HW I

EE 599: Mathematics of Data

University of Southern California

Assigned on: September 5, 2018 Due date: beginning of class on September 19, 2018

1- **Non-convex optimization.** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function that is not necessarily convex. Assume that $\boldsymbol{x}^*$ is a global minima of f and that for all $\boldsymbol{x} \in \mathbb{R}^n$ obeying $\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\ell_2} \leq R$. Also assume we have

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{1}{\alpha} \|\boldsymbol{x} - \boldsymbol{x}^*\|_{\ell_2}^2 + \frac{1}{\beta} \|\nabla f(\boldsymbol{x})\|_{\ell_2}^2 ,$$

for some $\alpha > 0$. Furthermore, suppose $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\ell_2} \leq R$, and assume $0 < \mu < 2/\beta$. Consider the following gradient descent update

$$\boldsymbol{x}_{\tau+1} = \boldsymbol{x}_\tau - \mu \nabla f(\boldsymbol{x}_\tau).$$

Prove that for all τ we have

$$\|\boldsymbol{x}_\tau - \boldsymbol{x}^*\|_{\ell_2}^2 \leq \left(1 - \frac{2\mu}{\alpha}\right)^\tau \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\ell_2}^2 .$$

Hint: If you really think you need a hint use the above assumed inequality in the very difficult! chain of equalities below

$$\|\boldsymbol{x}_{\tau+1} - \boldsymbol{x}^*\|_{\ell_2}^2 = \|\boldsymbol{x}_\tau - \mu \nabla f(\boldsymbol{x}_\tau) - \boldsymbol{x}^*\|_{\ell_2}^2 = \|\boldsymbol{x}_\tau - \boldsymbol{x}^*\|_{\ell_2}^2 - 2\mu \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \rangle + \mu^2 \|\nabla f(\boldsymbol{x}_t)\|_{\ell_2}^2$$

2- **Convergence to stationary points and the PL-inequality.** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L-smooth function (convex or nonconvex). Assume we are running gradient descent iterations of the form

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \mu \nabla f(\boldsymbol{x}_t).$$

(a) Show that if $\mu \leq 1/L$ and the function is bounded from below, then the norm of the gradient converges to zero i.e.

$$\lim_{t \rightarrow +\infty} \|\nabla f(\boldsymbol{x}_t)\|_{\ell_2} \rightarrow 0.$$

Hint: Start by using the famous (or infamous) inequality from class:

$$f(\boldsymbol{x} - \mu \nabla f(\boldsymbol{x})) \leq f(\boldsymbol{x}) - \mu \left(1 - \frac{\mu L}{2}\right) \|\nabla f(\boldsymbol{x})\|_{\ell_2}^2 .$$

(b) Under the same assumption of part (a) does $\boldsymbol{x}_t$ converges to a fixed point?

(b) Now assume the function in addition to being L-smooth also obeys the so called PL-inequality. That is for some $\gamma > 0$ we have

$$\|\nabla f(\boldsymbol{x})\|_{\ell_2}^2 \geq \gamma \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\right).$$

Here, $\boldsymbol{x}^*$ is a global optima (assume it exists). Show that if $\mu \leq 1/L$ then the following convergence guarantee holds

$$\left(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)\right) \leq \left(1 - \frac{\mu\gamma}{2}\right)\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\right).$$

3- **Logistic regression with momentum.** In this problem we will try to predict whether a patient with breast cancer has the malignant or benign variety. We are going to use logistic regression to do this. In logistic regression we have a data set of pairs $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \in \mathbb{R}^n$ represents the features and $y_i \in \{0, 1\}$ represents the output. For example, in our data $\boldsymbol{x}_i$ represent features extracted from a digitized image of a fine needle aspirate (FNA) of a breast mass. The output y_i represents the type of breast mass, 1 for malignant and 0 for benign. Using our training data we will try to fit a logistic model to our data of the form

$$(\hat{\boldsymbol{w}}, \hat{b}) = \underset{(\boldsymbol{w},b)}{\arg\min} \quad f(\boldsymbol{w}, b) := \sum_{i=1}^{N} -y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) + \log\left(1 + \exp\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_{\ell_2}^2. \quad (1)$$

Given a new image we again extract the features $\boldsymbol{x}$. We then use our trained model to determine the probability that the mass is malignant via

$$p = \frac{1}{1 + e^{-\left(\hat{\boldsymbol{w}}^T\boldsymbol{x} + \hat{b}\right)}}.$$

Our final prediction for the output i.e. malignant or benign will be based on

$$y = \begin{cases} 1 \text{ (malignant)} & \text{if } p \geq \frac{1}{2} \\ 0 \text{ (benign)} & \text{if } p < \frac{1}{2} \end{cases}$$

The data is available in a comma separated file called wdbc.data available in blackboard. Each line corresponds to a different patient and looks like

$$842302, 1, 17.99, 10.38, 122.8, 1001, 0.1184, \ldots.$$

The first entry is the patient i.d. and is irrelevant to us. The second entry is the output, 1 for malignant and 0 for benign and the rest are the extracted features. There is a total of 569 patients in this dataset each has 30 features. We will use 500 of these patients as our training data and other 69 patients as our test data set. We will train the model i.e. solve (1) using the training data set ($N = 500$) and then report our predictions on the test data set. This selection of train/test is at random. That is we pick 69 patients at random from the total of 569 patients and put them in the test category and put the remaining 500 in the training category. In all of our experiments we will perform 100 random partitions of the data for train/test and then report the final result as the average over these trials.

(a) First read in the data from wdbc.data, remove the patient i.d. and separate the features from the outputs. For matlab users you may find the csvread function useful.

(b) Before we proceed any further we must normalize our data. This is the first thing to do when dealing with real data. Often the right form of normalization is to make sure our data set is zero mean and our features have the same Euclidean norm. More specifically, calculate the mean vector of patient features across all of the data set. That is,

$$\bar{\boldsymbol{x}} = \frac{1}{569}\sum_{i=1}^{569} \boldsymbol{x}_i.$$

Now subtract the mean from each of the features. That is, set $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \bar{\boldsymbol{x}}$. Then normalize the data via $\boldsymbol{x}_i \leftarrow \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_{\ell_2}}$.

(c) Now that we have normalized data, partition it into train/test sets at random 100 times as discussed earlier. In each trial learn the model by solving (1) with $\lambda = 0.01$. To due this run gradient descent for $T = 500$ iterations and then use the trained model to make predictions on the test data and calculate the average error (average number of miss-classified patients on the test data) for each trial. Report the average over the 100 trials. The value of the step size you use does not matter too much. However, make sure that the algorithm has converged.

(d) Perform the experiment with the same step size you used before but now report the number of iterations it takes to get to an accuracy of 10^{-6} calculated via the first iteration t when the following inequality holds

$$\|\nabla f(\boldsymbol{w}_t, b_t)\|_{\ell_2}^2 \le 10^{-6}\left(1 + |f(\boldsymbol{w}_t, b_t)|\right). \tag{2}$$

The number you should report is the average of this number over the 100 trials.

(e) Perform the experiment of part (d) but now add a momentum term (1) using the heavy ball method and (2) using Nesterov's accelerated scheme. In both cases keep the same step size as part (d) but fine tune the momentum parameter to get the smallest number of iterations for convergence based on the stopping criteria (2) (again averaged over the 100 trials). Draw the convergence of the three algorithms gradient descent, heavy ball, Nesterov's accelerated scheme for one trial. That is, draw the ratio

$$\frac{\|\nabla f(\boldsymbol{w}_t, b_t)\|_{\ell_2}^2}{(1 + |f(\boldsymbol{w}_t, b_t)|)},$$

as a function of the iteration number $t = 1, 2, \ldots, 500$. Which algorithm would you use and why?