ECE 490 (Introduction to Optimization) – Homework 3

Due: 11:59pm, March 8th

Problem 1. Consider Newton's method with stepsize α, i.e.

$$x_{k+1} = x_k - \alpha(\nabla^2 f(x_k))^{-1}\nabla f(x_k),\ \alpha > 0.$$

(a) (15 points) Suppose we apply this method to the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^4$. Identify the range of α for which the method converges. Show that for this range of α, the convergence is "linear".

(b) (15 points) Suppose we choose $\alpha = 1$ and apply this method to the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \log(e^x + e^{-x})$. Note that f is convex with a unique minimum at $x^* = 0$. Show that the Newton's method for this function iterates as

$$x_{k+1} = x_k - \frac{e^{4x_k} - 1}{4e^{2x_k}}.$$

Run 5 steps of the above iteration with the following initializations: $x_0 = 1$ and $x_0 = 1.1$. You may use your favorite programming environment (Matlab, Python, etc). Report your iterates for both cases. Does Newton's method converge?

(c) (10 points) Consider the ridge regression problem $\min_{x\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \left\{(a_i^T x - b_i)^2 + \frac{\lambda}{2}||x||_2^2\right\}$, where $\lambda > 0$ and (a_i, b_i) (for $i = 1, 2, \cdots, n$) are given. If we run Newton's method with $\alpha = 1$ on this problem, what happens? Does it converge? If so, how many steps are needed to get an accurate solution?

Problem 2. We have studied the back-propagation algorithm for computing the gradient of the empirical loss corresponding to each data-point with respect to the weights of the neural network separately, with the understanding that the gradient of the total empirical loss J with respect to the weights is simply the sum of the gradients of the loss corresponding to each data-point. In this problem, you will develop a back-propagation algorithm for computing the gradient of J directly.

(a) (20 points) Derive the back-propagation algorithm for directly computing the gradients:

$$\frac{\partial J}{\partial W_{i,j}^{(m)}} \text{ and } \frac{\partial J}{\partial b_i^{(m)}} \quad \text{for all } i, j, \text{ and } m = 1, \cdots, M.$$

Hint: Using the chain rule:

$$\frac{\partial J}{\partial W_{i,j}^{(M)}} = \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(M)}[n]} \frac{\partial y_i^{(M)}[n]}{\partial W_{i,j}^{(M)}}.$$

(b) (10 points) Is there any computational advantage (or disadvantage) of running the back-propagation algorithm directly on J as opposed to running it on loss corresponding to the individual data-points? What is the advantage of the stochastic gradient descent (SGD) method?

Problem 3. Either prove the following statements or provide a counterexample.

(a) (5 points) The set $\{x \in \mathbb{R}^n : Ax = b\}$ is convex for all matrices $A \in \mathbb{R}^{k\times n}$ and $b \in \mathbb{R}^k$.

(b) (5 points) For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, the set $\{x \in \mathbb{R}^n : f(x) = r\}$ is convex for $r \in \mathbb{R}$.

(c) (10 points) For any positive semidefinite matrix $Q \in \mathbb{R}^{n\times n}$, the set $\{x \in \mathbb{R}^n : x^T Qx \le 1\}$ is convex.

(d) (10 points) Any μ-strongly convex function has to satisfy the following P-L inequality:

$$f(x) - f(x^*) \le \frac{1}{2\mu}||\nabla f(x)||^2,$$

where x^* is the global min of f.