Convergence of one hidden layer neural network with trainable output layer weights

Abhishek Panigrahi

November 2018

1 Introduction

This is a documentation of applying Simon Du's analysis of convergence of a single hidden layer neural network with fixed output layer weights to a a single hidden layer neural network, with trainable output layer weights.

2 Notation

Let the data samples be denoted by {X, Y}, where X denotes a set of data points and Y denotes the set of corresponding labels. Let the input layer weights be denoted by W and the output layer weights be denoted by a. The number of input dimension is denoted by d and number of neurons in the hidden layer is denoted by m. Let the number of samples available be denoted by n. Hence, the neural network can be given in functional form as

$$f(W, a, x) = \frac{1}{\sqrt{m}} \sum_{r \in [m]} a_r \sigma(W_r^T x)$$

where σ denotes the activation function, in this case relu.

Let the labels, denoted by y, have magnitude bounded by C. we use a l2 loss function to learn the neural network. (note that we can have the same derivation in binary classification with cross entropy loss). Hence, the loss function is given by

$$L(W, a, \{X, Y\}) = \sum_{i=1}^{n} \|f(W, a, x_i) - y\|^2$$

Let u denote the prediction vector, where u_i is equal to f(W, a, x_i). For representing a variable v at epoch k, we use superscript k ($v^{(k)}$).

3 Assumption

The neural network is trained in an alternative minimization procedure i.e. each training step is divided into 2 substeps, the output layer weights are trained in first substep, with the input layer weights fixed and vice versa for the second substep. This assumption has been taken to make the analysis easier.

4 Analysis

The analysis will follow an induction approach to calculate an upper bound on the ratio of loss value at epoch k to the loss value at epoch 0.

The next two subsections will derive a bound for gradient descent iteration k + 1.

4.1 Substep 1

$$a^{(k+1)} = a^{(k)} - \eta \frac{\partial L(W^{(k)}, a^{(k)})}{\partial a^{(k)}}$$

Let the new prediction vector be denoted by u_{mid}.

$$u_{mid}^{(k)} = f(W^{(k)}, a^{(k)}, X)$$

$$\begin{aligned} u_{mid,i}^{(k)} - u_i^{(k)} &= \tfrac{1}{\sqrt{m}} \textstyle\sum_{r\in[m]} \ (a_r^{(k+1)} - a_r^k) \ \sigma(W_r^{(k),T} x_i) \\ &= \tfrac{1}{\sqrt{m}} \textstyle\sum_{r\in[m]} \ \eta \ \frac{-\partial L}{\partial a_r^{(k)}} \ \sigma(W_r^{(k),T} x_i) \\ &= \tfrac{\eta}{m} \textstyle\sum_{j=1}^{n} \ (y_j - u_j^{(k)}) \ \sum_{r=1}^{m} \sigma(W_r^{(k),T} x_i) \ \sigma(W_r^{(k),T} x_j) \\ &= \eta \textstyle\sum_{j=1}^{n} \ (y_j - u_j^{(k)}) \hat{H}_{i,j} \end{aligned}$$

where $\hat{H}$ is a matrix of the form

$$\hat{H}_{i,j} = \frac{1}{m} \sum_{r=1}^{m} \sigma(W_r^{(k),T} x_i) \ \sigma(W_r^{(k),T} x_j)$$

Note that $\hat{H}$ is a positive semi definite matrix.

$$\Rightarrow (y - u^{(k)})^T (u_{mid}^{(k)} - u^{(k)}) = \eta (y - u^{(k)})^T \hat{H} (y - u^{(k)}) \geq 0 \tag{1}$$

Also,

$$\begin{aligned} |\hat{H}_{i,j}| &= \tfrac{1}{m} | \textstyle\sum_{r\in[m]} \sigma(W_r^{(k),T} x_i) \ \sigma(W_r^{(k),T} x_j)| \\ &\leq \tfrac{1}{m} \textstyle\sum_{r\in[m]} |\sigma(W_r^{(k),T} x_i)| \ |\sigma(W_r^{(k),T} x_j)| \\ &\leq \tfrac{1}{m} \textstyle\sum_{r\in[m]} |W_r^{(k),T} x_i| |W_r^{(k),T} x_j| \\ &\leq \tfrac{1}{m} \textstyle\sum_{r\in[m]} \|W_r^{(k)}\|_2^2 \end{aligned}$$

$$\begin{aligned} \Rightarrow \|\hat{H}\|_2^2 &\leq \|\hat{H}\|_F^2 \\ &= \textstyle\sum_{i,j} |\hat{H}_{i,j}|^2 \\ &= \frac{n^2}{m^2}(\textstyle\sum_{r\in[m]} \|W_r^{(k)}\|_2^2)^2 \end{aligned}$$

and

$$\|u_{mid}^{(k)} - u^{(k)}\|_2^2 \leq \eta^2 \|\hat{H}\|_2^2 \|y - u^{(k)}\|_2^2$$

That implies,

$$\|u_{mid}^{(k)} - u^{(k)}\|_2^2 \leq \eta^2 \frac{n^2}{m^2} \sum_{r\in[m]} (\sum_{r\in[m]} \|W_r^{(k)}\|_2^2)^2 \tag{2}$$

$$\begin{aligned} \|y - u_{mid}^{(k)}\|_2^2 &= \|(y - u^{(k)}) - (u_{mid}^{(k)} - u^{(k)})\|_2^2 \\ &= \|y - u_{mid}^{(k)}\|_2^2 + \|u_{mid}^{(k)} - u^{(k)}\|_2^2 - 2(y - u^{(k)})^T (u_{mid}^{(k)} - u^{(k)}) \end{aligned}$$

Using 2 and 1, we get

$$\|y - u_{mid}^{k}\|_2^2 = (1 + \eta^2 \frac{n^2}{m^2} (\sum_{r\in[m]} \|W_r^{(k)}\|_2^2)^2) \|y - u^{(k)}\|_2^2 \tag{3}$$

4.2 Substep 2

$$\begin{aligned} u_i^{(k+1)} - u_i^{(k)} &= \tfrac{1}{\sqrt{m}} \textstyle\sum_{r\in[m]} a_r^{(k+1)} \ (\sigma(W_r^{(k+1),T} x_i) - \sigma(W_r^{(k),T} x_i)) \\ &= \tfrac{1}{\sqrt{m}} \textstyle\sum_{r\in[m]} a_r^{(k+1)} \ (\sigma((W_r^{(k)} - \eta \frac{\partial L}{\partial W_r^{(k)}})^T x_i) - \sigma(W_r^{(k),T} x_i)) \end{aligned}$$

I follow the same steps that are being used in Du et al's discrete analysis. I will mention those steps in brief detail below to come up with equations for this substep.

- Let $A_{i,r}$, S_i be defined as follows

  $A_{i,r} = \{\exists w \colon \|w - w_r^{(0)}\|_2 < R \text{ and } 1_{w^T x_i > 0} \neq 1_{w_r^{(0),T} x_i > 0}\}$
  $S_i = \{r \in [m] | A_{i,r} = 0\}$
  $S_i^{\perp} = [m] \ S_i$

- Define I_1^i as

  $$I_1^i = \frac{1}{\sqrt{m}} \sum_{r\in S_i} a_r^{(k+1)} \ (\sigma(W_r^{(k+1),T} x_i) - \sigma(W_r^{(k),T} x_i))$$

  It can be simplified as

  $$I_1^i = \frac{\eta}{m} \sum_{j=1}^{n} \tilde{H}_{i,j}^{(k)} (y_j - u_{j,mid}^{(k)})$$

where

$$\tilde{H}_{i,j}^{(k)} = \frac{1}{m} \sum_{r \in S_i} a_r^{(k+1),2} 1_{w_r^{k,T} x_i > 0, w_r^{k,T} x_j > 0}$$

We also write $\tilde{H} = H - H^{\perp}$, as part of notation used by Du.

$$\tilde{H}_{i,j}^{\perp,(k)} = \frac{1}{m} \sum_{r \in S_i^{\perp}} a_r^{(k+1),2} 1_{w_r^{k,T} x_i > 0, w_r^{k,T} x_j > 0}$$

$$H_{i,j}^{(k)} = \frac{1}{m} \sum_{r \in [m]} a_r^{(k+1),2} 1_{w_r^{k,T} x_i > 0, w_r^{k,T} x_j > 0}$$

- Following Du's analysis, we get

$$\begin{aligned} |H_{i,j}^{\perp,(k)}| &= \tfrac{1}{m} |\tfrac{1}{m} \textstyle\sum_{r \in S_i^{\perp}} a_r^{(k+1),2} 1_{w_r^{k,T} x_i > 0, w_r^{k,T} x_j > 0}| \\ &\leq \tfrac{1}{m} \textstyle\sum_{r \in S_i^{\perp}} |a_r^{(k+1)}|^2 \\ &\leq \tfrac{1}{m} |S_i^{\perp}| \max_{r \in S_i^{\perp}} |a_r^{(k+1)}|^2 \end{aligned}$$

  That implies

$$\|H^{\perp}\|_2 \leq \tfrac{n}{m} \textstyle\sum_i |S_i^{\perp}| max_{r \in S_i^{\perp}} |a_r^{(k+1)}|^2$$

- Define I_2^i as

$$I_2^i = \frac{1}{\sqrt{m}} \sum_{r \in S_i^{\perp}} a_r^{(k+1)} \ (\sigma(W_r^{(k+1),T} x_i) - \sigma(W_r^{(k),T} x_i))$$

  Going by Du's analysis, I get

$$|I_2^i| \leq \eta \frac{\sqrt{n}}{m} max_{r \in S_i^{\perp}} (a_r^{(k+1)})^2 |S_i^{\perp}| \|y - u_k^{\perp}\|_2$$

- Also,

$$\begin{aligned} \|u^{(k+1)} - u_{mid}^{(k)}\|_2^2 &\leq \tfrac{\eta^2}{m} \textstyle\sum_{i=1}^n (\sum_{r \in [m]} a_r^{(k+1)} \| \frac{\partial L}{\partial W_r^{(k)}} \|)^2 \\ &\leq \tfrac{\eta^2}{m} \textstyle\sum_{i=1}^n \|a^{(k+1)}\|^2 \| \frac{\partial L}{\partial W^{(k)}} \|_2^2 \\ &\leq \eta^2 n^2 \|y - u^{(k)}\|^2 \| \|a^{(k+1)}\|^4 \end{aligned}$$

Combining all the steps above, I get

$$\begin{aligned} \|y - u^{(k+1)}\|_2^2 &= \|(y - u_{mid}^{(k)}) - (u^{(k+1)} - u_{mid}^{(k)})\|_2^2 \\ &= \|y - u_{mid}^{(k)}\|_2^2 + \|u^{(k+1)} - u_{mid}^{(k)}\|_2^2 - 2(y - u_{mid}^{(k)})^T \ (u^{(k+1)} - u_{mid}^{(k)}) \\ &= \|y - u_{mid}^{(k)}\|_2^2 - 2\eta (y - u_{mid}^{(k)})^T H^{(k)} (y - u_{mid}^{(k)}) \\ &\quad + 2\eta (y - u_{mid}^{(k)})^T H^{(k),\perp} (y - u_{mid}^{(k)}) - 2\eta (y - u^{(k)})^T I_2 + \|u^{(k+1)} - u_{mid}^{(k)}\|_2^2 \\ &\leq (1 - \eta\lambda_0 + (2\eta n^2 R + 2\eta n^{1.5} R) \ max_{r \in S_i^{\perp}} |a_r^{(k+1)}|^2 + \eta^2 n^2 \|a^{(k+1)}\|_2^4) \|y - u_{mid}^{(k)}\|_2^2 \\ &\leq (1 - \eta\lambda_0 + (2\eta n^2 R + 2\eta n^{1.5} R) \ max_{r \in S_i^{\perp}} |a_r^{(k+1)}|^2 + \eta^2 n^2 \|a^{(k+1)}\|_2^4) \\ &\quad (1 + \eta^2 \tfrac{n^2}{m^2} (\textstyle\sum_{r \in [m]} \|W_r^{(k)}\|_2^2)^2) \|y - u^{(k)}\|_2^2 \end{aligned}$$