}_{i=1}^n$ draw independently from uniform distribution on interval $[-2, 2]^d$. In this experiments, $m = 2, d = 3$. The iterations is 10^6 for $n = 5, 6, 7, 8$. The iterations is 2×10^7 for $n = 9, 10$. The learning rate is 0.25. The initialization scale is 10^{-20} . The generalization error is obtained by assessing 1000 points drawn independently from uniform distribution on interval $[-2, 2]^d$. We identify generalization error lower than 10^{-7} to be 10^{-7} , and regard it as successful recovery of the target function.

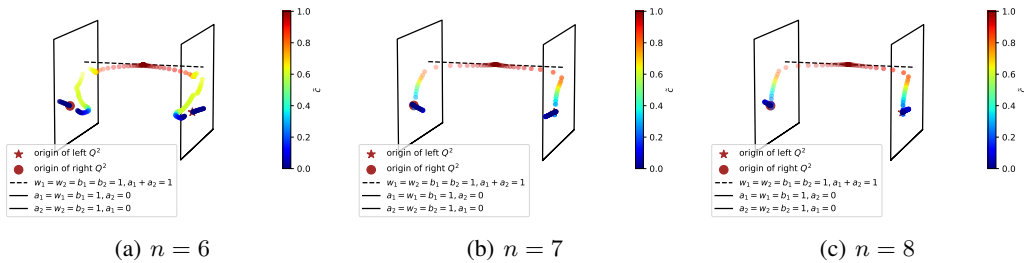


Figure 10: The experiment is same with 9. In this figure, the w_1 is first coordinate of \mathbf{w}_1 , w_2 is the first coordinate of \mathbf{w}_2 .

In Figure 4, gradient descent was employed as the training algorithm with a learning rate of 0.5. For $n = 2, 3, 4$, in all experiments training loss reaches 10^{-15} and then training stops. For $n = 5, 6$, the network is trained with iterations 10^7 . The train loss is shown in Figure 11.

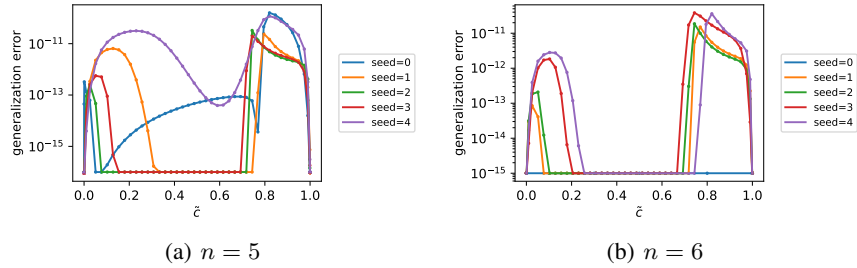


Figure 11: For $n = 5$, all networks are trained with learning rate 0.5 and iterations 10^7 . The training stops when loss reaches 10^{-15} . For $n = 6$, networks with seed= 1 to seed= 4 are trained with learning rate 0.5 and iterations 2×10^6 . The training stops when loss reaches 10^{-15} . Networks with seed= 0 are trained until loss reaches 10^{-15} .

In Figure 5, gradient descent with a learning rate of 0.01 is used for training. The iterations is 10^5 . Suppose θ_1 and θ_2 to be the initialization of the first and the second neuron, respectively. We transform θ_2 into $a\theta_2$, and choose appropriate value of a to keep $c = 0.5$ across all trials.

In Figure 6, gradient descent with a learning rate of 0.01 is used, and the initial weights are drawn from a Gaussian distribution with a mean of 0 and a standard deviation of 10^{-20} . In all experiments, train loss reaches 10^{-15} . For all width, we use 0 as random seed to generate Gaussian distribution for the initialization of parameters.

The details of experiments of Figure 7 to Figure 10 are in their captions.