# 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 $Q(x) = \frac{1}{2}x^\top Ax - b^\top x$.
- (2 point) For gradient descent method with constant learning rate $\eta$, what range should $\eta$ belong to? What is the optimal value of $\eta$, and what is the corresponding convergence rate?
- (1 points) For CG, how many iterations $T$ are needed to find $X_T = x_*$? Find values of $\alpha_1$, $\beta_1$, and $\alpha_2$.
- (1 point) For the Heavy-Ball method with constant $\eta$ and $\beta$. What's the optimal values of $(\eta, \beta)$ to achieve the fastest asymptotic convergence rate, and what is the corresponding convergence 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 $\|x_i\|_2 \leq 1$ for all $i$.

- (1 point) find the smoothness parameter $L$ of $f(w)$.
- (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 $\|x_i\|_2 \leq 1$.

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 $C$.

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

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

   - (1 point) Find a closed form solution of

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

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

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

   *Proof.* Let $z$ achieves 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 $\lambda$, $\gamma$, and the optimizer);
  2. Try some different $\lambda$ and discuss how $\lambda$ influences optimization and prediction respectively.