Homework 7 for Math 173A - Fall 2024

1. Suppose a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with $L = 4$ and satisfies the PL-property with parameter $\mu = 2$, i.e., $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f(x^*))$, where x^* is the global minimizer. Consider the gradient descent method for minimizing f. Suppose $x^{(0)}$ is the initialization such that $f(x^{(0)}) - f(x^*) \leq 5$. Determine the step size μ and the number of steps needed to satisfy

$$\left\| f\left(x^{(t)}\right) - f\left(x^*\right) \right\| \leq 10^{-4}.$$

2. Consider the following set in $\mathbb{R}^n$ for an integer $s > 0$:

$$B = \{x \in \mathbb{R}^n \mid x_i \geq 0, \text{ for } i = 1, \ldots, n \text{ and } x \text{ has at most } s \text{ nonzeros.}\}.$$

   (a) Find an expression for the orthogonal projection of a point $x \in \mathbb{R}^n$ onto B (No need for justification).

   (b) For the function

$$f(x) = \frac{1}{2}\|Ax - b\|^2,$$

   Write a projected gradient descent algorithm to solve

$$\min_{x \in \Omega} f(x)$$

   for $\Omega = B$, with B from part (a). You need to specify the gradient formula and the projection formula. You do not need to specify the step size for this problem.

   (c) Consider the function in (b) and suppose $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, and $s = 1$ for the set B in (a). Does the projected gradient method converge to the global minimizer for any initialization $x^{(0)}$ if the stepsize size $\mu \leq \frac{1}{8}$. Justify your answer.

3. **Coding Question:** Consider the optimization problem:

$$f(x) = \frac{1}{2}\|Ax - b\|^2, \tag{1}$$

where $A \in \mathbb{R}^{20 \times 50}$ and $b \in \mathbb{R}^{20}$ are from the dataset `HW7Q3.csv`. The file `HW7Q3.csv` contains the data A and b. The first 50 columns form the matrix A and the last column is the vector b. The vector b is generated by setting $b = Ax^*$ for a vector $x^* \in \mathbb{R}^{20}$ that has 2 nonzeros. Note the linear system $Ax = b$ is underdetermined and has a lot of solutions. Write a projected gradient method for the following optimization problem to find the x^*:

$$\text{minimize } f(x) \quad \text{s.t.} \quad x \text{ has at most 2 nonzeros.}$$

You can experiment the stepsize to make sure $f(x^{(t)})$ converge to 0. You need to submit the code, the plot of $f(x^{(t)}) - f(x^*) = f(x^{(t)})$, and the indices and values of the nonzero entries of x^* you found.

4. **Coding Question:** We will implement the SVM algorithm with gradient descent to classify two gaussians in 2D. The dataset is given in `HW7Q4.csv`.

   (a) In `HW7Q4.csv`, the first 100 rows are the data for cluster 1: $(x_i, y_i) \in \mathbb{R}^2 \times \mathbb{R}$, $i = 1, \ldots, 100$, with $y_i = 1$ always. The next 100 rows are the data for cluster 2: $(x_i, y_i) \in \mathbb{R}^2 \times \mathbb{R}$, $i = 101, \ldots, 200$, with $y_i = -1$ always. Create and turn in a scatter plot of the feature vectors, i.e., the x_is, colored by the label, i.e., y_is (blue for 1 and red for -1).

   (b) Create a function for the gradient of the loss

   $$L(w) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \max(0, 1 - y_i\langle x_i, w\rangle)$$

   $$\nabla L(w) = w + \sum_{i=1}^{n} -y_i x_i \cdot \mathbb{1}_{1-y_i\langle x_i, w\rangle > 0},$$

   where $\mathbb{1}_{1-y_i\langle x_i,w\rangle>0} = \begin{cases} 1 & \text{if } 1 - y_i\langle x_i, w\rangle > 0 \\ 0 & \text{else} \end{cases}$. Also, here $n = 200$. To compute the gradient, you'll have to compute an indicator of whether $1 - y_i\langle x_i, w\rangle$ is positive or negative at every point, and sum up the contribution of this term for all points where it's positive.

   (c) Setting the step size $\mu = 10^{-4}$ and starting at $w^{(0)} = (-1, 1)$, run 1000 iterations of gradient descent. You will create two plots.

      i. Plot the classification error (averaged over all the points) as a function of the iterations. The classification of x_i is determined by $\text{sign}(\langle x_i, w\rangle)$.

      ii. Plot the margin $\frac{2}{\|w\|}$ as a function of the iterations. This shows how much of a gap you have between the classes you've learned.

   (d) Create another scatter plot of your data, but this time color the points by the function $f(x_i) = 1 - y_i \cdot \langle x_i, w\rangle$. The numbers closest to 0 (positive numbers or largest negative numbers) will show you which points were "most important" in determining the classification.

Note: Here we're only defining a subspace classifier (i.e. the classifier goes through the origin). This is fine for our problem as the gaussians are on opposite sides of the origin. If you want to create an intercept term, simply append a vector of all 1's as a column of your data, and now your weight vector will be of dimension 3 instead of 2. This is done the same way as when running least squares regression.