이학석사 학위논문

Global Optimality in Deep Neural Networks with Regularization

(정칙화를 포함한 깊은 신경망에서의 전체적 최적성)

2019년 2월

서울대학교 대학원

수 리 과 학 부

고 영 진

Global Optimality in Deep Neural Networks with Regularization

(정칙화를 포함한 깊은 신경망에서의 전체적 최적성)

지도교수 강 명 주

이 논문을 이학석사 학위논문으로 제출함

2018년 10월

서울대학교 대학원

수 리 과 학 부

고 영 진

고 영 진의 이학석사 학위논문을 인준함

2018년 12월

위 원 장 _____ (인)

부 위 원 장 _____ (인)

위 원 _____ (인)

Global Optimality in Deep Neural Networks with Regularization

by

Youngjin Koh

A DISSERTATION

Submitted to the faculty of the Graduate School
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematical Sciences
Seoul National University
February 2019

Abstract

Global Optimality in Deep Neural Networks with Regularization

Youngjin Koh

Department of Mathematical Sciences
The Graduate School
Seoul National University

In recent years, Deep Neural Networks(DNNs) have shown a dramatic success in many domains. However, the theoretical reasons for explaining the performance remain elusive. One of the most important key issue is the error optimization problem. In general, minimizing the loss function of the DNNs is a non-convex problem, hence the optimization algorithms may fail to find the global minimum. In this paper, we introduce several conditions for a local minimum to be globally optimal. In particular, we provide the conditions in DNNs with regularization and suggest the efficient network structure and regularization function. We also apply the theoretical results to the practical DNNs.

Key words: Optimization, Global optimality, Regularization, Deep Learning
Student Number: 2016-20227

Contents

Chapter 1

Introduction

In recent years, Deep Neural Networks(DNNs) have shown a great practical performance in many application areas of machine learning such as computer vision, signal processing, pattern recognition and many other fields. However, despite its huge practical success, theoretical reasons of why neural networks perform well remains elusive.

One of the important theoretical challenge is the error optimization problem. In general, error function is not convex with respect to learning parameters in neural network, hence optimization algorithm could get stuck in a poor local minimum. Recently, several theoretical results for above challenge suggested. Bengio et al. [1] showed that the number of neurons in the hidden layer is not fixed, then the process of training a globally optimal neural network is analogous to selecting a finite number of hidden units. Dauphin et al. [4] applied random matrix theory to high-dimensional non-convex optimization. Using arguments from that, Choromanska [3] showed that, all local minima become increasingly close to being global minima under several assumptions. Janzamin et al. [6] proposed that with sufficient assumptions, polynomial-time training is possible. Further, Safran and Shamir [12] suggested the conditions that ensure a random initialization could be within the basin of a global minimizer.

In this thesis, we introduce several global optimality conditions of error optimization problem in neural networks, and provide the experimental results. First, we discuss about the case of deep linear neural network. By adapting

the results of Lu and Kawaguchi [11]'s and Yun et al. [13]'s, we introduce the assumption to avoid poor local minima, and provide necessary and sufficient conditions to determine a critical point to be global minimum in linear neural network. In addition, we expend the same discussion above to the case of deep nonlinear neural network with ReLU. We introduce the open problem of Choromanska [3]'s, and Kawaguchi [7]'s result. So we successfully reduce the error optimization problem in nonlinear network to that of linear model. Moreover, we discuss about the optimization problem with regularization. In this part, from the results of Haeffele and Vidal [5]'s, we suggest efficient network structure and regularization. We also provide practical results.

The content of this paper is organized as follows. We discuss on the error optimization problem in deep neural network in **Chapter 2**. We introduce the efficient network architecture with regularization in **Chapter 3**. The experimental results are provided in **Chapter 4**. Conclusions in **Chapter 5**.

Chapter 2

Global Optimality in Deep Neural Networks

2.1 Deep linear neural network

In this section, we will analyze the error optimization problem in deep linear neural network. First, we describe the notations for deep neural networks. Suppose we have m input-output training data pairs, where the dimensions of input data and output data are $d_x \geq 1$, $d_y \geq 1$, respectively. Let (X, Y) be the training data set with input $X \in \mathbb{R}^{d_x \times m}$ and output $Y \in \mathbb{R}^{d_y \times m}$. Suppose H be the number of hidden layers, where each hidden layer have width $d_1, \ldots d_H$. We denote the weight parameter matrices by W, where $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ with $1 \leq i \leq H+1$. For simplicity, let $d_0 = d_x$ and $d_{H+1} = d_y$.

Let $\Phi(W_1, \ldots, W_{H+1}) \in \mathbb{R}^{d_y \times m}$ be the output of deep neural network model. Φ can be arbitrary mapping. In this section, We assume that Φ is the output of feedforward deep linear neural network, as

$$\Phi(W) = W_{H+1} W_H \cdots W_2 W_1 X, \tag{2.1}$$

where $W{=}(W_1, \ldots, W_{H+1})$. We consider the optimization problem of summation of squared error loss over all data points,

$$\min_W L(W) = \frac{1}{2} \|\Phi(W_1, \ldots, W_{H+1}) - Y\|_F^2, \tag{2.2}$$

where $\|\cdot\|_F$ is the Frobenius norm.

We will now introduce several Theorems for deep linear neural network. To analyze (2.2), Lu and Kawaguchi [11] suggested the following optimization problem, which is equivalent to (2.2) in terms of the global minimum value.

$$\min_R F(R) = \|RX - Y\|_F^2 \quad s.t. \quad rank(R) \leq d_p, \tag{2.3}$$

where $R \in \mathbb{R}^{d_{H+1} \times d_0}$ and $p = \text{argmin}_{0 \leq i \leq H+1} d_i$. Unless $d_p = \min(d_{H+1}, d_0)$, (2.3) is non-convex. We can deduce the optimization problem (2.2) from (2.3) by following Theorem.

Theorem 2.1 (Lu and Kawaguchi [11]). *Suppose that X and Y have full rank. If $\bar{W} = (\bar{W}_1, \ldots, \bar{W}_{H+1})$ is a local minimum point of (2.2), then $\bar{R} = \bar{W}_{H+1} \cdots \bar{W}_1$ is a local minimum point of (2.3).*

Theorem 2.1 concludes that every local minimum in (2.2) corresponds to those in (2.3). Therefore, we can consider the optimization problem (2.3) only. The following Theorem shows that there is no poor local minimum in (2.3).

Theorem 2.2 (Lu and Kawaguchi [11]). *If X has full rank, then all local minima of (2.3) are global minima.*

By the results of Theorem 2.1 and Theorem 2.2, the following Theorem holds. That is, there is no poor local minima in deep linear neural network with a square error loss.

Theorem 2.3 (Lu and Kawaguchi [11]). *If X and Y have full rank, then all local minima of (2.2) are global minima.*

We only consider the Frobenius loss function in (2.2) here, however, this theorem holds for more general loss function which satisfies several conditions. Previously, Kawaguchi [7] proposed more strong properties for a deep linear neural network under some strong assumptions. Theorem 2.3 generalizes one of this property with fewer assumptions.

Theorem 2.3 states that every critical point is either a global minimum or a saddle point. Therefore, we cannot determine which critical point is global minimum. Yun et al. [13] proposed the conditions to distinguish between the

two with more strong assumptions. Here, following Theorems partition the domain of $L(W)$ into two sets which one set with only global minima of (2.2), and the other with saddle points.

Let $k = \min_{0 \le i \le H+1} d_i$ and $YX^T(XX^T)^{-1}X = U\Sigma V^T$ be the singular value decomposition of $YX^T(XX^T)^{-1}X \in \mathbb{R}^{d_y \times d_x}$.

Theorem 2.4 (Yun et al. [13]). *Suppose that $d_x \le m$, $d_y \le m$, and XX^T and YX^T have full ranks. Also, suppose the singular values of $YX^T(XX^T)^{-1}X$ are all distinct. If $k = \min\{d_x, d_y\}$, let*

$$\mathcal{V}_1 := \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_{H+1} \cdots W_1) = k\}. \tag{2.4}$$

Then, every critical point of L(W) in $\mathcal{V}_1$ is a global minimum. Moreover, every critical point of L(W) in $\mathcal{V}_1^c$ is a saddle point.

Theorem 2.5 (Yun et al. [13]). *Suppose that $d_x \le m$, $d_y \le m$, and XX^T and YX^T have full ranks. Also, suppose the singular values of $YX^T(XX^T)^{-1}X$ are all distinct. If $k < \min\{d_x, d_y\}$, let*

$$\mathcal{V}_2 := \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_{H+1} \cdots W_1) = k, \operatorname{col}(W_{H+1} \cdots W_{p+1}) = \operatorname{col}(\widehat{U})\}, \tag{2.5}$$

where $\widehat{U} \in \mathbb{R}^{d_y \times k}$ be a matrix consisting of the first k columns of U. Then, every critical point of L(W) in $\mathcal{V}_2$ is a global minimum. Moreover, every critical point of L(W) in $\mathcal{V}_2^c$ is a saddle point.

Note that Theorem 2.4 and Theorem 2.5 provide necessary and sufficient conditions to determine a critical point of $L(W)$ to be a global minimum. Therefore, we can easily determine if the critical point of error function is global optimum or not.

In this section, we have analyzed about the optimization problem (2.2) in deep linear neural network. Theorem 2.3 guarantee that there is no poor local minimum in deep linear neural network, and Theorem 2.4 and Theorem 2.5 gives the conditions to distinguish between the global minima and saddle points from the critical values.

2.2 Deep nonlinear neural network

We have obtained several properties for optimization problem of error function in deep linear network. Now we will extend the same discussion to deep nonlinear neural network. We use the same notation as for the deep linear neural network models. Typically, each layer of nonlinear neural network applies some form of linearity, followed by a nonlinear activation functions (e.g., max-pooling, sigmoid). We consider one of the most used activation function, rectified linear unit (ReLU).

Here, we assume that $\Phi(W_1, \ldots, W_{H+1}) \in \mathbb{R}^{d_y \times m}$ is the output of deep nonlinear neural network, as

$$\Phi(W) = q\sigma_{H+1}(W_{H+1}\sigma_H(W_H \cdots \sigma_2(W_2\sigma_1(W_1X)) \cdots))), \tag{2.6}$$

where $q \in \mathbb{R}$ is simply a normalization factor. $\sigma_i : \mathbb{R}^{d_i \times m} \to \mathbb{R}^{d_i \times m}$ is the element-wise rectified linear function:

$$\sigma_i \left(\begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{d_i 1} & \cdots & b_{d_i m} \end{bmatrix} \right) = \begin{bmatrix} \bar{\sigma}(b_{11}) & \cdots & \bar{\sigma}(b_{1m}) \\ \vdots & \ddots & \vdots \\ \bar{\sigma}(b_{d_i 1}) & \cdots & \bar{\sigma}(b_{d_i m}) \end{bmatrix}, \tag{2.7}$$

where $\bar{\sigma}(b_{ij}) = \max\{0, b_{ij}\}$, the rectified linear unit (ReLU). As the model $\Phi(W_1, \ldots, W_{H+1})$ can be represented a directed acyclic graph, Choromanska et al. [2] suggested the expression of (2.6) as

$$\Phi(W) = q \sum_{i=1}^{\Psi} X_i Z_i \prod_{k=1}^{H+1} w_i^{(k)}, \tag{2.8}$$

where Ψ is the total number of paths from the inputs to outputs in the directed acyclic graph. Moreover, $X_i \in \mathbb{R}$ denotes the entry of the input which is used in i-th path, and $w_i^{(k)} \in \mathbb{R}$ is the entry of W_k which is used in i-th path. Here, we have considered the rectified linear function, $Z_i \in \{0, 1\}$ represents whether the i-th path is active ($Z_i = 1$) or not ($Z_i = 0$).

Dauphin et al. [4] explained the connection between the loss function of neural networks and the theory of random Gaussian fields by providing experiments, and Choromanska et al. [3] discussed the theoretical results for the

existence of this connection. However, their results relied on several unrealistic assumptions, which were labeled A1p, A2p, A3p, A4p, A5p, A5u, A6u, and A7p. Choromanska et al. [3] suggested that implying their results with milder assumptions is an important open problem.

In Kawaguchi [7], he introduced the results that successfully discarded most of those assumptions. In his paper, he discarded A2p, A3p, A4p, A6u, and A7p, and used only A1p-m and A5u-m, which are weaker versions of assumptions A1p and A5u, respectively. Assumption A1p-m assumes that Z_i's in (2.8) are Bernoulli random variables with the same probability, that is, $\Pr(Z_i = 1) = \rho$ for all i. Assumption A5u-m is that the Z_i's are independent from the input X's and parameters w's. Therefore, under the assumptions A1p-m and A5u-m, we can notate $\mathbb{E}_Z[\Phi(W)] = q \sum_{i=1}^{\Psi} X_i \rho \prod_{k=1}^{H+1} w_i^{(k)}$.

For the squared error loss with expectation $L(W) = \frac{1}{2}\|\mathbb{E}_Z[\Phi(W) - Y]\|_F^2$, the following theorem holds:

Theorem 2.6 (Kawaguchi [7])**.** *Assume A1p-m and A5u-m. Let $q = \rho^{-1}$. Then the loss function of the deep nonlinear network L(W) can be reduced to that of the deep linear model $\bar{L}(W)$. Therefore, under the same conditions of Theorem 2.3, then all local minima are global minima.*

Theorem 2.6 provides the same results as theorem 2.3, which is in case of deep linear neural network, with milder assumptions compare to results of Choromanska [3]'s.

In this section, we have analyzed about the global optimality conditions in optimization problem of error function in deep neural network without any regularization terms. In next chapter, we will discuss on the optimization of error function, which contains the regularization terms.

Chapter 3

Deep Neural Networks with Regularization

3.1 Neural network with one hidden layer

We have discussed on the error optimization problem in deep neural network without regularization. However, to prevent overfitting, we typically designed neural network model with regularization term. Unfortunately, This regularization term makes the error optimization problem more complex. In this section, we will discuss about the global optimality conditions in the error optimization problem in deep linear neural network with regularization term.

We use the same notation as in Chapter 2. The most important difference between two chapters is the regularization term. Here, we consider the optimization problem :

$$\min_{W} L(W) + \lambda\Theta(W), \tag{3.1}$$

where $L(W)$ is a loss function and Θ is a regularization function designed to prevent overfitting. Note that we require both the loss function and regularization function to be convex on input data X.

First, we will discuss about the neural network with just one hidden layer. Let r be the dimension of hidden layer. We now denote as following: $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_1 \times m}$, $W_1 \in \mathbb{R}^{r \times d_0}$, $W_2 \in \mathbb{R}^{r \times d_1}$.

Under these conditions, Haeffele and Vidal [5] proposed the conditions

that every local minima in optimization problem is globally optimal. Before introduce the result, we need to define the *positive homogeneity*.

Definition 3.1. *Suppose $f : V \to W$ is a function between two vector spaces V and W over a field F, and k is an integer. If $f(\alpha v) = \alpha^k f(v)$ for all $\alpha > 0$, $\alpha \in F$ and $v \in V$, then f is said to be* ***positively homogeneous*** *of degree k.*

Theorem 3.2. *Suppose Φ and Θ are sums of positively homogeneous functions of the same degree. If one of the columns W_1 and W_2 is equals to zero, all local minima of* (3.1) *are global minima.*

Note that the ReLU, max(average)-pooling, and convolution are positively homogeneous. Therefore, we can easily design the network to be positively homogeneous. The important part is that Φ and Θ have same degree of homogeneity. Typically, we construct a neural network with one hidden layer as following:

$$\begin{aligned} \Phi(W_1, W_2) &= \sum_{i=1}^{r} \phi(W_1^i, W_2^i), \\ \Theta(W_1, W_2) &= \sum_{i=1}^{r} \theta(W_1^i, W_2^i), \end{aligned} \tag{3.2}$$

where W_1^i, W_2^i are the i-th columns of W_1 and W_2, respectively. Clearly, $\phi(w_1, w_2) = w_2^\top w_1 X$ and $\theta(w_1, w_2) = \|w_1\|^2 + \|w_2\|^2$ here. These ϕ and θ are positively homogeneous of degree 2, therefore, this network satisfies the conditions of Theorem 3.2.

Note that $\phi(w_1, w_2) = w_2^\top \bar{\sigma}(w_1 X)$ satisfies the conditions also, where $\bar{\sigma}$ is the rectified linear unit (ReLU). ReLU, max-pooling, and linear transformations are positively homogeneous of degree one, so these do not influence on the conditions of Theorem 3.2. However, sigmoid is not positively homogeneous, which could possibly explain the improved performance of ReLU compare to sigmoid.

3.2 Neural network with parallel structure

In this section, we extend our analyzation to network with r parallel sub-networks, which each sub-network have the same architecture. We use same notation as above. Let $(X,\ Y)$ be the training data set with input $X \in \mathbb{R}^{d_x \times m}$ and output $Y \in \mathbb{R}^{d_y \times m}$. Suppose r be the number of sub-networks and H be the number of hidden layers in each sub-network, where each hidden layer have width $d_1, \dots\ d_H$. We denote the weight parameter matrices in r-th sub-network by W^r, where $W_i^r \in \mathbb{R}^{d_i \times d_{i-1}}$ with $1 \leq i \leq H+1$. The maps Φ_r and Θ_r are defined as following:

$$\begin{aligned} \Phi_r(W_1, \dots, W_{H+1}) &= \sum_{i=1}^{r} \phi(W_1^i, \dots, W_{H+1}^i), \\ \Theta_r(W_1, \dots, W_{H+1}) &= \sum_{i=1}^{r} \theta(W_1^i, \dots, W_{H+1}^i). \end{aligned} \tag{3.3}$$

We assume ϕ and θ are positively homogeneous of same degree. To address the issue of non-convexity of (3.1), Haeffele and Vidal [5] suggested to define the *factorization regularization function* $\Omega_{\phi,\theta}(Z)$.

Definition 3.3 (Haeffele and Vidal [5])**.** *The* ***factorization regularization function*** $\Omega_{\phi,\theta}(Z)$ *is*

$$\begin{aligned} \Omega_{\phi,\theta}(Z) := \inf_{r \in \mathbb{N}^+} \inf_{W_r} & \sum_{i=1}^{r} \theta(W_1^i, \dots, W_{H+1}^i) \\ s.t. \quad & \Phi_r(W_1, \dots, W_{H+1}) = Z, \end{aligned} \tag{3.4}$$

with the additional condition that $\Omega_{\phi,\theta}(Z) = \infty$ *if* $Z \notin \bigcup_r \mathrm{Im}(\Phi_r)$

Theorem 3.4 (Haeffele and Vidal [5])**.** *The factorization regularization function* $\Omega_{\phi,\theta}(Z)$ *has the following properties*

1. $\Omega_{\phi,\theta}(Z)$ *is positively homogeneous of degree 1.*
2. $\Omega_{\phi,\theta}(Z)$ *is convex with respect to* Z*.*

Theorem 3.4 states that $\Omega_{\phi,\theta}(Z)$ is convex w.r.t. Z, thus, the infimum (3.3) exists when the number of sub-networks r is finite. We will now consider the convex problem, as

$$\min_{Z} F(Z) := L(Z) + \lambda \Omega_{\phi,\theta}(Z). \tag{3.5}$$

Here Z is the output of network, that is $Z = \Phi_r(W_1, \ldots, W_{H+1})$ and $\lambda > 0$. Haeffele and Vidal [5] provided the results to analyze the optimization problem.

$$\min_{W_r} f_r(W_1, \ldots, W_{H+1}) := L(W_1, \ldots, W_{H+1}) + \lambda \sum_{i=1}^{r} \theta(W_1^i, \ldots, W_{H+1}^i) \tag{3.6}$$

Theorem 3.5 (Haeffele and Vidal [5])**.** *Any local minimizer of* (3.6) *such that* $(W_1^{i_0}, \ldots, W_{H+1}^{i_0}) = (0, \ldots, 0)$ *for some* $i_0 \in \{1, \ldots, r\}$ *is a global minimizer of* (3.6)*. Moreover,* $Z = \Phi_r(W_1, \ldots, W_{H+1})$ *is a global minimizer of* (3.5)*.*

Proof. Since (3.5) is convex in Z by Theorem 3.4, the global optimality can be shown. By definition of $\Omega_{\phi,\theta}(Z)$, (3.5) lower bounds (3.6) for any $Z = \Phi_r(W_1, \ldots, W_{H+1})$. Therefore, this implies that every local minimum of (3.6) is a global minimum. □

This gives the result of Theorem 3.2 as a Corollary. In practice, when the size of neural network is large enough and ReLU or max-pooling is used, we can observe that many weight of network becomes zero. Theorem 3.5 possibly explains that phenomenon regarded as *dead neurons*. This Theorem also provides the following Corollary.

Corollary 3.6 (Haeffele and Vidal [5])**.** *If* $r > card(Z)$*, then from any initialization of network* $(V_1, \ldots, V_{H+1})$ *such that* $f_r(V_1, \ldots, V_{H+1}) < \infty$*, there exists a non-increasing path to a global minimizer of* $f_r(W_1, \ldots, W_{H+1})$*.*

Proof.
1. From given initialization, perform local descent to find a local minimum.
2. If that local minimum satisfies the conditions of Theorem 3.5, i.e., $(W_1^{i_0}, \ldots, W_{H+1}^{i_0}) = (0, \ldots, 0)$ for some $i_0 \in \{1, \ldots, r\}$, then by Theorem 3.5, That local minimum is a global minimum.

3. Else, there exists nonzero $\beta \in \mathbb{R}^r$ such that $\sum_{i=1}^{r} \beta_i \phi(W_1^i, \ldots, W_{H+1}^i) = 0$ by the assumption $r > card(Z)$. Let scale β so that $\min_i \beta_i = -1$ and set $W_k^i \leftarrow (1 + \beta_i)^{1/p} W_k^i$. Repeating this operation guarantees arriving at a point where one of the sub-networks is all 0.

□

Theorem 3.6 guarantees that by local descent, we can always find a global minimizer of $f_r(W_1, \ldots, W_{H+1})$ from any initialization $(V_1, \ldots, V_{H+1})$ of parallel network when the size of network is large enough. However, since several assumptions which are unrealistic to practical use, there is a limitations to use the result to design the network. In the next Chapter, we will design several neural network applying our results and observe the performances.

Chapter 4

Experiments

In this Chapter, we apply our theoretical discussions to practical deep neural network. We use the CIFAR10 [9] dataset. The MNIST [10] is also a very popular dataset, however we do not use the MNIST. The reason is that the performance of neural network on the MNIST dataset is originally very high, we could not compare to the performances well.

The CIFAR10 dataset is a database which consists of 50,000 examples for training, and 10,000 examples for test. Each example is a 32×32 color image in 10 classes. The classes contains airplane, automobile, bird, cat, deer, etc. This dataset is so popular and basic for the image classification in neural network, hence we could provide a practical result easily, and compare to the other network structures, or other works. We applied experiments by Tensorflow in Python 2 environment. We operates GeForce GTX 1080 to implement.

First, we compare to the performances in simple fully connected neural network. We observe the classification accuracy in neural network with various degree of homogeneity for network and regularization. In addition, we provide the performances in neural network with parallel sub-networks, which each sub-network have the same architecture.

4.1 Model of Neural Network

Here, we want to observe that the effect of degree of homogeneity for network and regularization. therefore, we construct four simple fully connected

# hidden layer	regularization			
	None	l_1	l_2	l_3
0	23.48	**24.87**	24.83	23.63
1	32.47	32.25	**33.01**	32.31
2	32.90	33.52	**34.54**	34.41
3	30.49	30.81	31.64	**31.87**

Table 4.1: Accuracy(%) on the CIFAR10 in simple fully connected networks. **Columns** : Applied regularization function. **Rows** : The number of hidden layers, that is, degree of homogeneity.

networks, which each have 0, 1, 2, and 3 hidden layers with 500 hidden neurons respectively. We do not contain any regularization function first, and we apply l_1, l_2, l_3 regularizations in regular sequence. We do not use dropout or other regularization, and any specific initialization also. For parallel networks, we design the network with parallel sub-network, which each sub-network is simple fully connected network with one hidden layer with 500 hidden neurons. The loss function is the cross-entropy and we use the Gradient-Descent. Every activation function in hidden layers is ReLU.

4.2 Results

Simple fully connected network

Table 4.1 gives the result. We do not use the Convolutional Neural Network (CNN) and just design the simple fully connected network, hence the performance of classification is not that high. However, it is enough to observe that the effect of degree of homogeneity for network and regularization. The network contains 2 hidden layers with 1_2 regularization provides the highest performance, 34.54%.

The number of hidden layers means the degree of homogeneity of network architecture. Generally, when the network architecture and regularization function have same degree of homogeneity, the highest performance is

# sub-networks	regularization			
	None	l_1	l_2	l_3
4	38.92	40.80	**42.82**	40.50
8	40.11	42.50	**43.17**	42.10
12	42.24	43.45	**44.79**	42.32

Table 4.2: Accuracy(%) on the CIFAR10 in networks with parallel sub-networks. **Columns** : Applied regularization function. **Rows** : The number of sub-networks.

observed. Since the CIFAR10 dataset classification is easy to overfit compare to the MNIST, the network without any regularization performs works. When the network contains 2 hidden layers, we expect that l_3 regularization gives the highest performance, however, l_2 gives the best. But it is negligible, since the difference of accuracy between l_2 and l_3 regularization.

Note that every network except the one with 1 hidden layer in this experiment actually do not satisfy the conditions of Theorem 3.2. This provides that even if the number of hidden layer is not one, it's important that the degree of degree of homogeneity for network and regularization to be same. Therefore, we can apply our theoretical results in deeper neural networks.

Network with parallel sub-networks

Table 4.2 gives the result. The network contains 12 sub-networks with 1_2 regularization provides the highest performance, 44.79%. In this case, every network have the same degree of homogeneity: 2. Therefore, we can easily predict that l_2 regularization performs best. In practice, for every network, the performances of l_2 regularization is the highest. Again, the results support our theoretical results efficiently. It's well known that the ensemble of network usually performs better compare to one. Haeffele and Vidal [5] suggested that possible explanation which the reason is that to satisfy the condition $r > card(Z)$ in Theorem 3.6, the number of sub-network is require to be large.

Note that again, every network except the one with 12 sub-networks in

this experiment actually do not satisfy the conditions of Theorem 3.6, which requires $r > card(Z)$. This provides that the condition about the degree of degree of homogeneity for network and regularization is stronger than other assumptions. Thus, we could apply our theoretical results in other structures of network.

We have provided the practical results which are applied several various conditions. Most of the results support our theoretical discussions well. However, since several conditions of our theoretical results are too tight for practical use, we do not provide the enough experimental results. In the next Chapter, we conclude our results and discuss about the limitation.

Chapter 5

Conclusion

In this thesis, we have discussed about the global optimality conditions of error optimization problem in deep neural networks. In Chapter 2, we proposed the necessary and sufficient conditions to determine a critical point of error function to be global minimum in deep linear neural network. Moreover, In deep nonlinear neural network with ReLU activation, we successfully reduce the error function to that of the deep linear model. In particular, we provided the sufficient conditions to guarantee that every local minimum to be a global minimum of error function, and suggested the network architecture to address the issue of non-optimal local minimum in deep neural network with regularization.

Furthermore, we provided the experimental results to support our discussions. We compared the deep neural network which satisfies the conditions of our theories, and which doesn't satisfy. In general, networks which have same degrees of positive homogeneity between the network mapping and the regularization function performs better. Therefore, the discussions in this paper and experimental results gives fine fine guidelines to design a network architectures and regularization.

Despite the above facts, we have several limitations of our discussions and results. First, we have assumed some unrealistic conditions to conclude the results. In Chapter 2, although Kawaguchi [7] successfully discarded most of those assumptions, there still remaining some of those.

Additionally, the network architecture is so limited for practical use. In

Chapter 3, we only discussed about networks which have only one hidden layer, or have parallel structure which each sub-network have the same architectures. We also assumed that $r > card(Z)$ in Corollary 3.6, which makes the size of network too large. It is also an important problem which is the fact that global optimality does not always give the improved performance.

Despite the limitations above, It's important to suggest the mathematical reasons for the performance of neural network. Our results possibly explain some phenomena, like dead neurons, improved performance of ReLU, and poor performance of classical regularizations, such as an l_1 or l_2 norms. Recently, Kawaguchi and Bengio [8] tried to suggest the similar discussions in ResNet which is not limited to neural network architecture. This theoretical attempts could provide more explanations about neural network, and improve the performances.

Bibliography

[1] Y. Bengio, N. Le Roux, P. Vincent, O. Delalleau, and P. Marcotte, *Convex neural networks*, In Neural Information Processing Systems, (2005), pp. 123–130.

[2] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun, *The loss surfaces of multilayer networks*, In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, (2015), pp. 192–214.

[3] A. Choromanska, Y. LeCun, and G. Ben Arous, *Open problem: The landscape of the loss surfaces of multilayer networks*, In Proceedings of the 28th Conference on Learning Theory, (2015), pp. 1756–1760.

[4] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, In Advances in Neural Information Processing Systems, (2014), pp. 2933–2941.

[5] B. D. Haeffele and R. Vidal, *Global optimality in neural network training*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), pp. 7331–7339.

[6] M. Janzamin, H. Sedghi, and A. Anandkumar, *Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods*, arXiv preprint arXiv:1506.08473, (2015).

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

[8] K. Kawaguchi and Y. Bengio, *Depth with nonlinearity creates no bad local minima in resnets*, arXiv preprint arXiv:1810.09038, (2018).

[9] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*, (2009).

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[11] H. Lu and K. Kawaguchi, *Depth creates no bad local minima*, arXiv preprint arXiv:1702.08580, (2017).

[12] I. Safran and O. Shamir, *On the quality of the initial basin in over-specified neural networks*, In International Conference on Machine Learning, (2016), pp. 774–782.

[13] C. Yun, S. Sra, and A. Jadbabaie, *Global optimality conditions for deep neural networks*, International Conference on Learning Representations (ICLR), (2018).

Appendix A

Proof of Main Theorem

A.1 Proof of Theorem 3.5

Lemma A.1. *If $(W_1, \ldots, W_{H+1})$ is a local minimum of $f_r(W_1, \ldots, W_{H+1})$, then for any $\beta \in \mathbb{R}^r$,*

$$\langle -\frac{1}{\lambda}\nabla_Z L(Z), \sum_{i=1}^{r} \beta_i \phi(W_1^i, \ldots, W_{H+1}^i)\rangle = \sum_{i=1}^{r} \beta_i \theta(W_1^i, \ldots, W_{H+1}^i). \quad \text{(A.1)}$$

Let

$$\Omega^{\circ}_{\phi,\theta}(U) := \sup_{(w_1,\ldots,w_{H+1})} \langle U, \phi(w_1, \ldots, w_{H+1})\rangle \quad s.t. \quad \theta(w_1, \ldots, w_{H+1}) \geq 1. \quad \text{(A.2)}$$

Note that

$$F(Z) = L(Z) + \lambda\Omega_{\phi,\theta}(Z) \leq f_r(W_1, \ldots, W_{H+1}) \quad \text{(A.3)}$$

by the definition of $\Omega_{\phi,\theta}(Z)$. Since (3.5) is convex, Z is a global minimum of $F(Z)$ if and only if

$$-\frac{1}{\lambda}\nabla_Z L(Z) \in \partial\Omega_{\phi,\theta}(Z). \quad \text{(A.4)}$$

Suppose $(W_1, \ldots, W_{H+1})$ is a local minimum of $f_r(W_1, \ldots, W_{H+1})$. Then for $\forall(U_1, \ldots, U_{H+1})_r$, $\exists\delta > 0$ such that for $\forall\epsilon \in (0, \delta)$, $f_r(W_1 + \epsilon^{1/p}U_1, \ldots, W_{H+1} +$

$\epsilon^{1/p}U_{H+1}) \geq f_r(W_1, \ldots, W_{H+1})$. Let

$$(U_1^j, \ldots, U_{H+1}^j) = \begin{cases} (0, \ldots, 0) & j \neq i_0 \\ (u_1, \ldots, u_{H+1}) & j = i_0 \end{cases}, \tag{A.5}$$

where $(W_1^{i_0}, \ldots, W_{H+1}^{i_0}) = (0, \ldots, 0)$. Then

$$\begin{aligned}
&L(Z) + \lambda \sum_{i=1}^{r} \theta(W_1^i, \ldots, W_{H+1}^i) \leq \\
&L(\Phi_r(W_1 + \epsilon^{1/p}U_1, \ldots, W_{H+1} + \epsilon^{1/p}U_{H+1})) + \\
&\quad \lambda \sum_{i=1}^{r} \theta(W_1^i + \epsilon^{1/p}U_1^i, \ldots, W_{H+1}^i + \epsilon^{1/p}U_{H+1}^i)) = \\
&L(\sum_{i \neq i_0} \phi(W_1^i, \ldots, W_{H+1}^i) + \phi(W_1^{i_0} + \epsilon^{1/p}U_1^{i_0}, \ldots, W_{H+1}^{i_0} + \epsilon^{1/p}U_{H+1}^{i_0})) + \\
&\quad \lambda \sum_{i \neq i_0} \theta(W_1^i, \ldots, W_{H+1}^i) + \lambda\theta(W_1^{i_0} + \epsilon^{1/p}U_1^{i_0}, \ldots, W_{H+1}^{i_0} + \epsilon^{1/p}U_{H+1}^{i_0}) = \\
&L(Z + \epsilon\phi(u_1, \ldots, u_{H+1})) + \lambda \sum_{i=1}^{r} \theta(W_1^i, \ldots, W_{H+1}^i) + \epsilon\lambda\theta(u_1, \ldots, u_{H+1}).
\end{aligned}$$

Therefore,

$$\epsilon^{-1}[L(Z + \epsilon\phi(u_1, \ldots, u_{H+1})) - L(Z)] \geq -\lambda\theta(u_1, \ldots, u_{H+1}). \tag{A.6}$$

Taking the limit as $\epsilon \searrow 0$,

$$\langle \phi(u_1, \ldots, u_{H+1}), \nabla_Z L(Z) \rangle \geq -\lambda\theta(u_1, \ldots, u_{H+1}). \tag{A.7}$$

Since $(u_1, \ldots, u_{H+1})$ is arbitrary,

$$\langle \phi(u_1, \ldots, u_{H+1}), -\frac{1}{\lambda}\nabla_Z L(Z) \rangle \leq \theta(u_1, \ldots, u_{H+1}) \tag{A.8}$$

$$\Leftrightarrow \quad \Omega^{\circ}_{\phi,\theta}(-\frac{1}{\lambda}\nabla_Z L(Z)) \leq 1. \tag{A.9}$$

By Lemma A.1, we get

$$\sum_{i=1}^{r} \theta(W_1^i, \ldots, W_{H+1}^i) = \langle Z, -\frac{1}{\lambda} \nabla_Z L(Z) \rangle. \tag{A.10}$$

This gives $-\frac{1}{\lambda} \nabla_Z L(Z) \in \partial \Omega_{\phi,\theta}(Z)$, concluding the result.

국문초록

최근 깊은 신경망이 여러 분야에서 매우 좋은 성능을 내고 있지만 깊은 신경망의 성능에 대한 이론적인 설명은 부족하다. 깊은 신경망의 손실함수를 최적화하는 것은 매우 중요한 문제이다. 일반적으로 손실함수가 가중매개변수에 대해 볼록하지 않기 때문에, 최적화 알고리즘의 전체적 최적성을 보장할 수 없다. 이 논문에서 우리는 깊은 신경망에서 손실함수의 극소점이 전체적으로 최적이 되는 조건을 알아본다. 추가적으로 우리는 정칙화가 포함된 깊은 신경망에서의 전체적 최적성에 대해 알아보고, 최적화 문제를 고려해 효과적인 신경망의 구조를 제시한다. 또한, 우리의 이론을 실험을 통해 확인한다.

주요어휘: 최적화, 전체적 최적성, 정칙화, 깊은 신경망
학번: 2016-20227