On Connected Sublevel Sets in Deep Learning

Quynh Nguyen

Abstract

We study sublevel sets of the loss function in training deep neural networks. For linearly independent data, we prove that every sublevel set of the loss is connected and unbounded. We then apply this result to prove similar properties on the loss surface of deep over-parameterized neural nets with piecewise linear activation functions.

1. Introduction

It is commonly observed in deep learning that over-parameterization sometimes can be helpful for optimizing neural networks. Theoretically, several recent work (Allen-Zhu et al., 2018b; Du et al., 2018; Zou et al., 2018) have also established convergence of gradient descent for non-linear networks under excessive over-parameterization regimes. For instance the above work require \( \Omega(N^4) \) neurons (or more) per hidden layer to guarantee that (stochastic) gradient descent converges to a global minimum with zero training error. This is an inspiring result given the hardness of the problem, however the question of which fundamental properties of the loss underpinning these successes remains unanswered. We are interested in the following question:

Is there any underlying geometric structure of the loss function that can “intuitively” supports for the success of local search algorithms like gradient descent under excessive over-parameterization regimes?

This paper sheds light on this question by showing that every sublevel set of the loss function is connected if the network has a sufficiently wide hidden layer. The key idea of the paper is to show that, for linearly independent training data and under a relatively mild condition on the architecture, every sublevel set of the loss is connected. This allows us to obtain similar results for deep and wide neural nets with arbitrary data. In particular, we first show that if one of the hidden layers has more neurons than the number of training samples, then the loss has no bad local valleys in the sense that there is a continuous path from anywhere in parameter space on which the loss is non-increasing and gets arbitrarily close to the minimal value of the loss. In a special case where the first hidden layer is a wide layer with twice more neurons than the number of training samples, then we show that every sublevel set is connected, and thus there is a unique global valley. All our results hold for deep fully connected networks with standard architecture, for arbitrary convex losses and strictly monotonic and/or piecewise linear activation functions such as Leaky-ReLU.

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 \( b_k \in \mathbb{R}^{m} \) the weight matrix and biases respectively of layer \( k \). By convention, we assume that \( n_0 = d \) and \( n_L = m \). Let \( \sigma : \mathbb{R} \rightarrow \mathbb{R} \) 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

\[
F_k = \begin{cases} 
X & k = 0 \\
\sigma(F_{k-1}W_k + 1_N b_k_{1:k}) & 1 \leq k \leq L - 1 \\
F_{L-1}W_L + 1_N b_L_{1:L} & k = L 
\end{cases}
\]

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

\[
\Phi(\theta) = \varphi(F_L(\theta))
\]

where \( \varphi : \mathbb{R}^{N \times m} \rightarrow \mathbb{R} \) is assumed to be any convex loss. Typical examples include the standard cross-entropy loss.
A key concept of this paper is the sublevel set of a function. ∑ is the ground-truth class of \( x_i \), and the standard square loss for regression \( \varphi(F_L) = \frac{1}{2} \| F_L - Y \|^2_F \) 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 \( \Phi \). Note that \( p^* \) is fully determined by the choice of \( \varphi(\cdot) \) and thus independent of the training data. We make no assumption on \( p^* \) in this paper but for most of practical losses as above one has \( p^* = 0 \). Below we list several assumptions on the activation function and will refer to them accordingly in our different results.

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

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

**Assumption 2.2** There does 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 \( x \in \mathbb{R} \).

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

We note that Lemma 2.3 implies that the set of activation functions which satisfy Assumption 2.2 is dense on the set of all continuous functions defined on a bounded interval. We make the following universal assumption on our 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 \( \alpha \)-level set of \( \Phi : \Omega \to \mathbb{R} \) is the preimage \( \Phi^{-1}(\alpha) = \{ \theta \in \Omega \mid \Phi(\theta) = \alpha \} \), and the \( \alpha \)-sublevel set of \( \Phi \) 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 \( AA^\dagger = I \), and if \( A \) has full column rank then it has a left inverse \( A^\dagger = (A^T)^{-1}A^\dagger \) with \( A^\dagger A = I \).

### 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, \( \text{rank}(X) = N \) and \( n_1 > \ldots > n_L \) where \( L \geq 2 \). Then the following hold:

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

We have the following decomposition of sublevel set: \( \Phi^{-1}((-\infty, \alpha]) = \bigcup_{\alpha > \alpha_i} \Phi^{-1}((\alpha_i, \alpha]) \). It follows that if \( \Phi \) 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 first statement of Theorem 3.2 implies that \( \Phi \) has a rather well-behaved loss surface, the second statement implies that it has no bounded valleys, regardless of whether these valleys contain a global minimum or not. This clearly also indicates that \( \Phi \) has no strict local minima/maxima. In the remaining 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 \( \left( (W_2, b_2), \ldots, (W_k, b_k), A \right) \in \mathbb{R}^{N \times m} \), there exists a continuous map \( h : \left( (W_2, b_2), \ldots, (W_k, b_k), A \right) \to \mathbb{R}^{N \times m} \) which satisfies the following:
\[ \Omega_2 \times \ldots \times \Omega_k^* \times \mathbb{R}^{N \times n_k} \text{ it holds that} \]
\[ F_k \left( h \left( (W_i, b_i)^{k}_{l=2}, A \right), (W_i, b_i)^{k}_{l=2} \right) = A. \]

2. For every \( \theta = (W^*_i, b^*_i)^{k}_{l=2} \) where all the matrices \( (W^*_i)^{k}_{l=2} \) have full rank, there is a continuous curve from \( \theta \) to \( \left( h \left( (W^*_i, b^*_i)^{k}_{l=2}, F_k(\theta) \right), (W^*_i, b^*_i)^{k}_{l=2} \right) \) on which the loss \( \Phi \) is constant.

**Proof:** For every \( (W_2, b_2), \ldots, (W_k, b_k), A) \in \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N \times n_k} \), let us define the value of the map \( h \) as
\[ h \left( (W_i, b_i)^{k}_{l=2}, A \right) = (W_1, b_1), \]
where \( (W_1, b_1) \) is given by the following recursive formula
\[
\begin{align*}
W^1_1 &= [X, 1_N]^T \sigma^{-1}(B_1), \\
B_l &= \left( \sigma^{-1}(B_{l-1}) - 1_N b^T_{l-1} \right) W^T_{l-1}, \quad \forall l \in [1, k - 2], \\
B_{k-1} &= \left( A - 1_N b^T_{k-1} \right) W^T_{k-1} \quad k = L.
\end{align*}
\]
By our assumption \( n_1 > \ldots > n_L \), it follows from the domain of \( h \) that all the matrices \( (W_l)^{k}_{l=2} \) have full column rank, and so they have a left inverse. Similarly, \( [X, 1_N] \) has full row rank due to our assumption that \( \text{rank}(X) = N \), and so it has a right inverse. Moreover \( \sigma \) 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 \( (W_2, b_2), \ldots, (W_k, b_k), A) \in \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N \times n_k} \). Since all the matrices \( (W_l)^{k}_{l=2} \) have full column rank and \( [X, 1_N] \) has full row rank, it holds that \( W^T_l W_l = I \) and \( [X, 1_N][X, 1_N]^T = I \) and thus we easily obtain from the above definition of \( h \) that
\[
\begin{align*}
B_1 &= \sigma \left( [X, 1_N] \begin{bmatrix} W^1_1 \\ b^T_1 \end{bmatrix} \right), \\
B_{l+1} &= \sigma \left( B_l W_{l+1} + 1_N b^T_{l+1} \right), \quad \forall l \in [1, k - 2], \\
A &= \left( \sigma(B_{k-1} W_k + 1_N b^T_k) \right) \quad k \in [2, L - 1].
\end{align*}
\]
One can easily check that the above formula of \( A \) is exactly the definition of \( F_k \) from (1) and thus it holds \( F_k \left( h \left( (W_i, b_i)^{k}_{l=2}, A \right), (W_i, b_i)^{k}_{l=2} \right) = A \) for every \( (W_2, b_2), \ldots, (W_k, b_k), A) \in \Omega_2^* \times \ldots \times \Omega_k^* \times \mathbb{R}^{N \times n_k} \).

2. Let \( G_l: \mathbb{R}^{N \times n_i} \to \mathbb{R}^{N \times n_i} \) be defined as
\[
G_l(Z) = \begin{cases}
ZW^T_l + 1_N (b^T_l)^T & l = L \\
\sigma \left( ZW^T_l + 1_N (b^T_l)^T \right) & l \in [2, L - 1].
\end{cases}
\]
For convenience, let us group the parameters of the first layer into a matrix, say \( U = [W^T_1, b_1]^T \in \mathbb{R}^{(d+1) \times n_1} \). 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-1} \circ G_{k-2} \circ \ldots \circ G_1(U), \quad G_1(U) = \sigma([X, 1_N] U), \quad U = [W^T, b^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 \( \theta \). Let \( A = F_k(\theta) \). By definition we have \( f(U^*) = A \) and thus \( U^* \in f^{-1}(A) \). Let us denote
\[
(W^h_1, b^h_1) = h \left( (W^*_1, b^*_1)^{k}_{l=2}, A \right), \quad U^h = [(W^h_1)^T, b^h_1]^T.
\]
By applying the first statement of the lemma to \((W^*_2, b^*_2), \ldots, (W^*_k, b^*_k), A)\) we have
\[
A = F_k \left( (W^h_1, b^h_1), (W^*_i, b^*_i)^{k}_{l=2} \right) = f(U^h)
\]
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 \((W^h_1, b^h_1)\)) 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 \( \text{range}(G_l) = \mathbb{R}^{N \times n_l} \) for every \( l \in [2, k] \) since all the matrices \( (W^*_l)^{k}_{l=2} \) have full column rank and \( \sigma(\mathbb{R}) = \mathbb{R} \) due to Assumption 2.1. Similarly, it follows from our assumption \( \text{rank}(X) = N \) that \( \text{range}(G_1) = \mathbb{R}^{N \times n_1} \). By standard rules of compositions, we have
\[
f^{-1}(A) = G^{-1}_1 \circ G^{-1}_2 \circ \ldots \circ G^{-1}_k(A),
\]
where all the inverse maps \( G^{-1}_i \) have full domain. It holds
\[
G^{-1}_k(A) = \begin{cases}
(A - 1_N b^T_k)(W^T_k)^T + \{ B \mid BW^*_k = 0 \} & k = L \\
\sigma^{-1}(A) - 1_N b^T_k \right) (W^T_k)^T + \{ B \mid BW^*_k = 0 \} & \text{else}
\end{cases}
\]
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 \( V := G^{-1}_2 \circ \ldots \circ G^{-1}_k(A) \) is connected. Lastly, we have
\[
G^{-1}_1(V) = [X, 1_N]^T \sigma^{-1}(V) + \{ B \mid [X, 1_N]B = 0 \}
\]
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 \(h \big( (W^*_1, b^*_1)_{l=2}^L, F_L(\theta) \big)\) belong to this solution set, there must exist a continuous path between them on which the loss \(\Phi\) is constant. \(\square\)

**Lemma 3.4** Let the conditions of Theorem 3.2 hold. Let \(\theta = (W_1, b_1)_{l=1}^L\) be any point in parameter space. Then there is a continuous curve which starts from \(\theta\) and ends at some \(\theta' = (W'_1, b'_1)_{l=1}^L\) so that \(\theta'\) satisfies Property 3.1 and the loss \(\Phi\) is constant on the curve.

**Proposition 3.5** (Evar & 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 \(L_\alpha\) be some sublevel set of \(\Phi\). Let \(\theta = (W_1, b_1)_{l=1}^L\) and \(\theta' = (W'_1, b'_1)_{l=1}^L\) be arbitrary points in \(L_\alpha\). 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 \(\theta''\) we mean the network output evaluated at \(\theta''\). The main idea is to construct two different continuous paths which simultaneously start from \(\theta\) and \(\theta'\) and are entirely contained in \(L_\alpha\) (this is done by making the loss on each individual path non-increasing), and then show that they meet at a common point in \(L_\alpha\), which then implies that \(L_\alpha\) is a connected set.

First of all, we can assume that both \(\theta\) and \(\theta'\) 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 \(L_\alpha\). As \(\theta\) and \(\theta'\) satisfy Property 3.1, all the weight matrices \((W_1)_{l=2}^L\) and \((W'_1)_{l=2}^L\) 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 \(\theta\) and \(\theta'\) to the following points,

\[
\begin{align*}
\theta &= \left( h \left( (W_1, b_1)_{l=2}^L, F_L \right), (W_2, b_2), \ldots, (W_L, b_L) \right), \\
\theta' &= \left( h \left( (W'_1, b'_1)_{l=2}^L, F'_L \right), (W'_2, b'_2), \ldots, (W'_L, b'_L) \right)
\end{align*}
\]  

(3)

where \(h : \Omega_2^L \times \ldots \times \Omega_L^L \times \mathbb{R}^{N \times m} \rightarrow \Omega_1\) is the continuous map from Lemma 3.3 which satisfies

\[
F_L \left( h \left( (W_1, b_1)_{l=2}^L, A \right), (W_1, b_1)_{l=2}^L \right) = A, \quad \text{for every} \quad (W_1, b_1), \ldots, (W_L, b_L), A \in \Omega_2^L \times \ldots \times \Omega_L^L \times \mathbb{R}^{N \times m}.
\]

Next, we construct a continuous path starting from \(\theta\) 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 \(\theta'\). Indeed, by applying Proposition 3.5 to the pairs of full rank matrices \((W_1, W'_1)\) for every \(l \in [2, L]\), we obtain continuous curves \(W_2(\lambda), \ldots, W_L(\lambda)\) 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 \(c_l : [0, 1] \rightarrow \Omega^*_l\) be the curve of layer \(l\) defined as

\[
c_l(\lambda) = \left( W_l(\lambda), (1 - \lambda)b_l + \lambda b'_l \right).
\]

We consider the curve \(c : [0, 1] \rightarrow \Omega\) given by

\[
c(\lambda) = \left( h \left( (c_1(\lambda)_{l=2}^L, F_L \right), c_2(\lambda), \ldots, c_L(\lambda) \right).
\]

Then one can easily check that \(c(0) = \theta\) and \(c(1) = \theta'\) is continuous as all the functions \(h, c_2, \ldots, c_L\) are continuous. Moreover, we have \(c(\lambda) = c(\lambda)\) \(\in \Omega_2^L \times \ldots \times \Omega_L^L\) 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 \(\theta\) and has constant loss, we can reset \(\theta\) to the end point of this curve, by setting \(\theta = c(1)\), while keeping \(\theta'\) from (3), which together give us

\[
\begin{align*}
\theta &= \left( h \left( (W_1, b_1)_{l=2}^L, F_L \right), (W_2, b_2), \ldots, (W_L, b_L) \right), \\
\theta' &= \left( h \left( (W'_1, b'_1)_{l=2}^L, F'_L \right), (W'_2, b'_2), \ldots, (W'_L, b'_L) \right).
\end{align*}
\]

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

\[
\varphi(\tilde{Y}) \leq \min(\Phi(\theta), \Phi(\theta')).
\]

(5)

Consider the curve \(c_1 : [0, 1] \rightarrow \Omega\) defined as

\[
c_1(\lambda) = \left( h \left( (W_1, b_1)_{l=2}^L, (1 - \lambda)F_L + \lambda \tilde{Y} \right), (W_1, b_1)_{l=2}^L \right).
\]

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

\[
c_1(0) = \theta, \quad c_1(1) = \left( h \left( (W_1, b_1)_{l=2}^L, \tilde{Y} \right), (W_1, b_1)_{l=2}^L \right).
\]

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

\[
\Phi(c_1(\lambda)) = \varphi(F_L(c_1(\lambda))) = \varphi((1 - \lambda)F_L + \lambda \tilde{Y})
\]

and thus by convexity of \(\varphi\),

\[
\Phi(c_1(\lambda)) \leq (1 - \lambda)\varphi(F_L) + \lambda \varphi(\tilde{Y})
\]

\[
\leq (1 - \lambda)\Phi(\theta) + \lambda \Phi(\theta) = \Phi(\theta),
\]

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

\[
c_2(0) = \theta', \quad c_2(1) = \left( h \left( (W'_1, b'_1)_{l=2}^L, \tilde{Y} \right), (W'_1, b'_1)_{l=2}^L \right).
\]

(4)
So far, the curves $c_1$ and $c_2$ start at $\theta$ and $\theta'$ respectively and meet at the same point $c_1(1) = c_2(1)$. Overall, we have shown that starting from any two points in $L_\alpha$, 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 $L_\alpha$, and so $L_\alpha$ 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 $\Phi$ 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 $\theta = (W_t, b_t)_{t=1}^L \in C$. Similar as above, we first use Lemma 3.4 to find a continuous path from $\theta$ 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 $\theta' := \left( h\left((W_2, b_2), F_2(\theta)\right), (W_2, b_2), \ldots, (W_L, b_L) \right)$ where $h : \Omega_2 \to \Omega_1$ is a continuous map satisfying that $F_2\left(h\left((W_2, b_2), A\right), (W_1, b_1)_{t=2}^L \right) = A$,

for every point $(W_1, b_1)_{t=1}^L$ such that $W_2$ has full rank, and every $A \in \mathbb{R}^{\mathbb{R} \times n_2}$. Note that $\theta' \in C$ as the loss is constant on the above paths. Consider the following continuous curve $c(\lambda) = \left( h\left((\lambda W_2, b_2), F_2(\theta)\right), (W_2, b_2), \ldots, (W_L, b_L) \right)$ for every $\lambda \geq 1$. This curve starts at $\theta'$ since $c(1) = \theta'$. We have $F_2(c(\lambda)) = F_2(\theta)$ for every $\lambda \geq 1$ and thus the loss is constant on this curve, meaning that the entire curve belongs to $C$. Lastly, the curve $c[1, \infty)$ is unbounded as $\lambda$ goes to infinity, and thus $C$ has to be 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 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 $L^{s}_\alpha := \left\{ \theta \mid \Phi(\theta) < \alpha \right\}$. A bad local valley is a local valley on which the training loss $\Phi$ 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:

1. The loss $\Phi$ has no bad local valleys.

2. If $k \leq L - 2$ then every local valley of $\Phi$ 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 satisfies “every local minimum is a global minimum” would automatically contain no bad local valleys. Unfortunately this is not true in general. Indeed, Figure 4 shows two counter-examples where a function does not have any bad local minima, but bad local valleys still exist. The reason for this lies at the fact that bad local valleys generally need not contain any critical point though in theory they can have very large volume or even be unbounded. 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 to the second statement of Theorem 4.2, the first statement on one hand can guarantee absence of strict local minima, but on the other hand cannot rule out the possibility of non-strict bad local minima. This suggests that it might be desirable to have in practice both properties for the loss surface of neural nets, that is, there are no bad local valleys and every local minimum is a global minimum. Overall, the statements of Theorem 4.2 altogether imply that every local valley must be 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 would yield the result. Below we list several lemmas which are helpful for the proof of Theorem 4.2.

**Lemma 4.3** Let \( (F, W, \mathcal{I}) \) be such that \( F \in \mathbb{R}^{N \times n}, W \in \mathbb{R}^{n \times p}, \text{rank}(F) < n \) and \( \mathcal{I} \subset \{1, \ldots, n\} \) be a subset of columns of \( F \) so that \( \text{rank}(F(:, \mathcal{I})) = \text{rank}(F) \) and \( \bar{\mathcal{I}} \) the remaining columns. Then there exists a continuous curve \( c: [0, 1] \rightarrow \mathbb{R}^{n \times p} \) 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 \( \sigma: \mathbb{R} \rightarrow \mathbb{R} \) satisfies Assumption 2.2. Let \( S \subseteq \mathbb{R}^n \) be defined as \( S = \{ \sigma(v + b1_n) \mid b \in \mathbb{R} \} \). Then it holds \( \text{span}(S) = \mathbb{R}^n \).

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

**Proposition 4.5** Let \( f: \mathbb{R}^m \rightarrow \mathbb{R}^n \) be a continuous function. If \( U \subseteq \mathbb{R}^m \) 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 \( L^*_a = \Phi^{-1}((-\infty, \alpha)) \), for some \( \alpha > p^* \). By Proposition 4.5, \( L^*_a \) 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 \( \theta \) 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,

\[
F_k = [\sigma(v_1 + 1_Na_1), \ldots, \sigma(v_{n_k} + 1_Na_{n_k})].
\]

Suppose that \( F_k \) has low rank, otherwise we are done. Let \( r = \text{rank}(F_k) < N \leq n_k \) and \( \bar{\mathcal{I}} = \{1, \ldots, n_k\} \). Then there exists a continuous path with invariant loss (i.e., entirely contained inside \( C \)) to arrive at some point where \( F_kW_{k+1} \) 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

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

Let \( \text{col}(\cdot) \) denotes the column space of a matrix. Then \( \dim(\text{col}(F_k(:, \bar{\mathcal{I}}))) = r < N \). Since \( v_{j_1} \) has distinct entries, Lemma 4.4 implies that there must exist \( a_{j_1} \in \mathbb{R} \) so that \( \sigma(v_{j_1} + 1_Na_{j_1}) \notin \text{col}(F_k(:, \bar{\mathcal{I}})) \), because otherwise \( \text{Span}\{\sigma(v_{j_1} + 1_Na_{j_1}) \mid a_{j_1} \in \mathbb{R}\} \subseteq \text{col}(F_k(:, \bar{\mathcal{I}})) \) 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 \( \text{rank}(F_k) \) increases by 1. Since \( n_k \geq N \) by our assumption, it follows that \( p \geq N - r \) and thus one can choose \( \{a_{j_1}, \ldots, a_{n-k-r}\} \) in a similar way and finally obtain \( \text{rank}(F_k) = N \).

*Step 2:* Applying Theorem 3.2 to the subnetwork above \( k \).

Suppose that we have found from the previous step a point \( \theta = ((W^*_1, b^*_1)^{l_1}_{l=1}, (W^*_1, b^*_1)^{l_1}_{l=1}) \) so that \( F_1 \) has full rank. Let the function \( g : \Omega_{k+1} \times \ldots \times \Omega_L \rightarrow \mathbb{R}^L \) be defined as

\[
g\left((W_1, b_1)^{l_1}_{l=1}, (W_1, b_1)^{l_1}_{l=1}\right) = \Phi\left((W^*_1, b^*_1)^{l_1}_{l=1}, (W_1, b_1)^{l_1}_{l=1}\right).
\]

We recall that \( C \) is a connected component of \( L^*_a \). It holds

\[
g\left((W^*_1, b^*_1)^{l_1}_{l=1}\right) = \Phi(\theta) \leq \alpha.
\]

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 \( \text{rank}(F_k) = N \) 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 \( (W^*_1, b^*_1)^{l_1}_{l=1} \) be any point such that \( g\left((W^*_1, b^*_1)^{l_1}_{l=1}\right) \leq \epsilon \). Since both \( (W^*_1, b^*_1)^{l_1}_{l=1} \) and \( (W^*_1, b^*_1)^{l_1}_{l=1} \) belongs to the \( \alpha \)-sublevel set of \( g \), which is a connected set, there must exist a continuous path from \( (W^*_1, b^*_1)^{l_1}_{l=1} \) to \( (W^*_1, b^*_1)^{l_{k+1}}_{l=1} \) on which the value of \( g \) is not larger than \( \alpha \). This combined with (7) implies that there is also a continuous path from \( \theta = ((W^*_1, b^*_1)^{l_1}_{l=1}, (W^*_1, b^*_1)^{l_1}_{l=1}) \) to \( \theta' := ((W^*_1, b^*_1)^{l_1}_{l=1}, (W^*_1, b^*_1)^{l_{k+1}}_{l=1}) \).
On Connected Sublevel Sets in Deep Learning

\[ (W^*_l, b^*_l)_{l=1}^{k}, (W'_l, b'_l)_{l=k+1} \] on which the loss \( \Phi \) is not larger than \( \alpha \). Since \( C \) is connected, it must hold \( \theta' \in C \). Moreover, we have \( \Phi(\theta') = g((W'_l, b'_l)_{l=k+1}) \leq \epsilon \). Since \( \epsilon \) can be chosen arbitrarily small and close to \( p^* \), we conclude that the loss \( \Phi \) can be made arbitrarily small inside \( C \), and thus \( \Phi \) 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 \( L^*_a = \Phi^{-1}(\langle -\infty, \alpha \rangle) \). According the the proof of the first statement above, one can find a \( \theta = (W^*_l, b^*_l)_{l=1}^n \in C \) 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

\[ g((W_l, b_l)_{l=k+1}) = \Phi((W^*_l, b^*_l)_{l=1}^k, (W_l, b_l)_{l=k+1}). \]

By our assumptions, \( \sigma \) 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 \( \beta := g((W^*_l, b^*_l)_{l=1}^k) = \Phi(\theta) < \alpha \). Let \( E \) be a connected component of the level set \( g^{-1}(\beta) \) which contains \((W^*_l, b^*_l)_{l=1}^k\). Let \( D = \{(W_l, b_l)_{l=k+1} \mid (W_l, b_l)_{l=k+1} \in E\} \).

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 \( D \subseteq \Phi^{-1}(\beta) \subseteq L^*_a \), where the last inclusion follows from \( \beta < \alpha \). Moreover, we have \( \theta = (W^*_l, b^*_l)_{l=1}^k, (W'_l, b'_l)_{l=k+1} \in D \) and also \( \theta \in C \), it follows that \( D \subseteq C \) since \( C \) is already the maximal connected component of \( L^*_a \). 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 \( \Phi \) is connected. Moreover, every connected component of every level set of \( \Phi \) 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 minimum. 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 \( \Phi \) has no bounded local valleys nor strict local extrema. The proof of Theorem 5.1 relies on the following lemmas.

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

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

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

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

5.1. Proof of Theorem 5.1

Let \( \theta = (W_l, b_l)_{l=1}^k, \theta' = (W'_l, b'_l)_{l=1}^k \) be arbitrary point in some sublevel set \( L^*_a \). It is sufficient to show that there is a connected path between \( \theta \) and \( \theta' \) on which the loss is not larger than \( \alpha \). The output at the first layer is given by

\[ F_1(\theta) = \sigma([X, 1_N]|W^T_l, b^T_l), \]
\[ F_1(\theta') = \sigma([X, 1_N]|W'^T_l, b'^T_l). \]

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 \( \theta \) to some other point where the rank condition is fulfilled and the loss is invariant on the path, and so we can reset \( \theta \) to this new point. Similarly, we can assume that \( F_1(\theta') \) has full rank.
Next, by applying Lemma 5.3 to the tuple \( [X, 1_N], [W_1^T, b_1]_1^T, W_2, [W_2^T, b_1']_1^T \), and using the similar argument as above, we can drive \( \theta \) to some other point where the parameters of the first hidden layer agree with the corresponding values of \( \theta' \). 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 \( \theta' \) but \( \theta \) and thus \( F_1(\theta') \) still has full rank.

Once the first hidden layer of \( \theta \) and \( \theta' \) coincide, one can view the output of this layer, say \( F_1 := F_1(\theta) = F_1(\theta') \) with \( \text{rank}(F_1) = N \), 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 \( \Phi \) restricted to this subnetwork has connected sublevel sets, which implies that there is a connected path between \( (W_1, b_1)_{1 \leq 2} \) and \( (W_1', b_1')_{1 \leq 2} \), on which the loss is not larger than \( \alpha \). This indicates that there is also a connected path between \( \theta \) and \( \theta' \) in \( L_\alpha \) and so \( L_\alpha \) must be connected.

To show that every level set component of \( \Phi \) is unbounded, let \( \theta \in \Omega \) be an arbitrary point. Denote \( F_1 = F_1(\theta) \) and let \( I \subseteq \{1, \ldots, N\} \) be such that \( \text{rank}(F_1(\cdot; I)) = \text{rank}(F_1) \).

Since \( \text{rank}(F_1) \leq \min(N, n_1) < n_1 \), we can apply Lemma 4.3 to the tuple \( (F_1, W_2, I) \) to find a continuous path \( W_2(\lambda) \) which drives \( \theta \) to some other point where the output at 2nd layer \( F_1 W_2 \) is independent of \( F_1(\cdot; I) \). 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 \( \theta \). From that point, one can easily scale \( (W_1(\cdot; I), b_1(I)) \) 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 \( \Phi \) is unbounded.

\section{6. Extensions to ReLU Activation Function}

We discuss a possible approach to extend our previous results to the ReLU activation by removing Assumption 2.1.

Theorem 6.1 \textit{All the following hold under Assumption 2.2:}

1. If \( \min \{n_1, \ldots, n_{L-1}\} \geq N \) then the loss function \( \Phi \) has no bad local valleys.

2. If \( \min \{n_1, \ldots, n_{L-1}\} \geq 2N \) then every sublevel set of \( \Phi \) is connected.

It is clear that the conditions of Theorem 6.1 are far from practical settings, and theoretically they are also significantly stronger than that of Theorem 5.1 as it requires all the hidden layers to be sufficiently over-parameterized. Nevertheless, we note that the similar conditions have also been used by recent theoretical work (Allen-Zhu et al., 2018b; Du et al., 2018) in proving convergence guarantees of gradient descent methods. Theoretically, we find these results interesting as together they seem to suggest that Leaky-ReLU might lead to a much “easier” loss surface than ReLU.

\section{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; Haefele & Vidal, 2017; Safran & Shamir, 2016; 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; Soltankotabi, 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 study the relationship between the intrinsic dimension of neural networks and the presence/absence of spurious valleys. They 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.

More 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 \( n_k \geq N \). This condition has been recently relaxed to \( n_1 + \ldots + n_{L-1} \geq N \) 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 give us further insights and intuition on the underlying geometric structure of the optimization landscape.

\textbf{Conclusion.} We show that every sublevel set of the loss function in training a certain class of deep over-parameterized neural nets is connected and unbounded.
On Connected Sublevel Sets in Deep Learning

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., Vescghini, 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 ICLR, 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.

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.
A. Proof of Lemma 2.3

A function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is continuous piecewise linear with at least two pieces if it can be represented as

$$\sigma(x) = ax + b, \quad \forall x \in (x_{i-1}, x_i), \forall i \in [1, n + 1].$$

for some $n \geq 1$, $x_0 = -\infty < x_1 < \ldots < x_n < x_{n+1} = \infty$ and $(a_i, b_i)_{i=1}^{n+1}$. We can assume that all the linear pieces agree at their intersection and there are no consecutive pieces with the same slope: $a_i \neq a_{i+1}$ for every $i \in [1, n]$.

Suppose by contradiction that $\sigma$ does not satisfy Assumption 2.2, then there are non-zero coefficients $(\lambda_i, y_i)_{i=1}^{m}$ with $y_i \neq y_j (i \neq j)$ such that $\sigma(x) = \sum_{i=1}^{m} \lambda_i \sigma(x - y_i)$ for every $x \in \mathbb{R}$. We assume w.l.o.g. that $y_1 < \ldots < y_m$.

Case 1: $y_1 > 0$. For every $x \in (-\infty, x_1)$ we have $\sigma(x) = a_1x + b_1 = \sum_{i=1}^{m} \lambda_i (a_1(x - y_i) + b_1)$ and thus by comparing the coefficients on both sides we obtain $\sum_{i=1}^{m} \lambda_i a_1 = a_1$. Moreover, for every $x \in (x_1, \min(x_1 + y_2, x_1, x_2 + y_1))$ it holds $\sigma(x) = ax + b_2 = \sum_{i=1}^{m} \lambda_i (a_1(x - y_i) + b_1)$ and so $\sum_{i=1}^{m} \lambda_i a_1 = a_2$. Thus $a_1 = a_2$, which is a contradiction.

Case 2: $y_1 < 0$. By definition, for $x \in (-\infty, x_1 + y_1)$ we have $\sigma(x) = a_1x + b_1 = \sum_{i=1}^{m} \lambda_i (a_1(x - y_i) + b_1)$ and thus by comparing the coefficients we obtain

$$\sum_{i=1}^{m} \lambda_i a_1 = a_1. \quad (8)$$

For $x \in (x_1 + y_1, \min(x_1 + y_2, x_1, x_2 + y_1))$ it holds

$$\sigma(x) = a_1x + b_1 = \lambda_1(a_2(x - y_1) + b_2) + \sum_{i=2}^{m} \lambda_i (a_1(x - y_i) + b_1)$$

and thus by comparing the coefficients we have

$$\lambda_1 a_2 + \sum_{i=2}^{m} \lambda_i a_1 = a_1.$$

This combined with (8) leads to $\lambda_1 a_1 = \lambda_1 a_2$, and thus $a_1 = a_2$ (since $\lambda_1 \neq 0$) which is a contradiction.

One can prove similarly for ELU (Clevert et al., 2016)

$$\sigma(x) = \begin{cases} \lambda \alpha x & x \geq 0 \\ \alpha (e^{\alpha x} - 1) & x < 0 \end{cases} \quad \text{where } \alpha > 0.$$

Suppose by contradiction that there exist non-zero coefficients $(\lambda_i, y_i)_{i=1}^{m}$ with $y_i \neq y_j (i \neq j)$ such that $\sigma(x) = \sum_{i=1}^{m} \lambda_i \sigma(x - y_i)$, and assume w.l.o.g. that $y_1 < \ldots < y_m$. If $y_m > 0$ then for every $x \in (\max(0, y_m), y_m)$ it holds

$$\sigma(x) = x = \lambda_0 \alpha (e^{\alpha x - y_m} - 1) + \sum_{i=1}^{m-1} \lambda_i (x - y_i)$$

$$\Rightarrow e^x = \frac{x e^{y_m} - \sum_{i=1}^{m-1} \lambda_i (x - y_i) e^{y_m}}{\lambda_0 \alpha} + e^{y_m}$$

which is a contradiction since $e^x$ cannot be identical to any affine function on any open interval. Thus it must hold that $y_m < 0$. But then for every $x \in (y_m, 0)$ we have

$$\sigma(x) = \alpha (e^{\alpha x} - 1) = \sum_{i=1}^{m} \lambda_i (x - y_i)$$

$$\Rightarrow e^x = \frac{1}{\alpha} \sum_{i=1}^{m} \lambda_i (x - y_i) + 1$$

which is a contradiction for the same reason above.

B. Proof of Proposition 2.7

Pick some $a, b \in f(A)$ and let $x, y \in A$ be such that $f(x) = a$ and $f(y) = b$. Since $A$ is connected, there is a continuous curve $r : [0, 1] \rightarrow A$ so that $r(0) = x, r(1) = y$. Consider the curve $f \circ r : [0, 1] \rightarrow f(A)$, then it holds that $f(r(0)) = a, f(r(1)) = b$. Moreover, $f \circ r$ is continuous as both $f$ and $r$ are continuous. Thus it follows from Definition 2.6 that $f(A)$ is a connected.

C. Proof of Proposition 2.8

Let $x, y \in U + V$ then there exist $a, b \in U$ and $c, d \in V$ such that $x = a + c, y = b + d$. Since $U$ and $V$ are connected sets, there exist two continuous curves $p : [0, 1] \rightarrow U$ and $q : [0, 1] \rightarrow V$ such that $p(0) = a, p(1) = b$ and $q(0) = c, q(1) = d$. Consider the continuous curve $r(t) := p(t) + q(t)$ then we have $r(0) = a + c = x, r(1) = b + d = y$ and $r(t) \in U + V$ for every $t \in [0, 1]$. This implies that every two elements in $U + V$ can be connected by a continuous curve and thus $U + V$ must be a connected set.

D. Proof of Lemma 3.4

The idea is to make one weight matrix full rank at a time while keeping the others fixed (except the first layer). Each step is done by following a continuous path which leads to a new point where the rank condition is fulfilled while keeping the loss constant along the path. Each time when we follow a continuous path, we reset our starting point to the end point of the path and proceed. This is repeated until all the matrices $\{W_l\}_{l=2}^L$ have full rank.

Step 1: Make $W_2$ full rank. If $W_2$ has full rank then we proceed to $W_3$. Otherwise, let $\text{rank}(W_2) = r < n_2 < n_1$. Let $\mathcal{I} \subset \{1, \ldots, n_1\}, |\mathcal{I}| = r$ denote the set of indices of linearly independent rows of $W_2$ so that $\text{rank}(W_2(\mathcal{I}, :)) = r$. Let $\hat{\mathcal{I}}$ denote the remaining rows of $W_2$. Let $E \in \mathbb{R}^{(n_1-r) \times r}$ be a matrix such that $W_2(\mathcal{I}, :) = EW_2(\hat{\mathcal{I}}, :)$ and $P \in \mathbb{R}^{n_1 \times n_1}$ be a permutation matrix which permutes the rows of $W_2$ according to $\hat{\mathcal{I}}$ so that we can write

$$PW_2 = \begin{bmatrix} W_2(\mathcal{I}, :) \\ W_2(\hat{\mathcal{I}}, :) \end{bmatrix}.$$
We recall that $F_1(\theta)$ is the output of the network at the first layer, evaluated at $\theta$. Below we drop $\theta$ and just write $F_1$ as it is clear from the context. By construction of $P$, we have

$$F_1 P^T = [F_1(:, I), F_1(:, \bar{I})].$$

The first step is to turn $W_1$ into a canonical form. In particular, the set of all possible solutions of $W_1$ which realizes the same the output $F_1$ at the first hidden layer is characterized by $X^T (\sigma^{-1}(F_1) - 1_N b_1^T) + \ker(X)$ where we denote, by abuse of notation, $\ker(X) = \{ A \in \mathbb{R}^{d \times n_1} \mid X A = 0 \}$. This solution set is connected because $\ker(X)$ is a connected set and the Minkowski-sum of two connected sets is known to be connected, and so there exists a continuous path between every two solutions in this set on which the output $F_1$ is invariant. Obviously the current $W_1$ and $X^T (\sigma^{-1}(F_1) - 1_N b_1^T)$ are elements of this set, thus they must be connected by a continuous path on which the loss is invariant. So we can assume now that $W_1 = X^T (\sigma^{-1}(F_1) - 1_N b_1^T)$.

Next, consider the curve:

$$W_1(\lambda) = X^T (\sigma^{-1}(A(\lambda)) - 1_N b_1^T),$$

$$A(\lambda) = [F_1(:, I) + \lambda F_1(:, \bar{I}) E, (1 - \lambda) F_1(:, \bar{I})] P.$$

This curve starts at $\theta$ since $W_1(0) = W_1$, and it is continuous as $\sigma$ has a continuous inverse by Assumption 2.1. Using $X X^T = I$, one can compute the pre-activation output (without bias term) at the second layer as

$$\sigma(X W_1(\lambda) + 1_N b_1^T) W_2 = A(\lambda) W_2 = F_1 W_2,$$

which implies that the loss is invariant on this curve, and so we can take its end point $W_1(1)$ as a new starting point:

$$W_1 = X^T (\sigma^{-1}(A) - 1_N b_1^T),$$

$$A = [F_1(:, I) + F_1(:, \bar{I}) E, 0] P.$$

Now, the output at second layer above, given by $A W_2$, is independent of $W_2(\bar{I}, :)$. So we can easily change $W_2(\bar{I}, :)$. Thus one can easily change the matrix $W_2(\bar{I}, :)$. So that $W_2$ has full rank while still keeping the loss invariant.

Step 2: Using induction to make $W_3, \ldots, W_L$ full rank.

Let $\theta = (W_i, 0_{i-1})_{i=2}^L$ be our current point. Suppose that all the matrices $(W_i)_{i=2}^L$ have full rank for some $k \geq 2$ then we show below how to make $W_{k+1}$ full rank. We write $F_k$ to denote $F_k(\theta)$. By the second statement of Lemma 3.3, we can follow a continuous path (with invariant loss) to drive $\theta$ to the following point:

$$\theta := \left( h \left( (W_i, b_i)_{i=2}^k, F_k \right), (W_i, b_i)_{i=2}^k \right)$$

where $h : \Omega_2 \times \ldots \times \Omega_2 \times \mathbb{R}^{N \times n_k}$ is the continuous map from Lemma 3.3 which satisfies for every $A \in \mathbb{R}^{N \times n_k}$,

$$F_k \left( h \left( (W_i, b_i)_{i=2}^k, A \right), (W_i, b_i)_{i=2}^k \right) = A.$$

Now, if $W_{k+1}$ already has full rank then we are done, otherwise we follow the similar steps as before. Indeed, let $r = \text{rank}(W_{k+1}) < n_{k+1} < n_k$ and $I \subset \{ 1, \ldots, n_k \}$, $|I| = r$ the set of indices of $r$ linearly independent rows of $W_{k+1}$. Then there is a permutation matrix $P \in \mathbb{R}^{n_k \times n_k}$ and some matrix $E \in \mathbb{R}^{(n_k - r) \times r}$ so that

$$P W_{k+1} = \begin{bmatrix} W_{k+1}(I, :) \mid W_{k+1}(\bar{I}, :) \end{bmatrix}, \quad W_{k+1}(I, :) = EW_{k+1}(I, :).$$

Moreover it holds

$$F_k P^T = [F_k(:, I), F_k(:, \bar{I})].$$

Consider the following curve $c : [0, 1) \to \Omega$ which continuously update $(W_1, b_1)$ while keeping other layers fixed:

$$c(\lambda) = \left( h \left( (W_i, b_i)_{i=2}^k, A(\lambda) \right), (W_2, b_2), \ldots, (W_L, b_L) \right),$$

where $A(\lambda) = [F_k(:, I) + \lambda F_k(:, \bar{I}) E, (1 - \lambda) F_k(:, \bar{I})] P$.

It is clear that $c$ is continuous as $h$ is continuous. One can easily verify that $c(0) = \theta$ by using (12) and (9). The pre-activation output (without bias term) at layer $k + 1$ for every point on this curve is given by

$$F_k(c(\lambda)) W_{k+1} = A(\lambda) W_{k+1} = F_k W_{k+1}, \quad \forall \lambda \in [0, 1],$$

where the first equality follows from (10) and the second follows from (11) and (12). As the loss is invariant on this curve, we can take its end point $c(1)$ as a new starting point:

$$\theta := \left( h \left( (W_i, b_i)_{i=2}^k, A \right), (W_2, b_2), \ldots, (W_L, b_L) \right),$$

where $A = [F_k(:, I) + F_k(:, \bar{I}) E, 0] P$.

At this point, the output at layer $k + 1$ as mentioned above is given by $A W_{k+1}$, which is independent of $W_{k+1}(\bar{I}, :)$. Thus one can easily change the matrix $W_{k+1}(\bar{I}, :)$. So that $W_{k+1}$ has full rank while leaving the loss invariant.

Overall, by induction we can make all the weight matrices $W_2, \ldots, W_L$ full rank by following several continuous paths on which the loss is constant, which finishes the proof.

E. Proof of Lemma 4.3

Let $r = \text{rank}(F) < n$. Since $\bar{I}$ contains $r$ linearly independent columns of $F$, the remaining columns must lie on their span. In other words, there exists $E \in \mathbb{R}^{r \times (n - r)}$ so that $F(:, \bar{I}) = F(:, I) E$. Let $P \in \mathbb{R}^{n \times n}$ be a permutation matrix which permutes the columns of $F$ according to $\bar{I}$ so that we can write $F = [F(:, I), F(:, \bar{I})] P$. Consider the continuous curve $c : [0, 1] \to \mathbb{R}^{n \times P}$ defined as

$$c(\lambda) = P^T \begin{bmatrix} W(I, :) + \lambda E W(\bar{I}, :) \mid (1 - \lambda) W(\bar{I}, :) \end{bmatrix}, \forall \lambda \in [0, 1].$$
It holds $c(0) = P^T \begin{bmatrix} W(I,:) \\ W(\tilde{I},:) \end{bmatrix} = W$. For every $\lambda \in [0,1]$:

$$Fc(\lambda) = [F(:,I), F(:,\tilde{I})] PP^T \begin{bmatrix} W(I,:) + \lambda EW(\tilde{I},:) \\ (1-\lambda)W(\tilde{I},:) \end{bmatrix} = F(:,I)W(I,:) + F(:,\tilde{I})W(\tilde{I},:) = FW.$$ 

Lastly, we have

$$Fc(1) = [F(:,I), F(:,\tilde{I})] PP^T \begin{bmatrix} W(\tilde{I},:) + EW(\tilde{I},:) \\ 0 \end{bmatrix} = F(:,I)W(\tilde{I},:) + F(:,\tilde{I})EW(\tilde{I},:)$$

which is independent of $F(:,\tilde{I})$.

### F. Proof of Lemma 4.4

Suppose by contradiction that $\dim(\text{Span}(S)) < n$. Then there exists $\lambda \in \mathbb{R}^n$, $\lambda \neq 0$ such that $\lambda \perp \text{Span}(S)$, and thus it holds $\sum_{i=1}^n \lambda_i \sigma(v_i + b) = 0$ for every $b \in \mathbb{R}$. We assume w.l.o.g. that $\lambda_1 \neq 0$ then it holds

$$\sigma(v_1 + b) = -\sum_{i=2}^n \frac{\lambda_i}{\lambda_1} \sigma(v_i + b), \quad \forall b \in \mathbb{R}.$$ 

By a change of variable, we have

$$\sigma(c) = -\sum_{i=2}^n \frac{\lambda_i}{\lambda_1} \sigma(c + v_i - v_1), \quad \forall c \in \mathbb{R},$$

which contradicts Assumption 2.2. Thus $\text{Span}(S) = \mathbb{R}^n$.

### G. Proof of Lemma 5.2

Let $F = \sigma(XW + 1_Nb^F) \in \mathbb{R}^{N \times n}$. If $F$ already has full rank then we are done. Otherwise let $r = \text{rank}(F) < N \leq n$. Let $I$ denote a set of column indices of $F$ so that $\text{rank}(F(:,I)) = r$ and $\tilde{I}$ the remaining columns. By applying Lemma 4.3 to $(F,V,\tilde{I})$, we can find a continuous path $V(\lambda)$ so that we will arrive at some point where $FV(\lambda)$ is invariant on the path and it holds at the end point of the path that $FV$ is independent of $F(:,\tilde{I})$. This means that we can arbitrarily change the values of $W(:,\tilde{I})$ and $b(\tilde{I})$ without affecting the value of $Z$, because any changes of these variables are absorbed into $F(:,\tilde{I})$ which anyway has no influence on $FV$. Thus it is sufficient to show that there exist $W(:,\tilde{I})$ and $b(\tilde{I})$ for which $F$ has full rank. Let $p = n - r$ and $\tilde{I} = \{j_1, \ldots, j_p\}$. Let $A = XW$ then $A(:,\tilde{I}) := [a_{j_1}, \ldots, a_{j_p}] = XW(:,\tilde{I})$. By assumption $X$ has distinct rows, one can choose $W(:,I)$ so that each $a_j \in \mathbb{R}^N$ has distinct entries. Then we have

$$F(:,\tilde{I}) = [\sigma(a_{j_1} + 1_Nb_{j_1}), \ldots, \sigma(a_{j_p} + 1_Nb_{j_p})].$$

Let $\text{col}(\cdot)$ denotes the column space of a matrix. It holds $\dim(\text{col}(F(:,\tilde{I}))) = r < N$. Since $a_{j_1}$ has distinct entries, Lemma 4.4 implies that there must exist $b_{j_1} \in \mathbb{R}$ so that $\sigma(a_{j_1} + 1_Nb_{j_1}) \notin \text{col}(F(:,I))$, because otherwise $\text{Span}([\sigma(a_{j_1} + 1_Nb_{j_1}) \ | \ b_{j_1} \in \mathbb{R}]) \subset \text{col}(F(:,I))$ whose dimension is strictly smaller than $N$, which contradicts Lemma 4.4. So it means that there is $b_{j_1} \in \mathbb{R}$ so that $\text{rank}(F)$ increases by 1. By assumption $n \geq N$ it follows that $p \geq N - r$, and thus we can choose $\{b_{j_2}, \ldots, b_{j_{N-r}}\}$ similarly to obtain $\text{rank}(F) = N$.

### H. Proof of Lemma 5.3

We need to show that there is a continuous path from $(W,V)$ to $(W',V')$ for some $V' \in \mathbb{R}^{n \times p}$, so that the output function, defined by $Z := \sigma(XW)V$, is invariant along the path. Let $F = \sigma(XW) \in \mathbb{R}^{N \times n}$ and $F' = \sigma(XW')$. It holds $Z = FW$. Let $I$ resp. $I'$ denote the maximum subset of linearly independent columns of $F$ resp. $F'$ so that $\text{rank}(F(:,I)) = \text{rank}(F(:,I')) = N$, and $\tilde{I}$ and $\tilde{I}'$ be their complements. By the rank condition, we have $|I| = |I'| = N$. Since $\text{rank}(F) = N < n$, we can apply Lemma 4.3 to the tuple $(F,V,\tilde{I})$ to arrive at some point where the output $Z$ is independent of $F(:,\tilde{I})$. From here, we can update $W(:,\tilde{I})$ arbitrarily so that it does not affect $Z$ because any change to these weights only lead to changes on $F(:,\tilde{I})$ which however has no influence on $Z$. So by taking a direct line segment from the current value of $W(:,\tilde{I})$ to $W'(I',i)$, we achieve $W(:,\tilde{I}) = W'(I',i)$. We refer to this step below as a copy step. Note here that since $n \geq 2N$ by assumption, we must have $|\tilde{I}| \geq |I'|$. Moreover, if $|\tilde{I}| > |I'|$ then we can simply ignore the redundant space in $W(:,\tilde{I})$.

Now we already copy $W'(I',i)$ into $W(:,\tilde{I})$, so it holds that $\text{rank}(F(:,\tilde{I})) = \text{rank}(F'(I',i')) = N$. Let $K = I' \cap I$ and $J = I' \cap \tilde{I}$ be disjoint subsets so that $I' = K \cup J$. Suppose w.l.o.g. that the above copy step has been done in such a way that $W(:,\tilde{I} \cap I') = W'(I',\tilde{I'})$. Now we apply Lemma 4.3 to $(F,V,\tilde{I})$ to arrive at some point where $Z$ is independent of $F(:,\tilde{I})$, and thus we can easily obtain $W(:,I \cap I') = W'(I',\tilde{I'})$ by taking a direct line segment between these weights. So far, all the rows of $W'(I',K \cup J)$ have been copied into $W(I',I')$ at the right positions so we obtain that $W(:,I') = W'(I',I')$. It follows that $\text{rank}(F(I',i')) = \text{rank}(F'(I',i')) = N$ and thus we can apply Lemma 4.3 to $(F,V',I')$ to arrive at some other point where $Z$ is independent of $F(:,I')$. From here we can easily obtain $W(:,I') = W'(I',I')$ by taking a direct line segment between these variables. Till now we already have $W = W'$. Moreover, all the paths which we have followed leave the output $Z$ invariant.

### I. Proof of Theorem 6.1

**Case 1:** $\min \{n_1, \ldots, n_{L-1}\} \geq N$. Let $\theta = (W_t,b_t)_{t=1}^T$ be an arbitrary point of some strict sublevel set $L^*_\alpha$, for some $\alpha > p^*$. We will show that there is a continuous descent path
starting from \( \theta \) on which the loss is non-increasing and gets arbitrarily close to \( p^* \). Indeed, for every \( \epsilon \) arbitrarily close to \( p^* \) and \( \epsilon \leq \alpha \), let \( \hat{Y} \in \mathbb{R}^{N \times m} \) be such that \( \varphi(\hat{Y}) \leq \epsilon \). Since \( X \) has distinct rows, \( n_1 \geq N \), and the activation \( \sigma \) satisfies Assumption 2.2, an application of Lemma 5.2 to \( (X, W_1, b_1, W_2) \) shows that there is a continuous path with constant loss which leads \( \theta \) to some other point where the output at the first hidden layer is full rank. So we can assume w.l.o.g. that it holds for \( \theta \) that rank \( F_1 = N \). By assumption \( n_1 \geq N \) and \( F_1 \in \mathbb{R}^{N \times n_1} \), it follows that \( F_1 \) must have distinct rows, and thus by applying Lemma 5.2 again to \( (F_1, W_2, b_2, W_3) \) we can assume w.l.o.g. that rank \( F_2 = N \). By repeating this argument to higher layers using our assumption on the width, we can eventually arrive at some \( \theta = (W_i, b_i)_{i=1}^L \) where rank \( F_{L-1} = N \). Thus there must exist \( W_{L-1}^* \in \mathbb{R}^{n_{L-1} \times m} \) so that \( F_{L-1}W_{L-1}^* = \hat{Y} - 1_Nb_l^T \). Consider the line segment \( W_L(\lambda) = (1-\lambda) W_L + \lambda W_{L}^* \), then it holds by convexity of \( \varphi \) that

\[
\Phi\left( (W_i, b_i)_{i=1}^{L-1}, (W_L(\lambda), b_L) \right) = \varphi(F_{L-1}W_L(\lambda) + 1_NB_l^T) \\
= \varphi\left( (1-\lambda)(F_{L-1}W_L + 1_NB_l^T) + \lambda(F_{L-1}W_{L-1}^* + 1_NB_l^T) \right) \\
\leq (1-\lambda)\varphi(F_L) + \lambda\varphi(\hat{Y}) \\
< (1-\lambda)\alpha + \lambda\epsilon \leq \alpha.
\]

Thus the whole line segment is contained in \( L_\alpha \). By plugging \( \lambda = 1 \) we obtain \( \left( (W_i, b_i)_{i=1}^{L-1}, (W_L, b_L) \right) \in L_\alpha \).

Moreover, it holds \( \Phi\left( (W_i, b_i)_{i=1}^{L-1}, (W_L, b_L) \right) = \varphi(\hat{Y}) \leq \epsilon \). As \( \epsilon \) can be chosen arbitrarily close to \( p^* \), we conclude that \( \Phi \) can be made arbitrarily close to \( p^* \) in every strict sublevel set which implies that \( \Phi \) has no bad local valleys.

**Case 2:** \( \min\{n_1, \ldots, n_{L-1}\} \geq 2N \). Our first step is similar to the first step in the proof of Theorem 5.1, which we repeat below for completeness. Let \( \theta = (W_i, b_i)_{i=1}^{L-1}, \theta' = (W'_i, b'_i)_{i=1}^{L-1} \) be arbitrary points in some sublevel set \( L_\alpha \). It is sufficient to show that there is a connected path between \( \theta \) and \( \theta' \) on which the loss is not larger than \( \alpha \). In the following, we denote \( F_k \) and \( F'_k \) as the output at a layer \( k \) for \( \theta \) and \( \theta' \) respectively. The output at the first layer is:

\[
F_1 = \sigma([X, 1_N], [W_1^T, b_1]^T), \\
F'_1 = \sigma([X, 1_N], [W'_1^T, b'_1]^T).
\]

By applying Lemma 5.2 to \( (X, W_1, b_1, W_2) \) and \( (X, W'_1, b'_1, W'_2) \) we can assume w.l.o.g. that both \( F_1 \) and \( F'_1 \) have full rank, since otherwise there is a continuous path starting from each point and leading to some other point where the rank condition is fulfilled and the network output at second layer is invariant on the path. Once \( F_1 \) and \( F'_1 \) have full rank, we can apply Lemma 5.3 to \( ([X, 1_N], [W_1^T, b_1]^T, W_2, [W_1^T, b'_1]^T) \) in order to drive \( \theta \) to some other point where the parameters of the first layer are all equal to the corresponding ones of \( \theta' \). So we can assume w.l.o.g. that \( (F_1, b_1) = (F'_1, b'_1) \).

Once the network parameters of \( \theta \) and \( \theta' \) coincide at the first hidden layer, we can view the output of this layer, which is equal for both points (i.e., \( F_1 = F'_1 \)), as the new training data for the subnetwork from layer 2 till layer \( L \). Same as before, we first apply Lemma 5.2 to \( (F_1, W_2, b_2, W_3) \) and \( (F'_1, W_2, b'_2, W'_3) \) to drive \( \theta \) and \( \theta' \) respectively to other new points where both \( F_2 \) and \( F'_2 \) have full rank. Note that this path only acts on \( (W_2, b_2, W_3) \) and thus leaves everything else below layer 2 invariant, in particular we still have \( F_1 = F'_1 \). Then we can apply Lemma 5.3 again to the tuple \( ([F_1, 1_N], [W_2^T, b_2]^T, W_3, [W_2^T, b'_2]^T) \) to drive \( \theta \) to some other point where \( (W_2, b_2) = (W'_2, b'_2) \).

By repeating the above argument to the last hidden layer, we can make all network parameters of \( \theta \) and \( \theta' \) coincide for all layers, except the output layer. In particular, the path that each \( \theta \) and \( \theta' \) has followed has invariant loss. The output of the last hidden layer for these points is \( A := F_{L-1} = F'_{L-1} \).

The loss at these two points can be rewritten as

\[
\Phi(\theta) = \varphi\left([A, 1_N], \begin{bmatrix} W_L \\ b_L^T \end{bmatrix} \right), \\
\Phi(\theta') = \varphi\left([A, 1_N], \begin{bmatrix} W'_L \\ b'_L^T \end{bmatrix} \right).
\]

Since \( \varphi \) is convex, the line segment

\[
(1-\lambda) \begin{bmatrix} W_L \\ b_L^T \end{bmatrix} + \lambda \begin{bmatrix} W'_L \\ b'_L^T \end{bmatrix}
\]

must yield a continuous descent path between \( (W_L, b_L) \) and \( (W'_L, b'_L) \), and so the loss of every point on this path cannot be larger than \( \alpha \). Moreover, this path connects \( \theta \) and \( \theta' \) together, and thus \( L_\alpha \) has to be connected.