Exercise 1.8.

P[red] = 0.9

P[green] = 0.1

$\frac{red}{10} \leq 0.1$

$red \leq 1$

Hence we have the problem:

P[$red \leq 1$] = P[red = 1] + P[red = 0]

Since we have 10 balls in our sample space, we got the following:

P[red = 1] = $b(x; n, p) = \binom{10}{1} 0.9^1 (0.1)^9$

P[red = 0] = 0.1^{10}

P[$red \leq 1$] = $\binom{10}{1} 0.9^1 (0.1)^9 + 0.1^{10}$

= $10 \times 0.9 \times 10^{-9} + 10^{-10}$

= $10^{-9} \times (9 + 0.1)$

= 9.1 x 10^{-9}

Exercise 1.9.

$P[|\nu - \mu| > \epsilon] \leq 2e^{-2\varepsilon^2 N}$

Since, $\mu = 0.9$ and $\nu \leq 0.1$, we get:

$0.9 \geq |\nu - \mu| \geq 0.8$

We take $\varepsilon = 0.8$

$P[|\nu - \mu| > \epsilon] \leq 2e^{-2*0.8^2*10}$

$P[|\nu - \mu| > \epsilon] \leq 5.52 \times 10^{-6}$

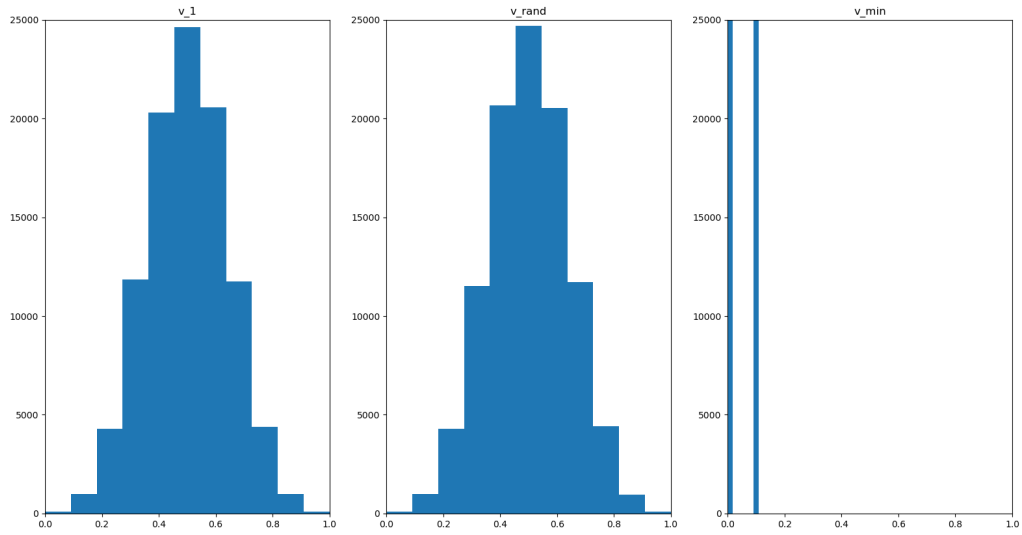
This means that the the probability of getting a bad value is smaller than 5.52×10^{-6}. This corresponds to the answer of the first problem, which is that the probability is **9.1×10^{-9};** it is much smaller than the bound.

Exercise 1.10.

a)

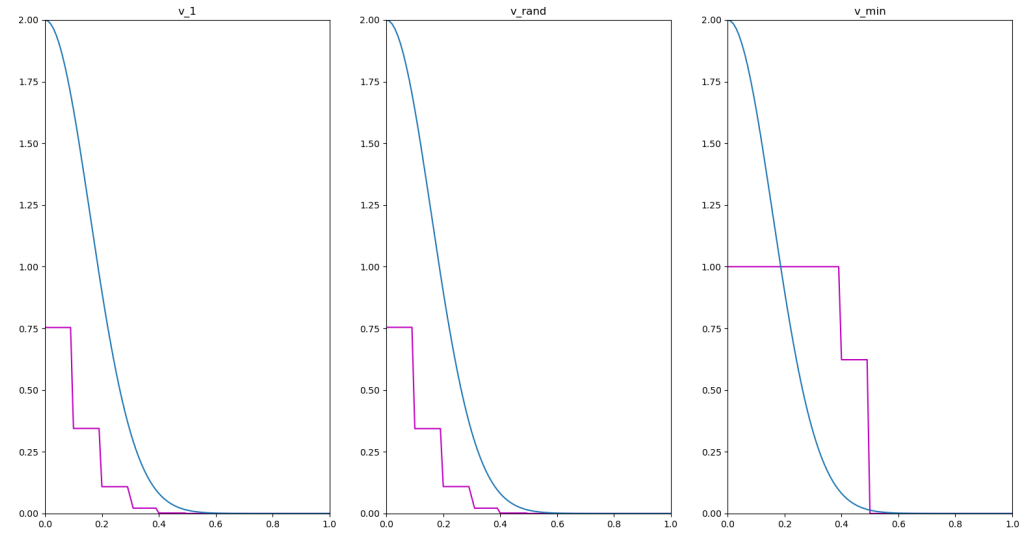
Since each coin is fair, $\mu = 0.5$

b) exercise1.10b.py (take about 5s to run, depending on PC)



The leftmost histogram is the v_1 ; the middle one is the v_{rand}; the rightmost one is the v_{min}. The x-axis is the v, the y-axis is the number of times the corresponding v appears.

c) exercise1.10c.py (take about 10s to run, depending on PC)



The leftmost diagram is the v_1 ; the middle one is the v_{rand}; the rightmost one is the v_{min}. The blue line in each diagram is the Hoeffding bound; the magenta line in each diagram is the $P[|v-\mu| > \varepsilon]$ as a function of ε. The x-axis is the epsilon; the y-axis is the probability.

d)
c_1 and c_{rand} obey the Hoeffding, while c_{min} does not.
c_{min} does not obey because we did not pick it randomly, we explicitly pick the minimum head frequency coin; the result will not be randomly distributed.
c_{rand} obeys because it is selected randomly. For c_1 , it also obeys since we never know how many heads it got in each term since it is generated randomly.

e)
1000 coins represents the 1000 bins on figure 1.10. The c_{min} is basically simulating the rarest case, since we intentionally pick 100000 minimum frequency head coins which breaks the rule of Hoeffding inequality.

The c_1 randomly picks 100000 tosses resulting from X generated in bin 1, which corresponds to the situation when only one bin (or h) exists. c_{rand} picks 100000 tosses result from X from a random bin, which is the multiple bin problem.

Overall, the experiment shows that once there are multiple bins (h), the situation becomes more complex, the Hoeffding inequality may no longer hold as c_{min} shown.

Exercise 1.11.

a)

No. We do not know anything outside of D; it is possible that S performs well on the training set D but worse than C outside of the D, so there is no guarantee that S will still perform better outside.

b)

Yes. Although all y_n in D is +1, we do not know anything outside of the training set, it is still possible that these 25 +1s are the only +1s outside of the training set. In this case, C is definitely better than S.

c)

Since we have P[f(x) = +1] = 0.9 on X, this means that from input space, the probability of +1 is 0.9.

Since S will always choose the hypothesis that agrees most with D, if there are 13 or more data from the given 25 training samples that have f(x) = +1, S will choose h1; we have the following equation

P[13 or more +1 in training data] = $\sum_{k}^{25} \binom{25}{k} (0.9)^{k} (1 - 0.9)^{25-k} \approx 0.999 \approx 1$, so

at this probability, S will choose h1 while C will choose h2. Under this situation, S will definitely perform better than C; because most of X is +1.

The probability that S will produce a better hypothesis when p=0.9 is almost 1.

d)

There is no value.

Since S will always select the hypothesis that agrees most with D, the only case that S will not likely generate a better hypothesis is when p = 0.5. In this case, both S and C will have the same performance. Neither of them is better. Otherwise, S will always be likely to generate better hypotheses than C; because S will choose the better matched hypothesis based on the data set, which has the high probability to imply X according to the Hoeffding inequality. As a result, S will also have a high probability to provide a better hypothesis.

Exercise 1.12.
c) is the best promise.
If I produce a successful hypothesis g, then I must have satisfied both of the two step approach to learning: $E_{out}(g) \approx E_{in}(g)$ and $E_{in}(g) \approx 0$ (implies $E_{out}(g) \approx 0$). In this case, we could conclude that we have g estimates f well out of sample with small error rate.

Otherwise, I could declare my failure once $E_{in}(g)$ is far larger than 0.

Problem 1.3.

a)

When correctly classified, the $w^{*T}x_n$ will produce the same sign as y_n.

Hence, their product will always be positive, so it can be concluded that $y_n(w^{*T}x_n) > 0$ for $1 \leq n \leq N$.

Therefore, $\rho = \min_{1 \leq n \leq N} y_n(w^{*T}x_n) > 0$.

b)

$$w^T(t)w^* \geq w^T(t-1)w^* + min_{1 \leq n \leq N} y_n(w^{*T}x_n)$$

According to update rule we have:

$$(w(t-1) + y(t-1)x(t-1))w^* \geq w^T(t-1)w^* + min_{1 \leq n \leq N} y_n(w^{*T}x_n)$$

$$w^T(t-1)w^* + y(t-1)\ w^{*T}x(t-1)\ \geq w^T(t-1)w^* + min_{1 \leq n \leq N} y_n(w^{*T}x_n)$$

$$y(t-1)\ w^{*T}x(t-1)\ \geq min_{1 \leq n \leq N} y_n(w^{*T}x_n)$$

LHS must >= RHS since RHS is the smallest possible $y_n(w^{*T}x_n)$

Hence, we proved that $w^T(t)w^* \geq w^T(t-1)w^* + \rho$, where $\rho = \min_{1 \leq n \leq N} y_n(w^{*T}x_n)$.

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

Induction:

For t = 1, we have

$w^T(1)w^* \geq \rho$, using $w^T(t)w^* \geq w^T(t-1)w^* + \rho$, we get $w^T(1)w^* \geq w^T(0)w^* + \rho$ which is $w^T(t)w^* \geq \rho$. Hence, $w^T(1)w^* \geq \rho$is true.

Assume the statement holds true when t = t ($w^T(t)w^* \geq t\rho$), prove for t+1 the statement holds true, which is $w^T(t+1)w^* \geq (t+1)\rho$

Using $w^T(t)w^* \geq w^T(t-1)w^* + \rho$ and t = t ($w^T(t)w^* \geq t\rho$), we get

$w^T(t+1)w^* \geq w^T(t)w^* + \rho \geq t\rho + \rho = (t+1)\rho$; the statement holds true for t+1

Hence, we can conclude that $w^T(t)w^* \geq t\rho$ is true.

c)

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

Since $||a||^2 = a \bullet a = a^T a$, we get

$w(t) \bullet w(t) \leq w^T(t-1)w(t-1) + x^T(t-1)x(t-1)$

According to the update rule we get:

$[w(t-1) + y(t-1)x(t-1)] \bullet [w(t-1) + y(t-1)x(t-1)] \leq w^T(t-1)w(t-1) + x$

$w^T(t-1)w(t-1) + y(t-1)w^T(t-1)x(t-1) + y(t-1)w^T(t-1)x(t-1) + y^2(t-1)$

$\leq w^T(t-1)w(t-1) + x(t-1)^T x(t-1)$

Since $y^2(t-1)$= 1 we get

$w^T(t-1)w(t-1) + y(t-1)w^T(t-1)x(t-1) + y(t-1)w^T(t-1)x(t-1) + x^T(t-1)$

$\leq w^T(t-1)w(t-1) + x(t-1)^T x(t-1)$

$y(t-1)w^T(t-1)x(t-1) + y(t-1)w^T(t-1)x(t-1) \leq 0$

Since x is not correctly classified, we have $y(t-1)w^T(t-1)x(t-1) \leq 0$, the above inequality is true.

d)

Induction:

Prove that for t = 1 the statement holds true.

$||w(1)||^2 \leq R^2$

$||w(0) + y(0)x(0)||^2 \leq max_{1 \leq n \leq N} \left|\left|x_n\right|\right|^2$ (update rule)

Since $w(0) = \overline{0}$ we have $||y(0)x(0)||^2 \leq max_{1 \leq n \leq N} \left|\left|x_n\right|\right|^2$

$y^2(0)x^T(0)x(0) \leq max_{1 \leq n \leq N} x_n{}^T x_n$

Since $y^2(0) = 1$, the inequality become $x(0)^T x(0) \leq max_{1 \leq n \leq N} x_n{}^T x_n$

This is true because the maximum of $x_n{}^T x_n$ must larger or equal to the $x(0)^T x(0)$.

Hence, we can conclude for t = 1 the statement is true.

Assume for t = t ($||w(t)||^2 \leq tR^2$) the statement holds true, prove that for t = t+1 the statement also holds true.

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

From c), t = t ($||w(t)||^2 \leq tR^2$) and $R = max_{1 \leq n \leq N} \left|\left|x_n\right|\right|^2$ we have:

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

Hence, we can conclude that the statement $||w(t)||^2 \leq tR^2$ is true.

e)

From b) and d), we have

$w^T(t)w^* \geq t\rho$ and $||w(t)||^2 \leq tR^2$

Apply square root to both side of d), we got

$||w(t)|| \leq R\sqrt{t}$, combining these two inequality we get,

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

Hence, $\frac{w^T(t)w^*}{||w(t)||} \geq \frac{t\rho}{R\sqrt{t}} \rightarrow \frac{w^T(t)w^*}{||w(t)||} \geq \sqrt{t}\frac{\rho}{R}$

Since $\frac{w^T(t)w^*}{||w(t)||\,||w^*||} \leq 1 \rightarrow w^T(t)w^* \leq ||w(t)||\,||w^*||$, we have

$\frac{||w(t)||\,||w^*||}{||w(t)||} \geq \sqrt{t}\frac{\rho}{R} \rightarrow ||w^*|| \geq \sqrt{t}\frac{\rho}{R}$ apply square on both side, we get

$||w^*||^2 \geq t\frac{\rho^2}{R^2} \rightarrow t \leq ||w^*||^2\frac{R^2}{\rho^2}$

Proved that $t \leq \frac{R^2||w^*||^2}{\rho^2}$

Problem 1.7.

a)

Since ν= 0, we have k = 0, the problem became

$\mu = 0.05$:

1 coin: P[0 | 10, 0.05] = $\binom{10}{0} 0.05^0 (1 - 0.05)^{10} = (0.95)^{10}$

1000 coins: $1 - (1 - (1 - 0.05)^{10})^{1000} \approx 1$

1000000 coins: $1 - (1 - (1 - 0.05)^{10})^{1000000} \approx 1$

$\mu = 0.8$:

1 coin: P[0 | 10, 0.8] = $\binom{10}{0} 0.8^0 (1 - 0.8) = (0.2)^{10}$

1000 coins: $(P[0 \mid 10, 0.8])^{1000} = 1 - (1 - (1 - 0.8)^{10})^{1000} \approx 1.02 \times 10^{-4}$

1000000 coins: $(P[0 \mid 10, 0.8])^{1000000} =$

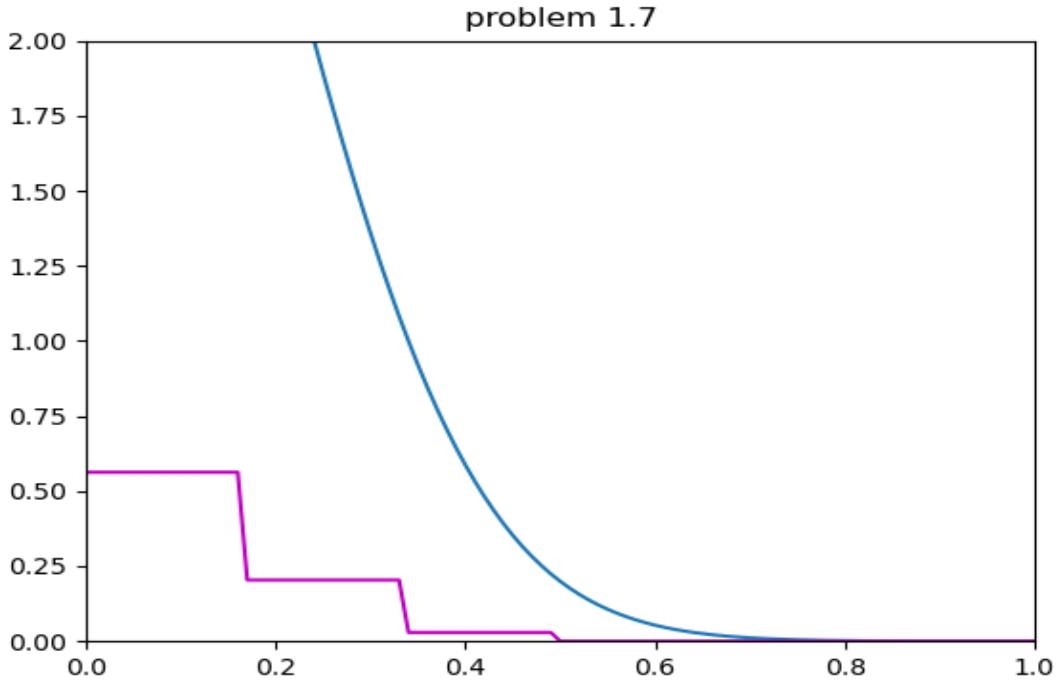
$1 - (1 - (1 - 0.8)^{10})^{10000} \approx 9.73 \times 10^{-2}$

b)

N = 6

$P[max_i |v_i - u_i| > \varepsilon] = P[|v_1 - u_1| > \varepsilon \ or \ |v_2 - u_2| > \varepsilon]$

$P[|v_1 - u_1| > \varepsilon] + P[|v_2 - u_2| > \varepsilon] - P[|v_1 - u_1| > \varepsilon] P[|v_2 - u_2| > \varepsilon]$



$\leq 4e^{-12\varepsilon^2} - 4e^{-24\varepsilon^2}$

The blue line is the bound $4e^{-12\varepsilon^2}$ for $P[max_i|v_i - u_i| > \varepsilon]$ as a function of ε. The magenta line is the $P[max_i|v_i - u_i| > \varepsilon]$ as a function of ε.