Diverse Neural Network Learns True Target Functions

Bo Xie
Georgia Tech

Yingyu Liang
Princeton University

Le Song
Georgia Tech

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 target function? If yes, what are the key factors contributing to such nice optimization properties?

In this paper, we answer these questions by analyzing one-hidden-layer neural networks with ReLU activation, and show that despite the non-convexity, neural networks with diverse units have no spurious local minima. We bypass the non-convexity issue by directly analyzing the first order optimality condition, and show that the loss can be made arbitrarily small if the minimum 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.

1 Introduction

Neural networks are a powerful class of nonlinear functions which have been successfully deployed in a variety of machine learning tasks. In the simplest form, neural networks with one hidden layer are linear combinations of nonlinear basis functions (units),

$$f(x) = \sum_{k=1}^{n} v_k \sigma(w_k^\top x) \tag{1}$$

where $\sigma(w_k^\top x)$ is a basis function with weights w_k, and v_k is the corresponding combination coefficient. Learning with neural networks involves adapting both the combination coefficients and the basis functions at the same time, usually by minimizing the empirical loss

$$L(f) = \frac{1}{2m} \sum_{l=1}^{m} \ell(y_l, f(x_l)) \tag{2}$$

with first-order methods such as (stochastic) gradient descent. It is believed that basis function adaptation is a crucial ingredient for neural networks to achieve more compact models and better performance [Barron, 1993, Yang et al., 2014].

However, the empirical loss minimization problem involved in neural network training is non-convex with potentially numerous local minima and saddle points. This makes formal analysis of training neural networks very challenging. Given the empirical success of neural networks, a sequence of important and urgent scientific questions need to be investigated: Can neural networks corresponding to stationary points of the empirical loss learn the true target function? If the answer is yes, then what are the key factors contributing to their nice optimization properties? Based on these understandings, can we design better regularization schemes and learning algorithms to improve the training of neural networks?

In this paper, we provide partial answers to these questions by analyzing one-hidden-layer neural networks with rectified linear units (ReLU) in a least-squares regression setting. We show that neural networks with diverse units have no spurious local minima. More specifically, we show that the training loss of neural networks decreases in proportion to $\|\partial L/\partial W\|^2 / s_m^2(D)$ where $\partial L/\partial W$ is the gradient and $s_m(D)$ is the minimum singular value of the extended feature matrix D (defined in Section 3.1). The minimum singular value is lower bounded by two terms, where the first term is related to the spectrum property of the kernel function associate with the activation $\sigma(\cdot)$, and the second term quantifies the diversity of the units, measured by the

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

classical notion of geometric discrepancy of a set of vectors. Essentially, the slower the decay of the spectrum, the better the optimization landscape; the more diverse the unit weights, the more likely stationary points will result in small training loss and generalization error.

We bypass the hurdle of non-convexity by directly analyzing the first order optimality condition of the learning problem, which implies that there are no spurious local minima 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. More specifically,

- We identify and separate two factors in the minimum singular value: 1) an ideal spectrum that is related to the kernel of the activation function and an ideal configuration of diverse unit weights; 2) deviation from the ideal spectrum measured by how far away actual unit weights are from the diverse configuration. This new perspective reveals benign conditions in learning neural networks.
- We characterize the deviation from the ideal diverse weight configuration using the concept of discrepancy, which has been extensively studied in the geometric discrepancy theory. This reveals an interesting connection between the discrepancy of the weights and the training loss of neural networks. Therefore, it serves as a clean tool to analyze and verify the learning and the generalization ability of the networks.

Whenever possible, we corroborate our theoretical analysis with numerical simulations. These numerical results include computing and verifying the relationship between the discrepancy of a learned neural network and the minimum singular value. Additionally, we measure the effects on the discrepancy with and without regularization. In all these examples, the experiments match with the theory nicely and they accord with the practice of using gradient descent to learn neural networks.

2 Related work

Kernel methods have many commonalities with one-hidden-layer neural networks. The random feature perspective [Rahimi and Recht, 2009, Cho and Saul, 2009] views kernels as linear combinations of nonlinear basis functions, similar to neural networks. The difference between the two is that the weights are random in kernels while in neural networks they are learned. Using learned weights leads to considerable smaller models as shown in [Barron, 1993]. However it is a non-convex problem and it is difficult to find the global optima. *e.g.*, one-hidden-layer networks are NP-complete to learn in the worst case [Blum and Rivest, 1993]. We will make novel use of techniques from kernel methods to analyze learning in neural networks.

The empirical success of training neural networks with simple algorithms such as gradient descent has motivated researchers to explain their surprising effectiveness. In [Choromanska et al., 2015], 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. The analyzed polynomial network is different from the actual neural network being used which typically contains ReLU nowadays. Moreover, the analysis does not lead to a generalization guarantee for the learned neural network.

A similar work shows that all local optima are also global optima in linear neural networks [Kawaguchi, 2016]. However their analysis for nonlinear neural networks hinges on independence of the activation patterns from the actual data, which is unrealistic. Some other works try to argue that gradient descent is not trapped in saddle points [Lee et al., 2016, Ge et al., 2015], as was suggested to be the major obstacle in optimization [Dauphin et al., 2014]. There is also a seminal work using tensor method to avoid the non-convex optimization problem in neural network [Janzamin et al., 2015]. However, the resulting algorithm is very different from typically used algorithms where only gradient information of the empirical loss L is used.

[Soudry and Carmon, 2016] is the closest to our work, which shows that zero gradient implies zero loss for all weights except an exception set of measure zero. However, this is insufficient to guarantee low training loss since small gradient can still lead to large loss. Furthermore, their analysis does not characterize the exception set and it is unclear a priori whether the set of local minima fall into the exception set.

Some recent works [Mariet and Sra, 2015, Xie et al., 2015, Littwin and Wolf, 2016] also focused on promoting diversity in neural network weights however their results are not concerned with guaranteeing global optima.

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.[1]

[1] It is possible to relax this assumption to sub-gaussian rotationally invariant distributions, but we use the current assumption for simplicity.

Furthermore, we consider the following hypothesis class:

$$\mathcal{F} = \left\{ \sum_{k=1}^{n} v_k \sigma(w_k^\top x) : v_k \in \{\pm 1\}, \sum_{k=1}^{n} \|w_k\| \le C_W \right\}$$

where $\sigma(\cdot) = \max\{0, \cdot\}$ 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,

$$f(x) = \sum_{k=1}^{n} v_k \sigma(w_k^\top x) = \sum_{k=1}^{n} \frac{v_k}{|v_k|} \sigma(|v_k| w_k^\top x). \quad (3)$$

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 $\{x_l, y_l\}_{l=1}^m \subseteq \mathbb{R}^d \times \mathbb{R}$, we want to find a function $f \in \mathcal{F}$ which minimizes the following least-squares loss function

$$L(f) = \frac{1}{2m} \sum_{l=1}^{m} \left(y_l - f(x_l) \right)^2. \quad (4)$$

Typically, gradient descent over $L(f)$ is used to learn all the parameters in f, and a solution with small gradient is returned at the end.[2] However, adjusting the bases $\{w_k\}$ leads to a non-convex optimization problem, and there is no theoretical guarantee on the training and test performance.

Our primary goal is to identify conditions under which there are no spurious local minima. In particular, let $W := (w_1^\top, w_2^\top, \ldots, w_k^\top)^\top$ be the column concatenation of the unit parameters. We need to identify a set $\mathcal{G}_W$ such that when gradient descent outputs a solution $W \in \mathcal{G}_W$ with the gradient norm $\|\partial L / \partial W\|$ smaller than ϵ, then the training and test errors can be bounded by ϵ. Ideally, $\mathcal{G}_W$ should have clear characterization that can be easily verified, and should contain most W in the parameter space (especially those solutions obtained in practice).

On notation, we will use c, c' or C, C' to denote constants and its value may change from line to line.

3.1 First order optimality condition

In this section, we will rewrite the set of first order optimality conditions for minimizing the empirical loss L. This rewriting motivates the direction of our later analysis. More specifically, the gradient of the empirical loss w.r.t. w_k is

$$\frac{\partial L}{\partial w_k} = \frac{1}{m} \sum_{l=1}^{m} \left(f(x_l) - y_l \right) v_k \sigma'(w_k^\top x_l) x_l, \quad (5)$$

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

$$D = \begin{pmatrix} v_1 \sigma'(w_1^\top x_1) x_1 & \cdots & v_1 \sigma'(w_1^\top x_m) x_m \\ \cdots & \cdots & \cdots \\ v_k \sigma'(w_k^\top x_1) x_1 & \cdots & v_k \sigma'(w_k^\top x_m) x_m \\ \cdots & \cdots & \cdots \\ v_n \sigma'(w_n^\top x_1) x_1 & \cdots & v_n \sigma'(w_n^\top x_m) x_m \end{pmatrix},$$

and the residual as

$$r = \frac{1}{m} \left(f(x_1) - y_1, \cdots, f(x_m) - y_m \right)^\top.$$

Then we have

$$\frac{\partial L}{\partial W} := \left(\frac{\partial L}{\partial w_1}^\top, \ldots, \frac{\partial L}{\partial w_n}^\top \right)^\top = D\, r. \quad (6)$$

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 $\|\partial L / \partial W\| \le \epsilon$, 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 value of D, we have

$$\|r\| \le \frac{1}{s_m(D)} \left\| \frac{\partial L}{\partial W} \right\|. \quad (7)$$

We can see that $s_m(D)$ needs to be large enough for the residual to be small. Thus it is important to identify conditions to lower bound $s_m(D)$ away from zero, which will be the focus of the paper.

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. More specifically, for an activation function $\sigma(w^\top x)$, we can define the following kernel function

$$g(x, y) = \mathbb{E}_w [\sigma'(w^\top x) \sigma'(w^\top x) \langle x, y \rangle] \quad (8)$$

where $\mathbb{E}_w$ is over w uniformly distributed on a sphere.

In particular, for ReLU, the kernel has a closed form

$$g(x, y) = \left(\frac{1}{2} - \frac{\arccos \langle x, y \rangle}{2\pi} \right) \langle x, y \rangle. \quad (9)$$

In fact, it is a dot-product kernel and its spectrum can be obtained through spherical harmonic decomposition:

$$g(x, y) = \sum_{u=1}^{\infty} \gamma_u \phi_u(x) \phi_u(y), \quad (10)$$

where the eigenvalues are ordered $\gamma_1 \ge \cdots \ge \gamma_m \ge \cdots \ge 0$ and the bases $\phi_u(x)$ are spherical harmonics. The m-th eigenvalue γ_m will be related to $s_m(D)$.

For each spherical harmonic of order t, there are $N(d, t) = \frac{2t+d-2}{n} \begin{pmatrix} t+d-3 \\ t-1 \end{pmatrix}$ basis functions sharing the same

[2] Note that even though ReLU is not differentiable, we can still use its sub-gradient by defining $\sigma'(u) = \mathbb{I}[u \ge 0]$.

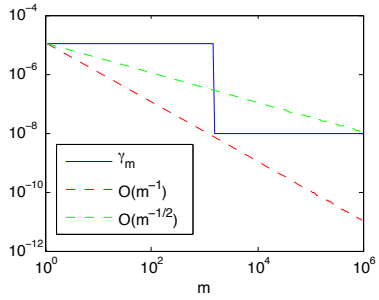

Figure 1: The spectrum decay of the kernel associated with ReLU. We set $d = 1500$. It is decays slower than $O(1/m)$ for a large range of m.

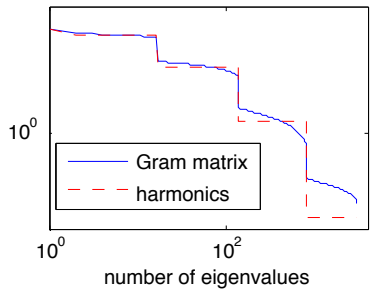

Figure 2: The spectrum of a Gram matrix concentrates around the spherical harmonic spectrum of the kernel.

eigenvalue. Therefore, the spectrum has a step like shape where each step is of length $N(d, t)$. Especially, for high dimensional input x, the number of such basis functions with large eigenvalues can be very large. Figure 1 illustrates the spectrum of the kernel for $d = 1500$, and it is about $\Omega(m^{-1})$ for a large range of m. For more details about the decomposition, please refer to Appendix A.

Such step like shape also appears in the Gram matrix associated with the kernel. Figure 2 compares the spectra of the kernel of $d = 15$ and the corresponding Gram matrix with $m = 3000$. We can see the spectrum of the Gram matrix closely resembles that of the kernel. Such concentration phenomenon underlies the reason why the spectrum of $D^\top D$ is closely related to the corresponding kernel.

3.3 Weight discrepancy

Another key factor in the analysis is the diversity of the unit weights, measured by its geometric discrepancy [Bilyk and Lacey, 2015]. Given a set of n points $W = \{w_k\}_{k=1}^n$ on the unit sphere $\mathbb{S}^{d-1}$, the discrepancy of W w.r.t. a measurable set $S \subseteq \mathbb{S}^{d-1}$ is defined as

$$\text{dsp}(W, S) = \frac{1}{n} |W \cap S| - A(S), \tag{11}$$

where $A(S)$ is the normalized area of S (i.e., the area of the whole sphere $A(\mathbb{S}^{d-1}) = 1$). $\text{dsp}(W, S)$ quantifies the difference between the empirical measure of S induced by W and the measure of S induced by a uniform distribution.

By defining a collection $\mathcal{S}$ of such sets, we can summarize the difference in the empirical measure induced by W versus the uniform distribution over the sphere. More specifically, we will focus on the set of slices, each defined by a pair of inputs $x, y \in \mathbb{S}^{d-1}$, *i.e.*,

$$\mathcal{S} = \left\{S_{xy} : x, y \in \mathbb{S}^{d-1}\right\}, \text{ where}$$
$$S_{xy} = \left\{w \in \mathbb{S}^{d-1} : w^\top x \geq 0, w^\top y \geq 0\right\}. \tag{12}$$

Essentially, each S_{xy} defines a slice-shaped area on the sphere which is carved out by the two half spaces $w^\top x \geq 0$ and $w^\top y \geq 0$.

Based on the collection $\mathcal{S}$, we can define two discrepancy measures relevant to ReLU units. L_∞ discrepancy of W w.r.t. $\mathcal{S}$ is defined as

$$L_\infty(W, \mathcal{S}) = \sup_{S \in \mathcal{S}} |\text{dsp}(W, S)|, \tag{13}$$

and the L_2 discrepancy as

$$L_2(W, \mathcal{S}) = \sqrt{\mathbb{E}_{x,y} \text{dsp}(W, S_{xy})^2} \tag{14}$$

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. The more diverse the points, the smaller the discrepancy.

For our analysis, we slightly generalize the discrepancy for w_k's not on unit sphere, by setting

$$\text{dsp}(\{w_k\}_k, S) = \text{dsp}(\{w_k/\|w_k\|\}_k, S). \tag{15}$$

4 Main results

Our first result is a bound on the smallest singular value of D based on the spectrum of the activation kernel in (9) and the discrepancy of the weights in (13) and (14). Recall that γ_m is the m-th eigenvalue of the kernel in (9), and we define β such that $\gamma_m = \Omega(m^{-\beta})$ for some $\beta < 1$.

Theorem 1 *With probability* $\geq 1 - m \exp(-m\gamma_m/8) - 2m^2 \exp\left(-4 \log^2 d\right) - \delta$, *the following holds. For any* $\xi > 0, \eta > 0$, *and any* W *with* $L_2(W) = \tilde{O}\left(n^{-\xi} d^{-\eta}\right)$,

$$s_m(D)^2 \geq \Omega(nm^{1-\beta}) - \tilde{O}\left(\frac{n^{1-\xi/2} m^{3/4}}{d^{(1+\eta)/2}}\right) - \tilde{O}\left(\frac{nm^{1/2}}{d^{1/2}}\right) - \tilde{O}\left(\frac{n^{1-\xi} m}{d^{1/2+\eta}}\right).$$

Here the notation $\tilde{\Omega}$ hides logarithmic terms $\log d \log \frac{1}{\delta}$.

The theorem is stated in its general form. It bounds the smallest singular value in terms of the n, d, m and two parameters ξ, η quantifying how large $L_2(W)$ is. It is instructive to consider an interesting special case, with concrete values of ξ and η.

Corollary 2 *Suppose* $n = \tilde{\Omega}(m^\beta)$, $d = \tilde{\Omega}(m^\beta)$. *Suppose* $\xi = \eta = 1/4$ *so that* $L_2(W) = \tilde{O}(n^{-1/4} d^{-1/4})$. *Then with probability at least* $1 - cm^{-\log m} - \delta$,

$$s_m(D)^2 \geq cm$$

for some constant $c > 0$.

In the above corollary, the parameters $n = \tilde{\Omega}(m^\beta)$ and $d = \tilde{\Omega}(m^\beta)$ match the practical setting. It is also worth noting that when n becomes larger, the singular value increases as well. This accords with some recent empirical and theoretical observation that the loss function is more benign in large overspecified networks (*e.g.*, [Shamir, 2016]).

The setting that $L_2(W) = \tilde{O}(1/(nd)^{1/4})$ is also common. In fact, this is true for any parameters W in a large subset $\mathcal{G}_W$ of all the parameters satisfying $\sum_k \|w_k\| \leq C_W$; see Section 5.2 for details. So under typical settings, the minimum singular value of D is bounded away from zero.

Finally, it is interesting to compare the theorem to the results in [Soudry and Carmon, 2016], which shows that D is full rank with probability one under small perturbations. However, full-rankness alone is not sufficient since its smallest singular value could be extremely small leading to possibly huge training loss. Instead, we directly bound the smallest singular value and relate it to the activation and the diversity of the weights.

Equipped with the bound on the singular value, we are ready to bound the training loss and the generalization error.

Theorem 3 *Suppose the data are bounded $\|x\| \leq 1$ and $|y| \leq Y$. Then with probability $\geq 1 - m\exp(-m\gamma_m/8) - 2m^2\exp\left(-4\log^2 d\right) - \delta$, the following holds. For any W with $L_2(W) = \tilde{O}\left(n^{-\xi}d^{-\eta}\right)$ such that*

$$m^\beta \leq \tilde{O}\left\{d^{(1+\eta)/2}n^{\xi/2}m^{1/4}, d^{1/2}m^{1/2}, n^\xi d^{1/2+\eta}\right\},$$

we have

$$\frac{1}{2m}\sum_{\ell=1}^{m}(f(x_\ell) - y_\ell)^2 \leq \frac{cm^\beta}{n}\left\|\frac{\partial L}{\partial W}\right\|^2,$$

$$\frac{1}{2}\mathbb{E}(f(x) - y)^2 \leq \frac{cm^\beta}{n}\left\|\frac{\partial L}{\partial W}\right\|^2 + c'(C_W + Y)^2\sqrt{\frac{1}{m}\log\frac{1}{\delta}}.$$

The theorem is also in the general form. It shows that when the weights are diverse (*i.e.*, with good discrepancy), the training loss is proportional to the squared norm of the gradient. This implies a local minimum leads to a global minimum and the neural network learns the target function. The generalization error has an additional term $\tilde{O}\left(1/\sqrt{m}\right)$. So in this case, a neural network trained with sufficiently many data points generalizes well.

We have a corresponding result for Corollary 2 concerning the errors. Let $\mathcal{F}_W = \{W : \sum_k \|w_k\| \leq C_W\}$ denote the feasible set of W's. Then by Theorem 3 and Lemma 9 in Section 5.2 which bounds $L_2(W)$, we have

Corollary 4 *Let $0 < \delta, \delta' < 1$, and suppose $n = \tilde{\Omega}(m^\beta)$ and $d = \tilde{\Omega}(m^\beta)$. Then there exists a set $\mathcal{G}_W \subseteq \mathcal{F}_W$ that takes up $1 - \delta'$ fraction of measure of $\mathcal{F}_W$ such that with probability at least $1 - cm^{-\log m} - \delta$, for any $W \in \mathcal{G}_W$,*

$$\frac{1}{2m}\sum_{\ell=1}^{m}(f(x_\ell) - y_\ell)^2 \leq c\left\|\frac{\partial L}{\partial W}\right\|^2,$$

$$\frac{1}{2}\mathbb{E}(f(x) - y)^2 \leq c\left\|\frac{\partial L}{\partial W}\right\|^2 + c'(C_W + Y)^2\sqrt{\frac{1}{m}\log\frac{1}{\delta}}.$$

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

By the corollary, when we obtain an solution $W \in \mathcal{G}_W$ with gradient $\|\partial L/\partial W\|^2 \leq \epsilon$, the training loss is bounded by $O(\epsilon)$, and the generalization error is $O(\epsilon + 1/\sqrt{m})$. This essentially means that although non-convex, the loss function is well behaved, and there are no spurious local minima 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, which by our analysis, has small error.

Analysis roadmap Theorem 1 is proved in Section 5, $L_2(W)$ and $\mathcal{G}_W$ are characterized in Section 5.2, and the proof sketch of Theorem 3 is provided in Section 6. Due to space limit, we describe the proof sketch for the lemmas and provide the remaining proofs in the appendix.

Here we describe the high level intuition for bounding the minimum singular value. It 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 $G_n = D^\top D/n$. 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

$$\lambda_m(G_n) \geq \underbrace{\lambda_m(G)}_{\text{I. ideal spectrum}} - \underbrace{\|G - G_n\|}_{\text{II. discrepancy}}$$

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

For the first term in the lower bound, we observe that G has a

particular nice form: $G(i,j) = g(x_i, x_j)$, the kernel defined in (9). 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, when w_k's are indeed from the uniform distribution over the sphere, this can be bounded by concentration bounds. It turns out that when w_k's are not too far away from that, it is still possible to do so. Therefore, we use the geometric discrepancy to measure the diversity of the weights, and show that when they are sufficiently diverse, $\|G - G_n\|$ is small. 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.

5 Bounding the smallest singular value

Our key technical lemma is a lower bound on the smallest singular value of the extended feature matrix D.

Lemma 5 *With probability* $\geq 1 - m\exp(-m\gamma_m/8) - 2m^2\exp\left(-4\log^2 d\right) - \delta$, *we have*

$$s_m(D)^2 \geq nm\gamma_m/2 - cn\rho(W), \tag{16}$$

where

$$\begin{aligned}\rho(W) = &\frac{\log d}{\sqrt{d}}\sqrt{L_\infty(W)L_2(W)}m\left(\frac{4}{m}\log\frac{1}{\delta}\right)^{1/4} \\ &+ \frac{\log d}{\sqrt{d}}mL_\infty(W)\sqrt{\frac{4}{3m}\log\frac{1}{\delta}} \\ &+ \frac{\log d}{\sqrt{d}}mL_2(W) + L_\infty(W).\end{aligned} \tag{17}$$

Proof [Proof of Theorem 1] First, $|\mathrm{dsp}(W,S)| \leq 2$ for any set W and slice S, so by definition $|L_\infty(W)| \leq 2$. Next, By the assumption in the theorem, $L_2(W) = \tilde{O}(n^{-\xi}d^{-\eta})$. Plugging these into Lemma 5 completes the proof. ■

Lemma 5 is meaningful only when $cn\rho(W)$ is small compared to $nm\gamma_m/2$. This requires $L_2(W)$ to be sufficiently small. In the following we will first provide the proof sketch of Lemma 5, and then bound that $L_2(W)$ in Section 5.2.

5.1 Proof of Lemma 5

To prove Lemma 5, it is sufficient to bound the smallest eigenvalue of $G_n = D^\top D/n$. Note that $v_k \in \{-1, 1\}$, so $v_k^2 = 1$, and thus the (i,j)-th entry of G_n is

$$G_n(i,j) = \frac{1}{n}\sum_{k=1}^{n}\sigma'(w_k^\top x_i)\sigma'(w_k^\top x_j)\langle x_i, x_j\rangle. \tag{18}$$

For ReLU, $\sigma'(w^\top x)$ does not depend on the norm of w so without loss of generality, we assume $\|w\| = 1$. Consider a related matrix G whose (i,j)-th entry is defined as

$$G(i,j) = \mathbb{E}_w\left[\sigma'(w^\top x_i)\sigma'(w^\top x_j)\langle x_i, x_j\rangle\right]. \tag{19}$$

where w is a random variable distributed uniformly over the unit sphere. Since $\sigma'(w^\top x) = \mathbb{I}\left[w^\top x \geq 0\right]$, we have a closed form expression for $G(i,j)$:

$$\begin{aligned}G(i,j) &= \mathbb{E}_w\left[\mathbb{I}(w^\top x_i \geq 0)\mathbb{I}(w^\top x_j \geq 0)\right]\langle x_i, x_j\rangle \\ &= \left(\frac{1}{2} - \frac{\arccos\langle x_i, x_j\rangle}{2\pi}\right)\langle x_i, x_j\rangle.\end{aligned} \tag{20}$$

Note that $G(i,j) = g(x_i, x_j)$ where g is the kernel defined in (9). This allows us to reason about the eigenspectrum of G, denoted as $\lambda_1(G) \geq \ldots \geq \lambda_m(G)$.

Therefore, our strategy is to first bound $\lambda_m(G)$ in Lemma 6 and then bound $|\lambda_m(G) - \lambda_m(G_n)|$ in Lemma 7. Combining the two immediately leads to Lemma 5.

First, consider $\lambda_m(G)$. We consider a truncated version of spherical harmonic decomposition:

$$g^{[m]}(x_i, x_j) = \sum_{u=1}^{m}\gamma_u\phi_u(x_i)\phi_u(x_j)$$

and the corresponding matrix $G^{[m]}$. On one hand, it is clear that $\lambda_m(G) \geq \lambda_m(G^{[m]})$. On the other hand, $G^{[m]} = AA^\top$ where A is a random matrix whose rows are

$$A^i := [\sqrt{\gamma_1}\phi_1(x_i), \ldots, \sqrt{\gamma_m}\phi_m(x_i)].$$

Next, we bound $\lambda_m(G^{[m]})$ by matrix Chernoff bound [Tropp, 2012], and it is better than previous work [Braun, 2006]. This leads to the following lemma.

Lemma 6 *With probability at least* $1 - m\exp(-m\gamma_m/8)$,

$$\lambda_m(G) \geq m\gamma_m/2.$$

Next, bound $|\lambda_m(G) - \lambda_m(G_n)|$. By Weyl's theorem, this is bounded by $\|G - G_n\|$. To simplify the notation, denote

$$E_{i,j} = \mathbb{E}_w[\sigma'(w^\top x_i)\sigma'(w^\top x_j)] - \frac{1}{n}\sum_{k=1}^{n}\sigma'(w_k^\top x_i)\sigma'(w_k^\top x_j).$$

Then $G(i,j) - G_n(i,j) = \langle x_i, x_j\rangle E_{ij}$, and thus

$$\begin{aligned}&\|G - G_n\| \\ &= \sup_{\|\alpha\|=1}\sum_{i,j}\alpha_i\alpha_j\langle x_i, x_j\rangle E_{ij} \\ &\leq \sup_{\|\alpha\|=1}\sqrt{\sum_{i\neq j}\alpha_i^2\alpha_j^2}\sqrt{\sum_{i\neq j}\langle x_i, x_j\rangle^2 E_{ij}^2} + \max_i|E_{ii}| \\ &\leq \frac{c\log d}{\sqrt{d}}\sqrt{\sum_{i\neq j}E_{ij}^2} + \max_i|E_{ii}|\end{aligned} \tag{21}$$

where the last inequality holds with high probability since x_i's are uniform over the unit sphere and thus we can apply sub-gaussian concentration bounds.

Note that $\sum_{i\neq j}E_{ij}^2/(m(m-1))$ is a U-statistics where the summands are dependent and typical concentration inequality for *i.i.d.* entries does not apply. Instead we use a Bernstein inequality for U-statistics [Peel et al., 2010] to

show that with probability at least $1-\delta$, it is bounded by

$$\mathbb{E}_{\{x_1,x_2\}}E_{12}^2 + \sqrt{\frac{4\Sigma^2}{m}\log\frac{1}{\delta}} + \frac{4B^2}{3m}\log\frac{1}{\delta} \tag{22}$$

where $B = \max_i |E_{ii}|$ and $\Sigma^2 = \mathbb{E}\left[E_{12}^4\right] - \left(\mathbb{E}\left[E_{12}^2\right]\right)^2$.

The key observation is that the quantities in the above lemma are related to discrepancy:

$$\max_{i,j} E_{ij} \le L_\infty(W), \tag{23}$$

$$\mathbb{E}_{x_1,x_2}\left[E_{12}^2\right] = (L_2(W))^2, \tag{24}$$

$$\Sigma^2 \le (L_2(W)L_\infty(W))^2. \tag{25}$$

This is because $\sigma'(w^\top x_i)\sigma'(w^\top x_j) = \mathbb{I}[w \in S_{x_ix_j}]$ and thus

$$\begin{aligned} E_{i,j} &= \mathbb{E}_w\mathbb{I}[w \in S_{x_ix_j}] - \frac{1}{n}\sum_{k=1}^n \mathbb{I}[w_k \in S_{x_ix_j}] \\ &= A(S_{x_ix_j}) - \frac{1}{n}\left|W \cap S_{x_ix_j}\right| \\ &= -\mathrm{dsp}(W, S_{x_ix_j}). \end{aligned}$$

Plugging (23)-(25) into (22) and (21), we have

Lemma 7 *The following inequality holds with probability at least* $1 - 2m^2\exp\left(-\log^2 d\right) - \delta$,

$$\|G_n - G\| \le c\rho(W) \tag{26}$$

where $\rho(W)$ *is as defined in Lemma 5.*

5.2 Characterizing the discrepancy

In this subsection, we present a bound for $L_2(W)$ and show that the $\mathcal{G}_W$ defined in the following covers most W's:

$$\mathcal{G}_W = \left\{W : (L_2(W))^2 \le c_g\left(\sqrt{\frac{\log d}{nd}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{1}{\delta}\right)\right\} \tag{27}$$

for $0 < \delta < 1$ and a proper constant $c_g > 0$.[3]

First we provide a closed form for L_2 discrepancy of slices defined in (12). The proof is provided in the appendix.

Theorem 8 *Suppose* $W = \{w_i\}_{i=1}^n \subseteq \mathbb{S}^{d-1}$.

$$(L_2(W))^2 = \frac{1}{n^2}\sum_{i,j=1}^n k(w_i, w_j)^2 - \mathbb{E}_{u,v}\left[k(u,v)^2\right]$$

where $\mathbb{E}_{u,v}$ *is over* u *and* v *uniformly distributed on* $\mathbb{S}^{d-1}$ *and the kernel* $k(\cdot,\cdot)$ *is*

$$k(u,v) = \frac{\pi - \arccos\langle u, v\rangle}{2\pi}.$$

The closed form is simple and intuitive. The kernel $k(w_i, w_j)$ measures how similar two units are. The discrepancy is the difference between the average pairwise similarity and the expected one over uniform distribution.

[3]The constant c_g is the constant in Lemma 9. δ will be clear from the context where $\mathcal{G}_W$ is used.

Given the theorem, we now show that $(L_2(W))^2$ can be small. We use the probabilistic method, *i.e.*, show that if w_k's are sampled from $\mathbb{S}^{d-1}$ uniformly at random, then with high probability W falls into $\mathcal{G}_W$. The key observation is that with random W, $(L_2(W))^2$ is the difference between a U-statistics and its expectation, which can be bounded by concentration inequalities. Formally,

Lemma 9 *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,*

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

Alternatively, the theorem means that $\mathcal{G}_W$ defined in Eqn 27 covers most W's. This is because $L_2(W)$ is independent of the length of w_k's, it is sufficient to show that $L_2(W)$ is small for $W \in \mathcal{G}_W \cap \mathbb{S}^{d-1}$.

6 Final bound on generalization error

Here we prove Theorem 3 using Theorem 1. Suppose an algorithm such as gradient descent on the loss function $L = \frac{1}{2m}\sum_l (y_l - f(x_l))^2$ gives a solution satisfying the assumption and with small gradient $\|\partial L/\partial W\|$. Using Eqn (7), we have $\|r\| \le \|\partial L/\partial W\|/s_m(D)$. By Theorem 1 and the assumption in Theorem 3, with high probability $s_m^2(D) = \Omega(nm^{1-\beta})$. This implies the training loss

$$\frac{1}{m}\sum_l (f(x_l) - y_l)^2 = m\|r\|^2 \le \frac{cm^\beta}{n}\left\|\frac{\partial L}{\partial W}\right\|^2.$$

The generalization error can be derived using McDiamid's inequality and Rademacher complexity. First, we need an upper bound on the difference of the loss for two data points for the McDiamid's inequality. Since $\|x\|_2 \le 1$ and $\sum_k \|w_k\|_2 \le C_W$, we have $|f| \le C_W$. Thus

$$\begin{aligned} &|l(y, f(x)) - l(y', f(x'))| \\ \le\ &\frac{1}{2}\max\left\{(y - f(x))^2, (y' - f(x'))^2\right\} \le Y^2 + C_W^2 \end{aligned}$$

where in the last inequality we use the fact that the true function $|y| \le Y$. Next, we use the composition rules to compute the Rademacher complexity. Since the complexity of linear functions $\left\{w^\top x : \|w\|_2 \le b_W, \|x\|_2 \le 1\right\}$ is $b_W/\sqrt{m}$ and $\sigma(\cdot)$ is 1-Lipschitz, and $\sum_k \|w_k\|_2 \le C_W$, the complexity $\mathcal{R}_m(\mathcal{F}) \le C_W/\sqrt{m}$. Composing it with the loss function, and applying the bound in [Bartlett and Mendelson, 2002], we get the final generalization bound.

Corollary 4 then follows from Theorem 3 and Lemma 9. More details of the proof are in Appendix E.

7 Numerical evaluation

In this section, we further investigate numerically the effects of gradient descent on the discrepancy and the effects of regularizing the weights using discrepancy measure.

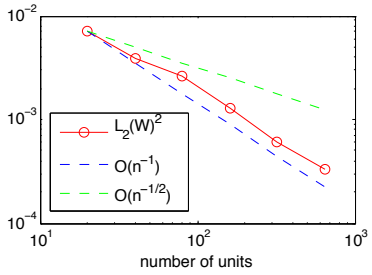

Figure 3: Discrepancy of W obtained after gradient descent. We perform gradient descent and compute discrepancy for the returned solution. The red curve corresponds to such solutions with different n. It scales similarly to the bound for uniform W as in Lemma 9.

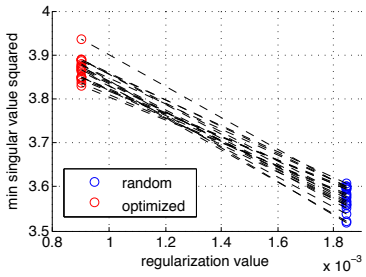

Figure 4: Effect of regularization. The blue dots represent random weights and the red dots linked with dashed black lines represent weights optimized by minimizing $R(w)$. Smaller regularization values correspond to larger minimal singular values.

7.1 Discrepancy and gradient descent

One limitation of the analysis is that we have not analyzed how to obtain a solution $W \in \mathcal{G}_W$ with small gradient. The theoretical analysis of gradient descent is left for future work. Meanwhile we provide some numerical results supporting our claims.

Although the set $\mathcal{G}_W$ contains most W's, it is still unclear whether the solutions given by gradient descent lie in the set. We design experiments to investigate this issue. The ground truth input data are of dimension $d = 50$ and true function consists of $n = 50$ units. We use networks of different n to learn the true function and perform SGD with batch size 100 and learning rate 0.1 for 5000 iterations. Figure 3 shows how $(L_2(W))^2$ changes as a function of n. It is slightly worse than $O(n^{-1})$ but scales better than $O(n^{-1/2})$, suggesting (stochastic) gradient descent outputs solutions with reasonable discrepancy.

7.2 Regularization

To reinforce solutions with small discrepancy, we propose a novel regularization term to minimize L_2 discrepancy:

$$R(W) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(w_i, w_j)^2. \tag{28}$$

It is essentially L_2 discrepancy without the constants.

To verify the effectiveness of the regularization term, we explore the relationship between the regularization and the minimum singular value. We first generate 20 random W's, all with $n = 100$ and $d = 100$, and compute their discrepancy and singular values using $m = 3000$. Then we optimize $R(W)$ and compare the quantities after optimization. The result is presented in Figure 4. We can see smaller regularization value corresponds to larger singular value.

We also conducts experiments to compare training and test errors with and without regularization. The ground truth data are of $d = 100$ and $n = 100$. We learn the true function by SGD with learning rate 0.1, momentum 0.9 and a total of 300,000 iterations. The regularization coefficients are chosen from $\{1, 0.1, 0.01, 0.001\}$ and the best results are reported. We use neural networks of size $n \in \{100, 150, 200, 300\}$ and for each n we repeat five times with different random seeds. The result is summarized in Table 1. Regularization leads to lower training and test errors for most settings. Even in the case where the un-regularized one performs better, the errors are all small enough (within the same range as standard deviation), suggesting the noise begins to dominate.

Table 1: Comparison of performance with/without regularization (all numbers are of unit 10^{-5}). The true function is generated with $d = 100$ and $n = 100$. To learn the function, we use networks with different n.

	$n = 100$		$n = 150$	
	train	test	train	test
no-reg	15.42(5.86)	14.80(5.36)	1.79(0.45)	1.86(0.50)
reg	**11.32(1.77)**	**10.63(1.58)**	**1.07(0.84)**	**1.13(0.99)**
	$n = 200$		$n = 300$	
	train	test	train	test
no-reg	**0.38(0.27)**	**0.44(0.35)**	0.39(0.39)	0.44(0.40)
reg	0.50(0.51)	0.58(0.59)	**0.10(0.05)**	**0.12(0.07)**

8 Conclusion

We have analyzed one-hidden-layer neural networks and identified novel conditions when local minima become global minima despite the non-convexity of the loss function. The key factors are the spectrum of the kernel associated with the activation function and the diversity of the units measured by discrepancy. Based on the insights, we have also proposed a novel regularization term that promotes unit diversity and achieves better generalization.

Acknowledgements. Y. L. was supported in part by NSF grants CCF-1527371, DMS-1317308, Simons Investigator Award, Simons Collaboration Grant, and ONRN00014-16-1-2329. L.S. was supported in part by NSF IIS1218749, NIH BIGDATA 1R01GM108341, NSF CAREER IIS-1350983, NSF IIS-1639792 EAGER, ONR N00014-15-1-2340, Nvidia and Intel.

References

[Barron, 1993] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945.

[Bartlett and Mendelson, 2002] Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

[Bilyk and Lacey, 2015] Bilyk, D. and Lacey, M. T. (2015). One bit sensing, discrepancy, and stolarsky principle. *arXiv preprint arXiv:1511.08452*.

[Blum and Rivest, 1993] Blum, A. L. and Rivest, R. L. (1993). Training a 3-node neural network is np-complete. In *Machine learning: From theory to applications*.

[Braun, 2006] Braun, M. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *JMLR*, 7:2303–2328.

[Cho and Saul, 2009] Cho, Y. and Saul, L. K. (2009). Kernel methods for deep learning. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 342–350.

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

[Dauphin et al., 2014] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*.

[Ge et al., 2015] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points – online stochastic gradient for tensor decomposition. *arXiv preprint arXiv:1503.02101*.

[Janzamin et al., 2015] Janzamin, M., Sedghi, H., and Anandkumar, A. (2015). Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *CoRR abs/1506.08473*.

[Kawaguchi, 2016] Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems (NIPS)*.

[Krotov and Hopfield, 2016] Krotov, D. and Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *CoRR*, abs/1606.01164.

[Lee et al., 2016] Lee, J., Simchowitz, M., Jordan, M., and Recht, B. (2016). Gradient descent only converges to minimizers. In *Proceedings of the Annual Conference on Learning Theory (COLT)*.

[Littwin and Wolf, 2016] Littwin, E. and Wolf, L. (2016). The multiverse loss for robust transfer learning. In *CVPR*.

[Mariet and Sra, 2015] Mariet, Z. and Sra, S. (2015). Diversity networks. *arXiv preprint arXiv:1511.05077*.

[Müller, 2012] Müller, C. (2012). *Analysis of spherical symmetries in Euclidean spaces*, volume 129. Springer Science & Business Media.

[Peel et al., 2010] Peel, T., Anthoine, S., and Ralaivola, L. (2010). Empirical bernstein inequalities for u-statistics. In *Advances in Neural Information Processing Systems*, pages 1903–1911.

[Rahimi and Recht, 2009] Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Neural Information Processing Systems*.

[Shamir, 2016] Shamir, O. (2016). Distribution-specific hardness of learning neural networks. *arXiv preprint arXiv:1609.01037*.

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

[Tropp, 2012] Tropp, J. A. (2012). User-friendly tools for random matrices: An introduction. Technical report, DTIC Document.

[Vershynin, 2010] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

[Xie et al., 2015] Xie, P., Deng, Y., and Xing, E. (2015). On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110*.

[Yang et al., 2014] Yang, Z., Moczulski, M., Denil, M., de Freitas, N., Smola, A. J., Song, L., and Wang, Z. (2014). Deep fried convnets. *CoRR*, abs/1412.7149.