IE 510 HW#6

TIANQI WU

1.

- Min f(x) s.t. $x \in X$
  - Unconstrained: Min f(x)
    - optimal condition
      - 1st order: $\nabla f(x^*) = 0$
      - 2nd order: $\nabla^2 f(x^*) \succeq 0$
    - algorithm
      - convergence rate
        - slow: GD
        - fast: HB, Nesterov's
      - decomposition
        - CD
        - SGD
      - Newton
        - BFGS
        - BB
      - better stepsize
  - Constrained
    - not explicit: Min f(x) st. $x \in X$, X convex
      - optimal condition: $\langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in X$
      - gradient projection
      - fixed point
      - ALM
      - ADMM
    - explicit
      - equality
        - 1st order
        - 2nd order
        - $\nabla f(x) + \lambda \nabla h(x) = 0$
          - feasible direction proof
          - Penalty proof
      - inequality
        - 1st order
        - 2nd order
        - $\nabla f(x) + \mu \nabla g(x) = 0$, $\mu \geq 0 \perp g(x) \leq 0$
          - Penalty proof
          - transform: equality.
    - convex
      - Assume f, g convex h linear
        - 1st order $\Rightarrow$ global min.
        - $\Updownarrow$ strong duality

2. (a) $\min_x x^T Q x + 2b^T x$

$\nabla f(x) = Qx + b$. Takes time $O(n^2)$.

$\frac{e_r}{e_0} \le (1-\frac{1}{k})^r \Rightarrow r \gtrsim k \log\frac{1}{\varepsilon}$

In total, the time complexity is $O(n^2 k \log\frac{1}{\varepsilon})$.

(b) If we ignore the term $\log\frac{1}{\varepsilon}$

The time complexity becomes $O(n^2 k)$

Then, if $k \gg n$, direct method is faster.

if $k \ll n$, GD method is faster

(c) When $n = 10$, $k = \frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)} = 265.651$

$n = 20$, $k = 3111$

$n = 50$, $k = 1.421 \times 10^5$

$n = 100$, $k = 4.89 \times 10^4$

It can be seen that k is much larger than n.

(d). Direct method will be faster than GD method for the problems generted in (c)

(e). GD will go through multiple stages and $k\log(\frac{1}{\varepsilon})$ only captures the speed of last stage. Each speed has $\frac{\lambda_{max}}{\lambda_k}$ and ends at roughly λ_k. If low-accuracy is wanted, $k\log(\frac{1}{\varepsilon})$ is not an issue since convergence behavior of GD roughly depends on eigenvalues above or around the accuracy. It is bad, however, if high accuracy is wanted.

3. (a) coordinate descent and gradient descent can solve the least square problem with some convergence guarantee.

(b). $\min_{w} \|Aw - b\|^2$ s.t. $w \geq 0$, $w \in \mathbb{R}^d$, $A \in \mathbb{R}^{n \times d}$

$= \min \sum_{i=1}^{n} \|A_i w - b_i\|^2 = \sum f_i(w)$

$= \min_{w, z_1 \dots z_n} \sum_{i=1}^{n} \|A_i z_i - b_i\|^2$ s.t. $z_i = w$, $w \geq 0$ $\forall i$

Lagraigian $= \underbrace{\sum f_i(z_i) + \langle \lambda, z_i \rangle}_{F(z_i)} - \underbrace{m \sum \langle \lambda_i, X \rangle}_{G(X)}$.

$= \|AX - b\|^2 + \lambda^T (z - x) + \frac{\rho}{2} \|z - x\|^2$

$$\begin{cases} X^{r+1} = \arg\min L_\rho (X, z^r, \lambda^r) \\ z^{r+1} = \arg\min L_\rho (X^{r+1}, z, \lambda^r) \\ \lambda^{r+1} = \lambda^r + \frac{\rho}{2} (z^r - x^r) \end{cases}$$

4 (a) $f(x) = \frac{1}{2}(x_1^2 - x_2^2) - 3x_2$

s.t. $x_2 = 0$.

$L(x, \lambda) = \frac{1}{2}(x_1^2 - x_2^2) - 3x_2 + \lambda(x_2 - 0)$

$\nabla_x L(x^*, \lambda^*) = 0 \Rightarrow \begin{cases} x_1 = 0 \Rightarrow x_1 = 0 \\ -x_2 - 3 + \lambda = 0 \Rightarrow \lambda = 3. \end{cases}$

$\nabla_\lambda L(x^*, \lambda^*) = 0 \Rightarrow x_2 = 0$.

Hence, $x^* = (0, 0)$, $\lambda^* = 3$

(b) $L_c(x, \lambda) = \frac{1}{2}(x_1^2 - x_2^2) - 3x_2 + \lambda(x_2 - 0) + \frac{c}{2}(x_2 - 0)^2$

$x_1(\lambda, c) = 0$ $\quad$ $x_2(\lambda, c) = -x_2 - 3 + \lambda + cx_2 = 0$

$= \frac{3-\lambda}{c-1}$

When $k = 0, 1, 2$, $c^k = 10, 10^2, 10^3$

For quadratic penalty with $\lambda^k = 0$:

when $k=0$, $x_2 = \frac{3-\lambda^k}{c^k-1} = \frac{3-0}{10-1} = \frac{3}{9}$

$k=1$, $x_2 = \frac{3}{100-1} = \frac{3}{99}$

$k=2$, $x_2 = \frac{3}{999}$

For Multipliers with $\lambda^0 = 0$ $\quad$ $\lambda^k \to 3 = \lambda^*$

when $k=0$, $x_2 = \frac{3-\lambda^k}{c-1} = \frac{3}{c-1}$

$k=1$, $x_2 = \frac{3-\lambda^1}{c-1} = 0$

$k=2$, $x_2 = \frac{3-\lambda^2}{c-1} = 0$.

when $k \to \infty$ the quadratic penalty method will have $x_2 \to 0 = x_2^*$

And the result from the two methods will be the same.