Homework 3 for Math 173B - Winter 2025

Read me:
For this assignment, you will need to download the MNIST data set. You may find it from the original source here:

http://yann.lecun.com/exdb/mnist/

or in convenient .csv file format here:

https://pjreddie.com/projects/mnist-in-csv/

In the last assignment you considered the setting where you have data $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, $i = 1..., N$. That is, each $x_i \in \mathbb{R}^d$ is associated with a class label y_i where y_i is either 1 or -1. You assumed the model $y_i = \text{sign}(w^T x_i + z_i)$ where z_i are independent random variables drawn from a certain distribution, and you derived the cost function

$$F(w) = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + e^{-w^T x_i y_i}\right). \tag{1}$$

whose minimizer is the best w that fits your data. You also found an SGD algorithm for minimizing this function.

In this assignment, you will apply what you learned in assignment 2, on the MNIST data-set consisting of images of handwritten digits. The goal, once you optimize for w, is to classify new images $x \in \mathbb{R}^d$, using the function $y = \text{sign}(w^T x)$, which in some sense is the best you can hope to do given your model.

Questions:

(0) You must submit all your computer codes as part of this assignment. In particular, for each question, your code must be presented as part of your answer.

(1) For this question, use the first **2000 training data points for each of the digits** 1 **and** 2, to form the pairs $(x_i, y_i) \in \mathbb{R}^{784} \times \{-1, 1\}, i = 1, ..., 1000$. Assign the label $y_i = 1$ to the 1 digits, and the label $y_i = -1$ to the 2 digits.

   (a) Implement and run an SGD algorithm, with step-size $\alpha = 10^{-5}$, to optimize the function (1) associated with this setup. You should
   - run your algorithm for at least T=2000 iterations (but more iterations if you can),
   - provide a plot showing the value of $\log(F(w))$ at each iteration. Note that you are plotting the log of the function, to enhance visibility.

   (b) Comment on the resulting plot. In particular, address the following:
   - Does the value of $F(w)$ decrease with every iteration?
   - Does your algorithm seem to converge to a fixed w^*?
   - Explain whether your answers to these questions are consistent with the theory we discussed in class (and in the notes). **You should refer to specific results as part of your explanation.**

* Include any other observations you find interesting.

(c) Now, use the w you found from part (a) to classify the first 500 *test* data points associated to each of the 1 and 2 handwritten digits. Recall that you need to use the function $y = \text{sign}(w^T x)$ to classify. What was the classification error rate associated with the two digits on the test data (this should be a number between 0 and 1)? What was it on the training data?

(2) Repeat question (1) with SGD with a decreasing step-size $\alpha_t = 10^{-4} \times \sqrt{\frac{1}{t+1}}$.

(3) Repeat question (1) with *gradient descent* with a decreasing step-size $\alpha_t = 10^{-4} \times \sqrt{\frac{1}{t+1}}$.

(4) Comment on

(a) the difference between the computational complexity of GD vs SGD.

(b) the optimization results for GD and those for the two flavors of SGD you implemented. In particular, compare the function values $F(w)$ at the final iteration of each of the algorithms. Which got closer to the global minimum?

(c) the test classification accuracy for GD and those for the two flavors of SGD you implemented.