Is the Skip Connection Provable to Reform the Neural Network Loss Landscape?

Lifu Wang[1] , **Bo Shen**[1*] , **Ning Zhao**[2] and **Zhiyuan Zhang**[1]
[1]Department of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education Beijing Jiaotong University, Beijing, China
[2]Department of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China
State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
{Lifu_Wang, bshen, n_zhao, zhangzhiyuan}@bjtu.edu.cn

Abstract

The residual network is now one of the most effective structures in deep learning, which utilizes the skip connections to "guarantee" the performance will not get worse. However, the non-convexity of the neural network makes it unclear whether the skip connections do provably improve the learning ability since the nonlinearity may create many local minima. In some previous works [Freeman and Bruna, 2016], it is shown that despite the non-convexity, the loss landscape of the two-layer ReLU network has good properties when the number m of hidden nodes is very large. In this paper, we follow this line to study the topology (sub-level sets) of the loss landscape of deep ReLU neural networks with a skip connection and theoretically prove that the skip connection network inherits the good properties of the two-layer network and skip connections can help to control the connectedness of the sub-level sets, such that any local minima worse than the global minima of some two-layer ReLU network will be very "shallow". The "depth" of these local minima are at most $O(m^{(\eta-1)/n})$, where n is the input dimension, $\eta < 1$. This provides a theoretical explanation for the effectiveness of the skip connection in deep learning.

1 Introduction

Although deep learning has achieved great success in almost all the fields of machine learning, understanding the abilities of deep learning theoretically is still a hard problem. A neural network with a large number of hidden nodes has been proved to have strong expressive powers [Barron and A.R., 1993], but the non-convexity makes the model hard to be learned. The pioneering work in [Krizhevsky *et al.*, 2012] utilized ReLU to improve the performance of deep networks, but ReLU is insufficient to train very deep ones. Resnet [He *et al.*, 2016a; He *et al.*, 2016b] is the most efficient structure after Alexnet, which utilizes skip connections to let the performance not get worse as the number of the layers increasing, yet due to the non-convexity of the loss function, a rigorous analysis of this property is not easy. There are lots of questions about the effect of this structure. For example, will the skip connections eliminate bad local minima created by the nonlinearity in residual blocks?

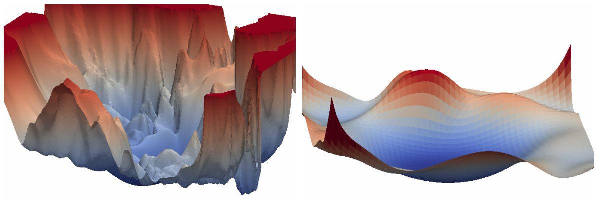
(a) Without skip connections (b) Skip connections network

Figure 1: The loss surfaces of ResNet-56 projected into 3d with/without skip connections.[1]

The problem of local minima is one of the most important questions in the theoretical study of neural networks. Gradient descent based algorithms will stop near the area of saddle points and local minima. By adding noise, it is possible to escape strictly saddle points and shallow local minima [Zhang *et al.*, 2017], but distinguishing a bad but deep local minimum from a global one can be very hard [Jin *et al.*, 2018]. If the loss of the suboptimal local minima created by the nonlinearity and multi-layer structures is very large, the performance of such networks may be very bad. However it has been shown that bad local minima are common in the non-over-parameterized two-layer networks [Safran and Shamir, 2018], so that it is generally hard to guarantee there is no bad local minima. Thus two questions arise naturally:

1.Since the loss in deep learning is not convex, what does the loss landscape of the deep neural network look like?

2.Since the residual network is now very successful, is the deep neural network with skip connections provably good such that more layers will not create more bad local minima?

There are some important works on these two questions. In the study of the landscape of the neural network, one of the most important properties is the "mode connectivity"[Kuditipudi *et al.*, 2019] of networks, which has been proved in

[*] Coresponding Author

[1] This figure is plotted using ParaView and the visualization method in [Li *et al.*, 2018]. Their code is available at https://github.com/tomgoldstein/loss-landscape.

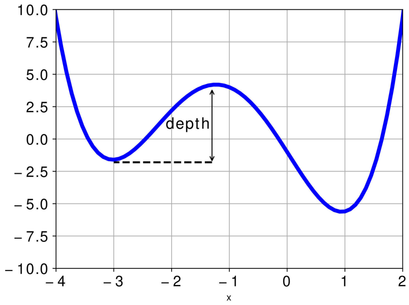

Figure 2: The depth of local minima (defined in Theorem 4)

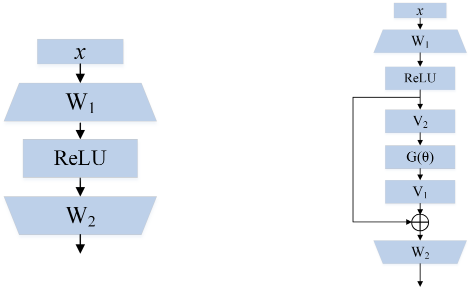

Figure 3: Two-layer network and skip connection network

theory in [Freeman and Bruna, 2016], and shown by experiments in [Garipov *et al.*, 2018; Draxler *et al.*, 2018]. In these works, they showed that the sub-level sets of the loss function are nearly connecting, and the local minima found by gradient descent can be connected by some continuous paths in the parameter space and the loss on the path is almost const, such that the landscape has very good properties. One can guess these properties are true for any neural network, but it is only proved for two-layer ReLU networks in [Freeman and Bruna, 2016], and it's hard to go beyond this case.

Another important step is taken by the work in [Shamir, 2018], where the author compares the two models:

$$y = \mathbf{W}^T(x + \mathbf{V}g(\theta, x)), \tag{1}$$

$$y = \mathbf{W}^T x. \tag{2}$$

It is obvious that (1) has stronger expressive powers than (2). Surprisingly, the work in [Shamir, 2018] shows that all the local minima(with MSE loss) created by $g(\theta, x)$ in (1) are provably no worse than the global minimum of convex model (2). However, their method is heavily dependent on the convexity of (2), thus it is hard to be generalized to more general cases. For an arbitrary neural network with skip connections, whether the skip connections can eliminate bad local minimum worse than shallower networks is still a open problem.

In the empirical aspect, the work in [Li *et al.*, 2018] proposed a visualization method and showed that the skip connection does make the landscape smoother, and the loss landscape has nearly convex behaviors. We plot the loss surface of Resnet-56 with and without skip connections in Figure 1, then we can see the ResNet-56 with skip connections does have smoother and better loss landscape. On the other hand, it is shown in [Veit *et al.*, 2016] that after removing the nonlinear parts in residual layer leaving the skip connection only, the performance will not drop too much. Thus one may guess that residual paths have ensemble-like behaviors. However, due to the non-convexity of the neural network, the principle is hard to be analyzed rigorously in theory.

In this work, we study these problems in the perspective of the sub-level sets of the loss landscape as in [Freeman and Bruna, 2016] to estimate the "depth" of local minima. In [Freeman and Bruna, 2016], the authors studied the two-layer ReLU network, and proved that the "Energy Gap" ε satisfies $\varepsilon \approx O(m^{-\frac{1}{n}})$ where m is the width and n is the dimension of the input data, therefore in the large width case, the landscape of the two-layer ReLU network has nearly convex behaviors. Following this line, in this paper, we compare the two networks (We use ξ to denote the parameters of the network, and σ is the ReLU activation function):

$$f_1(\xi) = \mathbf{W}_2[\sigma(\mathbf{W}_1 x) + \mathbf{V}_1 g(\theta, \mathbf{V}_2\sigma(\mathbf{W}_1 x))], \tag{3}$$

$$f_2(\xi) = \mathbf{W}_2\sigma(\mathbf{W}_1 x), \tag{4}$$

where $g(\theta, x)$ is a deep neural network with ReLU activation function. The form of f_1 is similar to the "pre-activation" Resnet in [He *et al.*, 2016b]. The structure of these two networks are showed in Figure 3.

We summarize our main results in the following theorem:

Theorem 1. *(informal)Supposing X, Y are bounded random variables, $L(\xi)$ is a convex function and $R(\xi)$ is the regularization term, and f_1 is defined as in (3), $\boldsymbol{W}_1 \in \mathbb{R}^{m\times n}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d_Y \times m}$, $0 < \eta < 1$, $l \le m^{\eta}$, $||\boldsymbol{W}||_0$ is the number of non-zero column vectors in $\boldsymbol{W}$,*

$$e(l) = \inf_{||\boldsymbol{W}_2||_0 \le l, ||\boldsymbol{W}_{1,i}||_2 = 1} \mathbb{E}\, L(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1 x), y) + \kappa\, ||\boldsymbol{W}_2||_1, \tag{5}$$

$$F_1(\xi) = \mathbb{E}\, L(f_1(\xi), y) + \kappa R(\xi), \tag{6}$$

for every strict local minimum F^* of F_1 with $F^* \ge e(l)$, the depth of it is at most $O(m^{\frac{\eta-1}{n}})$.

This result shows that for a suitable loss function, although f_1 is a multi-layer nonlinear network, by virtue of the skip connection, roughly all the local minima worse than the global minimum of the two-layer network f_2 are very "shallow". The depths of these local minima are controlled by $\varepsilon = O(m^{\frac{\eta-1}{n}})$, so that if m is very large, there is almost no bad strict local minima worse than $e(l)$. From the well known universal approximation theorem, the expressive power of the two-layer network with ReLU is very strong(this can be easily poved using Hahn-Banach theorem), such that $||\mathbf{W}_2^*\sigma(\mathbf{W}_1^* X) - Y||^2 \to 0$ as $l \to \infty$ for any functon $Y = f(X)$ under very mild conditions. So this result in fact describes the depth of nearly all the local minima if m is very large.

2 Related Work

The global geometry of the deep neural network loss surfaces has been widely concerned for a long time. The characteristics of deep learning are high dimension and non-convex, which make the model hard to be analyzed. The loss surface for deep linear networks was firstly studied in [Kawaguchi, 2016]. It is shown that all the local minima are global, and all the saddle points are strict. Although the expressive power of the deep linear network is the same as the single-layer one, the loss of the deep linear network is still non-convex. This work pointed out that the linear product of matrices will generally only create saddle points rather than bad local minima. The first rigorous and positive works on non-linear networks are in [Tian, 2017] and [Du *et al.*, 2018]. In these works, it is shown that, for a single-hidden-node ReLU network, under a very mild assumption on the input distribution, the loss is one point convex in a very large area. However, for the networks with multi-hidden nodes, the authors in [Safran and Shamir, 2018] pointed out that spurious local minima are common and indicated that an over-parameterization (the number of hidden nodes should be large) assumption is necessary. The loss surface of the two-layer over-parameterized network was studied in [Du and Lee, 2018] and [Mahdi *et al.*, 2018]. They showed that the over-parameterization helps two-layer networks to eliminate all the bad local minima, yet their methods required the unrealistic quadratic activation function (By using Hahn-Banach theorem, it is easy to show there are some functions which cannot be approximated by such networks). A different way to consider the landscape is in [Freeman and Bruna, 2016], which studied neural networks with ReLU and required the number of hidden nodes increases exponentially with the dimension of the input. They showed that if the number of the hidden nodes is large enough, the sub-level sets of the loss will be nearly connected.

The loss landscape of nonlinear skip connection networks was studied in [Shamir, 2018; Kawaguchi and Bengio, 2019; Yun *et al.*, 2019]. In these papers, it is shown that all the local minima created by the nonlinearity in the residual layer will never be worse than the linear model, yet their methods heavily rely on the convexity of $l(\mathbf{W}x)$ hence very hard to be generalized to more general residual networks. In our work, we focus on more realistic network structure $f = \mathbf{W}_2[\sigma(\mathbf{W}_1x) + \mathbf{V}_1g(\theta, \mathbf{V}_2\sigma(\mathbf{W}_1x))]$, and give a similar result as [Shamir, 2018] and extend the work on skip connection network [Shamir, 2018; Kawaguchi and Bengio, 2019] to more general non-linear cases. The structure we study is similar to the work in [Allenzhu and Li, 2019], but we focus on the global geometry of the loss rather than the gradient descent behaviors near neural tangent kernels area. The techniques we use are closely related with the theorems in [Freeman and Bruna, 2016] and our results are much stronger to apply to arbitrarily multilayered residual units with ReLU activation function and a large class of convex functions.

3 Preliminaries: Connectedness of Sub-level Sets and the Depth of Local Minima

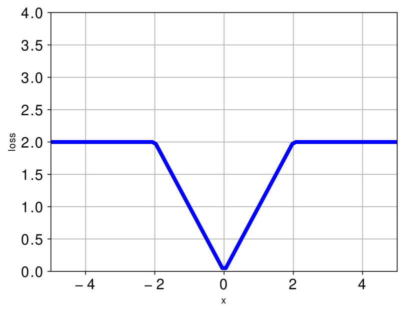

Figure 4: Loss function with non-strict local minima

The loss surface of the model is closely related to the solvability of the optimization problem, and the sub-level set method is a very important tool to study the loss landscape. We consider the risk minimization of the loss:

$$F(\xi) = \mathbb{E}_{X,Y\sim P}\, L(f(X,\xi),Y) + \kappa R(\xi). \tag{7}$$

The sub-level set of $F(\xi)$ is defined as:

$$\Omega_F(\lambda) = \{\xi; F(\xi) \le \lambda\}. \tag{8}$$

In the case that F is a convex function, we know that for any ξ_A, ξ_B, if $\xi(t) = (1-t)\xi_A + t\xi_B$, we have $F(\xi(t)) \le \max(F(\xi_A), F(\xi_B))$, so the sub-level sets for all λ are connected. If F is a function such that all the local minima are global(not need to be convex), for any ξ_A, ξ_B we can find a continuous path $\xi_1(t), \xi_2(t)$ with $F(\xi_1(t)), F(\xi_2(t))$ decreasing, then $\xi_1(0) = \xi_A$, $\xi_1(1) = \xi^*$, $\xi_2(0) = \xi_B$, $\xi_2(1) = \xi^*$, so that we can produce a path $\xi(t)$ with $F(\xi(t)) \le \max(F(\xi_A), F(\xi_B))$ by splicing the two paths together. Conversely, if the sub-level sets are connected, we can get some information about the strict local minima of the loss function:

Theorem 2. *(Proposition 2.1[Freeman and Bruna, 2016]) Supposing $\Omega_F(\lambda)$ is connected for all λ, any strict local minimum ξ^*, i.e. satisfying that there is a small disk $D = dist(\xi^*, \cdot) \le \varepsilon$ such that for all all the points $\xi' \in D$, $F(\xi') > F(\xi^*)$, is a global minimum.*

Note that this theorem cannot exclude the existence of bad non-strict local minima. Figure 4 is an example that all the sub-level sets are connected, but bad non-strict local minima exist.

In the case not all the sub-level sets are connected, sub-level sets still help us to understand the local minima. In fact we have:

Theorem 3. *Supposing $\Omega_F(\lambda)$ is connected for all $\lambda \ge \lambda_0$, all the strict local minima ξ^* satisfy $F(\xi^*) \le \lambda_0$*

The theorem can be proved by showing that there is a decreasing path from any ξ_A to ξ^* where $F(\xi^*) = \lambda_0$. From the condition of this theorem, there is a continuous path $\xi(t)$ from any ξ_A to ξ^* with $F(\xi(t)) \le F(\xi_A)$ and $F(\xi^*) = \lambda_0$. Supposing there is a $t_0 > 0$ such that on the path $F(\xi(t))$, the part $0 \le t \le t_0$ is decreasing, due to the fact $\Omega_F(F(\xi(t_0)))$ is connected, there must be a new path $\xi_1(t)$ such that $F(\xi_1(t)), 0 \le t \le t_0 + \varepsilon_0$ is decreasing, and this process can be extended continuously in this way.

There is also a theorem about the depth of local minima:

Theorem 4. *Suppose for any $\xi_A, \xi_B \in \Omega_F(\lambda)$, there is a smallest constant $\varepsilon > 0$ (the depth) and a continuous path $\xi(t)$ connecting ξ_A to ξ_B such that $F(\xi(t)) \leq \lambda + \varepsilon$. Then for any strict local minimum ξ^* , we have a value λ_1 and a continuous path $\xi_1(t)$ such that $\xi_1(t)$ connects it to a point ξ', with ξ' not belonging to the same connected component as ξ^* in $\Omega_F(\lambda_1)$ and $\max_t(F(\xi_1(t)) - F(\xi^*)) = \varepsilon$.*

This theorem can be proved by directly constructing such a path as in Theorem 3 from the conditions of this theorem. It is easy to see that if there is sub-level set $\Omega_F(\lambda_1)$ such that θ_A, θ_B are not in the same connected component, there is no decreasing path connecting the two points, so that ε is the depth, which measures the difficulty to jump out a local minimum.

Remark 3.1. *As in [Freeman and Bruna, 2016], the sub-level set is defined to be a closed, thus compact set. Under this definition, completely flat areas (c.f. Figure 4) will not influence the connectedness of such sub-level sets. And when we consider the connectedness of loss sub-level sets, it is sufficient to construct a path outside a zero-measure set. In fact, suppose there is a zero-measure set S. By adding small perturbation, there is a continuous path $\gamma(t)$ with $\gamma(t) \in \Theta \setminus S$ for almost all t, where $\Theta \setminus S$ is the parameter space outside S. We suppose $F(\gamma(t)) \leq \lambda + \varepsilon$ for almost all t and ε can be arbitrarily small. Due to the continuity of the F and $\bigcap_i(-\infty, \lambda+\varepsilon_i] = (-\infty, \lambda]$ where ε_i is a monotone decreasing sequence to 0, we have $F(\gamma(t)) \leq \lambda$ for all t and $\gamma(t)$ connects ξ_A and ξ_B.*

In the two-layer ReLU network case, the depth of the local minima is given in [Freeman and Bruna, 2016]:

Theorem 5. *(Theorem 2.4 in [Freeman and Bruna, 2016]) Consider the loss function $F(\boldsymbol{W}_1, \boldsymbol{W}_2) = \mathbb{E}_{X,Y\sim P}|Y - \boldsymbol{W}_2\sigma(\boldsymbol{W}_1 X)|^2$, where $X \in \mathbb{R}^n, Y \in \mathbb{R}, \boldsymbol{W}_1 \in \mathbb{R}^{m\times n}, \boldsymbol{W}_2 \in \mathbb{R}^{1\times m}$, and σ is the ReLU activation function. For any $\xi_A, \xi_B \in \Omega_F(\lambda)$, there is a continuous path $\xi(t)$ connecting ξ_A and ξ_B with $F(\xi(t)) \leq \max(\lambda, \varepsilon) + O(\alpha)$, where*

$$\varepsilon = \inf_{l.\alpha} \max(e(l), \delta_{W_1}(m,0,m), \delta_{W_1}(m-l,\alpha,m)),$$

$l = m^\eta, \alpha = m^{\frac{\eta-1}{n}}, \eta < 1$, $e(l)$ is the minimum approximation error using l hidden nodes, $\delta_{W_1}(m-l,\alpha,m) \sim O(\alpha)$, $\delta_{W_1}(m,0,m) \leq \lambda$.

In the two-layer ReLU case, this shows the depth of all the local minima worse than $e(l)$ is at most $O(m^{-\frac{1}{n}})$.

4 Warm up: One-layer Case

In this section, we consider the loss landscape in the linear case:

$$f(\mathbf{W}, \mathbf{V}, \theta, x) = \mathbf{W}(x + \mathbf{V}g(\theta, x)), \tag{9}$$

where $x \in \mathbb{R}^{d_x}, \mathbf{W} \in \mathbb{R}^{d_y\times d_x}, g(\theta,x) \in \mathbb{R}^{d_z}, \mathbf{V} \in \mathbb{R}^{d_x\times d_z}$. with loss

$$F(\mathbf{W}, \mathbf{V}, \theta) = \mathbb{E}_{x,y\sim P}\, L(f(\mathbf{W}, \mathbf{V}, \theta, x), y). \tag{10}$$

In the case y is a scalar and l is the MSE loss function, this has been studied in [Shamir, 2018]. The result is improved in [Kawaguchi and Bengio, 2019]. And under a weaker condition, we have a new theorem about the sub-level sets:

Theorem 6. *Supposing $w \to \mathbb{E}_{x,y}\, L(wx,y)$ is a function such that the sub-level sets are connected for all $w \in \mathbb{R}^{d_y\times d_x+d_z}$ and $\{x \in \mathbb{R}^{d_x+d_z}, y \in \mathbb{R}^{d_y}\}$, consider the input $\{x \in \mathbb{R}^{d_x}, y\}$ and two models:*

$$f_1(\boldsymbol{W}) = \boldsymbol{W}(x + \boldsymbol{V}g(\theta, x)), \tag{11}$$

$$f_2(\boldsymbol{W}) = \boldsymbol{W}x. \tag{12}$$

Let $F_1 = \mathbb{E}_{x,y\sim P}\, L(f_1(x), y)$, $F_2 = \mathbb{E}_{x,y\sim P}\, L(f_2(x), y)$. Assuming $d_y \leq \min\{d_x, d_z\}$, for any parameter θ_A, θ_B and $\lambda \in \mathbb{R}$ with $F(\theta_{\{A,B\}}) \leq \lambda$, there exists a continuous path $\gamma(t)$ such that $\gamma(0) = \theta_A$, $\gamma(1) = \theta_B$, and

$$F(\gamma(t)) \leq \max(\lambda, F_w^*)$$

where

$$F_w^* = \inf_W F_2(W).$$

Proof. From lemma 7, all the sub-level sets of $f(\mathbf{W}, \mathbf{V})$ are connected. So there is a path connecting it to $(\mathbf{W}^*, \mathbf{V} = 0, \theta)$ with the loss bounded by the endpoints. Note that $F_w^* = \inf_W F_2(W) = \mathbf{W}^*(x + \mathbf{0}g(\theta, x))$, our claim follows. □

Lemma 7. *For any distribution $\{x \in \mathbb{R}^{d_x+d_z}, y \in \mathbb{R}^{d_y}\} \sim P$ and $d_y \leq \min\{d_x, d_z\}$, supposing all the sub-level sets of function $F(\mathbf{Z}) = \mathbb{E}_{x,y}\, L(\mathbf{Z}x, y)$ are connected, the sub-level sets of function $F(\boldsymbol{W}, \boldsymbol{V}) = \mathbb{E}_{x,y}\, L([\boldsymbol{W}, \boldsymbol{W}\boldsymbol{V}]x, y)$ are also connected.*

Proof. Let $F(\mathbf{Z}) = \mathbb{E}_{x,y}\ L(\mathbf{Z}x, y)$. For any $\mathbf{Z}_1, \mathbf{Z}_2$, since the sub-level sets are connected, there is a continuous path $\mathbf{Z}(t) \in \mathbb{R}^{d_y\times(d_x+d_z)}$ such that $\mathbf{Z}(0) = \mathbf{Z}_1$ and $\mathbf{Z}(1) = \mathbf{Z}_2$, with $F(\mathbf{Z}(t)) \leq \max(F(\mathbf{Z}_1), F(\mathbf{Z}_2))$. To prove the theorem, we need a path with $\mathbf{Z}(0) = [\mathbf{W}_1, \mathbf{W}_1\mathbf{V}_1], \mathbf{Z}(1) = [\mathbf{W}_2, \mathbf{W}_2\mathbf{V}_2]$ and $\mathbf{Z}(t) = [\mathbf{W}(t), \mathbf{W}(t)\mathbf{V}(t)]$.

Note that $d_y \leq \min\{d_x, d_z\}$ so the sets $rank(\mathbf{W}(t)) \neq d_y$ have zero measure(since they are closed in Zariski topology). Following the discussion in Remark 3.1, we only need to prove in the case $rank(\mathbf{W}(t)) = d_y$. Let $\mathbf{Z}_b(t)$ be the last d_z columns of $\mathbf{Z}(t)$. We set $\mathbf{V}(t) = \mathbf{W}^+(t)\mathbf{Z}_b(t)$, where $\mathbf{W}^+(t)$ is the Moore-Penrose pseudoinverse of $\mathbf{W}(t)$, then $\mathbf{V}(t)$ is continuous and $\mathbf{W}(t), \mathbf{V}(t)$ is the required path. □

Remark 4.1. *In the case $d_y > \min\{d_x, d_z\}$, $\boldsymbol{W}\boldsymbol{V}$ will always be a low rank matrix, so this proof is invalid. However, if the loss function is convex and the eigenvalues of the Hessian matrix are bounded by a and b with $\frac{a}{b} - 1$ small, the sub-level sets will still be connected. This can be proved using the methods in [Barber and Ha, 2018; Ha* et al.*, 2018]. Since this is not the main target in this paper, we omit it.*

5 Main Results

5.1 Assumptions and Preliminary Lemmas

Assumption 1. *$L(w, y)$ is a convex function for w, $(x, y) \sim P$ are bounded, and $R(\xi)$ is the regularization term. There is a* *constant C such that all the strict local minima of the loss*

$$\begin{aligned} F(\xi) =& \mathbb{E}_{x,y\sim P}\, L(\boldsymbol{W}_2[\sigma(\boldsymbol{W}_1 x) + \boldsymbol{V}_1 g(\theta, \boldsymbol{V}_2\sigma(\boldsymbol{W}_1 x))], y) \\ & + \kappa R(\xi) \\ =& \mathbb{E}_{x,y\sim P}\, L(\boldsymbol{W}_2[\sigma(\boldsymbol{W}_1 x) + \boldsymbol{V}_1 g(\theta, \boldsymbol{V}_2\sigma(\boldsymbol{W}_1 x))], y) \\ & + \kappa(\sum_i (||w_{2,i}||_1 + ||v_{2,i}||_1)||w_{1,i}||_2 + ||\boldsymbol{W}_2\boldsymbol{V}_1||_1 \\ & + \sum_i ||\theta_i||_F), \end{aligned} \tag{13}$$

satisfying

$$\begin{aligned} &\max(\sum_i ||w_{2,i}||_1||w_{1,i}||_2, \sum_i ||v_{2,i}||_1||w_{1,i}||_2, \\ &||\boldsymbol{W}_2\boldsymbol{V}_1||_1, ||\theta_i||_F) \leq C \end{aligned} \tag{14}$$

where $w_{2,i}$, $v_{2,i}$ are the ith column vector of $\boldsymbol{W}_2$ and $\boldsymbol{V}_2$, $w_{1,i}$ is the ith row vector of $\boldsymbol{W}_1$.

Assumption 2. *L is locally Lipschitzian:*

$$|L(x_1, y) - L(x_2, y)| \leq L_0||x_1 - x_2||_F, \tag{15}$$

when $||x_1||_F, ||x_2||_F \leq C$.

An example satisfying these assumptions is the MSE loss. In fact we have:

Theorem 8. *Suppose $g(\theta, x)$ is an arbitrary multi-layer ReLU network, with the loss*

$$\begin{aligned} F(\xi) =& \mathbb{E}_{x,y\sim P}\, ||\boldsymbol{W}_2[\sigma(\boldsymbol{W}_1 x) + \boldsymbol{V}_1 g(\theta, \boldsymbol{V}_2\sigma(\boldsymbol{W}_1 x))] - y||_F^2 \\ & + \kappa R(\xi), \end{aligned} \tag{16}$$

For any points with $\nabla F = 0$, there is a constant C such that these assumptions are satisfied.

Proof. It is trivial that Assumption 2 is satisfied. We only need to prove Assumption 1. In this case, all the activation functions in this network are ReLU, so that we can write all the matrix parameter $\mathbf{W}$ as $\mathbf{W} = t_W\mathbf{E}_W$. where $||\mathbf{E}_W||_1 = 1$ and $t_W = ||\mathbf{W}||_1$, $\sigma(t_W\mathbf{E}_W) = t_W\sigma(\mathbf{E}_W)$. We fix all $\mathbf{E}_W$ and the loss has the form (We only need to consider the case $d_y = 1$ since if $d_y > 1$, it can be reduced to $\sum_i |y_i - x_i|^2$):

$$\begin{aligned} F(t,\theta) = \mathbb{E}_{x,y\sim P}\, &|(t + tv_1v_2\prod_i \theta_i)wx - y|^2 \\ &+ \kappa(|tw| + \sum_i |\theta_i| + |tv_1| + |tv_2|), \end{aligned} \tag{17}$$

where t, w, θ_i are the corresponding variables. We have:

$$\begin{aligned} t\nabla_t F =& 2\mathbb{E}_{x,y\sim P}\, [(t + tv_1v_2\prod_i \theta_i)wx - y][(t + tv_1v_2\prod_i \theta_i)wx] \\ & + \kappa|tw| + \kappa|tv_1| + \kappa|tv_2| \\ =& 0. \end{aligned} \tag{18}$$

Note that $|tw| + |tv_1| + |tv_2| > 0$, $\mathbb{E}_x\, [(t + tv_1v_2\prod_i \theta_i)wx - y][(t + tv_1v_2\prod_i \theta_i)wx] < 0$ We have $\mathbb{E}\, [(t + tv_1v_2\prod_i \theta_i)wx]^2 \leq \mathbb{E}\, |[(t + tv_1v_2\prod_i \theta_i)wx]y|$ so $\kappa(|tw| + |tv_1| + |tv_2|) \sim \mathbb{E}\, |(t + tv_1v_2\prod_i \theta_i)wx| \sim O(\mathbb{E}|y|)$. This proof also applies to other variables, so our claim follows.

□

Before proving the main theorem, we need two key lemmas from [Freeman and Bruna, 2016]:

Lemma 9. *Considering a matrix $W \in \mathbb{R}^{m\times n}$, which is equal to give m vectors, and $0 < \eta \leq 1$, there is a a collection Q_m of at last m^η vectors such that for any $v_1, v_2 \in Q_m$, $\angle v_1, v_2 \leq 2\varepsilon_{m,\eta} = 2m^{\frac{\eta-1}{n}}$.*

Lemma 10. *Given w_1, w_2 with $||w_1|| = ||w_2|| = 1$, $\angle w_1, w_2 \leq \alpha$, and σ is the ReLU activation function, we have $\mathbb{E}_x||\sigma(w_1x) - \sigma(w_2x)||^2 \leq 4||\Sigma_x||\alpha^2$, where $\Sigma_X = \mathbb{E}_{X\sim P}\, XX^T \in \mathbb{R}^{n\times n}$*

These lemmas are from the proof of corollary 2.5 and proposition 2.3 in [Freeman and Bruna, 2016] respectively.

5.2 Main Theorem

Theorem 11. *Consider a distribution $\{x \in \mathbb{R}^n, y \in \mathbb{R}^{d_y}\} \sim P$, σ the ReLU activation function, and a neural network with a skip connection:*

$$f(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{V}_1, \boldsymbol{V}_2, \theta, x) = \boldsymbol{W}_2[\sigma(\boldsymbol{W}_1 x) + \boldsymbol{V}_1 g(\theta, \boldsymbol{V}_2\sigma(\boldsymbol{W}_1 x))], \tag{19}$$

with L a function satisfying assumption 1 and 2. Assume $g(\theta, x)$ is a neural network with ReLU activation functions and $R(\xi)$ is the regular term with $R(\xi) = \sum_i ||w_{2,i}||_1||w_{1,i}||_2 + \sum_i ||v_{2,i}||_1||w_{1,i}||_2 + ||\boldsymbol{W}_2\boldsymbol{V}_1||_1 + \sum_i ||\theta_i||_F$. Let

$$F(\xi) = \mathbb{E}_{x,y\sim P}\, L(f(\xi, x), y) + \kappa R(\xi),$$

then we have: For any $\eta < 1, l < m^\eta$, ξ_A, ξ_B and $\lambda \in R$ satisfying $F(\xi_{\{A,B\}}) \leq \lambda$, there exists a continuous path $\gamma(t)$ such that $\gamma(0) = \xi_A$, $\gamma(1) = \xi_B$, and

$$F(\gamma(t)) \leq \max(F_w^*, \lambda) + O(m^{\frac{\eta-1}{n}}),$$

where m is the dimension of $\boldsymbol{W}_1 \in R^{m\times n}$ and

$$F_w^* = \inf_{||W_{1,i}||_2=1, ||W_2||_0 \leq l} \mathbb{E}_{X,Y\sim P}\, l(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1 X), Y) + \kappa||\boldsymbol{W}_2||_1. \tag{20}$$

Remark 5.1. *Our regularization term is chosen to be compatible with the loss function, and it is also compatible with the two-layer linear network. It can be replaced by using good initialization [Glorot and Bengio, 2010] or if we only consider a specific bounded area.*

Proof. Supposing $\mathbf{W}_1 \in \mathbb{R}^{m\times n}, \mathbf{W}_2 \in \mathbb{R}^{d_y\times m}, \mathbf{V}_2 \in \mathbb{R}^{d_g\times m}, \mathbf{V}_1 \in \mathbb{R}^{n\times d_o}$, we need to construct a path $\gamma(t)$ from $(\mathbf{W}_{1,A}, \mathbf{W}_{2,A}, \mathbf{V}_{1,A}, \mathbf{V}_{2,A}, \theta_A)$ to $(\mathbf{W}_{1,B}, \mathbf{W}_{2,B}, \mathbf{V}_{1,B}, \mathbf{V}_{2,B}, \theta_B)$. Note that we only need to construct a path $\gamma_1(t)$ from any $(\mathbf{W}_{1,A}, \mathbf{W}_{2,A}, \mathbf{V}_{1,A}, \mathbf{V}_{2,A}, \theta_A)$ to $(\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{V}_1^*, \mathbf{V}_2^*, \theta^*)$ with $F(\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{V}_1^*, \mathbf{V}_2^*, \theta^*) = F^*$ and show that $F(\gamma_1(t)) \leq \max(F(\gamma_1(0)), F^*) + O(m^{\frac{\eta-1}{n}})$ because the second half of the path $\gamma(t)$ is the inverse of the first half. So we need to construct the following parts:

1.$(\mathbf{W}_{1,A}, \mathbf{W}_{2,A}, \mathbf{V}_{1,A}, \mathbf{V}_{2,A}, \theta_A)$ to $(\mathbf{W}_{1,l}, \mathbf{W}_{2,l}, \mathbf{V}_{1,l}, \mathbf{V}_{2,l}, \theta_l)$.

On this path, the norms of all matrices are reduced without increasing the loss in virtue of the regularization term, such that the $g(\theta, x)$ and $||\mathbf{W}_2\mathbf{V}_1||$ are bounded.

2. $(\mathbf{W}_{1,l}, \mathbf{W}_{2,l}, \mathbf{V}_{1,l}, \mathbf{V}_{2,l}, \theta_l)$ to $(\mathbf{W}_{1,l}, \mathbf{W}_{2,s}, \mathbf{V}_{1,s}, \mathbf{V}_{2,s}, \theta_l)$.

On this path, $\mathbf{W}_{2,s}$ is a $(m - l)$-term approximation using perturbed atoms to minimize $\mathbb{E}_x||\mathbf{W}_{2_s}\sigma(\mathbf{W}_{1,l}x) -$

$\mathbf{W}_{2,l}\sigma(\mathbf{W}_{1,l}x)||_F$, and $\mathbf{V}_{2,s}$ is a $(m-l)$-term approximation to minimize $\mathbb{E}_x||\mathbf{V}_{2,s}\sigma(\mathbf{W}_{1,l}x) - \mathbf{V}_{2,l}\sigma(\mathbf{W}_{1,l}x)||_F^2$. The loss increasing along this path is roughly bounded by $O(m^{\frac{\eta-1}{n}})$.

3. $(\mathbf{W}_{1,l}, \mathbf{W}_{2,s}, \mathbf{V}_{1,s}, \mathbf{V}_{2,s}, \theta_l)$ to $(\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{0}, \mathbf{0}, \mathbf{0})$.

On this path, $(\mathbf{W}_1^*, \mathbf{W}_2^*)$ are the parameters of l-term approximation:

$$\begin{aligned}(\mathbf{W}_1^*, \mathbf{W}_2^*) = argmin_{||w_{1,i}||_2=1, ||\mathbf{W}_2||_0 \le l} \\ \mathbb{E}_{x,y\sim P}\, L(Y, \mathbf{W}_2\sigma(\mathbf{W}_1 x)) + \kappa||\mathbf{W}_2||_1.\end{aligned} \tag{21}$$

$\mathbf{W}_{2,s}$ is $m-l$ sparse with l zero columns and $\mathbf{W}_2^*$ has no-zero values only on these l columns. The loss along that path will be upper bounded by λ and $e(l)$.

The construction of these path is described as follow:

Step 1: Consider the loss:

$$\begin{aligned}\mathbb{E}\, L(Y, \mathbf{W}_1[\sigma(\mathbf{W}_1 x) + \mathbf{V}_1 g(\theta, \mathbf{V}_2\sigma(\mathbf{W}_1 x), x)]) \\ + R(\theta, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V}_1, \mathbf{V}_2).\end{aligned} \tag{22}$$

Since the Assumption 1 is satisfied, there is a constant C independent of m and a continuous path without increasing the loss such that

$$\begin{aligned}\max(\sum_i ||w_{2,i}||_1 ||w_{1,i}||_2, \sum_i ||v_{2,i}||_1 ||w_{1,i}||_2, \\ ||\mathbf{W}_2\mathbf{V}_1||_1, \sum_i ||\theta_i||_F) \le C,\end{aligned} \tag{23}$$

and $\forall i, ||w_{1,i}||_2 = 1$ where $w_{1,i}$ is the ith row vector of $\mathbf{W}_1$. Thus $||g(\theta, x)||_F$ will be G_0 Lipschitzian for x. Then fix $\mathbf{W}_1$, θ and $\mathbf{V}_2$, and consider the path $\mathbf{W}_2(t), \mathbf{V}_1(t)$ to the nearest local minimum. The loss will not increase on this path.

Step 2: The $m-l$ spare parameter matrix $\mathbf{W}_{2,s}$ and $\mathbf{V}_{2,s}$ are constructed as follow: Find a set Q_m of row vectors in $\mathbf{W}_{1,l}$ as in Lemma 9 and select a vector v at the jth row such that for any $v_i \in Q_m$ at the ith row of W_1, $\angle v_i, v \le 2\varepsilon_{m,\eta}$. $\mathbf{W}_{2,s}$ is constrcuted by setting the rows corresponding to the vectors in Q_m to be zero, and for the corresponding column of $W_{2,s,i} = 0, W_{2,s,j} = \sum_i W_{2,l,i} + W_{2,l,j}$. The construction of $\mathbf{V}_{2,s}$ is similar. Consider the paths $\mathbf{W}_2(t)$,$\mathbf{V}_2(t)$ constructed as follow:

$$\mathbf{W}_2(t) = [\alpha_1, \alpha_2 ... \bar{\alpha}_{i_1}, \alpha_{i_1+1} ... \widetilde{\alpha}_j \bar{\alpha}_{i_2} ...],$$

$$\mathbf{V}_2(t) = [\beta_1, \beta_2 ... \bar{\beta}_{i_1}, \beta_{i_1+1} ... \widetilde{\beta}_j \bar{\beta}_{i_2} ...].$$

where $i_1, i_2 ...$ are the indexes corresponding to the rows of vectors in Q_m and $\bar{\alpha}_j, \bar{\beta}_j$ are the ones corresponding to the rows of vectors in the jth row. We have $\bar{\alpha}_{i_1} = (1-t)\alpha_{i_1}$, $\bar{\beta}_{i_1} = (1-t)\beta_{i_1}$ and $\widetilde{\alpha}_j = \alpha_j + \sum_k t\alpha_{i_k}$, $\widetilde{\beta}_j = \beta_j + \sum_k t\beta_{i_k}$. This process will not increase the regularization loss.

For $i \in Q_m$, $j \in Q_2$, let $\sigma(\mathbf{W}_{1,l}x)_i = \sigma(\mathbf{W}_{1,l}x)_j + n_i$.

$$\begin{aligned}||\mathbf{W}_2(t)\sigma(\mathbf{W}_{1,l}x) - \mathbf{W}_{2,s}\sigma(\mathbf{W}_{1,l}x)||_F \le ||\sum_k \alpha_{i_k} n_{i_k}||_F \\ ||\mathbf{V}_2(t)\sigma(\mathbf{W}_{1,l}x) - \mathbf{V}_{2,s}\sigma(\mathbf{W}_{1,l}x)||_F \le ||\sum_k \beta_{i_k} n_{i_k}||_F.\end{aligned} \tag{24}$$

Let $\mathbf{V}_1(t) = \mathbf{W}_2^+(t)\mathbf{W}_2(0)\mathbf{V}_1(0)$, where $\mathbf{W}_2^+(t)$ is the Moore-Penrose pseudoinverse. Note that the set $rank\ \mathbf{W}_2 \ne d_y$ has zero measure. We only need to consider the case $rank\ \mathbf{W}_2(t) = d_y$, then $\mathbf{W}_2(t)\mathbf{V}_1(t) = \mathbf{W}_2(0)\mathbf{V}_1(0)$. To estimate the loss, note that L is locally Lipschitzian, we have:

$$\begin{aligned}&\mathbb{E}\ |L(f_1(\mathbf{V}_1(t), \mathbf{V}_2(t), \mathbf{W}_2(t), x), y) \\ &- L(f_1(\mathbf{V}_1(0), \mathbf{V}_2(0), \mathbf{W}_2(0), x), y)| + \kappa|R(\xi(t)) - R(\xi(0))| \\ &\le \mathbb{E}\ L_0||\sum_k \alpha_{i_k} n_{i_k}||_F + L_0||\mathbf{W}_2(0)\mathbf{V}_1(0)||G_0||\sum_k \beta_{i_k} n_{i_k}||_F \\ &\sim O(\mathbb{E}_{x,y}\ |n|) \sim O(\varepsilon_{m,\eta}).\end{aligned} \tag{25}$$

Step 3: $\mathbf{W}_{2,s}$ and $\mathbf{V}_{2,s}$ are $m-l$ sparse, so changing the l rows in $\mathbf{W}_{1,l}$ will not influence the loss. We consider a path to change these l rows to be same as $\mathbf{W}_1^*$ and the loss on this path is constant. The second step is to construct a path $(\mathbf{W}_2(t), \mathbf{W}_1(t), \mathbf{V}_1(t))$ with $\mathbf{W}_2(0) = \mathbf{W}_{2,s}, \mathbf{W}_2(1) = \mathbf{W}_2^*$, $\mathbf{W}_2(1)\mathbf{V}_1(1) = 0$. As the proof of Lemma 7, since the loss is convex for the final layer, the loss on this path is bounded by the two endpoints, and note that:

$$\begin{aligned}&F(\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{V}_1^*, \mathbf{V}_2^*, \theta^*) = F(\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{V}_1^* = \mathbf{0}, \mathbf{V}_2^* = \mathbf{0}, \theta^* = \mathbf{0}) \\ &\le \inf_{||w_{1,i}||_2=1, ||\mathbf{W}_2||_0 \le l} \mathbb{E}_{X,Y\sim P}\ L(Y, \mathbf{W}_2\sigma(\mathbf{W}_1 X)) + \kappa||\mathbf{W}_2||_1.\end{aligned} \tag{26}$$

The properties are satisfied. □

5.3 Discussion

Theorem 11 is a generalization of Theorem 5 in [Freeman and Bruna, 2016] and the linear case Theorem 6. As pointed out in section 3, this shows that for all local minima worse than the global minimum F^* of two-layer networks with $l = m^\eta$ hidden nodes, the depth is bounded by $O(m^{\frac{\eta-1}{n}})$, so that as $m \to \infty$, $\Omega_F(\lambda)$ is nearly connected if $\lambda > F^*$. Benefitting from the strong expressiveness of the two-layer ReLU network, for almost all the learning problem, F^* in this theorem will be much better than that in the linear case Theorem 6.

6 Conclusion

In this paper, we studied the loss landscape of the multi-layer nonlinear neural network with a skip connection and consider the connectedness of sub-level sets by constructing a path and estimating the loss on this path. The main theorem reveals that by virtue of the skip connection, under mild conditions all the local minima worse than the global minimum of the two-layer ReLU network will be very shallow, such that the "depths" of these local minima are at most $O(m^{\frac{\eta-1}{n}})$, where $\eta < 1$, m is the number of the hidden nodes in the first layer, and n is the dimension of the input data. This result shows that despite the non-convexity of the non-linear networks, skip connections provably help to reform the loss landscape, and in the over-parametrization($m \to \infty$) case, nearly all the strict local minima are no worse than the global minimum of the two-layer ones. Our results provide a theoretical explanation of the effectiveness of the skip connections and take a step to understand the mysterious effectiveness of deep learning networks.

Acknowledgments

The authors thank Wenyu Zhang for his great help. This research was funded by the National Key Research and Development Program of China, grant number 2018YFC0831300.

References

[Allenzhu and Li, 2019] Zeyuan Allenzhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*, pages 9015–9025, 2019.

[Barber and Ha, 2018] Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7(4):755–806, 2018.

[Barron and A.R., 1993] Barron and A.R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3), 1993.

[Draxler *et al.*, 2018] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A Hamprecht. Essentially no barriers in neural network energy landscape. *International conference on machine learning*, pages 1308–1317, 2018.

[Du and Lee, 2018] Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *International conference on machine learning*, pages 1328–1337, 2018.

[Du *et al.*, 2018] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn. *International conference on machine learning*, 2018.

[Freeman and Bruna, 2016] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *International conference on machine learning*, 2016.

[Garipov *et al.*, 2018] Timur Garipov, Pavel Izmailov, D-mitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, pages 8789–8798, 2018.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[Ha *et al.*, 2018] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between stationary points for rank constraints versus low-rank factorizations. *arXiv: Optimization and Control*, 2018.

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

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *European conference on computer vision*, pages 630–645, 2016.

[Jin *et al.*, 2018] Chi Jin, Lydia T Liu, Rong Ge, and Michael I Jordan. On the local minima of the empirical risk. *Advances in neural information processing systems*, pages 4896–4905, 2018.

[Kawaguchi and Bengio, 2019] Kenji Kawaguchi and Yoshua Bengio. Depth with nonlinearity creates no bad local minima in resnets. *Neural Networks*, 118:167–174, 2019.

[Kawaguchi, 2016] Kenji Kawaguchi. Deep learning without poor local minima. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 586–594. Curran Associates, Inc., 2016.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(2), 2012.

[Kuditipudi *et al.*, 2019] Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in Neural Information Processing Systems*, pages 14574–14583, 2019.

[Li *et al.*, 2018] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems 31*, pages 6389–6399. Curran Associates, Inc., 2018.

[Mahdi *et al.*, 2018] Soltanolkotabi Mahdi, Javanmard Adel, and Lee Jason D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

[Safran and Shamir, 2018] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *International conference on machine learning*, pages 4430–4438, 2018.

[Shamir, 2018] Ohad Shamir. Are resnets provably better than linear predictors. *Advances in neural information processing systems*, pages 507–516, 2018.

[Tian, 2017] Yuandong Tian. Symmetry-breaking convergence analysis of certain two-layered neural networks with relu nonlinearity. *International conference on learning representations*, 2017.

[Veit *et al.*, 2016] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in Neural Information Processing Systems*, pages 550–558, 2016.

[Yun *et al.*, 2019] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Are deep resnets provably better than linear predictors. *Advances in Neural Information Processing Systems*, 2019.

[Zhang *et al.*, 2017] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. *Conference on Learning Theory*, 2017.