Solutions to Homework 6 (graded)
CS-526 Learning Theory

Exercise 1

a) Fix $A, B \in \mathcal{S}_n^+$ and $\alpha \in [0,1]$. Let $\mathbf{e} \in \mathbb{R}^n$ a unit-norm eigenvector of $\alpha A + (1-\alpha)B$ associated to the maximum eigenvalue, i.e., $(\alpha A + (1-\alpha)B)\mathbf{e} = \lambda_{\max}(\alpha A + (1-\alpha)B)\mathbf{e}$ and $\|\mathbf{e}\| = 1$. We have:

$$\begin{aligned} f(\alpha A + (1-\alpha)B) = \mathbf{e}^T(\alpha A + (1-\alpha)B)\mathbf{e} &= \alpha \mathbf{e}^T A \mathbf{e} + (1-\alpha)\mathbf{e}^T B \mathbf{e} \\ &\leq \alpha \lambda_{\max}(A) + (1-\alpha)\lambda_{\max}(B) \\ &= \alpha f(A) + (1-\alpha) f(B)\,. \end{aligned}$$

This shows that f is convex.
b) Let $A \in \mathcal{S}_n^+$. A subgradient of f at A is a matrix $V \in \mathbb{R}^{n \times n}$ that satisfies:

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

Consider any $\mathbf{e} \in \mathbb{R}^n$ which is a unit-norm eigenvector of A associated to the maximum eigenvalue, i.e., $A\mathbf{e} = \lambda_{\max}(A)\mathbf{e}$ and $\|\mathbf{e}\| = 1$. Then for all $B \in \mathcal{S}_n^+$:

$$\begin{aligned} f(A) = \lambda_{\max}(A) = \mathbf{e}^T A \mathbf{e} = \mathbf{e}^T B \mathbf{e} + \mathbf{e}^T(A-B)\mathbf{e} &\leq \lambda_{\max}(B) + \mathbf{e}^T(A-B)\mathbf{e} \\ &= f(B) + \mathrm{Tr}(\mathbf{e}^T(A-B)\mathbf{e}) \\ &= f(B) + \mathrm{Tr}((A-B)^T \mathbf{e}\mathbf{e}^T)\,. \end{aligned}$$

In the last equality we used that $(A-B)^T = A - B$ and that the trace is preserved by cyclic permutations. We see that $\mathbf{e}\mathbf{e}^T$ satisfies the definition of a subgradient: $\mathbf{e}\mathbf{e}^T \in \partial f(A)$.

Exercise 2

a) $\min_{\|\mathbf{w}\| \leq \|\mathbf{w}^*\|} f(\mathbf{w}) \leq f(\mathbf{w}^*) \leq 0$ because $\forall i \in [m] : y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1$. Suppose there exists $\mathbf{w}$ satisfying both $\|\mathbf{w}\| \leq \|\mathbf{w}^*\|$ and $f(\mathbf{w}) < 0$. Then $\mathbf{w}$ can be slightly modify to obtain a vector $\tilde{\mathbf{w}}$ such that $\|\tilde{\mathbf{w}}\| < \|\mathbf{w}^*\|$, while still having $f(\tilde{\mathbf{w}}) \leq 0$. It contradicts $\mathbf{w}^*$'s definition, hence $\min_{\|\mathbf{w}\| \leq \|\mathbf{w}^*\|} f(\mathbf{w}) \geq 0$. It proves $\min_{\|\mathbf{w}\| \leq \|\mathbf{w}^*\|} f(\mathbf{w}) = 0$.

b) If $f(\mathbf{w}) < 1$ then $\forall i \in [m] : y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle > 0$, i.e., $\mathbf{w}$ separates the examples.

c) For all $i \in [m]$ the gradient of $f_i : \mathbf{w} \mapsto 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$ is $-y_i \mathbf{x}_i$. Applying Claim 14.6, we get that a subgradient of f at $\mathbf{w}$ is given by $-y_{i^*}\mathbf{x}_{i^*}$ where $i^* \in \arg\max_{i \in [m]} \{1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}$.

d) The algorithm is inialized with $\mathbf{w}^{(1)} = 0$. At each iteration, if $f(\mathbf{w}^{(t)}) \geq 1$ then it chooses $i^* \in \arg\min_{i \in [m]} \{y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle\}$ and updates $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_{i^*} \mathbf{x}_{i^*}$. Otherwise, if

$f(\mathbf{w}^{(t)}) < 1$, $\mathbf{w}^{(t)}$ separates all the examples and we stop. To analyze the speed of convergence of the subgradient algorithm, first notice that $\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \rangle - \langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle = \eta y_{i^*} \langle \mathbf{w}^*, \mathbf{x}_{i^*} \rangle \geq \eta$. Therefore, after performing T iterations, we have

$$\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle = \langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle - \langle \mathbf{w}^*, \mathbf{w}^{(1)} \rangle = \sum_{t=1}^{T} \langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \rangle - \langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle \geq \eta T \,. \tag{1}$$

Besides, $\|\mathbf{w}^{(t+1)}\|^2 = \|\mathbf{w}^{(t)}\|^2 + \eta^2 y_{i^*}^2 \|\mathbf{x}_i\|^2 + 2\eta y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_{i^*} \rangle \leq \|\mathbf{w}^{(t)}\|^2 + \eta^2 R^2$. The last inequality follows from $\|\mathbf{x}_i\| \leq R$ and $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_{i^*} \rangle \leq 0$ (we update only if $f(\mathbf{w}^{(t)}) \geq 1$). Then

$$\|\mathbf{w}^{(T+1)}\| \leq \eta R \sqrt{T} \,. \tag{2}$$

Combining Cauchy-Schwarz inequality, (**??**) and (**??**), we obtain

$$1 \geq \frac{\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle}{\|\mathbf{w}^{(T+1)}\| \|\mathbf{w}^*\|} \geq \frac{\sqrt{T}}{R \|\mathbf{w}^*\|} \,. \tag{3}$$

The subgradient algorithm must stop in less than $R^2 \|\mathbf{w}^*\|^2$ iterations. We see that η does not affect the speed of convergence.

e) The algorithm is almost identical to the Batch Perceptron algorithm with two modifications. First, the Batch Perceptron updates with any example for which $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$, while the current algorithm chooses the example for which $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle$ is minimal. Second, the current algorithm employs the parameter η. However, the only difference with the case $\eta = 1$ is that it scales $\mathbf{w}^{(t)}$ by η.

Exercise 3

a) Assume that A has the singular value decomposition $U \Lambda V^T$. Plugging this into the expression $I - \alpha A^T A$ we see that $I - \alpha A^T A$ has the singular value decomposition $V \Lambda' V^T$, where Λ' is of dimension $n \times n$ and has the singular values $1 - \alpha \sigma_i^2$. For the given choice of α all these singular values are non-negative and the largest is $1 - \alpha \sigma_{\min}^2(A) = 1 - \frac{\sigma_{\min}^2(A)}{\sigma_{\max}^2(A)}$.
b) We get

$$\nabla f(\mathbf{x}) = A^T(A\mathbf{x} - \mathbf{b}) = A^T A(\mathbf{x} - \mathbf{x}^*),$$

where we used the fact that A has full column rank so that $A\mathbf{x}^* = b$. Hence GD can be rewritten as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha A^T A(\mathbf{x}^t - \mathbf{x}^*). \tag{4}$$

c) Subtracting $\mathbf{x}^*$ from both sides of (**??**) gives

$$\mathbf{x}^{t+1} - \mathbf{x}^* = \mathbf{x}^t - \mathbf{x}^* - \alpha A^T A(\mathbf{x}^t - \mathbf{x}^*) = (I - \alpha A^T A)(\mathbf{x}^t - \mathbf{x}^*).$$

By taking norms we obtain

$$\begin{aligned} ||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &\leq \sigma_{\max}(I - \alpha A^T A) ||\mathbf{x}^t - \mathbf{x}^*||_2 \\ &= (1 - \alpha \sigma_{min}(A)^2) ||\mathbf{x}^t - \mathbf{x}^*||_2. \end{aligned}$$