Homework 6 (2^{nd} graded homework): due Monday 17th April
CS-526 Learning Theory

Exercise 1

$\mathcal{M}_n(\mathbb{R})$ 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 2 (adapted from 14.3, *Understanding Machine Learning*)

Let $S = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)) \in (\mathbb{R}^d \times \{-1, +1\})^m$. Assume that there exists $\mathbf{w} \in \mathbb{R}^d$ such that for every $i \in [m]$ we have $y_i\langle \mathbf{w}, \mathbf{x}_i\rangle \geq 1$, and let $\mathbf{w}^\star$ be a vector that has the minimal norm among all vectors that satisfy the preceding requirement. Let $R = \max_i \|\mathbf{x}_i\|$. Define a function $f(\mathbf{w}) = \max_{i\in[m]}(1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle)$.

a) Show that $\min_{\mathbf{w}:\|\mathbf{w}\|\leq\|\mathbf{w}^\star\|} f(\mathbf{w}) = 0$.

b) Show that any $\mathbf{w}$ for which $f(\mathbf{w}) < 1$ separates the examples in S.

c) Show how to calculate a subgradient of f.

d) Describe a subgradient descent algorithm for finding a $\mathbf{w}$ that separates the examples. Show that the number of iterations T of your algorithm satisfies

$$T \leq R^2\|\mathbf{w}^*\|^2.$$

Hint: it is a good idea to take a look at the Batch Perceptron algorithm in Section 9.1.2. for the analysis.
e) (Not graded) Compare your algorithm to the Batch Perceptron algorithm.

Exercise 3

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 σ_{max} and σ_{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 $\alpha = 1/\sigma_{\text{max}}(A)^2$.

a) Show that $\sigma_{\text{max}}(I - \alpha A^T A) = 1 - \alpha\sigma_{\text{min}}(A)^2 = 1 - \frac{\sigma_{\text{min}}(A)^2}{\sigma_{\text{max}}(A)^2}$.

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

c) Show that the procedure converges as

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