Homework 6 (Graded): due Monday, April 22, 2024
CS-526 Learning Theory

Problem 1. Gradient Descent for Positive Semi-definite Matrices
Let $X, Y \in \mathbb{R}^{n\times n}$ be $n\times n$ real matrices and $A, B \in \mathbb{R}^{n\times n}$ be $n\times n$ real symmetric and positive definite matrices. Let $F : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ the function $F(X) = \frac{1}{2}\mathrm{Tr}X^T BX$.

1. Show that $F(X) \geq 0$ for any X.
2. Compute the second derivative of
$$f(s) = \mathrm{Tr}(sX^T + (1-s)Y^T)B(sX + (1-s)Y)$$
for $s \in [0,1]$ and deduce that F is a convex function.
3. Deduce the inequality $F(Y) - F(X) \geq \mathrm{Tr}X^T B(Y - X)$. Is F Lipschitz ?
4. Consider now the function $G : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ with $G(X) = \frac{1}{2}\mathrm{Tr}(X - I)^T A(X - I)$ where I is the identity matrix. Define $L(X) = F(X) + G(X)$.
   (a) Write down the gradient descent algorithm for L. Call X_t the updated matrix at time t.
   (b) Assume that the operator norm $\|X_t\| \leq M$ stays bounded uniformly in t. Show that
$$\|\frac{1}{T}\sum_{t=1}^{T} X_t - (B + A)^{-1}A\| \leq \frac{2M}{\eta T}\|(B + A)^{-1}\|$$

Problem 2. Gradient Descent.
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex Lipshitz continuous differentiable function with Lipshitz constant $\rho > 0$. Let S be a real symmetric strictly positive-definite $d \times d$ matrix with smallest eigenvalue $\lambda_{\min} > 0$. We consider a gradient descent iteration for $t \geq 1$ and step size $\eta > 0$:
$$x^{t+1} = x^t - \eta S^{-1}\nabla f(x^t) \tag{1}$$
with initial condition $x^1 = 0$. Further, define $x^* = \mathrm{argmin}_{\|x\|\in B(0,R)} f(x)$, where $B(0, R)$ is the ball of radius R.

1. Show that if we choose the step size $\eta = \frac{R\sqrt{\lambda_{\max}\lambda_{\min}}}{\rho\sqrt{T}}$ after T iterations we have
$$f\big(\frac{1}{T}\sum_{t=1}^{T} x^t\big) - f(x^*) \leq \frac{\rho R}{\sqrt{T}}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

Hint: recall that in class we proved this statement when $S = I$ the identity matrix. Here you can use an eigenvalue decomposition $S^{-1} = U^T \Lambda^{-1} U$. The following is also useful:

$$\left\langle \underline{\nabla} f\left(x^t\right), x^t - x^*\right\rangle = \left\langle U \nabla f\left(x^t\right), U x^t - U x^*\right\rangle = \sum_{k=1}^{d} (U\nabla f)_k (x^t) \left(U x^t - U x^*\right)_k$$

Justify why these steps can be used.

Problem 3. (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 $f(\mathbf{w}) = \max_{i \in [m]} (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$.

1. Show that $\min_{\mathbf{w}: \|\mathbf{w}\| \leq \|\mathbf{w}^\star\|} f(\mathbf{w}) = 0$.
2. Show that any $\mathbf{w}$ for which $f(\mathbf{w}) < 1$ separates the examples in S.
3. Show how to calculate a subgradient of f.
4. 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.
5. (Not graded) Compare your algorithm to the Batch Perceptron algorithm.