University of Illinois **Spring 2023**

ECE 490: Problem Set 3

Due: Tuesday February 28 through Gradescope by 11 AM
Reading: Lecture Notes 8-11; Secs 1.4, 3.1, 3.3, 4.1 and 4.2 of text.

1. **[Linear Convergence in Newton's Method]**
   Consider Newton's method with stepsize α, i.e.,

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

   Suppose we apply this method to the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = |x|^3$.

   (a) Identify the range of α for which the method converges.

   (b) Show that for this range of α, the convergence is "linear".

   (c) Explain why the fact that the covergence is not "superlinear" for $\alpha = 1$ does not contradict the convergence result for Newton's method given in Lecture 8.

2. **[Sensitivity to Initialization in Newton's Method]**
   Consider Newton's method with stepsize $\alpha = 1$, i.e.,

   $$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k).$$

   Suppose we apply this method to the function $f : \mathbb{R} \to \mathbb{R}$ given by

   $$f(x) = \sqrt{x^2 + 1.1}.$$

   (a) Find the stationary points of f and derive the recursion of Newton's method in closed form.

   (b) Run 5 steps of the iteration of part (a) 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, and comment on your results.

3. **[Backpropagation Algorithm]**
   In lecture 9, we 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 NN 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) Derive the back-propagation algorithm for directly computing the gradients:

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

   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) 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?

4. **[Moore-Penrose Psuedo-inverse]**
   Consider the problem of minimizing the function $f(x) = \|x\|^2$ subject to the constraint $x \in \mathcal{S}$, where

   $$\mathcal{S} = \{x \in \mathbb{R}^n : \; Ax = b\}.$$

   Here A is an $m \times n$ matrix, with $m \leq n$ and $AA^\top$ being invertible. Note that $\mathcal{S}$ is closed and convex (make sure you know how to prove this), and f is (strictly) convex on $\mathcal{H}$.

   Show that minimizer x^* is given by
   $$x^* = A^\top (AA^\top)^{-1} b.$$

5. **[Projection]**
   Consider the set
   $$\mathcal{S} = \{y \in \mathbb{R}^n : \; Ay = b\}$$
   where A is an $m \times n$ matrix with $m \leq n$, and $AA^\top$ is invertible. Show that the projection of $z \in \mathbb{R}^n$ on $\mathcal{S}$ is given by
   $$y^* = z - A^\top (AA^\top)^{-1}(Az - b).$$

   Hint: You need to show that $y^* \in \mathcal{S}$, and that it satisfies the necessary and sufficient condition for being the projection (see Lec 10 notes).

6. **[Constrained Optimization]**
   Consider the function $f : \mathbb{R}^n_+ \to \mathbb{R}$ defined by:

   $$f(x) = \sum_{i=1}^{n} x_i \ln x_i.$$

   The set $\mathbb{R}^n_+$ consists of all vectors with non-negative components. Define $0 \log(0)$ to be 0. Suppose we wish to solve the following constrained optimization problem:

   $$\min_{x \in \mathcal{S}} f(x)$$

   where the constraint set is given by:

   $$\mathcal{S} = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, \text{ and } x_i \geq 0, i = 1, 2, \ldots, n \right\}.$$

   (a) Establish that f is a strictly convex function on $\mathcal{S}$.

   (b) Use the first-order necessary and sufficient condition for optimality to establish that

   $$x^* = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right)$$

   is the *unique* minimizer of f on $\mathcal{S}$.