1. (20 points) Assume that we have M slot machines in front of us. Each machine has an unknown probability of μ_m for returning one coin, and a probability of $1-\mu_m$ for returning no coin. For each of the time step $t = 1, 2, \ldots$, assume that we pull the machine $m = ((t-1) \bmod M) + 1$. After some $t > M$ time steps, we'd have pulled machine m for N_m times, and collected c_m coins from machine m. Note that $N_m \geq 1$ because $t > M$. Consider the following one-sided Hoeffding's inequality (which is slightly different from what we taught in class)

$$P\left(\mu > \nu + \epsilon\right) \leq \exp(-2\epsilon^2 N),$$

where ν, μ, ϵ, N have been defined in our class. Use the inequality above to prove that when given a fixed machine m and a fixed δ with $0 < \delta < 1$,

$$P\left(\mu_m > \frac{c_m}{N_m} + \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}}\right) \leq \delta t^{-2}.$$

$\mathcal{U} = \mu_m$

$\nu = \frac{c_m}{N_m}$

$N = N_m$

$\epsilon = \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}} \Rightarrow \ln t^2 - 2N\epsilon^2 = \ln\delta$

$-2N\epsilon^2 = \ln(\delta t^{-2})$

$\exp(-2N\epsilon^2) = \delta t^{-2}$

$\Rightarrow P(\mathcal{U} > \nu + \epsilon) \leq \exp(-2N\epsilon^2)$ #

2. (20 points) Continuing from Problem 1, prove that when $M \geq 2$, for all slot machines $m = 1, 2, \ldots, M$ and for all $t = M+1, M+2, \ldots$, with probability at least $1-\delta$,

$$\mu_m \leq \frac{c_m}{N_m} + \sqrt{\frac{\ln t + \ln M - \frac{1}{2}\ln\delta}{N_m}}.$$

You can use the magical fact that

$$\sum_{t=1}^{\infty} t^{-2} = \frac{\pi^2}{6}.$$

$$\varepsilon = \sqrt{\frac{\ln t + \ln M - \frac{1}{2}\ln\delta}{N_m}}$$

$$N_m \varepsilon^2 = \ln\left(\frac{Mt}{\delta^{\frac{1}{2}}}\right)$$

$$\exp(N_m \varepsilon^2) = Mt/\delta^{\frac{1}{2}}$$

$$\exp(-2N_m\varepsilon^2) M^2 = \delta t^{-2}$$

$$P\left(\mu_m \leq \frac{c_m}{N_m} + \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}}\right) = 1 - P\left(\mu_m > \frac{c_m}{N_m} + \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta}{N_m}}\right)$$

$$> 1 - \delta t^2$$

$$P\left(\mu_m \leq \frac{c_m}{N_m} + \sqrt{\frac{\ln t - \frac{1}{2}\ln\delta + \ln M}{N_m}}\right) > 1 - \delta t^{-2} M^{-2} \qquad (M^2 \geq 4)$$

$$> 1 - \frac{1}{4}\delta t^{-2} \qquad \left(\sum t^{-2} = \frac{\pi}{6}\right)$$

$$> 1 - \frac{\pi^2}{24}\delta$$

$$> 1 - \delta \quad \#$$

3. (20 points) Next, we illustrate what happens with multiple bins. Consider a special lottery game as follows. The game operates by having four kinds of lottery tickets placed in a big black bag, each kind with the same (super large) quantity. Exactly eight numbers $1, 2, \ldots, 8$ are written on each ticket. The four kinds are

- A: all even numbers are colored orange, all odd numbers are colored green
- B: all even numbers are colored green, all odd numbers are colored orange
- C: all small numbers (1-4) are colored orange, all big numbers (5-8) are colored green
- D: all small numbers (1-4) are colored green, all big numbers (5-8) are colored orange

Every person is expected to draw five tickets from the bag. A small price of 1450 is given if the five tickets contain "some number" that is purely green. What is the probability that such an event will happen?

1, 3 will are both even and small numbers ⇒ same color

2, 4 / 5, 7 / 6, 8 are same color for the same reason

when 1, 3 are all green on 5 cards : every cards are type A or D : $\frac{30}{4^5}$

$\frac{5!}{4!1!} \times 2 + \frac{5!}{3!2!} \times 2 = 30$

same for 2, 4 / 5, 7 / 6, 8

when 1, 3 / 2, 4 are all green on 5 cards : every cards are type D : $\frac{1}{4^5}$

same for 2, 4 / 6, 8 , 1, 3 / 5, 7 , 5, 7 / 6, 8
(all even) (all odd) (all big)

⇒ $\frac{30 \times 4}{4^5} + \frac{1 \times 4}{4^5} = \frac{1 \times 4}{4^5} = \frac{31}{256}$ #

4. (20 points) Continuing from Problem 3, a bigger price of three piggy banks will be given if the five tickets contain five green 2's. What is the probability that such an event will happen?

Hint: Each number can be viewed as a "hypothesis" and the drawn tickets can be viewed as the data. The E_{out} of each hypothesis is simply $\frac{1}{2}$ (You are welcome. ;-)). Problem 4 asks you to calculate the BAD probability for hypothesis 2; Problem 3 asks you to calculate the BAD probability for all hypotheses, taking the dependence into consideration.

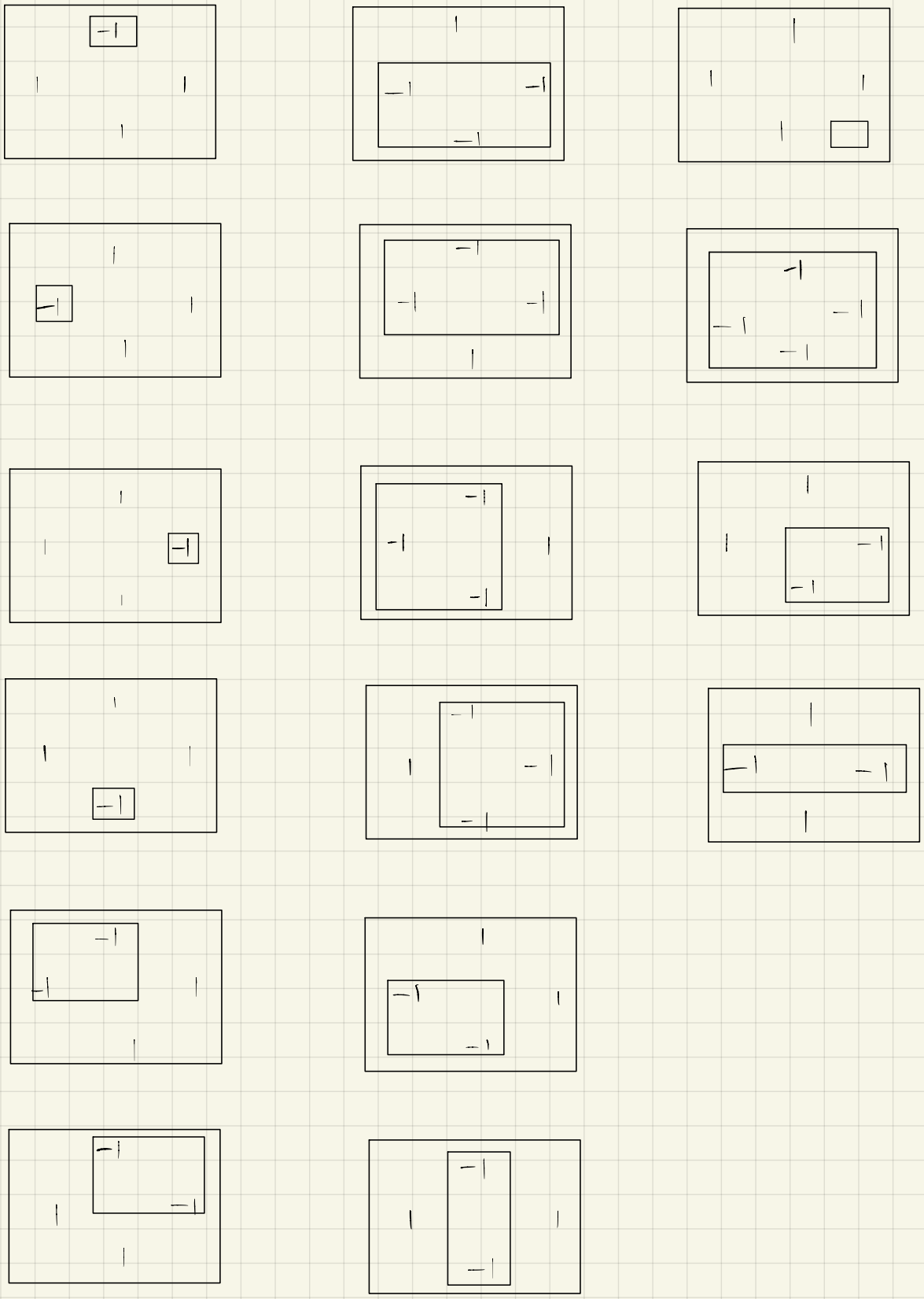
2, 4 are all green : $\frac{30}{45}$

2, 4 / 1, 3 are all green : all type D : $\frac{1}{45}$

2, 4 / 6, 8 are all green : all type B : $\frac{1}{45}$

$\frac{30}{45} + \frac{1}{45} + \frac{1}{45} = \frac{32}{45} = \frac{1}{32}$ #

5. (20 points) Consider the "negative rectangle" hypothesis set for $\mathcal{X} = \mathbb{R}^2$, which includes any hypothesis that returns -1 when $\mathbf{x}$ is within an axis-parallel rectangle and $+1$ elsewhere. Show that some set of 4 input vectors can be shattered by the hypothesis set. That is, the VC dimension of the hypothesis set is no less than 4.



there is 2^4 dichotomy for four inputs $\Rightarrow$ can be shatter by hypothesis set and VC dimension is no less than 4 #

6. (20 points) Consider a hypothesis set $\mathcal{H}$ for $\mathcal{X} = \mathbb{R}$ containing hypothesis with $2M + 1$ ($M \geq 1$) parameters. Each hypothesis $h(x)$ in $\mathcal{H}$ are defined by $s, a_1, b_1, a_2, b_2, \ldots, a_M, b_M$ that satisfies

- $s \in \{+1, -1\}$
- $a_m < b_m$, for $1 \leq m \leq M$;
- $b_m < a_{m+1}$, for $1 \leq m \leq M - 1$,

with

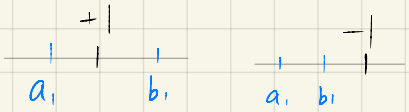
$$h_{s,\mathbf{a},\mathbf{b}}(x) = \begin{cases} s, \text{ if } a_m \leq x \leq b_m \text{ for some } 1 \leq m \leq M \\ -s, \text{ otherwise} \end{cases}$$

What is the VC dimension of $\mathcal{H}$? Prove your answer.

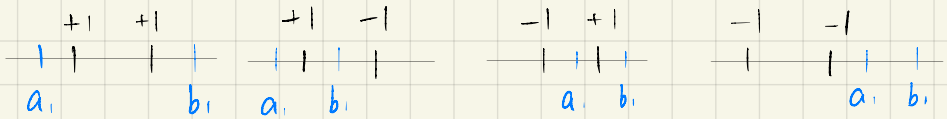
Hint: The positive intervals introduced in Lecture 5 correspond to $s = +1$ with $M = 1$.

when $M=1$ $s=+1$

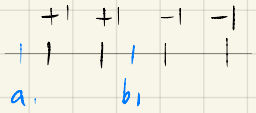
$N=1$



$N=2$

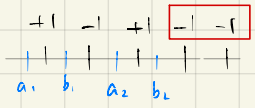


$N=4$

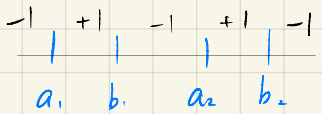


$\Rightarrow$ there must be at least 2 adjacent input with the same sign $\Rightarrow$ can't be shattered $\Rightarrow$ dvc = 3

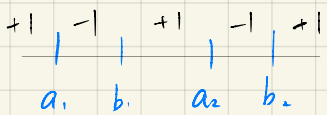
$\Rightarrow$ as long as we can make N inputs = +1 or -1 individually ($\cdots +1, -1, +1, -1, +1 \cdots$) seperated by a_m, b_m we can shatter N inputs

for $M=2$, 

$\Rightarrow$ can't be shattered



for $s=+1$, $M=2$



for $s=-1$, $M=2$

$\Rightarrow$ shatter 5 point $= 2M+1$ point

with $2M+1$ s. $a_1, b_1 \cdots a_m, b_m$ parameters,

we can shatter at most $(2M+1)$ inputs with $a_1, b_1 \cdots a_m, b_m$

$\Rightarrow$ dvc $= (2M+1)$

in terms of degree of freedom: dvc $\approx$ free variables, which is $(2M+1)$ in this case

7. (20 points) What is the growth function of origin-passing perceptrons on $\mathcal{X} = \mathbb{R}^2$? Those perceptrons are

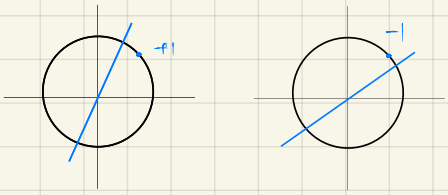
$$\mathcal{H}_0 \quad = \quad \{h\colon h(\mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2) \quad \textit{i.e.} \text{ perceptrons that pass the origin}\}$$

Prove your answer.

Hint: Consider putting your input vectors on the unit circle.

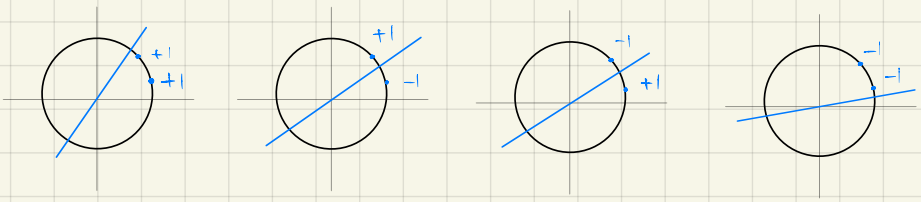
we normalize every x to put input on the unit circle

when $N=1$



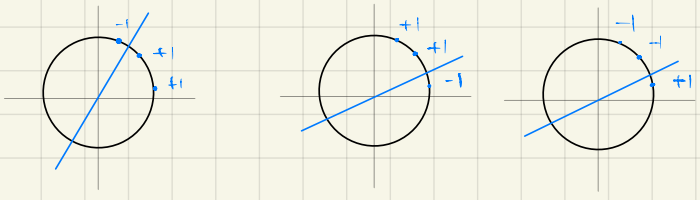
$\Rightarrow$ 2 dichotomy $= 2N$

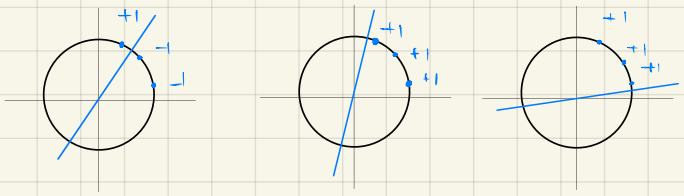
$N = 2$



$\Rightarrow$ 4 dichotomy $= 2N$

when $N = 3$





$\Rightarrow$ 6 dichotomy $= 2N$

An origin-passing perceptrons on unit circle can be viewed as positive and negative rays on 1D. the growth function with N inputs is $2(N-1) + 2 = 2N$

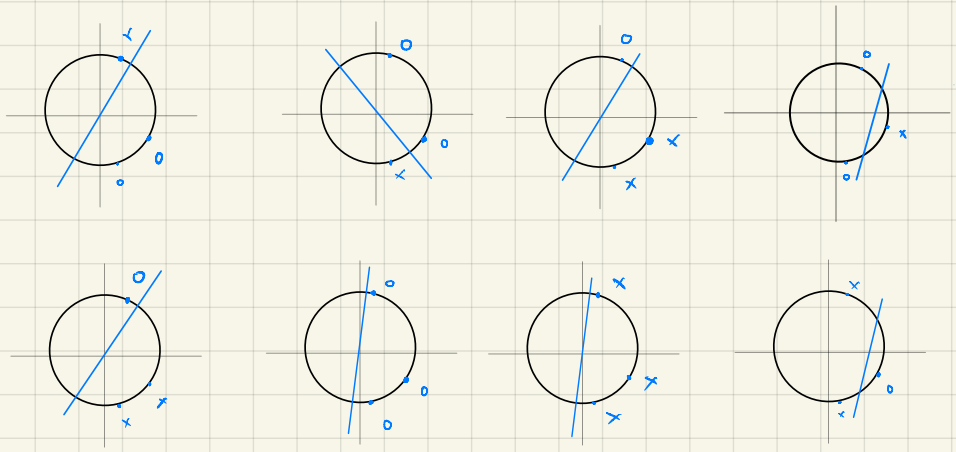
↳ all +1 or all -1

8. (20 points) For $\mathcal{X} = \mathbb{R}^2$, consider a hypothesis set $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$ that is a union of two types of perceptrons:

$$\begin{aligned} \mathcal{H}_0 &= \{h\colon h(\mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2) \quad \textit{i.e.} \text{ perceptrons that pass the origin}\} \\ \mathcal{H}_1 &= \{h\colon h(\mathbf{x}) = \text{sign}(w_1(x_1 - 1) + w_2(x_2 - 1)) \quad \textit{i.e.} \text{ perceptrons that pass } (1, 1)\} \end{aligned}$$

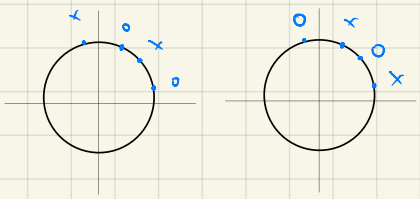
What is the VC dimension of $\mathcal{H}$? Prove your answer.

when $N = 3$



$N = 3$ can be shatter $\Rightarrow$ $V_{dc} \geq 3$

when $N = 4$, it's hard to find 2^4 dichotomy, if we can find two dichotomy that pass through two arbitrary points (looser constraint) can't shatter 4 inputs. we can say that two dichotomy that pass through $(0,0)$, $(1,1)$ can't shatter 4 inputs.



there is no dichotomy pass through an arbitrary points

can make $o \times o \times$, $\times o \times o$ on a unit circle

$\Rightarrow$ we can't shatter 4 input even with dichotomy that pass through an arbirary points

$\Rightarrow$ $H_0 = H_0 \cup H_1$ can't shatter $N = 4$

$\Rightarrow$ $d_{vc} = 3$ #

9. (20 points) In class, we taught about the learning model of "positive and negative rays" (which is simply one-dimensional perceptron) for one-dimensional data. The model contains hypotheses of the form:

$$h_{s,\theta}(x) = s \cdot \text{sign}(x - \theta).$$

You can take $\text{sign}(0) = -1$ for simplicity but it should not matter much for the following problems. The model is frequently named the "decision stump" model and is one of the simplest learning models. As shown in class, for one-dimensional data, the VC dimension of the decision stump model is 2.

In the following problems, you are asked to play with decision stumps on an artificial data set. First, start by generating a one-dimensional data by the procedure below:

a) Generate x by a uniform distribution in $[-1, 1]$.

b) Generate y by $y = \text{sign}(x) +$ noise, where the noise flips the sign with 10% probability.

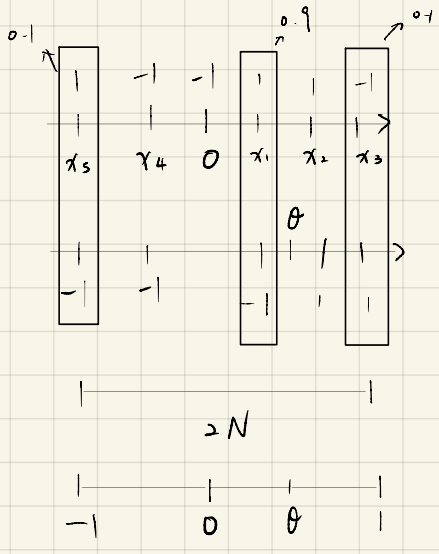
With the (x, y) generation process above, prove that for any $h_{s,\theta}$ with $s \in \{-1, +1\}$ and $\theta \in [-1, 1]$,

$$E_{\text{out}}(h_{s,\theta}) = 0.5 - 0.4s + 0.4s \cdot |\theta|.$$

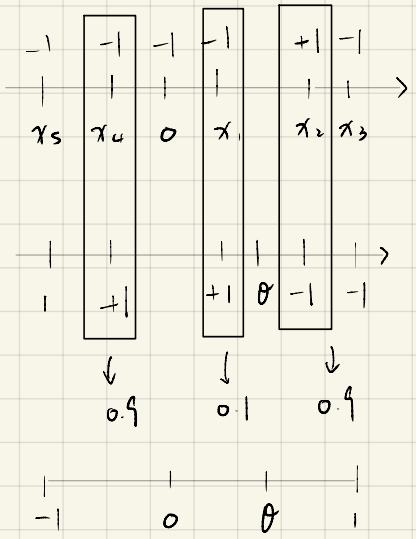
$$y = \begin{cases} \text{sign}(x) & P = 0.9 \\ -\text{sign}(x) & P = 0.1 \end{cases}$$

$$h_{s,\theta}(x) = s \cdot \text{sign}(x - \theta)$$

$s = +1$



$s = -1$



$$\Rightarrow P(h_{\theta,s}(x) \neq y) \begin{cases} 0.1 & x \in [-1, 0] \\ 0.9 & x \in [0, \theta] \\ 0.1 & x \in [\theta, 1] \end{cases}$$

$$\Rightarrow P(h_{\theta,s}(x) \neq y) \begin{cases} 0.9 & x \in [-1, 0] \\ 0.1 & x \in [0, \theta] \\ 0.9 & x \in [\theta, 1] \end{cases}$$

to sum up, $P(h_{\theta,s}(x) \neq y) = \begin{cases} 0.5 - 0.4s & x \in [-1, 0] \\ 0.5 + 0.4s & x \in [0, \theta] \\ 0.5 - 0.4s & x \in [\theta, 1] \end{cases}$

$\theta > 0$:

$$\frac{(0.5-0.45)\times 1\times N + (0.5+0.45)\times \frac{\theta}{1}\times N + (0.5-0.45)\times \frac{1-\theta}{1}\times N}{2N}$$

$$= \frac{0.5-0.45}{2} + \frac{(0.5+0.45)\theta}{2} + \frac{(0.5-0.45)(1-\theta)}{2}$$

$$= 0.5 - 0.45 + 0.25\theta + 0.25\theta - 0.25\theta + 0.25\theta$$

$$= 0.5 - 0.45 + 0.45\theta \quad (\theta > 0)$$

$\theta < 0$

$$\frac{(0.5-0.45)\frac{1+\theta}{1}\times N + (0.5+0.45)\frac{-\theta}{1}\times N + (0.5-0.45)\times 1\times N}{2N}$$

$$= 0.5 - 0.45 - 0.45\theta \quad (\theta < 0)$$

$$\Rightarrow \quad E_{in} = 0.5 - 0.45 + 0.45|\theta|$$

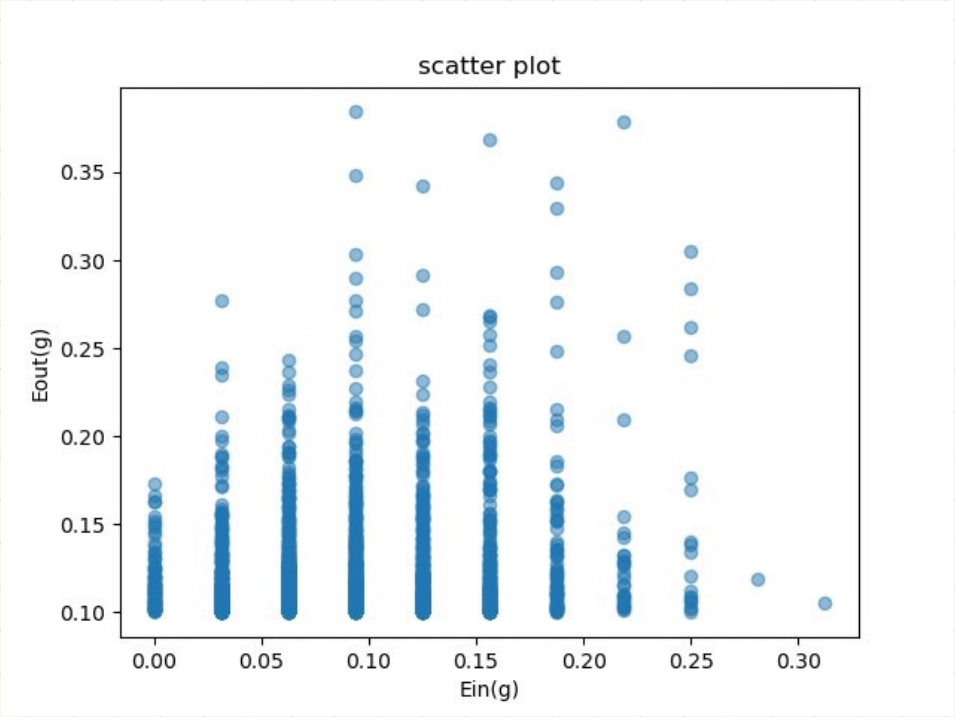
by Hoeffding's inequality, when N is large enough

$E_{in} \approx E_{out} \Rightarrow E_{out} = 0.5 - 0.45 + 0.45|\theta|$ #

10. (20 points, *) In fact, the decision stump model is one of the few models that we could minimize E_{in} efficiently by enumerating all possible thresholds. In particular, for N examples, there are at most $2N$ dichotomies (see the slides for positive rays), and thus at most $2N$ different E_{in} values. We can then easily choose the hypothesis that leads to the lowest E_{in} by the following decision stump learning algorithm.

> (1) sort all N examples x_n to a sorted sequence $x'_1, x'_2, \ldots, x'_N$ such that $x'_1 \leq x'_2 \leq x'_3 \leq \ldots \leq x'_N$
>
> (2) for each $\theta \in \{-1\} \cup \{\frac{x'_i + x'_{i+1}}{2} : 1 \leq i \leq N-1 \text{ and } x'_i \neq x'_{i+1}\}$ and $s \in \{-1, +1\}$, calculate $E_{\text{in}}(h_{s,\theta})$
>
> (3) return the $h_{s,\theta}$ with the minimum E_{in} as g; if multiple hypotheses reach the minimum E_{in}, return the one with the smallest $s \cdot \theta$.
>
> (*Hint: CS-majored students are encouraged to think about whether the second step can be carried out efficiently, i.e. $O(N)$, using dxxxxc pxxxxxxxxg instead of the naive implementation of $O(N^2)$.*)

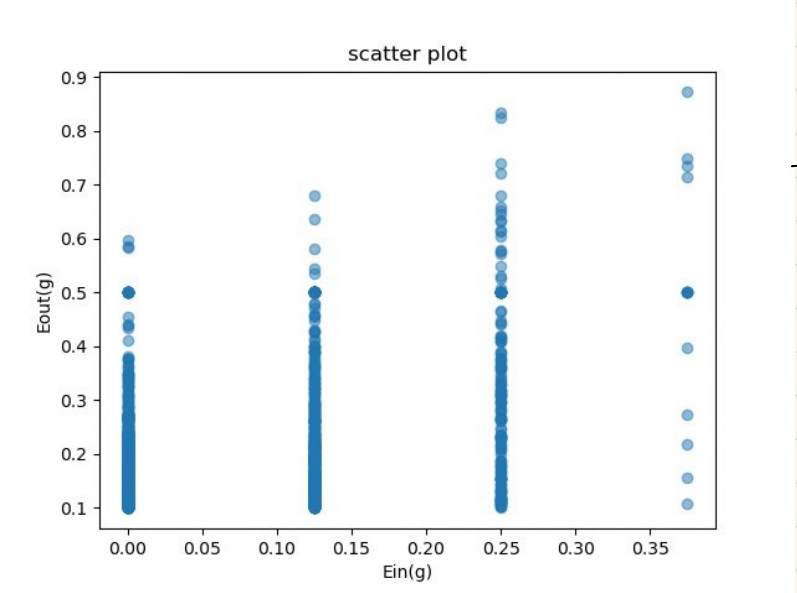
Generate a data set of size 32 by the procedure above and run the one-dimensional decision stump algorithm on the data set to get g. Record $E_{\text{in}}(g)$ and compute $E_{\text{out}}(g)$ with the formula in Problem 9. Repeat the experiment 2000 times. Plot a scatter plot of $(E_{\text{in}}(g), E_{\text{out}}(g))$, and calculate the median of $E_{\text{out}}(g) - E_{\text{in}}(g)$.




```
0.039114459302097 14
```

→ median

11. (20 points, *) Repeat Problem 10, but generate a data set of size 8 by the procedure instead. Plot a scatter plot of $(E_{\text{in}}(g), E_{\text{out}}(g))$, and calculate the median of $E_{\text{out}}(g) - E_{\text{in}}(g)$. Compare the scatter plot and the median value with those of Problem 10. Describe your findings.

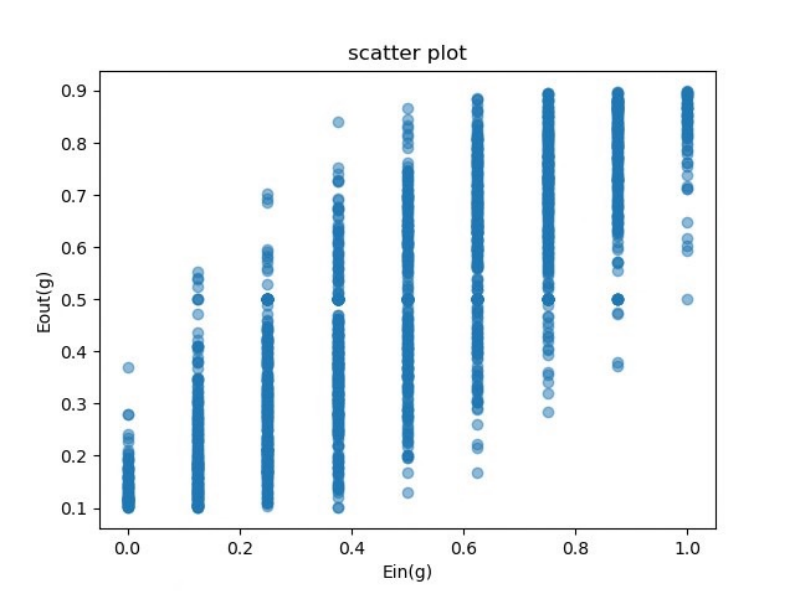


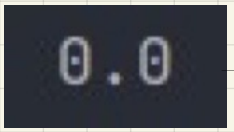
→ E is likely to be 4 values, which is less than N = 32

```
0.1224958501878991
```

→ median is larger than N = 32

12. (20 points, *) Repeat Problem 11, generate a data set of size 8 by the procedure above. Instead of running the decision stump algorithm, return a randomly chosen $h_{s,\theta}$ as g, with s uniformly sampled from $\{-1, +1\}$ and θ uniformly sampled from $[-1, 1]$. Record $E_{\text{in}}(g)$ and compute $E_{\text{out}}(g)$ with formula in Problem 9. Repeat the experiment 2000 times. Plot a scatter plot of $(E_{\text{in}}(g), E_{\text{out}}(g))$, and calculate the median of $E_{\text{out}}(g) - E_{\text{in}}(g))$. Compare the scatter plot and the median value with those of Problem 11. Describe your findings.





→ median

E is likely to be more values than problem 11, when Ein becomes bigger. Eout becomes larger too. Ein, Eout are positive correlation.

the median of Eout - Ein is much lower than problem 11.

Bonus: Perceptrons that Pass Special Points

13. (Bonus 20 points) Consider $\mathcal{H}$ being perceptrons in $\mathcal{X} = \mathbb{R}^d$. It is known, by the so-called Cover's Theorem, that the growth function is

$$m_{\mathcal{H}}(N) = 2\sum_{i=0}^{d} \binom{N-1}{i}.$$

See, for instance,

`https://web.mit.edu/course/other/i2course/www/vision_and_learning/perceptron_notes.pdf`

for its proof.

Now, assume that we require the perceptrons to pass *all* k anchor points for $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_k$, each being in $\mathbb{R}^d$ with $0 \leq k < d$. We shall call those perceptrons $\tilde{\mathcal{H}}$. What is the growth function $m_{\tilde{\mathcal{H}}}(N)$? Prove your answer.

Note: Problem 7 is a special case for $k = 1$ *and* $\mathbf{a}_1 = \mathbf{0}$.

To pass through k anchor impose k constraints on the perceptrons in H', each constriant reduces the free parameter. namely reduce the dimensionality of H' To d-k. $m_{H'}$ can be derived similar to Cover's Theorm. but reduce the dimensionality to d-k. To label N point can be thought as to choose subset of N points. For subset of size i there are C_i^{N-1} ways to choose from N-1 points. Summing all subset we get $m_{H'}(N) = 2\sum_{i=0}^{d-k} \binom{N-1}{i}$ #