Notes on Deep Neural Networks

Clay L. McLeod — clay.l.mcleod@gmail.com

January 27, 2016

1 Neural Network Basics

1.1 Weight Initializations

1.1.1 Lecun's Distribution

Presented in [**?**, Sec 4.6], Lecun's distribution assumes a linear model and is based on the following argument: suppose we have a neural network activation layer that uses $tanh(X)$ as a nonlinear squashing function. In order for convergence to occur quickly, the weights should be initialized so that (1) the weights are not too small, causing the gradient function to be small and (2) the $tanh$ is not saturated (the weights are not too large), also causing the gradient function to be small. Assuming that the data is properly normalized, all we need to do is initialize the weights to have $\mu = 0$ and $\sigma = 1$. Thus, the equations for Lecun's distribution are as follows:

$$X \sim N\left[-\frac{1}{\sqrt{n_j}}, \frac{1}{\sqrt{n_j}} \right] \tag{1}$$

$$X \sim U\left[-\frac{3}{\sqrt{n_j}}, \frac{3}{\sqrt{n_j}} \right] \tag{2}$$

1.1.2 Glorot Distribution

Presented in [**?**, 4.2].

$$X \sim N\left[-\frac{\sqrt{2}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{2}}{\sqrt{n_j + n_{j+1}}} \right] \tag{3}$$

$$X \sim U \left[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}} \right] \tag{4}$$

1.1.3 He Distribution

Previous assumes linear activation function, which is not suitable for ReLU or its derivatives. Thus, the authors of [**?**][pg. 4] derive a theoretically sound initialization for networks utilizing the ReLU activation family.

$$X \sim N \left[-\sqrt{\frac{2}{n_j}}, \sqrt{\frac{2}{n_j}} \right] \tag{5}$$

$$X \sim U \left[-\sqrt{\frac{6}{n_j}}, \sqrt{\frac{6}{n_j}} \right] \tag{6}$$

2 Experiments

2.1 Momentum ReLU