IE510 Applied Nonlinear Programming Homework 3

Instructor name: Ruoyu Sun Student name: ____________________

Date assigned: Week 10 Monday 03/19/2018
Date due: Week 12 Tuesday 04/03/2018, at 3pm

Remark: The work your are submitting for this homework assignment must be your own, and suspiciously similar homework submitted by multiple individuals may be reported to the University for investigation.

1 Reading

- Reading: Textbook Section 1.5.2 "Incremental Gradient Method"
- Reading: Textbook Section 2.7 "Block Coordinate Descent Methods" (note that the theorem in 1999 version needs to be modified; see the correction in the Box file)
- If interested in more details about incremental gradient methods and SGD, you can read references listed in lecture slides

2 Problems

1. (60 points + 5 bonus points) Consider a non-singular matrix $A = (A_1, A_2, \ldots, A_n)$, where $A_i \in \mathbb{R}^{d\times 1}$ and $||A_i|| = 1$, for $\forall i$. We want to solve $\min_x ||Ax - b||^2$, s.t. $x \in \mathbb{R}^n$. Assume $b = 0$ for simplicity.

   (a) (25 points) Let $n = 100$. Consider the setting $A_{ij} \sim \text{Unif}[0, 1]$, where Unif[0,1] represents uniform distribution on the interval $[0, 1]$. After generating A, normalize A_i, s.t. $||A_i|| = 1$, $\forall i$(i.e. define $\widetilde{A_{ij}} = \frac{A_{ij}}{||A_i||}$, $\forall i, j$, and replace A by $\widetilde{A}$).
   Use random initial point $x^0 = (x_1^0, \ldots, x_n^0)$, where $x_i^0 \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, n$. Here $\mathcal{N}(0, 1)$ represents standard Gaussian distribution. Compare two algorithms randomized coordinate descent (R-CD) and gradient descent (GD) with stepsize $\frac{1}{L}$ by plotting function values v.s. epochs.
   Remark: One epoch of R-CD consists of n iterations.

   (b) (5 points) I claimed in class that R-CD is roughly $\frac{\lambda_{max}(A^TA)}{\lambda_{avg}(A^TA)}$ times faster than GD, where $\lambda_{avg}(A^TA)$ is the average eigenvalue of A^TA. Do your figures in (a) support this claim?

   (c) (30 points) Under the same setting as (a), compare SGD with various constant stepsizes, SGD with diminishing stepsizes and GD with stepsize $\frac{1}{L}$. Report and discuss your findings.

(d) (5 bonus points) Generate Q as $Q_{ii} = 1, \forall i$; $Q_{ij} = c, \forall i \neq j$. Here $0 \leq c \leq 1$ is a constant. Consider problem

$$\min_{x} \quad x^T Q x, \quad s.t. \quad x \in \mathbb{R}^n.$$

Compare 4 algorithms: cyclic CD(C-CD), randomized CD(R-CD), randomly permutated CD(RP-CD), GD with stepsize $\frac{1}{L}$. Run experiments for different n and c. What do you find? Can you make some general conjectures that explain your findings?

2. (40 points) Consider the following problem:

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}||M - xx^T||_F^2. \tag{$*$}$$

(a) (5 points) Describe how to use gradient descent to solve ($*$).

(b) (10 points) Describe how to use coordinate gradient descent to solve ($*$).

(c) (15 points) Describe how to use coordinate descent to solve ($*$).
Note: You can assume there exists a 3-rd order equation solver that returns all three complex roots of equation $az^3 + bz^2 + cz + d = 0$.

(d) (10 points) Describe how to use stochastic gradient method to solve ($*$). Please provide at least two versions of SGD.
Remark: You can use any update order in all algorithms.