Assignment 2

CS 431, Optimization: Theory and Algorithms

To be submitted on or before: March 5, 2022

1. Find all critical points of the following functions. Further classify them into local minima, maxima and saddle points. (2 marks)
   (a) $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$
   (b) $f(x_1, x_2) = \frac{1}{2}x_1^2 + x_1 x_2 - \frac{3}{2}x_2^2 + 2x_1 + 5x_2 + \frac{x_2^3}{3}$
   Optional: Verify your answers by visualizing the functions.

2. Find the range of values of α for which the following function is convex. Also find the range of values for which it is concave. When is it both convex and concave? (3 marks)

$$f(x) = 2x_1x_3 + 4x_2x_3 - x_1^2 - 2x_2^2 - 3x_3^2 - 2\alpha x_1 x_2$$

3. State true or false with reason. (3 marks)
   (a) The difference of two convex functions is convex.
   (b) Recall that a Bernoulli random variable takes a value 1 with probability p and 0 with probability $1 - p$ where $0 < p < 1$. The entropy function for this random variable is a function $H(p) = -(p \log p + (1 - p)\log(1 - p))$. Verify whether the statement '$H(p)$ is a convex function in p' is true or false, with reason.

4. You are given the following points $(x_i, y_i) \in \mathbb{R}^2 : \{(0, 0), (1, 3), (2, 7), (3, -1), (4, 0), (5, 5), (6, 10)\}$. You want to find the best cubic polynomial that fits these points, $p(x) = c_3x^3 + c_2x^2 + c_1x + c_0$. Recognize that this can be performed using a regularized notion of least squares - this is done by treating $\{c_0, c_1, c_2, c_3\}$ as decision variables in the following optimization problem for appropriate values of the matrix A.

$$\min_{c_0,c_1,c_2,c_3} ||A\mathbf{c} - y||^2 + \lambda||\mathbf{c}||^2$$

   where λ is a fixed parameter. (12 marks)
   (a) Write down the matrix $\mathbf{A}$.
   (b) Derive the value of the optimal solution using necessary and sufficient conditions.
   (c) Solve the problem using gradient descent with exact line search and report the values of $\mathbf{c}$ obtained for different values of λ say, $\lambda = 0, 1, 10$ and 1000. How many iterations did it take to converge in each case?
   (d) Solve the problem using Newton's method. Use $\lambda = 0, 1, 10$ and 1000. How many iterations did it take for each?
   (e) How do the values of the optimal $\mathbf{c}$ compare across different settings of λ?
   (f) Plot the polynomial $p(x)$ obtained for different values of λ. Also plot the given points (x_i, y_i) in the same plot. So you will have one plot with four lines one for each $\lambda \in \{0, 1, 10, 1000\}$. What conclusions can you make here on the four polynomials?

   PS: For all iterative optimization methods use your favourite stopping criteria. Report what you used.

5. In this question we will work on logistic regression, a commonly used machine learning technique for classification. You will implement the iterative methods learnt in class to build a classifier. While here is a brief summary, please read up on logistic regression from [Boyd and Vandenberghe, Convex Optimization, Chapter 7] to understand more.

   Suppose there are n observations $(\mathbf{x}_i, y_i)$ where the feature vector $\mathbf{x}_i \in \mathbb{R}^m$ and the label $y_i \in \{-1, 1\}$. Logistic regression models observations using a co-efficient vector $\mathbf{w} \in \mathbb{R}^m$, in the following manner,

$$y_i = \begin{cases} 1 & \text{with probability } \sigma(\mathbf{w}^\top \mathbf{x}_i) \\ -1 & \text{with probability } 1 - \sigma(\mathbf{w}^\top \mathbf{x}_i) \end{cases} \tag{1}$$

where $\sigma(a) = \frac{1}{1+e^{-a}}$. The conditional probability of assigning a label y_i to the instance $\mathbf{x}_i$ given the value of $\mathbf{w}$ is $\mathbb{P}(y_i|\mathbf{w}, \mathbf{x}_i) = \sigma(y_i \mathbf{w}^\top \mathbf{x}_i)$. The overall negative log likelihood of the given data collection is

$$l(\mathbf{w}) = \sum_{i=1}^{n} -\log(\sigma(y_i \mathbf{w}^\top \mathbf{x}_i)).$$

You need to find $\mathbf{w}$ that minimizes $l(\mathbf{w})$. (In general the $\mathbf{w}$ could also contain an intercept term, so that $\mathbf{w} = [w_0, w_1, \ldots, w_m] \in \mathbb{R}^{m+1}$ and every data point $\mathbf{x}$ is augmented with an additional component $x_0 = 1$).

A two dimensional dataset containing 2000 sample points is given. Each row corresponds to a sample where the first two entries in a row give the features (x_1, x_2) and the last entry gives the class label. Randomly split this dataset into a training set containing $n = 1400$ points and a test set containing the remaining 600 points. (20 marks)

(a) Is $l(\mathbf{w})$ a convex function over $\mathbf{w} \in \mathbb{R}^m$? Prove or disprove.

(b) Write the expression for the gradient $\nabla l(\mathbf{w})$ and Hessian $\nabla^2 l(\mathbf{w})$.

(c) Apply steepest descent to compute the best value of $\mathbf{w}$ that minimizes $l(\mathbf{w})$ over the training set. How many iterations did it take? Report the step size and stopping criteria you used.

(d) Plot how $l(\mathbf{w})$ changes in each iteration: X-axis should show iterations and Y axis should show $l(\mathbf{w})$.

(e) Provide a scatter plot of the training set. Show points having true class +1 and points in true class -1 with different colours/shapes. In this plot show the final decision boundary learnt $(\mathbf{w}^*)^\top \mathbf{x} = 0$ and also the initial decision boundary $\mathbf{w}_0^\top \mathbf{x} = 0$ using the starting point $\mathbf{w}_0$. Examples of such plots appear in Pattern Recognition and Machine Learning, Christopher Bishop, Chapter 4 (eg. Fig 4.4). Label the plots appropriately.

(f) With the learnt $\mathbf{w}^*$, perform a prediction on the test set and compute the mis-classification error,

$$\text{Misclassificiation error} \;=\; \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbf{1}(\hat{y}_i \neq y_i),$$

where $\hat{y}_i$ is the prediction you make using Eq (1), and y_i is the true label provided in the dataset. $\mathbf{1}(\hat{y}_i \neq y_i) = 1$ when $\hat{y}_i \neq y_i$ and 0 otherwise. $N_{test} = 4000$, the number of points in the test data set.

(g) Try five different starting points and run steepest descent. Report the starting points you used. Do you reach the same optimal point? Is the final value of the function $l(.)$ same always?

(h) Try various settings of step size and stopping criteria learnt in class. Which step size and stopping criteria would you recommend?

(i) *(Optional)* Derive the update rules for Newton's method. How many iterations does Newton's method take to converge?