CSCI 6100 Machine Learning From Data
Fall 2018

HOMEWORK 2

Daniel Southwick
661542908
southd@rpi.edu

Exercise 1.8

The number of samples is 10 and $v \leq 0.1$, so there either is no red marbles in the selected sample or there's only one red marbles. Thus the total Probability is:
$\binom{10}{0}$ x 0.9^0 x $0.1^10 + \binom{10}{0}$ x 0.9^1 x $0.1^9 = 9.1 \times 10^{-9}$.

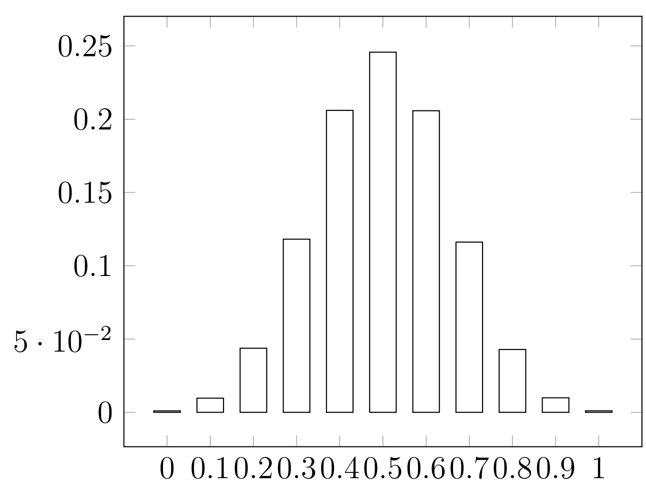
Exercise 1.9

Since $\mu = 0.9$, $\nu \leq 0.1$, we choose $\epsilon = 0.9 - 0.1 = 0.8$. Thus $\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N} = 2 \times e^{-2\times 0.8^2 \times 10} = 5.52 \times 10^{-6}$. Hoeffding Inequality provides an upper bound, so the result 5.52×10^{-6} is larger than the actual probability that was calculated in Exercise 1.8.

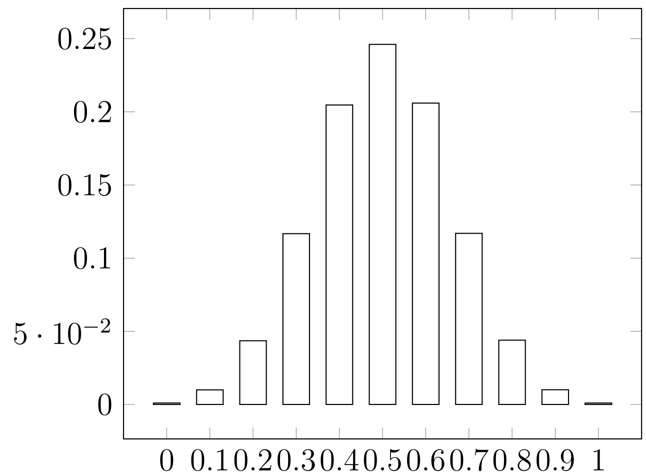
Exercise 1.10

(a) μ for all three coins are 0.5 since the fraction of heads for each coin is 0.5.
(b) The result is shown as follows:

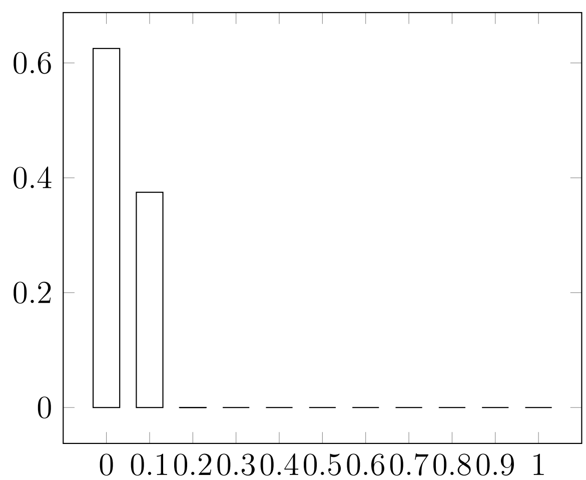
Histogram of the distribution of ν_1



Histogram of the distribution of ν_{rand}

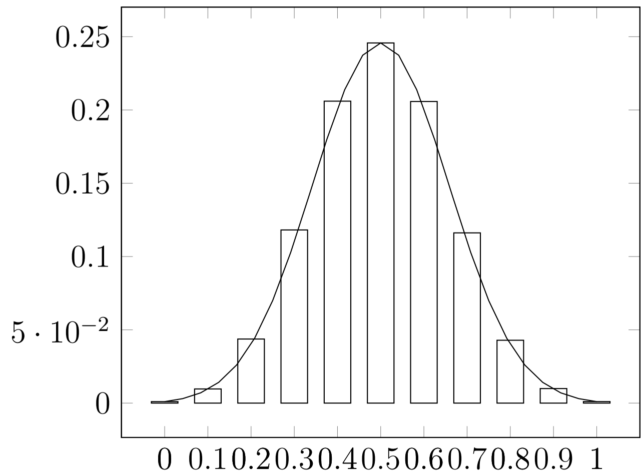


Histogram of the distribution of ν_{min}

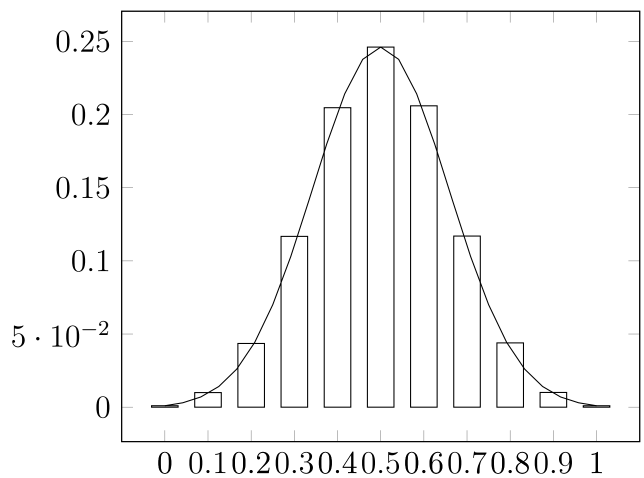


(c) The plots of the scaled estimators are as follows, we used $0.25 * exp(-20 * (x - 0.5)^2)$ instead of $2 * exp(-20 * (x - 0.5)^2)$ to visualize the pattern between the real distribution of the data and the estimators.

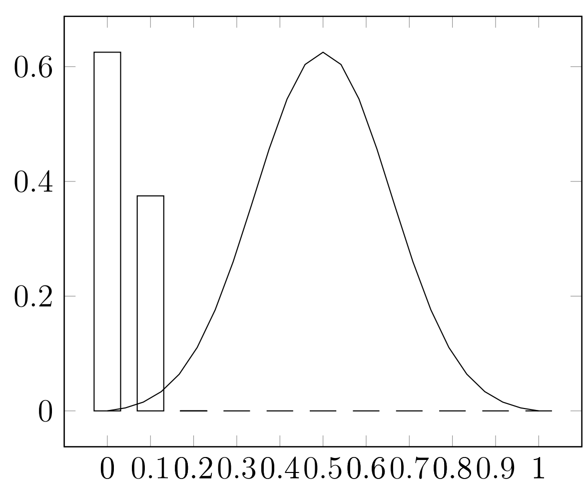
Scaled Estimator of the distribution of ν_1



Scaled Estimator of the distribution of ν_{rand}



Scaled Estimator of the distribution of ν_{min}



(d) Only coin c_1 and c_{rand} obey the Hoeffding bound, while c_{min} does not. c_1 follows the same pattern as c_{rand} since they are randomly chosen coins. They are both fixed coins during the simulation process, thus obeys the Hoeffding bound. However, coin c_{min} is not a fixed coin, it corresponds to the coin with minimum heads, since it's fixed before the data set was generated, the Hoeffding Inequality does not hold.

(e) c_1 and c_{rand} can represent multiple bins, in which each coin can be regarded as one bin, h is fixed before a data set is generated. c_{min} cannot represent bins, since it is not fixed, but it is closely related to the generated data set.

Exercise 1.11

(a) No. It cannot be guaranteed, there are 25 training examples, either case of $p > 0.5$, $p < 0.5$ or $p = 0.5$, the hypothesis produced by S does not performs better than random on any point outside $\mathbb{D}$.

(b) Yes, it is possible. If $p_{real} < 0.5$ and all training samples are labeled as $+1$, then the hypothesis produced by C performs better than S.

(c) Since $p = 0.9$, h_1 is better than h_2. So, if S produce a better hypothesis than C, S must pick h_1, which means there are at least 13 examples that agree with $f(x) = +1$. Therefore, probability of S produce a better hypothesis than C is equals to the probability of there are no less than 13 +1 with in the 25 examples given $p = 0.9$

$$\mathbb{P} = \sum_{i=13}^{25} \binom{25}{i} (0.9)^i (0.1)^{(25-i)} = 0.999999837$$

Exercise 1.12

(c) should be the best. The feasibility of learning is to verify $E_{out}(g) \approx E_{in}(g)$ and $E_{in}(g) \approx 0$. 4000 data points is too small for the above condition to hold. Even if $E_{in}(g) \approx 0$, there's no guarantee that $E_{out}(g) \approx E_{in}(g)$ with high probability.

Problem 1.3

(a) Since $\mathbf{w}^*$ is an optimal set of weights which separates the data , $\mathbf{w}^{*^T}\mathbf{x}_n$ and y_n has the same sign for n where $1 \leq n \leq N$. Thus, $\rho = \min\limits_{1\leq n\leq N} y_n(\mathbf{w}^{*^T}\mathbf{x}_n) > 0$ holds.

(b) Since we know the update rule is

$$w(t) = w(t-1) + y(t-1)x(t-1)$$

It follows that

$$\begin{aligned}
w^T(t)w^* &= w^T(t-1)w^* + w^* y(t-1)x(t-1) \\
&\geq w^T(t-1)w^* + min(y(t-1)(w^* x(t-1))) \\
&= w^T(t-1)w^* + \rho \\
&\geq w^T(t-2)w^* + \rho + \rho \\
&= w^T(t-2)w^* + 2\rho \\
&\geq w^T(t-3)w^* + 3\rho \\
&\geq ... \\
&\geq w^T(0)w^* + 6 * \rho
\end{aligned}$$

and we know $w(0) = 0$, thus

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

(c) Again, Since we know the update rule is

$$w(t) = w(t-1) + y(t-1)x(t-1)$$

It follows that

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

We know $w(t-1)$ misclassifies $x(t-1)$, thus

$$y(t-1)w^T(t-1)x(t-1) < 0$$

So,

$$\begin{gathered}||w(t)||^2 \leq ||w(t-1)||^2 + ||y(t-1)x(t-1)||^2 \\ = ||w(t-1)||^2 + ||x(t-1)||^2\end{gathered}$$

(d) From part (c), we have

$$\begin{gathered}||w(t)||^2 \leq ||w(t-1)||^2 + ||x(t-1)||^2 \\ \leq (||w(t-1)||^2 + ||x(t-1)||^2) + ||x(t-1)||^2 \\ \leq ... \\ \leq ||w(0)||^2 + \sum_{n=1}^{t-1} ||x(n)||^2 \\ = \sum_{n=1}^{t-1} ||x(n)||^2 \\ \leq tR^2\end{gathered}$$

and $R = \max_{1 \leq n \ leqN} ||\mathbf{x}_n||$

(e) From part (d), since

$$\begin{gathered}||w(t)||^2 \leq tR^2|| \\ ||w(t)|| \leq \sqrt{t}R \\ \frac{w^T(t)}{||w(t)||}w^* \geq \frac{t\rho}{\sqrt{t}R} \geq \sqrt{t}\frac{\rho}{R} \\ \sqrt{t} \leq \frac{R}{\rho}\frac{w^T(t)}{||w(t)||}w^*\end{gathered}$$

And due to Cauchy Inequality $\alpha^T\beta \leq ||\alpha||||\beta||$:

$$\begin{gathered}\sqrt{t} \leq \frac{R}{\rho}\frac{||w(t)||||w^*||}{||w(t)||} \\ \sqrt{t} \leq \frac{R||w^*||}{\rho} \\ t \leq \frac{R^2||w^*||^2}{\rho^2}\end{gathered}$$

Problem 1.7

(a)

For $\mu = 0.05$:

$$N = 1: \binom{10}{0} 0.05^0 (1-0.05)^{10} = 0.5987$$

$$N = 1{,}000: \mathbb{P} = 1 - (1 - \binom{10}{0} 0.05^0 (1-0.05)^{10})^{1000} = 1 - 2.6864 \times 10^{-397} \approx 1$$

$$N = 1{,}000{,}000: \mathbb{P} = 1 - (1 - \binom{10}{0} 0.05^0 (1-0.05)^{10})^{1000000} = 1 - 1.519 \times 10^{-396571} \approx 1$$

For $\mu = 0.8$:

$$N = 1: \binom{10}{0} 0.8^0 (1-0.8)^{10} = 1.024 \times 10^{-7}$$

$$N = 1{,}000: 1 - (1 - \binom{10}{0} 0.8^0 (1-0.8)^{10})^{1000} = 1.024 \times 10^{-4}$$

$$N = 1{,}000{,}000: 1 - (1 - \binom{10}{0} 0.8^0 (1-0.8)^{10})^{1000000} = 0.097$$

(b) The distribution of k is as follows:

Since only 2 coins, either $|v_1 - \mu_1| > \epsilon$ or $|v_2 - \mu_2| > \epsilon$.

1) $\epsilon \in [0, 1/6)$: coins are in $\{3\}$ thus $P = 1 - 0.3125^2 = 0.902$

2) $\epsilon \in [1/6, 2/6)$: coins are in $\{2, 3, 4\}$ $P = 1 - (0.2344 + 0.3125 + 0.2344)^2 = 0.390$

3) $\epsilon \in [3/6, 1)$: coins are in $\{1, 2, 3, 4, 5, 6\}$ all cases have been excluded so $P = 0$

k	0	1	2	3	4	5	6
$\|v-\mu\|$	3/6	2/6	1/6	0	1/6	2/6	3/6
$\mathbb{P}\|v-\mu\|$	0.0156	0.0938	0.2344	0.3125	0.2344	0.0938	0.0156

Probability of ϵ