

## hw2-Solution

● Graded

Student

Chiang Yi Jie

Total Points

233 / 260 pts

Question 1

Problem 1

■ 18 / 20 pts

✓ + 20 pts Correct.

+ 15 pts Generally correct, but has some minor logical errors, or lacks clear description.

+ 10 pts On the right path, but has some major logical problems.

- 2 pts Wrong selected page.

✓ - 2 pts Minor notation mistakes.

+ 0 pts Totally wrong or blank.



It should be  $\mu > \nu + \epsilon$

Question 2

Problem 2

10 / 20 pts

+ 20 pts Correct.

+ 15 pts Generally correct, but has some minor logical errors, or lacks clear description.

✓ + 10 pts Incomplete proof. (e.g. only consider given any fixed  $m$  and  $t$ ) / Some large problems with the solution.

+ 5 pts On a reasonable path, but has some major logical problems. / Very unclear proof.

- 2 pts Wrong selected page.

- 2 pts Minor notation mistake.

+ 0 pts Totally wrong or blank.

Question 3

Problem 3

20 / 20 pts

✓ + 20 pts Answer is correct.

+ 15 pts Answer has minor errors.

+ 10 pts Answer is directionally correct.

+ 5 pts Misunderstood 'ticket' but the result is reasonable.

+ 0 pts Answer is incorrect.

+ 0 pts Answer is incomprehensible.

- 2 pts Select wrong page.

Question 4

Problem 4

20 / 20 pts

✓ + 20 pts Answer is correct.

+ 10 pts Answer has some errors.

+ 0 pts Answer is incorrect.

- 2 pts Wrong page selection

Question 5

Problem 5

20 / 20 pts

✓ + 20 pts Correct.

- 2 pts Wrong selected page.

- 2 pts Minor notation mistakes. / Unclear in all "+1" situations.

- 5 pts A solution with minor mistakes.

- 5 pts Not clear enough.

- 10 pts A solution with major problems.

+ 0 pts Incorrect or blank.

Question 6

Problem 6

20 / 20 pts

✓ + 20 pts Answer is correct.

+ 15 pts Generally correct, but there's some minor logical errors, or lacks clear description.

+ 10 pts Answer is directionally correct but with some major logical problems or seriously lacks description.

+ 5 pts Answer is basically wrong but with some reasonable logical statements.

+ 0 pts Answer is incorrect.

– 2 pts Wrong selected pages.

Question 7

Problem 7

20 / 20 pts

✓ + 20 pts Correct answer.

+ 15 pts Generally correct, but with some minor mistakes or lacks clear explanation.

+ 10 pts Some logical problems or seriously lacks clear explanation.

+ 0 pts Wrong answer.

– 2 pts Wrong page selection.

Question 8

Problem 8

20 / 20 pts

✓ + 20 pts Answer correct.

+ 15 pts Generally correct, but with some minor mistakes or lacks clear explanation.

+ 10 pts Some logical problems or seriously lacks clear explanation.

+ 0 pts Wrong Answer.

– 2 pts Wrong page selection.

### Question 9

#### Problem 9

Resolved 20 / 20 pts

✓ + 4 pts Correct on  $s = 1$  and  $\theta > 0$ .

✓ + 4 pts Correct on  $s = -1$  and  $\theta > 0$ .

✓ + 4 pts Correct on  $s = 1$  and  $\theta \leq 0$ .

✓ + 4 pts Correct on  $s = -1$  and  $\theta \leq 0$ .

✓ + 4 pts Correct conclusion.

✓ - 2 pts Minor error in process or not clear.

- 2 pts Minor error in process or not clear.

+ 0 pts Wrong argument.

- 2 pts Minor error in process or not clear.

- 2 pts Minor error in process or not clear.

💬 + 2 pts Point adjustment

🔄 Regrade Request

Submitted on: Nov 06

由於沒有註解，想請問哪裡證明不夠完善

第一次批改沒看到討論 $\theta < 0$ 的情況，但在細看好有討論。批改錯誤，感謝您。

Reviewed on: Nov 06

### Question 10

#### Problem 10

20 / 20 pts

✓ + 10 pts Correct scatter plot.

+ 5 pts Partially correct plot.

✓ + 10 pts Correct median of difference (accept 0.03~0.04)

+ 0 pts Wrong median of difference.

+ 0 pts Wrong plot.

Question 11

Problem 11

20 / 20 pts

✓ + 5 pts Correct scatter plot.

+ 2.5 pts Partially correct plot.

✓ + 5 pts Correct median of difference. (accept 0.11~0.14)

✓ + 5 pts Reasonable findings in scatter plot.

+ 3 pts Unclear findings.

✓ + 5 pts Reasonable findings in median of difference.

+ 3 pts Unclear findings.

+ 0 pts Wrong answer.

Question 12

Problem 12

20 / 20 pts

✓ + 7 pts Correct scatter plot.

+ 4 pts Partially correct plot.

✓ + 7 pts Correct median of difference. (accept absolute value < 0.01)

✓ + 6 pts Reasonable findings.

- 2 pts Unclear reasoning or wrong reasoning.

- 2 pts Unclear reasoning or wrong reasoning.

+ 0 pts Wrong answer.

Question 13

Problem 13

5 / 20 pts

- 20 pts Wrong answer

- 0 pts Totally correct, no logical flaw.

- 5 pts A minor logical flaw exists.

- 10 pts A logical flaw exists.

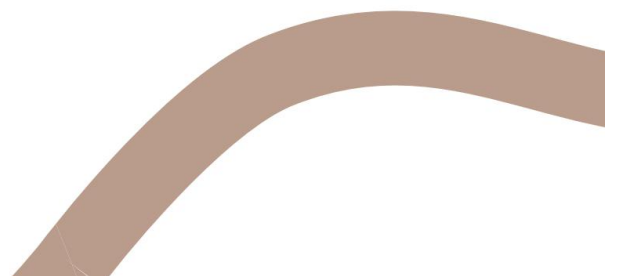
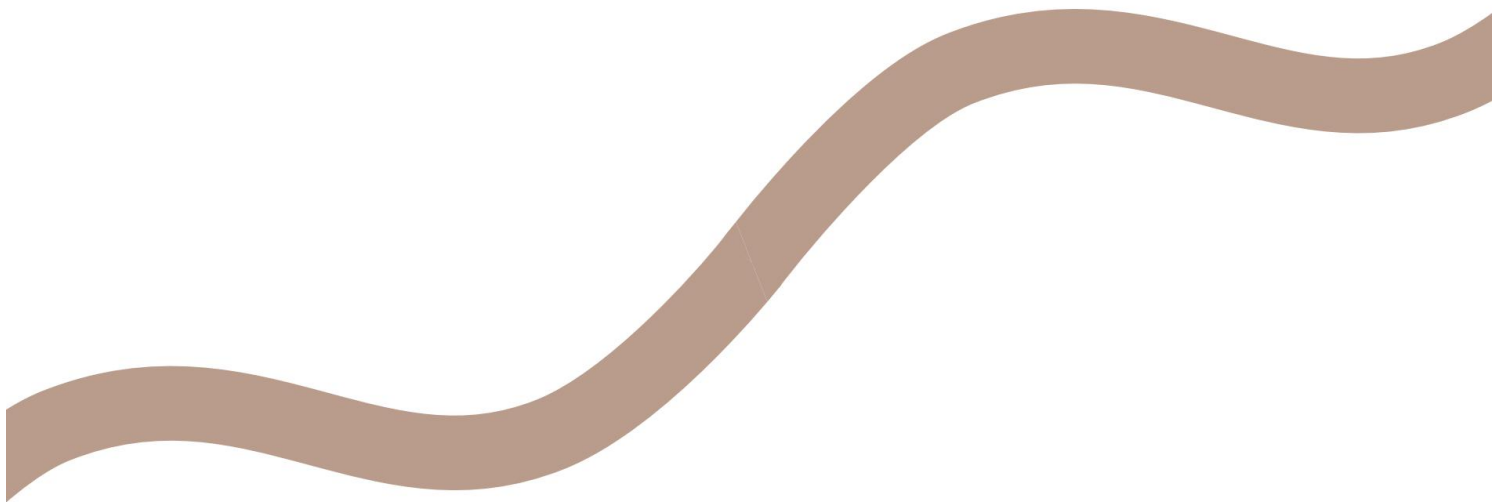
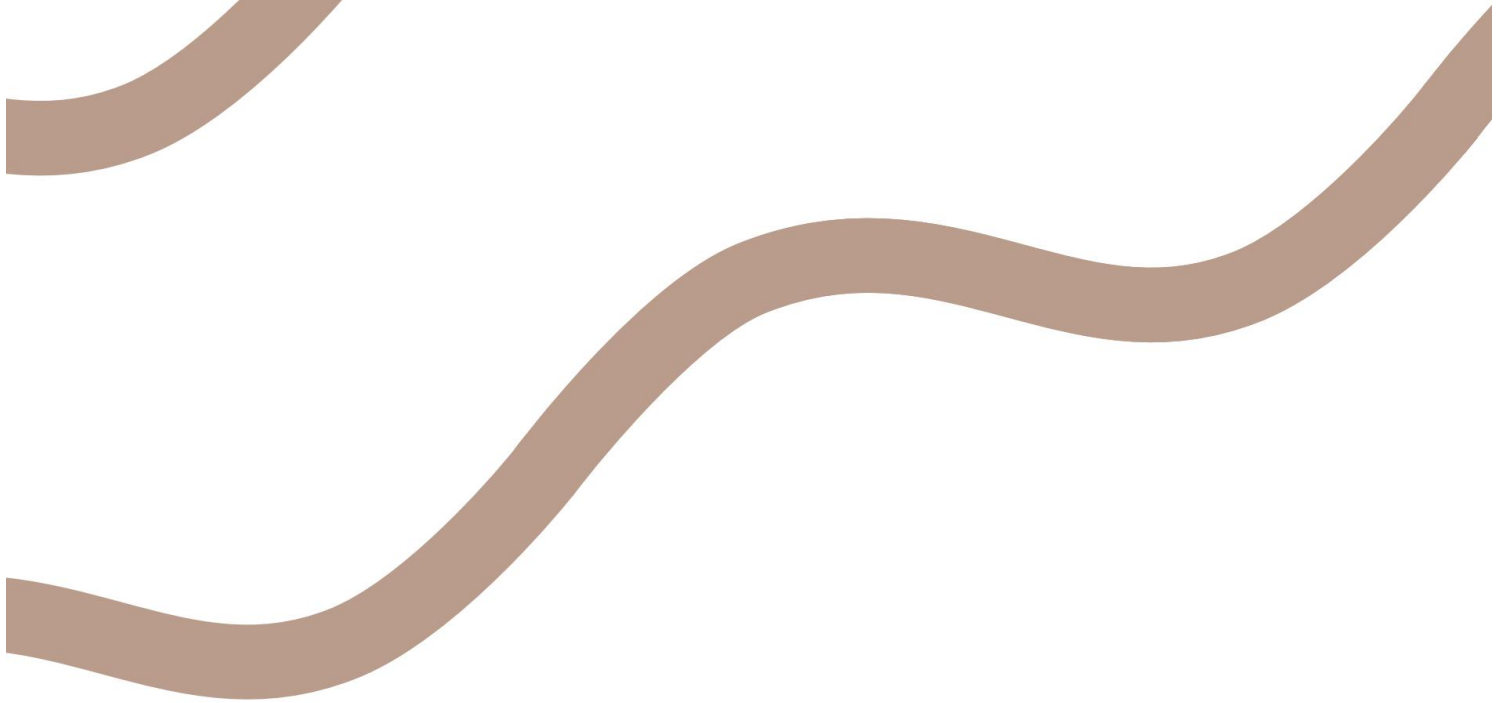
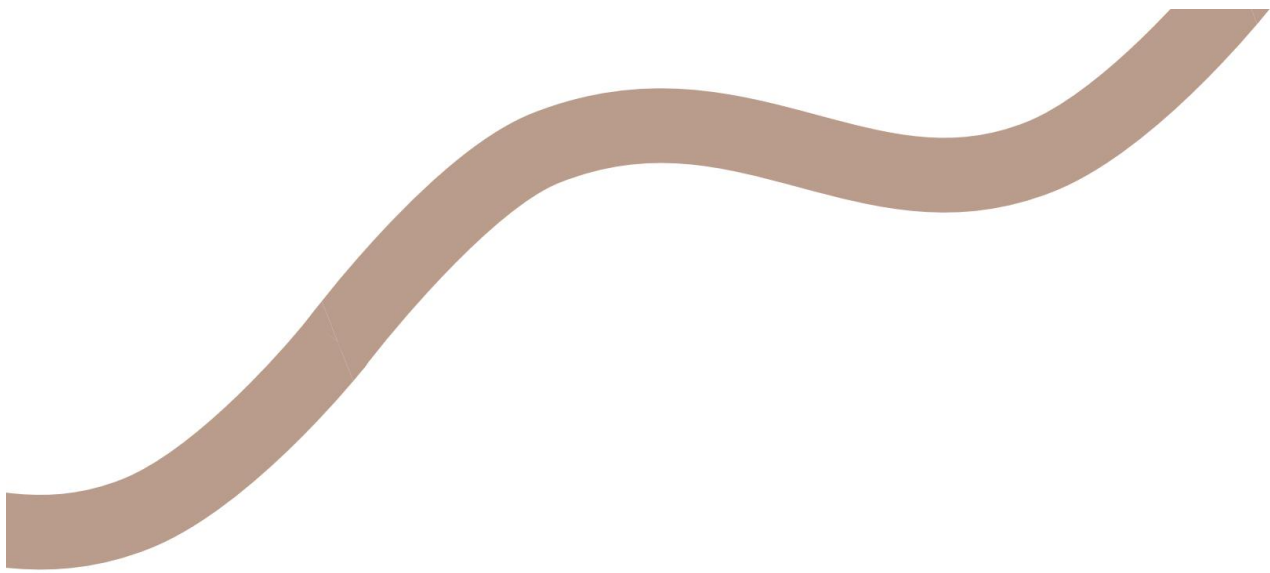
- 15 pts A major logical flaw exists.

✓ - 15 pts Without rigorous explanation.

- 2 pts Small typos exist.

1 In the bonus problem, we hope to see some level of rigor in the explanation for why the dimension is reduced by 'k' when we restrict the hyperplane to 'k' linearly independent anchor points (vectors).

No questions assigned to the following page.



Question assigned to the following page: [1](#)



每个 machine 都是一个 hypothesis

1. (20 points) Assume that we have  $M$  slot machines in front of us. Each machine has an unknown probability of  $\mu_m$  for returning one coin, and a probability of  $1 - \mu_m$  for returning no coin. For each of the time step  $t = 1, 2, \dots$ , 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)

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

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

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

$N_m$ : 第  $m$  个 slot machine 被拉了  $N_m$  次  
 $c_m$ : 第  $m$  个 slot machine 被拉了  $c_m$  次

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

$$\epsilon = \sqrt{\frac{\log t - \frac{1}{2} \log \delta}{N_m}}, \quad N = N_m, \quad \epsilon = \sqrt{\frac{\log t - \frac{1}{2} \log \delta}{N}}$$

$$\epsilon^2 = \frac{\log t - \frac{1}{2} \log \delta}{N}$$

$$N \epsilon^2 = \log t - \frac{1}{2} \log \delta$$

$$-2N \epsilon^2 = -2 \log t + \log \delta$$

$$e^{-2N \epsilon^2} = t^{-2} \delta$$

$$\rightarrow P(\mu_m > \nu + \epsilon) \leq e^{-2N \epsilon^2}$$

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



Question assigned to the following page: [2](#)

2. (20 points) Continuing from Problem 1, prove that when  $M \geq 2$ , for all slot machines  $m = 1, 2, \dots, M$  and for all  $t = M+1, M+2, \dots$ , 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  $\mu_m$ , 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.

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

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

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

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

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

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

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

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

$$> 1 - \delta$$



Questions assigned to the following page: [3](#) and [4](#)

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  $1, 2, \dots, 8$  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?

$$\begin{aligned}
 \text{if } (1, 2, 3, 4) \text{ green} &\rightarrow \text{all C} & C_5^5 &= 1 \\
 (5, 6, 7, 8) \text{ green} &\rightarrow \text{all D} & C_5^5 &= 1 \\
 (1, 3, 5, 7) \text{ green} &\rightarrow \text{all A} & C_5^5 &= 1 \\
 (2, 4, 6, 8) \text{ green} &\rightarrow \text{all B} & C_5^5 &= 1 \\
 (5, 7) \text{ green} &\rightarrow \text{A C C C C, A A C C C, A A A C C, A A A A C} & C_1^5 + C_2^5 + C_3^5 + C_4^5 &= 5 + 5 + \frac{5 \times 4}{2} \times 2 = 30 \\
 (1, 3) \text{ green} &\rightarrow \text{A D arrange} & & \text{the same above} = 30 \\
 (2, 4) \text{ green} &\rightarrow \text{B D arrange} & & = 30 \\
 (6, 8) \text{ green} &\rightarrow \text{B C arrange} & & = 30
 \end{aligned}$$

$$\text{The probability of "some number are all green" should be } \frac{1 \times 4 + 30 \times 4}{4^5} = \frac{84}{1024} = \frac{21}{256} \#$$

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.

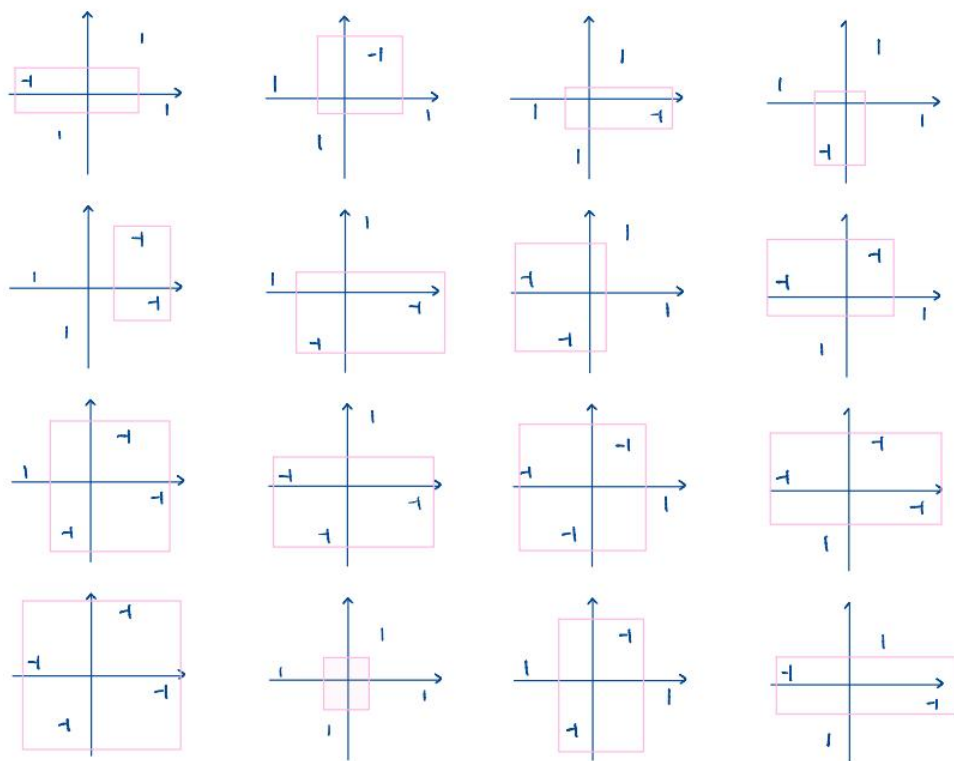
$$\begin{aligned}
 & \begin{matrix} (1, 2, 3, 4) & (2, 4, 6, 8) & (2, 4) \\ \uparrow & \uparrow & \\ 1 & 1 & 30 \end{matrix} \\
 P[ \text{five green 2's} ] &= \frac{1 + 1 + 30}{4^5} = \frac{32}{4^5} = \frac{1}{32} \#
 \end{aligned}$$





Question assigned to the following page: [5](#)

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.



Result:  $m_H(4) = 2^4$  for 4 inputs,  
 that is to say: 4 inputs can be shattered  
 dvc isn't less than 4.



Question assigned to the following page: [6](#)

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  $s, a_1, b_1, a_2, b_2, \dots, a_M, b_M$  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

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

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  $N=3$ ,  $M=1$ ,  $m_H(3)=8$ , shattered.

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \end{array} \rightarrow \{x_1, x_2, x_3\} = \{s, s, s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \end{array} \rightarrow \{x_1, x_2, x_3\} = \{s, s, -s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \end{array} \rightarrow \{x_1, x_2, x_3\} = \{s, -s, -s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \end{array} \rightarrow \{x_1, x_2, x_3\} = \{-s, s, -s\}, \quad s \in \{+1, -1\}$$

For  $N=4$ ,  $M=1$ ,  $m_H(4)=14$ , can't shattered.

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, s, s, s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, -s, s, s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, s, s, -s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, -s, -s, s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, s, -s, -s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, s, s, s\}, \quad s \in \{+1, -1\}$$

$$\begin{array}{c} a_1 \quad b_1 \\ | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{s, -s, s, s\}, \quad s \in \{+1, -1\}$$

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

That is, if  $h$  can shatter some  $N$  points  $\rightarrow h$  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  $N=5$ ,  $M=2$ ,  $m_H(5)=5$ , shattered

$$\begin{array}{c} a_1 \quad b_1 \quad a_2 \quad b_2 \\ | \quad | \quad | \quad | \quad | \\ x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \end{array} \rightarrow \{x_1, x_2, x_3, x_4\} = \{-s, s, -s, s, -s\}$$

using 4 parameters,  $\mathcal{H}$  could shatter 5 inputs.

The result can also refer to the conclusion in class:  
 parameters creates degree of freedom  $\rightarrow d_{vc} \approx \text{free parameters}$ .

In conclusion, hypothesis with  $(2M+1)$  parameters  $\begin{array}{c} \downarrow \quad \downarrow \quad \dots \quad \downarrow \quad \downarrow \\ a_1 \quad b_1 \quad \dots \quad a_M \quad b_M \end{array}$   $(2M+1)$  intervals.  
 $\rightarrow$  at most can shatter  $2M+1$  data  
 $\rightarrow d_{vc} = 2M+1$  #



Question assigned to the following page: [Z](#)

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

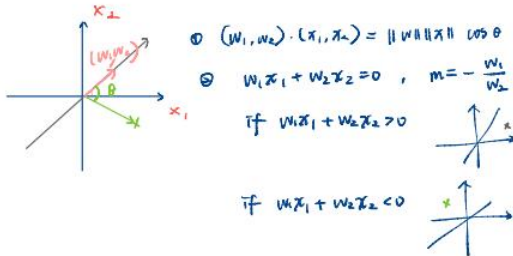
$$\mathcal{H}_0 = \{h: h(x) = \text{sign}(w_1 x_1 + w_2 x_2) \text{ i.e. perceptrons that pass the origin}\}$$

Prove your answer.

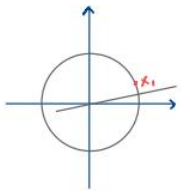
Hint: Consider putting your input vectors on the unit circle.

$w_1, w_2$  : weight

$x_1, x_2$  : variables (two characteristics)

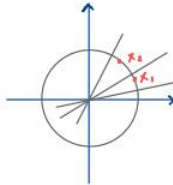


when  $N=1$ ,  $m_H(1)=2$



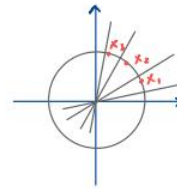
	$x_1$
$h_1$ :	0
$h_2$ :	x

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



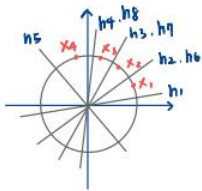
	$x_1$	$x_2$
$h_1$ :	0	0
$h_2$ :	0	x
$h_3$ :	x	x
$h_4$ :	x	0

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



	$x_1$	$x_2$	$x_3$
$h_1$ :	x	x	x
$h_2$ :	0	x	x
$h_3$ :	0	0	x
$h_4$ :	0	0	0
$h_5$ :	x	0	0
$h_6$ :	x	x	0

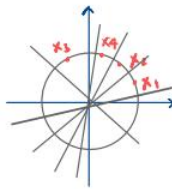
when  $N=4$ ,  $m_H(4)=8 = m_H(3)+2$



	$x_1$	$x_2$	$x_3$	$x_4$
$h_1$ :	x	x	x	x
$h_2$ :	0	x	x	x
$h_3$ :	0	0	x	x
$h_4$ :	0	0	0	x
$h_5$ :	0	0	0	0
$h_6$ :	x	0	0	0
$h_7$ :	x	x	0	0
$h_8$ :	x	x	x	0

when  $N=4$ ,

$x_4$  is added in the middle.



	$x_1$	$x_2$	$x_3$	$x_4$
$h_1$ :	x	x	x	x
$h_2$ :	0	x	x	x
$h_3$ :	0	0	x	x
$h_4$ :	0	0	0	x
$h_5$ :	0	0	0	0
$h_6$ :	x	0	0	0
$h_7$ :	x	x	0	0
$h_8$ :	x	x	x	0

如果  $x_N$  加在中间,

在  $H(N+1)$  里有  $m_H(N+1)$  种情形:

$x_N$  必须和左右两边的点同号,  
(也就是没有自由度);

$x_N$  可以有正负号的选择, 只有在左右不同号时,  
→ 多制造 2 种 hypothesis.

0,  $x_N, x$  or  $x, x_N, 0$

→  $m_H(N) = m_H(N-1) + 2$

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

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

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

which will create two more situation

$$\rightarrow m_H(3) + 2 = m_H(4)$$

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

Assume that  $m_H(N)=2N$

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

$$2(N-1) = 2N - 2, m_H(N) = 2N \text{ 成立.}$$

when  $N, m_H$

	$x_1$	$x_2$	...	$x_N$
	x	x	...	x
	0	x	...	x
	0	0	x	x
	0	0	0	x
	0	0	...	0
	x	0	...	x





Question assigned to the following page: [8](#)

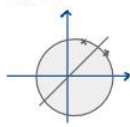
8. (20 points) For  $\mathcal{X} = \mathbb{R}^2$ , consider a hypothesis set  $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$  that is a union of two types of perceptrons:

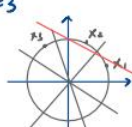
$$\mathcal{H}_0 = \{h: h(\mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2) \text{ i.e. perceptrons that pass the origin}\}$$

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

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

$d_{VC} \rightarrow$  find the min number of  $\mathbf{x}$  that could be shattered.

$N=2$   
  
 shattered  
 (Using proof of Q1)

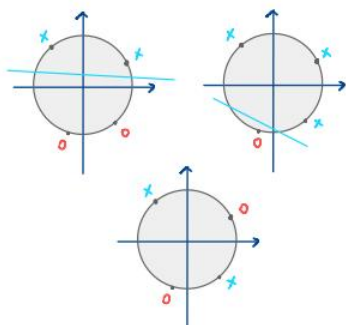
$N=3$   


$x_1$	$x_2$	$x_3$
0	0	0
x	0	0
x	x	0
x	x	x
0	x	x
0	0	x
x	0	x
0	x	0

shattered by  $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$

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)



$$\mathcal{H}_0'(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \{0000, x000, x00x, x0x0, 0xxx, 00xx, 000x, xxxx\}$$

$$\mathcal{H}_1'(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \{x00x, 0xx0, xx0x, 00x0\}$$

There aren't any  $\mathcal{H}$  such that  $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \{0xx0, x00x\}$

when  $N=4$ ,  $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$  couldn't be shattered by  $\mathcal{H}'$

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

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

$\rightarrow d_{VC} = 3$



Question assigned to the following page: [9](#)

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:

- Generate  $x$  by a uniform distribution in  $[-1, 1]$ .
- Generate  $y$  by  $y = \text{sign}(x) + \text{noise}$ , 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  $\theta \in [-1, 1]$ ,

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

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

$S = +1$

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

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

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

error since  $\hat{y} \neq y$  ( $P=0.1$ )

$S = -1$

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

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

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

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

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

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

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

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

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

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

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

in conclusion,  $E_{\text{out}}(h_{1,0}) = 0.5 + 0.4s + 0.4|\theta|$



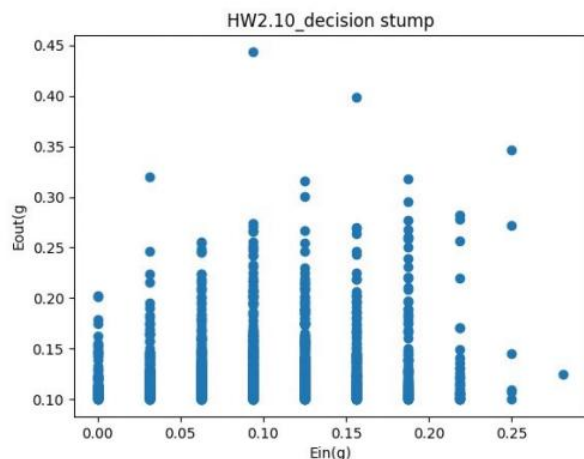
Question assigned to the following page: [10](#)



10. (20 points, \*) In fact, the decision stump model is one of the few models that we could minimize  $E_{\text{in}}$  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  $E_{\text{in}}$  values. We can then easily choose the hypothesis that leads to the lowest  $E_{\text{in}}$  by the following decision stump learning algorithm.

- (1) sort all  $N$  examples  $x_n$  to a sorted sequence  $x'_1, x'_2, \dots, x'_N$  such that  $x'_1 \leq x'_2 \leq x'_3 \leq \dots \leq x'_N$
  - (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  $E_{\text{in}}$  as  $g$ ; if multiple hypotheses reach the minimum  $E_{\text{in}}$ , 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 ~~direct partitioning~~ 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  $E_{\text{out}}(g) - E_{\text{in}}(g)$ .

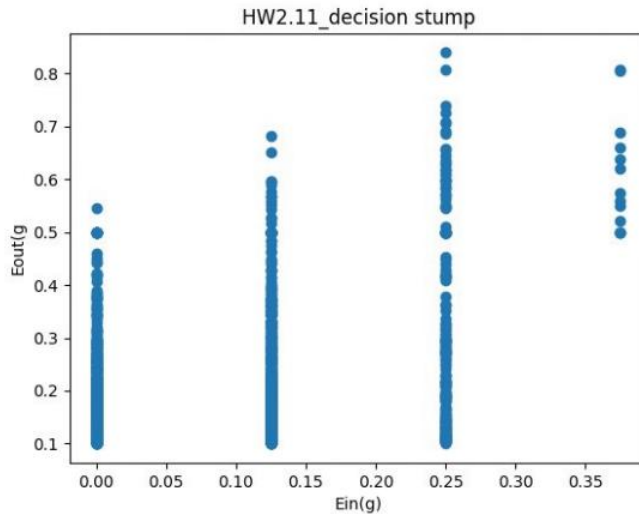


Median = 0.0383



Question assigned to the following page: [11](#)

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 Hoeffding's inequality, which indicates that when  $N$  become larger ( $N$  = the number of data),  $E_{in}$  will be closer to  $E_{out}$ .

Focusing on the right side of the plot, when  $E_{in} > 0.35$ ,  $E_{out} > 0.4$ ; to sum up, if  $E_{in}$  is large,  $E_{out}$  is highly possible to be large.

In most of cases,  $E_{in}$  is slightly smaller than  $E_{out}$ , which also explain the meaning of Both.

$E_{in}$  = the mean of hypothesis makes error on known data;

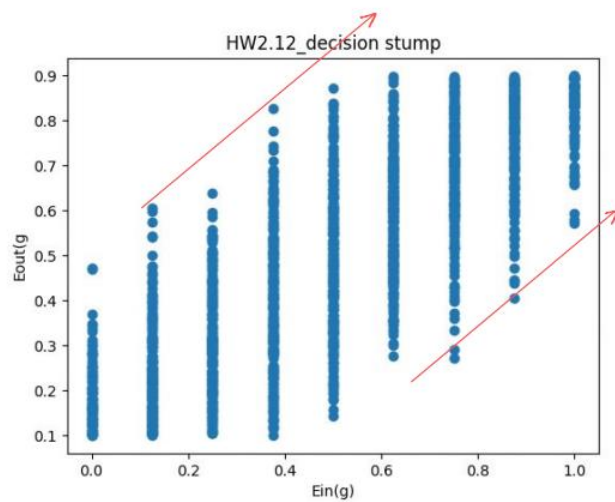
$E_{out}$  = the expectation of hypothesis making error on unknown data.

It is more difficult for hypothesis to predict unknown data.



Question assigned to the following page: [12](#)

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  $\theta$  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,  $E_{\text{out}}$  has positive correlation to  $E_{\text{in}}$

$E_{\text{in}} = 0, 0 < E_{\text{out}} < 0.5$

$E_{\text{in}} = 1.0, 0.5 < E_{\text{out}} < 1.0$





Question assigned to the following page: [13](#)

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/12course/www/vision\\_and\\_learning/perceptron\\_notes.pdf](https://web.mit.edu/course/other/12course/www/vision_and_learning/perceptron_notes.pdf)

for its proof.

Now, assume that we require the perceptrons to pass *all*  $k$  anchor points for  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ , 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  $\mathbf{a}_1 = \mathbf{0}$ .*

- 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.  
 $\rightarrow$  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  $\rightarrow \binom{N-1}{\hat{n}}$  ways.

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



No questions assigned to the following page.





No questions assigned to the following page.







No questions assigned to the following page.



