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 \(\langle A, B \rangle = \text{Tr}(A^T B)\). The induced norm is the Euclidian (or Frobenius) norm, i.e.,

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

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

\[ f : S^+_n \to [0, +\infty) \quad A \mapsto \lambda_{\text{max}}(A). \]

a) Show that \(f\) is convex.

b) Find a subgradient \(V \in \partial f(A)\) for any \(A \in S^+_n\).

*Hint:* A subgradient of \(f\) at \(A\) is a matrix \(V \in \mathbb{R}^{n \times n}\) that satisfies:

\[ \forall B \in S^+_n : f(B) \geq f(A) + \text{Tr}((B - A)^T V). \]

**Exercise 2 (adapted from 14.3, *Understanding Machine Learning*)**

Let \(S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathbb{R}^d \times \{-1, +1\})^m\). Assume that there exists \(w \in \mathbb{R}^d\) such that for every \(i \in [m]\) we have \(y_i \langle w, x_i \rangle \geq 1\), and let \(w^*\) be a vector that has the minimal norm among all vectors that satisfy the preceding requirement. Let \(R = \max_i \|x_i\|\). Define a function \(f(w) = \max_{i \in [m]} (1 - y_i \langle w, x_i \rangle)\).

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

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

c) Show how to calculate a subgradient of \(f\).

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

\[ T \leq R^2 \|w^*\|^2. \]

*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.
Algorithm 1: SGD with adaptive learning rate

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

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 $w^* \in \arg \min \{w : \|w\| \leq B f(w)\}$. Assume that SGD is run for $T$ iterations with $\eta_t = \frac{B}{\rho \sqrt{t}}$. Assume also that for all $t$, $\mathbb{E} \|v_t\|^2 \leq \rho^2$. Then

$$\mathbb{E}_v \{f(\bar{w})\} - f(w^*) \leq \alpha \cdot \frac{\rho B}{\sqrt{T}}$$

**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 $x = (x_1,x_2,\ldots,x_n) \in \{0,1\}^n$,

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

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

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