Homework 6 (2^{nd} graded homework)
CS-526 Learning Theory

Exercise 1

For the following convex functions, explain how to calculate a subgradient at a given $\mathbf{x}$.

1. $\forall x \in \mathbb{R}^n : f(x) = \max_{1 \le i \le m}(a_i^T x + b_i)$, where $\forall i \in \{1, \dots, m\} : (a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$.
2. $\forall x \in \mathbb{R}^n : f(x) = \max_{1 \le i \le m} |a_i^T x + b_i|$.
3. $\forall x \in \mathbb{R}^n : f(x) = \sup_{t \in [0,1]} p(t, x)$, where $p(t, x) = x_1 + x_2 t + \cdots + x_n t^{n-1}$.

Exercise 2

We recall the definition of a strongly convex function: A function f is λ-strongly convex if for all w, u and $\alpha \in (0, 1)$ we have:

$$f(\alpha w + (1-\alpha)u) \le \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2 .$$

Theorem 14.11 in the textbook is a refined bound for Stochastic Gradient Descent (SGD) when the function f is strongly convex. The proof of this theorem relies on the following claim (Claim 14.10 in *Understanding Machine Learning*):

If f is λ-strongly convex then for every w, u and $v \in \partial f(w)$ we have

$$\langle w - u, v \rangle \ge f(w) - f(u) + \frac{\lambda}{2}\|w-u\|^2$$

Prove this claim.

Exercise 3

$\mathcal{M}_n(\mathbb{R})$ is the Hilbert space of $n \times n$ real matrices endowed with the inner product $\langle A, B \rangle = \mathrm{Tr}(A^T B)$. The induced norm is the Euclidian (or Frobenius) norm, i.e.,

$$\|A\| = \sqrt{\mathrm{Tr}(A^T A)} = \left(\sum_{i,j=1}^{n} (A_{ij})^2 \right)^{1/2} .$$

Consider the cone of $n \times n$ symmetric positive semi-definite matrices, denoted $\mathcal{S}_n^+ \subseteq \mathcal{M}_n(\mathbb{R})$. For all $A \in \mathcal{S}_n^+$, $\lambda_{\max}(A)$ is the maximum eigenvalue associated to A. We define

$$f : \begin{array}{ccc} \mathcal{S}_n^+ & \to & [0, +\infty) \\ A & \mapsto & \lambda_{\max}(A) \end{array} .$$

a) Show that f is convex.
b) Find a subgradient $V \in \partial f(A)$ for any $A \in \mathcal{S}_n^+$.
Hint: A subgradient of f at A is a matrix $V \in \mathbb{R}^{n\times n}$ that satisfies:

$$\forall B \in \mathcal{S}_n^+ : f(B) \geq f(A) + \mathrm{Tr}\big((B-A)^T V\big).$$

Exercise 4

Consider the following Least Squares optimization problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}||A\mathbf{x}-\mathbf{b}||_2^2,$$

where $b \in \mathbb{R}^m$, A is a full column rank matrix in $\mathbb{R}^{m\times n}, n \leq m$ and there exists a solution to the linear system $A\mathbf{x} = \mathbf{b}$. Let $\sigma_{\max}$ and $\sigma_{\min}$ be the largest and the smallest singular values of A and consider the gradient descent method

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha\nabla f(\mathbf{x}^t)$$

with a fixed step size $\alpha = 1/\sigma_{\max}(A)^2$.
a) Show that $\sigma_{\max}(I - \alpha A^T A) = 1 - \alpha\sigma_{\min}(A)^2 = 1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}$.

b) Calculate the gradient $\nabla f(\mathbf{x})$ and rewrite the GD using this gradient.

c) Show that the procedure converges as

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 \leq (1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2})||\mathbf{x}^t - \mathbf{x}^*||_2.$$

Exercise 5

Consider a dataset given by $S = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ satisfies $\|x_i\| = 1$, and $y_i \in \mathbb{R}$ for all $1 \leq i \leq n$. Let X be the matrix with x_i's as its rows. Assume that the smallest eigenvalue of the matrix $X^T X$ is $\mu > 0$. We consider the 'linear noiseless setting', where we assume that there exists a $\beta^* \in \mathbb{R}^d$ such that $y_i = x_i^T\beta^*$ for all $i \leq i \leq n$. We want to find β^* by minimizing the loss function

$$L(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(\beta, x_i, y_i) = \frac{1}{n}\sum_{i=1}^n (x_i^T\beta - y_i)^2.$$

1. Show that for any $\beta, \beta' \in \mathbb{R}^d$,

$$L(\beta') - L(\beta) \geq (\beta' - \beta)^T\nabla L(\beta) + \frac{\mu}{n}\|\beta' - \beta\|^2.$$

2. Consider the following stochastic gradient descent for minimizing the loss function L: At each step k, we sample i_k uniformly at random from $\{1, 2, \cdots, n\}$ independent of the previous steps and do the SGD step given by

$$\beta_{k+1} = \beta_k - \eta\nabla\ell(\beta_k, x_{i_k}, y_{i_k}).$$

Show that for sufficiently small η, we have

$$\mathbb{E}\|\beta_k - \beta^*\|^2 \leq \left(1 - \frac{2\eta\mu}{n}\right)^k \|\beta_0 - \beta^*\|^2.$$

Find the values of η for which the above convergence rate is satisfied.
Hint: First estimate the conditional expectation of $\|\beta_k - \beta^*\|^2$ given β_{k-1}.

3. Discuss the differences between the convergence result in question 2 and the convergence result for SGD discussed in class for convex functions with bounded stochastic gradients.