Exercise 1

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

$$f(\alpha A + (1 - \alpha)B) = e^T(\alpha A + (1 - \alpha)B)e = \alpha e^T A e + (1 - \alpha) e^T B e$$

$$\leq \alpha \lambda_{\max}(A) + (1 - \alpha) \lambda_{\max}(B)$$

$$= \alpha f(A) + (1 - \alpha) f(B).$$

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:

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

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

$$f(A) = \lambda_{\max}(A) = e^T A e = e^T B e + e^T (A - B) e \leq \lambda_{\max}(B) + e^T (A - B) e$$

$$= f(B) + \text{Tr}(e^T (A - B) e)$$

$$= f(B) + \text{Tr}((A - B)^T ee^T).$$

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

Exercise 2

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

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

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

d) The algorithm is initialized with $w^{(1)} = 0$. At each iteration, if $f(w^{(t)}) \geq 1$ then it chooses $i^* \in \arg \min_{i \in [m]} \{y_i \langle w^{(t)}, x_i \rangle\}$ and updates $w^{(t+1)} = w^{(t)} + \eta y_i x_{i^*}$. Otherwise, if
\( f(w^{(t)}) < 1, \ w^{(t)} \) separates all the examples and we stop. To analyze the speed of convergence of the subgradient algorithm, first notice that \( \langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle = \eta y_i \langle w^*, x_i \rangle \geq \eta. \) Therefore, after performing \( T \) iterations, we have
\[
\langle w^*, w^{(T+1)} \rangle = \langle w^*, w^{(T+1)} \rangle - \langle w^*, w^{(1)} \rangle = \sum_{t=1}^{T} \langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle \geq \eta T. \tag{1}
\]

Besides, \( \|w^{(t+1)}\|^2 = \|w^{(t)}\|^2 + \eta^2 y_i^2 \|x_i\|^2 + 2\eta y_i \langle w^{(t)}, x_i \rangle \leq \|w^{(t)}\|^2 + \eta^2 R^2. \) The last inequality follows from \( \|x_i\| \leq R \) and \( y_i \langle w^{(t)}, x_i \rangle \leq 0 \) (we update only if \( f(w^{(t)}) \geq 1 \)). Then
\[
\|w^{(T+1)}\| \leq \eta R \sqrt{T}. \tag{2}
\]

Combining Cauchy-Schwarz inequality, (1) and (2), we obtain
\[
1 \geq \frac{\langle w^*, w^{(T+1)} \rangle}{\|w^{(T+1)}\| \|w^*\|} \geq \frac{\sqrt{T}}{R \|w^*\|}. \tag{3}
\]

The subgradient algorithm must stop in less than \( R^2 \|w^*\|^2 \) iterations. We see that \( \eta \) 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 \( y_i \langle w^{(t)}, x_i \rangle \leq 0 \), while the current algorithm chooses the example for which \( y_i \langle w^{(t)}, x_i \rangle \) is minimal. Second, the current algorithm employs the parameter \( \eta \). However, the only difference with the case \( \eta = 1 \) is that it scales \( w^{(t)} \) by \( \eta \).

**Exercise 3**

We prove the following Theorem:

**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 \( t \), \( \mathbb{E}\|\mathbf{v}_t\|^2 \leq \rho^2 \). Then
\[
\mathbb{E}_{\mathbf{v}_{1:T}}[f(\mathbf{w})] - f(w^*) \leq \frac{3\rho B}{\sqrt{T}}.
\]

**Proof.** By Jensen’s inequality, we have:
\[
\mathbb{E}_{\mathbf{v}_{1:T}}[f(\mathbf{w})] - f(w^*) \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) - f(w^*) \right]. \tag{4}
\]

As \( \forall t : \mathbb{E}[\mathbf{v}_t | w^{(t)}] \in \partial f(w^{(t)}) \), we can reproduce what is done in Theorem 14.8 to get the inequality:
\[
\mathbb{E}_{\mathbf{v}_{1:T}} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) - f(w^*) \right] \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[ \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, \mathbf{v}_t \rangle \right]. \tag{5}
\]

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 $\eta_t$. For all $t \in \{1, \ldots, T\}$:

$$
\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{\eta_t} \langle w^{(t)} - w^*, \eta_t v_t \rangle = \frac{1}{2\eta_t} (\|w^{(t)} - w^*\|^2 - \|w^{(t)} - w^* - \eta_t v_t\|^2 + \eta_t^2 \|v_t\|^2)
$$

$$
= \frac{1}{2\eta_t} (\|w^{(t)} - w^*\|^2 - \|w^{(t+1)} - w^*\|^2 + \eta_t^2 \|v_t\|^2)
$$

$$
\leq \frac{1}{2\eta_t} (\|w^{(t)} - w^*\|^2 - \|w^{(t+1)} - w^*\|^2 + \frac{\eta_t}{2} \|v_t\|^2). \quad (6)
$$

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

$$
\|\pi_H(w^{(t+1)}) - w^*\| = \|\pi_H(w^{(t+1)}) - \pi_H(w^*)\| \leq \|w^{(t+1)} - w^*\|.
$$

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

$$
\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \sum_{t=1}^{T} \frac{1}{2\eta_t} (\|w^{(t)} - w^*\|^2 - \|w^{(t+1)} - w^*\|^2 + \frac{\eta_t}{2} \|v_t\|^2)
$$

$$
= \frac{1}{2\eta_1} \|w^{(1)} - w^*\|^2 + \sum_{t=1}^{T-1} \frac{1}{\eta_{t+1}} \|w^{(t+1)} - w^*\|^2 \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t+1}} \right) + \frac{\eta_T}{2} \|v_t\|^2
$$

$$
\leq \frac{1}{2\eta_1} \|w^{(1)} - w^*\|^2 + \sum_{t=1}^{T-1} \frac{1}{\eta_{t+1}} \|w^{(t+1)} - w^*\|^2 \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t+1}} \right) + \sum_{t=1}^{T} \frac{\eta_t}{2} \|v_t\|^2
$$

$$
\leq 2B^2 \left( \frac{1}{\eta_1} + \sum_{t=1}^{T-1} \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \sum_{t=1}^{T} \frac{\eta_t}{2} \|v_t\|^2
$$

$$
= \frac{2B^2}{\eta_1} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|v_t\|^2. \quad (7)
$$

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

$$
\mathbb{E}_{v_{1:T}} \left[ \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \right] \leq \frac{2B^2}{T\eta_1} + \sum_{t=1}^{T} \frac{\eta_t}{2T} \mathbb{E}\|v_t\|^2 \leq \frac{2\rho B}{\sqrt{T}} + \frac{\rho^2}{2T} \sum_{t=1}^{T} \eta_t. \quad (8)
$$

The last inequality follows from the assumption $\mathbb{E}\|v_t\|^2 \leq \rho^2$ and $\eta_t$’s definition. Besides

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

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

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

It concludes the proof. \qed
Exercise 4

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

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

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

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

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

4