CSCI 4100 Assignment 2

Boliang Yang 661541863

1. Exercise 1.8 in LFD

We have occurrence of red to be 0 or 1 since $v \leq 0.1$ and there are 10 marbles in the sample, the red occurrence is $v \cdot 10 \leq 1$.

$$P[0\ occurrence] = P[v = 0] = (1 - 0.9)^{10} \times \binom{10}{0} = 1 \times 10^{-10}$$

$$P[1\ occurrence] = P[v = 0.1] = (1 - 0.9)^{9} \times 0.9 \times \binom{10}{1} = 9 \times 10^{-9}$$

$$P[v \leq 0.1] = P[v = 0] + P[v = 0.1] = 1 \times 10^{-10} + 9 \times 10^{-9} = 9.1 \times 10^{-9}$$

2. Exercise 1.9 in LFD

Since $v \leq 0.1$, $\mu = 0.9$ and $|v - \mu| > \epsilon$, then $\epsilon \geq 0.8$

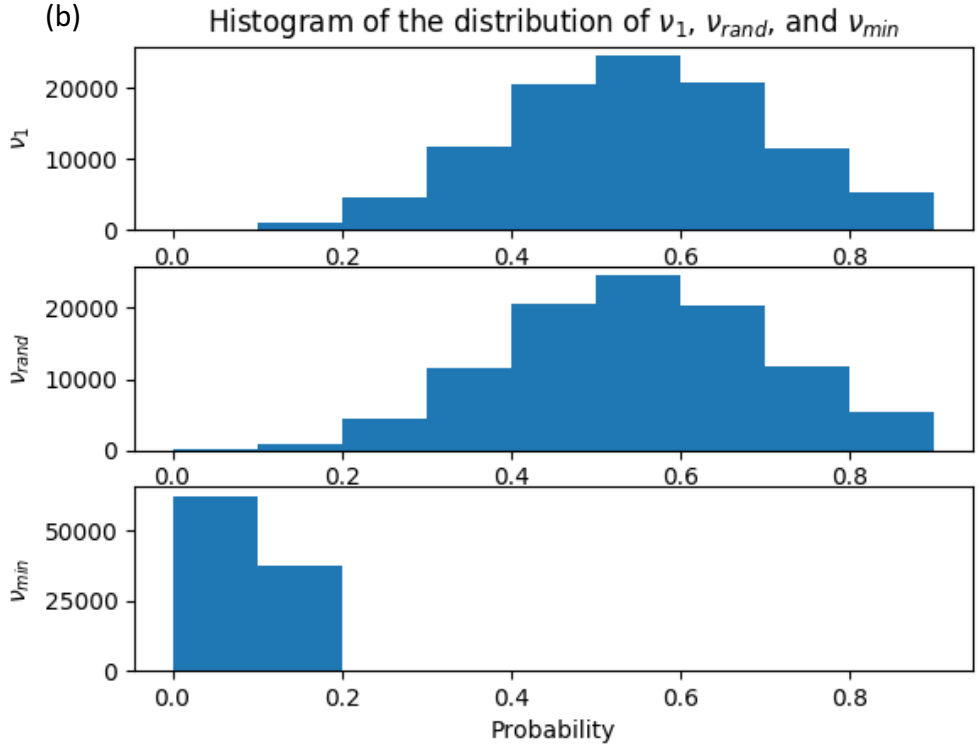
$$P[|v - \mu| > \epsilon] \leq 2e^{-2N\epsilon^2} \leq 2e^{-2 \times 10 \times 0.8^2} = 5.52 \times 10^{-6}$$

The answer here is large comparing to the answer in 1.8.

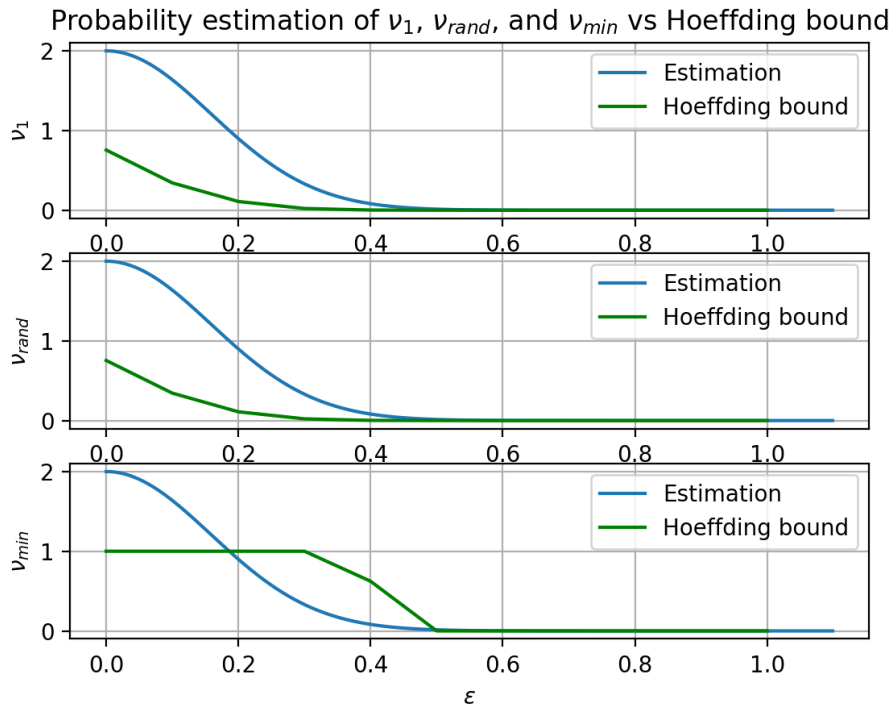
3. Exercise 1.10 in LFD

(a) The μ for all three coins is 0.5.

(b)



(c)



(d) v_1 and v_{rand} obey the Hoeffding bound, v_{min} doe not.

Essentially v_1 and v_{rand} are randomly chosen coins, and v_{min} was not randomly selected but specifically chosen. To be applicable to the Hoeffding bound, the selection must be random. Therefore, v_1 and v_{rand} obey the Hoeffding bound, but v_{min} doe not.

(e) If we assume that we do not know the probability of heads μ as in Figure 1.10 that we don't know the fraction of red marbles, different coins are just like different bins and our hypothesis set are all 1000 coins, then the problem is similar to the bins problem. We pick out the bins for v_1, v_{rand}, v_{min}, then the hypothesis are h_1, h_{rand}, h_{min} respectively. Because h_1 and h_{rand} are randomly chosen and independent, then they obey the Hoeffding bound. And h_{min} does not obey because we always choose it that has the least error and it is not independent.

4. Exercise 1.11 in LFD

(a) No. Because the given dataset D could contain all the training examples with the same value, for example y = +1 for all in D, then it cannot guarantee to perform better

than random on any point outside D. Nothing is guaranteed outside the data set D.

(b) Yes, it is possible. It is possible that all points outside D have y = -1 then C would work better than S. Again, nothing is guaranteed outside the D.

(c) If $p = 0.9$, then D should contain more than $\frac{25}{2}$ examples with y = +1:

$$\mathrm{P[S\ will\ produce\ better\ hypothesis\ than\ C]} = \sum_{i=13}^{25} \binom{25}{i} \times 0.9^i \times 0.1^{25-i} = 0.99999984$$

(d) If $E_{in}(h_1) < 0.5$, then S would choose h_1. If C works better outside D, then $E_{out}(h_1) = 1 - p > 0.5$. Then $E_{out}(h_1) - E_{in}(h_1) = 1 - p - E_{in}(h_1) > 0.5 - p = \epsilon$ and $p < 0.5$.

If $E_{in}(h_2) < 0.5$, then S would choose h_2. If C works better outside D, then $E_{out}(h_2) = p > 0.5$. Then $E_{out}(h_1) - E_{in}(h_1) = p - E_{in}(h_1) > p - 0.5 = \epsilon$ and $p > 0.5$.

Then $P[|E_{in}(g) - E_{out}(g)| > \epsilon] > 0.5$ and since $P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N} = 2\times2\times e^{-2|0.5-p|^2\times25}$, we have $0.5 < 2\times2\times e^{-2|0.5-p|^2\times25}$. Therefore, $0.2961 < p < 0.7039$ and $p \neq 0.5$.

5. Exercise 1.12 in LFD

The best I can promise her is (c). Nothing is guaranteed outside the data set. Depending on the problem, we may need a very large hypothesis set to deal with the training data. Because the data set is fixed to 4000 points, a good hypothesis will lead to high in E_{out}. Then we declare that we failed. Or we may find the function g by just trying a few times, then $E_{out}(g)$ is approximate to $E_{in}(g)$. And if $E_{in}(g)$ approximates to 0 then we can return a hypothesis g with high probability that it will approximate f well out of sample.

6. Problem 1.3 in LFD

(a) For separable data, the optimal set of weights w^* will separate the data such that $y_n = sign(\boldsymbol{w}^{*T}\boldsymbol{x}_n)$. Then for all $1 \leq n \leq N$, we have $y_n(\boldsymbol{w}^{*T}\boldsymbol{x}_n) > 0$. Thus, $\rho = \min_{1\leq n\leq N} y_n(\boldsymbol{w}^{*T}\boldsymbol{x}_n) > 0$.

(b) Since $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)$, then $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* = (\boldsymbol{w}(t-1) + y(t-$

$1)\boldsymbol{x}(t-1))^T\boldsymbol{w}^* = \boldsymbol{w}^{\boldsymbol{T}}(t-1)\boldsymbol{w}^* + \big(y(t-1)\boldsymbol{x}(t-1)\big)^T\boldsymbol{w}^*$.

Since $0 < \rho = \min_{1\le n\le N} y_n(\boldsymbol{w}^{*T}\boldsymbol{x_n}) \le y_{t-1}\boldsymbol{w}^{*\boldsymbol{T}}\boldsymbol{x}_{t-1} = (y(t-1)\boldsymbol{x}(t-1))^{\boldsymbol{T}}\boldsymbol{w}^*$,

then $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge \boldsymbol{w}^{\boldsymbol{T}}(t-1)\boldsymbol{w}^* + \rho$. Now we show $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge t\rho$ by induction.

Base case: When $t = 1, \boldsymbol{w}^{\boldsymbol{T}}(0) = \boldsymbol{0}$, then $\boldsymbol{w}^{\boldsymbol{T}}(1)\boldsymbol{w}^* \ge \boldsymbol{w}^{\boldsymbol{T}}(0)\boldsymbol{w}^* + \rho = \rho$

And $\boldsymbol{w}^{\boldsymbol{T}}(1)\boldsymbol{w}^* \ge \rho$ satisfy $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge t\rho$ when $t = 1$

Induction step:

Let $k \in N$ and $k > 1$ and assume $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge t\rho$ is ture for $t = k$.

at $t = k+1$: $\boldsymbol{w}^{\boldsymbol{T}}(k+1)\boldsymbol{w}^* \ge \boldsymbol{w}^{\boldsymbol{T}}(k)\boldsymbol{w}^* + \rho \ge k\rho + \rho = (k+1)\rho$

Thus, $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge t\rho$ holds for $t = k+1$.

By induction, we show that $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge t\rho$.

(c) $\|\boldsymbol{w}(\boldsymbol{t})\|^2 = \|\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}(\boldsymbol{t})\| = \Big\|\big(\boldsymbol{w}(t-1) + y(t-1)\boldsymbol{x}(t-1)\big)^T\big(\boldsymbol{w}(t-1) + y(t-1)\boldsymbol{x}(t-1)\big)\Big\| = \|\boldsymbol{w}^{\boldsymbol{T}}(t-1)\boldsymbol{w}(t-1) + 2\boldsymbol{w}^{\boldsymbol{T}}(t-1)y(t-1)\boldsymbol{x}(t-1) + \boldsymbol{x}^{\boldsymbol{T}}(t-1)\boldsymbol{x}(t-1)\|$

Since $\boldsymbol{x}(t-1)$ is misclassified, $then\ y(t-1)\ \boldsymbol{w}^T(t-1)\boldsymbol{x}(t-1) < 0$

Then together with $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \ge \|\boldsymbol{x}+\boldsymbol{y}\|$, we have $\|\boldsymbol{w}(t)\|^2 < \|\boldsymbol{w}^{\boldsymbol{T}}(t-1)\boldsymbol{w}(t-1) + \boldsymbol{x}^{\boldsymbol{T}}(t-1)\boldsymbol{x}(t-1)\| \le \|\boldsymbol{w}^{\boldsymbol{T}}(t-1)\boldsymbol{w}(t-1)\| + \|\boldsymbol{x}^{\boldsymbol{T}}(t-1)\boldsymbol{x}(t-1)\| = \|\boldsymbol{w}(t-1)\|^2 + \|\boldsymbol{x}(t-1)\|^2$.

Thus, $\|\boldsymbol{w}(t)\|^2 \le \|\boldsymbol{w}(t-1)\|^2 + \|\boldsymbol{x}(t-1)\|^2$.

(d) Prove by induction.

Base case: When t = 0, $\|\boldsymbol{w}(t)\|^2 = 0 \le 0\times R^2$.

Induction step: Assume $\|\boldsymbol{w}(t)\|^2 \le tR^2$ for t = k where $k \in N$.

At $t = k+1$, $\|\boldsymbol{w}(k+1)\|^2 \le \|\boldsymbol{w}(k)\|^2 + \|\boldsymbol{x}(k)\|^2 \le kR^2 + \|\boldsymbol{x}(k)\|^2$.

Since $R = \max_{1\le n\le N} ||\boldsymbol{x}_n||$, then $R^2 \ge \|\boldsymbol{x}(k)\|^2$. Then $\|\boldsymbol{w}(k+1)\|^2 \le kR^2 + \|\boldsymbol{x}(k)\|^2 \le kR^2 + R^2 = (k+1)R^2$. Thus, $\|\boldsymbol{w}(t)\|^2 \le tR^2$ holds for $t = k+1$.

By induction, we prove that $\|\boldsymbol{w}(t)\|^2 \le tR^2$, where $R = \max_{1\le n\le N} ||\boldsymbol{x}_n||$.

(e) Since $\|\boldsymbol{w}(t)\|^2 \le tR^2$, then $\|\boldsymbol{w}(t)\| \le \sqrt{t}R$ and it follows $0 < \frac{1}{\sqrt{t}R} \le \frac{1}{\|\boldsymbol{w}(t)\|}$.

Since $\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^* \ge t\rho$, then $\frac{\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^*}{\|\boldsymbol{w}(t)\|} \ge \frac{t\rho}{\sqrt{t}R} = \sqrt{t}\cdot\frac{\rho}{R}$, and it follows $\frac{\left(\boldsymbol{w}^{\boldsymbol{T}}(t)\boldsymbol{w}^*\right)^2}{\|\boldsymbol{w}(t)\|^2} \ge$

$\frac{t\rho^2}{R^2}$. Then $t \leq \frac{\left(\boldsymbol{w}^T(t)\boldsymbol{w}^*\right)^2 R^2}{\|\boldsymbol{w}(t)\|^2\rho^2}$. Since $\boldsymbol{x}^T\boldsymbol{y} = ||\boldsymbol{x}||||\boldsymbol{y}||cos\theta \leq ||\boldsymbol{x}||||\boldsymbol{y}||$, then $\boldsymbol{w}^T(t)\boldsymbol{w}^* \leq ||\boldsymbol{w}^T(t)||||\boldsymbol{w}^*||$. Thus, we have $t \leq \frac{\left(\boldsymbol{w}^T(t)\boldsymbol{w}^*\right)^2 R^2}{\|\boldsymbol{w}(t)\|^2\rho^2} \leq \frac{\left||\boldsymbol{w}^T(t)|\right|^2 ||\boldsymbol{w}^*||^2 R^2}{\|\boldsymbol{w}(t)\|^2\rho^2} = \frac{R^2||\boldsymbol{w}^*||^2}{\rho^2}$

7. Problem 1.7 in LFD

(a) For $\mu = 0.05$: The probability of getting no head with 1 coin is $\binom{10}{0}(1-0.05)^{10} = 0.5987$

For 1000 coins: $P[\text{at least 1 of 1000 with 0 head}] = 1 - P[\text{all coins at lest one head}] = 1 - (1-0.5987)^{1000} \approx 1$

For 1000000 coins: $P = 1 - (1-0.5987)^{1000000} \approx 1$

For $\mu = 0.8$: The probability of getting no head with 1 coin is $\binom{10}{0}(1-0.8)^{10} = 1.024\times10^{-7}$.

For 1000 coins: $P[\text{at least 1 of 1000 with 0 head}] = 1 - P[\text{all coins at lest one head}] = 1 - (1-1.024\times10^{-7})^{1000} = 1.023\times10^{-4}$

For 1000000 coins: $P = 1 - (1-1.024\times10^{-7})^{1000000} = 0.09733$

(b) Since the two coins are independent, then using $P[A\ or\ B] = P[A] + P[B] - P[A\ and\ B] = P[A] + P[B] - P[A]P[B]$ and N =6 we have: for hoeffding bounds, $P[|v_1 - \mu_1| > \epsilon\ or\ |v_2 - \mu_2| > \epsilon] = P[|v_1 - \mu_1| > \epsilon] + P[|v_2 - \mu_2| > \epsilon] - P[|v_1 - \mu_1| > \epsilon]P[|v_2 - \mu_2| > \epsilon] \leq 2\times2e^{-2N\epsilon^2} = 4e^{-12\epsilon^2}$