Homework 5 (2^{nd} graded homework): 26 March 2019
CS-526 Learning Theory

The deadline is Tuesday, April 9 2019. Please hand in your homework during the lecture (April 8) or the exercise session (April 9). **No scan of handwritten homework is accepted**.

Exercise 1 (adapted from J. Duchi)

$\mathcal{M}_n(\mathbb{R})$ 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 $\min_{\mathbf{w}:\|\mathbf{w}\| \leq \|\mathbf{w}^\star} f(w) = 0$.
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) (Ungraded) Compare your algorithm to the Batch Perceptron algorithm.

Exercise 3 (adapted from 14.4, *Understanding Machine Learning*)

Algorithm 1: SGD with adaptive learning rate

---

parameters: T
initialize: $\mathbf{w}^{(1)} = 0$
for $t = 1 \ldots T$ **do**
  Choose a random vector $\mathbf{v}_t$ s.t. $\mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(w^{(t)})$
  Set $\eta_t = {}^{B}/_{\rho\sqrt{t}}$
  Set $\mathbf{w}^{(t+1/2)} = \mathbf{w}^{(t)} - \eta_t \mathbf{v}_t$.
  Set $\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{y}:\|\mathbf{y}\| \le B} \|\mathbf{w}^{(t+1/2)} - \mathbf{y}\|$.
end
output: $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$

Prove the following theorem on the above algorithm and specify the constant $\alpha > 0$.

Theorem 1. *Let $B, \rho > 0$. Let f be a convex function and let $\mathbf{w}^\star \in \arg\min_{\mathbf{w}:\|\mathbf{w}\| \le B} f(\mathbf{w})$. Assume that SGD is run for T iterations with $\eta_t = \frac{B}{\rho\sqrt{t}}$. Assume also that for all t, $\mathbb{E}\|\mathbf{v}_t\|^2 \le \rho^2$. Then*

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \le \alpha \cdot \frac{\rho B}{\sqrt{T}}$$

Exercise 4 (6.3 from *Understanding Machine Learning*)

Let $\mathcal{X}$ be the Boolean hypercube $\{0,1\}^n$. For a set $I \subseteq \{1, 2, \ldots, n\}$ we denote a parity function h_I as follows. On a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$,

$$h_I(\mathbf{x}) = \sum_{i \in I} x_i \mod 2\,.$$

(That is, h_I computes parity of bits in I.) What is the VC-dimension of the class of all such parity functions,

$$\mathcal{H}_{n-parity} = \{h_I : I \subseteq \{1, 2, \ldots, n\}\}?$$