Management Sciences Topics: Convex Optimization
Homework 3

1. Problem 1

   We may change the step size to
$$\eta_k = \frac{2}{\mu(k+2)}$$
   as when $\mu > 0$ in SSG. Now we show that

   First, by three term lemma,
$$\langle x^{k+1} - x^k, g_x^k \rangle + \frac{1}{2\eta_k}\|x^{k+1} - x^k\|_2^2 \leq \langle x - x^k, g_x^k \rangle + \frac{1}{2\eta_k}\|x - x^k\|_2^2 - \frac{1}{2\eta_k}\|x - x^{k+1}\|_2^2.$$
   By strong convexity of f w.r.t. x,
$$f(x^k, y) - f(x^*, y) \leq \langle g_x^k, x^k - x^* \rangle - \frac{\mu}{2}\|x^k - x^*\|_2^2.$$
   Then by mean inequality,
$$\begin{aligned} f(x^k, y) - f(x^*, y) &\leq \langle x^* - x^k, g_x^k \rangle + \frac{1}{2\eta_k}\|x^* - x^k\|_2^2 - \frac{1}{2\eta_k}\|x^* - x^{k+1}\|_2^2 - \frac{1}{2\eta_k}\|x^{k+1} - x^k\|_2^2 - \frac{\mu}{2}\|x^k - x^*\|_2^2 \\ &\leq \frac{1}{2}\eta_k\|g_x^k\|_2^2 + (\frac{1}{2\eta_k} - \frac{\mu}{2})\|x^k - x^*\|_2^2 - \frac{1}{2\eta_k}\|x^* - x^{k+1}\|_2^2 \\ &\leq \frac{1}{2}\eta_k M^2 + (\frac{1}{2\eta_k} - \frac{\mu}{2})\|x^k - x^*\|_2^2 - \frac{1}{2\eta_k}\|x^* - x^{k+1}\|_2^2 \\ &= \frac{1}{\mu(k+2)}M^2 + \frac{\mu k}{4}\|x^k - x^*\|_2^2 - \frac{\mu(k+2)}{4}\|x^* - x^{k+1}\|_2^2. \end{aligned}$$
   Then
$$\sum_{k=0}^{K-1}(k+1)(f(x^k, y) - f(x^*, y)) \leq \sum_{k=0}^{K-1}\frac{k+1}{\mu(k+2)}M^2 + \sum_{k=0}^{K-1}\frac{\mu k(k+1)}{4}\|x^k - x^*\|_2^2 - \sum_{k=0}^{K-1}\frac{\mu(k+1)(k+2)}{4}\|x^* - x^{k+1}\|_2^2,$$
   and
$$\frac{k(k+1)}{2}(f(\bar{x}, y) - f(x^*, y)) \leq \frac{KM^2}{\mu},$$
   so
$$f(\bar{x}, y) - f(x^*, y) \leq \frac{2M^2}{(k+1)\mu}.$$
   Similarly, we have
$$f(x, \bar{y}) - f(x, y^*) \leq \frac{2M^2}{(k+1)\mu}.$$
   Lets $y = \bar{y}$ in the first equation and $x = x^*$ in the second equation, by adding these two equations, we get
$$f(\bar{x}, \bar{y}) - f(x^*, y^*) \leq \frac{4M}{(K+1)\mu}.$$
   Then the algorithm has a convergence rate of $O(\frac{1}{k})$.

2. Problem 2. By Nesterov smoothing,

$$h_\mu(x) = \max_{\|y_g\|_2 \le 1} \tilde{h}(y) = \max_{\|y_g\|_2 \le 1} \lambda y^\top Cx - \frac{1}{2}\mu\|y\|_2^2,$$

where the index g is for all groups, and C is the group matrix.

Take partial derivative w.r.t. y,

$$\frac{\partial}{\partial y}\tilde{h}(y) = \lambda Cx - \mu y.$$

Then the maximizer is

$$y_g^* = S_2(\frac{\lambda(Cx)_g}{\mu}),$$

where S_2 is the projection operator onto the unit L_2 ball. Then

$$h_\mu(x) = \lambda y^{*\top} Cx - \frac{1}{2}\mu\|y^*\|_2^2,$$

and from Theorem 1 of https://arxiv.org/ftp/arxiv/papers/1202/1202.3708.pdf,

$$\nabla h_\mu(x) = C^\top y^*,$$

and the Lipschitz constant for ∇h_μ is $L = \frac{1}{\mu}\|C\|_2^2$.

Now we consider the problem as follows:

$$f = g + h,$$

where

$$g = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^\top x)) + \lambda y^{*\top} Cx - \frac{1}{2}\|y^*\|_2^2,$$

and

$$h = 1_{\mathcal{X}}(x).$$

Then the proximal of ηh is just the projection operator w.r.t. $\mathcal{X}$.

The code for this problem is `accelerated_proximal_gradient.m` and `overlapped_group_regularized_logistic_regression_smooth.m`. The objective value of average iterate is shown in Figure 1.

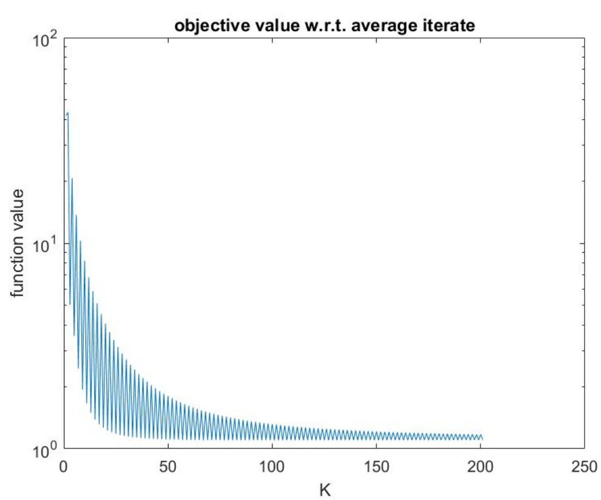

Figure 1: Objective value for problem 2.

3. Problem 3.

   The problem can be written as

$$\min_x \max_y \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-b_i a_i^\top x)) + \lambda y^\top Cx,$$

   where C is the group matrix. Then pick

$$g_x \in \partial_x f = \frac{1}{n}\sum_{i=1}^{n} \frac{-b_i a_i^\top \exp(-b_i a_i^\top x)}{1+\exp(-b_i a_i^\top x)} + \lambda C^\top y,$$

   and

$$g_y \in -\partial_y(-f) = \lambda Cx.$$

   The code for this problem is `primal_dual_subgradient.m` and `overlapped_group_regularized_logistic_regression_primal_dual.m`. The objective value of average iterate is shown in Figure 2.

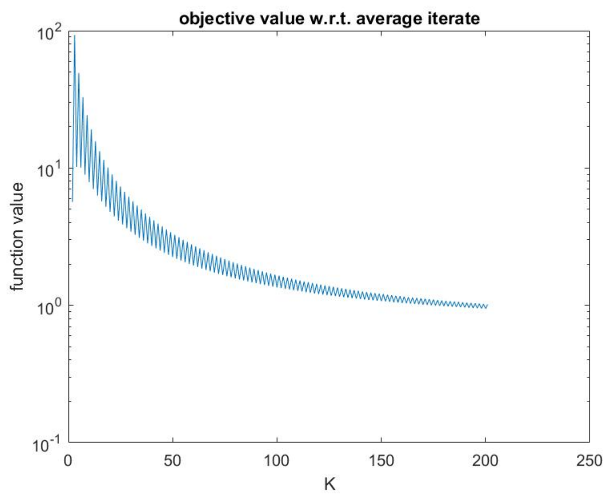

Figure 2: Objective value for problem 3.