CSCI 5525: Machine Learning (Fall'17)

Homework 4, Due 12/14/17

(30 points) This problem considers optimization methods for deep learning based on a mini-batch $\{(\mathbf{x}_1, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ of size m in each iteration. Let θ denote the parameter vector of the deep learning model.

(a) (20 points) Describe the Adagrad algorithm for iteratively computing the parameter vector θ. What limitations of the basic stochastic gradient descent approach does the Adagrad algorithm address and how? What are the limitations of the Adagrad approach for deep learning?

(b) (10 points) Explain why gradient based methods for learning parameters in a deep learning model may face the problem of exploding or vanishing gradients. Do you expect to see such problems more in the deeper layers closer to the input or shallower layers closer to the output? Briefly justify your answer.

2. (20 points) Read the paper http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf. Clearly explain the optimization hypthesis and the regularization hypothesis and explain how the experiments provide evidence for or against each hypothesis.

(30 **points** Extra credit) Implement stacked denoising autoencoder pretraining for MNIST following http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf. Read the paper, focused on section 5.1.2. Implement in Tensorflow. Submit write-up, code and a plot like Figure 11 for the MNIST dataset ONLY for 3 layers with and without pretraining.