Optimization for Machine Learning (Homework #1)

Assignment date: Sep 19
Due date: Oct 3 (noon)

Read Chapter 2.5 (Convex Duality)

Theoretical Problems (10 points)

1. (1 points) Let $f(x) = \|x\|_1 + \|x\|_2^4/4$, where $x \in \mathbb{R}^d$. Find its conjugate $f^*(x)$.

   Solution. We have
   $$f^*(x) = \sup_u \left[\sum_{i=1}^d u_i x_i - |u_i| - u_i^2 \|u\|_2^2/4 \right].$$
   Therefore at the optimal u:
   $$x_i = \text{sign}(u_i) + \|u\|_2^2 u_i.$$
   Let $z_i = \max(|x_i| - 1, 0)\text{sign}(x_i)$, then We have
   $$\|u\|_2^2 u_i = z_i,$$
   which implies that
   $$u_i = z_i / \|z\|_2^{2/3},$$
   and
   $$f^*(x) = \frac{3}{4} \left(\sum_{i=1}^d \max(0, |x_i| - 1)^2 \right)^{2/3}.$$
   □

2. (2 points) Let $x \in \mathbb{R}$ and $y \in \mathbb{R}_+$. Is $f(x, y) = x^2/y$ a convex function? Prove your claim.

   Solution. $f(x, y)$ is a convex function. This is because
   $$\nabla^2 f(x, y) = \begin{bmatrix} 2/y & -2x/y^2 \\ -2x/y^2 & 2x^2/y^3 \end{bmatrix}$$
   is positive semi-definite (one eigenvalue is 0, the other is positive). □

3. (2 points) Consider the convex set $C = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$. Given $y \in \mathbb{R}^d$, compute the projection $\text{proj}_C(y)$.

   Solution. The projection $\bar{x}$ is the solution of
   $$\bar{x} = \arg\min_x \|x - y\|_2^2 \quad \text{subject to } \|x\|_\infty \leq 1.$$
   The solution should satisfy
   $$\bar{x}_j = \begin{cases} \min(y_j, 1) & \text{if } y_j > 0 \\ \max(y_j, -1) & \text{if } y_j \leq 0 \end{cases}.$$
   □

4. (3 points) Compute $\partial f(x)$ for the following functions of $x \in \mathbb{R}^d$

- $f(x) = \|x\|_2$
  Solution. If $x \neq 0$, then $\partial f(x) = x/\|x\|_2$.
  If $x = 0$, then $g \in \partial\|x\|_2$ if and only if $g^\top x \leq \|x\|_2$ for all x. This means that $\|g\|_2 \leq 1$. Therefore
  $$\partial_x \|x\|_2|_{x=0} = \{g : \|g\|_2 \leq 1\}.$$
  □

- $f(x) = \mathbb{1}(\|x\|_\infty \leq 1)$
  Solution. We have
  $$\partial \mathbb{1}(\|x\|_\infty \leq 1) = [g_i] \quad g_i = \begin{cases} 0 & |x_i| < 1 \\ \mu_i & x_i = 1 \\ -\mu_i & x_i = -1 \end{cases}, \text{ for } \mu_i \geq 0$$
  □

- $f(x) = \|x\|_2 + \|x\|_\infty$
  Solution. When $x = 0$,
  $$\partial\|x\|_\infty = \{g : \|g\|_1 \leq 1\}.$$
  Thus,
  $$\partial f(x) = \{g_1 + g_2 : \|g_1\|_2 \leq 1, \|g_2\|_1 \leq 1\}.$$
  When $x \neq 0$,
  $$\partial\|x\|_\infty = \mathrm{CO}\{\mathrm{sign}(x_j)e_j : |x|_j = \|x\|_\infty\},$$
  and thus
  $$\partial f(x) = \frac{x}{\|x\|_2} + \mathrm{CO}\{\mathrm{sign}(x_j)e_j : |x|_j = \|x\|_\infty\}$$
  □

5. (3 points) Consider the square root Lasso method. Given $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$, we want to find $w \in \mathbb{R}^d$ to solve

$$[w_*, \xi_*] = \arg\min_{w,b,\xi} \left[\|Xw - y\|_2 + \lambda \sum_{j=1}^{d} \xi_j \right], \tag{1}$$
$$\text{subject to } \xi_j \geq w_j, \quad \xi_j \geq -w_j \quad (j = 1, \ldots, d). \tag{2}$$

Lasso produces sparse solutions. Define the support of the solution as

$$S = \{j : w_{*,j} \neq 0\}.$$

Write down the KKT conditions under the assumption that $Xw_* \neq y$. Simplify in terms of $S, X_S, X_{\bar{S}}, y, w_S$. Here X_S contains the columns of X in S, $X_{\bar{S}}$ contains the columns of X not in S, and w_S contains the nonzero components of w_*.

Solution. Consider the Lagrangian function

$$L(w, \xi, \mu, \nu) = \left[\|Xw - y\|_2 + \lambda \sum_{j=1}^{d} \xi_j \right] + \sum_j \mu_j (w_j - \xi_j) + \sum_j \nu_j (-w_j - \xi_j).$$

For notation simplicity, we denote the optimal solution w_* by w, and the KKT conditions are

- $\mu_j(w_j - \xi_j) = 0$ and $\nu_j[-w_j - \xi_j] = 0$ and $\mu_j \geq 0$ and $\nu_j \geq 0$.
- $\xi_j \geq w_j$ and $\xi_j \geq -w_j$
- $\nabla_{w,\xi} L(w, \xi, \mu, \nu) = 0$.

From $\nabla_w L(w, \xi, \mu, \nu) = 0$, we obtain for all j:

$$X_j^\top (Xw - y) + (\mu_j - \nu_j)\|Xw - y\|_2 = 0.$$

From $\nabla_\xi L(w, \xi, \mu, \nu) = 0$, we obtain for all j:

$$\lambda = \mu_j + \nu_j.$$

We consider three cases:

- $w_j > 0$: since $\xi_j \geq w_j$, we have $-w_j - \xi_j < 0$, and thus $\nu_j = 0$ and $\mu_j = \lambda$. From $\mu_j(w_j - \xi_j) = 0$, we obtain $\xi_j = w_j$.
- $w_j < 0$: similarly, we have $\mu_j = 0$, and $\nu_j = \lambda$, a nd $\xi_j = -w_j$.
- $w_j = 0$: since $\lambda = \mu_j + \nu_j$, we have either $\mu_j\ neq 0$ or $\nu_j \neq 0$, and thus $\xi_j = 0$. Since $0 \leq \nu_j, \nu_j \leq \lambda$, we have $\mu_j - \nu_j \in [-\lambda, \lambda]$.

In summary, we have the following conditions:

$$X_S^\top (Xw - y) + \lambda \mathrm{sign}(w_S)\|Xw - y\|_2 = 0.$$

and

$$\|X_{\bar{S}}^\top (Xw - y)\|_\infty \leq \lambda \|Xw - y\|_2.$$

□

Programming Problem (4 points)

We consider ridge regression problem with randomly generated data. The goal is to implement gradient descent and experiment with different strong-convexity settings and different learning rates.

- Use the python template "prog-template.py", and implement functions marked with '# implement'.
- Submit your code and outputs. Compare to the theoretical convergence rates in class, and discuss your experimental results.

Solution. see "prog-solution.py" □