CS260 Project Report: About SGD Convergence Analysis of NN with Non-linear Activation

Song Jiang 605222804*
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
songjiang@cs.ucla.edu

Yewen Wang 905229899*
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
wyw10804@cs.ucla.edu

Zhiping Xiao 604775684*
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
patricia.xiao@cs.ucla.edu

Abstract

Neural networks are increasingly popular during the past few years, with promising performance on various tasks in diverse fields. Recently, SGD based method are widely explored and became a standard training method for Neural Networks.

In this work, we tried to analyze the convergence of SGD on 1-hidden-layer feedforward networks with Non-linear activation and "identity mapping" structure under the constrain that the input obeys normal distribution. We also conduct related experiments to give an intuitive support of our theorem.

Also, we made multiple attempts including exploring SGD's convergence on NN with multiple layers or with special structures or with other activation functions. Discussions about our attempts and corresponding problems we met will be provided in this report.

Note: Our experiment code will be submitted as supplementary materials on CCLE. Also, there are still some unclear parts, and hopefully we can fix those parts later. Most parts of our detailed proof could be provided(handwritting version) if needed.

1 Introduction

Neural networks are increasingly popular during the past few years, with promising performance on various tasks in many fields. Stochastic gradient descent is applied in countless experiments, resulting in satisfying outcome. However, with different network architectures, there are various landscapes and some of them contains bad local minima and sandle points. It still remains not that clear when can SGD guarantee a convergence to global minima and there still lacks solid complete theoretical guarantees that SGD can have good performance when finding the desired weights for such neural networks.

To bridge this gap, in this project, we conducted a comprehensive literature review about theoretical guarantee for SGD convergence analysis of one-hidden-layer neural networks, summarized its related theorems and proof methods, Also, followed the work of [13], we tried to prove the convergence of stochastic gradient descent on feed-forward one-hidden-layer neural network with "identity mapping" structure and Non-linear activation function under the constraint that the network's input satisfies normal distribution. What is more, we took several other attempts including deepening the neural

The author list is sorted according to alphabet.

network, changing the structure of neural networks(to Resnet[9, 8] and DenseNet[10]), or focusing on other non-convex problems. Also, we conducted several auxiliary experiments to provide an intuitive support of our theorem.

The rest of this report is developed as follows: Section 2 provides a thoroughly literature review about SGD convergence analysis on NNs and NN with identity mapping structure. Section 3 illustrates how we formulate our problem and gives our main theorem. Section 4 is an overview of proofs in which we discuss the general proof roadmap and some key difficulties for the proof. Section 5 shows some experiment results which could be an intuitive support for our theorems. Section 6 concludes other attempts we made and discusses possible reasons why we failed in those attempts. Section 7 performs as a summary of our work.

2 Literature Review

Generally, the related works can be divided in to three types, the first type is about SGD convergence analysis for one-hidden-layer neural networks, the second type is the work expanding the network to multiple layers and analyze those deep neural networks, the last type is focusing on neural networks with an advanced structure – identity mapping. In this literature review section, we tried to summarize and elaborate these three types of related work.

2.1 Stochastic Gradient Descent on One-hidden-layer Neural Network

Recently, stochastic gradient descent has been widely leveraged in seeking the optimal parameters of neural network. While proving its convergence to the global minima, problem might occurs because even for the simplest setting where the neural network has only one hidden layer and is a feed forward network, SGD will probably get stuck at local minima or saddle points. To deal with such unsatisfactory situations, various solutions are brought by previous works.

Some works may simplifies the model by ignoring the activation functions and turn the network into a linear deep neural network. The work of Kenji in 2016 is a representative one of this type[11], it first proves that for deep linear neural networks with any depth and any weights and squared loss function, the loss is always non-convex and every local minima would be a global minima, then, this work proves for deep nonlinear neural networks, we can always conduct a reduction to linear model under the independence assumption, thus it can still have those properties they proved before for linear models. With these properties, SGD for such kind of problem can achieve convergence to global minima.

Some works may rely on some unrealistic assumptions with which local minima would have some nice properties[6, 12, 14, 3]. For example, focusing on the simple non-trivia ReLU neural networks, [14] proves that spurious local minima is guaranteed when k, the size of weight per layer, is fixed within certain range. From this point of view, it provides an answer on why neural network are successfully trained even if the associated optimization problem has non-convexity, and what assumptions could mitigate this problem.

Some works design advanced initialization method to guarantee SGD's convergence[16]. In the work of [16] section 5.2, the author applied a Tensor Method to obtain a proper initialization point, with which SGD can converge to global minima.

Some works guarantees SGD's convergence on over-parameterization networks, that is, based on the assumption that the network is wide enough. For example, [5] proved that even with random initialization, first-order methods can achieve zero training loss even if the objective function is neither smooth or convex as long as the hidden layer is wide enough.

Some works rely on specific network structures to ensure SGD can converge to global minima[13]. This is also a work that is most closely to our project. In this work, it proves that with a special structure, "Identity Mapping", stochastic gradient descent will converge to the global minimum of two-layer neural network in polynomial number of steps. The "Identity Mappin" structure makes the network asymmetric and thus guarantees a unique global minimum.

2.2 Convergence Analysis of Deep Neural Network

Despite some success on theoretical analysis for 2-layer neural network, recently, people starts to tackle deeper neural networks and attempts to explore more about NN with multiple layers on the theoretical side. Though the convergence of SGD for deep neural networks still remains an open problem, there are already some existing works focus on such problem[2, 1, 17, 4].

[2] focuses on the convergence speed on Recurrent Neural Networks(RNN) and provides its related theoretical understanding. It proved that as long as the number of neurons is sufficiently large, even with random initialization, SGD can minimize regression loss in a linear convergence rate.

Another interesting work[4] gives theoretical analysis of gradient descent(instead of SGD) and proves rgar GD can reach the global optima for a deep over-parameterized neural network with residual connections. This work relies on such NN architecture to ensure the global optimality of the gradient descent algorithm.

Also, there's one recent work[1] that studies how SGD can get to global minima of deep neural networks(DNN). This work is also based on the assumption that the network is sufficiently wide, and can be applied to fully-connected neural networks, convolutional neural networks (CNN), and residual neural networks (ResNet).

2.3 Network with Identity Mapping

To improve the performance of neural networks, in some previous works, modifications of structures are made to construct stronger networks. Identity mapping, which is presented in [9], is such kind of modifications that gains widely application. It allows that signal could be directly propagated from one neural unit to any others, in both forward and backward passe, which guarantees information to flow unimpeded through the entire network.

Based on the idea of identity mapping, there are two representative neural network architecture that outperform standard NN in most benchmark tasks. "ResNet"[9, 8] is a deep CNN architecture that has impressive performance for tasks such as image classification, object detect, semantic segmentation. Instead of simply stacking layers, ResNet added identity mapping, thus the deeper the network is, the lower error rate it will get(since identity mapping could guarantee that the deeper on will have at least same training error rate as the shallower one). Inspired by ResNet,[10] brought up "DenseNet", which is an advanced convolutional network architecture that achieves the-state-of-art performance in image classification task on various benchmark datasets. This architecture comprises several "dense blocks" connected with convolution layers and pooling layers, each block is a convolutional feed-forward network with skip-connection(which is identity mapping) to flow information between earlier layers and later layers.

3 Main Theorem

In this section, we first puts how we formulate this problem and provide some preliminaries (including notation and related definitions), then we develop our main theorems.

3.1 Problem Formulation

In our project, we tried to analyze SGD's convergence of single-hidden-layer neural network with some non-linear activation. First, for the activation, a Parametric Rectified Linear Unit (PReLU)[7] is proposed which generalizes the traditional rectified unit,which need almost no extra cost for computation and overfitting risk while improving model fitting. Also, inspired by Residual Network[8], we add identity mapping. Thus, our network structure could be presented as Figure 1.

With this architecture, we can develop a formal form for this problem as follows:

Denote by $x \in \mathbb{R}^{n\times 1}$ the input of the network, denote by $W \in \mathbb{R}^{n\times n}$ the parameters forwarding message from input layer to hidden layer, denote by σ the activation function(which is PReLU in our case), the out put of our network architecture could be given as:

$$y(x, W) = ||\sigma(Wx + x)||_1 \tag{1}$$

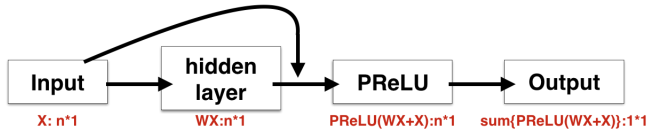


Figure 1: Network Architecture

where $||\cdot||_1$ denotes L1-norm which is the sum of all elements in "$\cdot$". Also, in our setting, input x should obey normal distribution, that is to say, $x \sim \mathcal{N}(0, I)$.

With the standard setting[15], we give the assumption that there exists a teaching network and a student network, both networks share the same structure which is a feed-forward 2 layer network with identity mapping and PReLU activation shown in Figure1. Denote by W^* the parameter(i.e. the weight) of teaching network, and W the parameter of student network. Our goal is to train the student network to learn the results given by teaching network. While training the student network, we apply quadratic loss, thus our objective function would be:

$$L(W) = \mathbb{E}_x[(y(x, W) - y(x, W^*))^2] \tag{2}$$

By applying equation 1, we will have:

$$L(W) = \mathbb{E}_x[(||\sigma(Wx + x)||_1 - ||\sigma(W^*x + x)||_1)^2] \tag{3}$$

And with algebra skills, denote by **e**$= e_1, ..., e_n$ the base vectors, and $W = (w_1, ..., w_n)$, $W^* = (w_1^*, ..., w_n^*)$. Then the above equation3, would be written as:

$$L(W) = \mathbb{E}_x[(\Sigma_i(\sigma(\langle w_i + e_i, x\rangle) - \sigma(\langle w_i^* + e_i, x\rangle)))^2] \tag{4}$$

By formulating the problem into this way, we find it can perfectly fit into a two phase framework brought up by [13] which was designed to analyze the convergence of SGD. In the two phase framework, SGD could be regarded as a 2 stages process. In the first stage, W may head towards wrong direction but with shrinking probability and will definitely get into a specific region(i.e. one-point region, which we will define in section3.2), which lead to stage 2. Then in this stage 2, W will for sure move towards correct direction, and thus finally converge to the target W^*.

3.2 Preliminaries

In this section, we present some definitions and notations that will be used throughout the analysis.

Basic Notations Basic notations appeared in the above sections remain the same as before. Plus, we denote $\theta(v_1, v_2)$ as the angle between vector v_1 and vector v_2, denote by $\bar{M}$ the column-normalized version of matrix M.

Definition1: One Point Strong Convexity[13] A function f(x) is called δ-one point strongly convex in domain D w.r.t. point x^* if $\forall x \in D$, $\langle -\nabla f(x), x^* - x\rangle > \delta||x^* - x||_2^2$.

Definition2: Auxiliary Function Denote $f_A = \Sigma_i(||e_i + w_i^*||_2 - ||e_i - w_i||_2)$ the main auxiliary function, and denote $f_A(i) = f_A - (||e_i + w_i^*||_2 - ||e_i - w_i||_2)$

Definition3: Auxiliary Matrix Denote $A = (W^* + I)\overline{W^* + I}^T - (W + I)\overline{W + I}^T$ the main auxiliary matrix, and denote $A(i) = A - ((e_i + w_i^*)\overline{(e_i + w_i^*)}^T - (e_i + w_i)\overline{(e_i + w_i)}^T)$

3.3 Theorems

Main Theorem While $x \sim \mathcal{N}(0, I)$, $||w||_2 \leq \gamma$ and $||W^*|| \leq \gamma^*$ are both bounded with some small constantγ, γ^*, SGD with small learning rate η and initial W_0(random/zero/standard all work) will converge to W^* within polynomial number of steps, in two phase.

This main theorem could be divided into two theorems, which are the theorem for phase 1 and phase 2 respectively.

Theorem for Phase 1 While $x \sim \mathcal{N}(0, I)$, $||w||_2 \leq \gamma$ and $||W^*|| \leq \gamma^*$ are both bounded with some small constantγ, γ^*, then the auxiliary function f_A will keep decreasing after every step until $f_A \leq \epsilon$ is small enough(which means we enter phase 2).

Theorem for Phase 2 While $x \sim \mathcal{N}(0, I)$, $||w||_2 \leq \gamma$ and $||W^*|| \leq \gamma^*$ are both bounded with some small constantγ, γ^*, and $f_A \leq \epsilon$, then $\langle -\nabla L(W), W^* - W \rangle > \delta ||W^* - W||_2^2$ with constant δ.

4 Overview of Proofs

This section gives an overview of proofs of our main theorem. We first provide a flowchart as a clear roadmap to our proof, then we illustrate the key point of each part of the proofs. About detailed proofs, due to time limitation, we didn't type out all our proofs within this latex document. We can provide a handwriting version of detailed proof if needed.

4.1 Generally: A 2-stage Process

To proof our theorem, we divide our proof into 2 stages. In the first stage, we proof that our W will definitely enter the one-point convexity region. To check if its a one-point convexity region, we calculate $\langle -\nabla L(W), W^* - W \rangle$ and proof it will be less than or equal to $\delta ||W^* - W||_2^2$. Here we have,

$$
\begin{aligned}
\nabla L(W)_j = 2\mathbb{E}_x[(\Sigma_i \sigma(\langle e_i + w_i, x \rangle) - \Sigma_i \sigma(\langle e_i + w_i^*, x \rangle))x\mathbb{1}_{\langle e_j + w_j, x \rangle \geq 0} \quad (5) \\
+ (\Sigma_i \sigma(\langle e_i + w_i, x \rangle) - \Sigma_i \sigma(\langle e_i + w_i^*, x \rangle))x\alpha\mathbb{1}_{\langle e_j + w_j, x \rangle < 0}] \quad (6)
\end{aligned}
$$

Then, after entering the one-point convexity region, we move to stage 2 and proof after each step, W will get closer to W^* with no exception.

4.2 For Phase I

Phase I aims at proving that, $\exists \gamma_0 \in (0, \gamma)$ s.t. If $||\mathbf{W}_0||_2, ||\mathbf{W}^*||_2 \leq \gamma_0$, and d has a constant lower bounded, η has a upper bound determined by γ and G the gradient, ϵ upper bounded by a term depending solely on γ, then the auxiliary function f_A will keep decreasing until it reaches $O(1)$, which is the condition of starting *Phase 2*.

The decreasing factor for each step is influenced by η and d, the total number of iterations needed is determined by η, and the final value of f_A after Phase I ends is decided by γ.

In order to prove the above claims and to give a clear bound of all terms, we use approximations.

The way we prove the decreasing trend of f_A is by first introducing an auxiliary variable $s = (\mathbf{W}^* - \mathbf{W})u$, where u is the all-one vector, and then express $s^{(t)}$ and $f_A^{(t)}$, using $s^{(t-1)}$ and $f_A^{(t-1)}$. Then we could solve the dynamics from the previous step to show that g, a potential function that is expected to depend on f_A, approaches to and stays around $O(\gamma)$.

The second task of Phase I is to use the conclusion that f_A decreases to prove that $||\mathbf{W}||$ remains small. This limitation will guarantee that once we move on to Phase II, there is no possibility of coming back to Phase I.

A and f_A are useful tools to help with constructing a matrix P to approximate $-\nabla L(W)$. The proof of that P is an appropriate approximation should also be provided in Phase I.

4.3 For Phase II

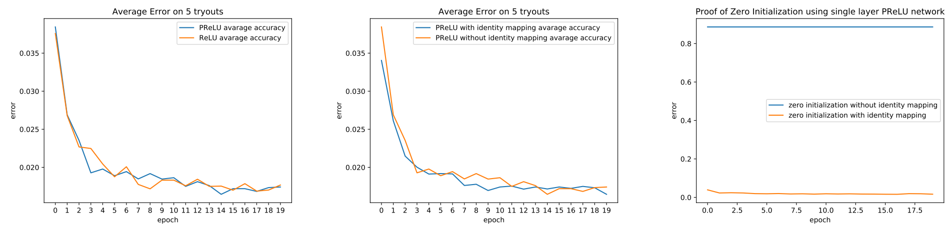
The goal of Phase II is to prove that SGD can obtain optimal parameters in the small region derived from Phase I. i.e., Prove that $\exists \gamma$ with a small enough f_A, s.t. $L(W)$ is a δ one point strongly convexity. The formal is shown as follow.

$$
\langle -\nabla L(W), W^* - W \rangle = \sum_{j=1}^{d} \langle -\nabla L(W)_j, w_j^* - w_j \rangle > \delta ||W^* - W||_F^2 \quad (7)
$$

Here we use Taylor expansion and control the higher order term, which is shown in figure 2 Then, lower bound each part of Taylor expansion separately. Note that when $W \approx W^*$, we will use

$$\left\langle \boxed{\text{constant}} + \boxed{\text{1 order}} + \boxed{\text{higher order}}, \quad W^* - W \right\rangle$$

Figure 2: Taylor expansion



(a) (b) (c)

Figure 3: **(a)**.Accuracy for NN with ReLU activation and PReLU activation, **(b)**.Accuracy Curve for NN architecture with and without identity mapping structure, **(c)**.Performance of NN with or without Identity Mapping while given Zero Initialization

joint Taylor expansion to overcome a large higher term. Then, we lower bound each part of Taylor expansion by $c * ||W^* - W||_F^2$, where is c is a small constant related to f_A. Note that the value of f_A is derived via Phase I.

5 Experiments

We designed a 4-stages experiment, each stage of experiment will perform as an auxiliary support for our main theorem.

5.1 Stage1: Show the Advancement of PReLU Activation

In the first stage of our experiment, we tried to show the advancement of PReLU activation even with one-hidden-layer neural network. Here we used MNIST handwritten digits dataset, trained with two layer NN without identity mapping, and with PReLU activation and ReLU activation respectively(so there are two different settings). We conducted each setting of experiments for 5 times and took the average of those tryouts. The curve of error rate we obtained is presented in Figure3(a).

For ReLU without identity mapping, random initialization, we get an average error rate of 0.01768. And for PReLU, without identity mapping, random initialization, we get an average error rate of 0.01742. The results indicates PReLU's advancement successfully.

5.2 Stage2: Validate Identity Mapping Helps in Improving Accuracy

In the second stage of our experiment, we tried to show the advancement of identity mapping with one-hidden-layer neural network. The dataset we applied remains the same as in stage 1, trained with two layer NN with and without identity mapping respectively, and with PReLU activation(so there are still two different settings). We conducted each setting of experiments for 5 times and took the average of those tryouts. The curve of error rate we obtained is presented in Figure3(b).

For PReLU without identity mapping, random initialization, we get an average error rate of 0.01742. And for PReLU with identity mapping, random initialization, we get an error rate of 0.01646. The result indicates identity mapping's advancement successfully.

5.3 Stage3: Validate Zero Initialization Works

In the third stage of our experiment, we tried to show the important role identity mapping plays while with zero initialization. Again, most the experiment settings remains the same as in stage 2, but the

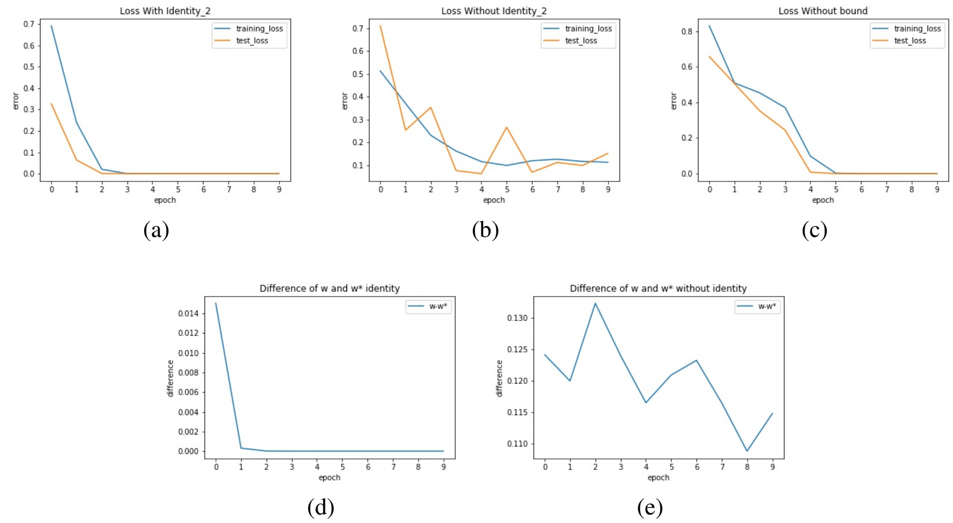


Figure 4: **(a)**.Loss with identity mapping and bound, **(b)**.Loss without Identity mapping, **(c)**.Loss without bound, **(d)**$||W - W^*||_2$ with identity mapping, **(e)**$||W - W^*||_2$ without identity mapping

for the initialization we use zeros instead of some random values. The curve of error rate we obtained is presented in Figure3(c).

For PReLU without identity mapping, zero initialization, we get an average error rate of 0.8865. And for PReLU with identity mapping, zero initialization, we get an error rate of 0.0165. This result sufficiently illustrates that with identity mapping, the zero initialization will still work(and almost have the same performance as random initialization).

5.4 Stage4: Global Minimum Convergence

Following the work of[15, 13], we utilized one teacher network and at least one student network, where teacher network knows the ground truth optima parameters $\mathbf{W}^*$, while the students will learn $\mathbf{W}$ via l_2 loss.

In this stage, we tried to show that, while $||W||_2$ and $||W^*||_2$ are bounded by some small value, when applying identity mapping, SGD can converge to global minima with zero error rate. Here we used random generalized input dataset of dimension $d = 10$ and size of 5000, both our teaching network and student network share the same structure of 2-layer with PReLU activation, for the training process of student network, we set learning rate as 0.1 and epoch as 10. We observed the loss with and without identity mapping, and with and without bound of $||W^*||_2$. For the bound, we set $||w^*||_2$ = 0.1 and $||w^*||_F$ = 1. To avoid coincidence, we conducted each setting of experiments for 5 times and took the average of those tryouts. The curves we obtained is presented in Figure4.

This result sufficiently presents that both the bound of $||W^*||_2$ and the identity mapping are crucial to SGD's convergence.

6 Other Attempts

We also took several steps in the SGD convergence analysis for neural networks, trying to see if this nice 2-phase framework is still applicable. Those attempts include: extending the proof framework to multiple layers neural networks, varying the network structures, and formulating several other non-convex problems and applying the proof framework to them. Though failed to achieve complete and solid proofs (or sometimes even failed to find a proper way to formulate the main theorem), we still want to discuss about those attempts and analyze why we failed in those attempts.

6.1 Deepen the Network

Our first step is to deepening the network by stacking multiple layers with identity mapping and trying to extend the 2 phase framework[13] to deep neural networks. Then, by doing so, the problem would be formulated as(take N-hidden-layers as example):

$$y(x, W) = ||\sigma((W_N + I)....\sigma((W_2 + I)\sigma((W_1 + I)x)))||_1 \tag{8}$$

Then, while dealing with L(W), we find it hard to formulate it in an elegant way like equation4 does, thus, by referring to the work[11], we tried to applied the way brought up by this work to turn our non-linear deep neural network into a linear form via reduction, thus it would be equipped with the following formula:

$$y(x, W) = ||\sigma((W_N...W_2W_1 + I)x))||_1 \tag{9}$$

which could be the exactly the same as the problem setting the 2-phase framework has. However, this would cause some problems. Because to simplify the network into a linear one, we ignored the activation functions by adding some constrains, which makes this problem less flexible, and thus made sure that all the global minima would be local minima, so that the proof of SGD's convergence to global minima no longer meaningful(since one of the key contribution to the convergence analysis is to proof it can avoid sticking at local minima).

According to the above statement, we found applying the two-stage framework on deep neural networks seems not to be an applicable idea.

6.2 Vary the Network Structures

We also tried to explore more about the convergence of SGD in diverse network structures. Based on the nice properties identity mapping has, naturally we came to ResNet and DenseNet, which are mentioned in section 2.3.

However, with such network structures, it would then turn to deep neural network, thus we encountered the same problems in section 6.1, where the 2-stage proof framework no longer works for this situation.

6.3 For Other Non-convex Problems

Since in previous sections of this report, we've already formulate this problem in a nice way in which network function and loss function could be represent with the same equation14. By simply changing the activation function σ, our problem will be different(not slightly) while still remains a non-convex problem. However, even though, and thus we encountered lots of difficulties while figuring out the corresponding auxiliary function and auxiliary matrix. Since for each different activation function, there's no uniform form of auxiliary function and auxiliary matrix(which means we need a different form of auxiliary function and auxiliary matrix), this attempt would turn to different direction and need diverse proof techniques corresponding with different activation. And we do not have enough time go down every branch of this path.

7 Conclusion

To sum up, in this project, our main work could be summarized into 4 parts, first, we conducted a comprehensive literature review including 3 types of related work(details are in section 2), second, we proved the convergence of SGD with one-hidden-layer neural network with PReLU activation and without constrains on initialization, third, though failed, we made several attempts on applying the nice 2-phase framework on NN with different layers, structures, or for different non-convex problems and analyzed why we can't get expected theoretical results, last, we conducted a 4-stage experiment to support our theory intuitively.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065*, 2018.

[3] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[4] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[5] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[6] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[11] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[12] Ilja Kuzborskij and Christoph H Lampert. Data-dependent stability of stochastic gradient descent. *arXiv preprint arXiv:1703.01678*, 2017.

[13] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[14] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[15] Yuandong Tian. Symmetry-breaking convergence analysis of certain two-layered neural networks with relu nonlinearity. 2017.

[16] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

[17] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.