ENME808E
HOMEWORK – 1

- Rachith Prakash

1. [d]
Explanation:
a) The first example talks about a case wherein the features of data are given to us. Exact coin specifications are given, thus there is nothing to learn, hence, no Machine Learning required.
b) Here, a set of labeled coins are given. This means we have input data that needs to be classified and we don't know how to classify them (features). This is where Machine Learning comes into picture. It is supervised learning – training data with labels i.e. output information for training data is present.
c) In this case, we don't tell the learning model whether the answer is right or wrong, but we give it a grade based on each output. The algorithm learns to get a good grade for every output. This type of learning is called Reinforcement Learning.

2. [a]
Explanation:
(i) and (iii) are defined by mathematical equations. Hence, no learning is required. However if (i) had included problem of vision, i.e. classifying based on image processing, it would have been a ML problem. Now, (ii) and (iv) cannot be defined by any equation explicitly. Hence, ML can be applied to solve these problems. This is assuming we have data and there exists a pattern.

3. [d]
Explanation:
Bayes' theorem!
Let A be an event of picking first ball as black, B be an event of picking second ball as black.
Now, from Law of Total Probability, P(A) = choosing bag 1*choosing black + choosing bag 2 * choosing black = 0.5*0.5 + 0.5*1 = 0.75. P(A) = 0.75. Now, given that A has occurred, we know we have a black ball. But, it could have been from either of the bag as both of them contains black balls. Thus, the probability of (second ball being black and first ball being black) is 0.5, according to Bayes theorem, P(B/A) = P(A and B)/P(A). Therefore, P(B/A) = 0.5/0.75 = 2/3.

4. [b]
Explanation:
The probability of v=0 is the probability of getting all marbles as green. Probability of getting one green marble is 0.45. Thus, probability of getting all green marbles are (0.45)^10 = 3.405e-4 in one sample.

5. [c]
Explanation:
P(v=0) = 0.0003405. Therefore, P(v!=0) = 1-0.000340506 = 0.999659494.
P(getting atleast one of sample with v=0) = 1 - P(not getting any sample as v=0)
P(v!=0) for 1000 experiments are independent events. Thus, for 1000 experiments, P(v!=0) = (0.999659494)^1000 = 0.711368802.
Thus, P(getting atleast one of sample with v=0) = 1- 0.711368802 = 0.288631198

6. [e]
Explanation:
All of them gives 3(1) + 2(3) + 1(3) + 0(1) = 12 as their scores.

Option/Score	3	2	1	0
a	111	011, 101, 110	001, 100, 010	000
b	000	001, 100, 010	011, 101, 110	111
c	001	000, 011, 101	010, 100, 111	110
d	110	010, 100, 111	000, 011, 101	001

NOTE: python file attached for the following questions:

Parameters chosen are as follows:
times PLA algorithm is run = 1000
random points on which classification is tested to calculate probability of misclassification = 1000
dimension of input = 2
No. of datapoints taken for training, N = 10 and 100

7. [b], average around 11

8. [c], average around 0.18

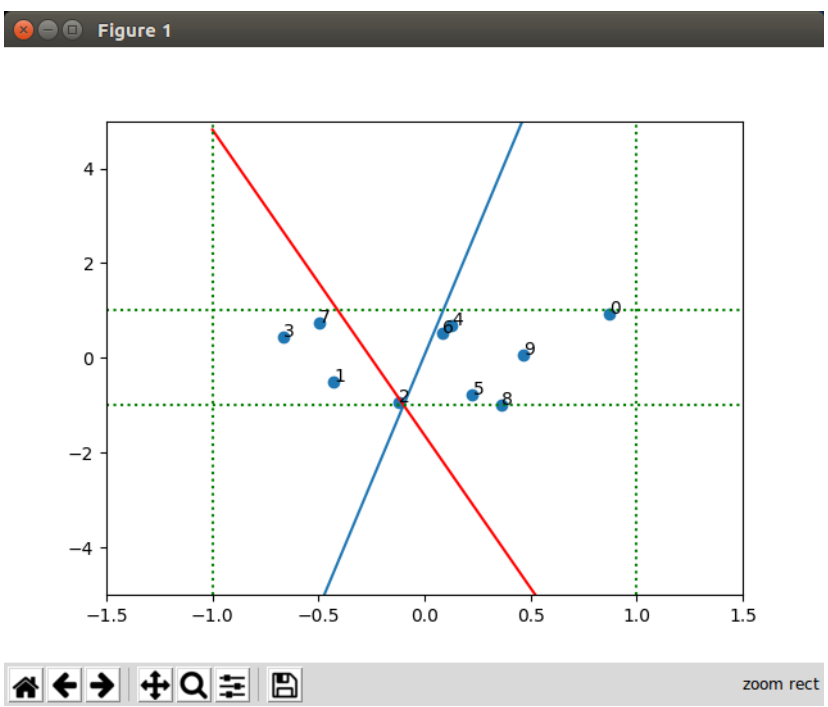
9. [b], average around 100

10. [b], average around 0.02

Some figures obtained are as shown below:

Blue line is f(x), red line is g(x). All cases are 0 misclassification scenarios.

N = 10



N = 10

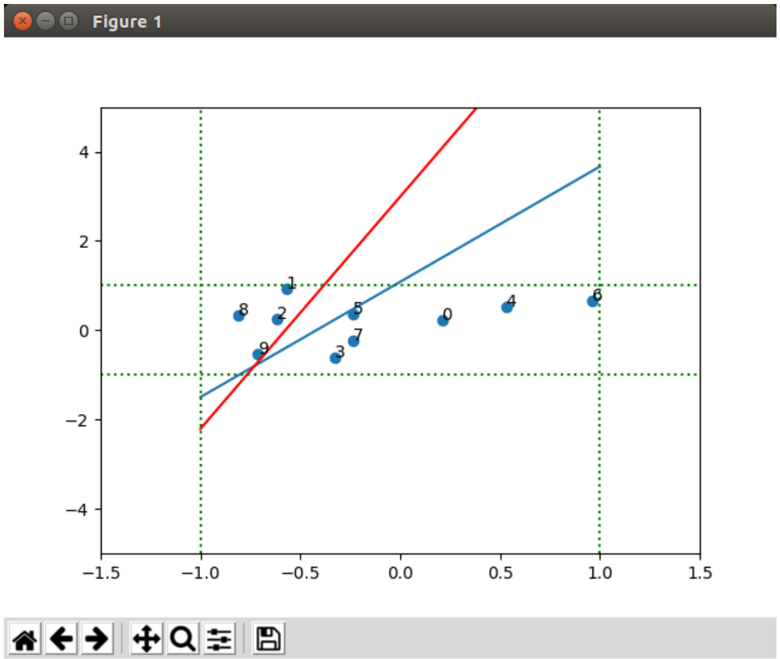


Figure 1
4
2
0
−2
−4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
0
1
2
3
4
5
6
7
8
9

N = 10

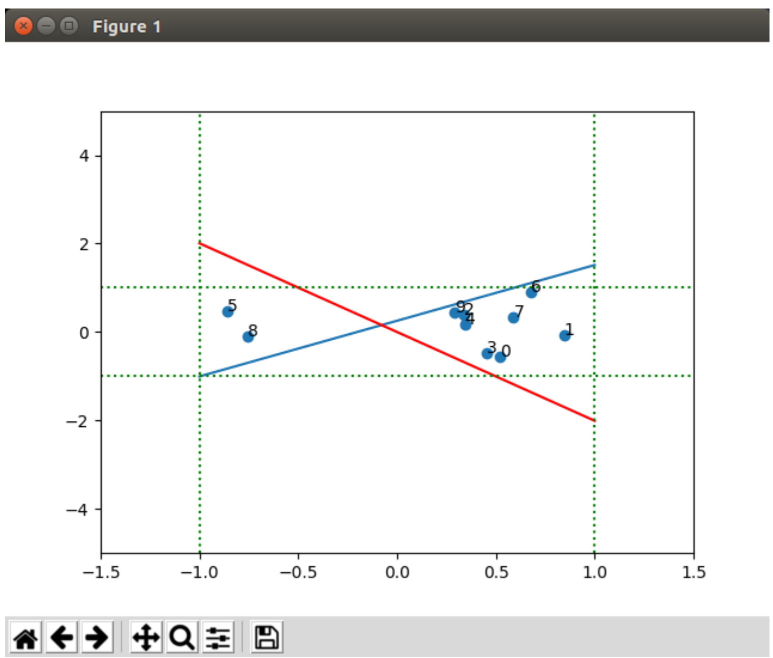
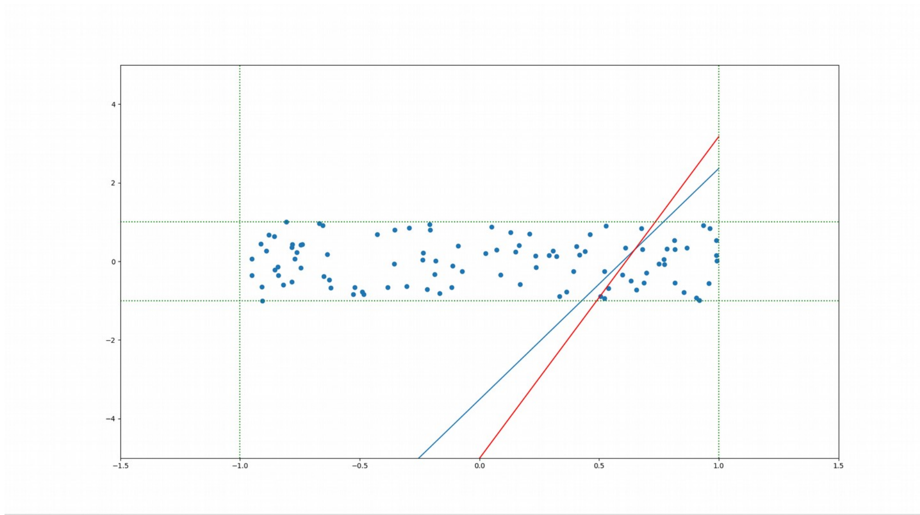
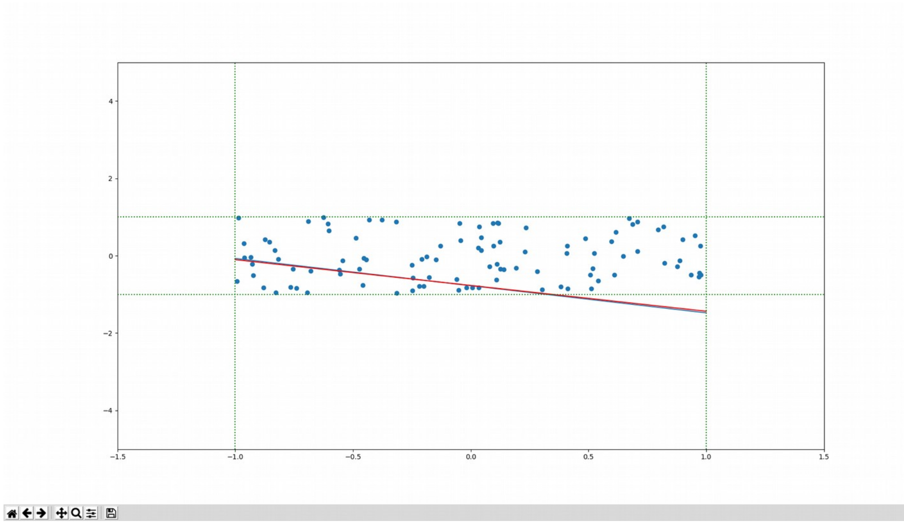


Figure 1
4
2
0
−2
−4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
9
2
4
6
7
1
3
0
5
8

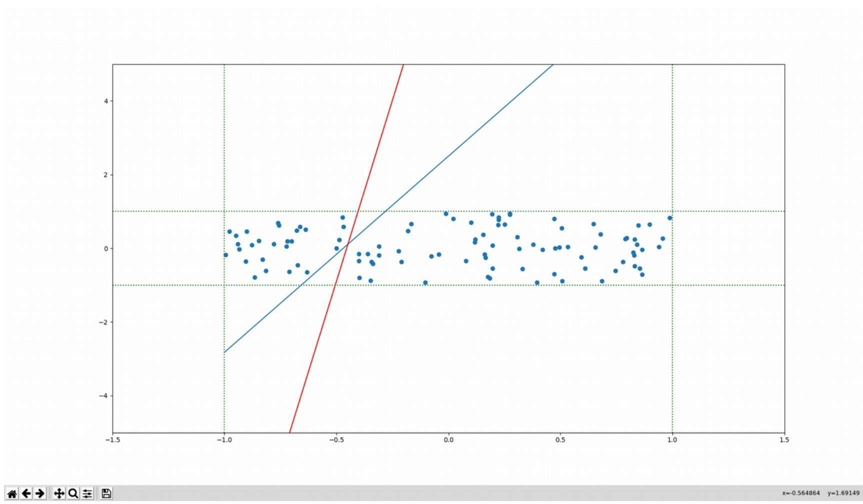
N = 100



N = 100



N = 100



APPENDIX:

Python code for **PLA** algorithm,

```python
import os
import sys
import matplotlib.pyplot as plt
import numpy as np

# Number of input points
N = 100
count = 0
cnt = 0

# Running PLA algorithm 1000times to get average
for _ in range(1000):

    p1 = [np.random.uniform(-1,1), np.random.uniform(-1,1)]
    p2 = [np.random.uniform(-1,1), np.random.uniform(-1,1)]

    # Find the equation of line
    a = p1[1]-p2[1]
    b = p2[0]-p1[0]
    d = -(a*p1[0]+b*p1[1])

    # Calculating slope and intercept
    m = -a/b
    c = -d/b

    y = []
    x = []

    for _ in range(N):

        xn = np.array([np.random.uniform(-1,1), np.random.uniform(-1,1)])
        x.append(xn)

        if m*xn[0]+c > xn[1]:
            y.append(1)
        else:
            y.append(-1)

    x = np.array(x).T
    y = np.array(y).reshape(1,N)

    fig, ax = plt.subplots()
    plt.plot(np.linspace(-1,1),m*np.linspace(-1,1)+c)
    ax.scatter(x[0],x[1])
    plt.xlim(-1.5,1.5)
    plt.ylim(-5,5)
    # for i in range(N):
        # ax.annotate(i, (x[0,i], x[1,i]))

    # Define weight vector according to size of d, x0 = 1, w0 = bias
    w = np.zeros((1,3))
    x = np.insert(x,0,np.ones((1,N)),axis=0)

    ite = 0
    val = False
    while not val:
        y_hat = np.matmul(w,x).reshape(1,N)

        classify = [1 if y_hat[0,i]>0 else -1 for i in range(N)]
```

```
        misclassified = [1 if y[0,i]!=classify[i] else 0 for i in range(N)]

        ind = [i for i in range(N) if misclassified[i]==1]

        if not len(ind):
            val = True
            break

        rn = np.random.randint(0,len(ind))

        w = w + x[:,ind[rn]] * y[0,ind[rn]]

        slope = -w[0,1]/w[0,2]
        intercept = -w[0,0]/w[0,2]

        ite += 1

    count += ite

    # Approximating the probability of misclassification on a random point
    # 1000 samples taken into consideration
    for _ in range(1000):
        p = [np.random.uniform(-1,1), np.random.uniform(-1,1)]
        if m*p[0]+c > p[1]:
            f = 1
        else:
            f = 0

        if slope*p[0]+intercept > p[1]:
            g = 1
        else:
            g = 0

        if f!=g:
            cnt += 1


    plt.plot(np.linspace(-1,1),slope*np.linspace(-1,1)+intercept, 'r')
    plt.axhline(y=1, color='g', linestyle=':')
    plt.axhline(y=-1, color='g', linestyle=':')
    plt.axvline(x=-1, color='g', linestyle=':')
    plt.axvline(x=1, color='g', linestyle=':')
    plt.show()

print "avg iterations: ", count/1000.0
print "avg probability of miscl: ", cnt/1000000.0
```

Problem - 1.3 :

w^* → optimal weight vector [we get this after converging]

$w(0) = 0$. , $w(t)$ is the weight vector at t^{th} iteration.

a) Given, $\rho = \min_{1 \le n \le N} y_n(w^{*T} x_n)$, s.t $\rho > 0$.

w^* is the optimal weight. This means this seperates the data correctly ∴ Classification is correct for $(w^{*T} x_n)$.

Since the PLA algorithm's classification algorithm function is

$$h(x) = \text{sig}(w^T x) : \begin{cases} +1 & w^T x > 0 \\ -1 & w^T x < 0 \end{cases}$$

for optimal weights w^*, $y_n = \pm 1$ is same as $h(x)$
i.e $h(x) > 0 \Rightarrow w^T x > 0$ and $h(x) < 0 \Rightarrow w^T x < 0$.

$$\therefore \quad y_n(w^{*T} x) > 0 \quad \forall n \in N$$

Case 1: $y_n = +1$, $w^{*T} x > 0 \Rightarrow y_n(w^{*T} x) > 0$
Case 2: $y_n = -1$, $w^{*T} x < 0 \Rightarrow y_n(w^{*T} x) > 0$

Thus, $\rho > 0$ //

b) The PLA algorithm is based on updating weight vector as follows:

$$w(t) = w(t-1) + y_n x_n$$

Taking transpose throughout the above equation, gives

$$w^T(t) = w^T(t-1) + y_n x_n^T$$

Scanned by CamScanner

Now, multiplying w^* vector throughout the equation gives

$$w^T(t)\,w^* = w^T(t-1)\,w^* + y_n \underbrace{x_n^T w^*}$$

$\langle x_n \cdot w^* \rangle = \langle w^* \cdot x_n \rangle$

$$\therefore\ w^T(t)\,w^* = w^T(t-1)\,w^* + y_n\, w^{*T} x_n \quad \text{—} \ 1$$

But, $\min_{1\le n\le N} y_n(w^{*T}x_n) = \rho$.

$\Rightarrow$ So, $\min_{1\le n\le N} y_n(w^{*T}x_n) \le y_n(w^{*T}x_n)$, $\forall n \in [1,N]$

$$\rho \le y_n(w^{*T}x_n) \quad - \ 2, \quad \forall n \in [1,N]$$

Using 2, 1 becomes, $\boxed{w^T(t)\,w^* \ge w^T(t-1)\,w^* + \rho}$,

For, $t=1$

$$w^T(1)\,w^* \ge w^T(0)\,w^* + \rho, \quad w(0) = 0$$

$$w^T(1)\,w^* \ge \rho.$$

For, $t=2$.

$$w^T(2)\,w^* \ge w^T(1)\,w^* + \rho.$$

$$\ge \rho + \rho$$

$$w^T(2)\,w^* \ge 2\rho$$

For, $t=3$,

$$w^T(3)\,w^* \ge w^T(2)\,w^* + \rho \ge 2\rho + \rho \ge 3\rho.$$

For, $t=t$, $w^T(t)\,w^* \ge w^T(t-1)\,w^* + \rho.$

$$\ge (t-1)\rho + \rho.$$

$$\boxed{w^T(t)\,w^* \ge t\rho} \quad - \ 6$$

Scanned by CamScanner

c) W.K.T $w(t) = w(t-1) + y_n(t-1)\, x_n(t-1)$ ↱ y_n corresponding to x_n

Here, $x_n(t-1)$ is misclassified by $w(t-1)$. Hence, $w(t)$ exists

$x_n(t-1) \to x(t-1)$ for simplicity.

Taking norm both sides,

$$\|w(t)\| = \|w(t-1) + y(t-1)\, x(t-1)\|$$
$$\Rightarrow \|w(t)\|^2 = \|w(t-1) + y(t-1)\, x(t-1)\|^2$$

W.K.T

$$\|a+b\|^2 = (a+b)^T(a+b)$$
$$= a^Ta + b^Tb + b^Ta + a^Tb$$
$$= \|a\|^2 + \|b\|^2 + 2a^Tb$$

$$\therefore \|w(t)\|^2 = \|w(t-1)\|^2 + \|\underbrace{y(t-1)}_{\pm 1}\, x(t-1)\|^2 + 2\, w^T(t-1)\, y(t-1)\, x(t-1)$$

$$\|w(t)\|^2 = \|w(t-1)\|^2 + \|x(t-1)\|^2 + 2\underbrace{y(t-1)\, w^T(t-1)\, x(t-1)}_{\leq 0}$$

Now, since $x(t-1)$ is misclassified by $w(t-1)$,

we have, $y(t-1)\left[w^T(t-1)\, x(t-1)\right] \leq 0$

Using this,

$$\underline{\|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2}$$

d). Let, $t = 1$.

$$\therefore \|w(1)\|^2 \leq \|w(0)\|^2 + \|x(0)\|^2$$
$$\|w(1)\|^2 \leq \|x(0)\|^2 \quad -(3)$$

$t = 2$

$$\|w(2)\|^2 \leq \|w(1)\|^2 + \|x(1)\|^2$$
$$\leq \|x(0)\|^2 + \|x(1)\|^2 \quad -(4)$$

Scanned by CamScanner

But, $R = \max\limits_{1 \le n \le N} \|x_n\|$

this means $\|x_n\| \le R, \ \forall n \in [1, N]$

$$\Rightarrow \|x_n\|^2 \le R^2, \ \forall n \in [1, N]$$

$$\therefore \|x(0)\|^2 \le R^2$$

$$\|x(1)\|^2 \le R^2$$

$$\vdots$$

$$\therefore (3) \Rightarrow \|w(1)\|^2 \le \|x(0)\|^2$$

$$\le R^2$$

$$(4) \Rightarrow \|w(2)\|^2 \le \|x(0)\|^2 + \|x(1)\|^2$$

$$\le R^2 + R^2$$

$$\le 2R^2$$

$$\vdots$$

$$\boxed{\|w(t)\|^2 \le tR^2} \quad \text{— (5)}$$

e) Take (6) $\rightarrow$ $w^T(t)\, w^* \geq t\rho$.

Divide by $\|w(t)\|$ as $\|w(t)\| \geq 0$, $\|w(t)\| \neq 0$ assuming $t>0$.

$$\therefore \quad \frac{w^T(t)\, w^*}{\|w(t)\|} \geq \frac{t\rho}{\|w(t)\|}, \quad t>0$$

But, $\|w(t)\|^2 \leq tR^2 \Rightarrow \|w(t)\| \leq \sqrt{t}\, R$

$$\therefore \frac{w^T(t) w^*}{\|w(t)\|} \geq \frac{t\rho}{\sqrt{t}R}$$ [Inequality still holds :)]

$$\boxed{\frac{w^T(t)\, w^*}{\|w(t)\|} \geq \frac{\sqrt{t}\,\rho}{R}} \quad -$$

Changing sides, $$\sqrt{t} \leq \frac{R}{\rho} \frac{w^T(t)\, w^*}{\|w(t)\|}$$

$$\leq \frac{R}{\rho} \frac{w^T(t)\, w^*\, \|w^*\|}{\|w(t)\|\, \|w^*\|}$$

$$\leq \frac{R}{\rho} \|w^*\| \underbrace{\frac{w^T(t)}{\|w(t)\|}}_{\text{unit vector}} \underbrace{\frac{w^*}{\|w^*\|}}_{\text{unit vector}} \quad - (7)$$

$$\Rightarrow \frac{w^T(t)}{\|w(t)\|} \frac{w^*}{\|w^*\|} = \hat{w}^T(t)\, \hat{w}^*$$

$$= \langle \hat{w}(t) \cdot \hat{w}^* \rangle$$

$= \cos\theta$, θ angle b/w $w(t)$ and w^*.

$-1 \le \cos\theta \le 1 \implies \cos^2\theta \le 1$

Squaring both sides of (7)

$$t^2 \le \frac{R^2}{\rho^2} \|w^*\|^2 \underbrace{\cos^2\theta}_{\le 1}$$

$$\therefore \boxed{t^2 \le \frac{R^2}{\rho^2} \|w^*\|^2}$$

<u>NOTE</u>: Part (b) and (d) can also be proved by Induction -

Part (b): Prove: $w^T(t)\, w^* \ge t\rho$

Assume: it is true for $t = k$

$$w^T(k) w^* \ge k\rho. \quad — (8)$$

For, $t=1$, $w^T(t)\, w^* \ge t\rho$

$w^T(1)\, w^* \ge \rho$

this is true as $w^T(t)\, w^* \ge w^T(t-1) w^* + \rho$

$t=1$, $w^T(1)\, w^* \ge \rho$

$\therefore$ $|||^{ly}$ for $t=k$

$w^T(k)\, w^* \ge w^T(k-1)\, w^* + \rho$

Assuming it is true for $t = k-1$,

$$w^T(k-1)\, w^* \ge (k-1)\rho.$$

$\therefore \quad w^T(k) w^* \geq (k-1)\rho + \rho$

$w^T(k)\ w^* \geq k\rho$

$\therefore$ the hypothesis is true for $t = 1, 2, \ldots, k-1, k, k+1 \ldots$

$\therefore$ It is true for $\forall t$.

(d) Prove: $\|w(t)\|^2 \leq tR^2$, $R = \max_{1 \leq n \leq N} \|x_n\|$

for $t=1$, $\|w(1)\|^2 \leq R^2$. This is true from the equation $\|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$

$$\downarrow$$

$$\|w(1)\|^2 \leq \|w(0)\|^2 + \|x(0)\|^2$$
$$\leq \|x(0)\|^2$$
$$\|w(1)\|^2 \leq R^2$$

Since, $R = \max_{1 \leq n \leq N} \|x_n\|$

$\|x_n\| \leq R$, $\forall n \in [1, N]$

$\therefore$ ~~...~~,

Now, $t=2$, $\|w(2)\|^2 \leq 2R^2$. This is also true as

$$\|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$$
$$\|w(2)\|^2 \leq \|w(1)\|^2 + \|x(1)\|^2$$
$$\leq R^2 + R^2$$
$$\|w(2)\|^2 \leq 2R^2$$

Assume it is true for $t = k$

$\therefore \|w(k)\|^2 \leq kR^2$

Now for $t = k+1$

$$\|w(k+1)\|^2 \leq \|w(k)\|^2 + \|x(k)\|^2$$
$$\leq kR^2 + R^2$$
$$\leq (k+1)R^2$$

Hence it is also true for $t = k+1$.

$\therefore$ By induction, it is true $\forall t$.