Learning Theory - Homework 2

Alexandru Mocanu, SCIPER 295172

9 April 2019

1 Exercise 1

Along the entire exercise, we consider $\|\cdot\|$ to be the spectral norm when applied to matrices.

a) We start by noting that S_n^+ is a convex set. Consider $A, B \in S_n^+$ and $\eta \in (0,1)$. Then, $\forall x \in \mathbb{R}^n$, we have that:

$$x^T[\eta A + (1-\eta)B]x = \eta x^T A x + (1-\eta) x^T B x \geq 0 \tag{1}$$

Therefore, $\eta A + (1-\eta)B \in S_n^+$, so S_n^+ is a convex set.

Now, consider $A, B \in S_n^+$ and $\eta \in (0,1)$. We know that $\lambda_{max}(A)$ is the spectral norm of matrix A. In other words, $\lambda_{max}(A) = \max\limits_x \frac{\|Ax\|}{\|x\|} = \max\limits_{x:\|x\|=1} \sqrt{x^T A^T A x}$. As the spectral norm is a norm, we have that

$$\|\eta A + (1-\eta)B\| \leq \|\eta A\| + \|(1-\eta)B\| = \eta\lambda_{max}(A) + (1-\eta)\lambda_{max}(B) \tag{2}$$

which is equivalent to $f(\eta A + (1-\eta)B) \leq \eta f(A) + (1-\eta) f(B)$, so f is convex.

b) Consider $v \in \mathbb{R}^n$ such that $\sum\limits_{i=1}^n v_i = 1$ and $A^T v = \lambda_{max}(A) v$. We construct matrix $V = [v, v, ..., v]$ and prove that it is a subgradient of f at A.

We see that $tr(A^T V) = tr([A^T v, A^T v, ..., A^T v]) = \lambda_{max}(A) tr(V) = \lambda_{max} \sum\limits_{i=1}^n v_i = \lambda_{max}(A)$. We also note that $B^T V = [B^T v, B^T v, ..., B^T v]$, so matrix $B^T V$ is 1-rank, having only one nonzero eigenvalue. $Vv = v$, so the only nonzero eigenvalue of V is 1.

Consider two $n \times n$ matrices A and B. Considering $x = \arg\max\limits_{y:\|y\|=1} \|ABy\|$, we have that:

$$\|AB\| = \|ABx\| \leq \max_{y:\|y\|=1} \|Ay\|\|Bx\| = \|A\|\|Bx\| \leq \|A\|\|B\|\|x\| = \|A\|\|B\| \tag{3}$$

Using the above fact for B and V, we have that $\|B^T V\| \leq \|B^T\|\|V\| = \lambda_{max}(B)$. This result along with the fact that the trace of a matrix is equal to the sum of its eigenvalues, leads us to:

$$f(A) + tr((B-A)^T V) \leq \lambda_{max}(A) + \lambda_{max}(B) - \lambda_{max}(A) = \lambda_{max}(B) = f(B), \tag{4}$$

so V is a subgradient of f at A.

2 Exercise 2

a) For w with the property presented in the statement, $1 - y_i x_i^T w \leq 0, \forall i \in [m]$, so $\max_i(1 - y_i x_i^T w) \leq 0$. Assume that $\max_i(1 - y_i x_i^T w^*) < 0$. Then $1 < y_i x_i^T w^*, \forall i \in [m]$, but then we can choose $w = \frac{w^*}{y_j x_j^T w^*}$, where $j = \arg\max_i(1 - y_i x_i^T w^*) = \arg\min_i y_i x_i^T w^*$ and this leads us to $y_j x_j^T w = 1$ and $y_i x_i^T w \geq 1, \forall i \neq j$, as $y_j x_j^T w^* \leq y_i x_i^T w^*, \forall i \in [m]$. This means that w^* is not the minimal norm solution which gives a contradiction.

Therefore, $\max_i(1 - y_i x_i^T w^*) = 1$, so $f(w^*) = 0$. Any other w with $\|w\| = \|w^*\|$ that has the property from the statement, also satisfies $f(w) = 0$. All other w with $\|w\| \leq \|w^*\|$ do not satisfy the property from the statement, so $\exists i \in [m]$ such that $y_i x_i^T w < 1$, so $f(w) > 0$.

In conclusion, $\min_{w : \|w\| \leq \|w^*\|} f(w) = 0$ with the minimum achieved at w^*.

b) If $f(w) < 1$, this means that $y_i x_i^T w > 0, \forall i \in [m]$. This implies that for for $y_i = 1$ we get $x_i^T w$ positive and for $y_i = -1$ we get $x_i^T w$ negative, so $sign(x_i^T w) = y_i, \forall i \in [m]$ and therefore w indeed separates all the examples in S.

c) In order for $g_1 \in \mathbb{R}^d$ to be a subgradient at w_1, we need

$$f(w_2) \geq f(w_1) + g_1^T(w_2 - w_1) \iff g_1^T(w_1 - w_2) \geq f(w_1) - f(w_2), \forall w_1, w_2 \in \mathbb{R}^d \tag{5}$$

We also have that

$$f(w_1) - f(w_2) = \max_i(1 - y_i x_i^T w_1) - \max_i(1 - y_i x_i^T w_2) = \min_i y_i x_i^T w_2 - \min_i y_i x_i^T w_1 \tag{6}$$

But we know that $\min_i y_i x_i^T w_2 - \min_i y_i x_i^T w_1 < y_j x_j^T(w_2 - w_1)$, where $j = \arg\min_i y_i x_i^T w_1$.

Therefore, choosing $g_1 = -y_j x_j$ with $j = \arg\min_i y_i x_i^T w_1$ leads to

$$g_1^T(w_1 - w_2) = y_j x_j^T(w_2 - w_1) \geq f(w_1) - f(w_2), \tag{7}$$

so g_1 is a subgradient of f at w.

d) As we have seen, $g = -y_j x_j$ with $j = \arg\min_i y_i x_i^T w$ gives a subgradient of f at w. The objective is to approach w^* that achieves perfect separation. We therefore propose algorithm 1 as a subgradient descent algorithm.

What remains is to prove that the number of iterations until convergence of the algorithm in case of linearly separable samples is $T \leq R^2\|w^*\|^2$. We will proceed as for the Batch Perceptron algorithm in section 9.1.2 of Understanding Machine Learning.

Data: A training set $(x_1, y_1), ..., (x_m, y_m)$
Initialize: $w^{(1)} = (0, 0, ..., 0)$
for $t = 1, 2, ...$ **do**
  if $\exists i\ s.t.\ y_i x_i^T w^{(t)} \leq 0$ **then**
    $j = \arg\min_i y_i x_i^T w^{(t)}$;
    $g = -y_j x_j$;
    $w^{(t+1)} = w^{(t)} - g$;
  end
  else
    output $w^{(t)}$
  end
end

Algorithm 1: Subgradient descent

We have w^* as defined in the statement. We will prove that after performing T iterations, the cosine of the angle between w^* and $w^{(T+1)}$ is at least $\frac{\sqrt{T}}{R\|w^*\|}$, or equivalently

$$\frac{\langle w^*, w^{(T+1)}\rangle}{\|w^*\|\|w^{(T+1)}\|} \geq \frac{\sqrt{T}}{R\|w^*\|} \tag{8}$$

Using the Cauchy-Schwarz inequality along with the inequality above, we get that $1 \geq \frac{\sqrt{T}}{R\|w^*\|}$, so $T \leq R^2\|w^*\|^2$. To prove that inequality 8 holds, we start by noting that $\langle w^*, w^{(T+1)}\rangle \geq T$. At the first iteration we have $w^{(1)} = (0, 0, ..., 0))$, so $\langle w^*, w^{(1)}\rangle = 0$, while at iteration t, if we update using sample (x_i, y_i), we get

$$\langle w^*, w^{(t+1)}\rangle - \langle w^*, w^{(t)}\rangle = \langle w^*, w^{(t+1)} - w^{(t)}\rangle = \langle w^*, y_i x_i\rangle = y_i\langle w^*, x_i\rangle \geq 1 \tag{9}$$

After T iterations, we therefore get

$$\langle w^*, w^{(T+1)}\rangle = \sum_{t=1}^{T} \left(\langle w^*, w^{(t+1)}\rangle - \langle w^*, w^{(t)}\rangle\right) \geq T \tag{10}$$

just like we wanted to prove. We now upper-bound $\|w^{(T+1)}\|$:

$$\|w^{(t+1)}\|^2 = \|w^{(t)} + y_i x_i\|^2 = \|w^{(t)}\|^2 + y_i^2\|x_i\|^2 + 2y_i\langle w^{(t)}, x_i\rangle \leq \|w^{(t)}\|^2 + R^2 \tag{11}$$

By progressively applying this inequality and by the fact that $\|w^{(1)}\| = 0$, we get that

$$\|w^{(T+1)}\|^2 \leq TR^2 \Rightarrow \|w^{(T+1)}\| \leq R\sqrt{T} \tag{12}$$

Combining equations 10 and 12, we get that

$$\frac{\langle w^*, w^{(T+1)} \rangle}{\|w^*\| \|w^{(T+1)}\|} \geq \frac{T}{\|w^*\| R\sqrt{T}} = \frac{\sqrt{T}}{R\|w^*\|} \tag{13}$$

This concludes our proof.

e) Unlike in the Batch Perceptron algorithm, at each step we choose the worst classified datapoint to update the weights.

3 Exercise 3

We need some preliminary statements before we venture with the actual proof.

We first show that $\|w^{(t+\frac{1}{2})} - w^*\|^2 \geq \|w^{(t+1)} - w^*\|^2$. In case $\|w^{(t+\frac{1}{2})}\| \leq B$, we have that $w^{(t+\frac{1}{2})} = w^{(t+1)}$, so we actually get equality. For $\|w^{(t+\frac{1}{2})}\| > B$, we get $w^{(t+1)} = w^{(t+\frac{1}{2})} \frac{B}{\|w^{(t+\frac{1}{2})}\|}$ such that $\|w^{(t+1)}\| = B$, so we proceed:

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

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

$$\|w^{(t+1)} - w^*\|^2 \leq \|w^{(t+\frac{1}{2})} - w^*\|^2 \iff B^2 + 2\left(1 - \frac{B}{\|w^{(t+\frac{1}{2})}\|}\right)(w^{(t+\frac{1}{2})})^T w^* \leq \|w^{(t+\frac{1}{2})}\|^2 \tag{14}$$

But, we have that $(w^{(t+\frac{1}{2})})^T w^* \leq B\|w^{(t+\frac{1}{2})}\|$, with equality for $\angle(w^{(t+\frac{1}{2})}, w^*) = 0$. Therefore:

$$B^2 + 2\left(1 - \frac{B}{\|w^{(t+\frac{1}{2})}\|}\right)(w^{(t+\frac{1}{2})})^T w^* \leq B^2 + 2\|w^{(t+\frac{1}{2})}\|B - 2B^2 = 2B\|w^{(t+\frac{1}{2})}\| - B^2 \leq \|w^{(t+\frac{1}{2})}\|^2 \iff$$

$$\iff 0 \leq (\|w^{(t+\frac{1}{2})}\| - B)^2 \tag{15}$$

The last inequality is obviously true, so we have that:

$$\|w^{(t+\frac{1}{2})} - w^*\|^2 \geq \|w^{(t+1)} - w^*\|^2 \tag{16}$$

Secondly, we prove that $\|w^{(t)} - w^*\|^2 \leq 4B^2$.

$$\|w^{(t)} - w^*\|^2 = \|w^{(t)}\|^2 + \|w^*\|^2 - 2(w^{(t)})^T w^* \leq 4B^2 \tag{17}$$

with equality for $\|w^{(t)}\| = \|w^*\| = B$ and $\angle(w^{(t)}, w^*) = \pi$. So, we indeed get

$$\|w^{(t)} - w^*\|^2 \leq 4B^2 \tag{18}$$

Thirdly, we prove by induction that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. For the base case we simply have that $\frac{1}{\sqrt{1}} \leq 2\sqrt{1}$, which is clearly true. We assume the statement to be true for T and prove it for T+1:

$$\sum_{t=1}^{T+1} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}+\frac{1}{\sqrt{T+1}} \leq 2\sqrt{T+1} \iff \frac{1}{\sqrt{T+1}} \leq 2(\sqrt{T+1}-\sqrt{T}) = \frac{2}{\sqrt{T}+\sqrt{T+1}} \tag{19}$$

The last inequality is true, so the assumption holds and we get:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} \tag{20}$$

We now proceed with the proof. By convexity, we have that

$$\mathbb{E}[f(\bar{w})] = \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} w^{(t)}\right)\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(w^{(t)})] \Rightarrow \mathbb{E}[f(\bar{w})]-f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(w^{(t)})-f(w^*)] \tag{21}$$

As $\mathbb{E}[v_t]$ is a subgradient of f at $w^{(t)}$, convexity also implies that

$$f(w^*) \geq f(w^{(t)})+\mathbb{E}[v_t]^T(w^*-w^{(t)}) \Rightarrow f(w^{(t)})-f(w^*) \leq \mathbb{E}[v_t]^T(w^{(t)}-w^*) \tag{22}$$

But we have that

$$\begin{aligned}\mathbb{E}[v_t]^T(w^{(t)}-w^*) = \mathbb{E}[v_t^T(w^{(t)}-w^*)] = \mathbb{E}\left[\frac{1}{\eta_t}(w^{(t)}-w^{(t+\frac{1}{2})})^T(w^{(t)}-w^*)\right] = \\ = \frac{1}{2\eta_t}\mathbb{E}[\|w^{(t)}-w^{(t+\frac{1}{2})}\|^2+\|w^{(t)}-w^*\|^2-\|w^{(t+\frac{1}{2})}-w^*\|^2]\end{aligned} \tag{23}$$

Using now inequality 16, we get

$$\mathbb{E}[v_t]^T(w^{(t)}-w^*) \leq \frac{1}{2}\eta_t\mathbb{E}[\|v_t\|^2]+\frac{1}{2\eta_t}\mathbb{E}[\|w^{(t)}-w^*\|^2-\|w^{(t+1)}-w^*\|^2] \tag{24}$$

Summing over t, this leads to

$$\sum_{t=1}^{T}\mathbb{E}[f(w^{(t)})-f(w^*)] \leq \frac{B}{2\rho}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\mathbb{E}[\|v_t\|^2]+\frac{\rho}{2B}\sum_{t=1}^{T}\sqrt{t}(\|w^{(t)}-w^*\|^2-\|w^{(t+1)}-w^*\|^2) \tag{25}$$

We note that

$$\sum_{t=1}^{T}\sqrt{t}(\|w^{(t)}-w^*\|^2-\|w^{(t+1)}-w^*\|^2) = \sum_{t=1}^{T}(\sqrt{t}-\sqrt{t-1})\|w^{(t)}-w^*\|^2 \leq 4B^2\sum_{t=1}^{T}(\sqrt{t}-\sqrt{t-1}) = 4B^2\sqrt{T} \tag{26}$$

where we used inequality 18.
From 21, 25, 26, 20 and using the fact that $\mathbb{E}[\|v_t\|^2] \leq \rho^2$, we get that

$$\mathbb{E}[f(\bar{w})]-f(w^*) \leq \frac{1}{T}\left(\frac{1}{2}2\rho B\sqrt{T}+\frac{\rho}{2B}4B^2\sqrt{T}\right) = 3\frac{\rho B}{\sqrt{T}} \tag{27}$$

so we satisfy the theorem with $\alpha = 3$.

4 Exercise 4

We notice that $|\mathcal{H}_{n-parity}| = 2^n - 1$, so we have VCdim$(\mathcal{H}_{n-parity}) < n$, as for a set of n points from $\mathcal{X}$ we would have 2^n different classifications to make.

Consider the ordered subset $S = (0, 0, 0, ..., 0, 1), (0, 0, 0, ..., 1, 0), ..., (0, 1, 0, ..., 0)$ for which element i contains only one 1 at the i-th least significant bit. Consider now a classification of the elements in S that contains 1 bits at positions $i_1 < i_2 < ... < i_m$. Choosing the subset $I = \{n+1-i_1, n+1-i_2, ..., n+1-i_m\}$ gives us the hypothesis function h_I which makes the desired classification, as element i_j from S contributes to the classification with its 1 from position $n+1-i_j$. For example, for $n = 4$ we have:

S	classification	I
0001	000	{1}
0010	001	{2}
0100	010	{3}
	011	{2, 3}
	100	{4}
	101	{2, 4}
	110	{3, 4}
	111	{2, 3, 4}

We have therefore found a subset of size $n - 1$ that is shattered by $\mathcal{H}_{n-parity}$, so VCdim$(\mathcal{H}_{n-parity}) = n - 1$.