On Connected Sublevel Sets in Deep Learning

Quynh Nguyen [1]

Abstract

This paper shows that every sublevel set of the loss function of a class of deep over-parameterized neural nets with piecewise linear activation functions is connected and unbounded. This implies that the loss has no bad local valleys and all of its global minima are connected within a unique and potentially very large global valley.

1. Introduction

It has been commonly observed in deep learning that over-parameterization can be helpful for optimizing deep neural networks. In particular, several recent work (Allen-Zhu et al., 2018b; Du et al., 2018; Zou et al., 2018) have shown that if "all the hidden layers" of a deep network have polynomially large number of neurons compared to the number of training samples and the network depth, then (stochastic) gradient descent converges to a global minimum with zero training error. While these theoretical guarantees are interesting conceptually, it remains largely unclear why this kind of simple local search algorithms can succeed given the non-convexity and NP-Hardness (worst-case) of the problem. We are instead interested in the following question:

Is there any underlying geometric structure of the loss function that can "intuitively" support for the success of (stochastic) gradient descent under excessive over-parameterization regimes?

In this paper, we shed light on this question by showing that every sublevel set of the loss is connected if "one of the hidden layers" is wide enough. Our key idea is first to prove that this property holds in general for linearly independent training data, and then extend such result to "arbitrary data" by using an additional condition on the wide layer. In particular, we first show that if one of the hidden layers has more neurons than the number of training samples, then the loss function has no bad local valleys in the sense that there is a continuous path from any starting point in parameter space on which the loss is non-increasing and gets arbitrarily close to its infimum (global minimum). Second, if the first hidden layer of the network has twice more neurons than the number of training samples, then we show that every sublevel set of the loss becomes connected. This is a stronger guarantee than before as it not only implies that the loss function does not contain any bad local valleys where gradient descent might get stuck, but also that all finite global minima (whenever they exist) are connected within a unique global valley. Our results hold for standard deep fully connected networks with arbitrary convex losses and piecewise linear activation functions. All missing proofs are moved to the appendix.

2. Background

Let N be the number of training samples and $X = [x_1, \ldots, x_N]^T \in \mathbb{R}^{N\times d}$ the training data with $x_i \in \mathbb{R}^d$. Let L be the number of layers of the network, n_k the number of neurons at layer k, d the input dimension, m the output dimension, and $W_k \in \mathbb{R}^{n_{k-1}\times n_k}$ and $b_k \in \mathbb{R}^{n_k}$ the weight matrix and biases respectively of layer k. By convention, we assume that $n_0 = d$ and $n_L = m$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous activation function specified later. The network output at layer k is the matrix $F_k \in \mathbb{R}^{N\times n_k}$ defined as

$$F_k = \begin{cases} X & k = 0 \\ \sigma\big(F_{k-1}W_k + \mathbf{1}_N b_k^T\big) & 1 \le k \le L-1 \\ F_{L-1}W_L + \mathbf{1}_N b_L^T & k = L \end{cases} \tag{1}$$

Let $\theta := (W_l, b_l)_{l=1}^L$ be the set of all parameters. Let Ω_l be the space of (W_l, b_l) for every layer $l \in [1, L]$, and $\Omega = \Omega_1 \times \ldots \times \Omega_L$ the whole parameter space. Let $\Omega_l^* \subset \Omega_l$ be the subset of parameters of layer l for which the corresponding weight matrix has full rank, that is $\Omega_l^* = \{(W_l, b_l) \mid W_l \text{ has full rank}\}$. In this paper, we often write $F_k(\theta)$ to denote the network output at layer k as a function θ, but sometimes we drop the argument θ and write just F_k if it is clear from the context. We also use the notations $F_k\Big((W_1, b_1), \ldots, (W_L, b_L)\Big), F_k\Big((W_1, b_1), (W_l, b_l)_{l=2}^L\Big)$. The training loss $\Phi : \Omega \to \mathbb{R}$ is defined as

$$\Phi(\theta) = \varphi(F_L(\theta)) \tag{2}$$

where $\varphi : \mathbb{R}^{N\times m} \to \mathbb{R}$ is assumed to be any convex loss. Typical examples include the standard cross-entropy loss

[1]Department of Mathematics and Computer Science, Saarland University, Germany. Correspondence to: Quynh Nguyen <quynh@cs.uni-saarland.de>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

$\varphi(F_L) = \frac{1}{N}\sum_{i=1}^N -\log\left(\frac{e^{(F_L)_{iy_i}}}{\sum_{k=1}^m e^{(F_L)_{ik}}}\right)$ where y_i is the ground-truth class of x_i, and the standard square loss for regression $\varphi(F_L) = \frac{1}{2}\left\|F_L - Y\right\|_F^2$ where $Y \in \mathbb{R}^{N\times m}$ is a given training output.

In this paper, we denote $p^* = \inf_{G\in\mathbb{R}^{N\times m}} \varphi(G)$ which serves as a lower bound on Φ. Note that p^* is fully determined by the choice of $\varphi(\cdot)$ and thus is independent of the training data. Please also note that we make no assumption about finiteness of p^* in this paper although for most of practical loss functions as mentioned above one has $p^* = 0$. We list below several assumptions on the activation function and will refer to them accordingly in our different results.

Assumption 2.1 *σ is strictly monotonic and $\sigma(\mathbb{R}) = \mathbb{R}$.*

Note that Assumption 2.1 implies that σ has a continuous inverse $\sigma^{-1} : \mathbb{R} \to \mathbb{R}$, which is satisfied for Leaky-ReLU.

Assumption 2.2 *There do not exist non-zero coefficients $(\lambda_i, a_i)_{i=1}^p$ with $a_i \neq a_j \,\forall\, i \neq j$ such that $\sigma(x) = \sum_{i=1}^p \lambda_i \sigma(x - a_i)$ for every $x \in \mathbb{R}$.*

Assumption 2.2 is satisfied for every piecewise linear activation functions except the linear one as shown below.

Lemma 2.3 *Assumption 2.2 is satisfied for any continuous piecewise linear activation function with at least two pieces such as ReLU, Leaky-ReLU, etc and for the exponential linear unit $\sigma(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$ where $\alpha > 0$.*

Through out the rest of this paper, we will make the following mild assumption on our training data.

Assumption 2.4 *All the training samples are distinct.*

A key concept of this paper is the sublevel set of a function.

Definition 2.5 *For every $\alpha \in \mathbb{R}$, the α-level set of $\Phi : \Omega \to \mathbb{R}$ is the preimage $\Phi^{-1}(\alpha) = \{\theta \in \Omega \mid \Phi(\theta) = \alpha\}$, and the α-sublevel set of Φ is given as $L_\alpha = \{\theta \in \Omega \mid \Phi(\theta) \leq \alpha\}$.*

Below we recall the standard definition of connected sets and some basic properties which are used in this paper.

Definition 2.6 (Connected set) *A subset $S \subseteq \mathbb{R}^d$ is called connected if for every $x, y \in S$, there exists a continuous curve $r : [0, 1] \to S$ such that $r(0) = x$ and $r(1) = y$.*

Proposition 2.7 *Let $f : U \to V$ be a continuous map. If $A \subseteq U$ is connected then $f(A) \subseteq V$ is also connected.*

Proposition 2.8 *The Minkowski sum of two connected subsets $U, V \subseteq \mathbb{R}^n$, defined as $U + V = \{u + v \mid u \in U, v \in V\}$, is a connected set.*

In this paper, $A^\dagger$ denotes the Moore-Penrose inverse of A. If A has full row rank then it has a right inverse $A^\dagger = A^T(AA^T)^{-1}$ with $AA^\dagger = \mathbb{I}$, and if A has full column rank then it has a left inverse $A^\dagger = (A^TA)^{-1}A^T$ with $A^\dagger A = \mathbb{I}$.

3. Key Result: Linearly Independent Data Leads to Connected Sublevel Sets

This section presents our key results for linearly independent data, which form the basis for our additional results in the next sections where we analyze deep over-parameterized networks with arbitrary data. Below we assume that the widths of the hidden layers are decreasing, i.e. $n_1 > \ldots > n_L$. Note that it is still possible to have $n_1 \geq d$ or $n_1 < d$. The above condition is quite natural as in practice (e.g., see Table 1 in (Nguyen & Hein, 2018)) the first hidden layer often has the most number of neurons, afterwards the number of neurons starts decreasing towards the output layer, which is helpful for the network to learn more compact representations at higher layers. We introduce the following property for a class of points $\theta = (W_l, b_l)_{l=1}^L$ in parameter space and refer to it later in our results and proofs.

Property 3.1 *W_l has full rank for every $l \in [2, L]$.*

Our main result in this section is stated as follows.

Theorem 3.2 *Let Assumption 2.1 hold, rank$(X) = N$ and $n_1 > \ldots > n_L$ where $L \geq 2$. Then the following hold:*

1. *Every sublevel set of Φ is connected. Moreover, Φ can attain any value arbitrarily close to p^*.*
2. *Every non-empty connected component of every level set of Φ is unbounded.*

We have the following decomposition of sublevel set: $\Phi^{-1}((-\infty, \alpha]) = \Phi^{-1}(\alpha) \cup \Phi^{-1}((-\infty, \alpha))$. It follows that if Φ has unbounded level sets then its sublevel sets must also be unbounded. We note that the reverse is not true, e.g. the standard Gaussian distribution function has unbounded sublevel sets but its level sets are bounded. Given that, the two statements of Theorem 3.2 together imply that every sublevel set of the loss must be both connected and unbounded. While the connectedness property of sublevel sets implies that the loss function is rather well-behaved, the unboundedness property of level sets intuitively implies that there are no bounded valleys in the loss surface, regardless of whether these valleys contain a global minimum or not. Clearly this also indicates that Φ has no strict local minima/maxima. In the remainder of this section, we will present the proof of Theorem 3.2. The following lemmas will be helpful.

Lemma 3.3 *Let the conditions of Theorem 3.2 hold. Given some $k \in [2, L]$. Then there is a continuous map $h : \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N\times n_k} \to \Omega_1$ which satisfy the following:*

1. *For every* $\Big((W_2,b_2),\ldots,(W_k,b_k),A\Big) \in \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N\times n_k}$ *it holds that* $F_k\Big(h\Big((W_l,b_l)_{l=2}^k,A\Big),(W_l,b_l)_{l=2}^k\Big)=A$.

2. *For every* $\theta=(W_l^*,b_l^*)_{l=1}^L$ *where all the matrices* $(W_l^*)_{l=2}^k$ *have full rank, there is a continuous curve from* θ *to* $\Big(h\Big((W_l^*,b_l^*)_{l=2}^k,F_k(\theta)\Big),(W_l^*,b_l^*)_{l=2}^L\Big)$ *on which the loss* Φ *is constant.*

Proof: For every $\Big((W_2,b_2),\ldots,(W_k,b_k),A\Big) \in \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N\times n_k}$, let us define the value of the map h as

$$h\Big((W_l,b_l)_{l=2}^k,A\Big)=(W_1,b_1),$$

where (W_1,b_1) is given by the following recursive formula

$$\begin{cases}\begin{bmatrix}W_1\\ b_1^T\end{bmatrix}=[X,\mathbf{1}_N]^\dagger\sigma^{-1}(B_1),\\ B_l=\Big(\sigma^{-1}(B_{l+1})-\mathbf{1}_N b_{l+1}^T\Big)W_{l+1}^\dagger,\ \forall\, l\in[1,k-2],\\ B_{k-1}=\begin{cases}(A-\mathbf{1}_N b_L^T)\,W_L^\dagger & k=L\\ \Big(\sigma^{-1}(A)-\mathbf{1}_N b_k^T\Big)W_k^\dagger & k\in[2,L-1]\end{cases}\end{cases}.$$

By our assumption $n_1>\ldots>n_L$, it follows from the domain of h that all the matrices $(W_l)_{l=2}^k$ have full column rank, and so they have a left inverse. Similarly, $[X,\mathbf{1}_N]$ has full row rank due to our assumption that *rank*$(X)=N$, and so it has a right inverse. Moreover σ has a continuous inverse by Assumption 2.1. Thus h is a continuous map as it is a composition of continuous functions. In the following, we prove that h satisfies the two statements of the lemma.

1. Let $\Big((W_2,b_2),\ldots,(W_k,b_k),A\Big) \in \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N\times n_k}$. Since all the matrices $(W_l)_{l=2}^k$ have full column rank and $[X,\mathbf{1}_N]$ has full row rank, it holds that $W_l^\dagger W_l=\mathbb{I}$ and $[X,\mathbf{1}_N][X,\mathbf{1}_N]^\dagger=\mathbb{I}$ and thus we easily obtain from the above definition of h that

$$\begin{cases}B_1=\sigma\Big([X,\mathbf{1}_N]\begin{bmatrix}W_1\\ b_1^T\end{bmatrix}\Big),\\ B_{l+1}=\sigma(B_lW_{l+1}+\mathbf{1}_N b_{l+1}^T),\ \forall\, l\in[1,k-2],\\ A=\begin{cases}B_{L-1}W_L+\mathbf{1}_N b_L^T & k=L,\\ \sigma(B_{k-1}W_k+\mathbf{1}_N b_k^T) & k\in[2,L-1].\end{cases}\end{cases}$$

One can easily check that the above formula of A is exactly the definition of F_k from (1) and thus it holds $F_k\Big(h\Big((W_l,b_l)_{l=2}^k,A\Big),(W_l,b_l)_{l=2}^k\Big)=A$ for every $\Big((W_2,b_2),\ldots,(W_k,b_k),A\Big)\in\Omega_2^*\times\ldots\times\Omega_k^*\times\mathbb{R}^{N\times n_k}$.

2. Let $G_l:\mathbb{R}^{N\times n_{l-1}}\to\mathbb{R}^{N\times n_l}$ be defined as

$$G_l(Z)=\begin{cases}ZW_L^*+\mathbf{1}_N(b_L^*)^T & l=L\\ \sigma\Big(ZW_l^*+\mathbf{1}_N(b_l^*)^T\Big) & l\in[2,L-1].\end{cases}$$

For convenience, let us group the parameters of the first layer into a matrix, say $U=[W_1^T,b_1]^T\in\mathbb{R}^{(d+1)\times n_1}$. Similarly, let $U^*=[(W_1^*)^T,b_1^*]^T\in\mathbb{R}^{(d+1)\times n_1}$. Let $f:\mathbb{R}^{(d+1)\times n_1}\to\mathbb{R}^{N\times n_k}$ be a function of (W_1,b_1) defined as

$$f(U)=G_k\circ G_{k-1}\ldots G_2\circ G_1(U),\ \text{where}$$
$$G_1(U)=\sigma([X,\mathbf{1}_N]U),\ U=[W_1^T,b_1]^T.$$

We note that this definition of f is exactly F_k from (1), but here we want to exploit the fact that f is a function of (W_1,b_1) as all other parameters are fixed to the corresponding values of θ. Let $A=F_k(\theta)$. By definition we have $f(U^*)=A$ and thus $U^*\in f^{-1}(A)$. Let us denote

$$(W_1^h,b_1^h)=h\Big(W_l^*,b_l^*)_{l=2}^k,A\Big),\quad U^h=[(W_1^h)^T,b_1^h]^T.$$

By applying the first statement of the lemma to $\Big((W_2^*,b_2^*),\ldots,(W_k^*,b_k^*),A\Big)$ we have

$$A=F_k\Big((W_1^h,b_1^h),(W_l^*,b_l^*)_{l=2}^k\Big)=f(U^h)$$

which implies $U^h\in f^{-1}(A)$. So far both U^* and U^h belong to $f^{-1}(A)$. The idea now is that if one can show that $f^{-1}(A)$ is a connected set then there would exist a connected path between U^* and U^h (and thus a path between (W_1^*,b_1^*) and (W_1^h,b_1^h)) on which the output at layer k is identical to A and hence the loss is invariant, which concludes the proof.

In the following, we show that $f^{-1}(A)$ is indeed connected. First, one observes that $\mathrm{range}(G_l)=\mathbb{R}^{N\times n_l}$ for every $l\in[2,k]$ since all the matrices $(W_l^*)_{l=2}^k$ have full column rank and $\sigma(\mathbb{R})=\mathbb{R}$ due to Assumption 2.1. Similarly, it follows from our assumption *rank*$(X)=N$ that $\mathrm{range}(G_1)=\mathbb{R}^{N\times n_1}$. By standard rules of compositions, we have

$$f^{-1}(A)=G_1^{-1}\circ G_2^{-1}\circ\ldots\circ G_k^{-1}(A).$$

where all the inverse maps G_l^{-1} have full domain. It holds

$$G_k^{-1}(A)=$$
$$\begin{cases}(A-\mathbf{1}_N b_L^T)(W_L^*)^\dagger+\{B\mid BW_L^*=0\} & k=L\\ \Big(\sigma^{-1}(A)-\mathbf{1}_N b_k^*\Big)(W_k^*)^\dagger+\{B\mid BW_k^*=0\} & \text{else}\end{cases}$$

which is a connected set in each case because of the following reasons: 1) the kernel of any matrix is connected, 2) the Minkowski-sum of two connected sets is connected by Proposition 2.8, and 3) the image of a connected set under a continuous map is connected by Proposition 2.7. By repeating the similar argument for $k-1,\ldots,2$ we conclude that $V:=G_2^{-1}\circ\ldots\circ G_k^{-1}(A)$ is connected. Lastly, we have

$$G_1^{-1}(V)=[X,\mathbf{1}_N]^\dagger\sigma^{-1}(V)+\{B\mid[X,\mathbf{1}_N]B=0\}$$

which is also connected by the same arguments above. Thus $f^{-1}(A)$ is a connected set.

Overall, we have shown in this proof that the set of (W_1, b_1) which realizes the same output at layer k (given the parameters of other layers in between are fixed) is a connected set. Since both (W_1^*, b_1^*) and $h\Big((W_l^*, b_l^*)_{l=2}^k, F_k(\theta)\Big)$ belong to this solution set, there must exist a continuous path between them on which the loss Φ is constant. □

The next lemma shows how to make the weight matrices full rank. Its proof can be found in the appendix.

Lemma 3.4 *Let the conditions of Theorem 3.2 hold. Let $\theta = (W_l, b_l)_{l=1}^L$ be any point in parameter space. Then there is a continuous curve which starts from θ and ends at some $\theta' = (W_l', b_l')_{l=1}^L$ so that θ' satisfies Property 3.1 and the loss Φ is constant on the curve.*

Proposition 3.5 *(Evard & Jafari, 1994) The set of full rank matrices $A \in \mathbb{R}^{m\times n}$ is connected for $m \neq n$.*

3.1. Proof of Theorem 3.2

1. Let L_α be some sublevel set of Φ. Let $\theta = (W_l, b_l)_{l=1}^L$ and $\theta' = (W_l', b_l')_{l=1}^L$ be arbitrary points in L_α. Let $F_L = F_L(\theta)$ and $F_L' = F_L(\theta')$. These two quantities are computed in the beginning and will never change during this proof. But when we write $F_L(\theta'')$ for some θ'' we mean the network output evaluated at θ''. The main idea is to construct two different continuous paths which simultaneously start from θ and θ' and are entirely contained in L_α (this is done by making the loss on each individual path non-increasing), and then show that they meet at a common point in L_α, which then implies that L_α is a connected set.

First of all, we can assume that both θ and θ' satisfy Property 3.1, because otherwise by Lemma 3.4 one can follow a continuous path from each point to arrive at some other point where this property holds and the loss on each path is invariant, meaning that we still stay inside L_α. As θ and θ' satisfy Property 3.1, all the weight matrices $(W_l)_{l=2}^L$ and $(W_l')_{l=2}^L$ have full rank, and thus by applying the second statement of Lemma 3.3 with $k = L$ and using the similar argument above, we can simultaneously drive θ and θ' to the following points,

$$\begin{aligned}\theta &= \Big(h\Big((W_l, b_l)_{l=2}^L, F_L\Big), (W_2, b_2), \dots, (W_L, b_L)\Big),\\ \theta' &= \Big(h\Big((W_l', b_l')_{l=2}^L, F_L'\Big), (W_2', b_2'), \dots, (W_L', b_L')\Big)\end{aligned} \tag{3}$$

where $h : \Omega_2^* \times \dots \times \Omega_L^* \times \mathbb{R}^{N\times m} \to \Omega_1$ is the continuous map from Lemma 3.3 which satisfies

$$F_L\Big(h\Big((\hat W_l, \hat b_l)_{l=2}^L, A\Big), (\hat W_l, \hat b_l)_{l=2}^L\Big) = A, \text{ for every} \tag{4}$$
$$\Big((\hat W_l, \hat b_l), \dots, (\hat W_L, \hat b_L), A\Big) \in \Omega_2^* \times \dots \times \Omega_L^* \times \mathbb{R}^{N\times n_k}.$$

Next, we construct a continuous path starting from θ on which the loss is constant and it holds at the end point of the path that all parameters from layer 2 till layer L are equal to the corresponding parameters of θ'. Indeed, by applying Proposition 3.5 to the pairs of full rank matrices (W_l, W_l') for every $l \in [2, L]$, we obtain continuous curves $W_2(\lambda), \dots, W_L(\lambda)$ so that $W_l(0) = W_l, W_l(1) = W_l'$ and $W_l(\lambda)$ has full rank for every $\lambda \in [0, 1]$. For every $l \in [2, L]$, let $c_l : [0, 1] \to \Omega_l^*$ be the curve of layer l defined as

$$c_l(\lambda) = \Big(W_l(\lambda), (1-\lambda)b_l + \lambda b_l'\Big).$$

We consider the curve $c : [0, 1] \to \Omega$ given by

$$c(\lambda) = \Big(h\Big((c_l(\lambda))_{l=2}^L, F_L\Big), c_2(\lambda), \dots, c_L(\lambda)\Big).$$

Then one can easily check that $c(0) = \theta$ and c is continuous as all the functions $h, c_2, \dots, c_l$ are continuous. Moreover, we have $\Big(c_2(\lambda), \dots, c_L(\lambda)\Big) \in \Omega_2^* \times \dots \times \Omega_L^*$ and thus it follows from (4) that $F_L(c(\lambda)) = F_L$ for every $\lambda \in [0, 1]$, which leaves the loss invariant on c.

Since the curve c above starts at θ and has constant loss, we can reset θ to the end point of this curve, by setting $\theta = c(1)$, while keeping θ' from (3), which together give us

$$\begin{aligned}\theta &= \Big(h\Big((W_l', b_l')_{l=2}^L, F_L\Big), (W_2', b_2'), \dots, (W_L', b_L')\Big),\\ \theta' &= \Big(h\Big((W_l', b_l')_{l=2}^L, F_L'\Big), (W_2', b_2'), \dots, (W_L', b_L')\Big).\end{aligned}$$

Now we note that the parameters of θ and θ' coincide at all layers except at the first layer. We will construct two continuous paths inside L_α, say $c_1(\cdot)$ and $c_2(\cdot)$, which starts from θ and θ' respectively , and show that they meet at a common point in L_α. Let $\hat Y \in \mathbb{R}^{N\times m}$ be any matrix so that

$$\varphi(\hat Y) \le \min(\Phi(\theta), \Phi(\theta')). \tag{5}$$

Consider the curve $c_1 : [0, 1] \to \Omega$ defined as

$$c_1(\lambda) = \Big(h\Big((W_l', b_l')_{l=2}^L, (1-\lambda)F_L + \lambda \hat Y\Big), (W_l', b_l')_{l=2}^L\Big).$$

Note that c_1 is continuous as h is continuous, and it holds:

$$c_1(0) = \theta, \quad c_1(1) = \Big(h\Big((W_l', b_l')_{l=2}^L, \hat Y\Big), (W_l', b_l')_{l=2}^L\Big).$$

It follows from the definition of Φ, $c_1(\lambda)$ and (4) that

$$\Phi(c_1(\lambda)) = \varphi(F_L(c_1(\lambda))) = \varphi((1-\lambda)F_L + \lambda \hat Y)$$

and thus by convexity of φ,

$$\begin{aligned}\Phi(c_1(\lambda)) &\le (1-\lambda)\varphi(F_L) + \lambda\varphi(\hat Y)\\ &\le (1-\lambda)\Phi(\theta) + \lambda\Phi(\theta) = \Phi(\theta),\end{aligned}$$

which implies that $c_1[0, 1]$ is entirely contained in L_α. Similarly, we can also construct a curve $c_2(\cdot)$ inside L_α which starts at θ' and satisfies

$$c_2(0) = \theta', \; c_2(1) = \Big(h\Big((W_l', b_l')_{l=2}^L, \hat Y\Big), (W_l', b_l')_{l=2}^L\Big).$$

So far, the curves c_1 and c_2 start at θ and θ' respectively and meet at the same point $c_1(1) = c_2(1)$.

Overall, we have shown that starting from any two points in L_α we can find two continuous curves so that the loss is non-increasing on each curve, and these curves meet at a common point in L_α, and so L_α has to be connected. Moreover, the point where they meet satisfies $\Phi(c_1(1)) = \varphi(\hat{Y})$. From (5), $\varphi(\hat{Y})$ can be chosen arbitrarily small, and thus Φ can attain any value arbitrarily close to p^*.

2. Let C be a non-empty connected component of some level set, i.e. $C \subseteq \Phi^{-1}(\alpha)$ for some $\alpha \in \mathbb{R}$. Let $\theta = (W_l, b_l)_{l=1}^L \in C$. Similar as above, we first use Lemma 3.4 to find a continuous path from θ to some other point where W_2 attains full rank, and the loss is invariant on the path. From that point, we apply Lemma 3.3 with $k = 2$ to obtain another continuous path (with constant loss) which leads us to $\theta' := \Big(h\Big((W_2, b_2), F_2(\theta)\Big), (W_2, b_2), \dots, (W_L, b_L)\Big)$ where $h : \Omega_2^* \to \Omega_1$ is a continuous map satisfying that

$$F_2\Big(h\Big((\hat{W}_2, \hat{b}_2), A\Big), (\hat{W}_l, \hat{b}_l)_{l=2}^L\Big) = A,$$

for every point $(\hat{W}_l, \hat{b}_l)_{l=1}^L$ such that $\hat{W}_2$ has full rank, and every $A \in \mathbb{R}^{N \times n_2}$. Note that $\theta' \in C$ as the loss is constant on the above paths. Consider the following continuous curve

$$c(\lambda) = \Big(h\Big((\lambda W_2, b_2), F_2(\theta)\Big), (\lambda W_2, b_2), \dots, (W_L, b_L)\Big)$$

for every $\lambda \geq 1$. This curve starts at θ' since $c(1) = \theta'$. Moreover $F_2(c(\lambda)) = F_2(\theta)$ for every $\lambda \geq 1$ and thus the loss is constant on this curve, meaning that the entire curve belongs to C. Lastly, since W_2 is full rank, the curve c is unbounded as λ goes to infinity, and thus C is unbounded. □

4. Large Width of One of Hidden Layers Leads to No Bad Local Valleys

In the previous section, we show that linearly independent training data essentially leads to connected sublevel sets. In this section, we show the first application of this result in proving absence of bad local valleys on the loss landscape of deep and wide neural nets with arbitrary training data.

Definition 4.1 *A local valley is a nonempty connected component of some strict sublevel set $L_\alpha^s := \{\theta \mid \Phi(\theta) < \alpha\}$. A bad local valley is a local valley on which the training loss Φ cannot be made arbitrarily close to p^*.*

The main result of this section is stated as follows.

Theorem 4.2 *Let Assumption 2.1 and Assumption 2.2 hold. Suppose that there exists a layer $k \in [1, L-1]$ such that $n_k \geq N$ and $n_{k+1} > \ldots > n_L$. Then the following hold:*

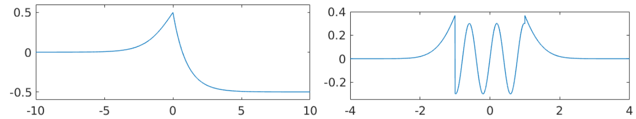

Figure 1. **Left**: an example function with exponential tails where local minima do NOT exist but local/global valleys still exist. **Right**: a different function which satisfies every local minimum is a global minimum, but bad local valleys still exist at both infinities (exponential tails) where local search algorithms easily get stuck.

1. *The loss Φ has no bad local valleys.*
2. *If $k \leq L-2$ then every local valley of Φ is unbounded.*

The conditions of Theorem 4.2 are satisfied for any strictly monotonic and piecewise linear activation function such as Leaky-ReLU (see Lemma 2.3). We note that for Leaky-ReLU and other similar homogeneous activation functions, the second statement of Theorem 4.2 is quite straightforward. Indeed, if one scales all parameters of one hidden layer by some arbitrarily large factor $k > 0$ and the weight matrix of the following layer by $1/k$ then the network output will be unchanged, and so every connected component of every level set (also sublevel set) must extend to infinity and thus be unbounded. However, for general non-homogeneous activation functions, the second statement is non-trivial.

The first statement of Theorem 4.2 implies that there is a continuous path from any point in parameter space on which the loss is non-increasing and gets arbitrarily close to p^*. At this point, one might wonder if a function satisfying "every local minimum is a global minimum" would automatically contain no bad local valleys. Unfortunately this is not true in general. Indeed, Figure 1 shows two counter-examples where a function does not have any bad local mimina, but bad local valleys still exist. The reason for this lies at the fact that bad local valleys generally need not contain any local minimum or even critical point though in theory they can have very large volume. Thus any pure results on global optimality of local minima with no further information on the loss would not be sufficient to guarantee convergence of local search algorithms to a global minimum, especially if they are initialized in such regions. Similar phenomenon has been observed by (Sohl-Dickstein & Kawaguchi, 2019).

We note that while the statements of Theorem 4.2 imply the absence of strict local minima and bounded local valleys, they do not rule out the possibility of non-strict bad local minima. Overall, the statements of Theorem 4.2 together imply that every local valley is an "unbounded" global valley in which the loss can attain any value arbitrarily close to p^*.

The high level proof idea for Theorem 4.2 is that inside every local valley one can find a point where the feature representations of all training samples are linearly independent at the wide hidden layer, and thus an application of Theorem 3.2 to the subnetwork from this wide layer till the output layer yields the result. We list below several technical lemmas which are helpful to prove Theorem 4.2.

Lemma 4.3 *Let $(F, W, \mathcal{I})$ be such that $F \in \mathbb{R}^{N\times n}, W \in \mathbb{R}^{n\times p}$, $rank(F) < n$ and $\mathcal{I} \subset \{1, \ldots, n\}$ be a subset of columns of F so that $rank(F(:, \mathcal{I})) = rank(F)$ and $\bar{\mathcal{I}}$ the remaining columns. Then there exists a continuous curve $c : [0, 1] \to \mathbb{R}^{n\times p}$ which satisfies the following:*

1. *$c(0) = W$ and $Fc(\lambda) = FW, \forall\, \lambda \in [0, 1]$.*
2. *The product $Fc(1)$ is independent of $F(:, \bar{\mathcal{I}})$.*

Lemma 4.4 *Given $v \in \mathbb{R}^n$ with $v_i \neq v_j \,\forall\, i \neq j$, and $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies Assumption 2.2. Let $S \subseteq \mathbb{R}^n$ be defined as $S = \{\sigma(v + b\mathbf{1}_n) \mid b \in \mathbb{R}\}$. Then it holds $Span(S) = \mathbb{R}^n$.*

We recall the following standard result from topology (e.g., see Apostol (1974), Theorem 4.23, p. 82).

Proposition 4.5 *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a continuous function. If $U \subseteq \mathbb{R}^n$ is an open set then $f^{-1}(U)$ is also open.*

4.1. Proof of Theorem 4.2

1. Let C be a connected component of some strict sublevel set $L^s_\alpha = \Phi^{-1}((-\infty, \alpha))$, for some $\alpha > p^*$. By Proposition 4.5, L^s_α is an open set and thus C must be open.

Step 1: Finding a point inside C where F_k has full rank. Let $\theta \in C$ be such that the pre-activation outputs at the first hidden layer are distinct for all training samples. Note that such θ always exist since Assumption 2.4 implies that the set of W_1 where this does not hold has Lebesgue measure zero, whereas C has positive measure. This combined with Assumption 2.1 implies that the (post-activation) outputs at the first hidden layer are distinct for all training samples. Now one can view these outputs at the first layer as inputs to the next layer and argue similarly. By repeating this argument and using the fact that C has positive measure, we conclude that there exists $\theta \in C$ such that the outputs at layer $k-1$ are distinct for all training samples, *i.e.* $(F_{k-1})_{i:} \neq (F_{k-1})_{j:}$ for every $i \neq j$. Let V be the pre-activation output (without bias term) at layer k, in particular $V = F_{k-1}W_k = [v_1, \ldots, v_{n_k}] \in \mathbb{R}^{N\times n_k}$. Since F_{k-1} has distinct rows, one can easily perturb W_k so that every column of V has distinct entries. Note here that the set of W_k where this does not hold has measure zero whereas C has positive measure. Equivalently, C must contain a point where every v_j has distinct entries. To simplify notation, let $a = b_k \in \mathbb{R}^{n_k}$, then by definition,

$$F_k = [\sigma(v_1 + \mathbf{1}_N a_1), \ldots, \sigma(v_{n_k} + \mathbf{1}_N a_{n_k})]. \quad (6)$$

Suppose that F_k has low rank, otherwise we are done. Let $r = rank(F_k) < N \leq n_k$ and $\mathcal{I} \subset \{1, \ldots, n_k\}, |\mathcal{I}| = r$ be the subset of columns of F_k so that $rank(F_k(:, \mathcal{I})) = rank(F_k)$, and $\bar{\mathcal{I}}$ the remaining columns. By applying Lemma 4.3 to $(F_k, W_{k+1}, \mathcal{I})$, we can follow a continuous path with invariant loss (*i.e.* entirely contained inside C) to arrive at some point where $F_k W_{k+1}$ is independent of $F_k(: .\bar{\mathcal{I}})$. It remains to show how to change $F_k(:, \bar{\mathcal{I}})$ by modifying certain parameters so that F_k has full rank. Let $p = |\bar{\mathcal{I}}| = n_k - r$ and $\bar{\mathcal{I}} = \{j_1, \ldots, j_p\}$. From (6) we have

$$F_k(:, \bar{\mathcal{I}}) = [\sigma(v_{j_1} + \mathbf{1}_N a_{j_1}), \ldots, \sigma(v_{j_p} + \mathbf{1}_N a_{j_p})].$$

Let $\mathrm{col}(\cdot)$ denotes the column space of a matrix. Then $\dim(\mathrm{col}(F_k(:, \mathcal{I}))) = r < N$. Since v_{j_1} has distinct entries, Lemma 4.4 implies that there must exist $a_{j_1} \in \mathbb{R}$ so that $\sigma(v_{j_1} + \mathbf{1}_N a_{j_1}) \notin \mathrm{col}(F_k(:, \mathcal{I}))$, because otherwise $\mathrm{Span}\{\sigma(v_{j_1} + \mathbf{1}_N a_{j_1}) \mid a_{j_1} \in \mathbb{R}\} \in \mathrm{col}(F_k(:, \mathcal{I}))$ whose dimension is strictly smaller than N and thus contradicts Lemma 4.4. So we pick one such value for a_{j_1} and follow a direct line segment between its current value and the new value. Note that the loss is invariant on this segment since any changes on a_{j_1} only affects $F_k(:, \bar{\mathcal{I}})$ which however has no influence on the loss by above construction. Moreover, it holds at the new value of a_{j_1} that $rank(F_k)$ increases by 1. Since $n_k \geq N$ by our assumption, it follows that $p \geq N - r$ and thus one can choose $\{a_{j_2}, \ldots, a_{j_{N-r}}\}$ in a similar way and finally obtain $rank(F_k) = N$.

Step 2: Applying Theorem 3.2 to the subnetwork above k. Suppose that we have found from previous step a $\theta = ((W^*_l, b^*_l)_{l=1}^L) \in C$ so that F_k has full rank. Let $g : \Omega_{k+1} \times \ldots \times \Omega_L$ be given as

$$g\Big((W_l, b_l)_{l=k+1}^L\Big) = \Phi\Big((W^*_l, b^*_l)_{l=1}^k, (W_l, b_l)_{l=k+1}^L\Big) \quad (7)$$

We recall that C is a connected component of L^s_α. It holds $g\Big((W^*_l, b^*_l)_{l=k+1}^L\Big) = \Phi(\theta) \leq \alpha$. Now one can view g as the new loss for the subnetwork from layer k till layer L and F_k can be seen as the new training data. Since $rank(F_k) = N$ and $n_{k+1} > \ldots > n_L$, Theorem 3.2 implies that g has connected sublevel sets and g can attain any value arbitrarily close to p^*. Let $\epsilon \in (p^*, \alpha)$ and $(W'_l, b'_l)_{l=k+1}^L$ be any point such that $g\Big((W'_l, b'_l)_{l=k+1}^L\Big) \leq \epsilon$. Since both $(W^*_l, b^*_l)_{l=k+1}^L$ and $(W'_l, b'_l)_{l=k+1}^L$ belongs to the α-sublevel set of g, which is a connected set, there must exist a continuous path from $(W^*_l, b^*_l)_{l=k+1}^L$ to $(W'_l, b'_l)_{l=k+1}^L$ on which the value of g is not larger than α. This combined with (7) implies that there is also a continuous path from $\theta = \Big((W^*_l, b^*_l)_{l=1}^k, (W^*_l, b^*_l)_{l=k+1}^L\Big)$ to $\theta' := \Big((W^*_l, b^*_l)_{l=1}^k, (W'_l, b'_l)_{l=k+1}^L\Big)$ on which the loss Φ is not larger than α. Since C is connected, it must hold $\theta' \in C$. Moreover, we have $\Phi(\theta') = g\Big((W'_l, b'_l)_{l=k+1}^L\Big) \leq \epsilon$. Since

ϵ can be chosen arbitrarily small and close to p^*, we conclude that the loss Φ can be made arbitrarily small inside C, and thus Φ has no bad local valleys.

2. Let C be a local valley, which by Definition 4.1 is a connected component of some strict sublevel set $L_\alpha^s = \Phi^{-1}((-\infty, \alpha))$. According the the proof of the first statement above, one can find a $\theta = (W_l^*, b_l^*)_{l=1}^L \in C$ so that $F_k(\theta)$ has full rank. Now one can view $F_k(\theta)$ as the training data for the subnetwork from layer k till layer L. The new loss is defined for this subnetwork as

$$g\Big((W_l, b_l)_{l=k+1}^L\Big) = \Phi\Big((W_l^*, b_l^*)_{l=1}^k, (W_l, b_l)_{l=k+1}^L\Big).$$

By our assumptions, σ satisfies Assumption 2.1 and $n_{k+1} > \ldots > n_L$, thus the above subnetwork with the new loss g and training data $F_k(\theta)$ satisfy all the conditions of Theorem 3.2, and so it follows that g has unbounded level set components. Let $\beta := g\Big((W_l^*, b_l^*)_{l=k+1}^L\Big) = \Phi(\theta) < \alpha$. Let E be a connected component of the level set $g^{-1}(\beta)$ which contains $(W_l^*, b_l^*)_{l=k+1}^L$. Let $D = \Big\{\Big((W_l^*, b_l^*)_{l=1}^k, (W_l, b_l)_{l=k+1}^L\Big) \;\Big|\; (W_l, b_l)_{l=k+1}^L \in E\Big\}$. Then D is connected and unbounded since E is connected and unbounded. It holds for every $\theta' \in D$ that $\Phi(\theta') = \beta$, and thus $D \subseteq \Phi^{-1}(\beta) \subseteq L_\alpha^s$, where the last inclusion follows from $\beta < \alpha$. Moreover, we have $\theta = \Big((W_l^*, b_l^*)_{l=1}^k, (W_l^*, b_l^*)_{l=k+1}^L\Big) \in D$ and also $\theta \in C$, it follows that $D \subseteq C$ since C is already the maximal connected component of L_α^s. Since D is unbounded, C must also be unbounded, which finishes the proof. □

5. Large Width of First Hidden Layer Leads to Connected Sublevel Sets

In the previous section (Theorem 4.2), we show that if one of the hidden layers has more than N neurons then the loss function has no bad local valleys. In this section, we treat a special case where the first hidden layer has at least $2N$ neurons. Under such setting, the next theorem shows in addition that every sublevel set must be also connected.

Theorem 5.1 *Let Assumption 2.1 and Assumption 2.2 hold. Suppose that $n_1 \geq 2N$ and $n_2 > \ldots > n_L$. Then every sublevel set of Φ is connected. Moreover, every connected component of every level set of Φ is unbounded.*

Theorem 5.1 shows a stronger result than Theorem 4.2 as it not only implies that there are no bad local valleys but also there is a unique global valley. Equivalently, all finite global minima (if exist) must be connected. This can be seen as a generalization of previous result (Venturi et al., 2018) from one hidden layer networks and square loss to arbitrary deep networks and convex losses. Interestingly, recent work (Draxler et al., 2018; Garipov et al., 2018) have shown that different global minima of several existing CNN architectures can be connected by a continuous path on which the loss has similar values. While our current results are not directly applicable to these models, we consider this as a stepping stone for such an extension in future work. Similar to previous results, the unboundedness of level sets as shown in the second statement of Theorem 5.1 implies that Φ has no bounded local valleys nor strict local extrema. The proof of Theorem 5.1 relies on the following lemmas.

Lemma 5.2 *Let $(X, W, b, V) \in \mathbb{R}^{N\times d} \times \mathbb{R}^{d\times n} \times \mathbb{R}^n \times \mathbb{R}^{n\times p}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ satisfy Assumption 2.2. Suppose that $n \geq N$ and X has distinct rows. Let $Z = \sigma(XW + \mathbf{1}_N b^T)\, V$. There is a continuous curve $c : [0,1] \to \mathbb{R}^{d\times n} \times \mathbb{R}^n \times \mathbb{R}^{n\times p}$ with $c(\lambda) = (W(\lambda), b(\lambda), V(\lambda))$ satisfying:*

1. *$c(0) = (W, b, V)$.*
2. *$\sigma\Big(XW(\lambda)) + \mathbf{1}_N b(\lambda)^T\Big)\, V(\lambda) = Z, \;\forall\, \lambda \in [0,1]$.*
3. *rank $\Big(\sigma\Big(XW(1) + \mathbf{1}_N b(1)^T\Big)\Big) = N$.*

Lemma 5.3 *Let $(X, W, V, W') \in \mathbb{R}^{N\times d} \times \mathbb{R}^{d\times n} \times \mathbb{R}^{n\times p} \times \mathbb{R}^{d\times n}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ satisfy Assumption 2.2. Suppose that $n \geq 2N$ and rank$(\sigma(XW)) = N$, rank$(\sigma(XW')) = N$. Then there is a continuous curve $c : [0,1] \to \mathbb{R}^{d\times n} \times \mathbb{R}^{n\times p}$ with $c(\lambda) = (W(\lambda), V(\lambda))$ which satisfies the following:*

1. *$c(0) = (W, V)$.*
2. *$\sigma(XW(\lambda))\, V(\lambda) = \sigma(XW)\, V, \quad \forall\, \lambda \in [0,1]$.*
3. *$W(1) = W'$.*

5.1. Proof of Theorem 5.1

Let $\theta = (W_l, b_l)_{l=1}^L, \theta' = (W_l', b_l')_{l=1}^L$ be arbitrary points in some sublevel set L_α. It is sufficient to show that there is a connected path between θ and θ' on which the loss is not larger than α. The output at the first layer is given by

$$F_1(\theta) = \sigma([X, \mathbf{1}_N][W_1^T, b_1]^T),$$
$$F_1(\theta') = \sigma([X, \mathbf{1}_N][W_1'^T, b_1']^T).$$

First, by applying Lemma 5.2 to (X, W_1, b_1, W_2), we can assume that $F_1(\theta)$ has full rank, because otherwise there is a continuous path starting from θ to some other point where the rank condition is fulfilled and the loss is invariant on the path, and so we can reset θ to this new point. Similarly, we can assume that $F_1(\theta')$ has full rank.

Next, by applying Lemma 5.3 to the tuple $\Big([X, \mathbf{1}_N], [W_1^T, b_1]^T, W_2, [W_1'^T, b_1']^T\Big)$, and using the similar argument as above, we can drive θ to some other point where the parameters of the first hidden layer agree with the corresponding values of θ'. So we can assume w.l.o.g. that $(W_1, b_1) = (W_1', b_1')$. Note that at this step we did not modify θ' but θ and thus $F_1(\theta')$ still has full rank.

Once the first hidden layer of θ and θ' coincide, one can view the output of this layer, say $F_1 := F_1(\theta) = F_1(\theta')$ with *rank*$(F_1) = N$, as the new training data for the subnetwork from layer 1 till layer L (given that (W_1, b_1) is fixed). This subnetwork and the new data F_1 satisfy all the conditions of Theorem 3.2, and so it follows that the loss Φ restricted to this subnetwork has connected sublevel sets, which implies that there is a connected path between $(W_l, b_l)_{l=2}^L$ and $(W_l', b_l')_{l=2}^L$ on which the loss is not larger than α. This indicates that there is also a connected path between θ and θ' in L_α and so L_α must be connected.

To show that every level set component of Φ is unbounded, let $\theta \in \Omega$ be an arbitrary point. Denote $F_1 = F_1(\theta)$ and let $\mathcal{I} \subset \{1, \ldots, N\}$ be such that *rank*$(F_1(:, \mathcal{I}))$ = *rank*(F_1). Since *rank*$(F_1) \le \min(N, n_1) < n_1$, we can apply Lemma 4.3 to the tuple $(F_1, W_2, \mathcal{I})$ to find a continuous path $W_2(\lambda)$ which drives θ to some other point where the output at 2nd layer $F_1 W_2$ is independent of $F_1(:, \bar{\mathcal{I}})$. Note that the network output at 2nd layer is invariant on this path and hence the entire path belongs to the same level set component with θ. From that point, one can easily scale $(W_1(:, \bar{\mathcal{I}}), b_1(\bar{\mathcal{I}}))$ to arbitrarily large values without affecting the output. Since this path has constant loss and is unbounded, it follows that every level set component of Φ is unbounded. $\square$

6. Extension to Other Activation Functions

One way to extend our results from the previous sections to other activation functions such as ReLU and exponential linear unit is to remove Assumption 2.1 from the previous theorems, as shown next.

Theorem 6.1 *Let $K = \min \{n_1, \ldots, n_{L-1}\}$. Then all the following hold under Assumption 2.2:*

1. *If $K \ge N$ then the loss Φ has no bad local valleys.*
2. *If $K \ge 2N$ then every sublevel set of Φ is connected.*

It is clear that the conditions of Theorem 6.1 are now much stronger than that of our previous theorems as all the hidden layers need to be wide enough. Nevertheless, it is worth mentioning that this kind of technical conditions (i.e. all the hidden layers are sufficiently wide) have also been used in recent work (Allen-Zhu et al., 2018b; Du et al., 2018; Zou et al., 2018) to establish convergence of gradient descent methods to a global minimum. From a theoretical standpoint, these results seem to suggest that Leaky-ReLU might in general lead to a much "easier" loss surface than ReLU.

7. Related Work

Many interesting theoretical results have been developed on the loss surface of neural networks (Livni et al., 2014; Choromanska et al., 2015; Haeffele & Vidal, 2017; Hardt & Ma, 2017; Xie et al., 2017; Yun et al., 2017; Lu & Kawaguchi, 2017; Pennington & Bahri, 2017; Zhou & Liang, 2018; Liang et al., 2018b;a; Zhang et al., 2018; Nouiehed & Razaviyayn, 2018; Laurent & v. Brecht, 2018). There is also a whole line of researches studying convergence of learning algorithms in training neural networks (Andoni et al., 2014; Sedghi & Anandkumar, 2015; Janzamin et al., 2016; Gautier et al., 2016; Brutzkus & Globerson, 2017; Soltanolkotabi, 2017; Soudry & Hoffer, 2017; Tian, 2018; Wang et al., 2018; Ji & Telgarsky, 2019; Arora et al., 2019; Allen-Zhu et al., 2018a; Bartlett et al., 2018; Chizat & Bach, 2018) and others studying generalization properties, which is however beyond the scope of this paper.

The closest existing result is the work by (Venturi et al., 2018) who show that if the number of hidden neurons is greater than the intrinsic dimension of the network, defined as the dimension of some function space, then the loss has no spurious valley, and furthermore, if the number of hidden neurons is greater than two times the intrinsic dimension then every sublevel set is connected. The results apply to one hidden layer networks with population risk and square loss. As admitted by the authors in the paper, an extension of such result, in particular the notion of intrinsic dimension, to multiple layer networks would require the number of neurons to grow exponentially with depth. Prior to this, (Safran & Shamir, 2016) showed that for a class of deep fully connected networks with common loss functions, there exists a continuous descent path between specific pairs of points in parameter space which satisfy certain conditions, and these conditions can be shown to hold with probability $1/2$ as the width of the last hidden layer goes to infinity.

Most closely related in terms of the setting are the work by (Nguyen & Hein, 2017; 2018) who analyze the optimization landscape of standard deep and wide (convolutional) neural networks for multiclass problem. They both assume that the network has a wide hidden layer k with $n_k \ge N$. This condition has been recently relaxed to $n_1 + \ldots + n_{L-1} \ge N$ by using flexible skip-connections (Nguyen et al., 2019). All of these results so far require real analytic activation functions, and thus are not applicable to the class of piecewise linear activations analyzed in this paper. Moreover, while the previous work focus on global optimality of critical points, this paper characterizes sublevel sets of the loss function which gives us further insights and intuition on the underlying geometric structure of the optimization landscape.

Conclusion. We have shown that every sublevel set of the training loss function of a certain class of deep over-parameterized neural networks is connected and unbounded.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers, 2018a. arXiv:1811.04918.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *arXiv:1811.03962*, 2018b.

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

Apostol, T. M. *Mathematical analysis*. Addison Wesley, 1974.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. *ICLR*, 2019.

Bartlett, P., Helmbold, D., and Long, P. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *ICML*, 2018.

Brutzkus, A. and Globerson, A. Globally optimal gradient descent for a convnet with gaussian inputs. *ICML*, 2017.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NIPS*, 2018.

Choromanska, A., Hena, M., Mathieu, M., Arous, G. B., and LeCun, Y. The loss surfaces of multilayer networks. *AISTATS*, 2015.

Clevert, D., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *ICML*, 2018.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *arXiv:1811.03804*, 2018.

Evard, J. C. and Jafari, F. The set of all mxn rectangular real matrices of rank-r is connected by analytic regular arcs. *Proceedings of American Mathematical Society*, 1994.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *NIPS*, 2018.

Gautier, A., Nguyen, Q., and Hein, M. Globally optimal training of generalized polynomial neural networks with nonlinear spectral methods. *NIPS*, 2016.

Haeffele, B. D. and Vidal, R. Global optimality in neural network training. *CVPR*, 2017.

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

Janzamin, M., Sedghi, H., and Anandkumar, A. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv:1506.08473*, 2016.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. *ICLR*, 2019.

Laurent, T. and v. Brecht, J. Deep linear networks with arbitrary loss: All local minima are global. In *ICML*, 2018.

Liang, S., Sun, R., Lee, J. D., and Srikant, R. Adding one neuron can eliminate all bad local minima. *arXiv:1805.08671*, 2018a.

Liang, S., Sun, R., Li, Y., and Srikant, R. Understanding the loss surface of neural networks for binary classification. In *ICML*, 2018b.

Livni, R., Shalev-Shwartz, S., and Shamir, O. On the computational efficiency of training neural networks. *NIPS*, 2014.

Lu, H. and Kawaguchi, K. Depth creates no bad local minima. *arXiv:1702.08580*, 2017.

Nguyen, Q. and Hein, M. The loss surface of deep and wide neural networks. *ICML*, 2017.

Nguyen, Q. and Hein, M. Optimization landscape and expressivity of deep cnns. *ICML*, 2018.

Nguyen, Q., Mukkamala, M. C., and Hein, M. On the loss landscape of a class of deep neural networks with no bad local valleys. In *ICLR*, 2019.

Nouiehed, M. and Razaviyayn, M. Learning deep models: Critical points and local openness. *ICLR Workshop*, 2018.

Pennington, J. and Bahri, Y. Geometry of neural network loss surfaces via random matrix theory. *ICML*, 2017.

Safran, I. and Shamir, O. On the quality of the initial basin in overspecified networks. *ICML*, 2016.

Sedghi, H. and Anandkumar, A. Provable methods for training neural networks with sparse connectivity. *ICLR Workshop*, 2015.

Sohl-Dickstein, J. and Kawaguchi, K. Eliminating all bad local minima from loss landscapes without even adding an extra unit. *arXiv:1901.03909*, 2019.

Soltanolkotabi, M. Learning relus via gradient descent. *NIPS*, 2017.

Soudry, D. and Hoffer, E. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *ICLR Workshop 2018*, 2017.

Tian, Y. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *ICML*, 2018.

Venturi, L., Bandeira, A. S., and Bruna, J. Spurious valleys in two-layer neural network optimization landscapes. *arXiv:1802.06384v2*, 2018.

Wang, G., Giannakis, G. B., and Chen, J. Learning relu networks on linearly separable data: Algorithm, optimality, and generalization. *arXiv:1808.04685*, 2018.

Xie, B., Liang, Y., and Song, L. Diverse neural network learns true target functions. *AISTAT*, 2017.

Yun, C., Sra, S., and Jadbabaie, A. Global optimality conditions for deep neural networks. *ICLR*, 2017.

Zhang, H., Shao, J., and Salakhutdinov, R. Deep neural networks with multi-branch architectures are less non-convex. *arXiv:1806.01845*, 2018.

Zhou, Y. and Liang, Y. Critical points of neural networks: Analytical forms and landscape properties. *ICLR*, 2018.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv:1811.08888*, 2018.