Solution 5 (2^{nd} graded homework): 26 March 2019
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, (1) and (2), 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. 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

We prove the following Theorem:

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

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{3\rho B}{\sqrt{T}}$$

Proof. By Jensen's inequality, we have:

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \right] . \tag{4}$$

As $\forall t : \mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$, we can reproduce what is done in Theorem 14.8 to get the inequality:

$$\mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \right] \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \right] . \tag{5}$$

We now have to prove an upper bound on the right-hand side of (5). This is similar to what is done in Lemma 14.10, except that we have to take into account the time-dependence of the steps η_t. For all $t \in \{1 \dots, T\}$:

$$\begin{aligned}
\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle = \frac{1}{\eta_t} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \eta_t \mathbf{v}_t \rangle &= \frac{1}{2\eta_t} \big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^\star - \eta_t \mathbf{v}_t\|^2 + \eta_t^2 \|\mathbf{v}_t\|^2 \big) \\
&= \frac{1}{2\eta_t} \big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1/2)} - \mathbf{w}^\star\|^2 + \eta_t^2 \|\mathbf{v}_t\|^2 \big) \\
&\leq \frac{1}{2\eta_t} \big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 \big) + \frac{\eta_t}{2} \|\mathbf{v}_t\|^2. \quad (6)
\end{aligned}$$

Let $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \leq B\}$. The last inequality follows from $\mathbf{w}^{(t+1)} = \pi_{\mathcal{H}}(\mathbf{w}^{(t+1/2)})$ and the 1-Lipschitzianity of $\pi_{\mathcal{H}}$ (see Homework 4, Exercise 4):

$$\|\pi_{\mathcal{H}}(\mathbf{w}^{(t+1/2)}) - \mathbf{w}^\star\| = \|\pi_{\mathcal{H}}(\mathbf{w}^{(t+1/2)}) - \pi_{\mathcal{H}}(\mathbf{w}^\star)\| \leq \|\mathbf{w}^{(t+1/2)} - \mathbf{w}^\star\| .$$

Summing the inequality (6) over t, we have:

$$\begin{aligned}
\sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle &\leq \sum_{t=1}^T \frac{1}{2\eta_t} \big(\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 \big) + \frac{\eta_t}{2} \|\mathbf{v}_t\|^2 \\
&= \frac{1}{2\eta_1} \|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \sum_{t=1}^{T-1} \frac{\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2}{2} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \\
&\qquad - \frac{1}{2\eta_T} \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2 + \sum_{t=1}^T \frac{\eta_t}{2} \|\mathbf{v}_t\|^2 \\
&\leq \frac{1}{2\eta_1} \|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \sum_{t=1}^{T-1} \frac{\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2}{2} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \sum_{t=1}^T \frac{\eta_t}{2} \|\mathbf{v}_t\|^2 \\
&\leq 2B^2 \left(\frac{1}{\eta_1} + \sum_{t=1}^{T-1} \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \sum_{t=1}^T \frac{\eta_t}{2} \|\mathbf{v}_t\|^2 \\
&= \frac{2B^2}{\eta_T} + \sum_{t=1}^T \frac{\eta_t}{2} \|\mathbf{v}_t\|^2 . \qquad (7)
\end{aligned}$$

Taking the expectation of inequality (7) and diving by T, we obtain:

$$\mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \right] \leq \frac{2B^2}{T\eta_T} + \sum_{t=1}^T \frac{\eta_t}{2T} \mathbb{E}\|\mathbf{v}_t\|^2 \leq \frac{2\rho B}{\sqrt{T}} + \frac{\rho^2}{2T} \sum_{t=1}^T \eta_t . \qquad (8)$$

The last inequality follows from the assumption $\mathbb{E}\|\mathbf{v}_t\|^2 \leq \rho^2$ and η_T's definition. Besides

$$\sum_{t=1}^T \eta_t = \frac{B}{\rho} \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq \frac{B}{\rho} \left(1 + \sum_{t=2}^T \int_{t-1}^t \frac{dx}{\sqrt{x}} \right) = \frac{B}{\rho} \left(1 + \int_1^T \frac{dx}{\sqrt{x}} \right) = \frac{B}{\rho} \big(2\sqrt{T} - 1 \big) .$$

Combining this last inequality with (4), (5) and (8), we finally obtain:

$$\mathbb{E}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{2\rho B}{\sqrt{T}} + \frac{\rho B}{2T} \big(2\sqrt{T} - 1 \big) \leq \frac{3\rho B}{\sqrt{T}} .$$

It concludes the proof. $\square$

Exercise 4

$\mathcal{H}_{n-parity}$ is a finite class, therefore (see paragraph 6.3.4):

$$\text{VCdim}(\mathcal{H}_{n-parity}) \leq \log_2 |\mathcal{H}_{n-parity}| = \log_2 2^n = n\,.$$

We now show that this upperbound on $\text{VCdim}(\mathcal{H}_{n-parity})$ is tight, i.e., there exists n points in $\{0,1\}^n$ that are shattered by $\mathcal{H}_{n-parity}$. Let $\mathbf{e}^{(j)} \in \{0,1\}^n$ be such that $\mathbf{e}^{(j)}_j = 1$ and $\forall i \neq j : \mathbf{e}^{(j)}_i = 0$. The subset $C = \{\mathbf{e}^{(j)}\}_{j=1}^n$ of n points is shattered by $\mathcal{H}_{n-parity}$. Indeed, given $(y_1, \ldots, y_n) \in \{0,1\}^n$, we can define $J = \{j \in \{1, \ldots, n\} : y_j = 1\}$ and see that:

$$\forall j \in \{1, \ldots, n\} : h_J(\mathbf{e}^{(j)}) = \sum_{i \in J} \mathbf{e}^{(j)}_i \mod 2 = \sum_{i=1}^{n} \mathbf{e}^{(j)}_i y_i \mod 2 = y_j\,.$$

Hence $\text{VCdim}(\mathcal{H}_{n-parity}) = n$.