ECE 595: Machine Learning II
Fall 2019
Instructor: Prof. Aly El Gamal

PURDUE
UNIVERSITY

Homework 3

Fall 2019
(Due: Nov. 1, 2019)

Introduction

This assignment is on deep neural network optimization, which is covered in Chapter 8 of the recommended text.

Exercises

1. (Stochastic Gradient Descent - Minibatch Size)

   Describe the major factors that contribute to choice of the minibatch size for stochastic gradient descent. What happens when we use a batch size of 1?

2. (Stochastic Gradient Descent - Randomness)

   a) During stochastic gradient descent, why should we select the minibatches randomly by shuffling the order of the dataset?

   b) For very large datasets, for example, datasets containing billions of examples in a data center, how is the random minibatch selection done in practice?

3. (Ill-Conditioning)

   a) Briefly explain what the phenomenon of Hessian ill-conditioning is.

   b) Discuss the impact of Hessian ill-conditioning on learning rate evolution with gradient descent.

4. (Saddle Points)

   a) For neural networks with a large parameter space, why do we expect saddle points to occur more frequently than local minima?

   b) Recall that second-order methods such as Newton's method are often designed to solve for a point in the parameter space where the gradient of the loss function is zero. Explain why the potential abundance of saddle points in the parameter space might explain the lack of success of second order methods in replacing gradient descent for optimizing neural networks. In your solution, make sure to discuss the behavior of gradient descent around saddle points.

5. (Stochastic Gradient Descent)

   a) Describe a sufficient condition on the learning rate for convergence of the stochastic gradient descent algorithm. What is a typical schedule for satisfying this condition in practice?

   b) What is the excess error? What is the order of decay of the excess error for stochastic gradient descent for convex and strictly convex objective functions? Provide an argument for why it is not worthwhile to pursue an optimization algorithm that has faster convergence than that of stochastic gradient descent.

6. (Momentum)

   a) Describe, with pseudocode, the stochastic gradient descent with momentum algorithm.

   b) Recall that there is a parameter $\alpha \in [0, 1)$ that controls how quickly the contributions of previous gradients to the current gradient step exponentially decay, and $\epsilon > 0$ is the learning rate. What is the effect of α on the behavior of the algorithm relative to ϵ?

   c) Describe the stochastic gradient descent with momentum that can be interpreted as a physical viscous drag. What problem with the base algorithm does viscous drag solve? Why is it preferable to both turbulent drag and dry friction?

   d) Explain, with pseudocode, the stochastic gradient descent with Nesterov momentum algorithm. What are its advantages?

7. (Parameter Initialization)

   a) Discuss the tradeoff between model optimization and regularization in the context of network weights initialization.

   b) Describe sparse initialization. What potential problem does it have with maxout units?

8. (Adaptive Learning Rate)

   a) Describe, with pseudocode, the AdaGrad algorithm.

   b) Describe, with psuedocode, the RMSProp with Nesterov momentum algorithm.

   c) Compare the above two optimization algorithms.

   d) Describe the Adam optimization algorithm. How is it different from the above-mentioned algorithms?

9. (Conjugate Gradients)

   a) Describe the conjugate gradients method, and compare it to steepest descent.

   b) What is a problem with computing the conjugate direction? What are the popular approximation algorithms for it?

10. (Limited-Memory BFGS) Discuss the limited-memory BFGS algorithm.

Bonus

11. a) Explain why the update of weights in a layer can depend strongly on those in other layers. When is this problem severe?

    b) Explain batch normalization. Why is replacing the batch of hidden unit activations $\boldsymbol{H}$ with $\boldsymbol{\gamma}\boldsymbol{H}' + \boldsymbol{\beta}$ rather than simply the normalized $\boldsymbol{H}'$ useful?

12. a) Describe the meta-algorithm called Polyak averaging.

    b) Provide an example of greedy supervised pretraining.