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 $D_h(x, y)$.
  Solution. We have
  $$D_h(x, y) = \frac{x - y}{y} - \ln \frac{x}{y}$$
  □
- Consider function
  $$f(x) = 0.5x - \ln(x + 1).$$
  What's the smoothness and strong convexity parameters of f with respect to h?
  Solution. The smoothness and strong convexity parameters of f with respect to h are upper and lower bounds for
  $$\frac{D_f(x, y)}{D_h(x, y)}$$
  which can be achieved with lower and upper bounds for
  $$\frac{f''(x)}{h''(x)} = \frac{x^2}{(x + 1)^2} \in [0, 1).$$
  Therefore it is 0-strongly convex and 1-smooth. □
- 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.
  Solution.
  $$-\frac{1}{x_t} = -\frac{1}{x_{t-1}} - \eta_t(0.5 - \frac{1}{x_{t-1} + 1}).$$
  This gives
  $$x_t = \frac{x_{t-1}(x_{t-1} + 1)}{x_{t-1} + 1 + 0.5\eta_t(x_{t-1}^2 - x)}.$$
  □

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?

  Solution. The second order derivative with respect to x_j

  $$\geq 1 - \frac{\lambda\eta}{(1 + |x_j|)^2},$$

  which is non-negative (thus objective is convex) when

  $$\eta < 1/\lambda.$$

  If $\eta < 1/\lambda$, then it is easy to check the function isn't convex when $x = z$. $\square$

- Find the closed form solution for $\text{prox}_{\eta g}(z)$ when η is sufficiently small so that (1) is convex.

  Solution. Let y be the solution. For each j, if $|z_j| \leq \lambda\eta$, we know that the first order condition is satisfied at

  $$y_j = 0.$$

  Otherwise, if $z_j > \lambda\eta$, we have $y_j \geq 0$ and

  $$y_j - z_j + \frac{\lambda\eta \text{sign}(y_j)}{1 + |y_j|} = 0,$$

  which implies that

  $$y_j = \frac{z_j - 1 + \sqrt{(1 - z_j)^2 + 4(z_j - \lambda\eta)}}{2}$$

  is the unique solution. Similarly for $z_j < -\lambda\eta$.
  We can summarize all situations and obtain

  $$y_j = \text{sign}(z_j)\frac{|z_j| - 1 + \sqrt{(|z_j| - 1)^2 + 4\max(0, |z_j| - \lambda\eta)}}{2}.$$

  $\square$

- 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?

  Solution. The overall function is λ strongly convex. It has shown in the lecture that the proximal gradient method is equivalent to

  $$\tilde{f}(x) + \tilde{g}(x),$$

  where

  $$\tilde{f}(x) = f(x) - \frac{\lambda}{2}\|x\|_2^2, \qquad \tilde{g}(x) = g(x) + \frac{\lambda}{2}\|x\|_2^2.$$

  Both of these are convex. Hence the theory of proximal gradient implies linear convergence at rate of $(L/\lambda)\ln(1/\epsilon)$. $\square$

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.

  Solution. We have

  $$L(x,z,\lambda) = \sum_{i=1}^{n} f_i(z_i) + \frac{1}{2}\|z_0\|_2 + \sum_{i=0}^{n} \alpha_i^\top (A_i x - z_i).$$

  □

- Find the dual formulation $\phi_D(\alpha) = \min_{x,z} L(x,z,\alpha)$ in terms of f_i^*.

  Solution. Note that

  $$\phi_D(\alpha) = \sum_{i=1}^{n} -f_i^*(\alpha_i) \text{ subject to } \|\alpha_0\|_2 \le 1, \quad \sum_{i=1}^{n} A_i^\top \alpha_i = 0.$$

  □

- 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?

  Solution. Since

  $$\alpha_1 = -A_0 \alpha_0,$$

  we can eliminate α_1, and the dual objective becomes

  $$\phi_D(\alpha_0) = -f_1^*(-A_0^\top \alpha_0) \quad \text{subject to } \|\alpha_0\|_2 \le 1.$$

  We can get x via

  $$x = \nabla f_1^*(-A_0 \alpha_0).$$

  □

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 $f(x) + g(x)$.

  Solution. Given α. Let x be the solution to

  $$-\alpha = \nabla f(x) = Ax - b,$$

  which implies that $x = A^{-1}[-\alpha + b]$. Then

  $$\begin{aligned} f^*(-\alpha) =& -\alpha^\top x - f(x) = x^\top A x - b^\top x - \frac{1}{2} x^\top A x + b^\top x \\ =& \frac{1}{2} x^\top A x = \frac{1}{2}[-\alpha + b]^\top A^{-1}[-\alpha + b]. \end{aligned}$$

Similarly. The solution $x = [x_j]$ of $\sup_x[\alpha^\top x - g(x)]$ is

$$[x_j] = \begin{cases} (\alpha_j - \mu)/\lambda & \alpha_j > \mu \\ (\alpha_j + \mu)/\lambda & x_j < -\mu \\ 0 & \alpha_j \in [-\mu, \mu] \end{cases}$$

Therefore

$$g^*(\alpha) = \alpha^\top x - g(x) = \frac{\lambda}{2}\|x\|_2^2 = \frac{1}{2\lambda}\sum_{j=1}^{d}(|\alpha_j| - \mu)_+^2.$$

Therefore the dual is

$$-f^*(-\alpha) - g^*(\alpha) = -\frac{1}{2}[b - \alpha]^\top A^{-1}[b - \alpha] - \frac{1}{2\lambda}\sum_{j=1}^{d}(|\alpha_j| - \mu)_+^2.$$

□

- Find $\nabla f^*(\alpha)$ and $\nabla g^*(\alpha)$

  Solution. We have

  $$[\nabla g^*(\alpha)]_j = \frac{1}{\lambda}(|\alpha_j| - \mu)_+ \mathrm{sign}(\alpha_j)$$

  for $j = 1, \ldots, d$. We have

  $$\nabla f^*(\alpha) = A^{-1}(\alpha + b).$$

  □

- Write down a closed form formula for the dual ascent method.

  Solution. We have the update rule of

  $$\alpha_t = \alpha_{t-1} - \eta_t \left[A^{-1}(\alpha_{t-1} - b) + \frac{1}{2\lambda}(|\alpha_j| - \mu)_+ \mathrm{sign}(\alpha_j)\right]$$

  □

- Find the smoothness of f^* and g^* and explain how to set learning rate for dual ascent.

  Solution. The smoothness of f^* is $1/\lambda_{\min}(A)$, where $\lambda_{\min}(A)$ is the smallest eigenvalue of A. The smootheness of g^* is $1/\lambda$. Therefore the overall smoothness is

  $$L \leq \frac{1}{\lambda_{\min}(A)} + \frac{1}{\lambda},$$

  and the learning rate can be set according to $1/L$ as

  $$\eta_t \leq \frac{\lambda_{\min}(A)\lambda}{\lambda_{\min}(A) + \lambda}.$$

  □

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

  Solution. see "prog-solution.py" □

- (1 points) Discuss the behaviors of proximal and dual averaging algorithms in the experiments

  Solution. Acceleration helps. Proximal and dual algorithms perform similarly, both in terms of convergence and in terms of model sparsity. Sparsity difference should be more visible in stochastic algorithms. □

- (1 points) Discuss the impacts of different μ in the experiments

  Solution. Large μ implies larger nonsmoothness. convergence of gradient descent and accelerated gradient descent methods will be affected. The advantage of using proximal gradient methods become more clear with larger μ. □

- (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?

  Solution. One can reduce learning rate to reduce oscillation. □