TTIC 31230 Fundamentals of Deep Learning
Problem set 2
Due Thursday 11:59 pm, January 19

- Zip all your ipynb&pdf file with name PS2_yourfullname to: ttic.dl.win.2017@gmail.com.
- Late Submission: submitting late work will be penalized 10% per day, maximum three days delay allowed, no submission allowed after that.

This problem sets involves understanding and modifying a second version of the education framework (EdF) that supports minibatching. Everyone should start by installing Anaconda It is still only about four pages of python.

Problem 1. Consider

$$\ell_{\text{generalize}}(w) = \mathrm{E}_{(x,y)\sim D}\left[\ell(w,x,y)\right]$$

$$\ell_{\text{train}}(w) = \frac{1}{N}\sum_{i=0}^{N-1}\ell(w,x_i,y_i)$$

Assuming that the training data $(x_0,y_0),\ \ldots,\ (x_{N-1},y_{N-1})$ is drawn IID, and assuming that for all w, x and y we have

$$||\nabla_w \ell(w,x,y)|| < b,$$

give an upper bound on

$$||\nabla_w \ell_{\textbf{train}}(w) \quad - \nabla_w \ell_{\text{generalize}}(w)||$$

that holds with probability at least $1-\delta$ over the draw of the training data. This can be done using the formulas on the lecture slides for SGD.

Problem 2 We have provided you with a Jupyter notebook for MNIST using ReLU nonlinearities and minibatch size of 50. You are to extend the notebook with the following experiments.

a. Let B be the minibatch size. Note that the EDF implementation uses

$$w.\text{grad} = \frac{1}{B}\sum_{i=0}^{B-1}\nabla_w \ell(w,x_i,y_i).$$

$$w.\text{value} \texttt{ -= } \eta w.\text{grad}$$

we have provided the near optimal learning rate $\eta = 0.37$ for batch size 50, and assume that optimal $\eta^*(B) = 0.0056B{+}0.0659$. Please try the following four settings: $B = 10$ and $\eta = \eta^*(B)$, $B = 10$ and $\eta = 0.37$, $B = 100$ and $\eta = \eta^*(B)$, $B = 100$ and $\eta = 0.37$ and compare the training loss in four different settings.

b. In a Jupyter cell redefine SGD to use momentum. You should use the momentum equations.

$$w.\text{momentum} = \mu\ w.\text{momentum} + (1 - \mu)\ w.\text{grad}$$

and

$$w.\text{value} \texttt{ -= } \eta w.\text{momentum}.$$

where μ is the momentum parameter. For batch size $B = 50$, we have provided the near optimal $\eta = 0.37$, and we assume that optimal $\eta^*(B) = 0.0056B + 0.0659$ still holds. Please keep μ fixed in all cases and repeat the problem **a**.

c. Redefine SGD to use Adam, please keep $\beta_1 = 0.9$, $\beta_2 = 0.999$ fixed in all cases. We have provided the near optimal learning rate $\eta = 0.0015$ for batch size 50. Please tune η a little bit when $B = 10$ and $B = 100$, and compare the training loss with the case using fixed $\eta = 0.0015$. Describe what you have observed.

d.(extra credit) Explain you observation in problem **a** $\sim$ **c**, your explanation doesn't need to be very concrete and mathematical mature, just say what you have in your mind that might explain the observation.