CSE 417

alex.huang

January 2020

1

1.1

Since we know $\rho = \min y_n(W*^T x_n)$, if we can prove all the $y_n(W*^T x_n) > 0$, we are done. If $y_n = 1$, since $W*$ is an optimal set of weights, we know that $W*^T x_n > 0$ must always be true. On the other hand, if $y_n = -1$, we know that x_n is on the other side of $W*$, and $W*^T x_n < 0$, which makes their product positive. Therefore, $\rho > 0$ is always true.

1.2

We use induction on t. If $t = 1$, we know that $W(0) = 0$, which leads to $W^T(1)W* \geq \rho$, and we know $W(1) = y(1)x(1)$. Since $\rho = \min y_n(W*^T x_n)$, and we know that $W^T(1)W* = y(1)x(1)^T W*$, which is bigger than ρ since it is one of the terms in ρ.

If the result is true on t, which means that $W^T(t)W* \geq W^T(t-1)W*+\rho$. By our procedure, we know that $W(t+1) = W(t) + y(t)x(t)$, so we can write $W^T(t+1)W* = W^T(t)W*+y(t)x^T(t)W* \geq W^T(t)W*+\rho$, that proves the theorem.

If we apply the theorem many times, we will have $W^T(t)W* \geq W^T(t-1)W*+\rho \geq W^T(t-2)W*+2\rho \geq \ldots \geq t\rho$.

1.3

Since we already know that $W(t) = W(t-1)+y(t-1)x(t-1)$, squaring each side gets us $\|W(t)\|^2 = \|W(t-1)\|^2 + \|y(t-1)x(t-1)\|^2 + 2y(t-1)x^T(t-1)W(t-1)$. By the hint we know that $y(t-1)(W^T(t-1)x(t-1)) \leq 0$. So, if we remove a non-positive side from the right, we will get $\|W(t)\|^2 \leq \|W(t-1)\|^2 + \|x(t-1)\|^2$.

1.4

By the result from the last part, we have $\|W(t)\|^2 \leq \|W(t-1)\|^2 + \|x(t-1)\|^2 \leq \|W(t-2)\|^2 + \|x(t-1)\|^2 + \|x(t-2)\|^2 \leq \ldots \leq \|x(t-1)\|^2 + \|x(t-2)\|^2 + \ldots + \|x(0)\|^2$, which is smaller or equal than tR^2, since all the x_n term are not bigger than R.

1.5

By b we have $W^T(t)W* \geq t\rho$, and by d (square root on both sides) we have $\|W(t)\| \leq \sqrt{t}R$. $\frac{formulab}{formulad}$ gets us $\frac{W^T(t)}{\|W(t)\|}W* \geq \frac{t\rho}{\sqrt{t}R} = \sqrt{t}\frac{\rho}{R}$.

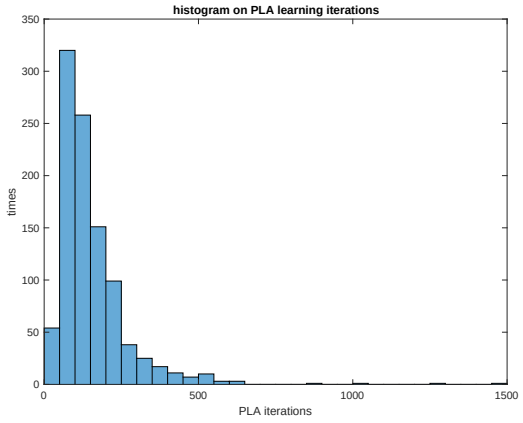
We also know that $\frac{W^T(t)W*}{\|W(t)\|\|W*\|} \leq 1$, since the upper half is vector dot multiplication, which also equals the lower half multiplied by $\cos\theta$.

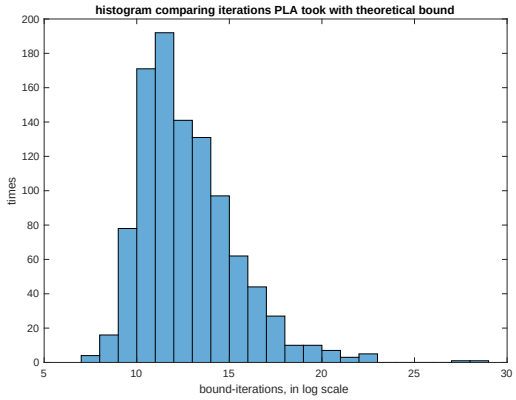
If we square the previous equation, we will get $t\frac{\rho^2}{R^2} \leq W*^2 \frac{W^T(t)^2}{\|W(t)\|^2}$. Divide the right by $(\frac{W^T(t)W*}{\|W(t)\|\|W*\|})^2$, we will get $t\frac{\rho^2}{R^2} \leq \|W*\|^2$, which means $t \leq \frac{\|W*\|^2R^2}{\rho^2}$. That proves the problem.

2

2.1

Below are the plots generated:





2.2

In the second chart, we can see that the log value of bound minus iterations is clustered around 9 to 17, with minor outliers close to 30. All the values are bigger than 5. Also, the chart skews to the right. The log difference is large indicates that the bound derived in the last question is a very loose bound, since we took the maximum value for R, and took the minimum value for ρ. Most of the times, we end up taking much less time than the bound predicted. However, it is still a very valid bound, since we can see that all values of bound is bigger than the number of actual iterations, which is consistent with our proof.

In the first chart we see that the number of iterations it took for us to learn via PLA mostly clusters in the region 0-500 (more than 90 percent of our attempts), with minor outliers up to 1500. We can therefore conclude that most the PLA gets to the goal in a relatively small amount of iterations, but we cannot be sure, since there are situations that PLA took a long time to finish.

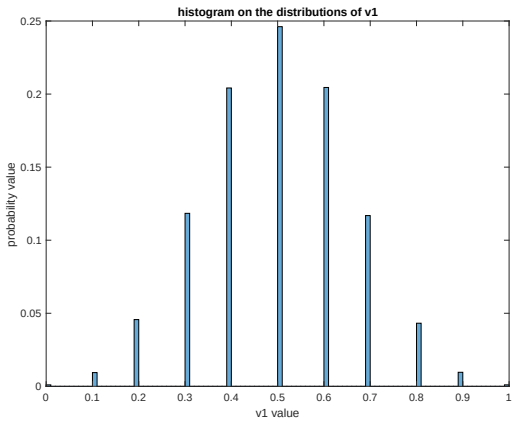
3

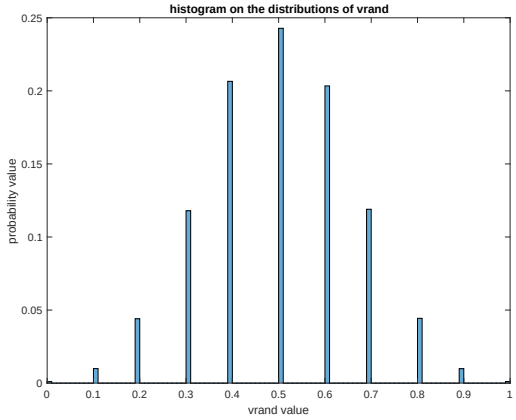
3.1

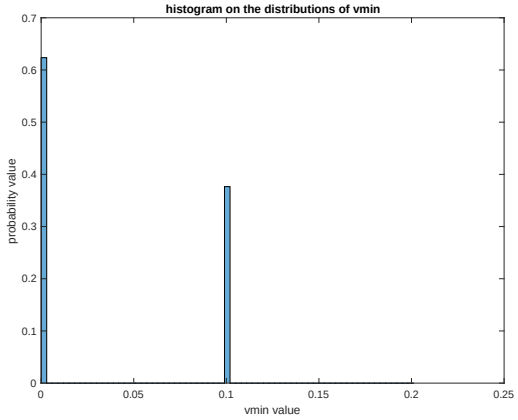
Since the three coins are fair coins, and μ is the probability of heads, we will get $\mu = 0.5$ for all three coins.

3.2

plots attached below:

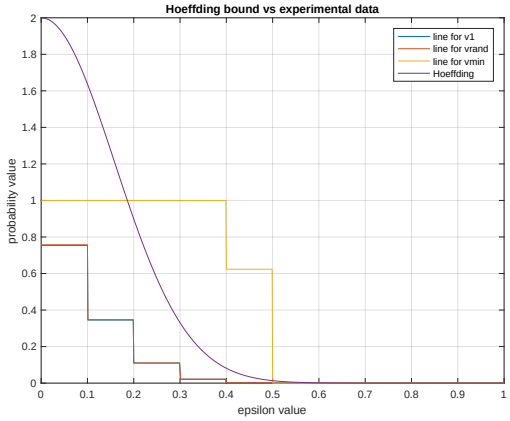



histogram on the distributions of vrand
probability value
vrand value
0
0.05
0.1
0.15
0.2
0.25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



3.3

see attached plot



3.4

v_1, v_{rand} obey the hoeffding bound, while v_{min} does not, since the function $P[|v - \mu| > \epsilon]$ for $vmin$ can be bigger than he Hoeffding bound. The reason for it is that the Hoeffding bound assumes "h is fixed before generating the data set". If I am allowed to change h after generating the data set, which in this case means choosign the minimum,

4

4.1

As t is a non-negative random variable, by definition we have $P[t \geq \alpha] = \int_{\alpha}^{\infty} P(x)dx$, with $P(x)$ being the probability density function. Also, by definition we have $E(t) = \int_{0}^{\infty} xP(x)dx \geq \int_{\alpha}^{\infty} xP(x)dx$. Since all values of x in $[\alpha, \infty)$ is greater or equal to α, it is easy to conclude that $\int_{\alpha}^{\infty} xP(x)dx \geq \int_{\alpha}^{\infty} \alpha P(x)dx = \alpha \int_{\alpha}^{\infty} P(x)dx$. Therefore, $E(t) \geq \alpha P[t \geq \alpha]$, which is equivalent

to $P[t \geq \alpha] \leq \frac{E(t)}{\alpha}$.

4.2

Since u is a random variable and μ, σ^2 is the mean and variance. By definition, we know that the expected value of $(u-\mu)^2$ is σ^2, and it is a valid, non-negative random variable. Therefore, we can plug it into the formula in part a, and get $P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$.

4.3

If we see u as a mean of $u_1 \ldots, u_N$, (since all the random variables have the same variance and mean) then its expected variance is know to be $\frac{\sigma^2}{N}$. Therefore, if $(u-\mu)^2$ is a random variable, its expected value is $\frac{\sigma^2}{N}$. Plugging this into our result in part a, and we get $P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$.

5

5.1

In this part we try to minimize $E_{in} = \sum_{n=1}^{N}(h-y_n)^2 = \sum_{n=1}^{N}(h^2+y_n^2-2hy_n) = Nh^2 + \sum_{n=1}^{N} y_n^2 - 2h(Nh_{mean})$. Since y_n is a fixed value, we want to minimize $Nh^2 - 2h(Nh_{mean})$, which is a parabola. If we take the derivative, we can see that the minimum of the parabola is at $h = h_{mean}$. Therefore, the problem is proven.

5.2

Suppose our estimate is not the sample median, then the number of data points smaller than it are different than the number of data points larger than it. If there are more data points smaller than it (we can use the same argument for the alternative situation), we can slightly decrease the estimate by $\epsilon > 0$ (a small value by our choice), so our absolute deviation from every point in the left decreases by ϵ, and deviation from points in the right increases by ϵ. The sum of total deviation, as a consequence, decreases since there are more points in the left. Therefore, our estimate is not the minimum value, which gives a contradiction. That proves the problem.

5.3

The first estimation goes to infinity. Since we now have $y_N \to \infty$, so does $h_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_n$. Therefore, the first estimation is going to be infinitely large.

The second estimation stays the same as long as $N > 1$. Since the median only need to make satisfy equal amount of data points are at most/at least the

median, regardless of individual values of data points, an outlier on the right does not affect the position of median. It will still stay in the middle data point (or between two middle data points).

Therefore, we can conclude that median is far less affected by an outlier in the data.

6

6.1

$m_H(N) = 2n$ in this part. For both positive and negative rays, as covered in class, there are $n+1$ positions to put the divider, which gets us $2n+2$, and we subtract two examples of double counting since we counted all points positive and all negative twice.

$m_H(1) = 2, m_H(2) = 4, m_H(3) = 6 < 8$, therefore the VC dimension is 2.

6.2

$m_H(N) = n(n-1) + 2$ in this part. We covered in class that positive intervals have $m_H(N) = \frac{n(n+1)}{2} + 1$, and if we double that, and minus the double counting ($2n$ because any situations similar to positive/negative ray we counted twice). That gives us the answer.

$m_H(1) = 2, m_H(2) = 4, m_H(3) = 8, m_H(4) = 14 < 16$, therefore the VC dimension is 3.

6.3

Let $x_D = \sqrt{x_1^2 + \ldots + x_d^2}$, and by definition, we see that this question is equivalent to choosing N points on a non-negative number line, and have the hypothesis set defined as a positive interval. If we choose discrete points on the number line, we will get $m_H(N) = \frac{n(n+1)}{2} + 1$.

$m_H(1) = 2, m_H(2) = 4, m_H(3) = 7$, therefore the VC dimension is 2.

7

7.1

$1 + N$ is possible. Let the hypothesis set defined as predicting -1 except for a single point on the dataset. Then we get $m_H(N) = 1 + N$ (could be $+1$ for any one point, or -1 for all points).

7.2

This growth function is equivalent to the one for positive interval. Therefore it is possible.

7.3

2^N is possible. Example include the convex set talked on class.

7.4

Both $2^{[\sqrt{N}]}$ and $2^{[\frac{N}{2}]}$ is not possible. We covered in class that if the growth function is smaller than 2^N (both are smaller here), it must be polynomial. They are not polynomial, and therefore are wrong.

7.5

It is not possible. Since $m_H(2) = 3$, we know that the VC dimension is 1, so therefore we conclude $m_H(N) \leq N^1 + 1$, which contradicts. Therefore, it is not possible.