Diversity Leads to Generalization in Neural Networks

Bo Xie\(^1\), Yingyu Liang\(^2\), Le Song\(^1\)
\(^1\)Georgia Institute of Technology
bo.xie@gatech.edu, lsong@cc.gatech.edu
\(^2\)Princeton University
yingyul@cs.princeton.edu

Abstract

Neural networks are a powerful class of functions that can be trained with simple gradient descent to achieve state-of-the-art performance on a variety of applications. Despite their practical success, there is a paucity of results that provide theoretical guarantees on why they are so effective. Lying in the center of the problem is the difficulty of analyzing the non-convex loss function with potentially numerous local minima and saddle points. Can neural networks corresponding to the stationary points of the loss function learn the true labeling function? If yes, what are the key factors contributing to such generalization ability?

In this paper, we provide answers to these questions by analyzing one-hidden-layer neural networks with ReLU activation, and show that despite the non-convexity, neural networks with diverse units can learn the true function. We bypass the non-convexity issue by directly analyzing the first order optimality condition, and show that the loss is bounded if the smallest singular value of the “extended feature matrix” is large enough. We make novel use of techniques from kernel methods and geometric discrepancy, and identify a new relation linking the smallest singular value to the spectrum of a kernel function associated with the activation function and to the diversity of the units. Our results also suggest a novel regularization function to promote unit diversity for potentially better generalization ability.

1 Introduction

In this paper, we show that despite the non-convexity, neural networks with diverse units can learn the true function. We bypass the hurdle of non-convexity by directly analyzing the first order optimality condition of the learning problem, which implies that the training loss is bounded if the minimum singular value of the extended feature matrix is large enough. Bounding the singular value is challenging because it entangles the nonlinear activation function, the weights and data in a complicated way. Unlike most previous attempts, we directly analyze the effect of nonlinearity without assuming independence of the activation patterns from actual data; in fact, the dependence of the patterns on the data and the unit weights underlies the key connection to activation kernel spectrum and the diversity of the units.

We have constructed a novel proof, which makes use of techniques from geometric discrepancy and kernel methods, and have identified a new relation linking the smallest singular value to the diversity of the units and the spectrum of a kernel function associated with the unit. Our results also suggest a novel regularization scheme to promote unit diversity for potentially better generalization.

2 Related work

In [3], the authors analyze the loss surface of a special random neural network through spin-glass theory and show that for many large-size networks, there is a band of exponentially many local optima, whose loss is small and close to that of a global optimum. A similar work shows that all local

We will focus on a special class of data distributions where the input is a stationary point has zero gradient, so if \( D \parallel (\text{SGD}) \). In other words, typically we only have \( \partial L/\partial W \parallel \leq \epsilon \), and we use the current assumption for simplicity. However, this is insufficient to guarantee low training loss since small gradient can still lead to large loss.

3 Problem setting and preliminaries

We will focus on a special class of data distributions where the input \( x \in \mathbb{R}^d \) is drawn uniformly from the unit sphere \( \mathbb{S}^n \) and assume the label \( |y| \leq Y \) where \( Y \) is a constant. We consider the following hypothesis class:

\[
\mathcal{F} = \left\{ \sum_{k=1}^{n} v_k \sigma (w_k^\top x) : v_k \in \{ \pm 1 \}, W = \{ w_k \} \in \mathcal{F}_W \right\}, \quad \text{where} \quad \mathcal{F}_W = \left\{ W = \{ w \} : \sum_{k=1}^{n} ||w_k|| \leq C_W \right\}
\]

and \( \sigma (\cdot) = \max \{ 0, \cdot \} \) is the rectified linear unit (ReLU) activation function, \( \{ w_k \} \) and \( \{ v_k \} \) are the unit weights and combination coefficients respectively, \( n \) is the number of units, and \( C_W \) is some constant. We restrict \( v_k \in \{ -1, 1 \} \) due to the positive homogeneity of ReLU. That is, the magnitude of \( v_k \) can always be scaled into the corresponding \( w_k \).

Given a set of \( i.i.d. \) training examples \( \{ x_l, y_l \}_{l=1}^{m} \subset \mathbb{R}^d \times \mathbb{R} \), we want to find a function \( f \in \mathcal{F} \) which minimizes the following least-squares loss function

\[
L(f) = \frac{1}{2m} \sum_{l=1}^{m} (y_l - f(x_l))^2.
\]

3.1 First order optimality condition

The gradient of the empirical loss w.r.t. \( w_k \) is\(^2\)

\[
\frac{\partial L}{\partial w_k} = \frac{1}{m} \sum_{l=1}^{m} (f(x_l) - y_l) v_k \sigma'(w_k^\top x_l) x_l,
\]

for all \( k = 1, \ldots, n \). We will express this collection of gradient equations using matrix notation. Define the “extended feature matrix” as

\[
D = \begin{pmatrix}
  v_1 \sigma'(w_1^\top x_1) x_1 & \cdots & v_1 \sigma'(w_1^\top x_m) x_m \\
  \vdots & \ddots & \vdots \\
  v_n \sigma'(w_n^\top x_1) x_1 & \cdots & v_n \sigma'(w_n^\top x_m) x_m
\end{pmatrix},
\]

and the residual as

\[
r = \frac{1}{m} (f(x_1) - y_1, \ldots, f(x_m) - y_m)^\top.
\]

Then we have

\[
\frac{\partial L}{\partial W} = \left( \frac{\partial L}{\partial w_1}, \ldots, \frac{\partial L}{\partial w_n} \right)^\top = D r.
\]

A stationary point has zero gradient, so if \( D \in \mathbb{R}^{dn \times m} \) has full column rank, then immediately \( r = 0 \), i.e., it is actually a global optimal. Since \( nd > m \) is necessary for \( D \) to have full column rank, we assume this throughout the paper. However, in practice, we will not have the gradient being exactly zero because, e.g., we stop the algorithm in finite steps or because we use stochastic gradient descent (SGD). In other words, typically we only have \( ||\partial L/\partial W|| \leq \epsilon \), and \( D \) being full rank is insufficient since small gradient can still lead to large loss. More specifically, let \( s_m(D) \) be the minimum singular

\(^1\)The restriction of input to the unit sphere is not too stringent since many inputs are already normalized. Furthermore, it is possible to relax uniform assumption to sub-gaussian rotationally invariant distributions, but we use the current assumption for simplicity.

\(^2\)Note that even though ReLU is not differentiable, we can use its sub-gradient \( \sigma'(u) = \mathbb{I} [u \geq 0] \).
value of $D$, we have
\[ \|r\| \leq \frac{1}{s_m(D)} \left\| \frac{\partial L}{\partial W} \right\|. \] (4)

### 3.2 Spectrum decay of activation kernel

We will later show that $s_m(D)$ is related to the decay rate of the kernel spectrum associated with the activation function. Specifically, for an activation function $\sigma(w^\top x)$, we can define a kernel function
\[ g(x, y) = \mathbb{E}_w[\sigma'(w^\top x)\sigma'(w^\top x)(x, y)] \]
where $\mathbb{E}_w$ is over $w$ uniformly distributed on a sphere. In fact, it is a dot-product kernel and we can decompose it with spherical harmonic decomposition:
\[ g(x, y) = \sum_{u=1}^{\infty} \gamma_u \phi_u(x) \phi_u(y), \]
where the eigenvalues are ordered $\gamma_1 \geq \cdots \geq \gamma_m \geq \cdots \geq 0$ and the bases $\phi_u(x)$ are spherical harmonics. The $m$-th eigenvalue $\gamma_m$ will be related to $s_m(D)$.

### 3.3 Weight discrepancy

Another key factor in the analysis is the diversity of the unit weights, measured by its geometric discrepancy $[2][3][4]$. Given a set of $n$ points $W = \{w_k\}_{k=1}^n$ on the unit sphere, the discrepancy of $W$ w.r.t. a measurable set $S_{xy} = \{w \in \mathbb{S}^{d-1} : w^\top x \geq 0, w^\top y \geq 0\}$ associated with a pair of points $(x, y)$ on the unit sphere is defined as
\[ \text{dsp}(W, S_{xy}) = \frac{1}{n} |W \cap S_{xy}| - A(S_{xy}), \]
where $A(S_{xy})$ is the normalized area of $S_{xy}$ (i.e., the area of the whole sphere $A(\mathbb{S}^{d-1}) = 1$).

Based on the collection of such sets $S = \{S_{xy} : x, y \in \mathbb{S}^{d-1}\}$, we can define two discrepancy measures relevant to ReLU units:
\[ L_\infty(W, S) = \sup_{S \in S} |\text{dsp}(W, S)|, \]
\[ L_2(W, S) = \frac{1}{n} \mathbb{E}_{x,y} [\text{dsp}(W; S_{xy})]^2 \]
where the expectation is taken over $x, y$ uniformly on the sphere. We use $L_\infty(W)$ and $L_2(W)$ as their shorthands. Both discrepancies measure how diverse the points $W$ are.

### 4 Main results

Recall that $\gamma_m$ is the $m$-th eigenvalue of the kernel in (6), and we define $\beta$ such that $\gamma_m = \Omega(m^{-\beta})$ for some $\beta < 1$.

**Theorem 1** Let $0 < \delta, \delta' < 1$, and suppose $n = \tilde{\Omega}(m^{\delta})$ and $d = \tilde{\Omega}(m^{\delta})$. Then there exists a set $G_W \subseteq F_W$ that takes up $1 - \delta'$ fraction of measure of $F_W$ such that with probability at least $1 - cm^{-\log m} - \delta$, for any $W \in G_W$,
\[ \frac{1}{2} \mathbb{E}(f(x; v, W) - y)^2 \leq c \left\| \frac{\partial L}{\partial W} \right\|^2 + c'(C_W + Y)^2 \sqrt{\frac{1}{m} \log \frac{1}{\delta}}. \]

Here $\tilde{\Omega}$ hides logarithmic terms $\log m \log \frac{1}{\delta} \log \frac{1}{\delta}$.

By the theorem, when we obtain an solution $W \in G_W$ with gradient $\|\partial L/\partial W\|^2 \leq \epsilon$, the generalization error is $O(\epsilon + 1/\sqrt{m})$. This essentially means that although non-convex, the loss function is well behaved, and learning is not difficult over this set. Furthermore, a randomly sampled set of weights $W$ are likely to fall into this set. This then suggests an explanation for the practical success of training with random initialization: after initialization, the parameters w.h.p. fall into the set, then stay inside during training, and finally get to a point with small gradient and small error.

**High level intuition** Due to space limitations, we only describe the high level intuition and proof sketch for bounding the minimum singular value. Please see the supplementary material for full proof.
We know where \( L \) is necessarily connected to the activation function and the diversity of the weights. For example, if \( \sigma'(t) \) is very small for all \( t \), then the smallest singular value is expected to be very small. For the weights, if \( d < m \) (the interesting case) and all \( w_k \)'s are the same, then \( D \) cannot have rank \( m \). If \( w_k \)'s are very similar to each other, then one would expect the smallest singular value to be very small or even zero. Therefore, some notion of diversity of the weights are needed.

The analysis begins by considering the matrix \( G_n = D^T D/n \). It suffices to bound \( \lambda_m(G_n) \), the \( m \)-th (and the smallest) eigenvalue of \( G_n \). To do so, we introduce a matrix \( G \) whose entries \( G(i, j) = \mathbb{E}_w[G_n(i, j)] \) where the expectation \( \mathbb{E}_w \) is taken assuming \( w_k \)'s are uniformly random on the unit sphere. The intuition is that when \( w \) is uniformly distributed, \( \sigma'(w^\top x) \) is most independent of the actual value of the \( x \), and the matrix \( D \) should have the highest chance of having large smallest singular value. We will introduce \( G \) as an intermediate quantity and subsequently bound the spectral difference between \( G_n \) and \( G \). Roughly speaking, we break the proof into two steps

\[
\lambda_m(G_n) \geq \lambda_m(G) - \| G - G_n \| 
\]

where \( \| G - G_n \| \) is the spectral norm of the difference.

For the first term, we observe that \( G \) has a particular nice form: \( G(i, j) = g(x_i, x_j) \), the kernel defined in \( \text{[6]} \). This allows us to apply the eigendecomposition of the kernel and positive definite matrix concentration inequality to bound \( \lambda_m(G) \), which turns out to be around \( m \gamma_m/2 \).

For the second term, we use the geometric discrepancy to characterize \( \| G - G_n \| \) and show it is small for most \( W \). In particular, the entries in \( G - G_n \) can be viewed as the kernel of some U-statistics, hence concentration bounds can be applied. The expected U-statistics turns out to be the \( (L_2(W))^2 \), which has a closed form and can be shown to be small.

Our key technical lemma is a lower bound on the smallest singular value of the matrix \( D \).

**Lemma 2** With probability \( \geq 1 - m \exp(-m \gamma_m/8) - 2m^2 \exp(-4 \log^2 d) \) \( - \delta \), we have

\[
(s_m(D))^2 \geq nm \gamma_m/2 - c\rho(W) ,
\]

where

\[
\rho(W) = \frac{m \log d}{\sqrt{d}} \left( \sqrt{L_\infty(W) L_2(W)} \left( \frac{4}{m} \log \frac{1}{\delta} \right)^{1/4} + L_\infty(W) \sqrt{\frac{4}{3m} \log \frac{1}{\delta} + L_2(W)} \right) + L_\infty(W).
\]

We know \( L_\infty(W) \) is bounded by \( 2 \), so the question is how large is \( L_2(W) \). In the next lemma, we show it is small for most random \( W \).

**Lemma 3** There exists a constant \( c_g \), such that for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \) over \( W = \{ w_i \}_{i=1}^n \) that are sampled from the unit sphere uniformly at random,

\[
(L_2(W))^2 \leq c_g \left( \sqrt{\frac{\log d}{nd} \log \frac{1}{\delta} + \frac{1}{n} \log \frac{1}{\delta}} \right).
\]

One limitation of the analysis is that we have not analyzed how to obtain a solution with both small discrepancy and small gradient. To reinforce solutions with small discrepancy, we propose a novel regularization term to minimize \( L_2 \) discrepancy. Detailed experiments are in the supplementary.

**5 Conclusion**

We have analyzed one-hidden-layer neural networks and identified novel conditions when they achieve small training and test errors despite the non-convexity of the loss function.

Although we focus on a least-square loss function and uniform input distribution, the analysis technique can be readily extended to other loss function and input distributions. At the moment, our analysis is still limited in the sense that it is independent of the actual algorithm. In the future work, we will explore the interplay between the discrepancy and gradient descent. In addition, we will further investigate the issue of designing an algorithm that guarantees good discrepancy thus small errors, possibly in a way similar to \( \text{[7]} \) in low-rank recovery problems.
References


