Optimization for Machine Learning (Homework #2)

Assignment date: Oct 3
Due date: Oct 17 (noon)

Theoretical Problems (11 points)

1. (5 points) Consider the quadratic objective function

$$Q(x) = 3x_1^2 + x_2^2 + 2x_1x_2 - x_1 - x_2$$

defined on $x = [x_1, x_2] \in \mathbb{R}^2$. Assume that we want to solve

$$x_* = \arg\min_x Q(x)$$

from $x_0 = 0$.

- (1 point) Find A and b so that $Q(x) = \frac{1}{2}x^\top Ax - b^\top x$.
  Solution. We have
  $$A = \begin{bmatrix} 6 & 2 \\ 2 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$
  □
- (2 point) For gradient descent method with constant learning rate η, what range should η belong to? What is the optimal value of η, and what is the corresponding convergence rate?
  Solution. The eigenvalues of A are $\lambda = 4 - 2\sqrt{2}$ and $L = 4 + 2\sqrt{2}$. We have $\eta \in (0, 2/L) = (0, 1 - 0.5\sqrt{2})$. The optimal value of $\eta = 1/4$ and convergence is $\rho = 0.5\sqrt{2}$. □
- (1 points) For CG, how many iterations T are needed to find $X_T = x_*$? Find values of α_1, β_1, and α_2.
  Solution. $T = 2$ is sufficient. $p_0 = r_0 = [1, 1]$, $q_0 = [8, 4]$, $\alpha_1 = 1/6$, $x_1 = [1/6, 1/6]$, $r_1 = [-1/3, 1/3]$, $\beta_1 = 1/9$, $p_1 = [-2/9, 4/9]$, $\alpha_1 = 3/4$. □
- (1 point) For the Heavy-Ball method with constant η and β. What's the optimal values of (η, β) to achieve the fastest asymptotic convergence rate, and what is the corresponding convergence rate?
  Solution. We can take $\eta = 1/(\sqrt{\lambda} + \sqrt{L})^2 = 1/(8 + 4\sqrt{2})$, and $\beta = (\sqrt{L} - \sqrt{\lambda})/(\sqrt{L} + \sqrt{\lambda}) = 1/(1 + \sqrt{2})$. β is the rate. □

2. (2 points) Consider the regularized logistic regression:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-w^\top x_i y_i)) + \frac{\lambda}{2}\|w\|_2^2$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. Assume $\|x_i\|_2 \le 1$ for all i.

- (1 point) find the smoothness parameter L of $f(w)$.

  Solution. We know that

$$\nabla^2 f(w) = \frac{1}{n}\sum_{i=1} \frac{1}{(1+\exp(-w^\top x_i y_i))(1+\exp(w^\top x_i y_i))} x_i x_i^\top + \lambda I \leq \frac{1}{4n}\sum_{i=1} x_i x_i^\top + \lambda I.$$

  The first term has largest eigenvalue of no more than 0.25 and thus the smoothness $L \leq 0.25 + \lambda$. The equality can be achieved at $w = 0$ and when all x_i are identical. □

- (1 point) find an estimate of Lipschitz constant G in the region $\{w : f(w) \leq f(0)\}$ which holds for all dataset $\{x_i\}$ such that $\|x_i\|_2 \leq 1$.

  Solution. The following is not the best estimate but sufficient for our purpose: we have

$$\frac{\lambda}{2}\|w\|_2^2 \leq f(0) = \ln 2.$$

  Therefore $\|w\|_2 \leq \sqrt{(2\ln 2)/\lambda}$.

$$\begin{aligned}\|\nabla f(w)\|_2 = & \left\| \frac{1}{n}\sum_{i=1} \frac{-x_i y_i}{1+\exp(w^\top x_i y_i)} + \lambda w \right\|_2 \\ \leq & \frac{1}{1+\exp(-\|w\|_2)} + \lambda \|w\|_2 \\ \leq & \frac{1}{1+\exp(-\sqrt{2\ln 2/\lambda})} + \sqrt{2\lambda \ln 2} \leq 1 + \sqrt{2\lambda \ln 2}.\end{aligned}$$

  □

3. (2 points) Consider training data (x_i, y_i) so that $\|x_i\|_2 \leq 1$ and $y_i \in \{\pm 1\}$, and we would like to solve the linear SVM problem

$$\min_w f(w) \triangleq \left[\frac{1}{n}\sum_{i=1}^{n}(1 - w^\top x_i y_i)_+ + \frac{\lambda}{2}\|w\|_2^2\right]$$

   using subgradient descent with $w_0 = 0$, and learning rate $\eta_t \leq \eta < 1/\lambda$.

   - (1 point) Let $C = \{w : \|w\|_2 \leq R\}$. Find the smallest R so that for all training data that satisfy the assumptions of the problem, subgradient descent without projection belongs to C.

     Solution. We have

$$\|w_t\|_2 \leq (1-\eta\lambda)\|w_{t-1}\|_2 + \eta.$$

     This implies that we can take $R = 1/\lambda$. □

   - (1 point) Find an upper bound of Lipschitz constant G of $f(w)$ in C.

     Solution. We have

$$G \leq 1 + \lambda\|w\|_2 \leq 2.$$

     □

4. (2 points) Given a nonsmooth function, we would like to find its smooth approximation.

   - (1 point) Find a closed form solution of

$$\phi_\gamma(x) = \min_z \left[(1-z)_+ + \frac{1}{2\gamma}(z-x)^2\right].$$

Solution. The solution z satisfies (with $\xi \in [0, -1]$)

$$\begin{cases} -1 + \frac{1}{\gamma}(z - x) = 0 & 1 - z > 0 \\ \xi + \frac{1}{\gamma}(z - x) = 0 & 1 - z = 0 \\ 0 + \frac{1}{\gamma}(z - x) = 0 & 1 - z < 0 \end{cases},$$

which implies

$$z = \begin{cases} x + \gamma & 1 - x > \gamma \\ 1 & 1 - x \in [0, \gamma] \\ x & 1 - x < 0 \end{cases}.$$

This implies that

$$\phi_\gamma(x) = \begin{cases} 1 - x - \frac{\gamma}{2} & 1 - x > \gamma \\ \frac{1}{2\gamma}(1 - x)^2 & 1 - x \in [0, \gamma] \\ 0 & 1 - x < 0 \end{cases}.$$

□

- (1 point) Let $x \in \mathbb{R}^d$, find

$$f(x) = \min_{z \in \mathbb{R}^d} \left[\|z\|_2 + \frac{1}{2\gamma}\|x - z\|_2^2 \right].$$

Solution. Let z achieve the minimum on the right hand side. If $\|x\|_2 \le \gamma$, then $z = 0$ is a solution, and

$$f(x) = \frac{1}{2\gamma}\|x\|_2^2.$$

If $\|x\|_2 > \gamma$, then $z = (1 - \gamma/\|x\|_2)x$, and

$$f(x) = \|x\|_2 - \gamma/2.$$

We thus obtain

$$f(x) = \begin{cases} \frac{1}{2\gamma}\|x\|_2^2 & \|x\|_2 \le \gamma \\ \|x\|_2 - 0.5\gamma & \text{otherwise} \end{cases}$$

□

Programming Problem (4 points)

We consider optimization with the smoothed hinge loss, and randomly generated data.

- Use the python template “prog-template.py”, and implement functions marked with ’# implement’.
- Submit your code and outputs. Please note that in real machine learning problems, strong-convexity settings are often associated with L2 regularization: it is necessary to choose a proper regularization to better estimate true w. Thus, you should
  1. choose the most proper setting mentioned in “prog-template.py” for real problems (including λ, γ, and the optimizer);
  2. Try some different λ and discuss how λ influences optimization and prediction respectively.

Solution. see “prog-solution.py”. □