Optimization for Machine Learning (Homework #3)

Assignment date: Oct 18
Due date: Nov 1 (noon)

Theoretical Problems (13 points)

1. (3 points) Consider $x \in \mathbb{R}_+$. Consider
$$h(x) = -\ln x$$
   - Find the Bregman divergence $D_h(x, y)$.
   - Consider function
   $$f(x) = 0.5x - \ln(x + 1).$$
   What's the smoothness and strong convexity parameters of f with respect to h?
   - Consider h-mirror descent method for solving $f(x)$. What's the formula to obtain x_t from x_{t-1} with a learning rate η_t.

2. (3 points) Consider a composite optimization problem
$$f(x) + g(x),$$
with non-convex regularizer $g(x)$ given by
$$g(x) = \lambda \sum_{j=1}^{d} \ln(1 + |x_j|),$$
for some $\lambda > 0$. Since logarithm grows slowly when $|x_j|$ increases, this regularizer leads to less bias than L_1 regularization. Assume we want to apply proximal gradient method
$$\text{prox}_{\eta g}\left[x - \eta \nabla f(x)\right]$$
to solve this problem, where the proximal optimization problem becomes
$$\text{prox}_{\eta g}(z) = \arg\min_x \left[\frac{1}{2}\|x - z\|_2^2 + \eta g(x)\right] \tag{1}$$
   - Show that when $\eta > 0$ is sufficiently small, then the proximal optimization problem (1) is strongly convex for all z. What is the range for such η?
   - Find the closed form solution for $\text{prox}_{\eta g}(z)$ when η is sufficiently small so that (1) is convex.
   - If $f(x)$ is L smooth and 2λ strongly convex. Does the proximal gradient algorithm converge? If so, how many iterations are needed to obtain an ϵ primal suboptimal solution?

3. (3 points) Consider the optimization problem

$$\sum_{i=1}^{n} f_i(A_i x) + \|A_0 x\|_2,$$

where A_i are matrices. We rewrite it as

$$\phi(x, z) = \sum_{i=1}^{n} f_i(z_i) + \|z_0\|_2 \quad \text{subject to } A_1 x - z_1 = 0, \ldots, A_n x - z_n = 0,\ A_0 x - z_0 = 0.$$

- Write down the Lagrangian function $L(x, z, \alpha)$ with multipliers $\alpha_1, \ldots, \alpha_n$ and α_0 corresponding to the $n+1$ linear constraints.
- Find the dual formulation $\phi_D(\alpha) = \min_{x,z} L(x, z, \alpha)$ in terms of f_i^*.
- If $n = 1$ and $A_1 = I$, write down the dual objective function in α_0 by eliminating α_1. Assume $f_1^*(\cdot)$ is smooth, how to get primal the primal variable x from dual α_0?

4. (4 points) Consider a symmetric positive definite matrix A, and let

$$f(x) = \frac{1}{2} x^\top A x - b^\top x, \quad g(x) = \frac{\lambda}{2}\|x\|_2^2 + \mu\|x\|_1.$$

- Find the Fenchel's dual of $f(x) + g(x)$.
- Find $\nabla f^*(\alpha)$ and $\nabla g^*(\alpha)$
- Write down a closed form formula for the dual ascent method.
- Find the smoothness of f^* and g^* and explain how to set learning rate for dual ascent.

Programming Problem (7 points)

- Download data in the mnist sub-directory (which contains class 1 (positive) versus 7 (negative)
- Use the python template "prog-template.py", and implement functions marked with '# implement'.
- (4 points) Complete codes and get the plots
- (1 points) Discuss the behaviors of proximal and dual averaging algorithms in the experiments
- (1 points) Discuss the impacts of different μ in the experiments
- (1 points) Moreover, can you reduce the degree of oscillations for GD by selecting other learning rates when the objective function is highly non-smooth?