About SGD Convergence Analysis of 2 layer NN with Non-linear Activation

Song Jiang, Yewen Wang, Zhiping Xiao

12/12/2018

Contents

Contents

Highlight

- Comprehensive literature review
- SGD Convergence analysis on on-hidden-layer NN with Non-linear activation
- Other attempts on SGD convergence analysis on NN with different layers, structures, or for different non-convex problems
- Auxiliary experiments

Contents

Background: Stochastic Gradient Descent on One-hidden-layer Neural Network

- Simplifies the model by ignoring activation functions and turn the NN into a linear one
- Rely on unrealistic assumptions with which can achieve some nice properties such as all local are global
- With fancy well-designed initialization method
- Extra condition that the network should be wide enough
- Rely on specific network structures
- ...

Background: Convergence Analysis of Deep Neural Network

- Still an open problem
- Almost all of them need the condition that the NN should be over-parameterized

Background: Network with Identity Mapping

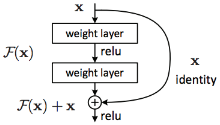

Figure 1: ResNet

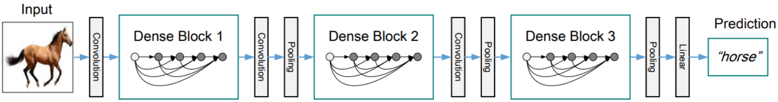

Figure 2: DenseNet

Contents

Problem Formulation: Network Architecture

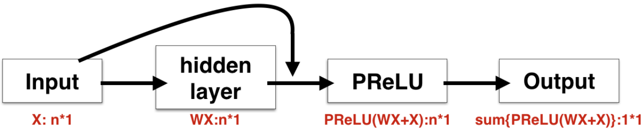

Figure 3: Network Architecture

- One-hidden-layer
- Extra identity mapping
- PReLU Activation
- Teaching network and student network

Problem Formulation: Objective Function

- $y(x, W) = ||\sigma(Wx + x)||_1$
- $L(W) = \mathbb{E}_x[(y(x, W) - y(x, W^*))^2]$
- $L(W) = \mathbb{E}_x[(||\sigma(Wx + x)||_1 - ||\sigma(W^*x + x)||_1)^2]$
- $L(W) = \mathbb{E}_x[(\Sigma_i(\sigma(\langle w_i + e_i, x\rangle) - \sigma(\langle w_i^* + e_i, x\rangle)))^2]$

Denote: $x \in \mathbb{R}^{n\times 1} \sim \mathcal{N}(0, I)$ is input, y is the output, $W \in \mathbb{R}^{n\times n}$ is weight, σ is the activation function, $||\cdot||_1$ is L1-norm, **e**$= e_1, ..., e_n$ are the base vectors, $W = (w_1, ..., w_n)$, $W^* = (w_1^*, ..., w_n^*)$.

Problem Formulation: Proof Framework

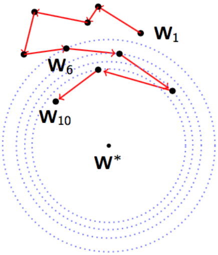

Figure 4: Dynamics

- 2-phase process
- Phase1: probability of going to wrong direction decrease
- Phase2: get closer to global optima after each step

Contents

Main: Preliminaries

- **One-point convexity**: A function f(x) is called δ-one point strongly convex in domain D w.r.t. point x^* if $\forall x \in D$, $\langle -\nabla f(x), x^* - x \rangle > \delta ||x^* - x||_2^2$
- **Auxiliary Function**: Denote $f_A = \Sigma_i(||e_i + w_i^*||_2 - ||e_i - w_i||_2)$ the main auxiliary function, and denote $f_A(i) = f_A - (||e_i + w_i^*||_2 - ||e_i - w_i||_2)$
- **Auxiliary Matrix**: Denote $A = (W^* + I)\overline{W^* + I}^T - (W + I)\overline{W + I}^T$ the main auxiliary matrix, and denote $A(i) = A - ((e_i + w_i^*)\overline{(e_i + w_i^*)}^T - (e_i + w_i)\overline{(e_i + w_i)}^T)$

Main: Main Theorem

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

Main: Overview of Proof

- Proof SGD will converge to global optima following the 2-phase process!
- Phase 1: auxiliary function decrease, and get into one point convex region
- Phase 2: after every step, get closer to $||W^*||$

Main: Overview of Proof for Phase 1

Goal: Prove that $\exists \gamma_0 \in (0, \gamma)$ s.t. if

- $\|\mathbf{W}_0\|_2, \|\mathbf{W}^*\|_2 \leq \gamma_0$
- d lower bounded by a constant, η with a upper bound determined by γ and G (gradient), ϵ with upper bound depending on γ

then: f_A **will keep decreasing** until reaches *Phase 2* (auxiliary function decreases to $O(1)$).
The decreasing factor for each step depends on η, d, number of iterations depend on η, and the upper bound of f_A after *Phase 1* is decided by γ.

Main: Overview of Proof for Phase 1

Solution: approximation, introduces an auxiliary variable $s = (\mathbf{W}^* - \mathbf{W})u$, u is the all-one vector.

1. Show that *Phase 1* will reach *Phase 2*:
   1.1 Calculate the update for f_A and s. Expected to have an approximation of $s^{(t)}$ and $f_A^{(t)}$, each depends on both $s^{(t-1)}$ and $f_A^{(t-1)}$
   1.2 Solve the dynamics from the above step to show that f_A approaches to and stays around $O(\gamma)$.
2. Show that there's **NO WAY BACK** from *Phase 2* to *Phase 1*.
   - (f_A decreases $\Rightarrow$) $\|\mathbf{W}\|_2$ remains small
3. Justify the form of A and f_A
   - Prove that using g and A we could successfully formulate an approximation matrix $\mathbf{P}$ which approximates $-\Delta L(\mathbf{W})$.

Main: Overview of Proof for Phase 2

Goal: Prove that $\exists\gamma$, with a small enough f_A, s.t. $L(W)$ is a δ one point strongly convexity. i.e., $\langle -\nabla L(W), W^* - W\rangle = \sum_{j=1}^{d}\langle -\nabla L(W)_j, w_j^* - w_j\rangle > \delta ||W^* - W||_F^2$
Solution: Use Taylor expansion and control the higher order term.

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

Then, lower bound each part of Taylor expansion. Note that when $W \approx W^*$, we will use joint Taylor expansion to overcome a large higher term.

Contents

Experiments: Stage 1

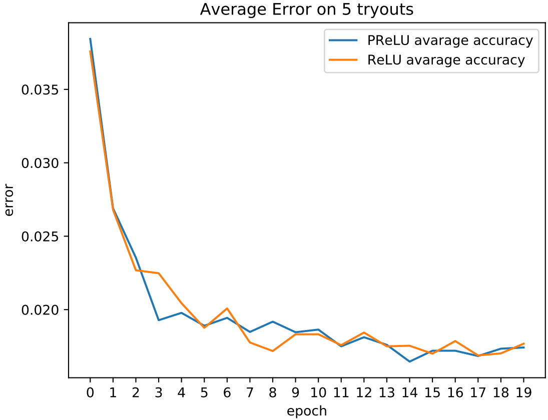

Figure 5: Accuracy for NN with ReLU activation and PReLU activation

Experiments: Stage 2

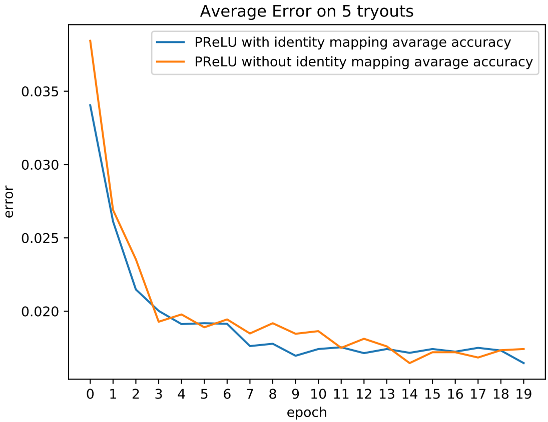

Figure 6: Accuracy Curve for NN architecture with and without identity mapping structure

Experiments: Stage 3

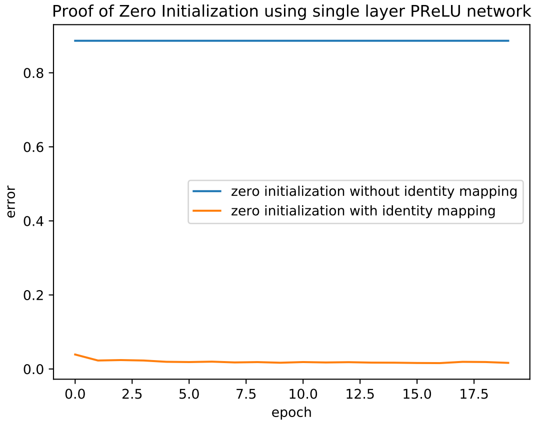

Figure 7: Performance of NN with or without Identity Mapping while given Zero Initialization

Experiments: Stage 4

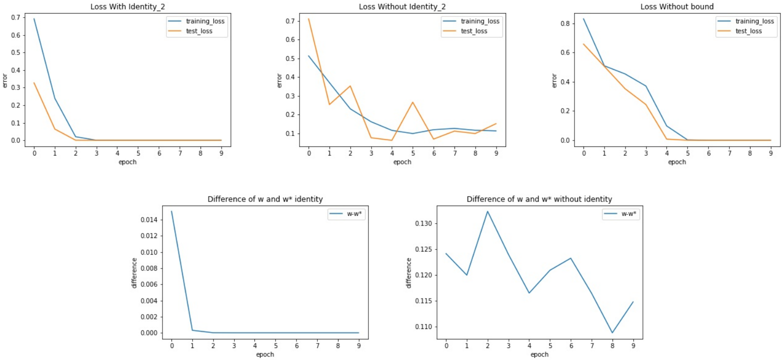

Figure 8: **(a)**Loss with identity mapping and bound, **(b)**Loss without Identity mapping, **(c)**Loss without bound, **(d)**$||W - W^*||_2$ with identity mapping, **(e)**$||W - W^*||_2$ without identity mapping

Contents

Attempts: Deepen the Network

- $y(x, W) = ||\sigma(W_N....\sigma(W_2\sigma((W_1x)))||_1$
- Turn to linear? $y(x, W) = ||\sigma(W_N...W_2W_1x))||_1$
- Not applicable!

Attempts: Vary the Network Structures

- ResNet..DenseNet..?
- Common constrain: over-parameterized!
- $y(x, W) = ||\sigma((W_N + i_N I)....\sigma((W_2 + i_2 I)\sigma((W_1 + i_1 I)x)))||_1$ where i_j is 0 or 1 indicating if this layer has an identity mapping
- Still not applicable

Attempts: Several Non-convex Problems

- When σ varies, become different non-convex problem
- Slightly change..?
- No! A lot of work needed including redefine auxiliary matrix and auxiliary function, thus will lead to totally different proof method for each stage!

THANKS $\|$ FREE TO ASK

- **Comprehensive literature review**: 3 types
- **SGD Convergence analysis**: on-hidden-layer NN with PReLU activation
- **Several attempts**: NN with different layers, structures, or for different non-convex problems
- **Auxiliary experiments**: 4 stage