Assignment 2

Name: Mohaiminul Al Nahian
RIN: 662026703
Course no.: CSCI 6100

1 Exercise 1.8

Here $\nu \leq 0.1$ means that there will at most be 0 or 1 red balls in a sample of N=10
So,

$$\begin{aligned} P(\nu \leq 0.1) &= P(\nu = 0) + P(\nu = 1) \\ &= \binom{10}{0}(1-\mu)^{10} + \binom{10}{1}\mu^1(1-\mu)^9 \\ &= 0.1^{10} + 10 \times 0.9 \times 0.1^9 \\ &= 9.1 \times 10^{-9} \end{aligned}$$

2 Exercise 1.9

For P($\nu \leq 0.1$) we have P($|\nu - \mu| \geq 0.8$)
So, for Hoeffding inequality we may get the probability in limiting case, i.e, $P(|\nu - \mu| > 0.8)$, where $\epsilon = 0.8$

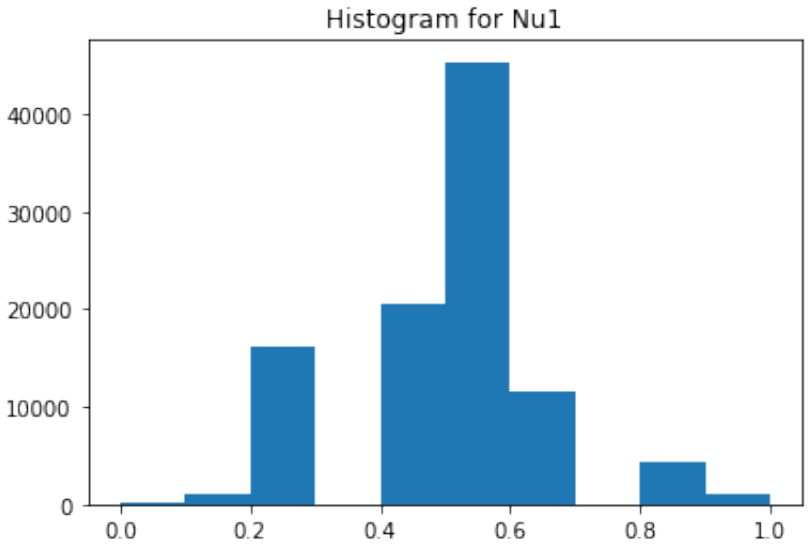
$$\begin{aligned} P(|\nu - \mu| > \epsilon) &\leq 2 \times e^{-2\epsilon^2 N} \\ &= 2 \times e^{-2*0.8^2*10} \\ &= 5.52 \times 10^-6 \end{aligned}$$

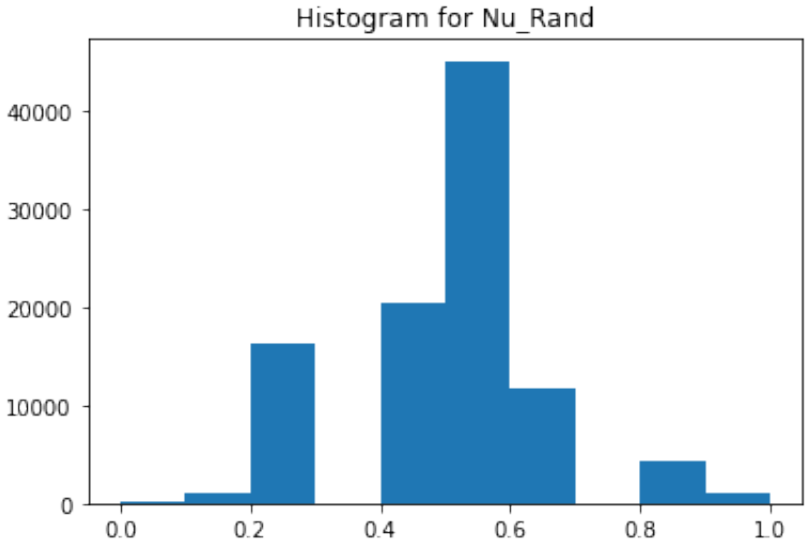
So, $P(|\nu - \mu| > \epsilon) \leq 5.52 \times 10^-6$
Here, the value is much larger than the probability calculated in Exercise 1.8 (9.1×10^{-9}).In this case, the result is a probability and not tightly bounded.

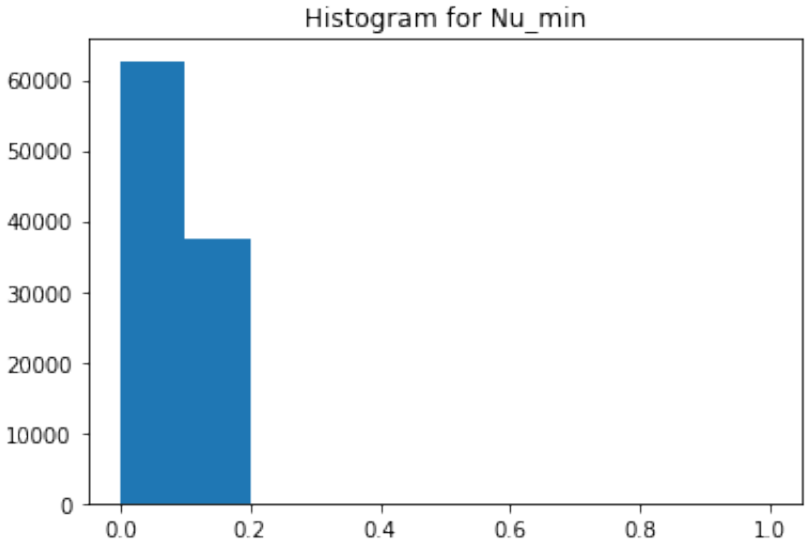
3 Exercise 1.10

(a) Since the coins are fair, the μ for 3 coins is 0.5

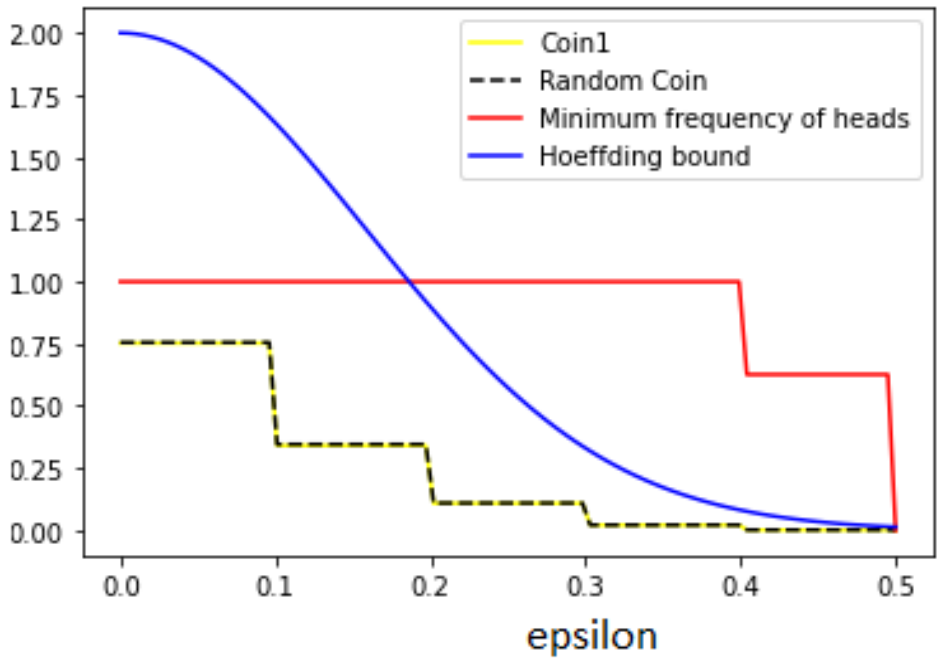
(b) For 100,000 runs the histograms for ν_1, ν_{rand} and ν_{min} are given below







(c) Estimates for $P(|\nu - \mu| > \epsilon)$ together with Hoeffding bound is shown below:



(d) The first coin and random coin of the experiments are always below the Hoeffding bound line and hence follows the bound. The coin with minimum frequency does not follow the Hoeffding bound.
The reason is that for the first two cases, the hypothesis has been fixed before the experiment. But for the coin with minimum frequecy of heads, we have to flip all the coins first and then choose the coin with minimum frequency of heads, which violates the hoeffding bound condition of choosing h before experiment.

(e) The 1000 coins can be considered as 1000 bins. Choosing the First coin and one random coin is like choosing the bin before generating the data. But for choosing the coin with minimum frequency of heads is like choosing a bin after dataset has been generated, just like learning algorithm picks a hypothesis g based on dataset

4 Exercise 1.11

(a) S cannot guarantee a better performance than a random function.
For example, if D has majority outputs +1, then S will choose h_1, and output will always be +1. Random function will have 50% as +1 and 50% as -1. In this case, if f has more outputs as -1 outside D, then S (h_1) fails to predict the outputs correctly, whereas random function will be correct 50% of the time. So, it will be better than S.

(b) It is possible for C to perform better than S. If majority outputs of examples in D is +1, then S will choose h_1 (all +1) and C will choose h_2 (all -1). If f has majority outputs as -1 outside D, then C will perform better than S in this case.

(c) If p=0.9, then h_1 is better than h_2. So, S will choose h_1 and there must be at least 13 points among the 25 whose outputs are +1. So, probability that S will be better than C is

$$P = \sum_{k=13}^{25} \binom{25}{k} 0.9^k (1-0.9)^{25-k} = 0.999999984$$

(d) There is no exact value of p for which it is more likely that C will perform better than S. Because S chooses the hypothesis that better fits samples of D and it is likely that samples will approximate the real results of f. So, S should be better than C always.

5 Exercise 1.12

The best option would be to choose (c), because
(i) If we can learn and produce a g, we may show that $E_{in}(g) \approx 0$ and also Hoeffding inequality $P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$ could hold as we have a more or less large data size N=4000 and we may ascertain that $E_{in}(g) \approx E_{out}(g)$
(ii) Or, we may conclude that the target function f is so complicated that we will get $E_{in}(g) \neq 0$, and I shall declare that I have failed

6 Problem 1.3

(a) Since w^* is the optimal set of weights so, y_n and $w^T x_n$ has the same sign. So, $y_n w^T x_n > 0$
Since $\rho = min_{1 \leq n \leq N}\ y_n w^T x_n$, so,

$$\rho > 0$$

(b) From weight update rule,

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

We know that, $y(t-1)x^T(t-1)w^* \geq \rho$ as $\rho = min_{1 \leq n \leq N}\ y_n w^T x_n$, so

$$w^T(t)w^* \geq w^T(t-1)w^* + \rho$$
$$[Showed]$$

We above may write,

$$w^T(1)w* \geq w^T(0)w^* + \rho$$
$$or, w^T(1)w* \geq \rho$$
$$again, w^T(2)w* \geq w^T(1)w^* + \rho$$
$$\geq 2\rho$$

So, proceeding like this, we may conclude,

$$w^T(t)w* \geq t\rho$$
$$[showed]$$

(c) From weight update rule,

$$w(t) = w(t-1) + y(t-1)x(t-1)$$
$$or, ||w(t)||^2 = ||w(t-1)||^2 + (y(t-1))^2||x(t-1)||^2 + 2y(t-1)w^T(t-1)x(t-1)$$

as x(t-1) is misclassified by w(t-1), the last term in the above equation is less than or equal to zero. Also $(y(t-1))^2 = 1$. So, we may write

$$||w(t)||^2 \leq ||w(t-1)||^2 + ||x(t-1)||^2$$
$$[showed]$$

(d) First, let us assume that $||w(t)||^2 \leq tR^2$

From c, we can write

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

Here, $R = max||x(n)||, so||x(t)||^2 \leq R^2$. So,

$$||w(t+1)||^2 \leq tR^2 + R^2$$
$$or, ||w(t+1)||^2 \leq (t+1)R^2$$

So, if $||w(t)||^2 \leq tR^2$ then $||w(t+1)||^2 \leq (t+1)R^2$ holds and we may conclude that

$$||w(t)||^2 \leq tR^2$$
$$[showed]$$

(e)From b, we know that $w^T w* \geq t\rho$ and from d

$$||w(t)||^2 \leq tR^2$$
$$or, ||w(t)|| \leq \sqrt{t}R$$
$$or, \frac{1}{||w(t)||} \geq \frac{1}{\sqrt{t}R}$$

multiplying it with the result of (b), we get

$$\frac{w^T w^*}{||w(t)||} \geq \frac{t\rho}{\sqrt{t}R}$$
$$or, \frac{w^T}{||w(t)||} w^* \geq \sqrt{t}\frac{\rho}{R}$$
$$[showed]$$

Again, from above

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

We know that, $w^T w^* \leq ||w||||w^*||$, so,

$$\sqrt{t} \leq \frac{||w(t)||||w^*||R}{||w(t)||\rho}$$
$$or, \sqrt{t} \leq \frac{||w^*||R}{\rho}$$
$$or, t \leq \frac{||w^*||^2 R^2}{\rho^2}$$
$$[showed]$$

7 Problem 1.7

(a) For 1 coin P($\nu = 0$)= $\binom{N}{0}\mu^0(1-\mu)^{N-0} = (1-\mu)^N$
So, for 1 Coin, $P(\nu > 0) = [1-(1-\mu)^N]$
So, for n coins $P(\nu > 0) = [1-(1-\mu)^N]^n$
Then, for n coins, $P(\nu = 0) = 1-[1-(1-\mu)^N]^n$

When, $\mu = 0.05$,
for, n=1, $P(\nu = 0) = 1-[1-(0.95)^{10}]^1 = 0.599$
for, n=1000, $P(\nu = 0) = 1-[1-(0.95)^{10}]^{1000} \approx 1$
for, n=1000000, $P(\nu = 0) = 1-[1-(0.95)^{10}]^{1000000} \approx 1$

When, $\mu = 0.8$,
for, n=1, $P(\nu = 0) = 1-[1-(0.2)^{10}]^1 = 1.024 \times 10^{-7}$
for, n=1000, $P(\nu = 0) = 1-[1-(0.2)^{10}]^{1000} = 0.0001024$
for, n=1000000, $P(\nu = 0) = 1-[1-(0.2)^{10}]^{1000000} = 0.0973$

(b)

$$\begin{aligned} P(max|\nu_i - \mu_i| > \epsilon) &= P(|\nu_1 - \mu_1| > \epsilon) or P(|\nu_2 - \mu_2| > \epsilon) \\ &= P(|\nu_1 - \mu_1| > \epsilon) + P(|\nu_2 - \mu_2| > \epsilon) - P(|\nu_1 - \mu_1| > \epsilon)P(|\nu_2 - \mu_2| > \epsilon) \\ &\leq P(|\nu_1 - \mu_1| > \epsilon) + P(|\nu_2 - \mu_2| > \epsilon) \\ &\leq 4e^{-2\epsilon^2 N} \end{aligned}$$

Here, $\mu = 0.5$

Since, there are two coins,

$$\begin{aligned} P(max|\nu_i - \mu_i| > \epsilon) &= 1 - P(max|\nu_i - \mu_i| \leq \epsilon) \\ &= 1 - [P(|\nu_1 - \mu| \leq \epsilon) and P(|\nu_2 - \mu| \leq \epsilon) \\ &= 1 - [P(|\nu_1 - \mu| \leq \epsilon)]^2 \end{aligned}$$

For, k=[0,1,2,3,4,5,6] we have $P(|\nu - \mu|)$=[0.0156,0.0938,0.2344,0.3125, 0.2344,0.0938, 0.0156]

Also, weh have $|\nu - \mu|$=[0.5, 0.333, 0.167, 0, 0.167,0.333, 0.5]

Now, when $0 < \epsilon < 0.167$, 4th value of $|\nu - \mu|$ (with a probability 0.3125) is less than ϵ and is excluded. In this range, $P(max|\nu_i - \mu_i| > \epsilon) = 1 - (0.3125)^2 = 0.902$

When, when $0.167 < \epsilon < 0.333$, 3rd, 4th and 5th value of $|\nu - \mu|$ (with a probability 0.2344, 0.3125 and 0.2344) is less than ϵ and is excluded.
In this range, $P(max|\nu_i - \mu_i| > \epsilon) = 1 - (0.2344 + 0.3125 + 0.2344)^2 = 0.39$

When, $0.333 < \epsilon < 0.5$ then 2nd, 3rd, 4th, 5th and 6th value of $|\nu - \mu|$ (with a probability 0.0938, 0.2344, 0.3125, 0.2344, 0.0938) is less than ϵ and is therefore excluded.
In this range, $P(max|\nu_i - \mu_i| > \epsilon) = 1 - (0.0938 + 0.2344 + 0.3125 + 0.2344 + 0.0938)^2 = 0.061$

$\epsilon > 0.5$, all values of $|\nu - \mu|$ is less than ϵ, so all are excluded, and $P(max|\nu_i - \mu_i| > \epsilon) = 0$

The plot is givent below: