CSE417: Introduction to Machine Learning
Problem Set 1

Problems

1. LFD Problem 1.3 2. Out of textbook 3. LFD Problem 1.7 4. LFD Problem 1.8

Q1: LFD Problem 1.3

(a)

Since w^* is the optimal set of weights, for all $1 \leq n \leq N$, $w^{*T}x_n$ must have same sign as y_n since a linear separation is achieved.
Therefore, for all $1 \leq n \leq N$, $y_n(w^{*T}x_n) > 0$, which suggests $min_{1 \leq n \leq N} y_n(w^{*T}x_n) > 0$, or $\rho > 0$

(b)

Given the update rule, $w(t+1) = w(t) + y(t)x(t)$, transpose both sides and multiply by w^*:

$$\begin{aligned}
w(t+1)^T &= w(t)^T + (y(t)x(t))^T \\
w(t+1)^T w^* &= w(t)^T w^* + (y(t)x(t))^T w^* \\
w(t+1)^T w^* &= w(t)^T w^* + y(t)(x(t)^T w^*) \rightarrow \text{ given y(t) is 1 * 1} \\
w(t+1)^T w^* &= w(t)^T w^* + y(t)(w^{*T} x(t)) \rightarrow \text{ given } w^{*T}x(t) \text{ is 1*1 so its symmetric}
\end{aligned}$$

And by the definition of $\rho = min_{1 \leq n \leq N} y_n(w^{*T}x_n)$, for all $1 \leq t \leq N$, we have $rho \leq y(t)(w^{*T}x(t)$.
Then we could conclude that:

$$w(t+1)^T w^* \geq w(t)^T w^* + \rho$$

Base Case:
t = 0, since we assumed w(0) = 0, we have:

$$w^T(0)w^* = 0 \geq 0\rho = 0$$

Induction:
for t = n, given that the inequality is valid:

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

and incorporate the inequality we deduced above:

$$\begin{aligned}
w^T(n+1)w^* &\geq w(n)^T w^* + \rho \\
w^T(n+1)w^* &\geq n\rho + \rho \\
w^T(n+1)w^* &\geq (n+1)\rho
\end{aligned}$$

so the inequality must also be valid for t = n + 1.
Thus proved.

(c)

Given the update rule, $w(t+1) = w(t) + y(t)x(t)$, transpose both sides and multiply by the original equation:

$$\begin{aligned} w(t+1) * w(t+1)^T &= (w(t) + (y(t)x(t)))(w(t)^T + (y(t)x(t))^T) \\ ||w(t+1)||^2 &= ||w(t)||^2 + 2y(t)x(t)w(t)^T + y(t)^2||x(t)||^2 \end{aligned}$$

Since for any y(t), we have $y(t)^2 = 1$, and $2y(t)x(t)w(t)^T < 0$ given y(t) represents a misclassified point, we have:

$$\begin{aligned} ||w(t+1)||^2 &\leq ||w(t)||^2 + y(t)^2||x(t)||^2 \\ ||w(t+1)||^2 &\leq ||w(t)||^2 + ||x(t)||^2 \end{aligned}$$

(d)

Proof:

Base case:
for t = 0, $||w(t)||^2 = 0 \leq 0 * R^2 = 0$

Induction:
Given that $||w(t)||^2 \leq tR^2$, we have:

$$\begin{aligned} ||w(t+1)||^2 &\leq ||w(t)||^2 + ||x(t-1)||^2 \\ ||w(t+1)||^2 &\leq ||w(t)||^2 + R^2 \\ ||w(t+1)||^2 &\leq ||w(t)||^2 + R^2 \\ ||w(t+1)||^2 &\leq ||w(t)||^2 + R^2 \\ ||w(t+1)||^2 &\leq tR^2 + R^2 \\ ||w(t+1)||^2 &\leq (t+1)R^2 \end{aligned}$$

□

(e)

Proof: Using the conclusion in part (b):

$$
\begin{aligned}
w^T(t)w^* &\geq (t)\rho \\
\frac{w^T(t)w^*}{||w(t)||} &\geq \frac{t\rho}{||w(t)||} \\
\frac{w^T(t)w^*}{||w(t)||} &\geq \frac{t\rho}{||w(t)||} \geq \frac{t\rho}{\sqrt{tR^2}} \rightarrow \text{given the conclusion in (d)} \\
\frac{w^T(t)w^*}{||w(t)||} &\geq \frac{t\rho}{\sqrt{tR^2}} \\
\frac{w^T(t)w^*}{||w(t)||} &\geq \frac{\sqrt{t}\rho}{R}
\end{aligned}
$$

Since $\frac{w^T(t)w^*}{||w(t)||||w^*||} = \cos\theta$ where θ represents the angle between $w^T(t)$ and w^*, we must have $\frac{w^T(t)w^*}{||w(t)||||w^*||} \leq 1$, then we could rewrite the inequality to be:

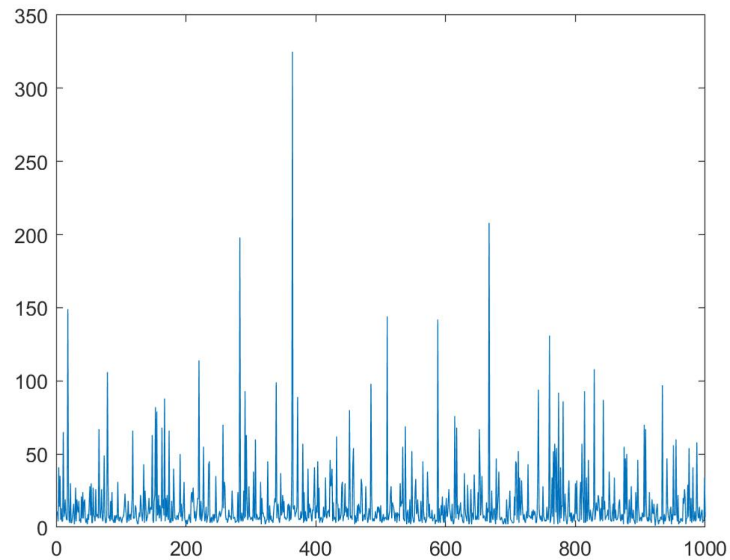
$$
\begin{aligned}
\sqrt{t} &\leq \frac{Rw^T(t)w^*}{||w(t)||\rho} \\
\sqrt{t}\frac{w^T(t)w^*}{||w(t)||||w^*||} &\leq \frac{Rw^T(t)w^*}{||w(t)||\rho} \rightarrow \text{Given That} \sqrt{t} \geq 0 \\
\frac{\sqrt{t}}{||w^*||} &\leq \frac{R}{\rho} \\
\sqrt{t} &\leq \frac{R||w^*||}{\rho} \\
t &\leq \frac{R^2||w^*||^2}{\rho^2}
\end{aligned}
$$

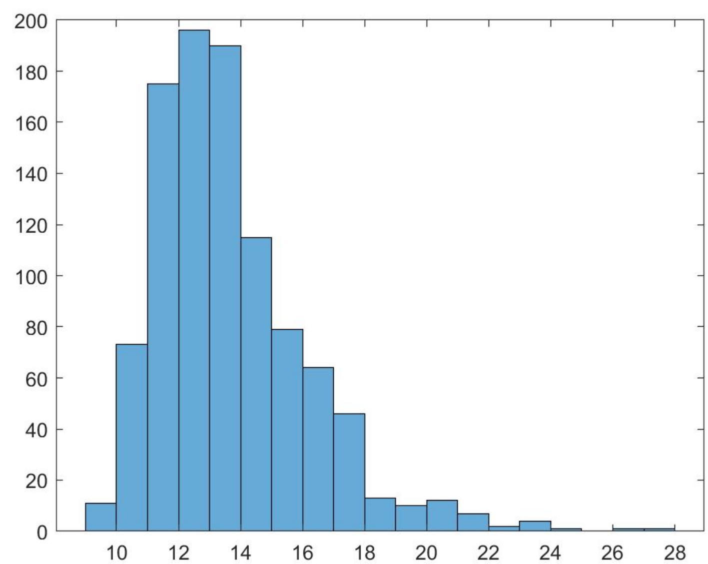
□

MATH493, PS1
Yushu Liu

Q2

The first plot is the number of operation for each iteration and the second plot is a histogram of log difference. Plot attached below:





Q3: LFD Problem 1.7

(a)

For $\mu = 0.05$

$P[0|10, 0.05] = 0.95^{10} = 0.5987$
1 Coin: P = 0.5987
10 Coins: P $= 1 - (1 - 0.95^{10})^{10} = 0.9998$
1000 Coins: P $= 1 - (1 - 0.95^{10})^{1000} = 1$
1000000 Coins: P $= 1 - (1 - 0.95^{10})^{1000000} = 1$

For $\mu = 0.8$

$P[0, |10, 0.8] = 0.2^{10} = 1.024e - 7$
1 Coin: P = 1.024e-7
10 Coins: P $= 1 - (1 - 0.2^{10})^{10} = 0.000001024$
1000 Coins: P $= 1 - (1 - 0.2^{10})^{1000} = .0001024$
1000000 Coins: P $= 1 - (1 - 0.2^{10})^{1000000} = 0.09733$

(b)

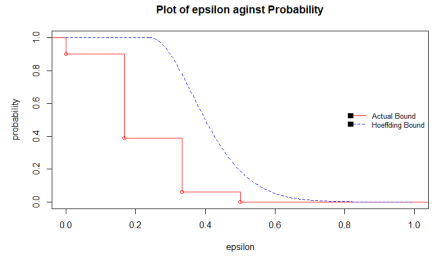
The result should be equivalent to P $= 1 - P(|v_1 - u_1| \leq \epsilon) * P(|v_2 - u_2| \leq \epsilon)$ for each ϵ, which produces a step function satisfy the following condition:

$$\begin{cases} P = 1, & \epsilon = 0 \\ P = 1 - P[3|6, 0.5]^2, & 0 < \epsilon \leq \frac{1}{6} \\ P = 1 - P[2, 3, 4|6, 0.5]^2, & \frac{1}{6} < \epsilon \leq \frac{1}{3} \\ P = 1 - P[1, 2, 3, 4, 5|6, 0.5]^2, & \frac{1}{3} < \epsilon \leq \frac{1}{2} \\ P = 0 & \epsilon > 0.5 \end{cases}$$

Using Hoeffding bound for 2 coins,

$$\begin{cases} P[max|vi - \mu_i| > \epsilon] \leq 1, & 2e^{-2N\epsilon^2} > 1 \\ P[max|vi - \mu_i| > \epsilon] \leq 1 - (1 - 2e^{-2N\epsilon^2})^2, & 2e^{-2N\epsilon^2} \leq 1 \end{cases}$$

Plot attached below:



Q4: LFD Problem 1.8

(a)

By definition, we have that $E(t) = \int_0^\infty \frac{tP(t)}{\alpha} dt$
Then for each $\alpha > 0$, we have:

$$\begin{aligned}
\frac{E(t)}{\alpha} &= \int_0^\infty \frac{tP(t)}{\alpha} dt \\
&= \int_\alpha^\infty \frac{tP(t)}{\alpha} + \int_0^\alpha \frac{tP(t)}{\alpha} dt \\
&\geq \int_\alpha^\infty \frac{tP(t)}{\alpha} dt \\
&\geq \int_\alpha^\infty \frac{\alpha P(t)}{\alpha} dt \\
&= \int_\alpha^\infty P(t) dt \\
&= P(t \geq \alpha)
\end{aligned}$$

Proved.

(b)

By definition, $E[(u-\mu)^2] = \sigma^2$, then using the conclusion in part (a), we have:

$$\begin{aligned}
P((u-\mu)^2 \geq \alpha) &\leq \frac{E[(u-\mu)^2]}{\alpha} \\
P((u-\mu)^2 \geq \alpha) &\leq \frac{\sigma^2}{\alpha}
\end{aligned}$$

(c)

By definition, for N iid random variables $(u_1, u_2, ...u_n)$ each with $E(u_i) = \mu$ and variance $Var(u_i) = \sigma^2$ for $0 < i \leq n$
Let $u = \frac{u_1+u_2+...+u_n}{n}$ we have that:

$$\begin{aligned} E[u] &= E[\frac{u_1 + u_2 + ... + u_n}{n}] \\ &= \frac{1}{N} E[(u_1 + u_2 + ... + u_n)] \\ &= \frac{1}{N}(E[u_1] + E[u_2] + ...E[u_n]) \\ &= \frac{1}{N}(N\mu) \\ &= \mu \end{aligned}$$

And that:

$$\begin{aligned} Var[u] &= Var[\frac{u_1 + u_2 + ... + u_n}{N}] \\ &= \frac{1}{N^2} Var[(u_1 + u_2 + ... + u_n)] \\ &= \frac{1}{N^2}(Var[u_1] + Var[u_2] + ...Var[u_n]) \\ &= \frac{1}{N^2}(n\sigma^2) \\ &= \frac{\sigma^2}{N} \end{aligned}$$

Then similar to that in part (b):

$$\begin{aligned} P((u-\mu)^2 \geq \alpha) &\leq \frac{E[(u-\mu)^2]}{\alpha} \\ P((u-\mu)^2 \geq \alpha) &\leq \frac{Var[u]}{\alpha} \\ P((u-\mu)^2 \geq \alpha) &\leq \frac{\sigma^2}{N\alpha} \end{aligned}$$

proved.