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 as $(t - 1) \bmod M) + 1$. 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 $0 < \delta < 1$,

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

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 $1-\delta$,

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

You can use the magical fact that

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

Hint: The fact that we can upper-bound all μ_m confidently and simultaneously by $\frac{c_m}{N_m}$ plus a deviation term is the core technique for deriving the so-called upper-confidence bound algorithm for multi-armed bandits, which is an important algorithm for the task of online and reinforcement learning. The actual algorithm differs from what we do in Problem 1 by pulling the machine with the largest upper confidence bound in each iteration, instead of periodically going through each machine. Those who are interested can certainly search for more about this.

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

$$N_m \varepsilon^2 = \log t + \log M - \frac{1}{2}\log\delta$$

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

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

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

$$P\left(\mu_m \leq \frac{c_m}{N_m} + \sqrt{\frac{\log t - \frac{1}{2}\log\delta + \log M}{N_m}}\right) \geq 1 - \delta t^{-2} M^{-2}, \quad M \geq 2$$

$$\geq 1 - \delta \cdot t^{-2} \cdot \frac{1}{4} \quad \cdot \sum t^{-2} = \frac{\pi^2}{6}$$

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

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

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?

If (1,2,3,4) green → all C , $C_8^5 = 1$

(5,6,7,8) green → all D , $C_8^5 = 1$

(1,3,5,7) green → all A , $C_8^5 = 1$

(2,4,6,8) green → all B , $C_8^5 = 1$

(3,7) green → AEEEE, AAEEE, AAAEE, AAAAE $C_1^4 + C_2^4 + C_3^4 + C_4^5 = 4+24+\frac{5\times4}{2}\times2 = 30$

(1,5) green → AD arrange , the same above = 30

(2,4) green → BD arrange " = 30

(6,8) green → BC arrange " = 30

The probability of "same number are all green" should be $\frac{1+1+30\times4}{4^5} = \frac{121}{1024} = \frac{31}{256}$ #

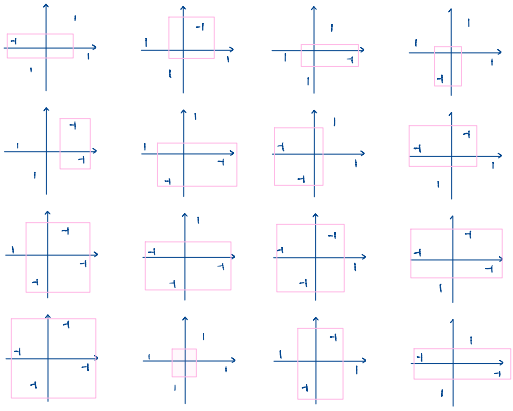
4. (20 points) Continuing from Problem 3, a bigger price of three piggy banks will be given if the five tickets contain five green 2's. What is the probability that such an event will happen?

Hint: Each number can be viewed as a "hypothesis" and the drawn tickets can be viewed as the data. The E_{val} 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.

P[five green 2's] = $\frac{\overset{(1,2,3,4)}{1} + \overset{(2,4,6,8)}{1} + \overset{(2,4)}{30}}{4^5} = \frac{32}{4^5} = \frac{1}{32}$ #

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.

↳ 2^4-is it.



Result: $m_{\mathcal{H}}(4) = 2^4$ for 4 inputs,
that is to say: 4 inputs can be shattered.
dvc isn't 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 $1 \leq m \leq M - 1$,

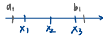
with

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

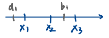
What is the VC dimension of $\mathcal{H}$? Prove your answer.

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

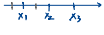
For $N=3$, $M=1$, $m_{\mathcal{H}}(3) = 8$, shattered.



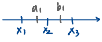
$\{x_1, x_2, x_3\} = \{s, s, s\}$, $s \in \{+1, -1\}$



$\{x_1, x_2, x_3\} = \{s, s, -s\}$, $s \in \{+1, -1\}$



$\{x_1, x_2, x_3\} = \{s, -s, -s\}$, $s \in \{+1, -1\}$

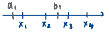


$\{x_1, x_2, x_3\} = \{-s, s, -s\}$, $s \in \{+1, -1\}$

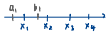
For $N=4$, $M=1$, $m_{\mathcal{H}}(4) = 14$, can't shattered.

$\{x_1, x_2, x_3, x_4\} = \{s, s, s, s\}$, $s \in \{+1, -1\}$

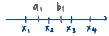
$\{x_1, x_2, x_3, x_4\} = \{s, s, s, -s\}$, $s \in \{+1, -1\}$



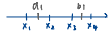
$\{x_1, x_2, x_3, x_4\} = \{s, s, -s, -s\}$, $s \in \{+1, -1\}$



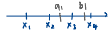
$\{x_1, x_2, x_3, x_4\} = \{s, s, s, s\}$, $s \in \{+1, -1\}$



$\{x_1, x_2, x_3, x_4\} = \{s, s, s, s\}$, $s \in \{+1, -1\}$



$\{x_1, x_2, x_3, x_4\} = \{s, -s, -s, s\}$, $s \in \{+1, -1\}$



$\{x_1, x_2, x_3, x_4\} = \{s, s, s, s\}$, $s \in \{+1, -1\}$

If h can shattered N points → $m_{\mathcal{H}}(N) = 2^N$, every x_i for $0 < i < N$ has two choices $\{s=+1\}$ or $\{s=-1\}$

That is, if h can shatter some N points → h has $(N-1)$ parameters to build N intervals such that x_i, x_{i+1} won't be forced to be in the same intervals.
x_i has the freedom to choose $\{s=+1\}$ or $\{s=-1\}$

For $N=5$, $M=2$, $m_{\mathcal{H}}(5) = 5$, shattered

$\{x_1, x_2, x_3, x_4\} = \{-s, s, -s, s, -s\}$

using 4 parameters, H could shatter 5 inputs.

The result can also refer to the conclusion in class: parameters creates degree of freedom → dvc ≈ free parameters.

In conclusion, hypothesis with $(2M+1)$ parameters

$(2M+1)$ intervals.
→ at most can shatter $2M+1$ data
→ dvc = $2M+1$ #

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

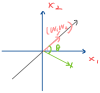
$$\mathcal{H}_0 = \{h\colon h(\mathbf{x}) = \text{sign}(w_1x_1 + w_2x_2) \quad \textit{i.e.}\text{ perceptrons that pass the origin}\}$$

Prove your answer.

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

w_1, w_2 : weight

x_1, x_2 : variables (two characteristics)



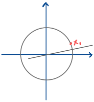
① $(w_1, w_2) \cdot (x_1, x_2) = \|w\| \|x\| \cos\theta$

② $w_1x_1 + w_2x_2 = 0$, $m = -\frac{w_1}{w_2}$

if $w_1x_1 + w_2x_2 > 0$

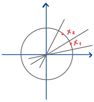
if $w_1x_1 + w_2x_2 < 0$

when $N=1$, $m_H(3)=2$



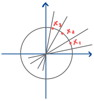
	x_1
h_1:	o
h_2:	x

when $N=2$, $m_H(2)=4$



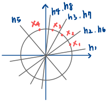
	x_1	x_2
h_1:	o	o
h_2:	o	x
h_3:	x	x
h_4:	x	o

when $N=3$. $m_H(3)=6$



	x_1	x_2	x_3
h_1:	x	x	x
h_2:	o	x	x
h_3:	o	o	x
h_4:	o	o	o
h_5:	x	o	o
h_6:	x	x	o

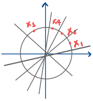
when $N=4$, $m_H(4) = 8 = m_H(3)+2$



	x_1	x_2	x_3	x_4
h_1:	x	x	x	x
h_2:	o	x	x	x
h_3:	o	o	x	x
h_4:	o	o	o	x
h_5:	o	o	o	o
h_6:	x	o	o	o
h_7:	x	x	o	o
h_8:	x	x	x	o

when $N=4$,

x_4 is added in the middle.



	x_1	x_2	x_4	x_3
h_1:	x	x	x	x
h_2:	o	x	x	x
h_3:	o	o	x	x
h_4:	o	o	o	x
h_5:	o	o	o	o
h_6:	x	o	o	o
h_7:	x	x	o	o
h_8:	x	x	x	o

如果 x_N 加在中間，

在 $H(N-1)$ 裡有 $m_H(N-1)$ 种情形：

x_N 必須要和左右兩边的其同號，

(也就是沒有自由度)；

x_N 可以有正負号的選擇只有在左右不同号時，

→ 多製造 2 种 hypothesis.

o, x_N, x or x, x_N, o

→ $m_H(N) = m_H(N-1)+2$

$N+1$, $m_H(N) = m_H(N-1)+2$

x_4 can be see as the point following x_3 → $\text{sign}(x_4) = \text{sign}(x_3)$, h_i for $i=1,2,\ldots,6$

when h_7, h_8 , $\text{sign}(x_4) = -\text{sign}(x_3)$,

which will create two more situation

→ $m_H(3)+2 = m_H(4)$

Using mathematical Induction , $m_H(3)=6$. $m_H(4)=8$

Assume that $m_H(N) = 2N$

$m_H(N-1) = m_H(N) - 2$

$2(N-1) = 2N-2$, $m_H(N) = 2N$ 成立.

→ when N , m_H

x_1	x_2	⋯	x_N	
x	x	___	x	N
o	x	___	x	
o	o	x ___	x	
o	o	o ___	x	
o	o	___	o	N
x	o	___	x	
	⋮			

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:

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

$\mathcal{H}_1 = \{h : h(\mathbf{x}) = \text{sign}(w_1(x_1 - 1) + w_2(x_2 - 1))$ *i.e.* perceptrons that pass $(1, 1)\}$

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

dvc → find the min number of x that could be shattered.

N=2

shattered

(Using proof of q9)

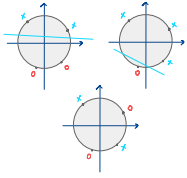
N=3

x_1	x_2	x_3
o	o	o
x	o	o
x	x	o
x	x	x
o	x	x
o	o	x
x	o	x
o	x	o

Shattered by $H = H_0 \cup H_1$

Next, we assume $\mathcal{H}' = \mathcal{H}_0' \cup \mathcal{H}_1'$ with no limitation on passing (0,0) and (1,1)

in other words, $\mathcal{H}'$ (looser) is the upper bound of $\mathcal{H}$ (limited)



$\mathcal{H}_0'(x_1, x_2, x_3, x_4) = \{$ oooo, xooo, xxoo, xxxo, oxxx, ooxx, ooox, xxxx $\}$

$\mathcal{H}_1'(x_1, x_2, x_3, x_4) = \{$ xoox, oxxo, xxox, ooxo $\}$

There aren't any $\mathcal{H}$ such that $\{x_1, x_2, x_3, x_4\} = \{$oxxo, xoox$\}$

when N=9, $\{x_1, x_2, x_3, x_4, x_5\}$ couldn't be shattered by $\mathcal{H}'$

→ the breakpoint k of $\mathcal{H}'$, k = 4

Since $\mathcal{H}'$ is the upper bound of $\mathcal{H}$ → $\mathcal{H}$ couldn't shatter N=4

→ dvc = 3 a

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.

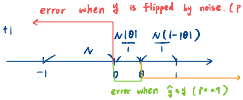
With the (x, y) generation process above, prove that for any $h_{s,\theta}$ with $s \in \{-1, +1\}$ and $\theta \in [-1, 1]$,

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

error when y is flipped by noise. ($P=0.1$)

$$P[y \neq \hat{y} \mid x] = \begin{cases} 0.1, & -1 \leq x < 0 \\ 0.9, & 0 \leq x \leq \theta \\ 0.1, & \theta < x \leq 1 \end{cases}$$

$s = +1$



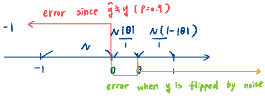
$$E_{in}(\theta, s) = \frac{0.1N + 0.9 \cdot (N\theta) + 0.1 \cdot N(1-\theta)}{2N}$$

$$= \frac{1N + 0.8N\theta}{2N} = 0.1 + 0.4\theta$$

error since $\hat{y} \neq y$ ($P=0.9$)

$$P[\hat{y} \neq y \mid x] = \begin{cases} 0.9, & -1 \leq x \leq 0 \\ 0.1, & 0 < x < \theta \\ 0.9, & \theta \leq x < 1 \end{cases}$$

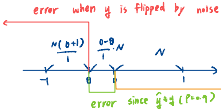
$s = -1$



$$E_{in}(\theta, s) = \frac{0.9N + 0.1 \cdot (N\theta) + 0.9N(1-\theta)}{2N} = \frac{1.8N + 0.8N\theta}{2N} = 0.9 + 0.4\theta$$

error when y is flipped by noise ($P=0.1$)

$$P[\hat{y} \neq y \mid x] = \begin{cases} 0.1, & -1 \leq x < \theta \\ 0.9, & \theta \leq x < 0 \\ 0.1, & 0 \leq x \leq 1 \end{cases}$$



$s = +1 \wedge \theta < 0$

$$E_{in}(\theta, s) = \frac{0.1N(\theta+1) + 0.9(-\theta)N + 0.1N}{2N} = \frac{0.2N + 0.8(-\theta)}{2N} = 0.1 + 0.4(-\theta)$$

$\rightarrow E_{in}(\theta, +1) = 0.1 + 0.4\theta$, $\theta > 0$ $\rightarrow$ $E_{in}(\theta, s) = 0.5 + 0.4s + 0.4|\theta|$

$E_{in}(\theta, -1) = 0.9 + 0.4\theta$, $\theta > 0$

$E_{in}(\theta, +1) = 0.1 + 0.4(-\theta)$, $\theta < 0$

By Hoeffdings inequality. for any fixed $h_{\theta,s}(x)$, if N is large enough

$E_{in}(h) \approx E_{out}(h)$

in conclusion, $E_{out}(h_{s,\theta}) = 0.5 + 0.4s + 0.4|\theta|$ #

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.

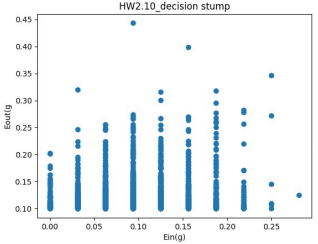
(1) sort all N examples x_n to a sorted sequence $x'_1, x'_2, \ldots, x'_N$ such that $x'_1 \leq x'_2 \leq x'_3 \leq \ldots \leq x'_N$

(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 $s \cdot \theta$.

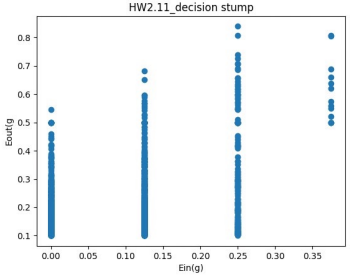
(*Hint: CS-majored students are encouraged to think about whether the second step can be carried out efficiently, i.e. $O(N)$, using* dzzzzzz pzzzzzzzzzg *instead of the naive implementation of $O(N^2)$.*)

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 $E_{\text{out}}(g) - E_{\text{in}}(g)$.



Median = 0.0383

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



Median = 0.1240

The median of data set =6 is smaller than the median of data set = 32.

The consequence is similar to Hoffoeding's inequality, which indicates that when N become larger (N= the number of data), Ein will be closer to Eout.

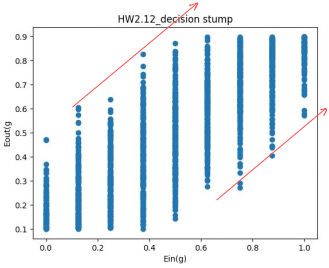
Focusing on the right side of the plot, when Ein>0.35, Eout >0.4;
to sum up, if Ein is large, Eout is highly possible to be large.

In most of cases, Ein is slightly smaller than Eout, which also explain the meaning if Both.

Ein = the mean of hypothesis makes error on known data;
Eout = the expectation of hypothesis making error on unknown data.

It is more difficult for hypothesis to predict unknown data.

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 = 0.0
Which is much lower than 2.10, 2.11

In this case, Eout has possitive correlation to Ein
Ein = 0, 0 < Eout < 0.5
Ein = 1.0, 0.5 < Eout <1.0

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 $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$, each being in $\mathbb{R}^d$ with $0 \leq k < d$. We shall call those perceptrons $\mathcal{H}$. What is the growth function $m_{\mathcal{H}}(N)$? Prove your answer.

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

① The requirement to pass k anchor points imposes k constraints on $\mathcal{H}'$, which reduces the number of "free parameters", thereby reducing its capacity to shatter sets.

→ Each k anchor points makes $d_{VC} = d \rightarrow d_{VC} = d-k$

② $m_{\mathcal{H}}(N)$ = the number of ways to choose subsets of those N points by using $\mathcal{H}'$ for each subset of size $\hat{n}$ $(0 \leq \hat{n} \leq d-k)$ if we want to choose subset from N-1 points → $\binom{N-1}{\hat{n}}$ ways.

→ $m_{\mathcal{H}}(N) = 2\sum_{\hat{n}=0}^{d-k} \binom{N-1}{\hat{n}}$