# COMS W4771: Machine Learning (sec:001) - Homework #5

Name: Geraldi Dzakwan (gd2551), Discussants: Deka (da2897), Patricia (pk2618)

December 9, 2019

## Problem 1

a. The proof will be divided into three steps to mimic the one in practice problem:

1. When the perception makes a mistake at time t, the update is $w^t = w^{t-1} + \eta yx$ (which is different from practice problem). Let's consider the update effect on $\overrightarrow{w}^{(t)}.\overrightarrow{w}^*$:

$$\overrightarrow{w}^{(t)}.\overrightarrow{w}^* = (\overrightarrow{w}^{(t-1)} + \eta y\overrightarrow{x}) * \overrightarrow{w}^* = \overrightarrow{w}^{(t-1)}.\overrightarrow{w}^* + \eta y\overrightarrow{x}\overrightarrow{w}^*$$

Because $\underset{(x,y)\in S}{min}\, yx^T w^* = 1$ (which is equivalent to the margin $\gamma$ in the practice problem but not normalized), then:

$$\overrightarrow{w}^{(t)}.\overrightarrow{w}^* \geq \overrightarrow{w}^{(t-1)}.\overrightarrow{w}^* + \eta(1)$$

$$\overrightarrow{w}^{(t)}.\overrightarrow{w}^* \geq \overrightarrow{w}^{(t-1)}.\overrightarrow{w}^* + 1/L^2$$

In other words, for every update, $\overrightarrow{w}^{(t)}.\overrightarrow{w}^*$ increases at least by $1/L^2$. Suppose T is the number of iteration. By induction (base case is $w^{(0)} = 0$), we will get:

$$\overrightarrow{w}^{(T)}.\overrightarrow{w}^* \geq \frac{T}{L^2} \text{ ... Eq 1.1}$$

2. Next, let's consider the update effect on $(\overrightarrow{w}^{(t-1)})^T\overrightarrow{w}^{(t)} = ||\overrightarrow{w}^{(t)}||^2$:

$$||\overrightarrow{w}^{(t)}||^2 = ||\overrightarrow{w}^{(t-1)} + \eta y\overrightarrow{x}||^2$$

$$||\overrightarrow{w}^{(t)}||^2 = ||\overrightarrow{w}^{(t-1)}||^2 + 2\eta y\overrightarrow{x}\overrightarrow{w}^{(t-1)} + \eta^2||y\overrightarrow{x}||^2$$

Notice that in this algorithm:

1. The updated training example satisfies $\eta yx^T w^{(t-1)} < 1$.
2. L is $\underset{(x,y)\in S}{max}||x||_2$ and $y^2$ is 1 (y is either -1 or +1).

Using above statements, we can have:

$$||\overrightarrow{w}^{(t)}||^2 \leq ||\overrightarrow{w}^{(t-1)}||^2 + 2\eta(1) + \eta^2(1)L^2$$
$$||\overrightarrow{w}^{(t)}||^2 \leq ||\overrightarrow{w}^{(t-1)}||^2 + 2(1/L^2) + (1/L^2)^2L^2$$
$$||\overrightarrow{w}^{(t)}||^2 \leq ||\overrightarrow{w}^{(t-1)}||^2 + \frac{3}{L^2}$$

In other words, for every update, $||\overrightarrow{w}^{(t)}||^2$ increases by at most $3/L^2$. By induction (base case is $w^{(0)} = 0$), we will get:

$$||\overrightarrow{w}^{(t)}||^2 \leq \frac{3T}{L^2} \text{ ... Eq 1.2}$$

3. From the prior two parts:

$$\frac{T}{L^2} \leq \overrightarrow{w}^{(t)}.\overrightarrow{w}^* \leq ||\overrightarrow{w}^{(t)}||.||\overrightarrow{w}^*|| \leq \frac{3T}{L^2}$$

because $a.b \leq ||a||\ ||b||$. Since $||\overrightarrow{w}^*||_2$ is not 1 (different from the practice problem), we would have:

$$\frac{T}{L^2} \leq \overrightarrow{w}^* \sqrt{\frac{3T}{L^2}}$$

Solve for T:

$$\frac{T^2}{L^4} \leq ||\overrightarrow{w}^*||_2^2(\frac{3T}{L^2})$$
$$\boldsymbol{T \leq 3||\overrightarrow{w}^*||_2^2 L^2}$$

b. Let's take two statements from Part (a):

1. The proof's claim:

$$T \leq 3||\overrightarrow{w}^*||_2^2 L^2 \text{ ... Eq 3.3}$$

2. Some part of the proof, Eq 1.2:

$$||\overrightarrow{w}^{(t)}||_2^2 \leq \frac{3T}{L^2}$$

Solve for T:

$$\frac{1}{3}||\overrightarrow{w}^{(t)}||_2^2 L^2 \leq T \text{ ... Eq 3.4}$$

From Eq 3.3 and Eq 3.4, we have:

$$\frac{1}{3}||\overrightarrow{w}^{(t)}||_2^2 L^2 \leq 3||\overrightarrow{w}^*||_2^2 L^2$$
$$||\overrightarrow{w}^{(t)}||_2^2 \leq 9||\overrightarrow{w}^*||_2^2$$
$$\boldsymbol{||\overrightarrow{w}^{(t)}||_2 \leq 3||\overrightarrow{w}^*||_2}$$

c. The margin definition in this case is:

1. For $w^*$:

$$m^* = \frac{\min\limits_{(x,y)\in S} yx^T w^*}{||w^*||_2} = \frac{1}{||w^*||_2} \text{ (by the definition of } w^*)$$

2. For $w^{(t)}$:

$$m^{(t)} = \frac{\min\limits_{(x,y)\in S} yx^T w^{(t)}}{||w^{(t)}||_2} = \frac{1}{||w^{(t)}||_2}$$

Notice that $\min\limits_{(x,y)\in S} yx^T w^{(t)} = 1$ because the Margin Perceptron terminates when $yx^T w^{(t)} \geq 1$ for all $x, y \in S$.

Because $||\overrightarrow{w}^{(t)}||_2 \leq 3||\overrightarrow{w}^*||_2$ and margin is normalized by the norm of the weight vector, then we can clearly see that:

$$m^{(t)} = \frac{1}{||w^{(t)}||_2} \geq \frac{1}{3||w^*||_2}$$

$$m^{(t)} \geq \frac{m^*}{3}$$

In other words, $m^{(t)}$ is as large as $m^*$ within a factor of three.

# Problem 2

a. $f(x) := -x^4$. $f(x)$ is twice differentiable. So, we can determine the convexity by looking at the value of $f''(x)$.

$$f'(x) = -4x^3$$

$$f''(x) = -12x^2$$

Because $f''(x) <= 0$ for $-\infty < x < \infty$, then $f(x) = x^4$ is concave.

b. $f(x) := x^3$. $f(x)$ is twice differentiable. So, we can determine the convexity by looking at the value of $f''(x)$.

$$f'(x) = 3x^2$$

$$f''(x) = 6x$$

$f(x)$ is neither concave nor convex because:

1. For $x < 0$, $f''(x) < 0$ so it is not convex for $x < 0$.
2. For $x > 0$, $f''(x) > 0$ so it is not concave for $x > 0$.

c. $f(x) := sin(x)$. $f(x)$ is twice differentiable. So, we can determine the convexity by looking at the value of $f''(x)$.

$$f'(x) = cos(x)$$

$$f''(x) = -sin(x)$$

$f(x)$ is neither concave nor convex because:

1. For $0 < x < \pi$, $f''(x) < 0$ so it is not convex for $0 < x < \pi$.
2. For $\pi < x < 2\pi$, $f''(x) > 0$ so it is not concave for $\pi < x < 2\pi$.

d. $f(x) := \frac{1}{2}x^T Ax + b^T x + c$. $f(x)$ is twice differentiable. So, we can determine the convexity by looking at the value of $f''(x)$.

$$f'(x) = \frac{1}{2}(2Ax) + b = Ax + b$$

$$f''(x) = A$$

To determine the convexity of $f(x)$, we need to check if A if positive semi definite (thus convex) or negative semi definite (thus concave).

$$|A - \lambda I| = 0$$

$$(2 - \lambda)^2 - 6^2 = 0$$

$$\lambda^2 - 4\lambda - 32 = 0$$

$$(\lambda - 8)(\lambda + 4) = 0$$

$$\lambda_1 = 8; \lambda_2 = -4$$

Because the two eigenvalues have different sign, then A is neither positive semi definite nor negative semi definite. Thus, $f(x)$ is neither convex nor concave.

e. $f(x) := \frac{1}{2}x^T(A - 2\lambda_1 v_1 v_1^T)x + b^T x + c$. We know that $f''(x) = A - 2\lambda_1 v_1 v_1^T$, the derivation steps are the same as that in part (d). Again, like in part (d), to determine the convexity of $f(x)$, we need to check if $A - 2\lambda_1 v_1 v_1^T$ if positive semi definite (thus convex) or negative semi definite (thus concave).

From part(d), we will pick the biggest eigenvalue, i.e. $\lambda_1 = 8$. Calculate the corresponding unit length eigenvector $v_1$, suppose it is denoted by [p, q]:

$$\begin{bmatrix} 2 - \lambda_1 & 6 \\ 6 & 2 - \lambda_1 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = 0$$

$$\begin{bmatrix} -6 & 6 \\ 6 & -6 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = 0$$

$$6\begin{bmatrix}-1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}p\\ q\end{bmatrix} = 0$$

The above basically says that $p = q$. Because $v_1$ is a unit length eigenvector:

$$\|v_1\|_2^2 = \sqrt{p^2+q^2} = 1$$
$$\sqrt{p^2+p^2} = \sqrt{2p^2} = 1$$
$$p = \frac{1}{\sqrt{2}} \text{ or } p = -\frac{1}{\sqrt{2}}$$

Then, the eigenvector is either $v_1 = (1/\sqrt{2}, 1/\sqrt{2})$ or $v_1 = (-1/\sqrt{2}, -1/\sqrt{2})$. But, they would yield the same $v_i v_i^T = \begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix}$. Hence, we have:

$$A - 2\lambda_1 v_1 v_1^T = \begin{bmatrix}2 & 6\\ 6 & 2\end{bmatrix} - 16\begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix} = \begin{bmatrix}-6 & -2\\ -2 & -6\end{bmatrix}$$

If we try to find the eigenvalue from the above matrix, we would get:

$$(-6-\lambda)^2 - (-2)^2 = 0$$
$$\lambda^2 + 12\lambda + 32 = 0$$
$$(\lambda+8)(\lambda+4) = 0$$
$$\lambda_1 = -8, \lambda_2 = -4$$

Because it has all negative eigenvalues, then the matrix is negative semi definite and hence f(x) is concave.

f. $f(x) := |b^T x|$. Suppose $x = (x_1, x_2)$ so that $f(x) = |b^T x| = |x_1 + 2x_2|$. We can write $f(x)$ as $f(x) = p(q(x))$ where:

1. $q(x) = b^T x = x_1 + 2x_2 \rightarrow (R^2 : R)$
2. $p(x) = |x| \rightarrow (R : R)$

Since $p(x)$ is convex and $q(x)$ is an affine map, then $f(x) = p(q(x))$ is also convex.

g. $f(x) := \sqrt{|x^T M x|}$. Suppose $x = (x_1, x_2)$, then:

$$f(x) = \sqrt{|x^T M x|} = \sqrt{|\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}-6 & -2\\ -2 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}|}$$

$$f(x) = \sqrt{|-6x_1^2 - 4x_1x_2 - 6x_2^2|}$$

Notice that $-6x_1^2 - 4x_1x_2 - 6x_2^2$ is always negative so that we can simplify $f(x)$:

$$f(x) = \sqrt{-(-6x_1^2 - 4x_1x_2 - 6x_2^2)} = \sqrt{6x_1^2 + 4x_1x_2 + 6x_2^2}$$

Furthermore, We can write $f(x)$ as $f(x) = p(q(x))$ where:

1. $q(x) = 6x_1^2 + 4x_1x_2 + 6x_2^2 \rightarrow (R^2 : R)$
2. $p(x) = \sqrt{x} \rightarrow (R : R)$

Take a look at $q(x)$. The Hessian matrix would be:

$$\begin{bmatrix} 12 & 4 \\ 4 & 12 \end{bmatrix}$$

If we try to find the eigenvalue from the above matrix, we would get:

$$(12 - \lambda)^2 - (4)^2 = 0$$
$$\lambda^2 - 24\lambda + 144 = 0$$
$$(\lambda - 8)(\lambda - 16) = 0$$
$$\lambda_1 = 8, \lambda_2 = 16$$

Since all the eigenvalues are positive, the Hessian matrix is positive semi definite and thus $q(x)$ is convex. Because $p(x)$ is just taking the square root from a non-negative function that is convex, then the resulting function will also be convex. Thus, $f(x)$ is convex.

h. $f(x) := (\sigma(b^T x) - 1)^2 + (\sigma(-b^T x))^2$. We know that:

$$1 - \sigma(a) = 1 - \frac{1}{1 + e^{-a}} = \frac{e^{-a}}{1 + e^{-a}} = \sigma(-a)$$

Thus:

$$f(x) := (\sigma(b^T x) - 1)^2 + (1 - \sigma(b^T x))^2$$
$$f(x) := 2(\sigma(b^T x) - 1)^2$$

Suppose $x = (x_1, x_2)$, then:

$$f(x) = 2(\sigma(x_1 + 2x_2) - 1)^2$$

$f(x)$ is neither concave nor convex because:

1. Take a look at the range: $[p, q] = [(0, 0); (-1, -1)]$. Using Jensen inequality with $\alpha = 1/2$:

$$f((1/2)p + (1/2)q) = f(-1/2, -1/2) = 1.3369$$
$$(1/2)f(p) + (1/2)f(q) = 1.1574$$

Because $f((1/2)p + (1/2)q) > (1/2)f(p) + (1/2)f(q)$, then for the range $[p, q]$, $f(x)$ is not convex.

2. Take a look at the range: $[p, q] = [(0, 0); (1, 1)]$. Using Jensen inequality with $\alpha = 1/2$:

$$f((1/2)p + (1/2)q) = f(1/2, 1/2) = 0.068$$
$$(1/2)f(p) + (1/2)f(q) = 0.252$$

Because $f((1/2)p + (1/2)q) < (1/2)f(p) + (1/2)f(q)$, then for the range $[p, q]$, $f(x)$ is not concave.

# Problem 3

a. I would prove the statement using induction on K. We divide the proof into the base case and the induction step:

1. The base case would be when K equals to 1, i.e. there is only one (left-most) line segment. The piece-wise linear function for a compact interval $[x_0, x_1]$, slope $s_1$ and intercept $f(x_0) = b$, can be expressed by:

$$f(x) = s_1(x - x_0) + b$$

Because for the interval $[x_0, x_1]$, $x \geq x_0 \rightarrow x - x_0 \geq 0$, we then can have an equivalent function $g(x)$ where:

$$g(x) = s_1\sigma(x - x_0) + b = f(x)$$

For $b$, we have three cases:

1. $b = 0$. $g(x) = s_1\sigma(x - x_0)$
2. $b > 0$. Consider the following equality:

$$b = (x - x_0 + b) - (x - x_0)$$

Again, because $x - x_0 \geq 0$, we can have:

$$b = \sigma(x - x_0 + b) - \sigma(x - x_0)$$

Thus:

$$g(x) = s_1\sigma(x - x_0) + \sigma(x - x_0 + b) - \sigma(x - x_0)$$
$$g(x) = (s_1 - 1)\sigma(x - x_0) + \sigma(x - (x_0 - b))$$

3. $b < 0$. Consider the following equality:

$$b = (x - x_0) - (x - x_0 - b)$$

Again, because $x - x_0 \geq 0$, we can have:

$$b = \sigma(x - x_0) - \sigma(x - x_0 - b)$$

Thus:

$$g(x) = s_1\sigma(x - x_0) + \sigma(x - x_0) - \sigma(x - x_0 - b)$$
$$g(x) = (s_1 + 1)\sigma(x - x_0) - \sigma(x - (x_0 + b))$$

Thus, now $g(x)$ satisfies to be a family of univariate two-layers ReLU as all the above three functions follow the general definition:

$$\sum_{i=1}^{m} a_i \sigma(w_i x + b_i)$$

where all $w_i$ are equal to 1 in this case.

Even though the base case is proved, for the sake of the induction step, we need to modify $g(x)$ so that for $x > x_1$, $g(x) = g(x_1)$ (it forms a flat line). We can do this by using the hint provided $\rightarrow \sigma(z) - \sigma(z-1)$ and a slight modification. We need to add another ReLU unit to $g(x)$ to be subtracted:

$$g(x) = g(x) - s_1 \sigma(x - x_1)$$

2. For the induction case, we start with a hypothesis that for $K = n$, there is a univariate two-layers ReLU network $g(x)$ that can satisfy $f(x) = g(x)$ for the compact interval $I = [x_0, x_n]$. Also, $g(x)$ has a property that $g(x) = g(x_n)$ for $x > x_n$ (as elaborated in the base case).

   Suppose $f'(x)$ is a new continuous piecewise affine function that is built by adding a new right-most line segment to $f(x)$ so now it spans from $x_0$ to $x_{n+1}$. Suppose $g'(x)$ is a function that is built by adding some other ReLU units to $g(x)$.

   We need to prove that for $K = n + 1$, there is a way to modify $g(x)$ so that $f'(x) = g'(x)$ for a new compact interval $I' = [x_0, x_{n+1}]$. Let $f'(x)$ be:

$$f'(x) = f(x) \text{ for } [x_0, x_n]$$

$$f'(x) = f(x_n) + s_{n+1}(x - x_n) \text{ for } (x_n, x_{n+1}]$$

   Because for $x_n < x \leq x_{n+1}$, $x - x_n > 0$, then we can have an equivalent function $g'(x)$ where:

$$g'(x) = g(x) + s_{n+1}\sigma(x - x_n) \text{ for } [x_0, x_{n+1}]$$

   In this case, $g'(x) = f'(x)$ for $[x_0, x_{n+1}]$ holds because we start with the hypothesis of $f(x) = g(x)$ for $[x_0, x_n]$ and we have showed that $g(x) = g(x_n) = f(x_n)$ for $x > x_n$ in the base case.

Both steps are proved. The statement is proved.

# Problem 4

a. First, I want to prove that $a_1^t = a_2^t$ for all $t \geq 1$. I am going to use induction to prove the statement. As usual, divide into two cases:

1. Base case, i.e. $t = 1$.

$$a_1^1 = a_1^0 - \eta \frac{\delta}{\delta a_1} \hat{R}(\theta^0)$$

$$a_1^1 = a_1^0 - \eta \frac{\delta}{\delta a_1} (\frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i)^2)$$

Using chain rule:

$$a_1^1 = a_1^0 - \frac{2\eta}{n} \sum_{i=1}^{n} ((\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) * \frac{\delta}{\delta a_1} (\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i))$$

For the right-most part, all terms not involving $a_1 \rightarrow a_2, a_3, ..., a_m, w_2, w_3, ..., w_m$ and $y_1, y_2, ..., y_n$ will be zero:

$$a_1^1 = a_1^0 - \frac{2\eta}{n} \sum_{i=1}^{n} ((\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) * \frac{\delta}{\delta a_1} (a_1 \sigma(w_1^{(0)^T} x_i)))$$

$$a_1^1 = a_1^0 - \frac{2\eta}{n} \sum_{i=1}^{n} ((\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) * (\sigma(w_1^{(0)^T} x_i)))$$

For $a_2^1$, we will get an equivalent result:

$$a_2^1 = a_2^0 - \frac{2\eta}{n} \sum_{i=1}^{n} ((\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) * (\sigma(w_2^{(0)^T} x_i)))$$

Since $a_1^0 = a_2^0$ and $w_1^0 = w_2^0$:

$$a_2^1 = a_1^0 - \frac{2\eta}{n} \sum_{i=1}^{n} ((\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) * (\sigma(w_1^{(0)^T} x_i)))$$

Thus, it is proved that $a_1^1 = a_2^1$.

2. Induction step. Suppose that at timestep k, $a_1^k = a_2^k$ and $w_1^k = w_2^k$. Using the same derivation step from the base case we we will get:

$$a_1^{k+1} = a_1^k - \frac{2\eta}{n} \sum_{i=1}^{n} ((\sum_{j=1}^{m} a_j^k \sigma(w_j^{(k)^T} x_i) - y_i) * (\sigma(w_1^{(k)^T} x_i)))$$

$$a_2^{k+1} = a_2^k - \frac{2\eta}{n}\sum_{i=1}^{n}((\sum_{j=1}^{m} a_j^k \sigma(w_j^{(k)^T} x_i) - y_i) * (\sigma(w_2^{(k)^T} x_i)))$$

Since $a_1^k = a_2^k$ and $w_1^k = w_2^k$:

$$a_2^{k+1} = a_1^k - \frac{2\eta}{n}\sum_{i=1}^{n}((\sum_{j=1}^{m} a_j^k \sigma(w_j^{(k)^T} x_i) - y_i) * (\sigma(w_1^{(k)^T} x_i)))$$

Thus, $a_1^{k+1} = a_2^{k+1}$ when $a_1^k = a_2^k$ and $w_1^k = w_2^k$. Hence, it is proved that $a_1^t = a_2^t$ for all $t \geq 1$.

Next, I want to prove that $w_1^t = w_2^t$ for all $t \geq 1$. I am going to use induction to prove the statement. As usual, divide into two cases:

1. Base case, i.e. $t = 1$.

$$w_1^1 = w_1^0 - \eta \nabla w_1 \hat{R}(\theta^0)$$

$$w_1^1 = w_1^0 - \eta \nabla w_1 (\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i)^2)$$

Using chain rule:

$$w_1^1 = w_1^0 - \frac{2\eta}{n}\sum_{i=1}^{n}((\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) \nabla w_1 (\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i))$$

For the right-most part, all terms not involving $w_1 \rightarrow w_2, w_3, ..., w_m, a_2, a_3, ..., a_m$ and $y_1, y_2, ..., y_n$ will be zero:

$$w_1^1 = w_1^0 - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i) \nabla w_1 (a_1^{(0)} \sigma(w_1^{(0)^T} x_i)))$$

$$w_1^1 = w_1^0 - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i)(a_1^{(0)} \sigma(x_i)))$$

For $w_2^1$, we will get an equivalent result:

$$w_2^1 = w_2^0 - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i)(a_2^{(0)} \sigma(x_i)))$$

Since $w_1^0 = w_2^0$ and $a_1^0 = a_2^0$:

$$w_2^1 = w_1^0 - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(0)} \sigma(w_j^{(0)^T} x_i) - y_i)(a_1 \sigma(w_1^{(0)^T} x_i)))$$

Thus, it is proved that $w_1^1 = w_2^1$.

2. Induction step.

   Suppose that at time k, $w_1^k = w_2^k$ and $a_1^k = a_2^k$. Using the same derivation step from the base case we we will get:

$$w_1^{k+1} = w_1^k - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(k)}\sigma(w_j^{(k)^T}x_i) - y_i)(a_1^{(k)}\sigma(x_i)))$$

$$w_2^{k+1} = w_2^k - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(k)}\sigma(w_j^{(k)^T}x_i) - y_i)(a_2^{(k)}\sigma(x_i)))$$

   Since $w_1^k = w_2^k$ and $a_1^k = a_2^k$:

$$w_2^{k+1} = w_1^k - \frac{2\eta}{n}(\sum_{i=1}^{n}(\sum_{j=1}^{m} a_j^{(k)}\sigma(w_j^{(k)^T}x_i) - y_i)(a_1^{(k)}\sigma(x_i)))$$

   Thus, $w_1^{k+1} = w_2^{k+1}$ when $a_1^k = a_2^k$ and $w_1^k = w_2^k$. Hence, it is proved that $w_1^t = w_2^t$ for all $t \geq 1$.

Finally, from the two proofs above, it is proved that $a_1^t = a_2^t$ and $w_1^t = w_2^t$ for all $t \geq 1$.

b. The proofs below are divided into two parts:

   1. First, prove for the expected scale of the hidden unit value.

      Suppose that $y = x^T w_j^{(0)}$ so that:

$$E[\sigma(x^T w_j^{(0)})^2] = E[\sigma(y)^2]$$

      We can use integral to calculate the expectation:

$$E[\sigma(y)^2] = \int_{-\infty}^{\infty} \max(0, y)^2 p(y) dy$$

      Because $\max(0, y) \geq 0$, we can disregard the case when $y < 0$:

$$E[\sigma(y)^2] = \int_{0}^{\infty} y^2 p(y) dy$$

      Since $y^2$ is symmetric around zero and so is $p(y)$ (because the mean of the weights is zero and the distribution is normal), we can use full domain $(-\infty, \infty)$ and take half of it:

$$E[\sigma(y)^2] = \frac{1}{2}\int_{-\infty}^{\infty} y^2 p(y) dy \text{ ... Eq. 4.b.1}$$

Since $x$ is not random and $w_j^{(0)}$ has zero mean, consider the fact that:

$$E[y] = E[x^T w_j^{(0)}] = 0 \text{ ... Eq. 4.b.2}$$

Thus, going back to Eq. 4.b.1 with $E[y] = 0$:

$$E[\sigma(y)^2] = \frac{1}{2}\int_{-\infty}^{\infty}(y - E[y])^2 p(y)dy$$

We can now get rid of the integral and replace it with expectation:

$$E[\sigma(y)^2] = \frac{1}{2}E[(y - E[y])^2]$$

$$E[\sigma(y)^2] = \frac{1}{2}E[y^2 - 2yE[y] + E[y]^2]$$

$$E[\sigma(y)^2] = \frac{1}{2}(E[y^2] - 2E[y]E[y] + E[y]^2])$$

$$E[\sigma(y)^2] = \frac{1}{2}(E[y^2] - E[y]^2]) = \frac{1}{2}\text{var}(y)$$

Finally, we can substitute $y$ with the original value:

$$E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{2}\text{var}(x^T w_j^{(0)})$$

Again, because $x$ is not random, then the variance of the inner product would be:

$$E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{2}x^T \text{cov}(w_j^{(0)})x$$

Since $E[w_j^{(0)}] = 0$ (zero mean random vector), then the covariance would be:

$$E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{2}x^T E[(w_j^{(0)})^T w_j^{(0)}]x = \frac{1}{2}x^T E[(w_j^{(0)})^2]x$$

Using the fact that $E[X^2] = \text{var}(X) + E[X]^2$ and plugging the variance of $w_j^{(0)}$, finally we will have:

$$E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{2}x^T(\text{var}(w_j^{(0)}) + E[w_j^{(0)}]^2)x$$

$$E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{2}x^T(\frac{2}{d}I + 0)x$$

$$E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{d}x^T Ix = \frac{1}{d}x^T x$$

$$\boldsymbol{E[\sigma(x^T w_j^{(0)})^2] = \frac{1}{d}\sum_{i=1}^{d} x_i^2}$$

2. Second, prove for the expected scale of the output value. Using the fact that $E[X^2] = var(X) + E[X]^2$:

$$E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = \text{var}(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)})) + E[\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)})]^2$$

Since $a_j^{(0)}$ has zero mean $\rightarrow E[a_j^{(0)}] = 0$, then $E[X]^2$ is zero:

$$E[\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)})]^2 = E[m * E[a_j^{(0)}] * \sigma(x^T w_j^{(0)})]^2 = 0$$

Thus, we are only left with the variance:

$$E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = \text{var}(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))$$

Because for different $j$, $a_j^{(0)}\sigma(x^T w_j^{(0)})$ is uncorrelated to each other (covariance equals zero), then:

$$E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = \sum_{j=1}^{m} \text{var}(a_j^{(0)}\sigma(x^T w_j^{(0)}))$$

Moreover, since $a_j^{(0)}$ and $\sigma(x^T w_j^{(0)})$ are also uncorrelated, we can have:

$$var(XY) = var(X)var(Y) + var(X)E[Y]^2 + var(Y)E[X]^2$$

where $X = a_j^{(0)}$ and $Y = \sigma(x^T w_j^{(0)})$. But, we know that:

1. $E[X]^2 = E[a_j^{(0)}]^2 = 0^2 = 0$
2. $E[Y]^2 = E[\sigma(x^T w_j^{(0)})]^2 = \sigma(E[(x^T w_j^{(0)})])^2$. From the previous part Eq. 4.b.2, we know that $E[(x^T w_j^{(0)})] = 0$ so $E[Y]^2 = \sigma(0)^2 = 0$

Thus, we are only left with $var(XY)$:

$$E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = \sum_{j=1}^{m}(\text{var}(a_j^{(0)}) * var(\sigma(x^T w_j^{(0)})))$$

Using some results/observations from the previous proof:

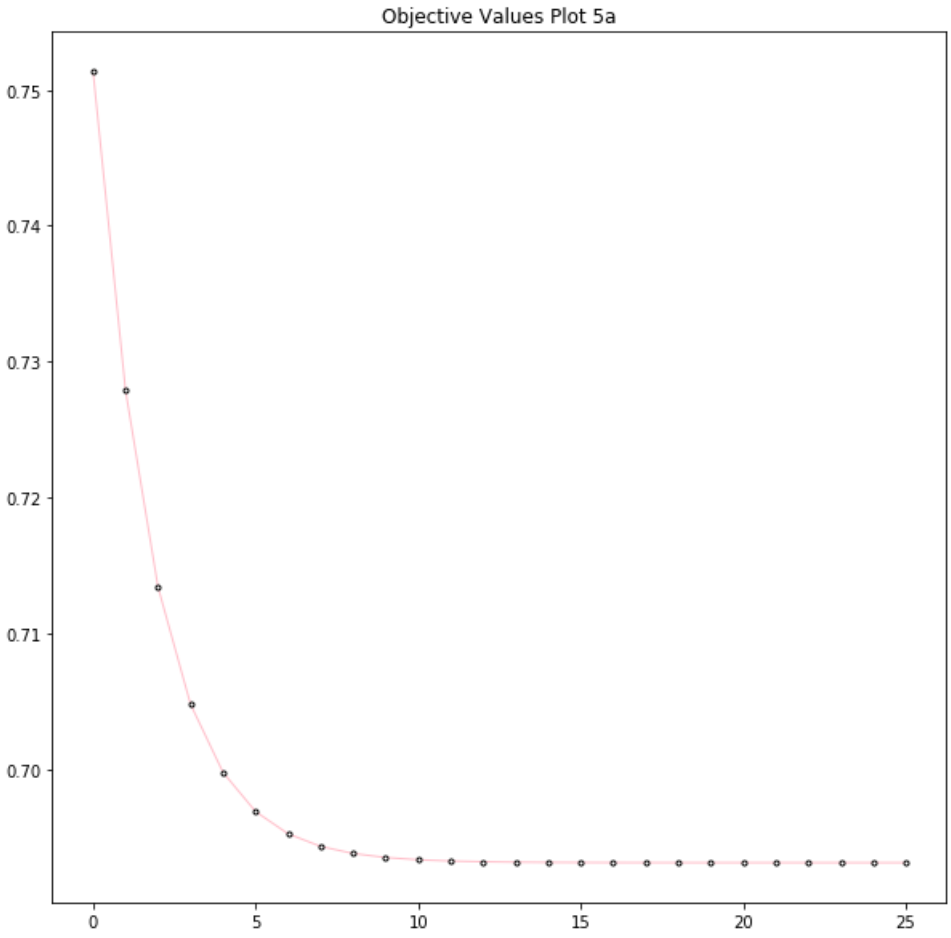
$$E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = \sum_{j=1}^{m}(\frac{1}{m} * \frac{1}{2} var(x^T w_j^{(0)}))$$

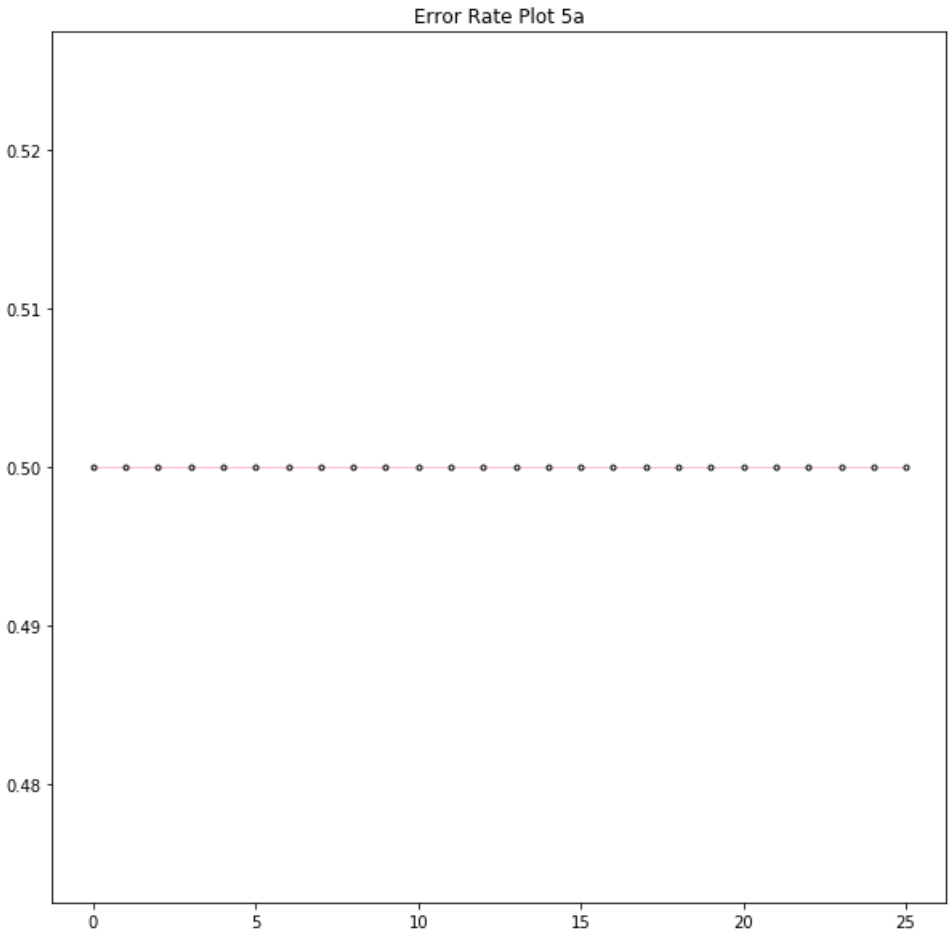
$$E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = \sum_{j=1}^{m}(\frac{1}{m} * \frac{1}{2} * \frac{2}{d}\sum_{i=1}^{d} x_i^2) = \sum_{j=1}^{m}(\frac{1}{m * d}\sum_{i=1}^{d} x_i^2)$$

$$\boldsymbol{E[(\sum_{j=1}^{m} a_j^{(0)}\sigma(x^T w_j^{(0)}))^2] = m(\frac{1}{m * d}\sum_{i=1}^{d} x_i^2) = \frac{1}{d}\sum_{i=1}^{d} x_i^2}$$

# Problem 5

a. Below are the objective value and error rate plot respectively:

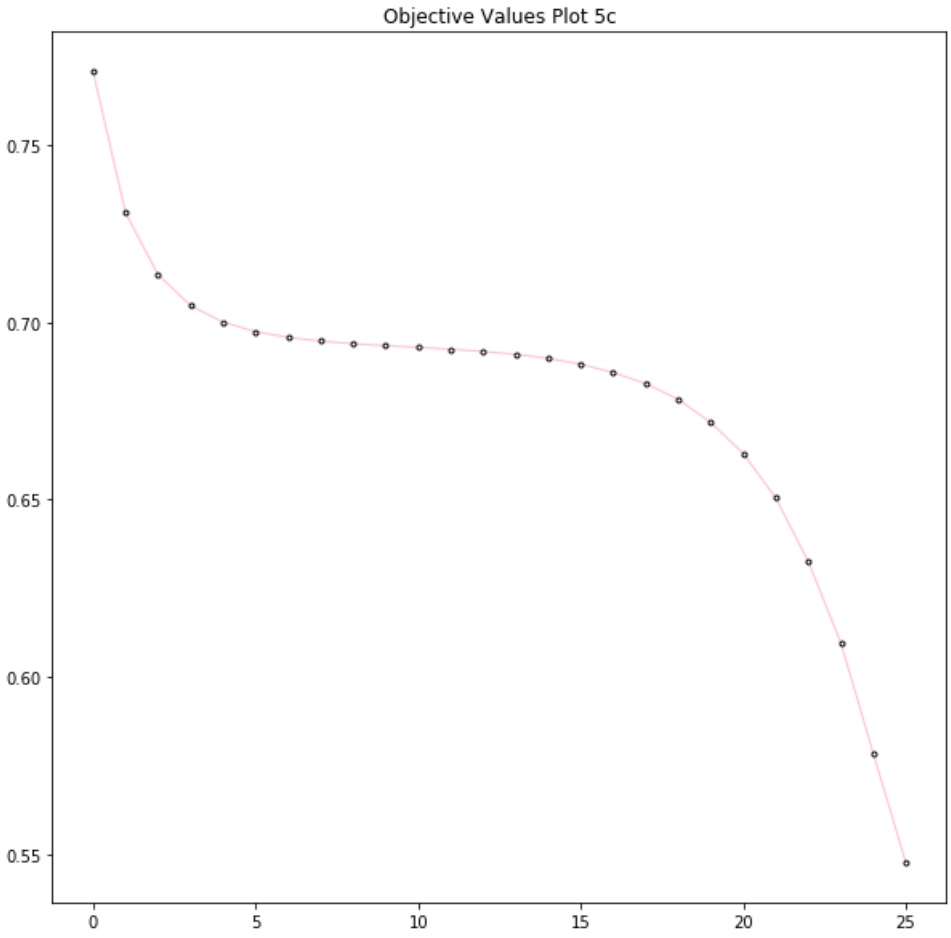


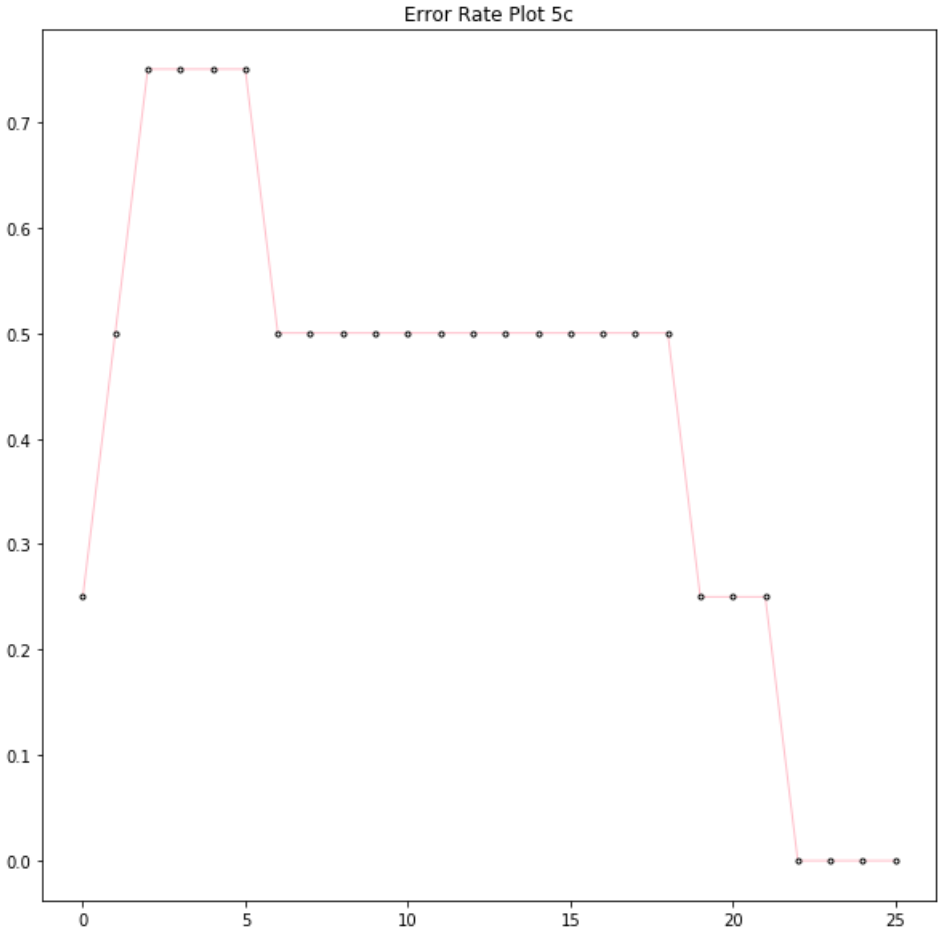



b. Below are the initial and final values:

1. Initial (Objective Value, Error Rate): (0.7513735294342041, 0.5)
2. Final (Objective Value, Error Rate): (0.6931471824645996, 0.5)

c. Below are the objective value and error rate plot respectively:

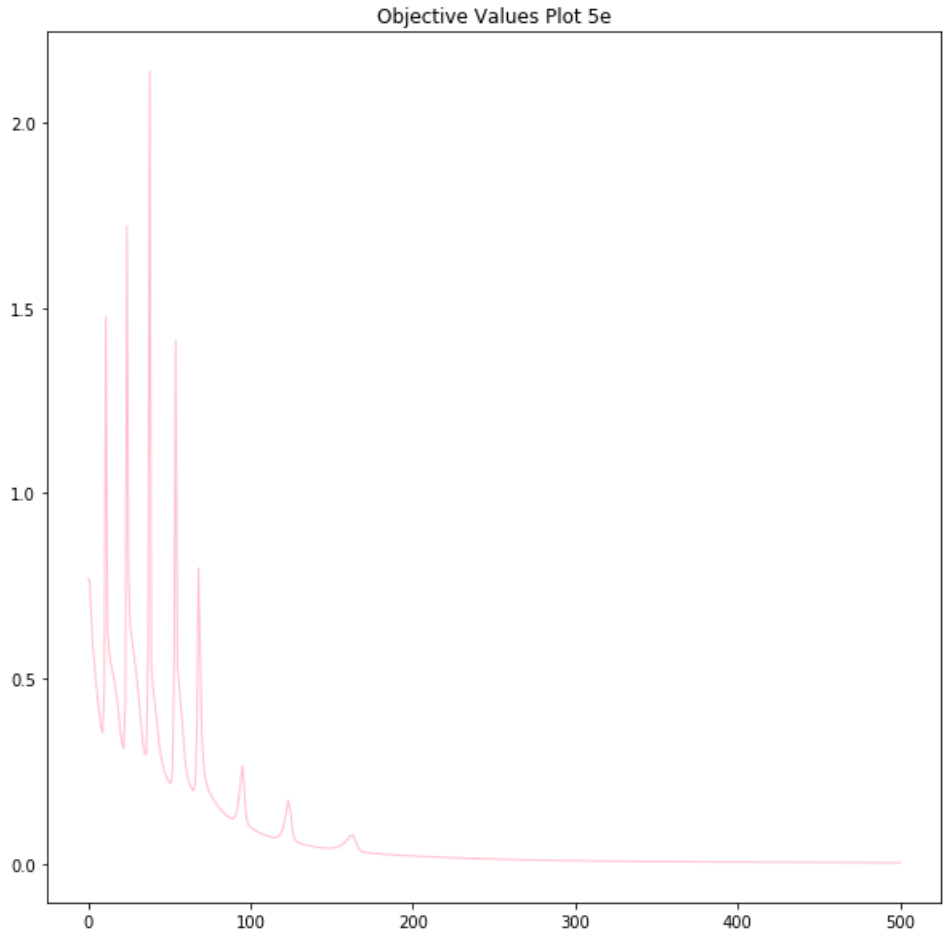


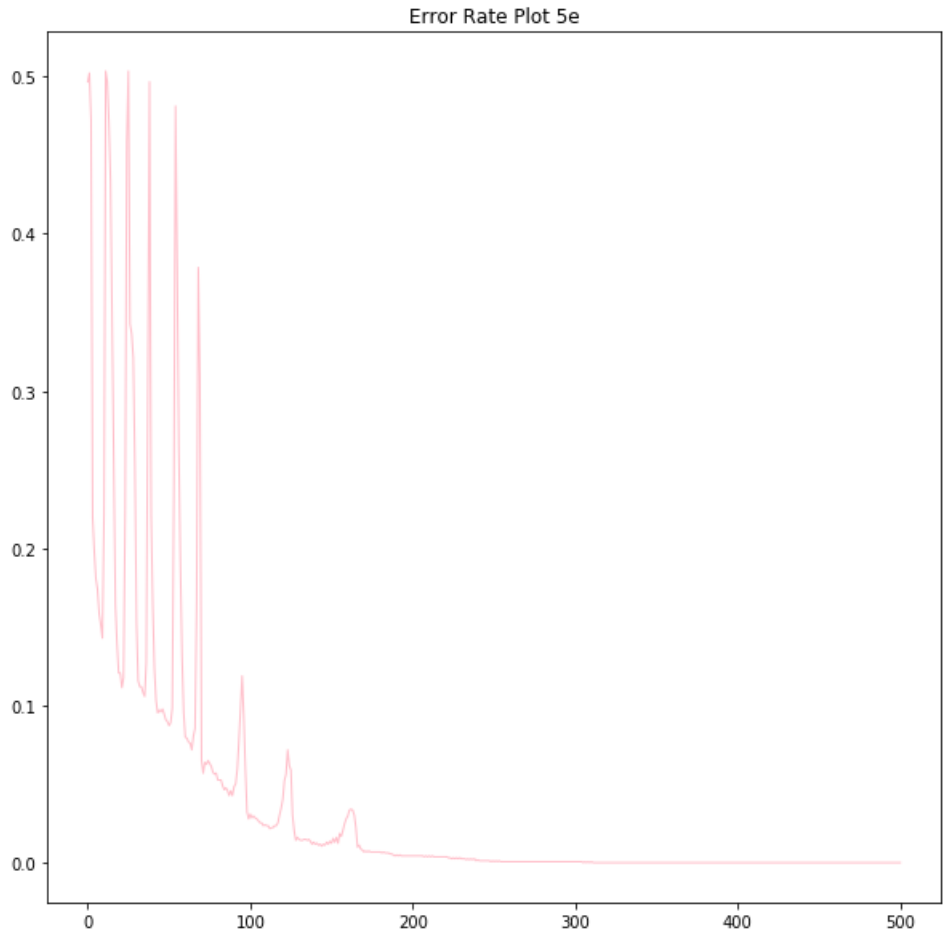



d. Below are the initial and final values:

1. Initial (Objective Value, Error Rate): (0.7708840370178223, 0.25)
2. Final (Objective Value, Error Rate): (0.5475983023643494, 0.0)

e. Below are the objective value and error rate plot respectively:



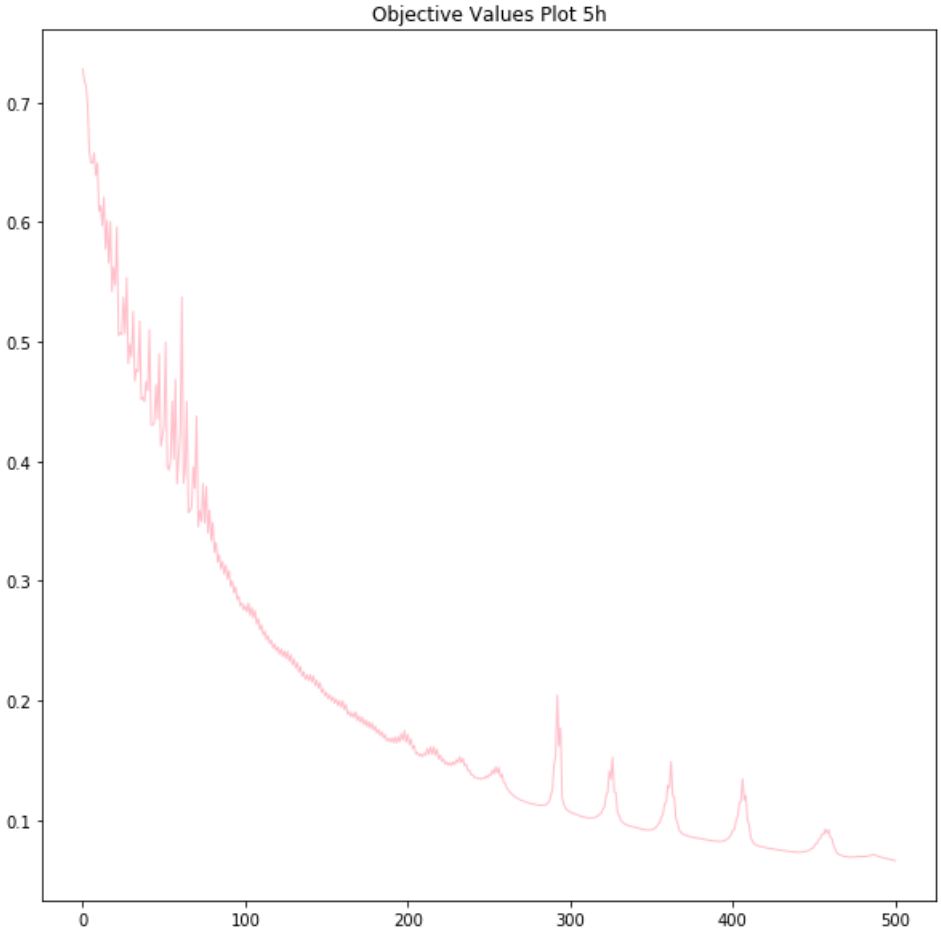


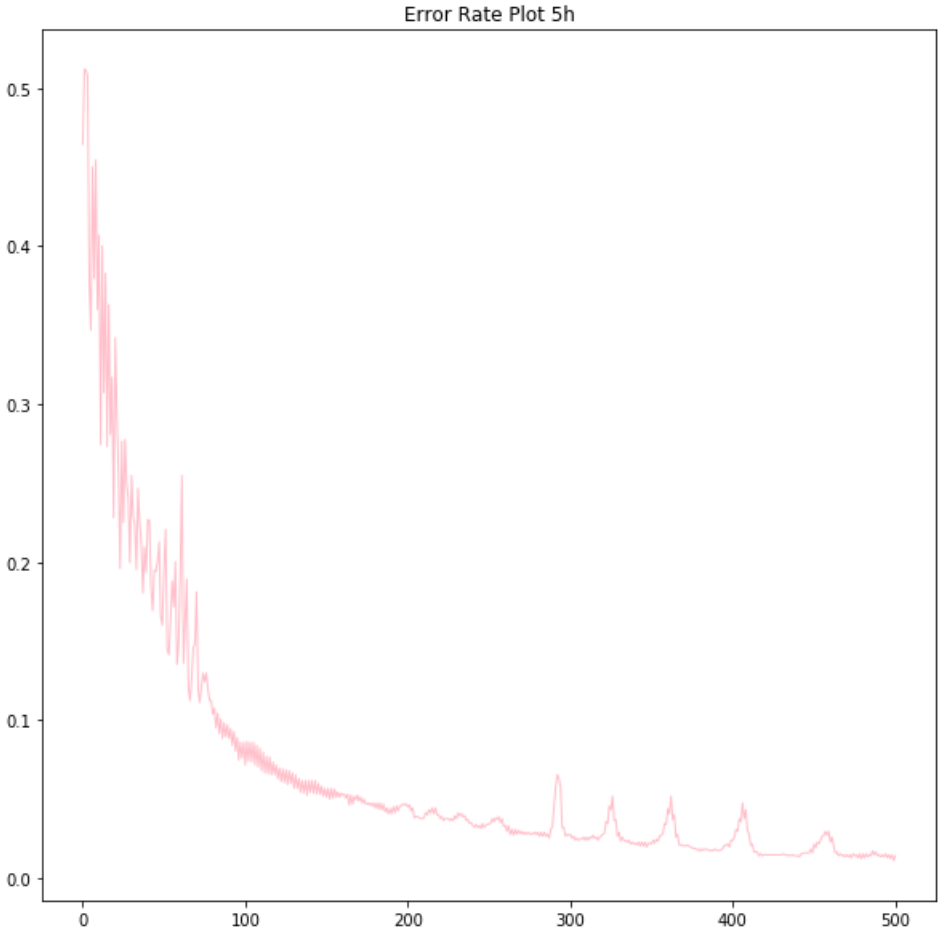

f. Below are the initial and final values:

1. Initial (Objective Value, Error Rate): (0.7686847448348999, 0.49659863114356995)
2. Final (Objective Value, Error Rate): (0.0024712553713470697, 0.0)

g. Final test error rate: **0.02777777798473835**

h. Below are the objective value and error rate plot respectively:






i. Below are the initial and final values:

1. Initial (Objective Value, Error Rate): (0.7276044487953186, 0.46444031596183777)
2. Final (Objective Value, Error Rate): (0.06854907423257828, 0.014223871752619743)

j. Final test error rate: **0.02222222276031971**

# Problem 6

a. The error rate of $\hat{f}_{MLE}$ on the same data is **0.1933**.

b. The training error rates are:

1. Training error rate on Cyan subpopulation: **0.168**.
2. Training error rate on Red subpopulation: **0.32**.

c. The classification discrepancy is **0.368** where:

1. $P_n(\hat{f}_{MLE}(X) = +1 | A = Cyan) = 0.548$
2. $P_n(\hat{f}_{MLE}(X) = +1 | A = Red) = 0.18$

d. The FNR discrepancy is **0.26** where:

1. $P_n(\hat{f}_{MLE}(X) = -1 | Y = +1, A = Cyan) = 0.06$
2. $P_n(\hat{f}_{MLE}(X) = -1 | Y = +1, A = Red) = 0.32$

e. My approach is basically following the hint from Part (b) description which says that "SAT" feature is not useful for the Red subpopulation. So, here are what I did:

1. I compute the mean of "SAT" values from all the training data (600 data), which is equal to 6.7244563.
2. I use this mean value to replace all "SAT" values in the Red subpopulation (100 data). In other words, now all the training data in Red subpopulation have "SAT" value equals 6.7244563.
3. When computing error rate, I use the original training data, i.e. the data which "SAT" values haven't been altered.

The reason that I didn't just drop the SAT field is that it would lower the performance on Cyan subpopulation. Also, I tried to replace the "SAT" values in the Red subpopulation with the mean of "SAT" values within the Red subpopulation itself. I thought initially that this would be better. But empirically, doing so didn't lower the error rate and FNR discrepancy. Performance achieved:

a. Training error rate: **0.11166** where:

1. Training error rate on Cyan subpopulation: **0.126**.
2. Training error rate on Red subpopulation: **0.04**.

b. FNR discrepancy: **0.018** where:

1. $P_n(\hat{f}_{MLE}(X) = -1 | Y = +1, A = Cyan) = 0.058$
2. $P_n(\hat{f}_{MLE}(X) = -1 | Y = +1, A = Red) = 0.04$

We can see that the error rate is cut by approx. 9% and the false negative rate for Red subpopulation falls sharply from 32% to 4%. Thus, it leaves the false negative discrepancy rate to only 1.8% from previously 26%.