Why Deep Learning?

李宏毅
Hung-yi Lee

larger $|\mathcal{H}|$

smaller $|\mathcal{H}|$

$L(h^{train}, \mathcal{D}_{all})$

large

small $L(h^{all}, \mathcal{D}_{all})$

small $L(h^{train}, \mathcal{D}_{all})$

large $L(h^{all}, \mathcal{D}_{all})$
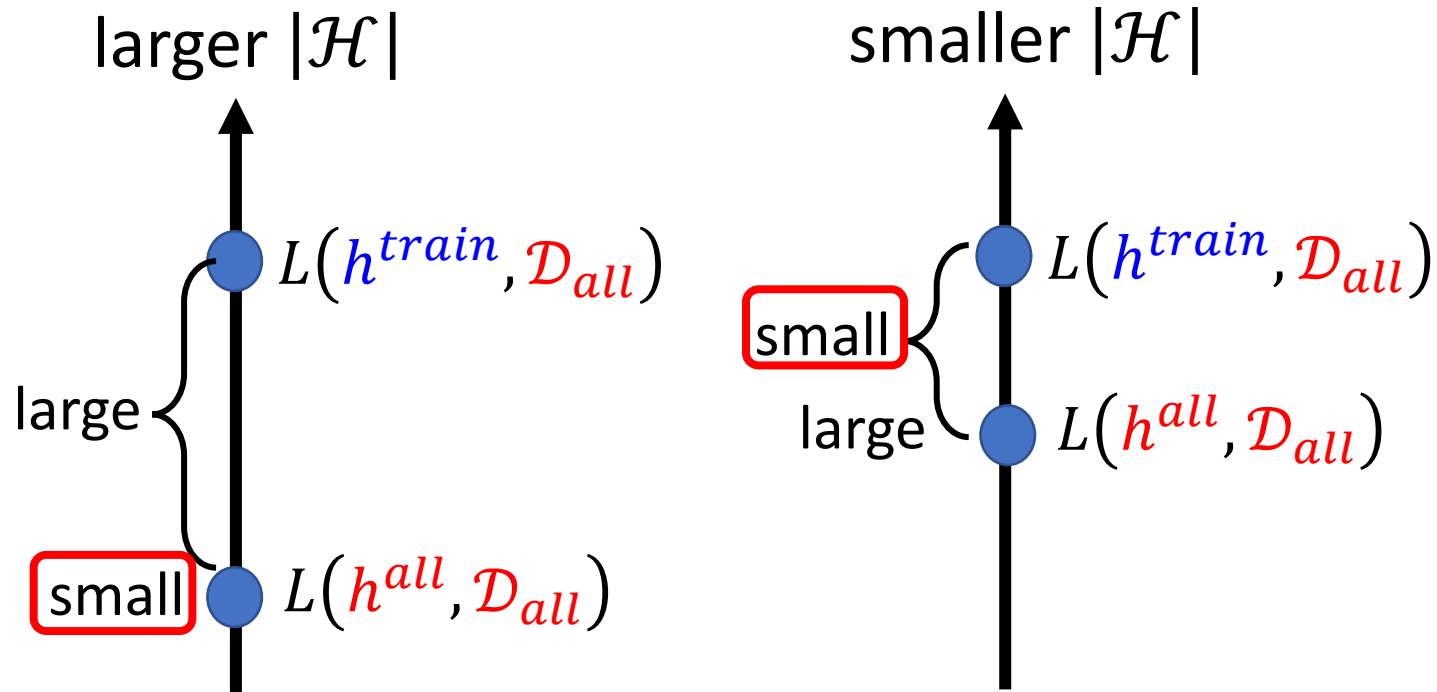
魚與熊掌可以兼得嗎？

$$h^{all} = arg \min_{h \in \mathcal{H}} L(h, \mathcal{D}_{all})$$
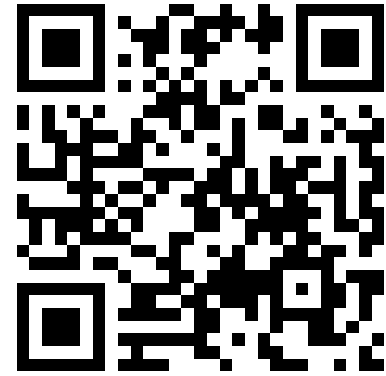
Still small loss

Small (fewer candidates)
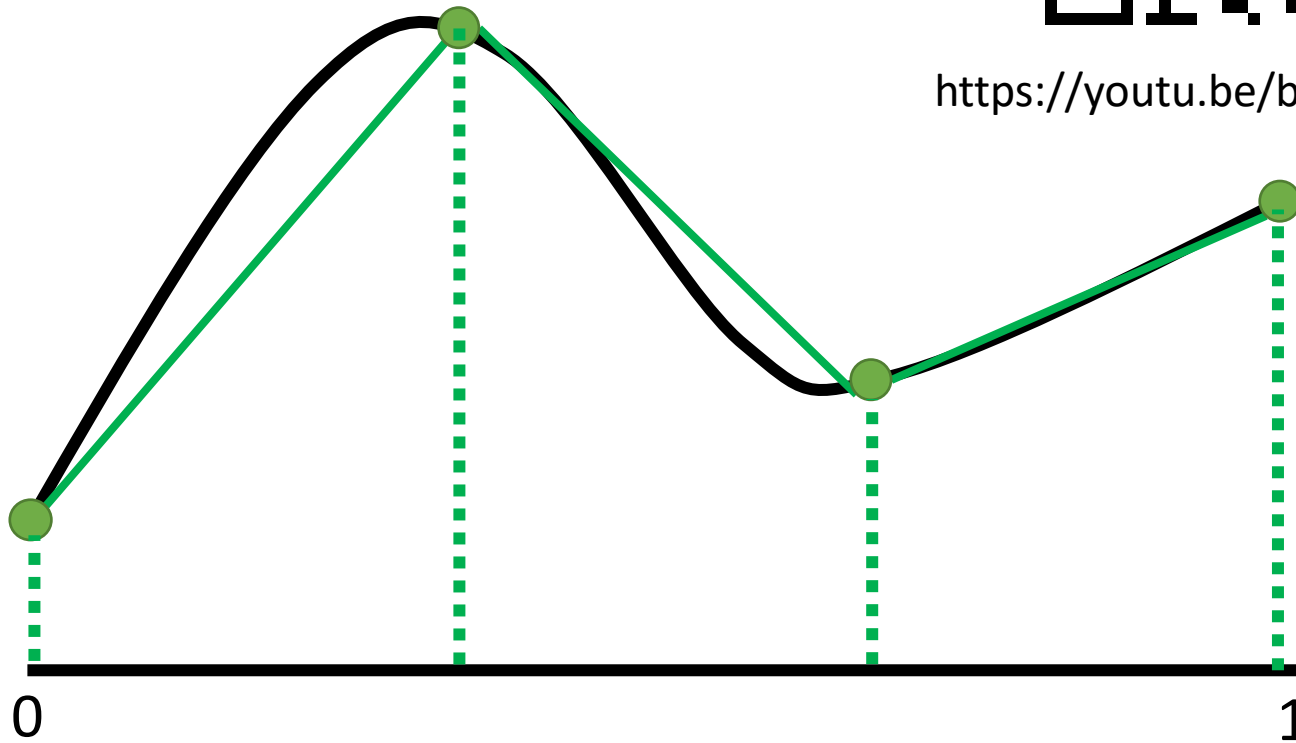
Review: Why Hidden Layer?
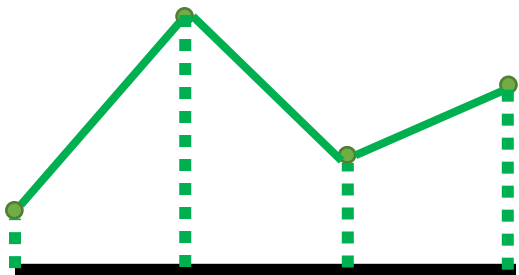
Piecewise Linear

0   1

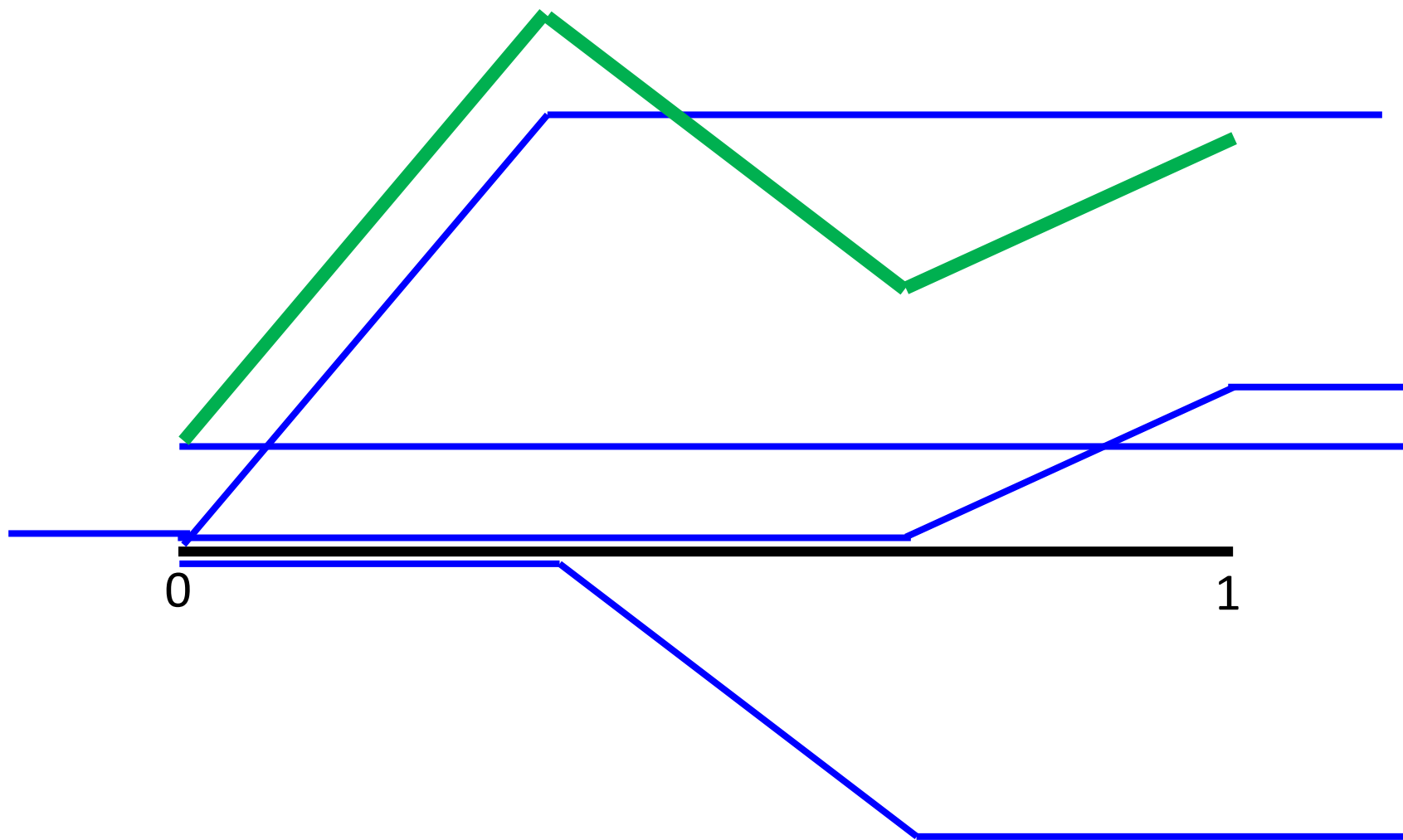We can have good approximation with sufficient pieces.
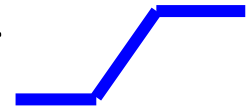
piecewise
linear = constant +
sum of a set of

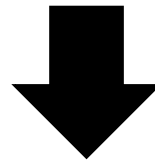0   1

Piecewise linear = constant + sum of a set of

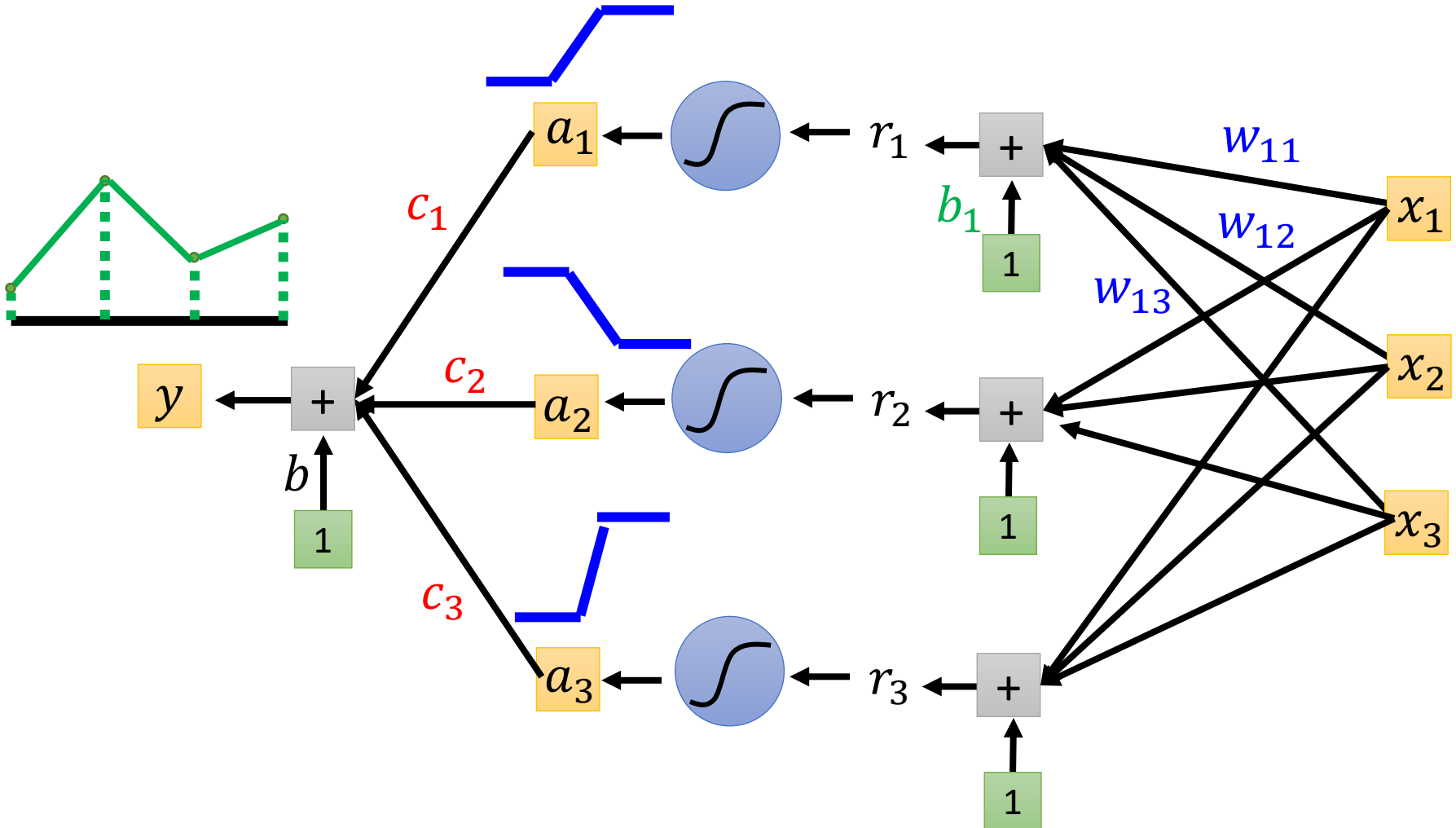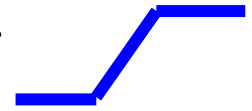How to represent this function?

Hard Sigmoid

Sigmoid Function

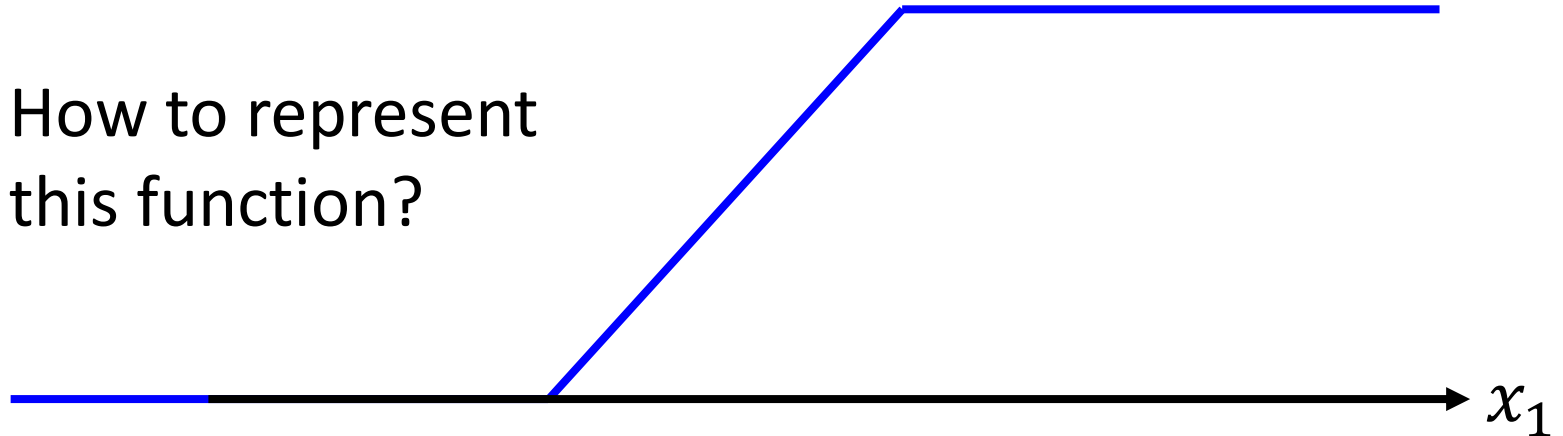$$y = c \frac{1}{1 + e^{-(b + wx_1)}}$$

$$= c \, sigmoid(b + wx_1)$$
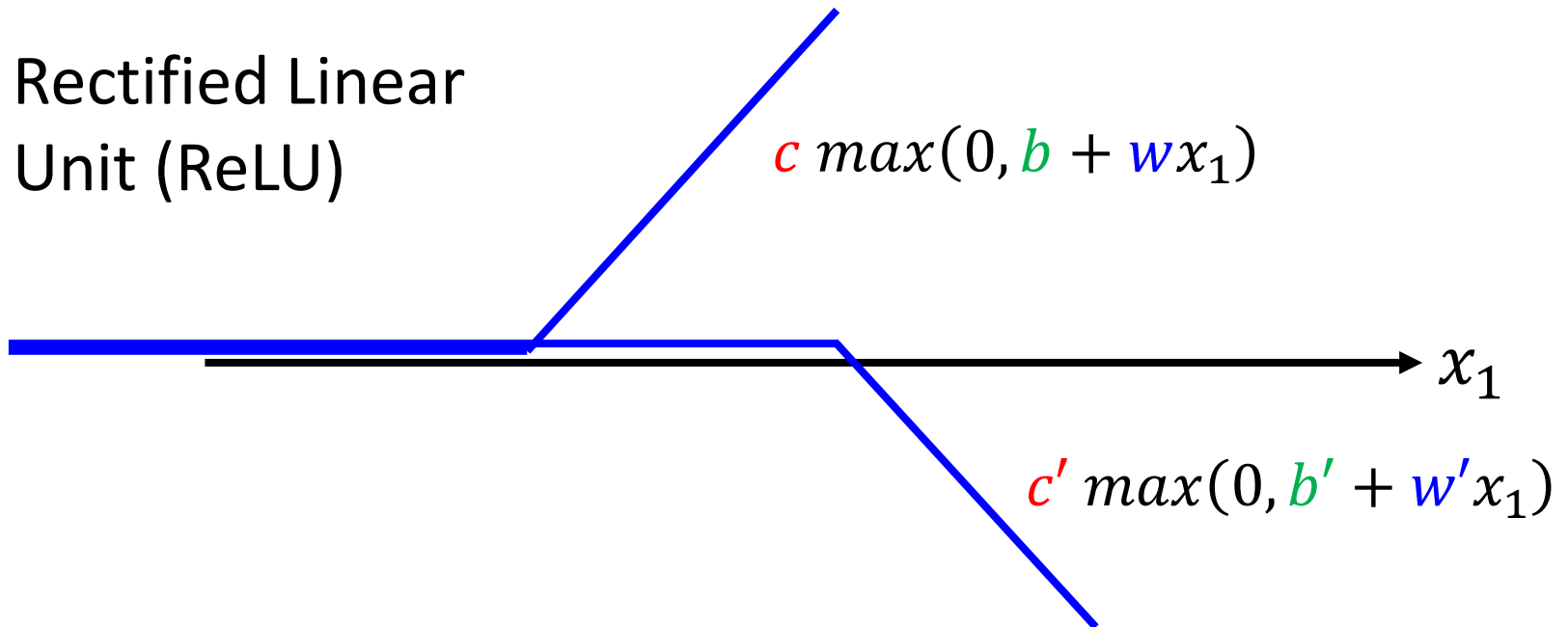
Piecewise linear = constant + sum of a set of

Hard Sigmoid → ReLU

How to represent
this function?

Rectified Linear
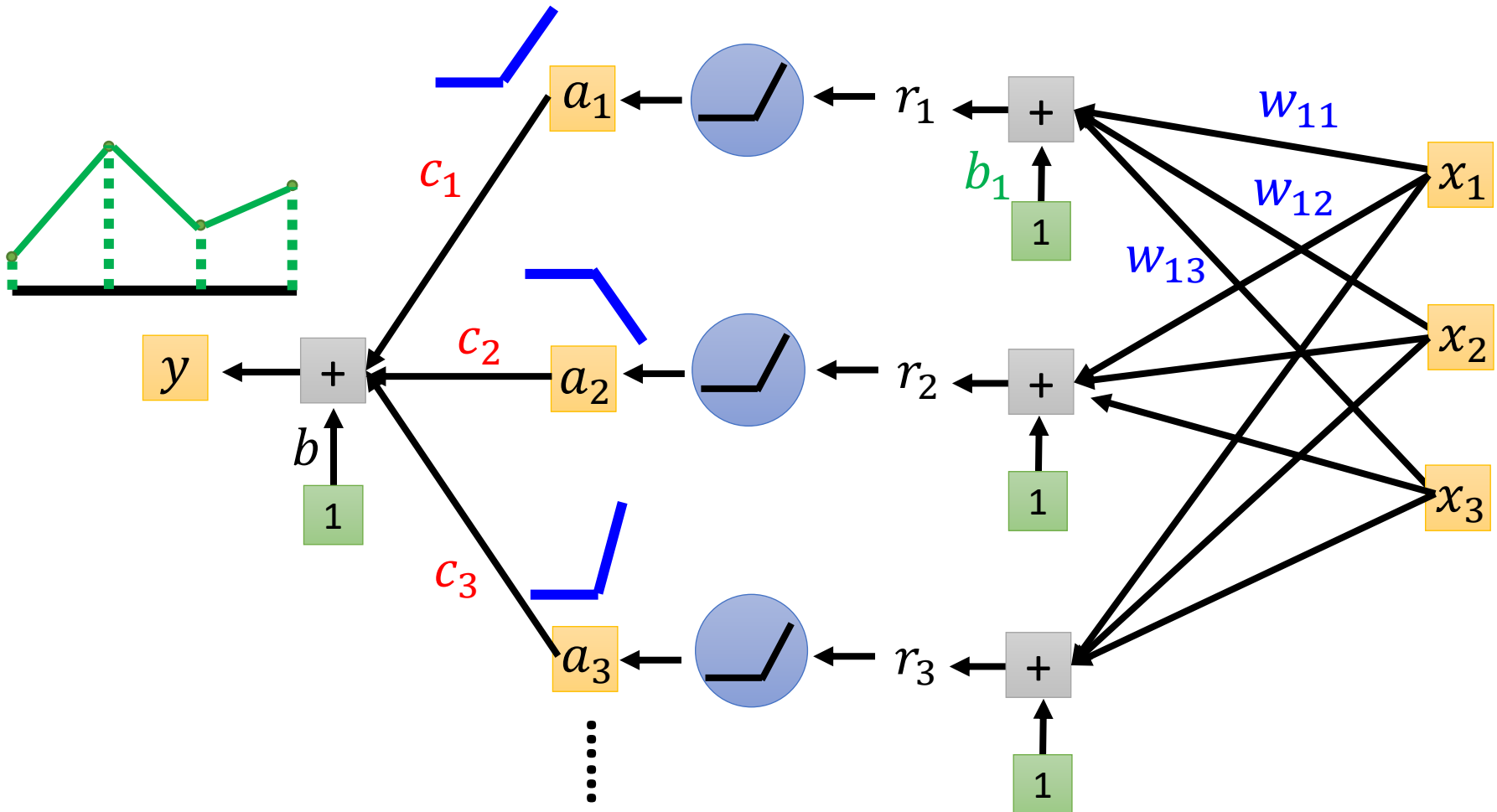Unit (ReLU)

$c\ max(0, b + wx_1)$

$c'\ max(0, b' + w'x_1)$

Piecewise linear $=$ constant $+$ sum of a set of

Why we want "***Deep***" network, not "***Fat***" network?

Deeper is Better?

Layer X Size	Word Error Rate (%)
1 X 2k	24.2
2 X 2k	20.4
3 X 2k	18.4
4 X 2k	17.8
5 X 2k	17.2
7 X 2k	17.1
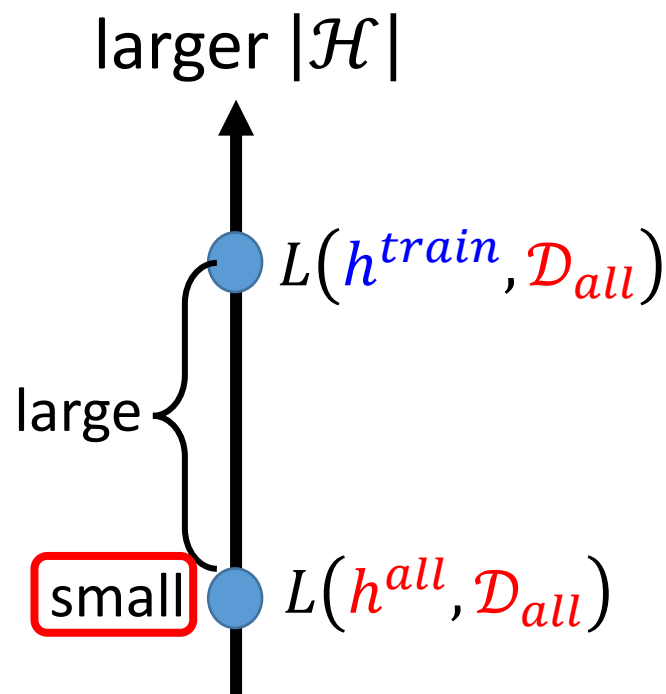
larger $|\mathcal{H}|$

$L\left(h^{train}, \mathcal{D}_{all}\right)$

large

small $\quad L\left(h^{all}, \mathcal{D}_{all}\right)$

Seide Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech.* 2011.

Fat + Short v.s. Thin + Tall

The same number of parameters

Which one is better?

Shallow

Deep
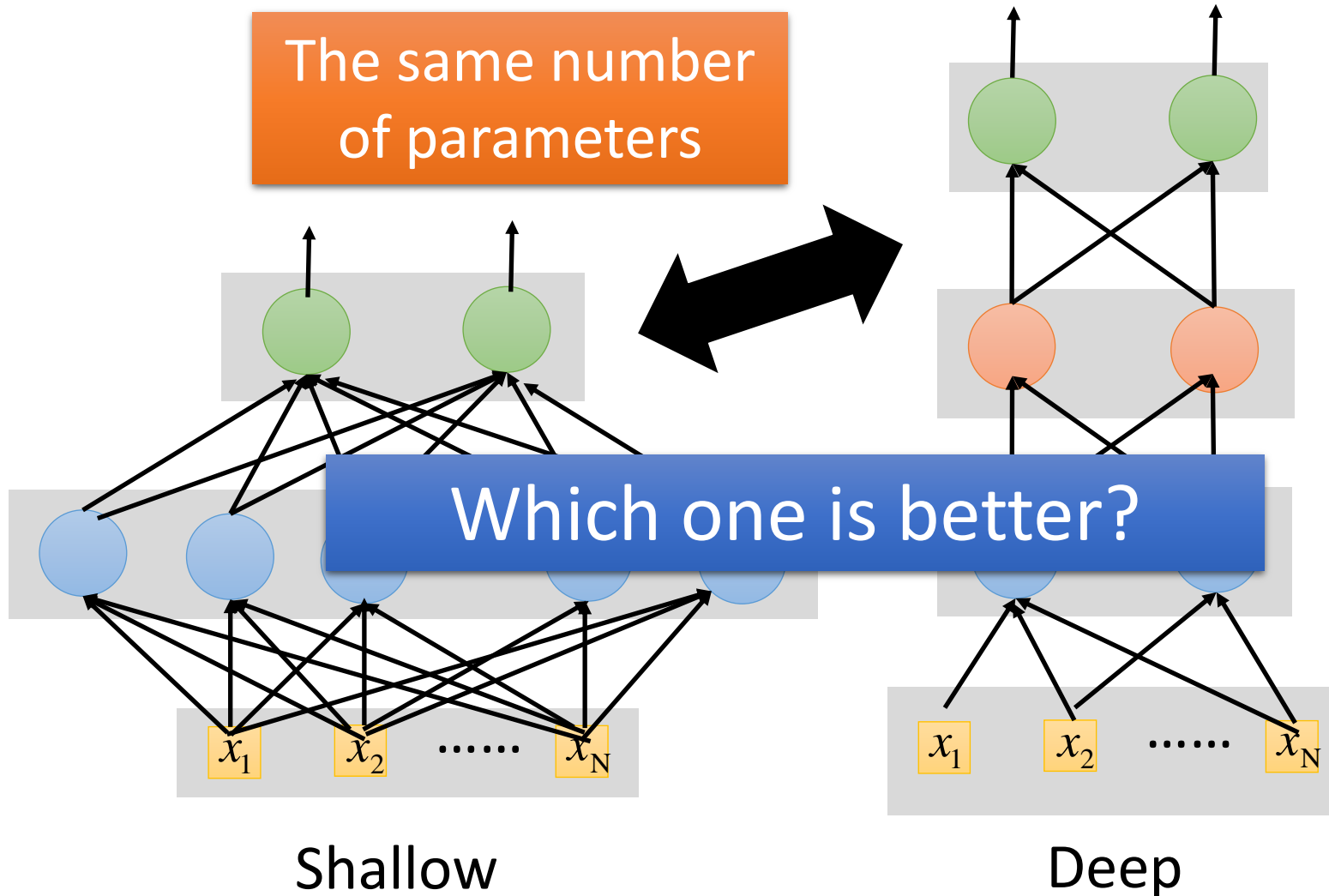
Fat + Short v.s. Thin + Tall

Layer X Size	Word Error Rate (%)	Layer X Size	Word Error Rate (%)
1 X 2k	24.2		
2 X 2k	20.4		Why?
3 X 2k	18.4		
4 X 2k	17.8		
5 X 2k	17.2 ⟷	1 X 3772	22.5
7 X 2k	17.1 ⟷	1 X 4634	22.6
		1 X 16k	22.1

Seide Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech*. 2011.
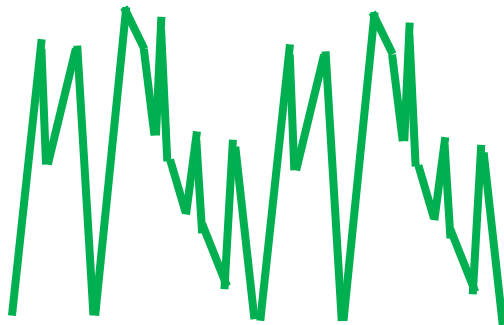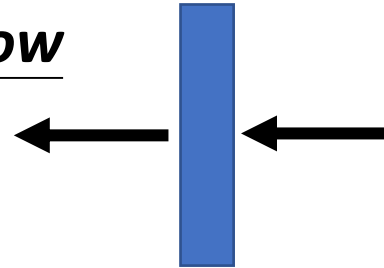
Why we need deep?
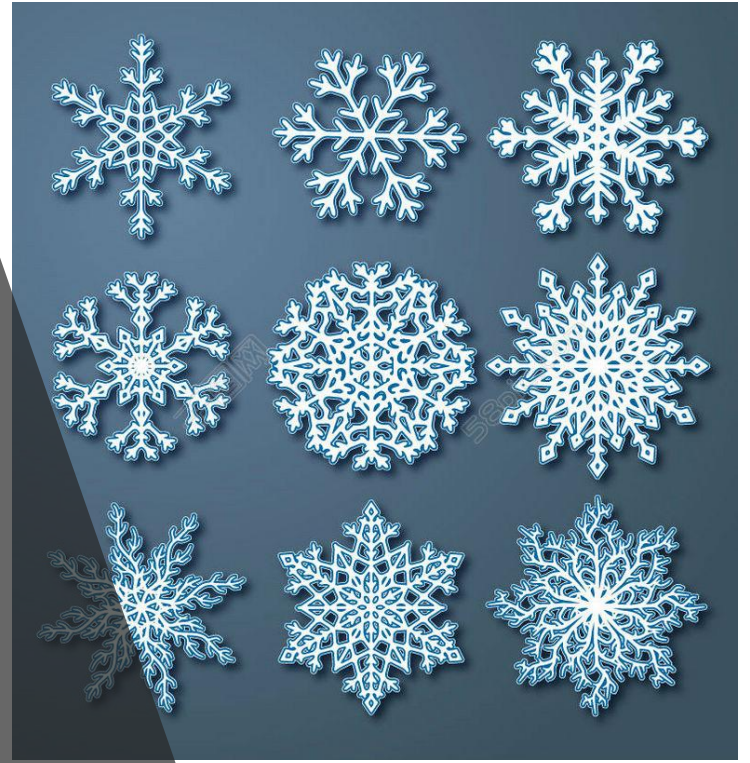
Why we need deep?

Yes, one hidden layer can represent any function.

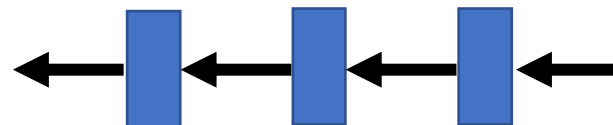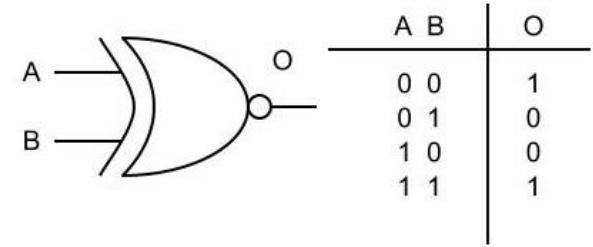However, using deep structure is more effective.
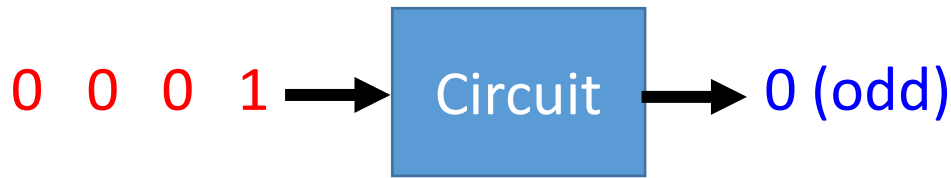
Shallow

Deep

More parameters

Analogy – Logic Circuits

- E.g., ***parity check***

1 0 1 0 → Circuit → 1 (even)

0 0 0 1 → Circuit → 0 (odd)

For input sequence with d bits,

Two-layer circuit need $O(2^d)$ gates.

XNOR

1 A
0 B
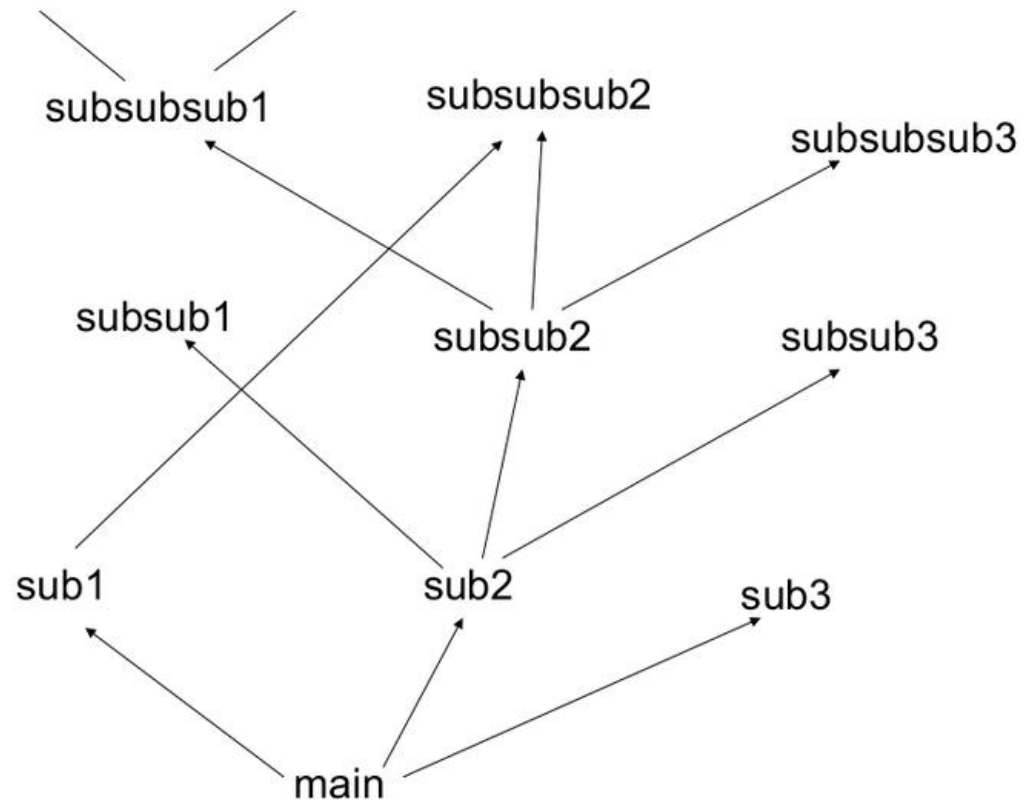1 C
0 D

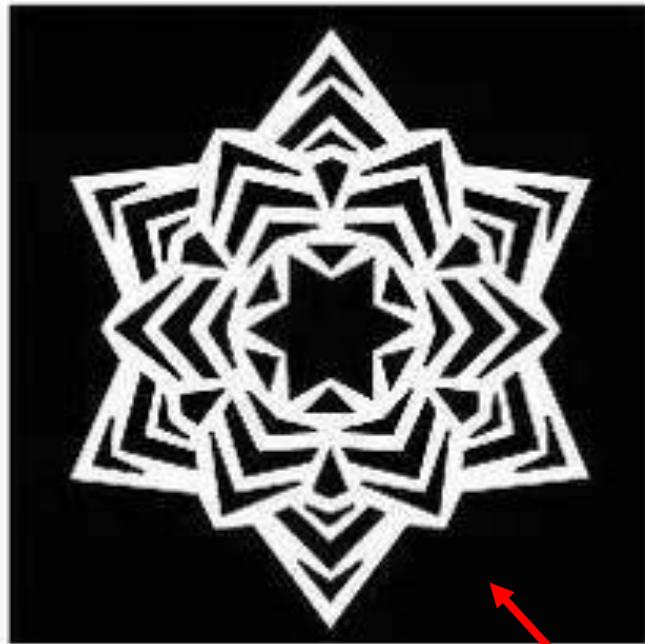0   0   1
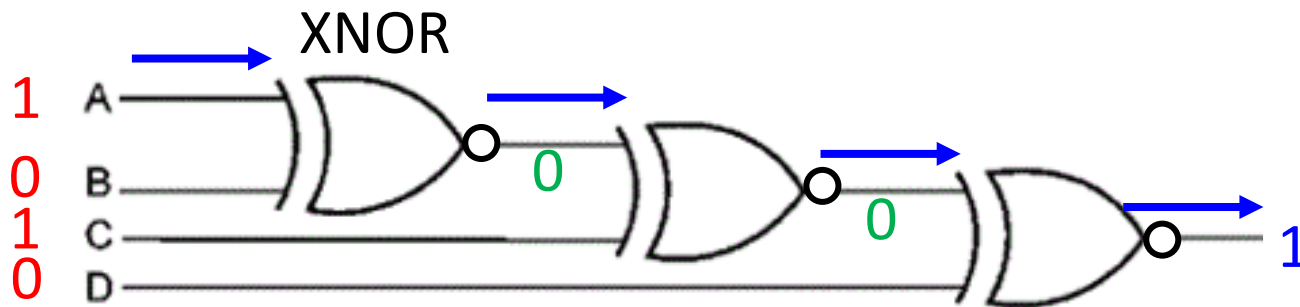
With multiple layers, we need only $O(d)$ gates.

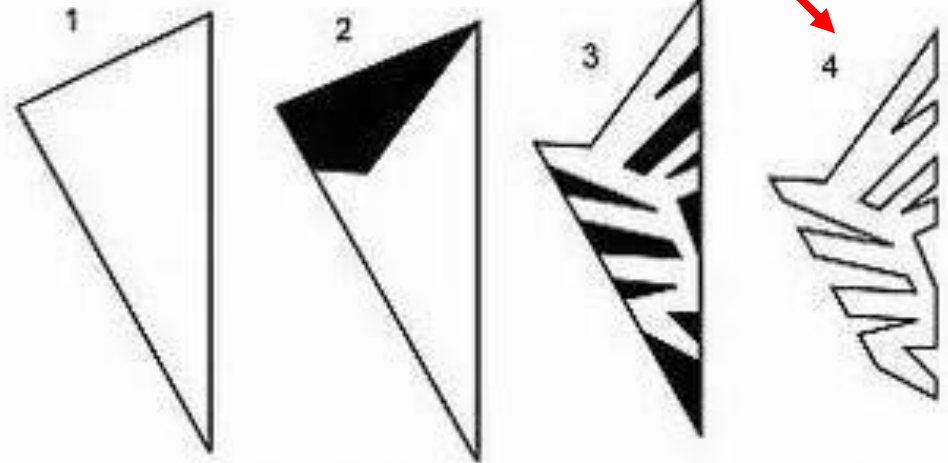# Analogy – Programming

Don't put everything in your main function.

More Analogy
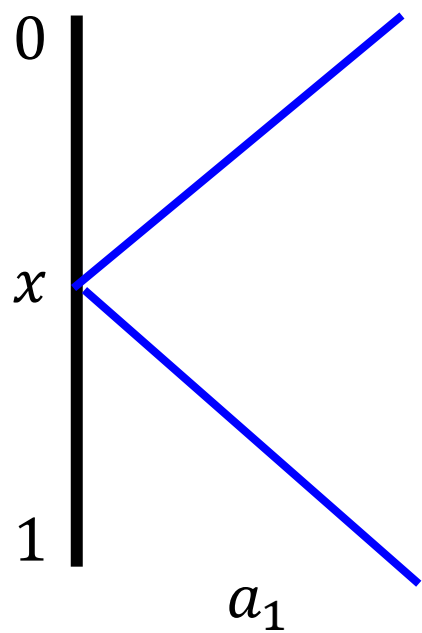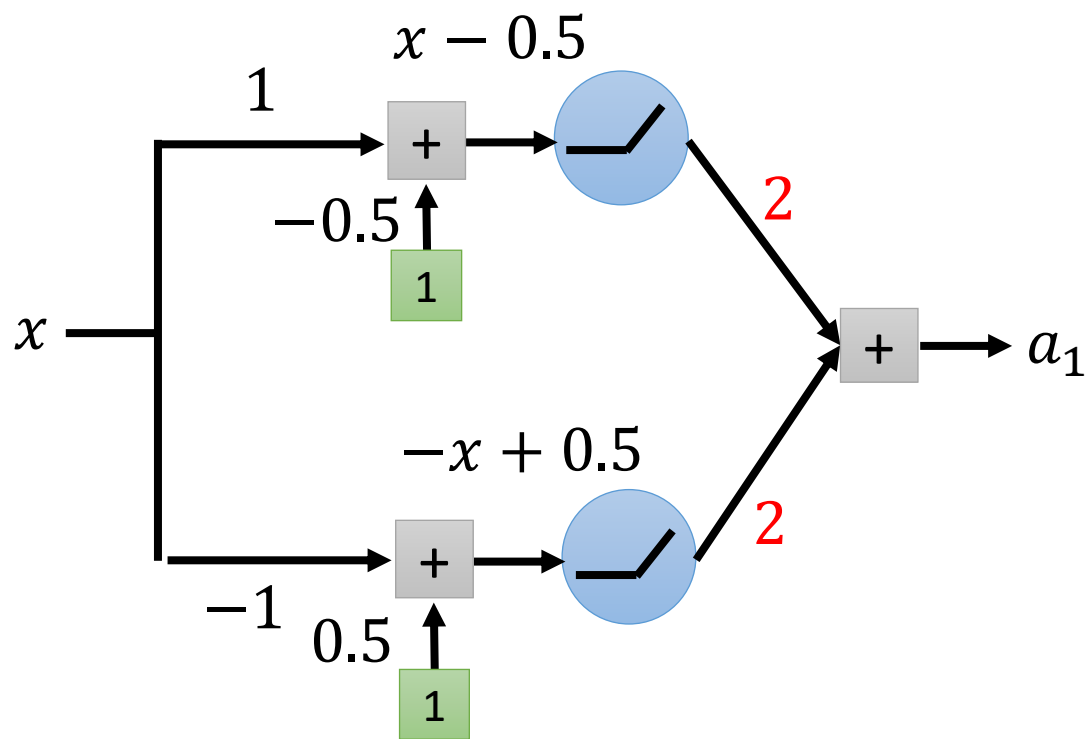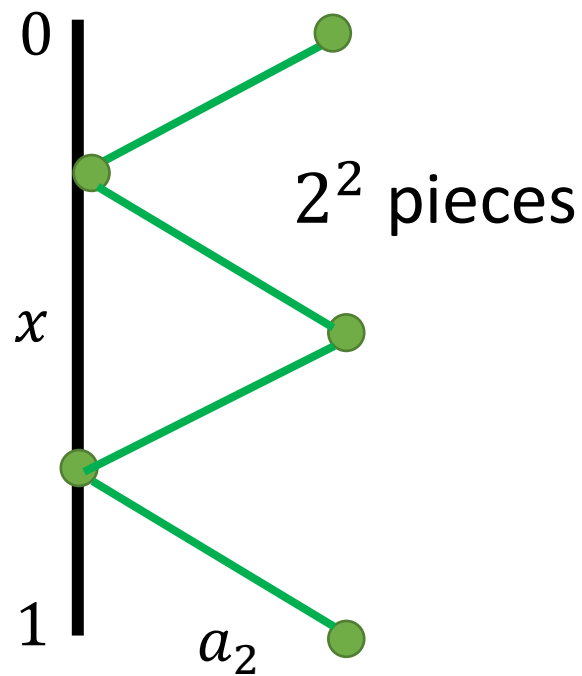

比較有效率

剪很多刀

- 以下三頁投影片上課時的數字有誤，已經將修改部的分套上紅色，感謝同學指出錯誤

$x - 0.5$

1

-0.5

1

x

2

$+$ → a_1

$-x + 0.5$

-1 0.5

1

2

0

x

1

a_1

$x - 0.5$

$a_1 - 0.5$

1

2

-0.5

1

-2

1

x

2

$+$

a_2

$-x + 0.5$

2

$-a_1 + 0.5$

-1

0.5

-2
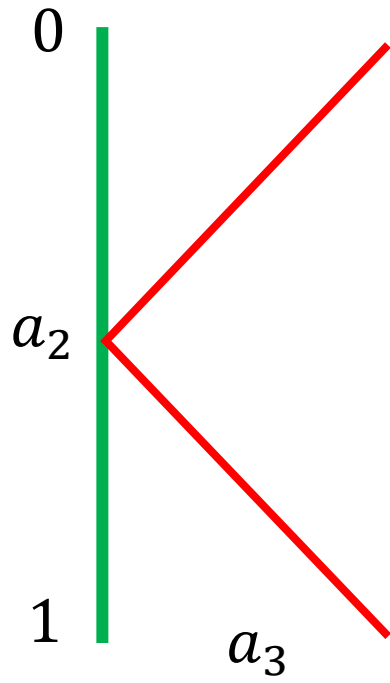
2

1

1

0

0

0

x

a_1

x

2^2 pieces

1

1

1

a_1

a_2

a_2

2³ pieces

Deep

2 neurons

x → → → → ..... → y

K layers ($2K$ neurons)

smaller $|\mathcal{H}|$

Shallow

2^K neurons

x → → y

larger $|\mathcal{H}|$

2^K pieces

0

x

1

y
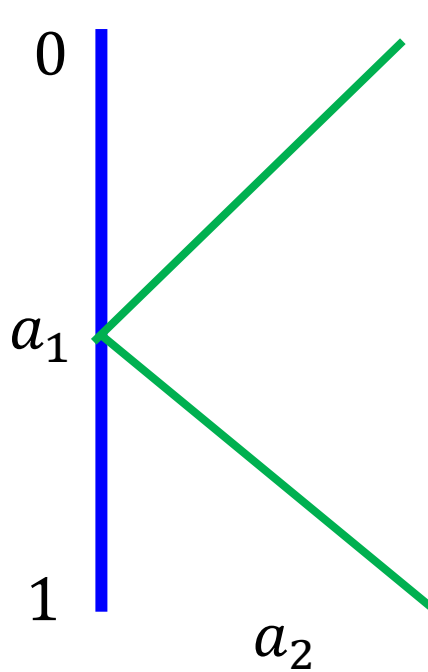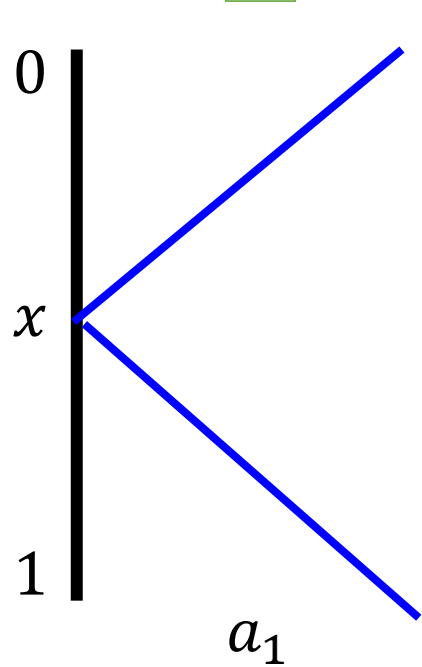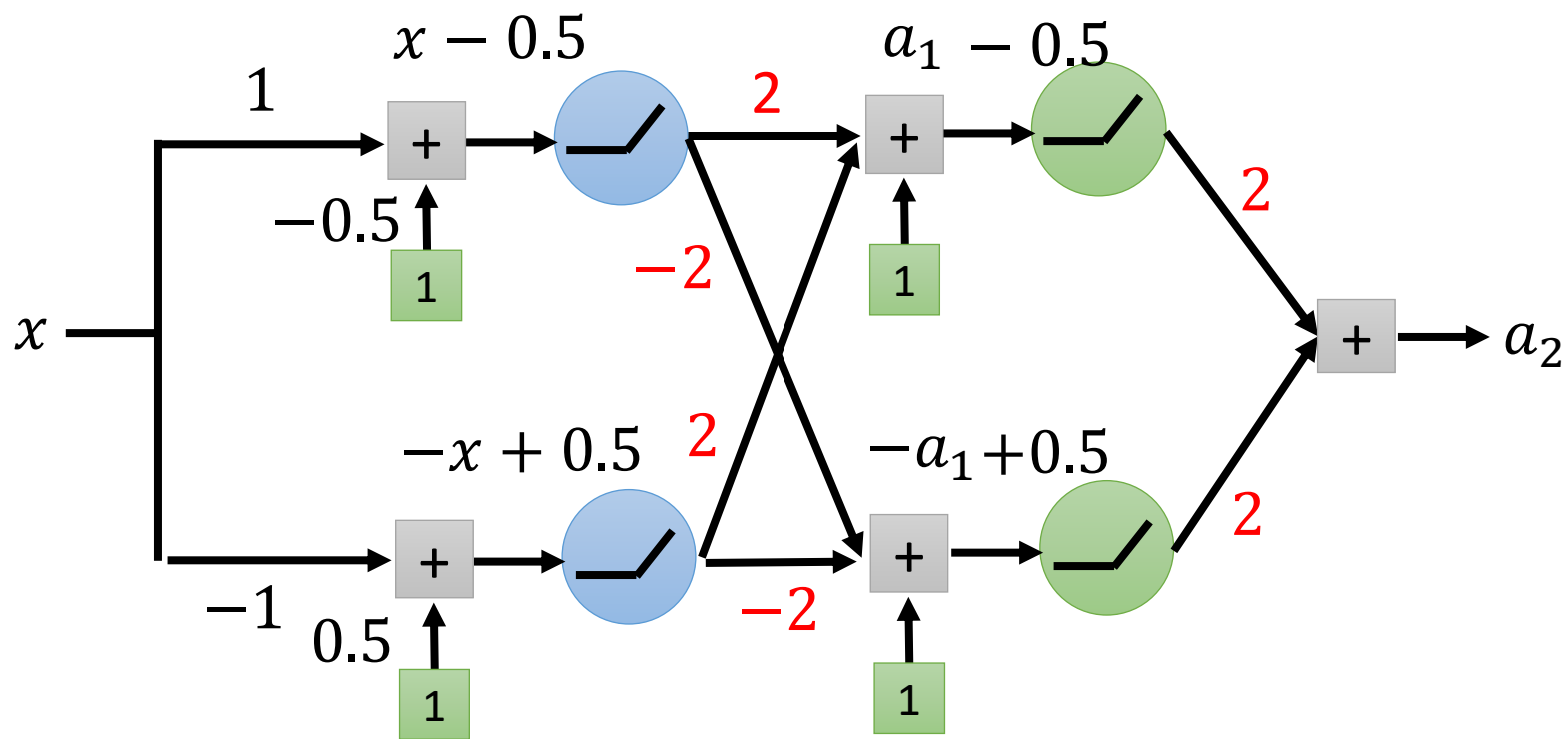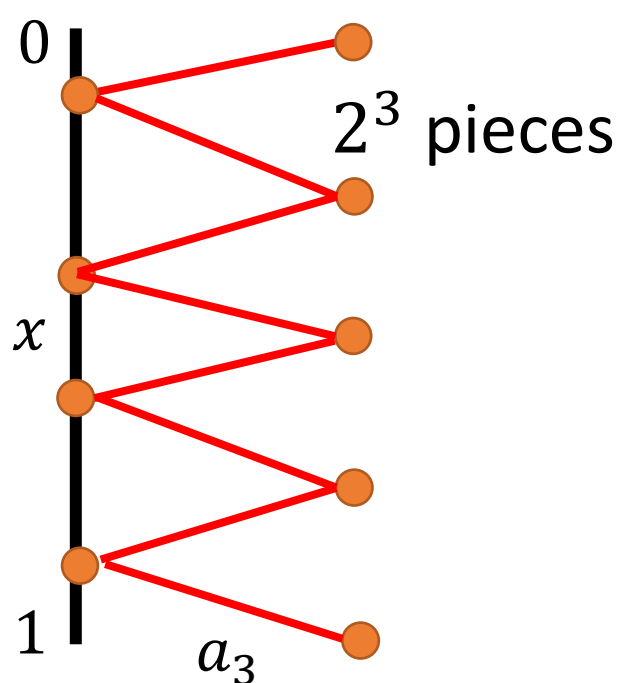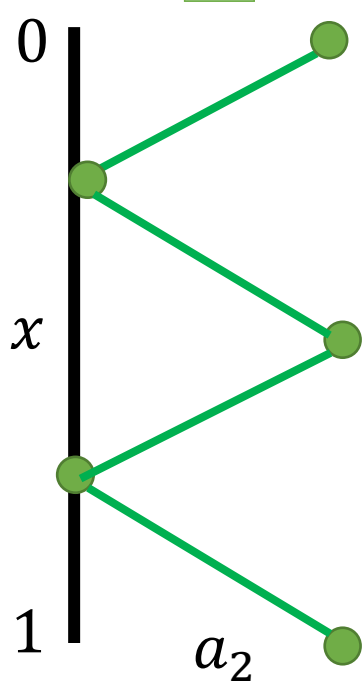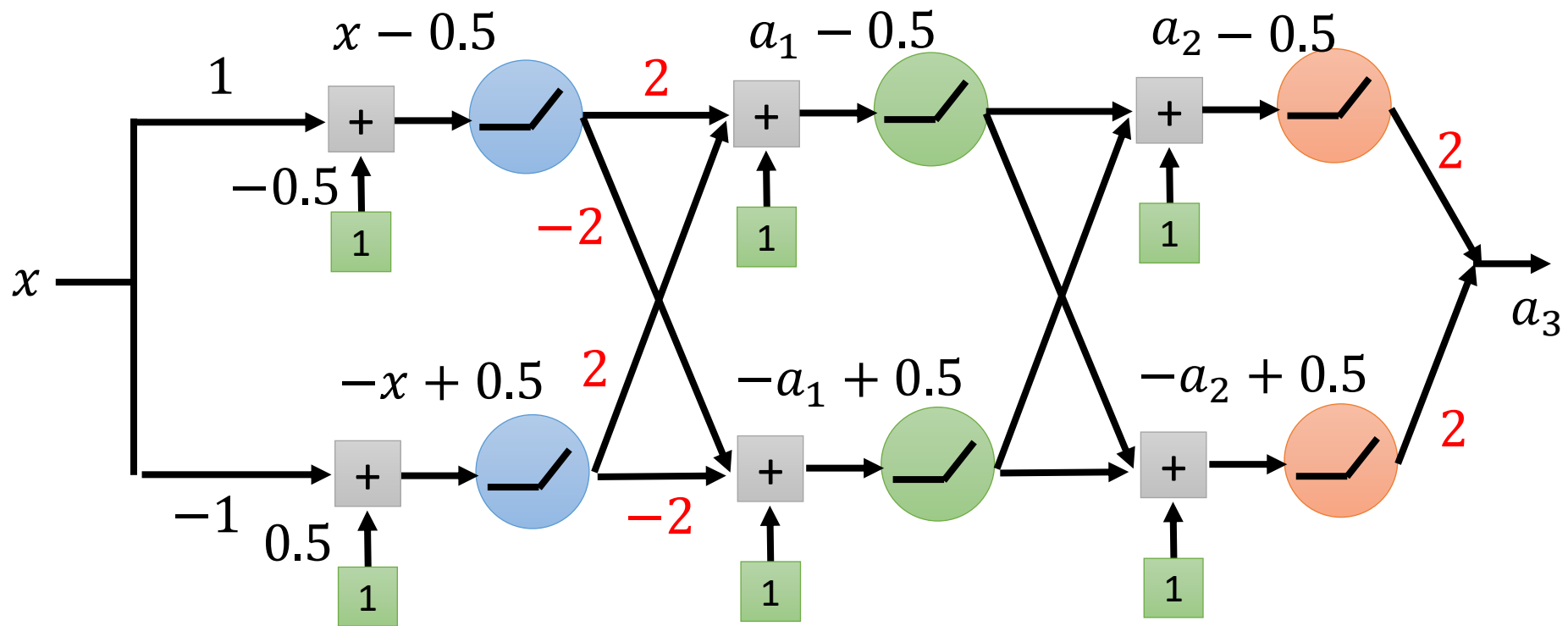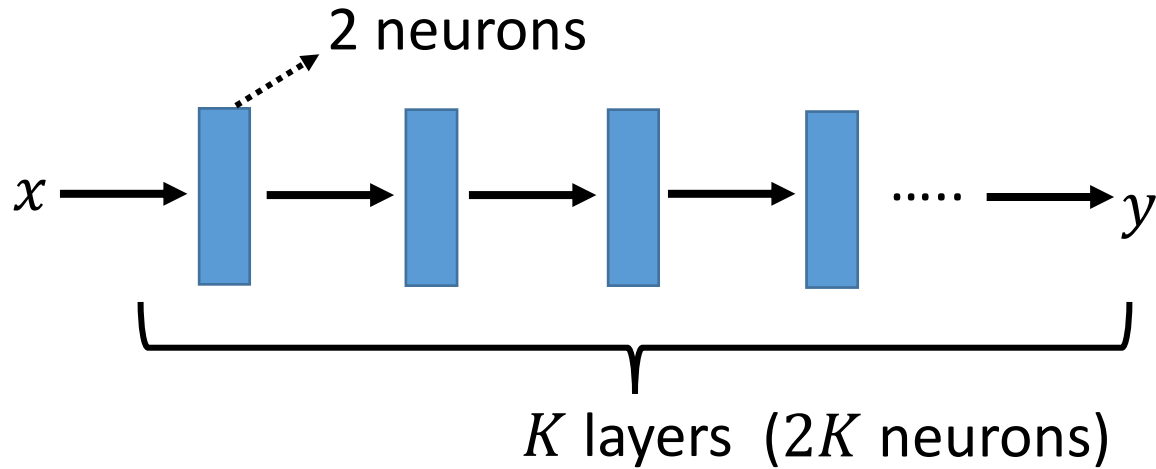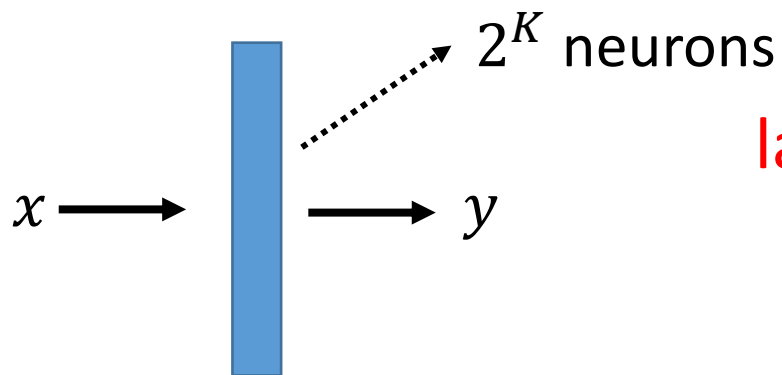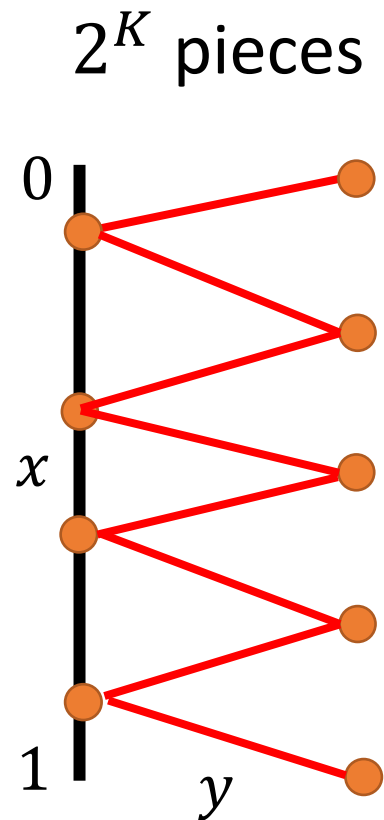
Thinks more ......

- Deep networks outperforms shallow ones when the required functions are <u>complex and regular.</u>

  Image, speech, etc. have this characteristics.

- Deep is exponentially better than shallow even when $y = x^2$.

https://youtu.be/FN8jclCrqY0

https://youtu.be/qpuLxXrHQB4

深度學習是一個讓
魚與熊掌可以兼得的方法

$$h^{all} = arg \min_{h \in \mathcal{H}} L(h, \mathcal{D}_{all})$$

Still small loss

Small (fewer candidates)