Intro to ML HW 1

Matthew DeSantis

2023-09-12

1: LFD Problem 1.3

(a)

w^* is the optimal set of weights. This means it categorizes every point correctly, which in turn means that if a point's (x_n) corresponding label (y_n) is negative (that is to say, negative 1), then $w^{*T}x_n$ will also be negative, and vice versa if the label is positive. To put it another way, the signs of y_n and $w^{*T}x_n$ will match for all $n \subseteq N$. Since they have the same sign, their product will always be positive, and thus will always be > 0. If all members of the set are > 0, then clearly the minimum is also > 0, and thus ρ is > 0.

(b)

Let's assume that $w^T(t)w^* \geq w^T(t-1)w^* + \rho$ holds true for $t = k$. If this is true, then it is also true for $t = k + 1$.

$w^T(t+1)w^* \geq w^T(t+1-1)w^* + \rho \longrightarrow w^T(t+1)w^* \geq w^T(t)w^* + \rho$. Note that $w(t)$ denotes the weight vector at step t in PLA, so $w(t)$ and $w(t-1)$ are just the weight vector at any arbitrary adjacent steps. $w(t+1)$ and $w(t)$ are exactly the same then, they are just the weight vector at any adjacent steps (where the left side is one ahead of the right). Therefore $w^T(t)w^* \geq w^T(t-1)w^* + \rho$ is equivalent to $w^T(t+1)w^* \geq w^T(t)w^* + \rho$, meaning that if $w^T(t)w^* \geq w^T(t-1)w^* + \rho$ is true for $t = k$, then it's true for $t = k+1$. The only remaining step to prove that it is generally true is to prove the base case, where $t = 1$.

$w^T(1)w^* \geq w^T(1-1)w^* + \rho \longrightarrow w^T(1)w^* \geq w^T(0)w^* + \rho$. $w^T(0)$ is just a zero vector, so $w^T(0)w^* = 0$.

$w^T(1)w^* \geq \rho$. At step 1 of the PLA algorithm, $w(1)$ is just equal to one of the points x_n.ρ is the smallest of all $w^{*T}x_n$, so it follows that it must be less than or equal to $w^T(1)w^*$, since that is the same as one of the $w^{*T}x_n$ picked at random.

With each subsequent $w^T(t)w^*$, the value of $w^T(t)w^*$ goes up by at least ρ, so $w^T(t)w^* \geq t\rho$.

(c)

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

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

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

we can drop the $y(t-1)$ from the term $||y(t-1)x(t-1)||^2$ because all y are either 1 or -1, only changing the sign, but the value is squared so the sign will end up positive.
$||w(t)||^2 = ||w(t-1)||^2 + ||x(t-1)||^2 + 2y(t-1)w(t-1)^Tx(t-1)$

$||w(t)||^2 \geq ||w(t-1)||^2 + ||x(t-1)||^2$
$2y(t-1)w(t-1)^Tx(t-1)$ is always negative because x is misclassified, so by dropping it we know the left side is greater than or equal to.

(d)

Using induction. At t=0, obviously $0 \geq 0$. Assuming $||w(t)||^2 \leq R^2$, at t+1

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

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

by the definition of R^2, $||x(t-1)||^2 \leq R^2$

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

$||w(t)||^2 \leq tR^2$

(e)

From b, $w^T(t)w^* \geq t\rho$

$\frac{w^T(t)w^*}{||w(t)||} \geq \frac{t\rho}{||w(t)||}$

$\frac{w^T(t)w^*}{||w(t)||} \geq \frac{t\rho}{||w(t)||}$

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

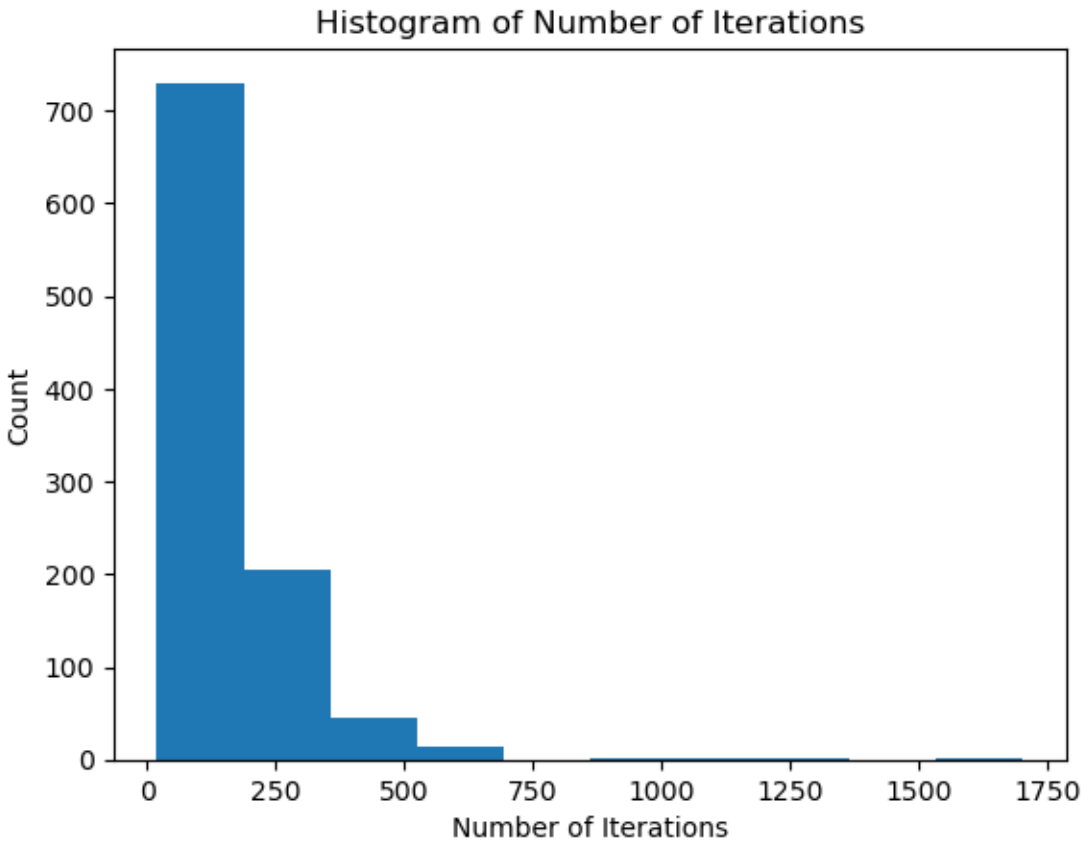
$\frac{w^T(t)w^*}{||w(t)||} \geq \sqrt{t}\frac{\rho}{R}$

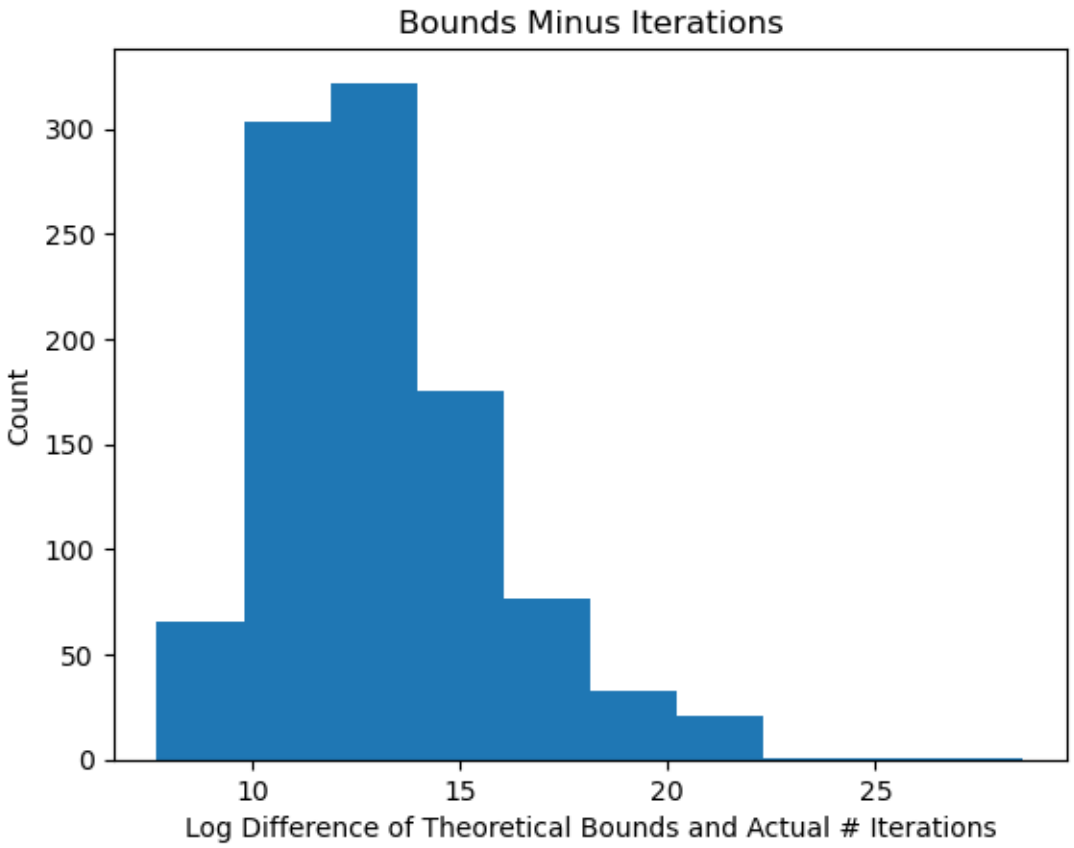
$\frac{w^T(t)w^*R}{||w(t)||\rho} \geq \sqrt{t}$

$\frac{(w^T(t)w^*)(w^T(t)w^{*T})R^2}{w(t)^Tw(t)\rho^2} \geq t$

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

2:


Histogram of Number of Iterations
Count
0
100
200
300
400
500
600
700
0
250
500
750
1000
1250
1500
1750
Number of Iterations


Bounds Minus Iterations
Count
300
250
200
150
100
50
0
10
15
20
25
Log Difference of Theoretical Bounds and Actual # Iterations

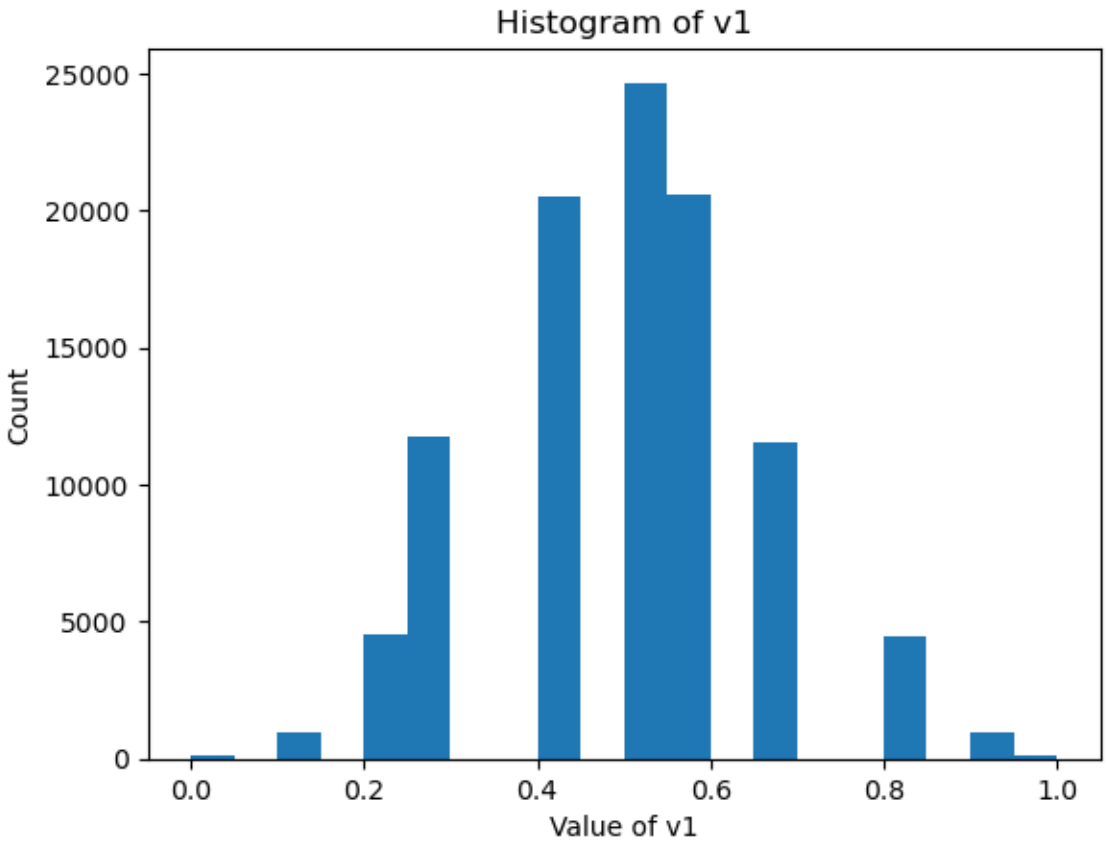
As can be seen in figure 2, the number of iterations it takes the PLA algorithm to successfully classify all points is always less than the bound, most often by a large number of iterations in the tens of thousands.

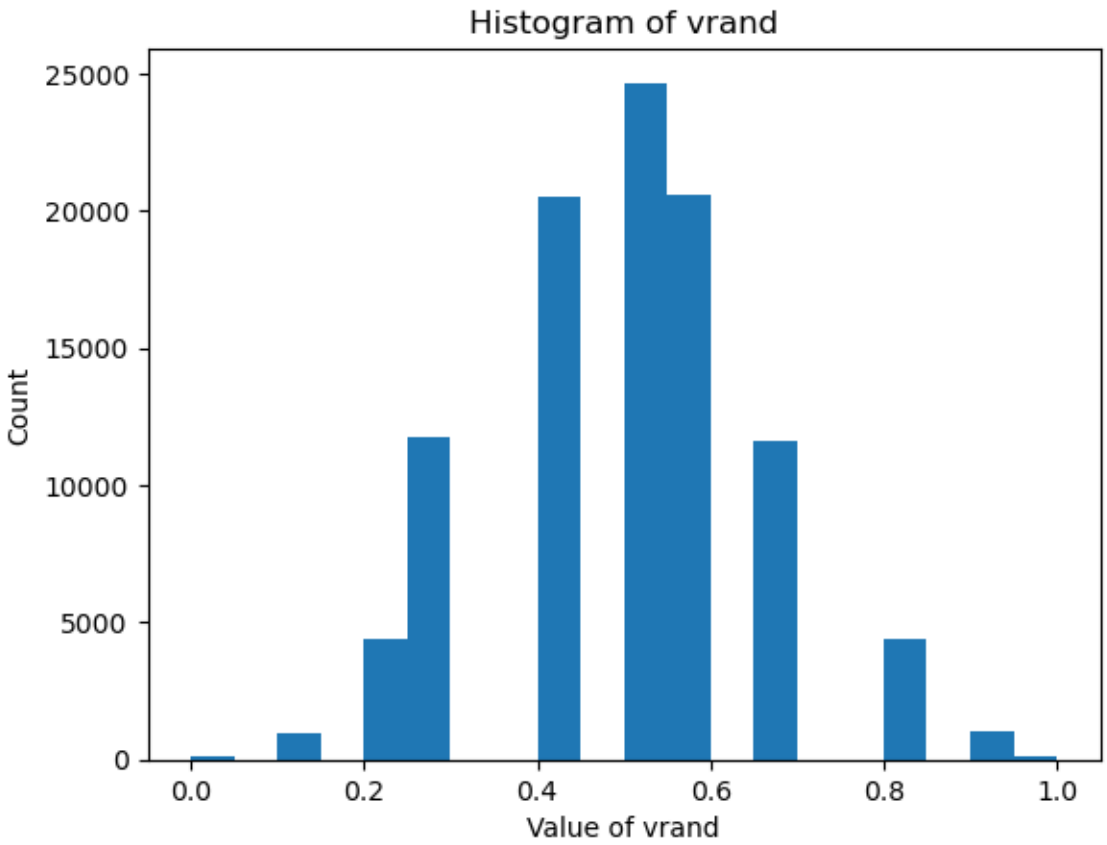
3: LFD Exercise 1.10 (a)-(d)

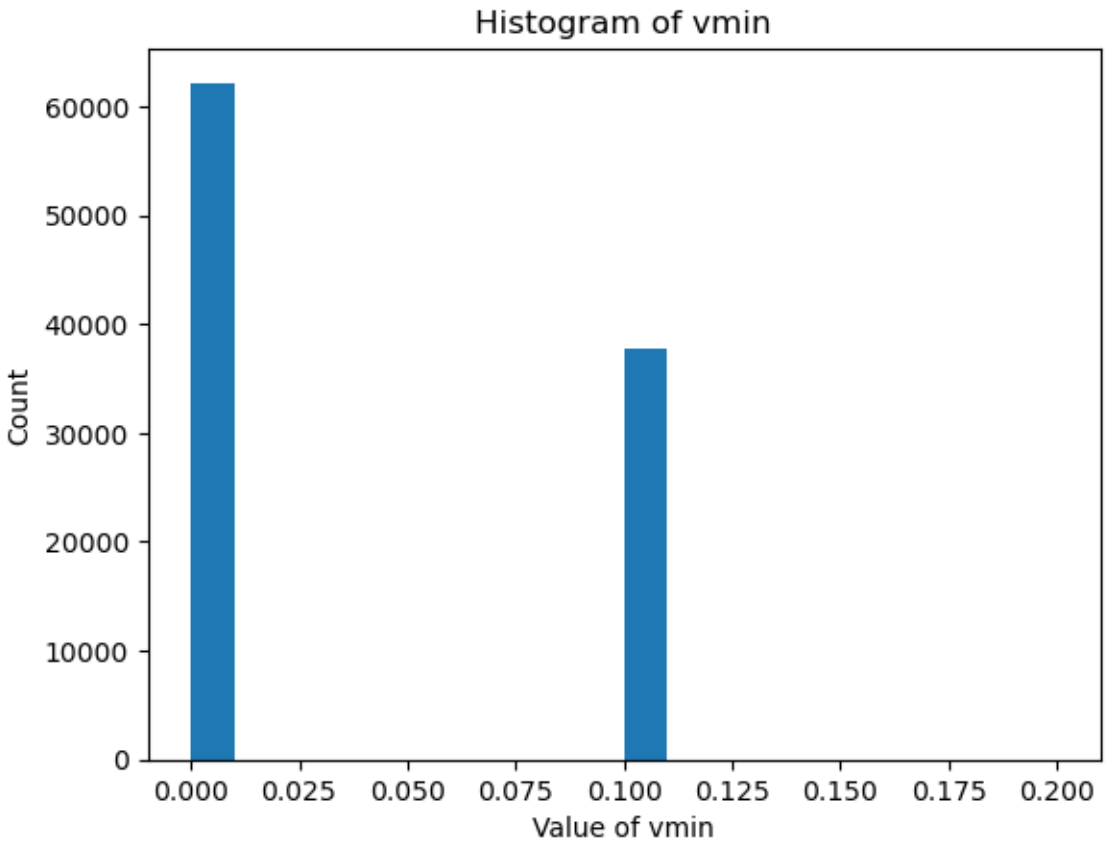
(a)

The expected value for each coin (assuming we code heads as 1 and tails as 0) is 0.5. This is true for all of the coins, so it's true for v_1, v_{rand}, and v_{min}.

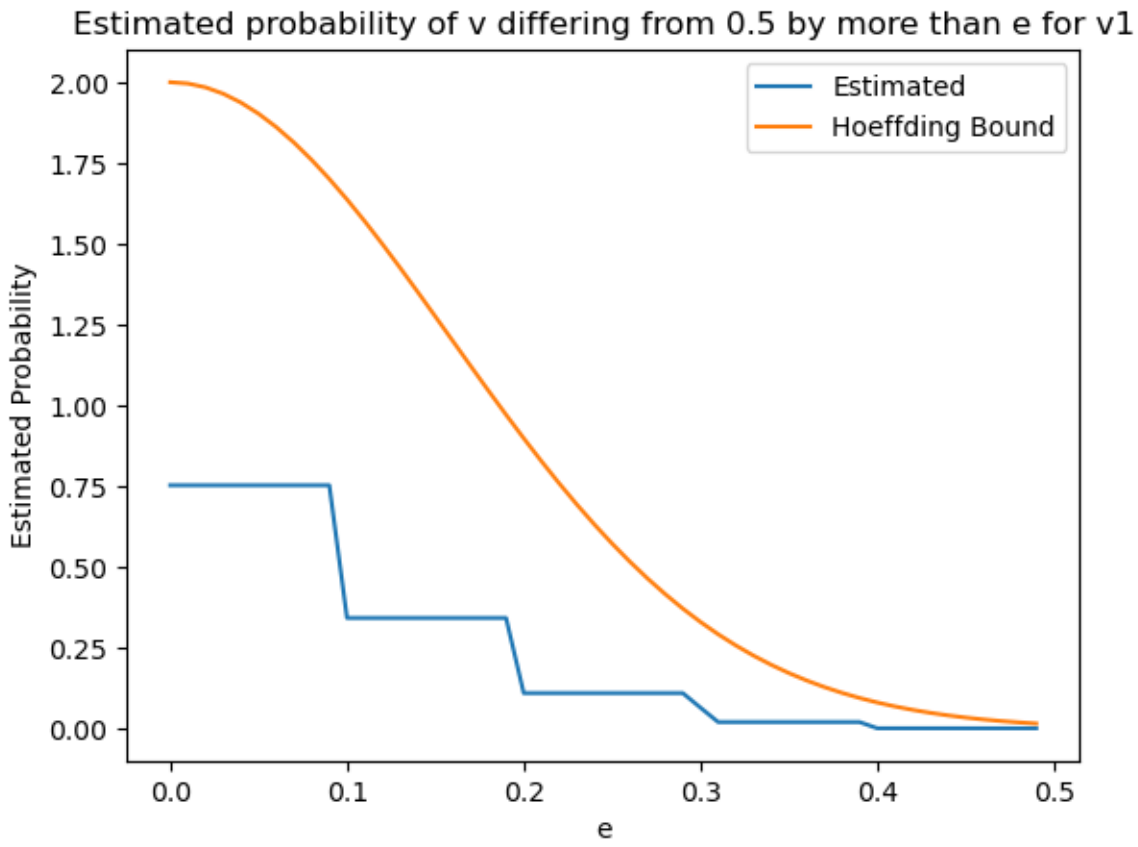
(b)

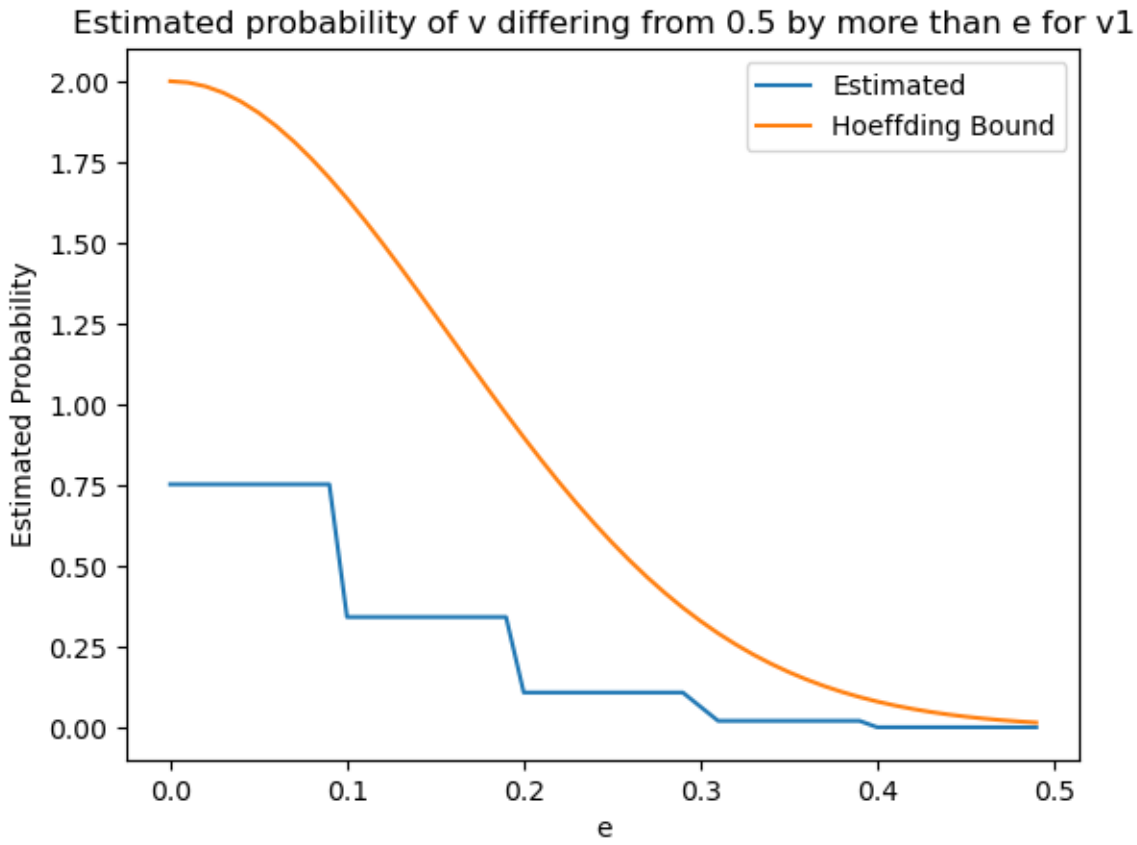


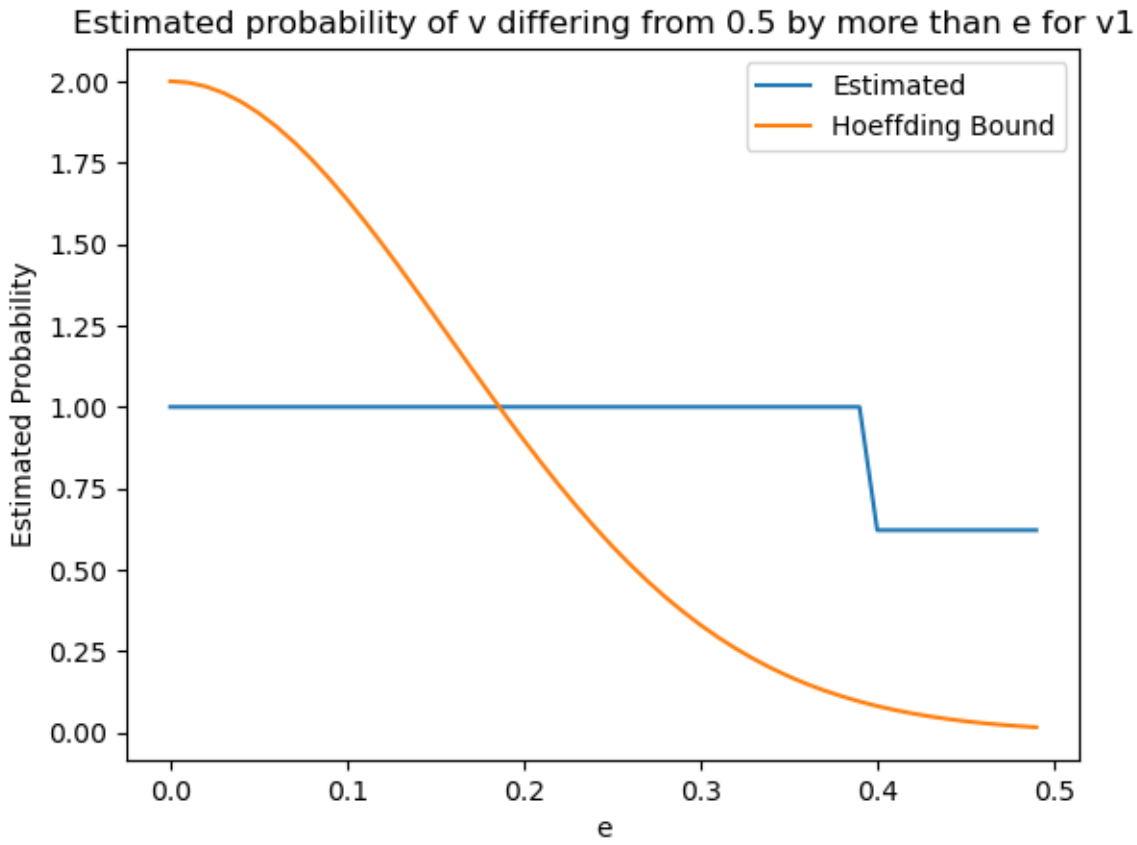

Histogram of vrand
25000
20000
15000
10000
5000
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
Value of vrand


Histogram of vmin
Count
60000
50000
40000
30000
20000
10000
0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Value of vmin

(c)


Estimated probability of v differing from 0.5 by more than e for v1
Estimated
Hoeffding Bound
Estimated Probability
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
e


Estimated probability of v differing from 0.5 by more than e for v1
Estimated
Hoeffding Bound
Estimated Probability
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
e


Estimated probability of v differing from 0.5 by more than e for v1
Estimated
Hoeffding Bound
Estimated Probability
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
e

(d)

As can be seen from the graphs, v1 and vrand obey the Hoeffding bound, but vmin doesn't. vmin doesn't obey the bound because with vmin, we are violating an assumption of Hoeffding's inequality by picking our hypothesis based on the results of our experiment, rather than "fixing" a hypothesis beforehand.

4: LFD Problem 1.8

(a)

Start by using the law of total expectation.

$P(t \geq \alpha)E(t|t \geq \alpha) + P(t < \alpha)E(t|t < \alpha) = E(t)$

t is non negative, so

$P(t \geq \alpha)E(t|t \geq \alpha) \leq E(t)$

α is also non negative, so

$\frac{P(t \geq \alpha)E(t|t \geq \alpha)}{a} \leq \frac{E(t)}{\alpha}$

$\frac{E(t|t \geq \alpha)}{a} P(t \geq \alpha) \leq \frac{E(t)}{\alpha}$

$\frac{E(t|t \geq \alpha)}{a}$ must be greater than or equal to 1, so

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

(b)

By part (a)

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

By definition of variance

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

(c)

u is the average of the N iid random variables $u_1, ..., u_N$, which have mean μ and variance σ^2. Therefore, u has mean μ and variance $\frac{\sigma^2}{N}$.

Therefore, the statement $P[(u - \mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$ is true by the result of (b)

5: LFD Problem 1.12

(a)

Let's minimize $\sum_{n=1}^{N}(h-y_n)^2$. To minimize, we take the derivative, set to zero, then solve.

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

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

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

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

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

$Nh - \sum_{n=1}^{N}y_n = 0$

$Nh = \sum_{n=1}^{N}y_n$

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

The second derivative is just 2, so we know this is a minimum rather than a maximum or an inflection point.

(b)

We'll take the same approach here.

$\frac{d}{dh}\sum_{n=1}^{N}|h-y_n|$

$\sum_{n=1}^{N}\frac{d}{dh}|h-y_n|$

$\sum_{n=1}^{N}sign(h-y_n) = 0$

This sum is only zero when the number of y_n greater than h is equal to the number that are less than h. That's the definition of the median. Taking the second derivative gives us zero, which implies an inflection point. However, this is due to the weirdness of the absolute value, it is in fact a minimum.

(c)

Adding an outlier would cause the mean to dramatically increase, going to infinity as the outlier does, but it would make no (or almost no) difference to the median, as the median only cares about the number of points higher or lower than it, not about how much higher or lower they are.

6: LFD Problem 2.3

(a)

$\mathrm{m(N)} = 2N$

$d_{VC} = 2$

(b)

$\mathrm{m(N)} = N^2 - N + 2$

$d_{VC} = 3$

(c)

$\mathrm{m(N)} = \frac{(N+1)(N)}{2} + 1$

$d_{VC} = 2$

7: LFD Problem 2.8

$2^{\sqrt{N}}$ cannot be a growth function as it is less than 2^N but is not polynomial.

$2^{\frac{N}{2}}$ cannot be a growth function for the same reason as above.

1+N is a possible growth function, as it is $\leq 2^N$ and returns an integer for all integer values of N.

$1 + N + \frac{N(N-1)}{2}$ is a possible growth function. N(N-1) is an even times an odd for all integer values of N, so it will always be an even number for all integer values of N, meaning that $\frac{N(N-1)}{2}$ will always be an integer and thus the expression as a whole will also always be an integer. It is also $\leq 2^N$ for all integer values of N.

2^N is obviously a valid growth function, it is the growth function of any hypothesis space which contains all possible dichotomies for all values of N.

$1 + N + \frac{N(N-1)(N-2)}{6}$ is not a valid growth function, as it does not obey the bound from theorem 2.4.