Over-Parameterized Deep Neural Networks Have No Strict Local Minima For Any Continuous Activations

Dawei Li * Tian Ding† Ruoyu Sun ‡

Nov 22, 2018

Abstract

In this paper, we study the loss surface of the over-parameterized fully connected deep neural networks. We prove that for any continuous activation functions, the loss function has no bad strict local minimum, both in the regular sense and in the sense of sets. This result holds for any convex and differentiable loss function, and the data samples are only required to be distinct in at least one dimension. Furthermore, we show that bad local minima do exist for a class of activation functions, so without further assumptions it is impossible to prove every local minimum is a global minimum.

1 Introduction

Recently, the application of deep neural networks [1] has led to a phenomenal success in various artificial intelligence areas, e.g., computer vision, natural language processing, and audio recognition. However, the theoretical understanding of neural networks is still limited. One of the main difficulties of analyzing neural networks is the non-convexity of the objective function, which may cause many local minima.

In practice, it is observed that when the number of parameters is sufficiently large, common optimization algorithms such as stochastic gradient descent (SGD) can achieve small training error [2–5]. These observations are often explained by the intuition that more parameters can smooth the landscape [4, 6]. Among various definitions of over-parameterization, a popular one is that the last hidden layer has more neurons than the number of training samples. Even under this assumption, it is yet unclear to what extent we can prove a rigorous result. For instance, can we prove that for any neuron activation function, every local minimum is a global minimum? If not, what exactly can we prove, and what can we not prove?

1.1 Main Contributions

In this paper, we study the multi-layer feed-forward neural networks where the number of neurons in the last hidden layer is no less than the number of data samples. The loss function can be any convex and differentiable function, and the data samples are only required to be distinct in at least

*Coordinated Science Laboratory, Department of ISE, University of Illinois at Urbana-Champaign, Urbana, IL. dawei2@illinois.edu.

†Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. dt016@ie.cuhk.edu.hk. The work is done while the author is visiting Department of ISE, Univeristy of Illinois at Urbana-Champaign.

‡Coordinated Science Laboratory, Department of ISE, University of Illinois at Urbana-Champaign, Urbana, IL. ruoyus@illinois.edu.

one dimension. The activation functions can be any continuous functions, which covers a wide range of practically used activation functions such as ReLU, leaky ReLU, sigmoid, etc.

Our main result is that for any fully connected deep neural networks and any continuous activation functions, the empirical loss is a weakly global function [7]. Weakly global functions are a class of continuous functions that admit no set-wise strict bad local minima, as illustrated in in Figure 1. This implies that the loss surface is well-behaved in two-fold. First, there is no strict bad local minimum, and therefore any suboptimal local minimum can only lie in a plateau. Second, any sub-optimal plateau cannot be the bottom of a basin on the loss surface. In other words, "truly bad" local minima that are surrounded by barrier do not exist.

One natural question is whether over-parametrization can eliminate all bad local minima, not just strict bad local minima. Unfortunately, We provide examples to show that non-strict bad local minima exist for a large class of activation functions. Therefore, without further assumptions such as restricting to a smaller class of activation functions, it is impossible to prove every local minimum is a global minimum.

The analytical framework in this paper is sketched as follows. First, we establish the result for a specific class of analytic activation functions, which constitute a dense set in the space of continuous functions. That is, we can use a sequence of activation functions in the considered class to uniformly approximate any continuous function. Based on this approximation and a recent theoretical result of [7], we manage to extend our result to all continuous activation functions.

1.2 Related Works and Discussions

The loss surface of single-hidden-layer neural networks has been extensively studied in recent years [8–24]. These works provide sufficient conditions under which local search algorithms will converge to the global optimum of the empirical loss. It can be roughly divided into two categories: non-global landscape analysis and global landscape analysis. For the first category, the result do not apply to all local minima. One typical conclusion is about the local geometry, i.e., in a small neighborhood of the global minima no bad local minima exist [20–22]. Another typical conclusion is that a subset of local minima are global minima [13, 25–29]. The presence of various conclusions reflects the difficulty of the problem: while analyzing the global landscape seems hard, we may step back and analyze the local landscape or a "majority" of the landscape. There are also a few works directly studying the loss surface of deep neural networks, but they either require linear activation functions [30–34], or require assumptions such as independence of ReLU activations [35].

The study of over-parameterized non-linear neural networks can be dated back to 1990's (e.g. [36]). Yu et al. [36] analyzed the landscape of over-parameterized single-hidden-layer neural networks [1], and it motivates the analysis of our paper. It is worth noting that a recent work [37] has simultaneously addressed similar issues to our work, by identifying a class of over-parameterized deep neural networks with no spurious valleys. Their work covers a rather broad range of network structures, but only holds for a limited family of activation functions which do not include ReLU, leaky ReLU and the Swish activation recently proposed in [38]. In contrast, our focus is on the neuron activation function, and our result holds for any continuous activation function, which of course includes ReLU, leaky ReLU and Swish. Thus, [37] and our work can be regarded as complementary works to each other, and they together shed light on the loss surface of deep and over-parameterized neural networks.

Another interesting related work is [23] which aims to understand for which neural network architecture and data, the landscape is nice. It shows that ReLU and leaky ReLU can cause bad local minima

[1] Note that their main result Theorem 3 is not rigorous; it claims that no suboptimal local minimum exists, but in fact their proof only implies no suboptimal strict local minimum exists, and we will give counter-examples to show that suboptimal local minima can exist under their setting.

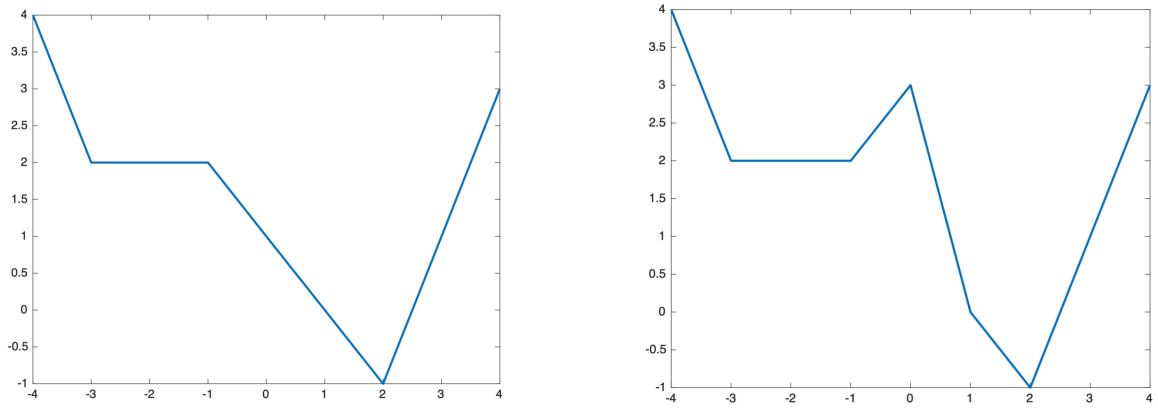

Figure 1: An example of a weakly global function (left) and a non-weakly-global function (right). Both functions have non strict bad local minima, consisting a plateau of $(-3, -1)$. The plateau in the right figure is the bottom of a basin, entailing a strict bad local minimum in the sense of sets.

for certain data distributions, while smooth versions of ReLU (e.g. SoftPlus) can eliminate bad local minima under the same setting. This seems to suggest that ReLU and leaky ReLU are bad at least in terms of optimization landscape. In this paper, we prove that with over-parameterization, ReLU and leaky ReLU are not "truly bad" in the sense that they do not cause setwise *strict* local minima. More specifically, all bad local minima must lie in plateaus of the type shown in the left part of Figure 1. Note that it is easy to verify that the local minima constructed in [23] lie in plateaus, but without the proof of this paper, it is not clear whether these plateaus are like the left part or the right part of Figure 1.

Finally, landscape analysis is just one part of the deep learning theory, which includes representation, algorithm convergence, optimization landscape and generalization. In terms of algorithm convergence, there is much recent interest in analyzing algorithms that escape saddle points for generic non-convex functions [39–42], since escaping saddle points can help converge to local minima. Converging to local minima itself is not that interesting, but will be very interesting if the hypothesis that all local minima are close to global minima holds for certain problems. Our study takes advantage of the structure of the neural networks, and is orthogonal to the research on escaping saddle points. In terms of generalization, many recent works [5, 43] try to understand why over-parameteriztion does not cause overfitting. This is a very interesting line of research, but its underlying assumption that over-parameterization can lead to small training error still requires rigorous justification. Again, our study is orthogonal to the research on the generalization error analysis of over-parameterized networks.

1.3 Paper Organization

This paper is organized as follows. In Section 2, we specify the network model considered in this paper. In Section 3, we present the main results and provide the proofs of the main theorems. Conclusions are presented in Section 5. The proofs of all the lemmas and propositions are provided in Appendix.

2 Preliminaries

In this paper, we study the deep fully connected neural networks with H hidden layers. Assume that the i-th hidden layer contains d_i neurons for $1 \leqslant i \leqslant H$, and the input and output layers contain d_0 and d_{H+1} neurons, respectively. Given an input sample x of dimension d_0, the output of the j-th

neuron of the i-th hidden layer, denoted by $t_{i,j}$, is given by

$$\begin{aligned} t_{1,j}(x) =& \sigma\left(\sum_{k=1}^{d_0} w_{1,j,k}x_k + b_{1,j}\right), \quad 1 \leqslant j \leqslant d_1 \\ t_{i,j}(x) =& \sigma\left(\sum_{k=1}^{d_{i-1}} w_{i,j,k}t_{i-1,k}(x) + b_{i,j}\right), \quad 1 \leqslant j \leqslant d_i, \quad 2 \leqslant i \leqslant H \end{aligned} \tag{1}$$

where $w_{i,j,k}$ is the weight from the k-th neuron of the $(i-1)$-th layer to the j-th neuron of the i-th layer, $b_{i,j}$ is the bias added to the j-th neuron of the i-th layer, and σ is the neuron activation function. The i-th output of the network, denoted by $t_{H+1,j}$, is given by

$$t_{H+1,j}(x) = \sum_{k=1}^{d_H} w_{H+1,j,k}t_{H,j}(x), \quad , 1 \leqslant j \leqslant d_H \tag{2}$$

where $w_{H+1,j,k}$ is the weight to the output layer, defined similarly to that of the hidden layers.

Consider a training dataset consisting of N samples. Denote the s-th sample by (x^s, y^s), $s = 1, \cdots, N$, where $x^s \in \mathbb{R}^{d_0}$ and $y^s \in \mathbb{R}^{d_H}$ are the input and output patterns, respectively. In what follows, we rewrite all the training samples in matrix forms, which allows us to represent the input-output relation of the neural network in a more compact way. Specifically, let $X \triangleq [x^1, x^2, \cdots, x^N] \in \mathbb{R}^{d_1 \times N}$ and $Y \triangleq [y^1, y^2, \cdots, y^N] \in \mathbb{R}^{d_{H+1} \times N}$ as the input and output data matrices, respectively. Then, we define $W_i \in \mathbb{R}^{d_{i-1} * d_i}$ as the weight matrix from the $i-1$-th layer to the i-th layer, and $\mathbf{b}_i \in \mathbb{R}^{d_i}$ as the bias vector of the i-th layer, and $T_i \in \mathbb{R}^{d_i \times N}$ as the output matrix of the i-th layer. The entries of each matrix are given by

$$\begin{aligned} (W_i)_{j,k} &= W_{i,j,k} \\ (\mathbf{b}_i)_j &= b_{i,j} \\ T_i(j,s) &= t_{i,j}(x^s) \end{aligned} \tag{3}$$

for $1 \leq i \leq H+1$, $1 \leq j \leq d_i$, $1 \leq k \leq d_{i-1}$ and $1 \leq s \leq N$. Based on the above definition, we can immediately rewrite the output of each layer as

$$\begin{aligned} T_1 &= \Phi\left(\begin{bmatrix} W_1 \\ \boldsymbol{b_1} \end{bmatrix} \begin{bmatrix} X & \mathbf{1} \end{bmatrix}\right), \\ T_i &= \Phi\left(\begin{bmatrix} W_i \\ \boldsymbol{b_i} \end{bmatrix} \begin{bmatrix} T_{i-1} & \mathbf{1} \end{bmatrix}\right), \quad i = 2, 3, \cdots, H, \\ T_{H+1} &= W_{H+1}T_H. \end{aligned}$$

where $\Phi(\cdot)$ is the operation that applies the activation function σ componentwise to the input matrix and outputs a matrix with the same size. That is, $(\Phi(A))_{ij} = \sigma(A_{ij})$ for any input matrix A.

In the rest of this paper, we simplify the feed-forward operation (2) by ignoring all the bias neurons, yielding

$$\begin{aligned} T_1 &= \Phi(W_1 X), \\ T_i &= \Phi(W_i T_{i-1}), \quad i = 2, 3, \cdots, H, \\ T_{H+1} &= W_{H+1}T_H. \end{aligned} \tag{4}$$

We note that this simplification does not affect our analysis, and therefore the main results also hold for feed-forward deep neural networks with bias. Let $W = (W_1, \cdots, W_{H+1})$ denote all the weights and define the empirical loss as

$$E(W) = l(Y, T_{H+1}) = l(Y, W_{H+1}T_H) \tag{5}$$

where l is the loss function. Then, the training problem of the considered network is to find W to minimize the empirical loss $E(W)$.

3 Main Results

3.1 Assumptions

In this section, we specify our assumptions on the training dataset, the loss functions, the over-parameterization, and the activation functions.

Assumption 1

A1 There exists some k such that $X_{ki} \neq X_{kj}, \forall i, j$;

A2 The loss function $l(Y, T_{H+1})$ is convex and continuous with respect to T_{H+1};

A3 $d_H \geqslant N$.

A4 The activation function σ is continuous.

Assumption A1 implies that the input data samples need to be distinguished with each other in one dimension. This can be always achieved if we allow an arbitrarily small perturbation on data. Assumption A2 is satisfied for almost all commonly-used loss functions, including quadratic, cross entropy, etc. Assumption A3 is the over-parameterization assumption, which only requires the last hidden layer to be wide. There is no assumption on the width of all other hidden layers. Assumption A4 is a very mild assumption on the neuron activation that it should be continuous.

3.2 Main Theorem

In this section, we present our main result on the absence of non-strict local minima for any fully connected deep over-parameterized networks. To this end, we first borrow some definitions from [7].

Definition 1 (Setwise Strict Local minimum) *We say that a compact subset $X \in S$ is a local minimum (respectively, strict local minimum) of $f : S \to \mathbb{R}$ in the sense of sets if there exists $\varepsilon > 0$ such that for all $x \in X$ and for all $y \in S \setminus X$ satisfying $\|x - y\|_2 \leqslant \varepsilon$, it holds that $f(x) \leqslant f(y)$ (respectively, $f(x) < f(y)$).*

Definition 1 generalizes the notion of local minimum and strict local minimum from the sense of points to the sense of sets. Any strict local minimum must be setwise strict local minimum, but not vice versa. Strict local minimum is a single point that is strictly smaller than any points around it, and thus at the bottom of a basin. A simple way to eliminate a strict local minimium is to reparametrize the problem (e.g. replace $z \in \mathbb{R}$ by $z_1 + z_2$) so that this single point becomes a line or a plateau, but this pleateau may still be the bottom of a prolonged basin and cannot be easily escaped from. Such a plateau, which we call setwise strict local minimum, contains a set of truely bad local minima.

Definition 2 (Weakly global function) *We say that $f : S \to \mathbb{R}$ is a weakly global function if it is continuous and all setwise strict local minima are setwise global minima.*

Definition 2 introduces an important class of continuous functions, termed weakly global functions, which admits no strict bad local minima in the sense of sets.

We are now ready to present our main theorem. The detailed proof of the theorem will be given in Section 4

Theorem 1 *Given a fully connected neural network with H hidden layers, activation function σ and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Suppose that Assumptions 1 holds. Then, $E(W)$ is a weakly global function.*

Theorem 1 states that the empirical loss function of an over-parameterized neural network is weakly global as long as the activation function is continuous. Note that the notion of weakly global function is distinct from that of "no bad local valleys" used in [37], but they both guarantee non-existence of strict bad local minimum. Formally, we have the following corollary.

Corollary 1 (Non-Existence of Strict Bad Local Minimum) *For the neural network considered in Theorem 1, there is no strict bad local minimum.*

Note that in a simple problem $\min_{u,v\in\mathbb{R}}(uv-1)^2$, it is easy to show no strict local minimum exists [2], not to mention *bad* strict local minimum. For matrix factorization problems $\min_{U,V}\|M-UV^T\|_F^2$, which can be viewed as 1-hidden-layer neural network problem, it is also easy to show no strict local minimum exists. However, non-linear neural networks are different from matrix factorization since the nonlinear activation functions eliminate a lot of freedom, and for most points there is no continuous symmetry. Discrete symmetry still exists as swapping two neurons do not change the output, but such symmetry only creates discrete copies of a point and thus does not eliminate strict local minimum. Thus, it is not very clear whether strict local minima exist. Consequently, the above result that there is no strict bad local minimum is nontrivial for a non-linear neural network problem.

3.3 Example of Non-strict Bad Local Minima

It is worth mentioning that our result does not guarantee non-existence of non-strict bad local minimum. In fact, for a large class of analytic activation functions satisfying Assumption 2, we can construct simple examples to show that non-strict bad local minimum can exist.

Proposition 1 *Assume that Assumption 1 is satisfied. Suppose σ is an analytic function satisfying Assumption 2. Moreover, there exists $t \neq 0$ and δ such that $\sigma(t) = 0$ and $\sigma(t') \leqslant 0$ for $t-\delta < t' < t+\delta$. Then there exists a network architecture with arbitrary width such that non-strict bad local minimum exists.*

Proof: Consider a two-layer neural network with one input neuron, one output neuron and d hidden neurons. Assume that the loss function is quadratic, which is represented as $l(\boldsymbol{w_1}, \boldsymbol{w_2}) = (y - \boldsymbol{w_2} \cdot \sigma(\boldsymbol{w_1}x))^2$, where $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ are weight vectors in $\mathbb{R}^d$. Now, for any data pair $x \neq 0, y \neq 0$, let $\boldsymbol{w_1^*} = (t/x, t/x, \cdots, t/x)$ and $\boldsymbol{w_2^*} = (\text{sign}(y), \text{sign}(y), \cdots, \text{sign}(y))$. Then $l(\boldsymbol{w_1^*}, \boldsymbol{w_2^*}) = y^2$. In addition, since $\boldsymbol{w_2} \cdot \sigma(\boldsymbol{w_1}x)$ is always non-positive in the neighborhood of $(\boldsymbol{w_1^*}, \boldsymbol{w_2^*})$, $l(\boldsymbol{w_1}, \boldsymbol{w_2})$ is always at least y^2 in the neighborhood of $(\boldsymbol{w_1^*}, \boldsymbol{w_2^*})$, implying that $(\boldsymbol{w_1^*}, \boldsymbol{w_2^*})$ is a local minimum. □

Proposition 1 implies that a neural network without strict local minimum can have non-strict local minimum even if it is arbitrarily wide. Note that the constructed counterexample satisfies Assumption 1, so l does not have strict local minimum. It is easy to see that $(\boldsymbol{w_1^*}, \boldsymbol{w_2^*})$ is a non-strict bad local minimum since $l(\boldsymbol{w_1^*}, \boldsymbol{w_2^*}) = l(\boldsymbol{w_1^*}, \boldsymbol{w_2})$ if $\boldsymbol{w_2}$ has the same sign with $\boldsymbol{w_2^*}$. This counterexample can also be generalized to neural networks with arbitrary N pairs of given data, given σ satisfying that there exists at least N points $t_1, \cdots, t_N$ such that $\sigma(t_i) = 0$ is a local maximum or minimum.

[2] Any nonzero local minimum (u, v) has the same objective value as $(\alpha u, v/\alpha)$ for any nonzero α, thus not a strict local minimum. In addition, $(u, v) = (0, 0)$ is not a local minimum

4 Proof of Theorem 1

Before presenting the proof, we briefly describe the proof sketch. First, we establish the result for a specific class of analytic activation functions. These analytic functions constitute a dense set in the space of continuous functions. In other words, for any continuous activation function, there exists a sequence of analytic functions in the considered class that uniformly converges to it. This also implies the compact convergence of the empirical loss function. Combining with the fact that the property of weakly global is preserved under compact convergence [7], we extend our result to all continuous activation functions and prove Theorem 1.

Step 1: Prove the result for a specific class of activation functions.

Assumption 2 (Special Activation Functions) *The activation function σ is analytic, and its first n derivatives at 0, i.e., $\sigma(0)$, $\sigma'(0)$, $\cdots$, $\sigma^{(n-1)}(0)$, are all non-zero.*

Assumption 2 covers many commonly used activation functions such as sigmoid and softplus, but it does not cover ReLU since it requires smoothness (as mentioned before, ReLU is covered by using the approximation trick). Based on this assumption, we have the following theorem, the proof of which is given in the next section.

Theorem 2 *Consider a fully connected neural network with H hidden layers, activation function σ and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Suppose that Assumption 1 and Assumption 2 hold. Then $E(W)$ is a weakly global function.*

Step 2: Show that the activation function in Assumption 2 can approximate any continuous function.

In order to extend Theorem 2 to all continuous activation functions without dealing them directly, we use a mathematical trick that approximates the continuous activation by a class of analytical functions.

Lemma 1 *For any continuous functions $f : \mathbb{R} \to \mathbb{R}$, there exists a sequence of functions $(f_k)_{k\in\mathbb{N}}$, all satisfying Assumption 2, such that f_k converges to f uniformly.*

Lemma 1 means that the analytic functions satisfying Assumption 2 constitute a dense set in the space of continuous function, which allows us to approximate a neural network with any continuous activation function by a sequence of neural networks under Assumption 2.

Step 3: Show that the the property of weakly global function is preserved under compact convergence.

Having built the relation between the the neural network with analytic activation functions and the neural network with continuous activation function, the last step is to show that the weakly global property is preserved under this relation. The following result is a modification of a result in [7].

Proposition 2 *Consider a sequence of functions $(f_k)_{k\in\mathbb{N}}$ and a function f, all from $S \subset \mathbb{R}^n$ to $\mathbb{R}$. If,*

$$f_k \to f \quad \textit{compactly} \tag{6}$$

and if f_k are weakly global functions on S, then f is a weakly global function on S.

Proposition 2 is slightly different from its original version in [7]: here we assume that f_k are weakly global functions instead of global functions. Nevertheless, we can still prove that f is weakly global by using similar techniques as in [7]. The detailed proof is provided in Appendix B.

Proof:(Proof of Theorem 1) We denote the considered network by $\mathcal{N}$. From Lemma 1, there exists a sequence of activation functions $(\sigma_k)_{k\in\mathbb{N}}$ that uniformly converges to σ. For each $k \in \mathbb{N}$, we construct

a neural network, denoted by $\mathcal{N}_k$, by replacing the activation function in $\mathcal{N}$ with σ_k. For all $\mathcal{N}_k$, we assume the training dataset to be identical to that of $\mathcal{N}$, i.e., X. We also denote the output matrix of the i-th hidden layer by $T_i^{(k)}$ and the empirical loss by

$$E_k(W) = l\left(Y, W_{H+1}T_H^{(k)}\right). \tag{7}$$

From Theorem 2, E_k is a weakly global function with respect to W, $\forall k \in \mathbb{N}$.

Consider the sequence of the empirical loss functions $(E_k)_{k\in\mathbb{N}}$. In what follows, we prove that E_k compactly converges to E. To this end, we first present the following lemma.

Lemma 2 *Consider two continuous functions $f : S \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$, where $S \in \mathbb{R}^m$ is a compact set. Suppose that there exists two sequences of functions $(f_k)_{k\in\mathbb{N}}$ and $(g_k)_{k\in\mathbb{N}}$, such that f_k uniformly converges to f on S, and g_k uniformly converges to g on $\mathbb{R}^n$. Then, $g_k \circ f_k$ converges to $g \circ f$ uniformly on S.*

Consider an arbitrary compact subset S in the space of W. For any $W \in S$, define $\tilde{t}_{i,j,s}^{(k)}(W) = (T_i^{(k)})_{j,s}$ and $\tilde{t}_{i,j,s}(W) = (T_i)_{j,s}$ for any $k \in \mathbb{N}$, $1 \le i \le H$, $1 \le j \le d_i$, and $1 \le s \le N$. That is, we rewrite the output of each neuron in the hidden layers as a function of W. We prove by induction that every sequence $(\tilde{t}_{i,j,s}^{(k)})_{k\in\mathbb{N}}$ converges to $\tilde{t}_{i,j,s}$ uniformly on S.

For $i = 1$, we have

$$\tilde{t}_{1,j,s}^{(k)}(W) = \sigma_k\left(\sum_{l=1}^{d_0}(W_1)_{j,l}X_{l,s}\right) \tag{8}$$

$$\tilde{t}_{1,j,s}(W) = \sigma\left(\sum_{l=1}^{d_0}(W_1)_{j,l}X_{l,s}\right). \tag{9}$$

Since σ_k uniformly converges to σ, $\tilde{t}_{1,j,s}^{(k)}$ also uniformly converges to $\tilde{t}_{1,j,s}$ on S for all $1 \le j \le d_1$, $1 \le s \le N$.

For $i > 1$, assume that $\tilde{t}_{i-1,j,s}^{(k)}$ uniformly converges to $\tilde{t}_{i-1,j,s}$ on S for all $1 \le j \le d_{i-1}$, $1 \le s \le N$. For the i-th layer, we have

$$\tilde{t}_{i,j,s}^{(k)}(W) = \sigma_k\left(\sum_{l=1}^{d_{i-1}}(W_i)_{j,l}\left(T_{i-1}^{(k)}\right)_{l,s}\right) = \sigma_k\left(\sum_{l=1}^{d_{i-1}}(W_i)_{j,l}\tilde{t}_{i-1,j,s}^{(k)}(W)\right) \tag{10}$$

$$\tilde{t}_{i,j,s}(W) = \sigma\left(\sum_{l=1}^{d_{i-1}}(W_i)_{j,l}(T_{i-1})_{l,s}\right) = \sigma\left(\sum_{l=1}^{d_{i-1}}(W_i)_{j,l}\tilde{t}_{i-1,j,s}(W)\right). \tag{11}$$

By the induction hypothesis, it is easy to show that $\sum_{l=1}^{d_{i-1}}(W_i)_{j,l}\tilde{t}_{i-1,j,s}^{(k)}(W)$ uniformly converges to $\sum_{l=1}^{d_{i-1}}(W_i)_{j-1,l}\tilde{t}_{i,j,s}(W)$ on S. It directly follows from Lemma 2 that $\tilde{t}_{i,j,s}^{(k)}(W)$ converges to $\tilde{t}_{i,j,s}(W)$. Therefore, we conclude that $\tilde{t}_{i,j,s}^{(k)}$ converges to $\tilde{t}_{i,j,s}$ uniformly on S for every $1 \le i \le H$, $1 \le j \le d_i$, and $1 \le s \le N$.

Now we consider the empirical loss

$$E_k(W) = l\left(Y, W_{H+1}T_H^{(k)}\right) \tag{12}$$

$$E(W) = l\left(Y, W_{H+1}T_H\right). \tag{13}$$

As every component of $T_H^{(k)}$ converges uniformly to the corresponding component of T_H, it can be shown that $W_{H+1}T_H^{(k)}$ converges uniformly to $W_{H+1}T_H$ on S. By Lemma 2, where we set both g_k and g to the loss function l, we have that E_k uniformly converges to E on S. Noting that S is an arbitrary compact subset in the space of W, the empirical loss E_k converges to E compactly on the space of W. Since $E_k(W)$ is a weakly global function for every $k \in \mathbb{N}$, by Proposition 2, $E(W)$ is also a weakly global function. We complete the proof. □.

4.1 Proof of Theorem 2

The proof of Theorem 2 consists of three steps. First, we show that for any W, we can perturb it to a point W' whose corresponding T_H is full rank. Second, we prove that starting from the perturbed point W', there exists a strictly decreasing path reaching the global infimum. Thus, we prove the first conclusion of Theorem 2. Finally, we show that the second conclusion is a natural consequence of the first one.

In order to present a rigorous proof, we introduce a useful lemma.

Lemma 3 *Given a fully connected neural network with H hidden layers, activation function σ and empirical loss function $E(W) = l(Y, W_{H+1}T_H)$. Let $\Omega = \{(W_1, \cdots, W_H) \mid rank(T_H) < \min\{d_H, N\}\}$. If Assumption 1 and Assumption 2 hold, Ω is a zero-measure set.*

Lemma 3 shows that the set of W that gives rise to a non-full-rank output matrix of the last hidden layer only constitute a zero-measure set. This result does not require any assumption on over-parameterization. In fact, this nice property is introduced by non-linearity of the activation function. Using this lemma, we now provide the formal proof of Theorem 2.

Proof:(Proof of Theorem 2)

We first prove that from any initial weight $W^o = (W_1^o, \cdots, W_{H+1}^o)$, there exists a strictly decreasing path reaching $\inf_W E(W)$ after an arbitrarily small perturbation. According to Lemma 3, all W's that entail a non-full-rank T_H only constitute a zero-measure set. Therefore, for any initial weight W^o and an arbitrarily small $\delta > 0$, there exists $\hat{W}^p = (W_1^p, W_2^p, \cdots, W_{H+1}^p) \in B(W_0, \delta)$ such that the corresponding T_H^p is full rank. Since $d_H \geqslant N$, we have $\text{rank}(T_H^p) = N$.

In what follows, we show that starting from W^p, there exists a strict decreasing path reaching $\inf_W E(W)$. Denote $\hat{W} = (W_1, \cdots, W_H)$, i.e., the weights in the first H layers. By the feed-forward operation (4), T_H is a function of $\hat{W}$. Thus, $E(W)$ can be rewritten as $l(Y, W_{H+1}T_H(\hat{W}))$. Since $l(Y, \hat{Y})$ is convex to $\hat{Y}$, for any W_{H+1}^1, W_{H+1}^2 and $\lambda \in [0, 1]$, we have

$$\begin{aligned} E(W) &= l\left(Y, \left(\lambda W_{H+1}^1 + (1-\lambda)W_{H+1}^2\right) T_H(\hat{W})\right) \\ &= l\left(Y, \lambda W_{H+1}^1 T_H(\hat{W}) + (1-\lambda)W_{H+1}^2 T_H(\hat{W})\right) \\ &\leqslant \lambda l\left(Y, W_{H+1}^1 T_H(\hat{W})\right) + (1-\lambda) l\left(Y, W_{H+1}^2 T_H(\hat{W})\right) \end{aligned} \tag{14}$$

Thus, with the weights of the first H layers fixed, $E(W)$ is convex with respect to W_{H+1}. This implies that starting from W^p, we can find a strict decreasing path reaching $\inf_{W_{H+1}} l(Y, W_{H+1}T_H(\hat{W}^p))$ by fixing $\hat{W} = \hat{W}^p$ and moving along W_{H+1}. Moreover, since $T_H(\hat{W}^p) \in \mathbb{R}^{d_H \times N}$ is full column rank, for any $\hat{Y} \in \mathbb{R}^{d_{H+1} \times N}$, there exists W_{H+1} such that $W_{H+1}T_H(\hat{W}^o) = \hat{Y}$, yielding

$$\inf_{W_{H+1}} l(Y, W_{H+1}T_H(\hat{W}^p)) = \inf_{\hat{Y}} l(Y, \hat{Y}) = \inf_{W} E(W). \tag{15}$$

Therefore, the constructed path is strictly decreasing towards $\inf_W E(W)$. We complete the proof of the first conclusion.

Now we prove by contraposition that $E(W)$ is a weakly global function. Assume in contrast that there exists a strict bad local minimum of $E(W)$ in the sense of sets, denoted by $\mathcal{W}$. Note by Definition 2, $\mathcal{W}$ is a compact set. Let $\mathcal{W}_\delta = \{W' \mid \inf_{W\in\mathcal{W}} \|W' - W\|_2 \leqslant \delta\}$, then there exists $\delta > 0$ such that for all $W \in \mathcal{W}$ and $W' \in \mathcal{W}_\delta \setminus \mathcal{W}$, $E(W) < E(W')$. Denote $\partial\mathcal{W}_\delta$ as the boundary of $\mathcal{W}_\delta$. Note that both $\mathcal{W}_\delta$ and $\partial\mathcal{W}_\delta$ are closed, there exists W^* such that $E(W^*) = \inf_{W'\in\partial\mathcal{W}_\delta} E(W')$. Moreover, $E(W^*) = \sup_{W\in\mathcal{W}} E(W) + \varepsilon$ for some $\varepsilon > 0$.

Consider an arbitrary point $W^o \in \mathcal{W}$. Since $E(W)$ is an analytic function, there exists $\delta > \delta_0 > 0$ such that for any $W' \in B(W^o, \delta_0)$, $|E(W') - E(W)| < \varepsilon/2$. According to the first conclusion, we can find $W^p \in B(W^o, \delta_0)$ such that there exists a strictly decreasing path from W^p to $\inf_W E(W)$. Since $\mathcal{W}$ is a bad local minimum, $\inf_{W\in\mathcal{W}_\delta} E(W) > \inf_W E(W)$. Therefore, the above strictly decreasing path starting from W^p must pass through the boundary $\partial\mathcal{W}_\delta$. However, $E(W^p) < E(W^o) + \varepsilon/2 < \sup_{W\in\mathcal{W}} E(W) + \varepsilon = E(W^*) = \inf_{W'\in\partial\mathcal{W}_\delta} E(W')$. This implies that the considered path can never be strictly decreasing, leading to a contradiction. Therefore, we conclude that there is no strict bad local minima in the sense of sets, and therefore $E(W)$ is a weakly global function. We complete the proof of the second conclusion. □

5 Conclusions

In this paper, we studied the loss surface of over-parameterized fully connected deep neural networks. We show that for all continuous activation functions, there is no strict bad local minima, and the non-strict bad local minima cannot lie in bad local valleys. We also show that for almost all analytic activation functions, there exists a strictly decreasing path to the global infimum if we allow an arbitrarily small perturbation at the initial point. This provides an intuitive explanation on why local search algorithms usually converges to the global minimum. Future research directions include exploiting our results to design efficient training algorithms for practical over-parameterized neural networks.

References

[1] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[2] Yoshua Bengio, Nicolas L Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *NIPS*, pages 123–130, 2006.

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

[4] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[5] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[6] David Lopez-Paz and Levent Sagun. Easing non-convex optimization with neural networks. 2018.

[7] C Josz, Y Ouyang, UC IEOR, RY Zhang, J Lavaei, and S Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. *NIPS*, 2018.

[8] S. S Du and J. D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

[9] R. Ge, J. D Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. *ICLR*, 2018.

[10] A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang. Learning polynomials with neural networks. In *ICML*, 2014.

[11] H. Sedghi and A. Anandkumar. Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*, 2014.

[12] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[13] B. D Haeffele and R. Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[14] A. Gautier, Q. N. Nguyen, and M. Hein. Globally optimal training of generalized polynomial neural networks with nonlinear spectral methods. In *NIPS*, pages 1687–1695, 2016.

[15] A. Brutzkus and A. Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[16] M. Soltanolkotabi. Learning relus via gradient descent. In *NIPS*, pages 2004–2014, 2017.

[17] D. Soudry and E. Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

[18] S. Goel and A. Klivans. Learning depth-three neural networks in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017.

[19] D. Boob and G. Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

[20] S. S. Du, J. D. Lee, and Y. Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

[21] K. Zhong, Z. Song, P. Jain, P. L Bartlett, and I. S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *ICLR*, 2017.

[22] Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with relu activation. In *NIPS*, pages 597–607, 2017.

[23] S. Liang, R. Sun, Y. Li, and R. Srikant. Understanding the loss surface of neural networks for binary classification. 2018.

[24] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561*, 2018.

[25] B. Haeffele, E. Young, and R. Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *ICML*, 2014.

[26] D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[27] Q. Nguyen and M. Hein. The loss surface and expressivity of deep convolutional neural networks. *arXiv preprint arXiv:1710.10928*, 2017.

[28] Q. Nguyen and M. Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

[29] O. Shamir. Are resnets provably better than linear predictors? *arXiv preprint arXiv:1804.06739*, 2018.

[30] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[31] K. Kawaguchi. Deep learning without poor local minima. In *NIPS*, pages 586–594, 2016.

[32] C D. Freeman and J. Bruna. Topology and geometry of half-rectified network optimization. *ICLR*, 2016.

[33] M. Hardt and T. Ma. Identity matters in deep learning. *ICLR*, 2017.

[34] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Global optimality conditions for deep neural networks. *arXiv preprint arXiv:1707.02444*, 2017.

[35] A. Choromanska, M. Henaff, M. Mathieu, G. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

[36] Xiao-Hu Yu and Guo-An Chen. On the local minima free condition of backpropagation learning. *IEEE Transactions on Neural Networks*, 6(5):1300–1303, 1995.

[37] Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. *arXiv preprint arXiv:1809.10749*, 2018.

[38] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. 2018.

[39] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

[40] J. D Lee, M. Simchowitz, M. I Jordan, and B. Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[41] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[42] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017.

[43] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[44] Wilfred Kaplan. Approximation by entire functions. *Michigan Math. J.*, 3(1):43–52, 1955.

[45] Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

A Proof of Lemma 1

The proof of Lemma 1 consists of two parts. In the first part we show that the function class specified by Assumption 2 is dense in the space of analytic functions. In the second part, following the fact that the space of analytic functions is a dense set in the space of continuous function, we prove that the function class specified by Assumption 2 is also dense in the space of continuous functions.

To prove the first part, we consider an arbitrary analytic function $g : \mathbb{R} \to \mathbb{R}$, and then construct a sequence of functions $(f_k)_{k\in\mathbb{N}}$, all satisfying Assumption 2, such that f_k converges to g uniformly.

Let

$$f_k(x) = g(x) + \frac{1}{s(k+1)} \left(\sin x + \cos x\right). \tag{16}$$

Clearly, f_k is analytic for any $k \in \mathbb{N}_+$ and $s \neq 0$. Further, we have

$$f_k^{(n)}(0) = g^{(n)}(0) + \frac{1}{s(k+1)}(-1)^n. \tag{17}$$

We next show that there exists $s \neq 0$ such that all f_k's satisfy Assumption 2. Consider the following two cases: (1) $g^{(n)}(0) = 0$ for all $0 \leq n \leq N$; and (2) $g^{(n)}(0) \neq 0$ for some $0 \leq n \leq N$.

Case 1: For any $s \neq 0$, since $g^{(n)}(0) = 0$, we have

$$f_k^{(n)}(0) = \frac{1}{s(k+1)}(-1)^n \neq 0 \tag{18}$$

for all $n = 0, 1, \cdots, N$. Thus, all f_k's satisfy Assumption 2.

Case 2: Since $g^{(n)}(0) \neq 0$ for at least one $n \in \{0, 1, \cdots, N\}$, we can define

$$\delta_{\min} = \min\left\{|g^{(n)}(0)| \mid 0 \leq n \leq N, \;\; g^{(n)}(0) \neq 0\right\} \tag{19}$$

i.e., the minimum non-zero absolute value of $g^{(n)}(0)$, $n = 0, 1, \cdots, N$. Clearly, $\delta_{\min} > 0$. Letting $s = 2/\delta_{\min}$, we have

$$f_k^{(n)}(0) = g^{(n)}(0) + \frac{\delta_{\min}}{2(k+1)}(-1)^n \tag{20}$$

For $g^{(n)}(0) = 0$, we have

$$f_k^{(n)}(0) = \frac{\delta_{\min}}{2(k+1)}(-1)^n \neq 0. \tag{21}$$

For $g^{(n)}(0) \neq 0$, we have

$$\begin{aligned}
\left|f_k^{(n)}(0)\right| &= \left|g^{(n)}(0) + \frac{\delta_{\min}}{2(k+1)}(-1)^n\right| && \text{(22a)}\\
&\geq \left|g^{(n)}(0)\right| - \left|\frac{\delta_{\min}}{2(k+1)}(-1)^n\right| && \text{(22b)}\\
&\geq \delta_{\min} - \frac{\delta_{\min}}{2(k+1)} && \text{(22c)}\\
&= \frac{\delta_{\min}(2k+1)}{2(k+1)} && \text{(22d)}\\
&> 0 && \text{(22e)}
\end{aligned}$$

where (22c) holds by the definition of $\delta_{\min}$ in (19). Therefore, all f_k's satisfy Assumption 2.

We now prove the uniform convergence of f_k for any $s \neq 0$. Specifically, for any $\epsilon > 0$, we have

$$|f_k(x) - g(x)| = \frac{1}{s(k+1)} |\sin x + \cos x| \leq \frac{\sqrt{2}}{s(k+1)} < \epsilon \tag{23}$$

for all $k > \sqrt{2}/(\epsilon s) - 1$ and $x \in \mathbb{R}$. Therefore, f_k converges uniformly to g.

We conclude that function class specified by Assumption 2 is dense in the space of analytic functions.

Now we come to the second part. By the Carleman Approximation Theorem [44], the space of analytic functions is dense in the space of continuous functions. That is, for any continuous function $f : \mathbb{R} \to \mathbb{R}$, there exists a sequence of analytic functions $(g_k)_{k\in\mathbb{N}}$ such that g_k converges to f uniformly. Following the idea of Cantor's diagonal argument, we can construct a sequence of functions satisfying Assumption 2, which also converges to f.

Note that each g_k is an analytic function. By the analysis in the first part, for each $k \in \mathbb{N}$, we can construct a sequence of functions $(f_j^{(k)})_{j\in\mathbb{N}}$, all satisfying Assumption 2, such that $f_j^{(k)}$ converges to g_k uniformly. Further, we can require that for each $k \in \mathbb{N}$,

$$\left|f_j^{(k)}(x) - g_k(x)\right| \leq \frac{1}{k+1}, \quad \forall x \in \mathbb{R}, \quad j \in \mathbb{N}. \tag{24}$$

In fact, if (24) is not satisfied, we can always delete a finite number of functions at the beginning of the sequence, so as to produce a new sequence that meet the requirement. Now considered the sequence $(f_k^{(k)})_{k\in\mathbb{N}}$. Since g_k converges to f uniformly, for any $\epsilon > 0$, there exists a $K_1 \in \mathbb{N}$ such that $|g_k(x) - f(x)| \leq \epsilon/2$ for any $k \geq K_1$ and $x \in \mathbb{R}$. Then, for any $k > \max\{K_1, 2/\epsilon - 1\}$, we have

$$\left|f_k^{(k)}(x) - f(x)\right| \leq \left|f_k^{(k)}(x) - g_k(x)\right| + |g_k(x) - f(x)| \leq \frac{1}{k+1} + \epsilon/2 \leq \epsilon. \tag{25}$$

Therefore, $f_k^{(k)}$ converges to f uniformly. Noting that f is an arbitrary continuous function from $\mathbb{R}$ to $\mathbb{R}$, We complete the proof.

B Proof of Proposition 2

Consider a sequence of weakly global functions f_k that converge compactly towards f. Since $S \subset \mathbb{R}^n$ and $\mathbb{R}^n$ is a compactly generated space, it follows that f is continuous. We proceed to prove that f is a weakly global function by contradiction. Suppose $X \subset S$ is a strict local minimum that is not global minimum. There exists $\epsilon > 0$ such that the uniform neighborhood $V := \{y \in S \mid \exists\, x \in X : \|x-y\|_2 \leqslant \epsilon\}$ satisfies $f(x) < f(y)$ for all $x \in X$ and for all $y \in V \setminus X$. Since f is continuous on the compact set X, it attains a minimal value on it, say $\inf_X f := \alpha + \inf_S f$ where $\alpha > 0$ since X is not a global minimum. Consider a compact set $V \subset K \subset S$ such that $\inf_K f \leqslant \alpha/2 + \inf_S f$. Since f is continuous on the compact set ∂V, it attains a minimal value on it, say $\inf_{\partial V} f := \beta + \inf_X f$ where $\beta > 0$ by strict optimality. Let $\gamma := \min\{\alpha/2, \beta\}$. For a sufficiently large value of k, compact convergence implies that $|f_k(y) - f(y)| \leqslant \gamma/3$ for all $y \in K$. Since the function f_k is compact on V, it attains a minimum, say $z' \in V$. Consider the compact set defined by $Z := \{z \in V \mid f(z) = f(z')\}$. Therefore, for any $z \in Z$,

$$f_k(z) \;\leqslant\; \gamma/3 + \inf_V f \;\leqslant\; \beta/3 + \inf_V f \;<\; 2\beta/3 + \inf_V f \tag{26}$$

$$\leqslant\; -\gamma/3 + \beta + \inf_V f \;\leqslant\; -\gamma/3 + \inf_{\partial V} f \;\leqslant\; \inf_{\partial V} f_k. \tag{27}$$

Thus, $z \in \mathrm{int}(V)$. So $Z \subseteq \mathrm{int}(V)$. Since both Z and ∂V are compact, we have $d(\partial V, Z) > 0$. We now proceed to show by contradiction that Z is a strict local minimum of f_k. Assume that for all $\epsilon' > 0$,

there exists $y' \in S \setminus Z$ satisfying $d(y', Z) \leqslant \epsilon'$ such that $f_k(z) \geqslant f_k(y')$ for some $z \in Z$. We can choose $\epsilon' < d(\partial V, Z)$ to guarantee that y' belongs to V since $Z \subseteq \mathrm{int}(V)$. The point y' then contradicts the strict minimality of Z on V. This means that $Z \in V$ is a strict local minimum of f_k. Now, observe that for any $z \in Z$,

$$\inf_K f_k \leqslant \gamma/3 + \inf_K f \leqslant \gamma/3 + \alpha/2 + \inf_S f \leqslant 2\alpha/3 + \inf_S f < 5\alpha/6 + \inf_S f \tag{28}$$

$$\leqslant \alpha - \gamma/3 + \inf_S f = -\gamma/3 + \inf_X f = -\gamma/3 + \inf_V f \leqslant \inf_V f_k \leqslant f_k(z). \tag{29}$$

Thus, Z is not a global minimum of f_k. This contradicts the fact that f_k is a weakly global function.

C Proof of Lemma 2

Let $D \subset \mathbb{R}^n$ be the domain of f on S. Since S is compact and f is continuous, D is also compact. Define

$$D' = \{z \in \mathbb{R}^n \mid \exists z_0 \in D, \quad ||z - z_0|| \leq 1\}. \tag{30}$$

Then, D' is also compact.

Since g is continuous, its restriction on D' is uniformly continuous. That is, for any $\epsilon > 0$, there exits $\delta > 0$ such that

$$|g(z_1) - g(z_2)| \leq \frac{\epsilon}{2}, \quad \forall z_1, z_2 \in D', \quad ||z_1 - z_2|| \leq \delta. \tag{31}$$

Further, since f_k converges to f uniformly on S, there exists $K_1 \in \mathbb{N}$ such that

$$||f_k(x) - f(x)|| \leq \min\{1, \delta\}, \quad \forall k \geq K_1, \quad x \in S. \tag{32}$$

Note that by the definition of D', (32) also implies $f_k(x) \in D'$ for all $k \geq K_1$ and $x \in S$. Also, as g_k uniformly converges to g, there exists $K_2 \in \mathbb{N}$ such that $|g_k(z) - g(z)| \leq \epsilon/2$ for all $k \geq K_2$ and $z \in \mathbb{R}^n$.

For any $k \geq \max\{K_1, K2\}$ and $x \in S$, we have

$$|g_k(f_k(x)) - g(f(x))| \leq |g_k(f_k(x)) - g(f_k(x))| + |g(f_k(x)) - g(f(x))| \tag{33a}$$

$$\leq \frac{\epsilon}{2} + |g(f_k(x)) - g(f(x))| \tag{33b}$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \tag{33c}$$

$$= \epsilon \tag{33d}$$

where (33c) follows from (31), (32), and the fact that $f_k(x), f_(x) \in D'$.

Therefore, we conclude that $g_k \circ f_k$ converges to $g \circ f$ uniformly on S. We complete the proof.

D Proof of Lemma 3

To prove Lemma 3, we first present several lemmas.

Lemma 4 *If Assumption 2 holds, then for any $x_1, \cdots, x_n$ such that $x_i \neq x_j, i \neq j$, the following matrix*

$$A = \begin{pmatrix} \sigma(0) & \sigma(0) & \cdots & \sigma(0) \\ x_1\sigma'(0) & x_2\sigma'(0) & \cdots & x_n\sigma'(0) \\ \vdots & \vdots & & \vdots \\ x_1^{n-1}\sigma^{(n-1)}(0) & x_2^{n-1}\sigma^{(n-1)}(0) & \cdots & x_n^{n-1}\sigma^{n-1}(0) \end{pmatrix}$$

is non-singular.

Proof: Notice that A is a Vandermonde matrix multiplied by $\sigma^{j-1}(0)$ to the j-th row. Since $\sigma^{j-1}(0) \neq 0$ according to Assumption 2, A is a non-singular matrix. □

Lemma 5 *Let $f(w) : \mathbb{R}^n \to \mathbb{R}$ be a real analytic function on $\mathbb{R}^n$. If f is not identically zero, then its zero set $\Omega = \{w \in \mathbb{R}^n \mid f(w) = 0\}$ has zero measure.*

Lemma 5 is the main result of [45]. It states that the zero set of an analytic function is either $\mathbb{R}^n$ or zero-measure.

Lemma 6 *Suppose that σ is an analytic function satisfying Assumption 2. Given $a, b \in \mathbb{R}^n$, let $\Omega = \{w \in \mathbb{R}^n \mid \sigma(a^\top w) = \sigma(b^\top w)\}$. If $a \neq b$, then Ω is of measure zero.*

Proof: Assume that Ω is not of zero measure. Since $\sigma(a^\top w) - \sigma(b^\top w)$ is an analytic function of w, Ω must be $\mathbb{R}^n$ according to Lemma 5. In the following, we show that this leads to a contradiction.

If $a = 0$ or $b = 0$, assume $a = 0$ without loss of generality. Since $b \neq 0$, there exists some w^0 such that $b^\top w^0 = 1$. Therefore, for any $\lambda \in \mathbb{R}$, we have

$$\sigma(\lambda) = \sigma(b^\top(\lambda w^0)) = \sigma(a^\top(\lambda w^0)) = \sigma(0).$$

Thus, σ is a constant function and therefore $\sigma' \equiv 0$, a contradiction to Assumption 2.

If $a \neq 0$ and $b \neq 0$, then the set of w that satisfies $a^\top w = 0$ or $b^\top w = 0$ is of measure zero. Since $a \neq b$, the set of w that satisfies $a^\top w = b^\top w$ is also of zero measure. Therefore, there exists some w^0 such that both $a^\top w^0$ and $b^\top w^0$ are non-zero as well as $a^T w^0 \neq b^T w^0$. Denote $a_0 = a^T w^0 and b_0 = b^T w^0$. Since $\Omega = \mathbb{R}^n$, we conclude that for any $\lambda > 0$, $\sigma(\lambda a_0) = \sigma(a^\top(\lambda w^0)) = \sigma(b^\top(\lambda w^0)) = \sigma(\lambda b_0)$. Note that $a_0 b_0 \neq 0$, $a_0 \neq b_0$, and $\sigma(a_0) = \sigma(b_0)$. Letting $\lambda \to 0$, we have

$$0 = \lim_{\lambda \to 0} \frac{\sigma(\lambda a_0) - \sigma(\lambda b_0)}{\lambda a_0 - \lambda b_0} = \sigma'(0)$$

where the second equality holds since σ is analytic. This also contradicts Assumption 2.

We conclude that for any $a \neq b$, Ω cannot be $\mathbb{R}^n$, and therefore must have zero measure. □

In the following we will provide a somewhat "hierarchical" proof for Lemma 3. Specifically, we first consider a special case – a two-layer neural network with one input neuron and N neurons in the hidden layer. We prove that the output matrix of the hidden layer has full column rank for almost all W. Then, we generalize the proof to two-layer networks with arbitrary number of input neurons. Finally, we show that Assumption A1 can be preserved for the input of every hidden layer, which allows us to prove Lemma 3 by induction.

We start by investigating the easiest case.

Proposition 3 (Rank-1 Two-Layer Case) *Consider a two-layer neural network with one input neuron and N neurons in the hidden layer. Given an activation function σ and $x \in \mathbb{R}^n$, let $\Omega = \{w \in \mathbb{R}^N \mid det(\sigma(wx^T)) = 0\}$. If Assumption A1-A2, and Assumption 2 hold, then Ω is a zero-measure set.*

Proof: We prove this result by induction on N. The conclusion is obvious when $N = 1$.

Since $f(w) \triangleq \det(wx^T)$ is an analytic function with respect to w, from Lemma 5 we know that Ω is either $\mathbb{R}^N$ or a zero-measure set. We now prove that Ω cannot be $\mathbb{R}^N$.

Assume on the contrary that $\Omega = \mathbb{R}^N$, i.e., $f(w) = 0$, $\forall w \in \mathbb{R}^N$. For $k \geqslant 0$, denote the k-th order partial derivative with respect to w_1 as

$$G_k(w) \triangleq \frac{\partial^k f(w)}{\partial w^k} = \det \begin{pmatrix} x_1^k \sigma^{(k)}(w_1 x_1) & x_2^k \sigma^{(k)}(w_1 x_2) & \cdots & x_N^k \sigma^{(k)}(w_1 x_N) \\ \sigma(w_2 x_1) & \sigma(w_2 x_2) & \cdots & \sigma(w_2 x_N) \\ \vdots & \vdots & & \vdots \\ \sigma(w_N x_1) & \sigma(w_N x_2) & \cdots & \sigma(w_N x_N) \end{pmatrix}$$

As $f(w) = 0$, $\forall w \in \mathbb{R}^N$, we have $G_k(w) = 0$, $\forall w \in \mathbb{R}^N$.

Denote $u_k = [\sigma(w_k x_1), \cdots, \sigma(w_k x_N)]^T, k = 2, \cdots, N$. We show there exist some $w_2, \cdots, w_N$ such that $u_2, \cdots, u_N$ are linearly independent. In fact, denote $\hat{u}_k = [\sigma(w_k x_1), \cdots, \sigma(w_k x_{N-1})]^T, k = 2, \cdots, N$, and $\hat{G} = [\hat{u}_2, \cdots, \hat{u}_N]$. According to the induction hypothesis, the set $\{(w_2, \cdots, w_N) \mid \det(\hat{G}_k) \neq 0\}$ is zero-measure in $\mathbb{R}^{N-1}$, implying that there exist some $w_2, \cdots, w_N$ such that $\hat{u}_2, \cdots, \hat{u}_N$ are linearly independent. This also implies that $u_2, \cdots, u_N$ are linearly independent.

Now we have found some $w_2, \cdots, w_N$ such that $u_2, \cdots, u_N$ are linearly independent. Fix $w_2, \cdots, w_N$ and let $w_1 = 0$. Denote the first row of G_k as a_k. Since $\det(G_k) = 0$, a_k must be a linear combination of $u_2, \cdots, u_N$ for any $k \geqslant 0$, so all a_k's lie in a $(N-1)$-dimension space. However, according to Lemma 4, the N vectors $a_0, \cdots, a_{N-1}$ are linearly independent, which is a contradiction.

Therefore we have proved that Ω cannot be $\mathbb{R}^N$, so it must be a zero-measure set. □

For general two-layer cases, we have the following result.

Proposition 4 (General Two-layer Case) *Consider a two-layer neural network where $X \in \mathbb{R}^{m \times N}$, $W \in \mathbb{R}^{d \times m}$, and $Y = \Phi(WX)$. Let $\Omega_1 = \{W \in \mathbb{R}^{d \times m} \mid rank(Y) < \min\{d, N\}\}$. If assumption A1, A2, and 2 hold, then Ω_1 is a zero-measure set.*

Proof: Let $\boldsymbol{w_i}^T$ and $\boldsymbol{x_j}$ be the i-th row of W and the $j-th$ column X, respectively. According to Assumption A1, we can assume without loss of generality that the first row of X has distinct entries, i.e., $(\boldsymbol{x_1})_1, \cdots, (\boldsymbol{x_N})_1$ are distinct from each other.

Notice that $Y \in \mathbb{R}^{d \times N}$. If $d < N$, we select the first d columns of Y and obtain a sub-matrix $\hat{Y} \in \mathbb{R}^{d \times d}$. Let $\Omega_1' = \{W \in \mathbb{R}^{d \times m} \mid \text{rank}(\hat{Y}) < d\}$. We can show that Ω_1' is a zero-measure set by applying a similar analysis to $\hat{Y}$ as in the proof of Proposition 3. The only change to make is that here we calculate the partial derivatives with respect to $(\boldsymbol{w_1})_1$. Notice that for any $W \in \Omega_1$, any d-by-d sub-matrix of Y should be singular. Therefore Ω_1 is a subset of Ω_1', so it should also be zero measure.

If $d \geqslant N$, we select the first N rows of Y and obtain a sub-matrix $\hat{Y} \in \mathbb{R}^{N \times N}$. Similarly, let $\hat{W} \in \mathbb{R}^{N \times m}$ be the first N rows of W. Let $\Omega_1' = \{\hat{W} \in \mathbb{R}^{N \times m} \mid \text{rank}(\hat{Y}) < N\}$. From Proposition 3 to $\hat{Y}$, Ω_1' is of measure zero in $\mathbb{R}^{N \times m}$. Note that for any $W \in \Omega_1$, the submatrix consisting of the first N rows is in Ω_1'. Thus, Ω_1 is of measure zero in $\mathbb{R}^{d \times m}$. □

Finally, we consider the general over-parameterized deep networks and accomplish the proof of Lemma 3.

Proof:(Proof of Lemma 3) Denote $W^{(i)} = (W_1, W_2, \cdots, W_i)$, i.e., the weights of the first i hidden layers. Define

$$\Omega_i = \{W^{(i)} \mid \text{rank}(T_i) < \min\{d_i, N\}\} \tag{34}$$

$$\hat{\Omega}_i = \{W^{(i)} \mid \forall j = 1, \cdots, d_i, \quad \exists k_j, s_j, \quad s.t. \ \ (T_i)_{j,k_j} = (T_i)_{j,s_j}\}. \tag{35}$$

Ω_i is the set of $W^{(i)}$ such that the output matrix of the i-th hidden layer is not full rank, which generalizes Ω_1 defined in Proposition 4. $\hat{\Omega}_i$ is the set of $W^{(i)}$ such that there exist identical entries

in every row of T_i. That is, for any $W^{(i)} \in \hat{\Omega}_i$, the resulting T_i, if regarded as an input data matrix, violates Assumption A1.

In the following, we prove by induction that $\Phi_i \triangleq \Omega_i \cup \hat{\Omega}_i$ is of measure zero for all $1 \leq i \leq H$.

We first consider the case with $i = 1$. By Proposition 4, Ω_1 is of measure zero . Further, note that $(T_1)_{j,k} = \sigma((\boldsymbol{w}_1)_j^\top \boldsymbol{x}_k)$, where $(\mathbf{w}_1)_j^T$ and $\mathbf{x}_k$ are the j-th row of W_1 and the k-th column of X, respectively. Noting that Assumption A1 guarantees that $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ are non-identical from each other, from Lemma 6, $\hat{\Omega}_1$ is of measure zero. As a result, $\Phi_1 = \Omega_1 \cup \hat{\Omega}_1$ is also of measure zero.

Now assume that Φ_{i-1} is of measure zero. Then Φ_i can be decomposed into

$$\begin{aligned}\Phi_i = &\left\{W^{(i)} \mid W^{(i-1)} \in \Phi_{i-1}, \quad W^{(i)} \in \Omega_i, \quad W^{(i)} \in \hat{\Omega}_i\right\} \\ &\cup \left\{W^{(i)} \mid W^{(i-1)} \notin \Phi_{i-1}, \quad W^{(i)} \in \Omega_i, \quad W^{(i)} \in \hat{\Omega}_i\right\} \end{aligned} \tag{36}$$

By the induction hypothesis, the first component of the set union in (36) has zero measure in the space of $W^{(i)}$. Moreover, for $\hat{W}^{(i-1)} \notin \hat{\Omega}$, the resulting T_{i-1}, if regarded as an input data matrix, satisfies Assumption A1. Following a similar procedure as in the case of $i = 1$, we obtain that the set of W_i satisfying $(\hat{W}^{(i-1)}, W_i) \in \Omega$ has zero measure in $\mathbb{R}^{d_i \times d_{i-1}}$. This implies that the second component of the set union in (36) also has zero measure in the space of $W^{(i)}$. Therefore, Φ_i is of measure zero for all $1 \leq i \leq H$.

Noting that $\Omega = \Omega_H$, we complete the proof of Lemma 3 □