1. **Problem 1.3**

a) Because $\vec{w}^{*T}$ is optimal, it correctly separates the data. This means that $h(x) = \text{sign}(\vec{w}^{*T}\vec{x}_n)$ correctly classifies the data as either positive or negative. y_n is the correct output of either -1 or +1, so $\text{sign}(\vec{w}^{*T}\vec{x}_n) = y_n$. Therefore, $y_n(\vec{w}^{*T}\vec{x}_n)$ will always be greater than 0 since positive · positive and negative · negative is always greater than 0.

b)
$$\vec{w}^T(t-1) = \vec{w}^T(t) - y_n(t-1)\vec{x}_n(t-1) \quad \text{update rule}$$
$$\vec{w}^T(t-1) + y_n(t-1)\vec{x}_n(t-1) = \vec{w}^T(t) \quad \text{rearrange}$$
$$\vec{w}^* \vec{w}^T(t-1) + \underbrace{\vec{w}^* y_n(t-1)\vec{x}_n(t-1)}_{\leq \rho \text{ since } \rho = \min_{1\leq n\leq N} \vec{w}^* y_n(t-1)\vec{x}_n(t-1)} = \vec{w}^* \vec{w}^T(t) \quad \text{multiply by } w^*$$
$$\vec{w}^*\vec{w}^T(t-1) + \rho \leq \vec{w}^*\vec{w}^T(t)$$

Base case: $t=0$, $\vec{w}(0) = 0 \rightarrow 0 \cdot \vec{w}^* \geq 0 \cdot \rho$ ✓

induction: Assume $\vec{w}^*\vec{w}^T(t-1) + \rho \geq (t-1)\rho + \rho$
$$\vec{w}^*\vec{w}^T(t-1) + \rho \geq t\rho$$
$$\text{Since } \vec{w}^*\vec{w}^T(t-1) + \rho \geq \vec{w}^*\vec{w}^T(t)$$
$$\vec{w}^*\vec{w}^T(t) \geq t\rho \text{ For all } t = 0, \ldots, t$$

c)
$$\|\vec{w}(t)\|^2 = \|\vec{w}(t-1) + y(t-1)\vec{x}(t-1)\|^2$$
$$= \|\vec{w}(t-1)\|^2 + \underbrace{2y(t-1)\vec{x}(t-1)\vec{w}(t-1)}_{\text{Less than or equal to } 0} + \|y(t-1)\vec{x}(t-1)\|^2$$
$$\leq \|\vec{w}(t-1)\|^2 + \|\underbrace{y(t-1)}_{y(t-1)^2 = 1}\vec{x}(t-1)\|^2$$
$$\|\vec{w}(t)\|^2 \leq \|\vec{w}(t-1)\|^2 + \|\vec{x}(t-1)\|^2$$

d) $R = \max_{1 \le n \le N} \|\vec{x}_n\|$

Base: if $t = 0$, $0 \le 0 \cdot R^2$ ✓

Assume: $\|\vec{w}(t-1)\|^2 + \|\vec{x}(t-1)\|^2 \le (t-1)R^2 + R^2$

$\|\vec{w}(t-1)\|^2 \le tR^2$

Since $\|\vec{w}(t)\|^2 \le \|\vec{w}(t-1)\|^2$ $\quad (\|\vec{x}(t-1)\|^2 > 0)$

$\|\vec{w}(t)\|^2 \le tR^2$ For all $t = 0, \dots, t$

e) $\sqrt{\|\vec{w}(t)\|^2} \le \sqrt{tR^2} \longrightarrow \|\vec{w}(t)\| \le \sqrt{t}R$

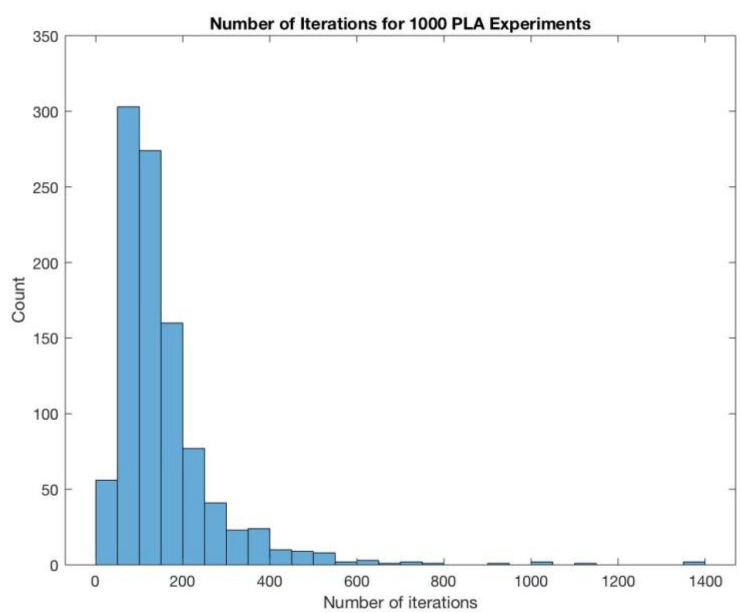
$$\frac{\vec{w}(t) w^*}{\|\vec{w}(t)\|} \ge \frac{t\rho}{\sqrt{t}R} \rightarrow \frac{w^T(t) w^*}{\|\vec{w}(t)\|} \ge \sqrt{t}\frac{\rho}{R}$$

$$\frac{(\vec{w}^T(t)\vec{w}^*)^2}{\|\vec{w}(t)\|^2} \ge t\frac{\rho^2}{R^2}$$

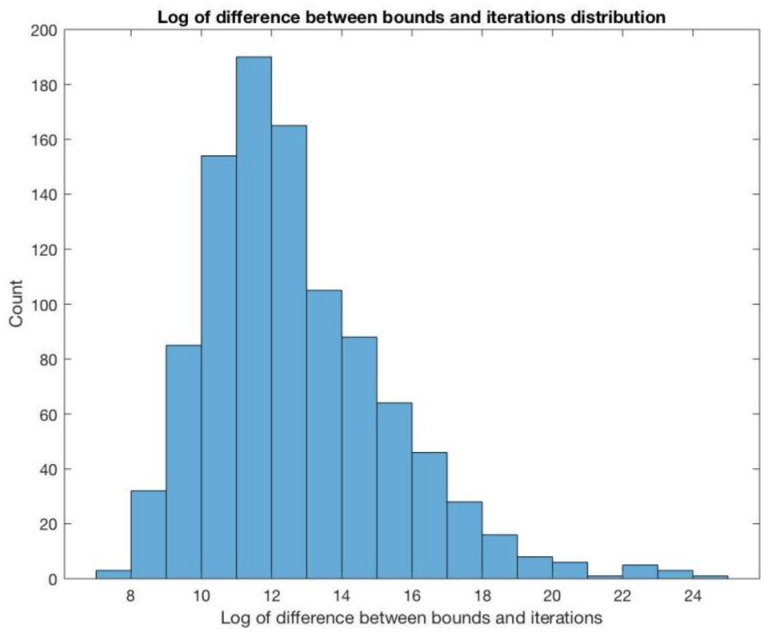
$$\underbrace{\boxed{\frac{(\vec{w}^T(t) \cdot \vec{w}^*)^2}{\|\vec{w}(t)\|^2\|\vec{w}^*\|^2}}}_{<1} \|\vec{w}^*\|^2 \cdot \frac{R^2}{\rho} \le t$$

$$\frac{R^2}{\rho^2}\|\vec{w}^*\|^2 \le t \quad \text{Q.E.D}$$

2.



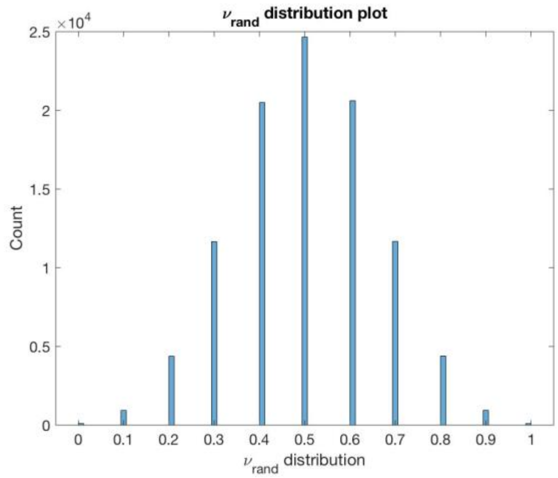
Most experiments take between about 50 and 250 iterations which makes sense since there are 100 vectors that need to be classified and the weight vector is updated based on one point each time. Occasionally the guessed weight vector might be way off or almost correct so it will take more or fewer iterations to guess the separating weight vector.



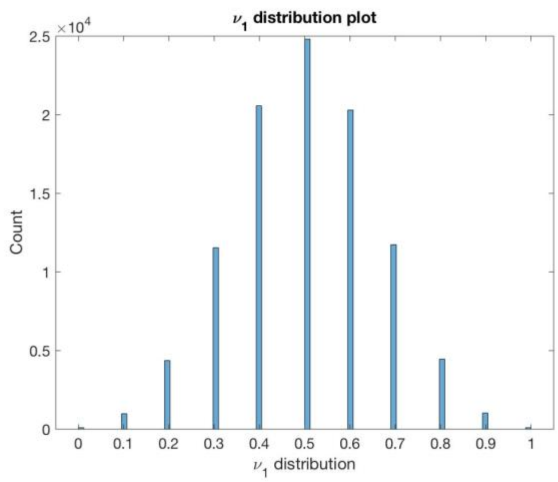
The log difference between the bounds and iterations suggests that the PLA algorithm converges quicker than the bound in the previous question would indicate.

3.

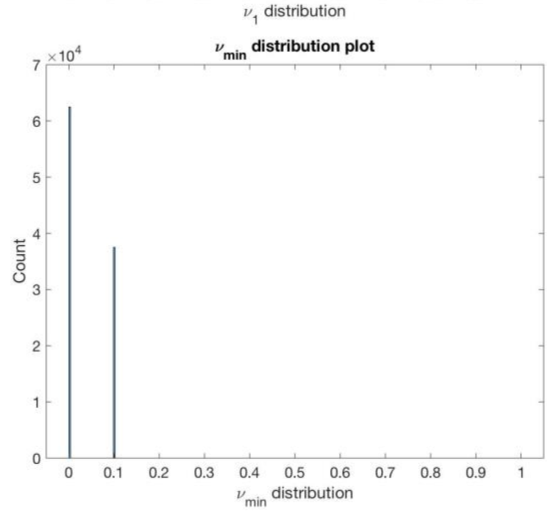
a. μ = 0.5 for all 3 fair coins
b.



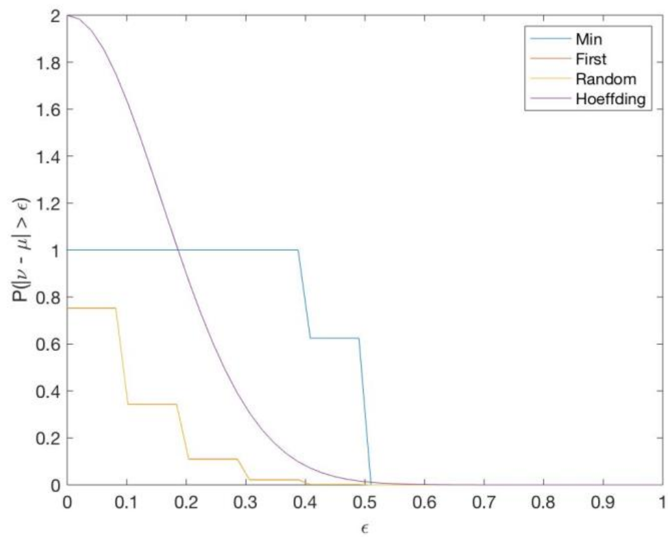
ν_{rand} distribution plot
$\times 10^4$
Count
ν_{rand} distribution
0
0.5
1
1.5
2
2.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9



ν_1 distribution plot
$\times 10^4$
Count
ν_1 distribution
0
0.5
1
1.5
2
2.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9



ν_{min} distribution plot
$\times 10^4$
Count
ν_{min} distribution
0
1
2
3
4
5
6
7
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

c. 

d. The first coin and the random coin obey the bound because they are random processes. The first coin counts as a random process because all coins are the same so the first one is identical to selecting any other coin. The min coin does not obey the bound because we choose a coin based on its properties after the experiment and it is not random, therefore the bound does not apply.

a) Define $Y = \begin{cases} \alpha, & t \geq \alpha \\ 0, & t < \alpha \end{cases}$

Referenced Math 493 notes: Convergence and Limit Theorems lecture from Fall 2019

$$Y \leq t \rightarrow E[Y] \leq E[t]$$

$$E[Y] = \alpha P(t \geq \alpha) \text{ so}$$

$$\alpha P(t \geq \alpha) \leq E[t]$$

$$P(t \geq \alpha) \leq \frac{E[t]}{\alpha}$$

b) Using eqn in a: $P[(u-\nu)^2 \geq \alpha] \leq \frac{E[(u-\nu)^2]}{\alpha}$

$$E[(u-\nu)^2] = \sigma^2$$

$$P[(u-\nu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$$

c) if $u = \frac{1}{N}\sum_{n=1}^{N} u_n$ then variance $= \frac{1}{N^2}\sum_{n=1}^{N} \sigma_n^2 = \frac{\sigma^2}{N}$

so $E[(u-\nu)^2] = \frac{\sigma^2}{N}$

$$P[(u-\nu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$$

4.

a) $E_{in}(h) = \sum_{n=1}^{N} (h - y_n)^2$

$$\frac{d E_{in}(h)}{dh} = 2 \sum_{n=1}^{N} (h - y_n) = 0$$

$$\cancel{2} \sum_{n=1}^{N} h - \cancel{2} \sum_{n=1}^{N} y_n = 0$$

$$N h_{mean} = \sum_{n=1}^{N} y_n$$

$$h_{mean} = \frac{1}{N} \sum_{n=1}^{N} y_n$$

b) $E_{in}(h) = \sum_{n=1}^{N} |h - y_n|$

$$\frac{d E_{in}(h)}{dh} = \sum_{n=1}^{N} \frac{h - y_n}{|h - y_n|} = 0$$

This equals 0 only when number of positive terms equals the number of negative terms so $h = med\{y_1, \dots, y_N\}$

$h_{med} = med\{y_1, \dots, y_N\}$

c) $h_{mean} \to \infty$

h_{med} stays the same (more stable)

5.

a) $m_H(N)$ For positive rays is $N+1$ (From class)
For negative rays, we also have $N+1$ possibilities, but all $+1$ and all -1 are covered by positive rays so negative rays adds $N-1$ dichotomies
$m_H(N) = N+1+N-1 = 2N$ $d_{vc} = 2$ since $m_H(3) = 6$ is the largest value where $m_H(N) \neq 2^N$

b) $m_H(N) = \binom{N+1}{2} + 1$ for positive intervals (class)
$m_H(N) = \binom{N+1}{2} + 1$ for negative rays as well
but we need to consider overlap.
$\frac{N^2}{2} + \frac{N}{2} + 1$. If $N=3$, we add 1 dichotomy.
If $N=4$, we add 3 more: [(+,-,-,+), (+,-,+,+), (+,+,-,-
If $N=5$, we add 6

For $N=3$, 6 possibilities are covered by both $(2N)$
For $N=4$, 8 possibilities covered by both $(2N)$...
Therefore $2 \cdot m_{H pos}(N) - 2N =$
$N^2 + N + 2 - 2N = N^2 - N + 2$
$m_H(N) = N^2 - N + 2$
$m_H(3) = 8$, $m_H(4) = 14$ so $d_{vc} = 3$

c) For two concentric circles, we are basically doing the same thing as positive rays, just in 2D space. The possibilities still remain the same though, so $m_H(N) = \frac{N^2}{2} + \frac{N}{2} + 1$
$d_{vc} = 2$ $m_H(3) = 7$

6.

''

2 possibilities: $d_{vc}(H) = \infty$ if $m_H(N) = 2^N$ for all N
or $m_H(N) \leq N^{d_{vc}} + 1$

$1+N$: $d_{vc} = 1$ so obviously, $N^1 + 1 = 1 + N$ for all N,
so $1+N$ is possible

$1+N+\frac{N(N-1)}{2}$: $d_{vc} = 2$ so $m_H(N)$ must be less than N^2+1
$\hookrightarrow 1 + \frac{N}{2} + \frac{N^2}{2}$ → obviously a possible growth
function as it will always be bounded by N^2+1

2^N: $d_{vc} = \infty$ and $m_H(N) = 2^N$ → 2^N is a possible
growth function

$2^{\lfloor\sqrt{N}\rfloor}$: $d_{vc} = 1$ so $m_H(N) \leq N+1$
$m_H(25) = 2^5 = 32 \geq 26$ so $2^{\lfloor\sqrt{N}\rfloor}$ is NOT
a possible growth function

$2^{\lfloor \frac{N}{2} \rfloor}$: $d_{vc} = 0$ since $2^{\lfloor\frac{1}{2}\rfloor} = 2^0 = 1 \neq 2^1$ so $m_H(N) \leq 2$
This obviously doesn't hold; NOT possible growth function

$1+N+\frac{N(N-1)(N-2)}{6}$: $d_{vc} = 1$ → $m_H(2) = 3$; $m_H(1) = 2$ so $m_H(N) \leq N+1$
$m_H(5) = 16 > 6$ so this function is NOT
a possible growth function

7.