计算机视觉
Computer Vision

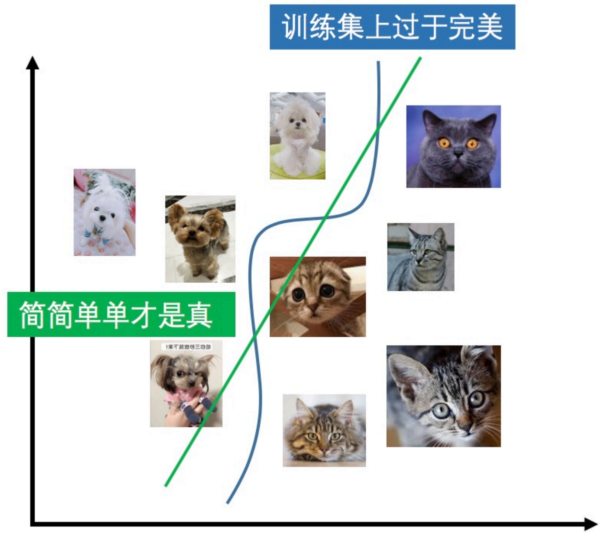
Lecture 4: 神经网络和反向传播

DaSE Data Science & Engineering

EAST CHINA NORMAL UNIVERSITY 華東師範大學

L03：损失函数和优化

- multiclass SVM loss (Hinge loss)：$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$
  ✓SVM分类器
- cross-entropy loss ：$L_i = -\log(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}})$
  ✓Softmax分类器（多类别逻辑回归）

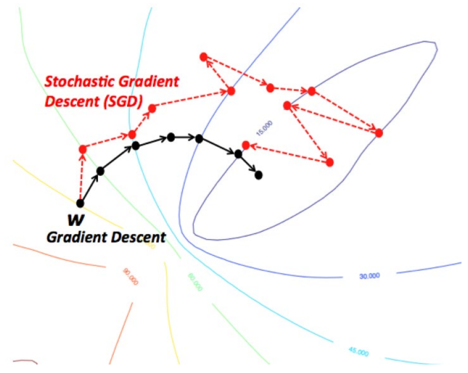


$$L(\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} l(f(\boldsymbol{W}, \boldsymbol{x}_i), y_i) + \lambda R(\boldsymbol{W})$$

正则化
- ✓ 提高模型泛化能力
- ✓ 调整权重分布的偏好

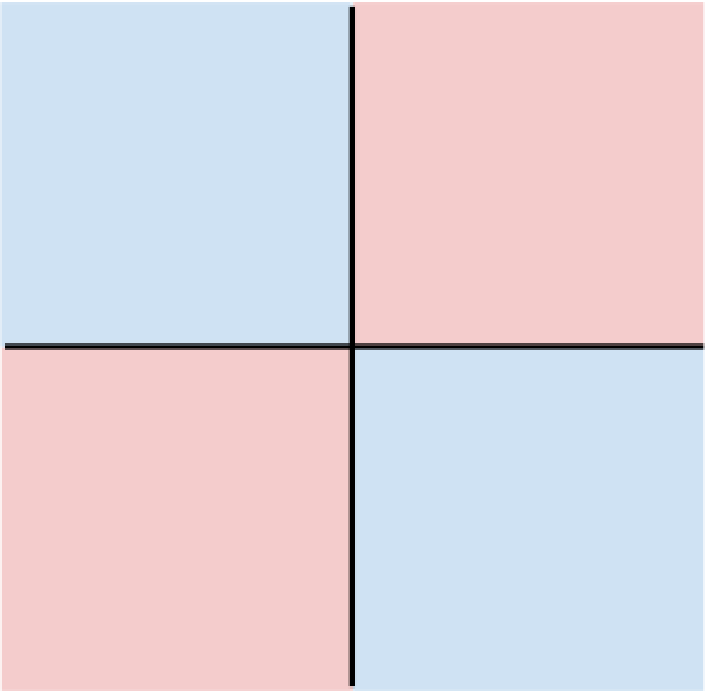
$\boldsymbol{W}$		$\boldsymbol{W}+h$ (第一维)	$\lim_{h\to 0}\frac{L(w+h)-L(w)}{h}$	$\nabla_{\boldsymbol{W}} L$
1.5	+0.001	1.501	(1.2369-1.2356)/0.001=1.3	1.3
3.3		3.3		?
2.1		2.1		?
-0.5		-0.5		?
0.2		0.2		?
…		…		…
Loss=1.2356		Loss=1.2369		

$\boldsymbol{W}$	偏导公式	$\nabla_{\boldsymbol{W}} L$公式	代入数据和$\boldsymbol{W}$	$\nabla_{\boldsymbol{W}} L$值
1.5		$\frac{\partial L(\boldsymbol{W})}{\partial w_1}$		1.3
3.3		$\frac{\partial L(\boldsymbol{W})}{\partial w_2}$		-5.8
2.1		$\frac{\partial L(\boldsymbol{W})}{\partial w_3}$		0
-0.5		$\frac{\partial L(\boldsymbol{W})}{\partial w_4}$		3.7
0.2		$\frac{\partial L(\boldsymbol{W})}{\partial w_5}$		-0.2
…		…		…
Loss=1.2356				



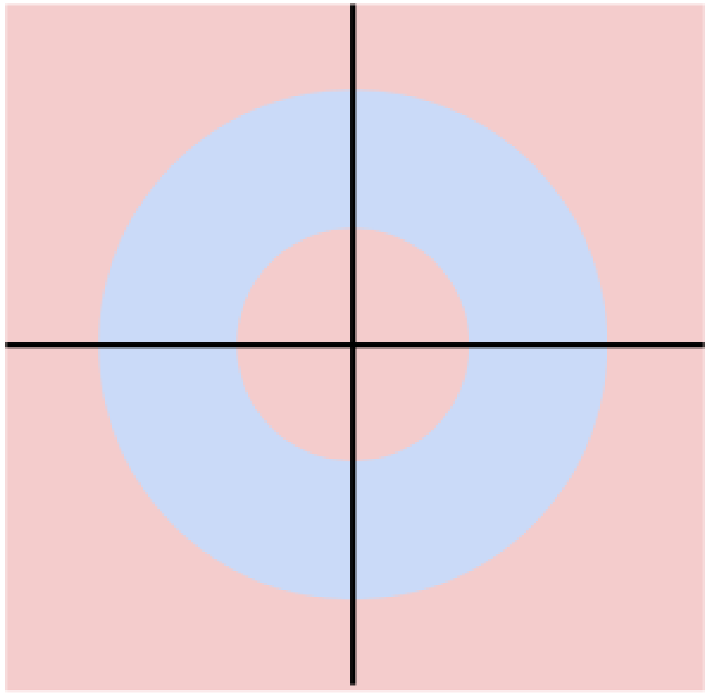
随机梯度下降

线性分类器无法处理的情况（L02）



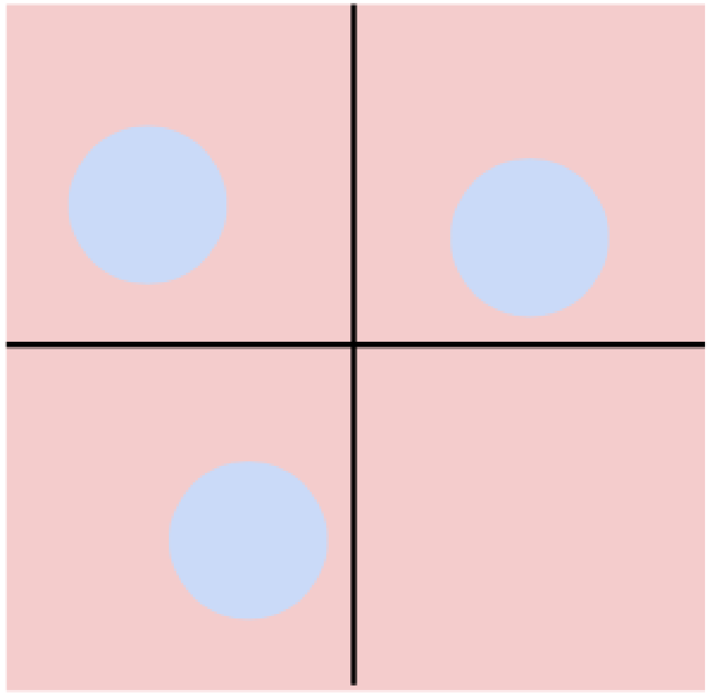
蓝色： $x1 \times x2 < 0$

红色： $x1 \times x2 > 0$



蓝色： $1 \leq L2\ norm < 2$

红色： 其他点



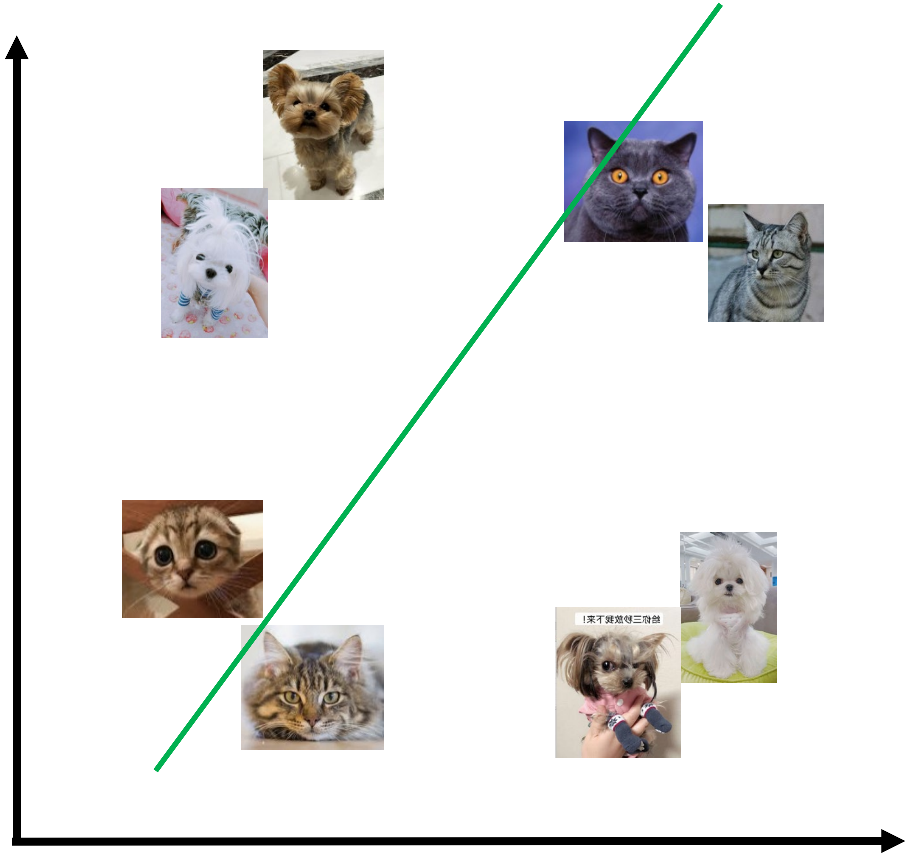
蓝色： 三个离散区域

红色： 其他点

线性分类器无法处理的情况（L02）

$$\begin{bmatrix} w1 & w2 & w3 & w4 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \\ x4 \end{bmatrix} + \boxed{b} = \boxed{\text{score (f(W,x))}}$$

$$f(W,x) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b$$



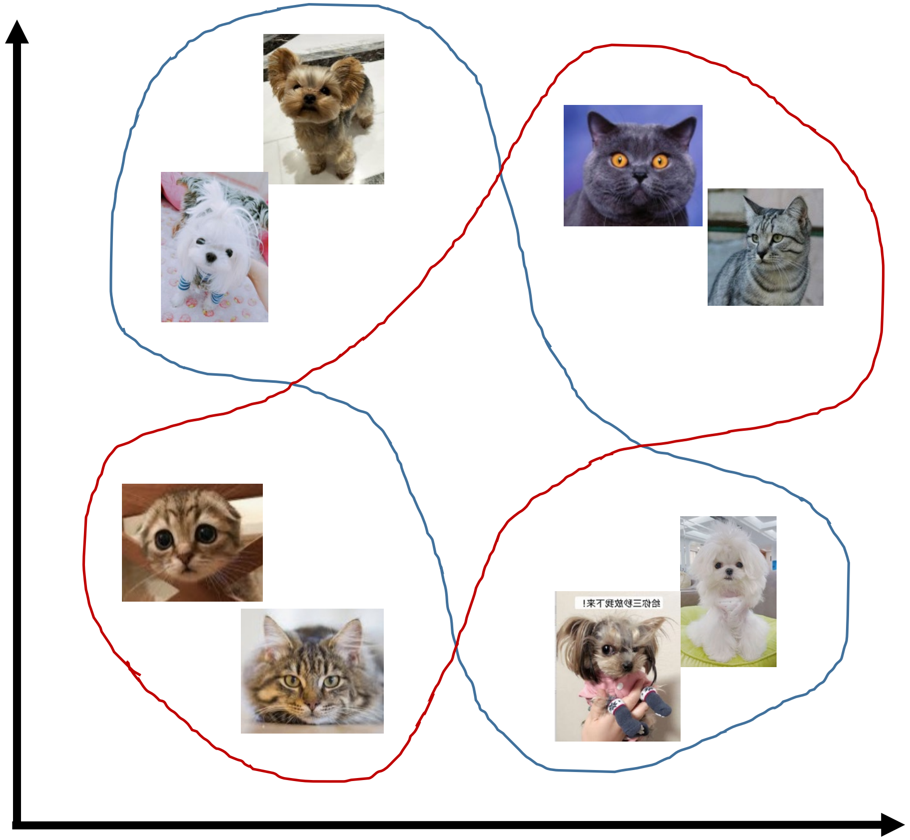
线性分类器无法处理的情况（L02）

w1 w2 w3 w4

x1
x2
x3
x4

\+ b = score (f(W,x))

$$f(W,x) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b$$



可能的方案：增加高阶多项式项

非线性函数

$$f(W,x) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b$$

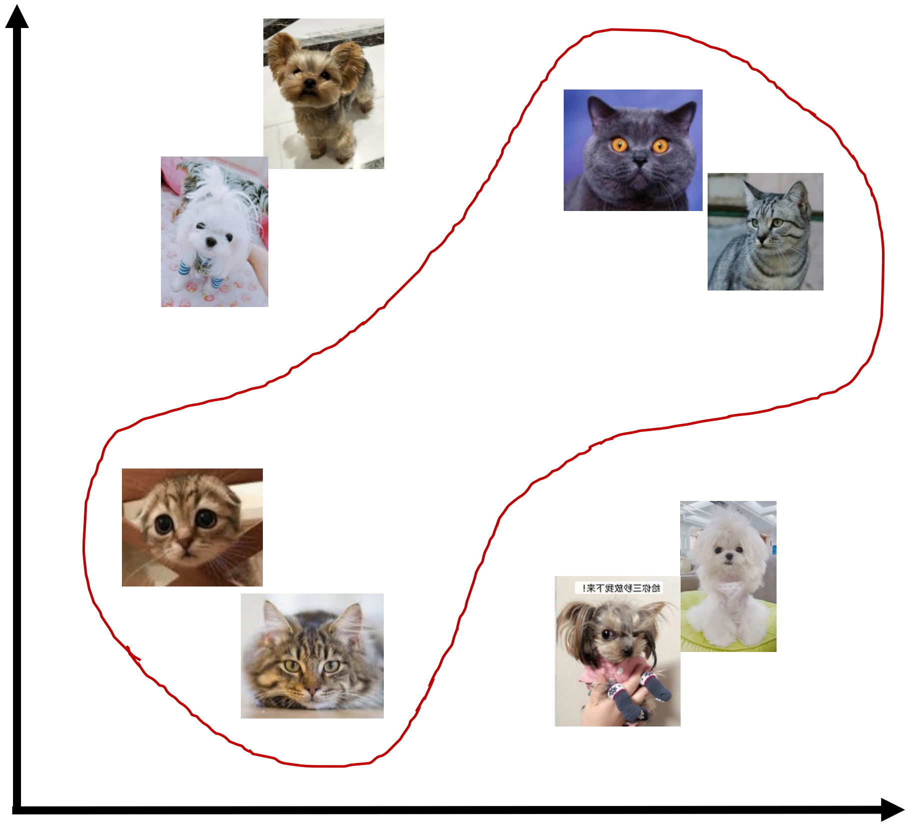
$$+w_{11}{x_1}^2 + w_{12}x_1x_2 + w_{13}x_1x_3 + w_{14}x_1x_4 + \cdots$$

$$+w_{111}{x_1}^3 + w_{112}{x_1}^2x_2 + \cdots$$

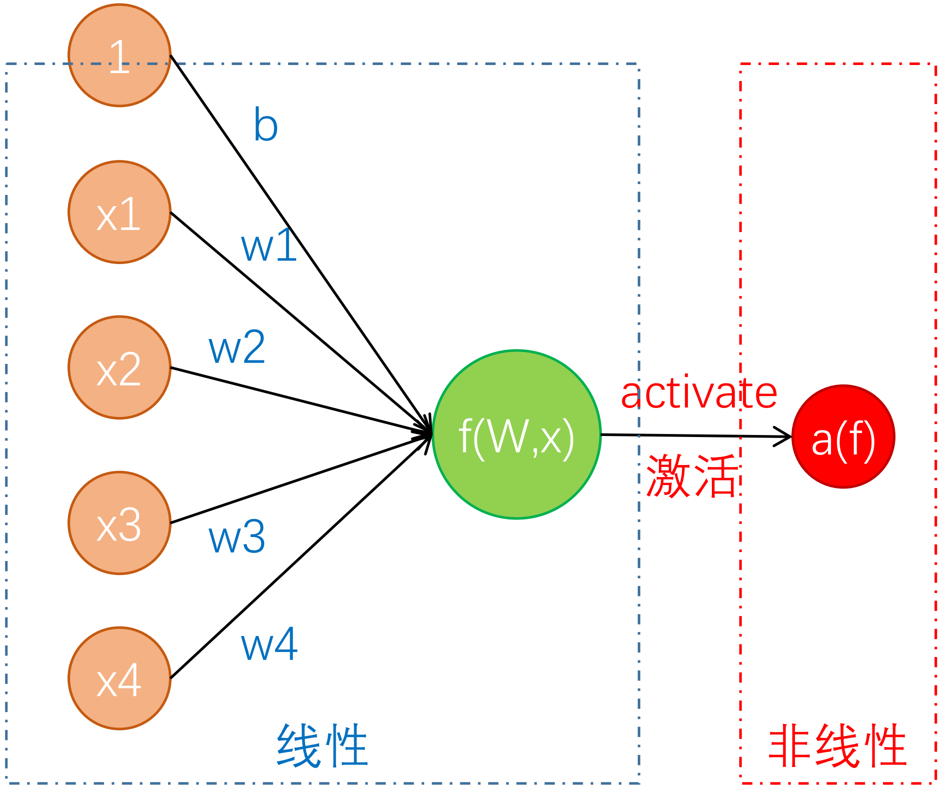
$$+w_{1111}{x_1}^4 + \cdots$$

Q：输入变量有n个，最多有多少参数?

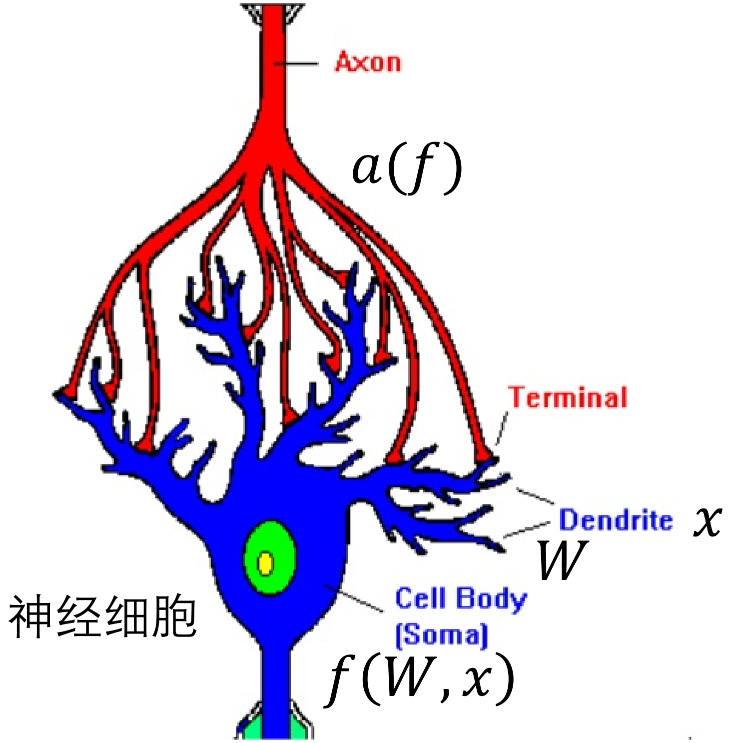
A：$O(n^1) + O(n^2) + O(n^3) + \cdots O(n^n)$!!!



神经网络（Neural Network）



a称为激活函数
(activation function)

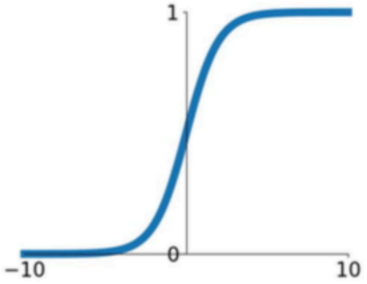


例如Sigmoid： $\sigma(x) = \frac{1}{1+e^{-x}}$

激活函数（Activation functions）

Sigmoid

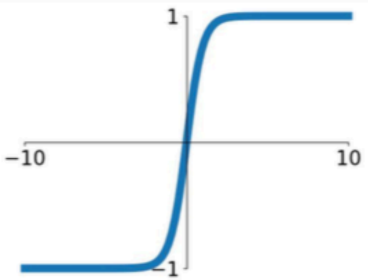
$\sigma(x) = \frac{1}{1+e^{-x}}$



较为通用

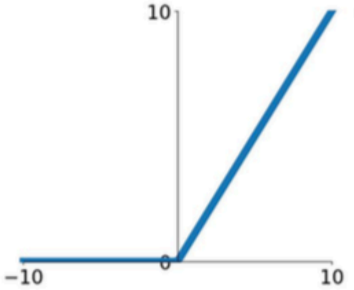
tanh

$\tanh(x)$



ReLU

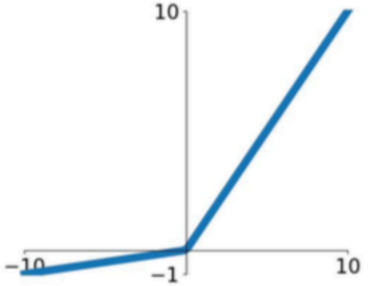
$\max(0, x)$



CV中最常用

Leaky ReLU

$\max(0.1x, x)$

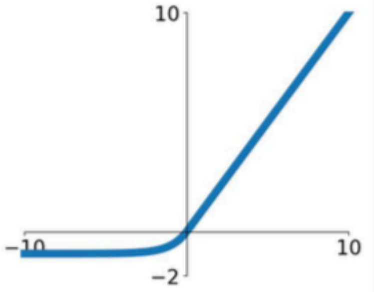


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

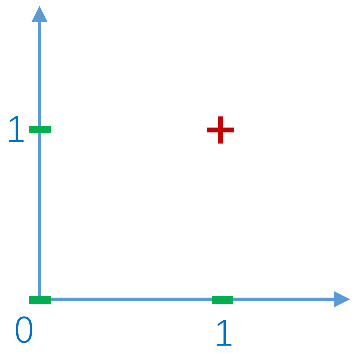
ELU

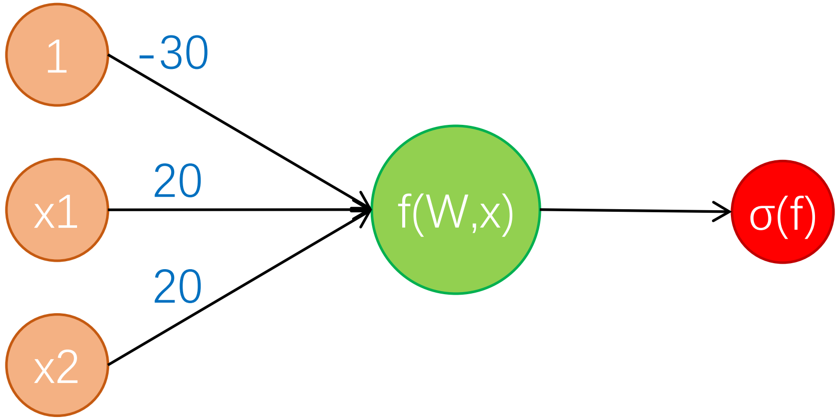
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

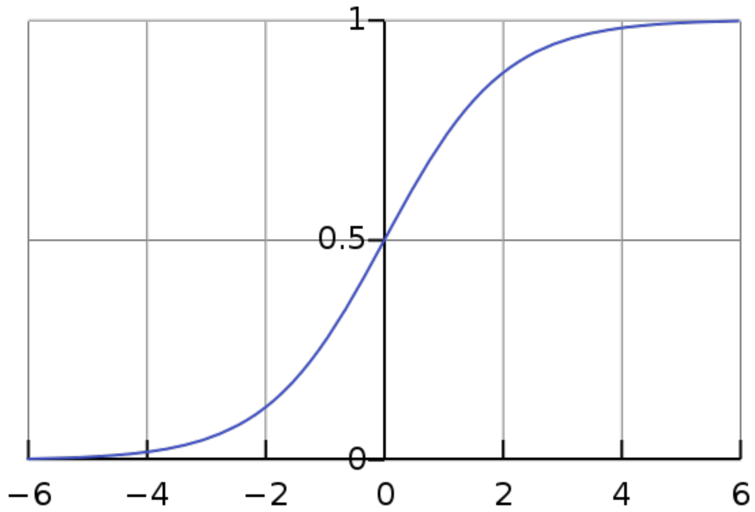


非线性例子

$$x_1, x_2 \in \{0,1\}$$
$$y = x_1 \text{ AND } x_2$$







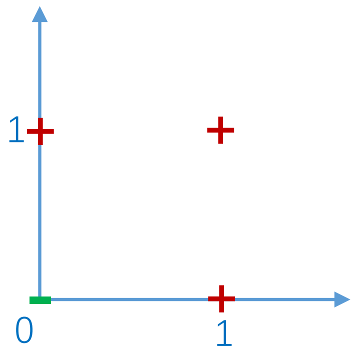
x1	x2	σ(f(W,x))
0	0	σ(-30)≈0
0	1	σ(-10)≈0
1	0	σ(-10)≈0
1	1	σ(10)≈1

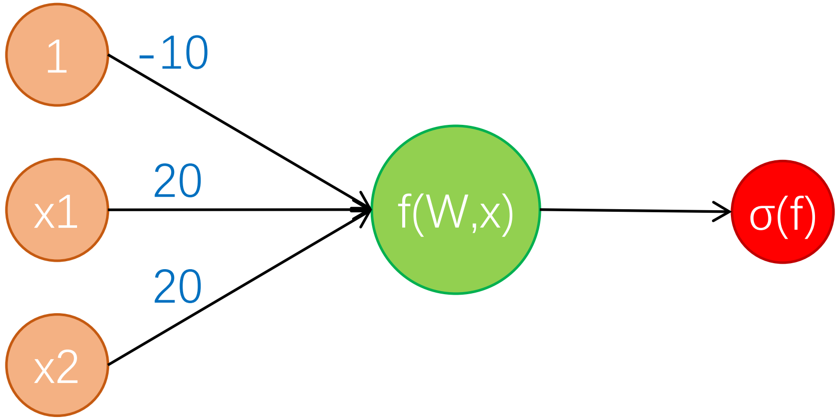
$$f(W,x) = 20x_1 + 20x_2 - 30$$

$$\sigma(f) = \frac{1}{1+e^{-f}}$$

非线性例子

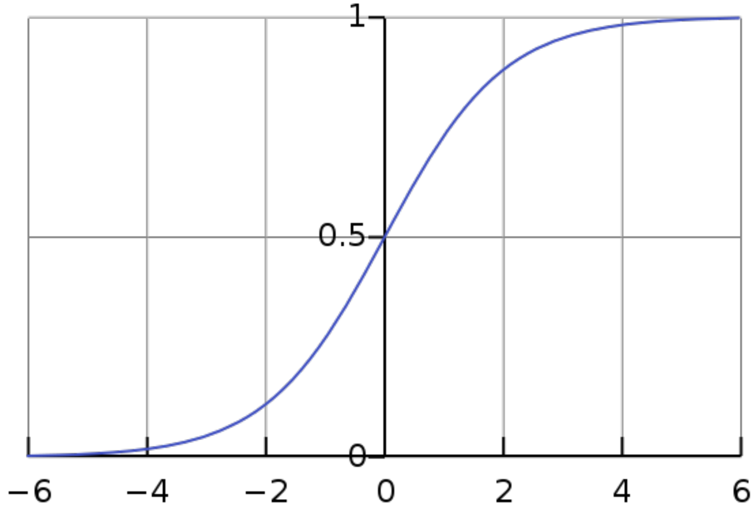
$$x_1, x_2 \in \{0,1\}$$
$$y = x_1 \text{ OR } x_2$$





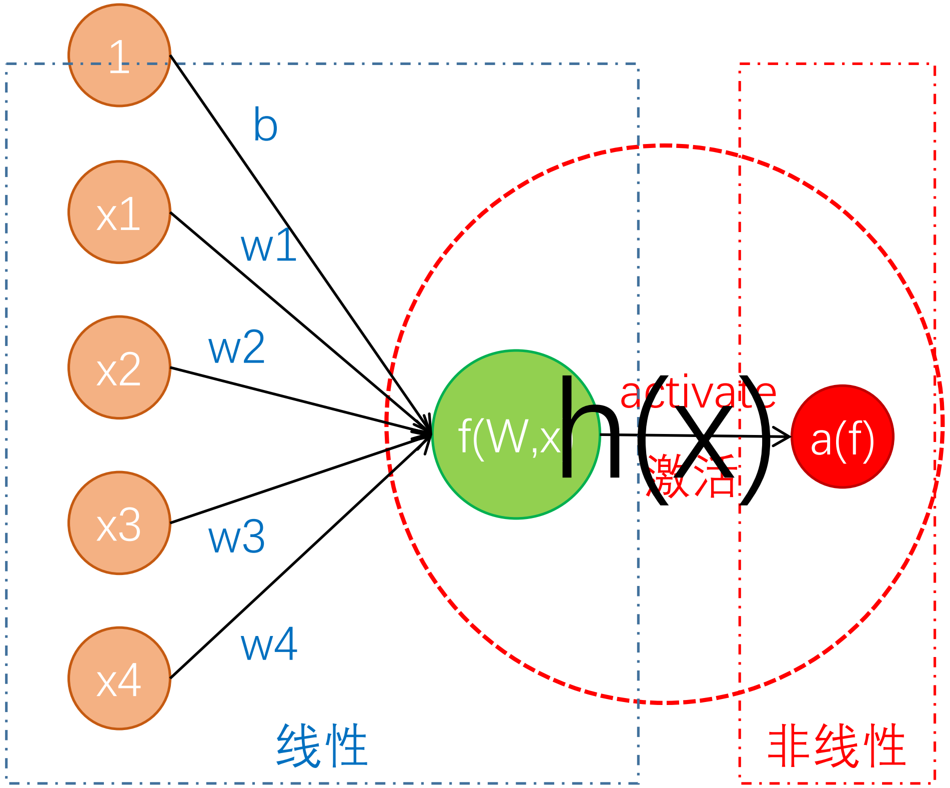
$$f(W,x) = 20x_1 + 20x_2 - 10$$

$$\sigma(f) = \frac{1}{1+e^{-f}}$$

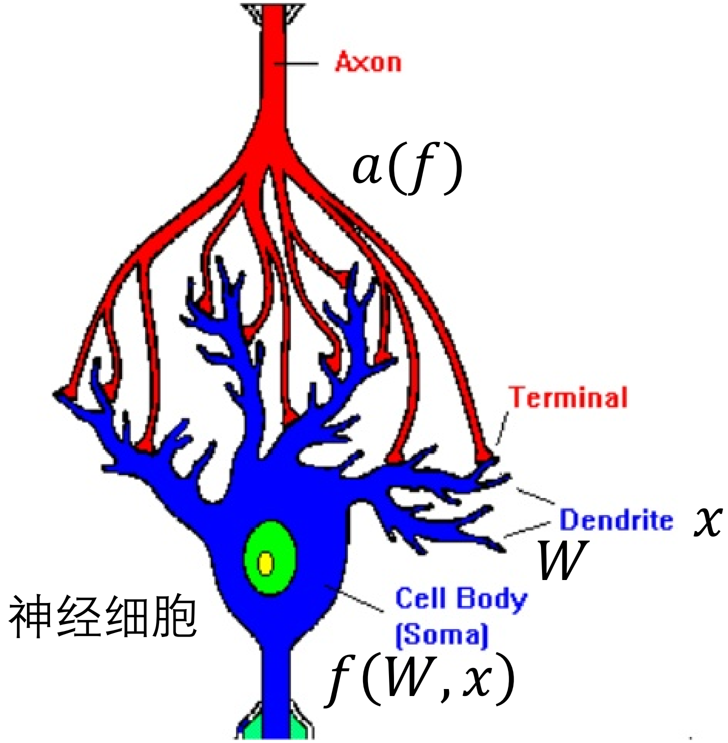


x1	x2	σ(f(W,x))
0	0	σ(-10)≈0
0	1	σ(10)≈1
1	0	σ(10)≈1
1	1	σ(30)≈1

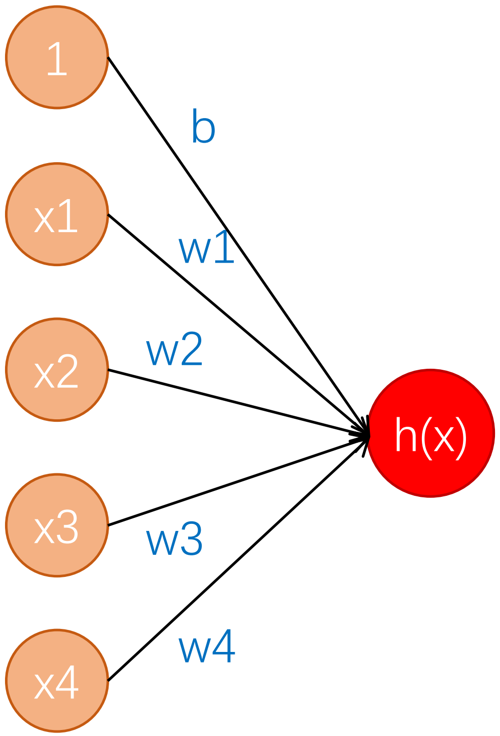
神经网络



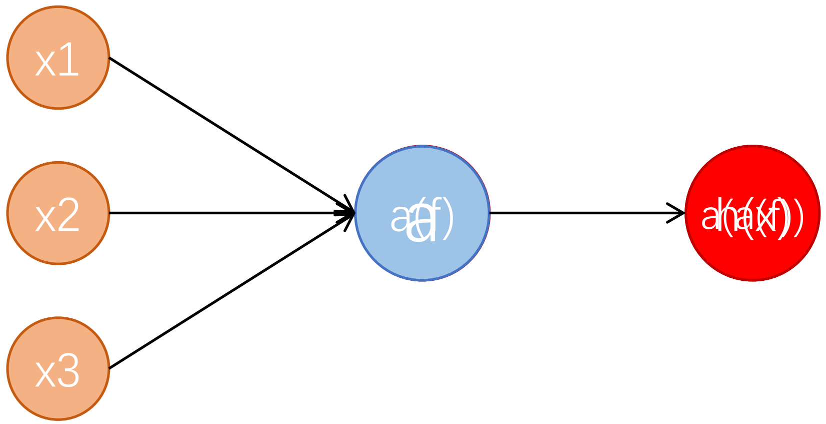
a称为激活函数
(activation function)



神经网络

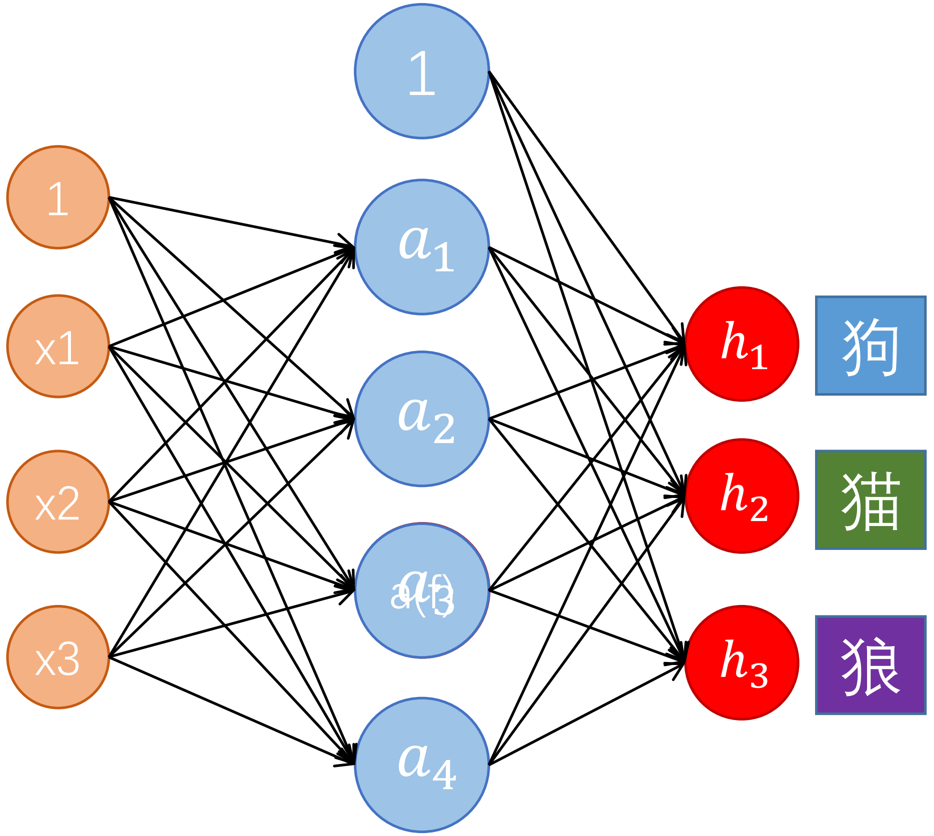


神经网络： hidden layer



input layer　hidden layer　output layer

神经网络



2-layer neural network

hidden layer

output layer

假设使用Sigmoid激活

$$a_1 = \sigma(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1)$$

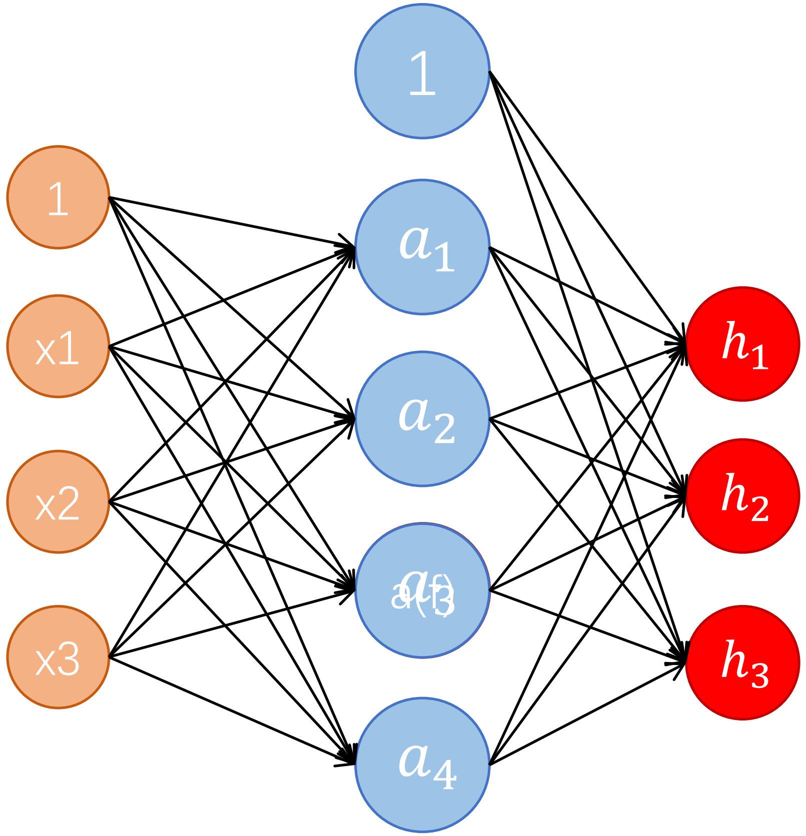
$$a_2 = \sigma(w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2)$$

$$a_3 = \sigma(w_{31}x_1 + w_{32}x_2 + w_{33}x_3 + b_3)$$

$$a_4 = \sigma(w_{41}x_1 + w_{42}x_2 + w_{43}x_3 + b_4)$$

$$h_1 = \sigma(w_{11}a_1 + w_{12}a_2 + w_{13}a_3 + w_{14}a_4 + b_1)$$

神经网络



2-layer neural network

假设使用Sigmoid激活

hidden layer

$$a_1 = \sigma(w_{11}^1 x_1 + w_{12}^1 x_2 + w_{13}^1 x_3 + b_1^1)$$

$$a_2 = \sigma(w_{21}^1 x_1 + w_{22}^1 x_2 + w_{23}^1 x_3 + b_2^1)$$

$$a_3 = \sigma(w_{31}^1 x_1 + w_{32}^1 x_2 + w_{33}^1 x_3 + b_3^1)$$

$$a_4 = \sigma(w_{41}^1 x_1 + w_{42}^1 x_2 + w_{43}^1 x_3 + b_4^1)$$

output layer

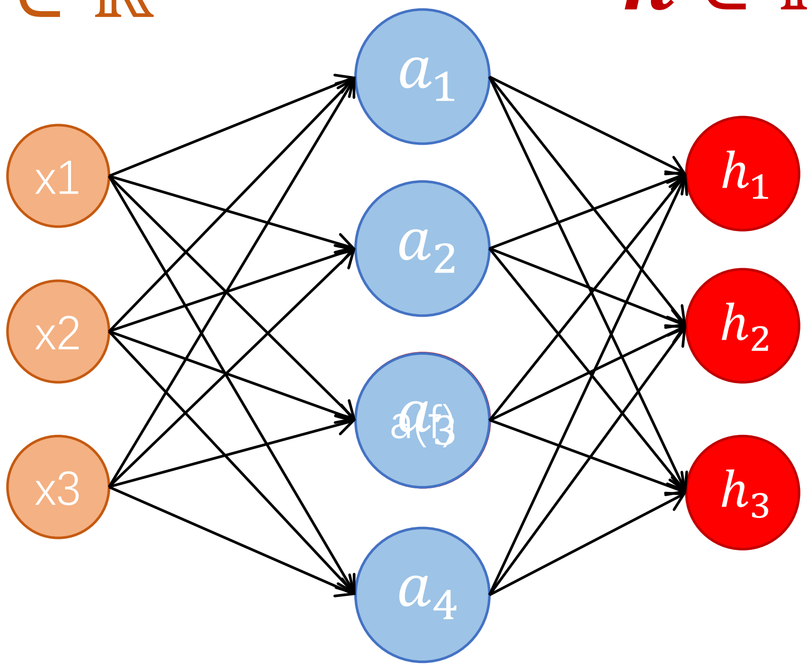
$$h_1 = \sigma(w_{11}^2 a_1 + w_{12}^2 a_2 + w_{13}^2 a_3 + w_{14}^2 a_4 + b_1^2)$$

$$h_2 = \sigma(w_{21}^2 a_1 + w_{22}^2 a_2 + w_{23}^2 a_3 + w_{24}^2 a_4 + b_2^2)$$

$$h_3 = \sigma(w_{31}^2 a_1 + w_{32}^2 a_2 + w_{33}^2 a_3 + w_{34}^2 a_4 + b_3^2)$$

神经网络

$\boldsymbol{x} \in \mathbb{R}^{D}$ $\boldsymbol{a} \in \mathbb{R}^{H}$ $\boldsymbol{h} \in \mathbb{R}^{C}$



input layer hidden layer output layer

2-layer neural network

假设使用Sigmoid激活

hidden layer

$$a_1 = \sigma(w_{11}^1 x_1 + w_{12}^1 x_2 + w_{13}^1 x_3 + b_1^1)$$

$$a_2 = \sigma(w_{21}^1 x_1 + w_{22}^1 x_2 + w_{23}^1 x_3 + b_2^1)$$

$$a_3 = \sigma(w_{31}^1 x_1 + w_{32}^1 x_2 + w_{33}^1 x_3 + b_3^1)$$

$$a_4 = \sigma(w_{41}^1 x_1 + w_{42}^1 x_2 + w_{43}^1 x_3 + b_4^1)$$

$$\boldsymbol{W}^{\mathbf{1}} \in \mathbb{R}^{H \times D}$$

output layer

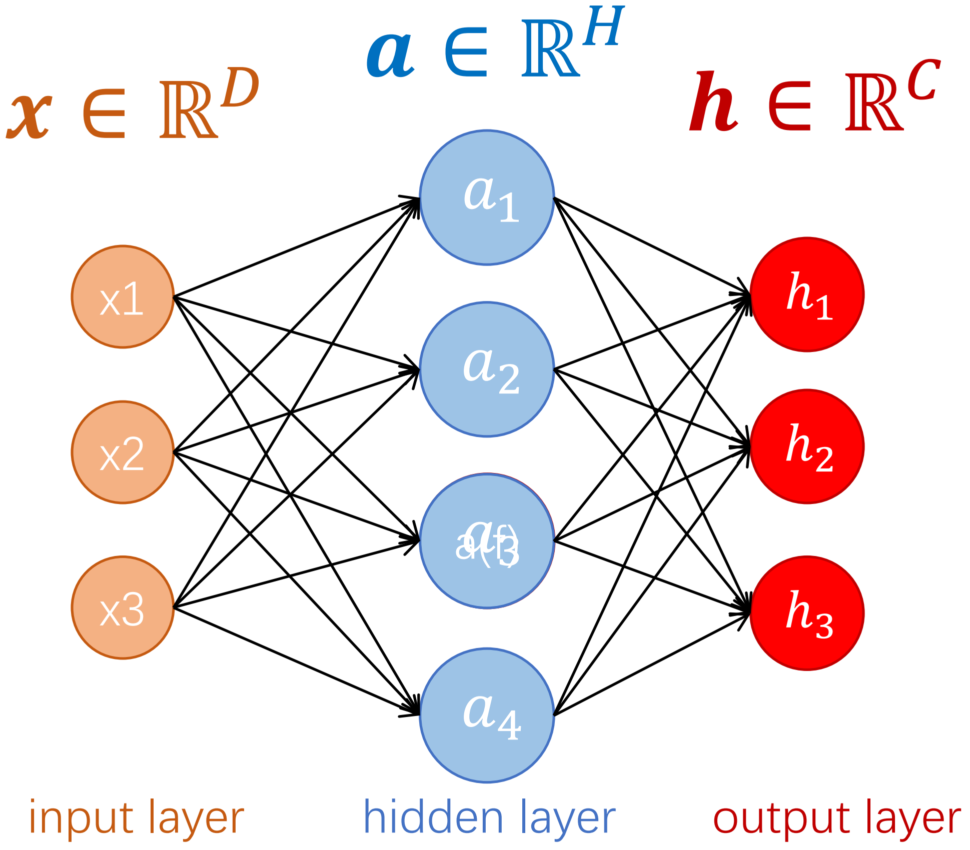
$$h_1 = \sigma(w_{11}^2 a_1 + w_{12}^2 a_2 + w_{13}^2 a_3 + w_{14}^2 a_4 + b_1^2)$$

$$h_2 = \sigma(w_{21}^2 a_1 + w_{22}^2 a_2 + w_{23}^2 a_3 + w_{24}^2 a_4 + b_2^2)$$

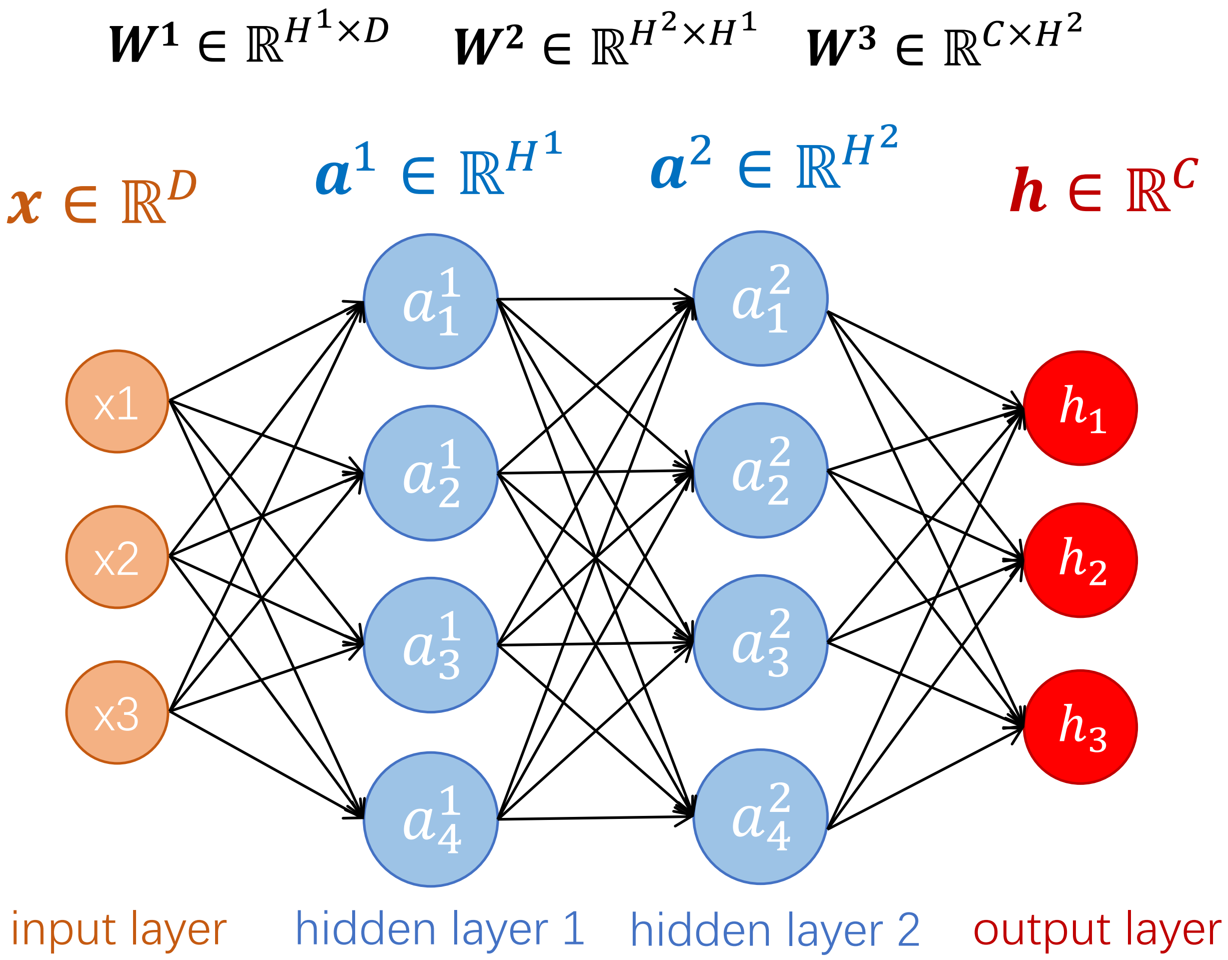
$$h_3 = \sigma(w_{31}^2 a_1 + w_{32}^2 a_2 + w_{33}^2 a_3 + w_{34}^2 a_4 + b_3^2)$$

$$\boldsymbol{W}^{\mathbf{2}} \in \mathbb{R}^{C \times H}$$

神经网络



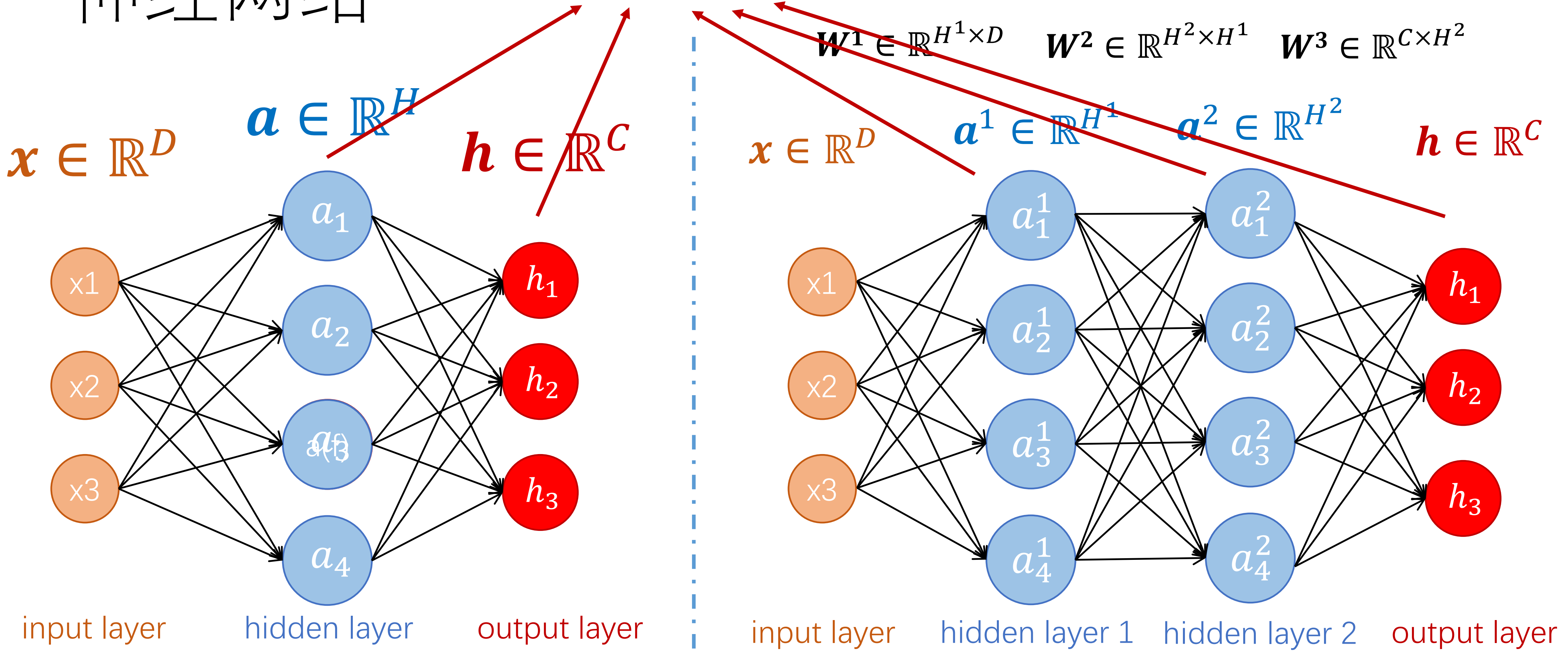
2-layer neural network



3-layer neural network

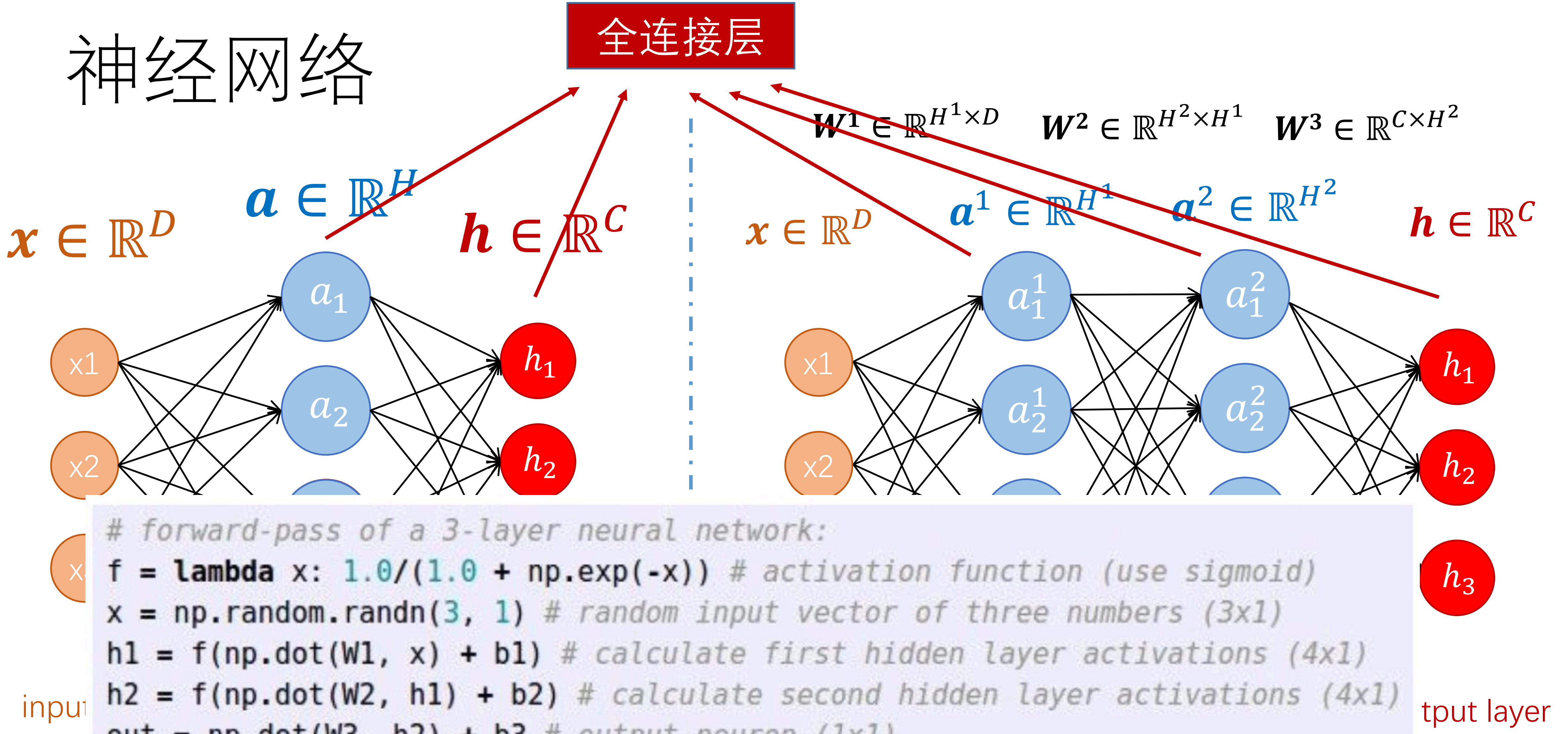
神经网络

全连接层



2-layer neural network **3-layer neural network**

神经网络

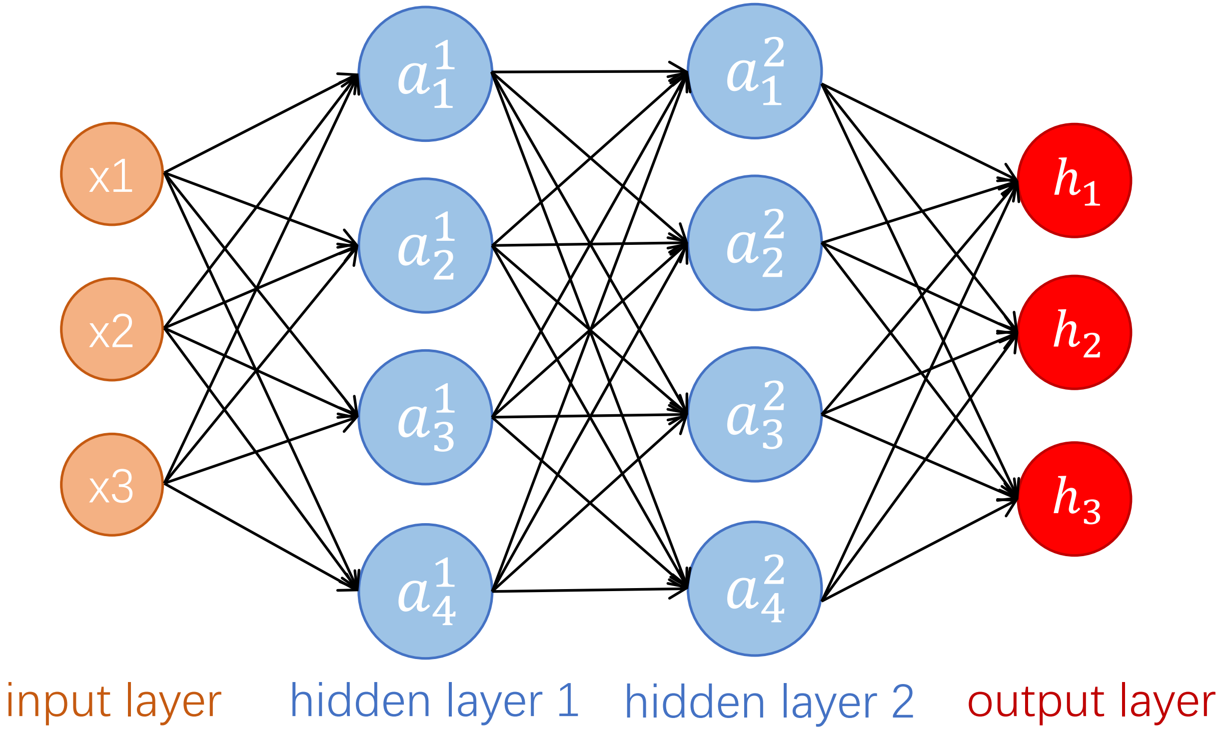



```python
# forward-pass of a 3-layer neural network:
f = lambda x: 1.0/(1.0 + np.exp(-x)) # activation function (use sigmoid)
x = np.random.randn(3, 1) # random input vector of three numbers (3x1)
h1 = f(np.dot(W1, x) + b1) # calculate first hidden layer activations (4x1)
h2 = f(np.dot(W2, h1) + b2) # calculate second hidden layer activations (4x1)
out = np.dot(W3, h2) + b3 # output neuron (1x1)
```

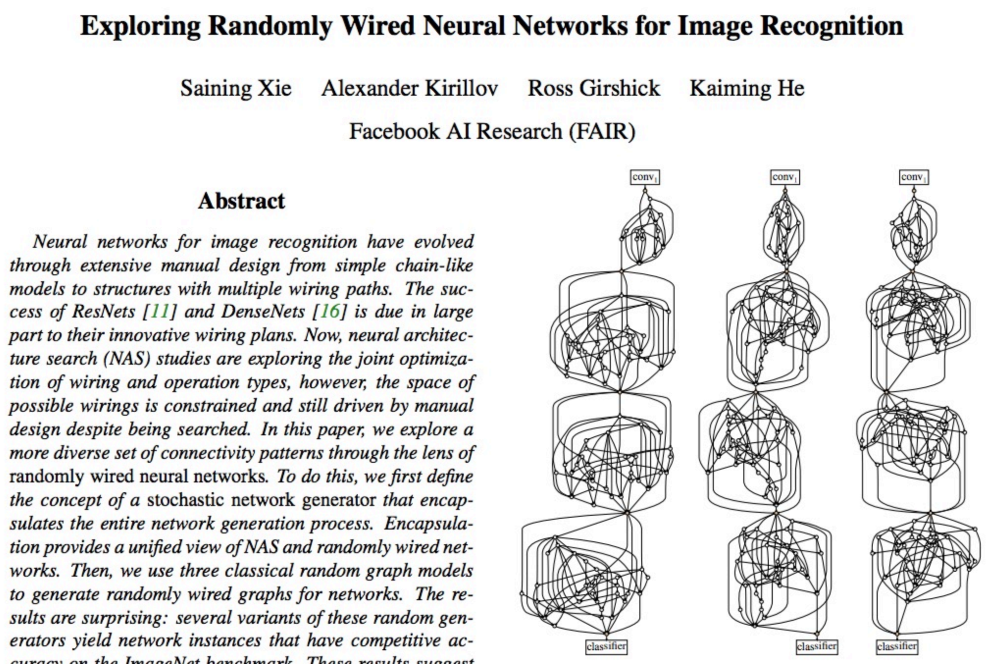
2-layer neural network

3-layer neural network

随机连接的神经网络



全连接神经网络



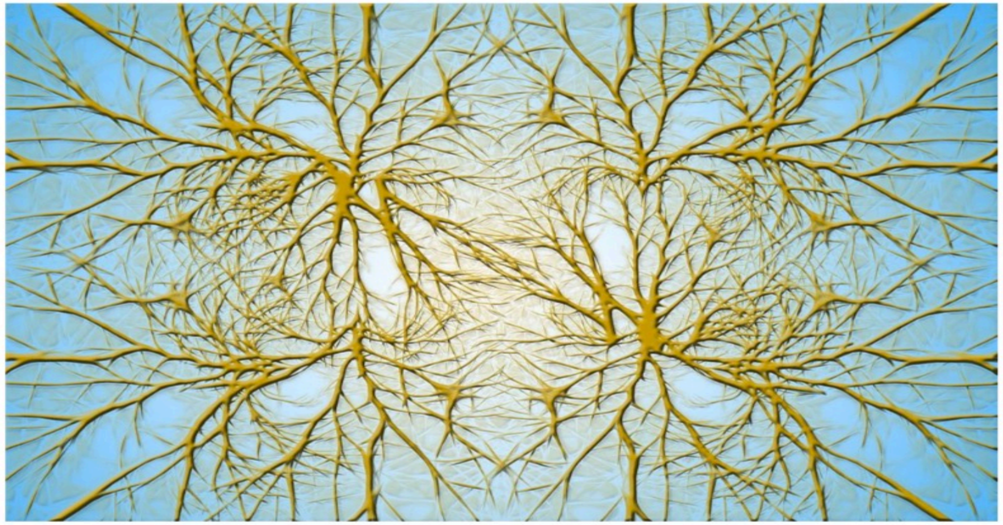
Exploring Randomly Wired Neural Networks for Image Recognition

Saining Xie Alexander Kirillov Ross Girshick Kaiming He

Facebook AI Research (FAIR)

Abstract

Neural networks for image recognition have evolved through extensive manual design from simple chain-like models to structures with multiple wiring paths. The success of ResNets [11] and DenseNets [16] is due in large part to their innovative wiring plans. Now, neural architecture search (NAS) studies are exploring the joint optimization of wiring and operation types, however, the space of possible wirings is constrained and still driven by manual design despite being searched. In this paper, we explore a more diverse set of connectivity patterns through the lens of randomly wired neural networks. To do this, we first define the concept of a stochastic network generator that encapsulates the entire network generation process. Encapsulation provides a unified view of NAS and randomly wired networks. Then, we use three classical random graph models to generate randomly wired graphs for networks. The results are surprising: several variants of these random generators yield network instances that have competitive accuracy on the ImageNet benchmark. These results suggest



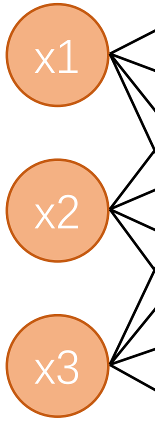
This image is CC0 Public Domain

随机连接的神经网络

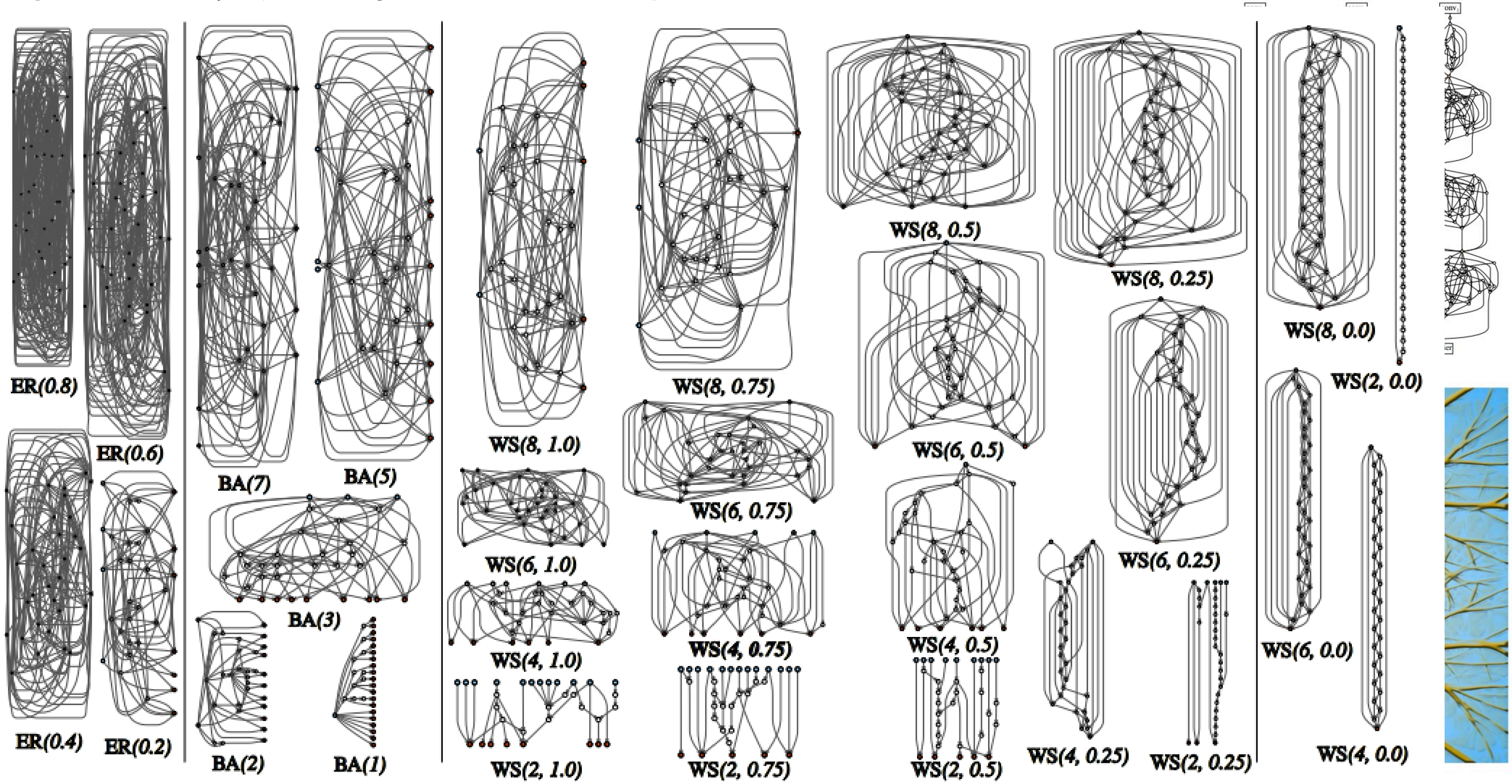
Exploring Randomly Wired Neural Networks for Image Recognition

Saining Xie Alexander Kirillov Ross Girshick Kaiming He

Facebook AI Research (FAIR)

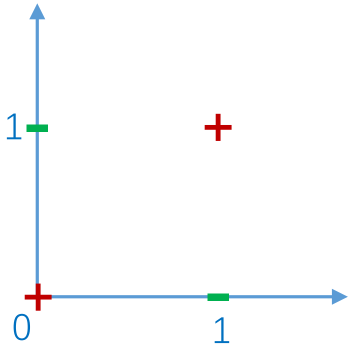


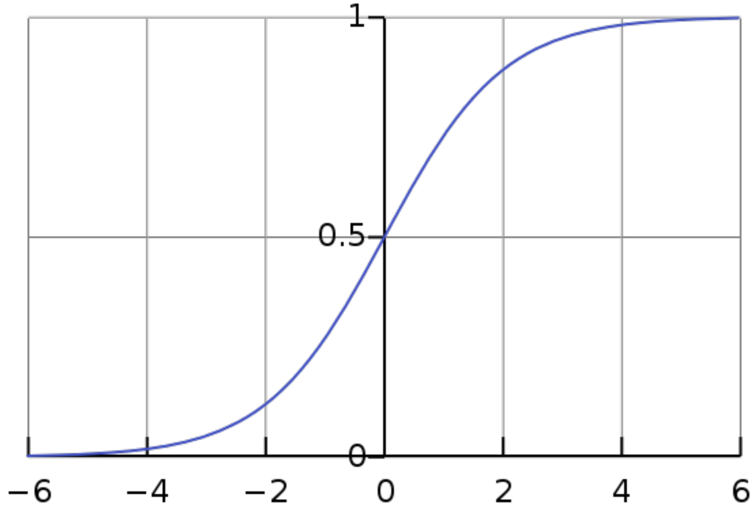
input laye

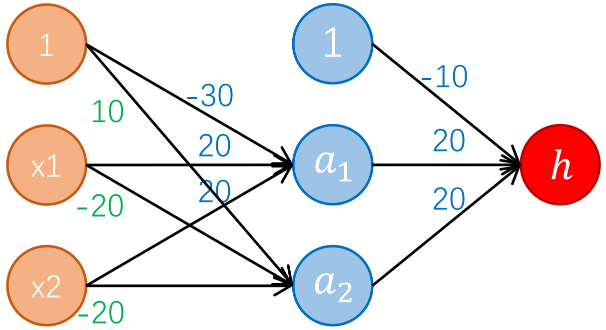


两层神经网络例子

$$x_1, x_2 \in \{0,1\}$$
$$y = x_1 \text{ XNOR } x_2$$



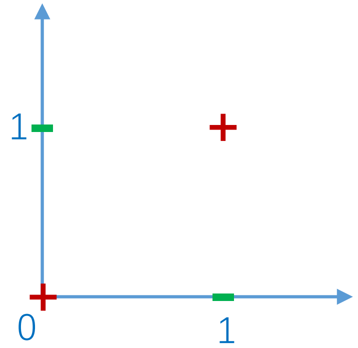


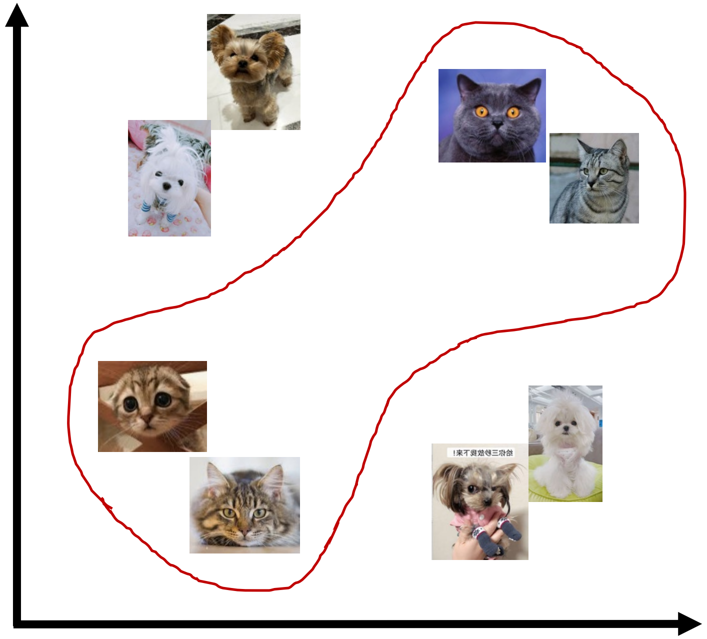


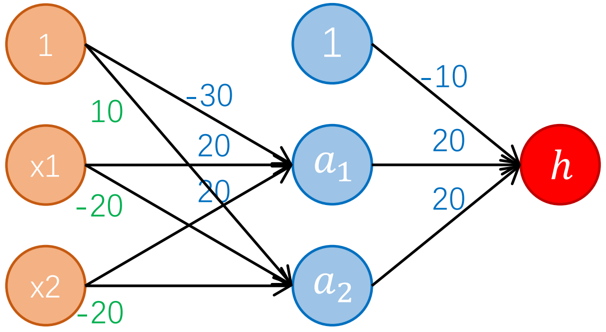
x1	x2	a_1	a_2	h
0	0	σ(-30)≈0	σ(10)≈1	σ(10)≈1
0	1	σ(-10)≈0	σ(-10)≈0	σ(-10)≈0
1	0	σ(-10)≈0	σ(-10)≈0	σ(-10)≈0
1	1	σ(10)≈1	σ(-30)≈0	σ(10)≈1

两层神经网络例子

$$x_1, x_2 \in \{0,1\}$$
$$y = x_1 \text{ XNOR } x_2$$



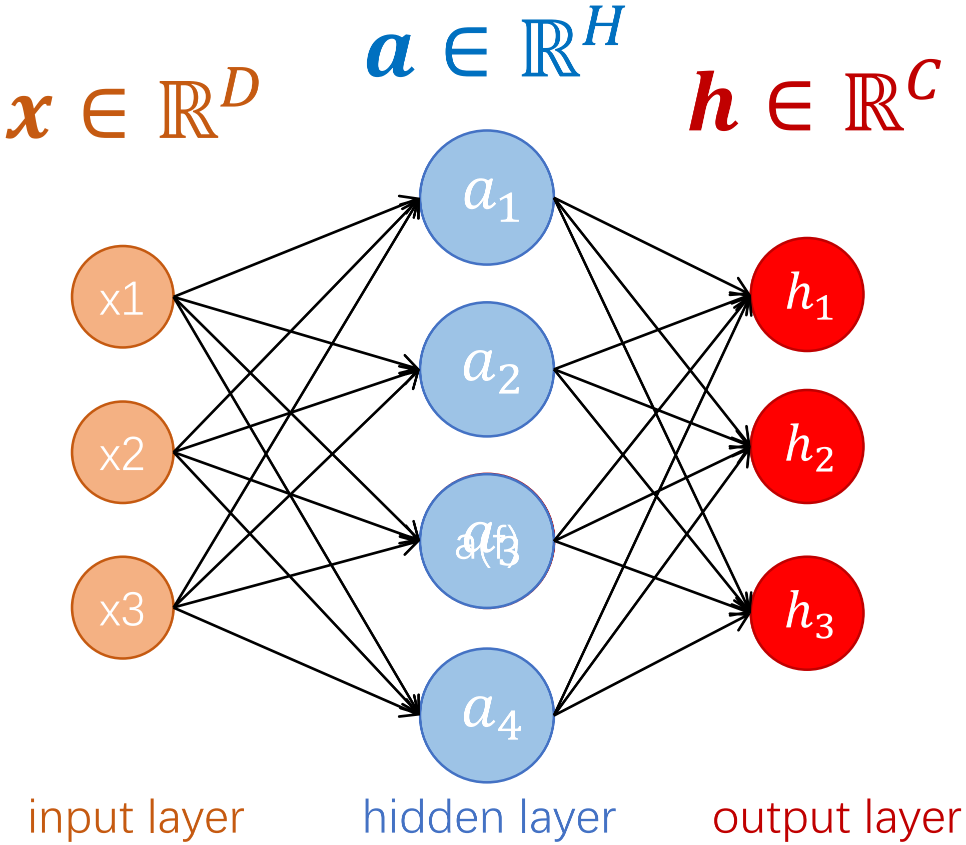




x1	x2	a_1	a_2	h
0	0	σ(-30)≈0	σ(10)≈1	σ(10)≈1
0	1	σ(-10)≈0	σ(-10)≈0	σ(-10)≈0
1	0	σ(-10)≈0	σ(-10)≈0	σ(-10)≈0
1	1	σ(10)≈1	σ(-30)≈0	σ(10)≈1

神经网络计算

训练集：$(x_i, y_i)_{i=1}^N$



2-layer neural network

ReLU $\max(0, x)$

$$\boldsymbol{a} = \max(0,\ f(\boldsymbol{x}; \boldsymbol{W^1})); \quad \boldsymbol{s} = f(\boldsymbol{a}, \boldsymbol{W^2});$$

Softmax

$$h_k = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Loss func

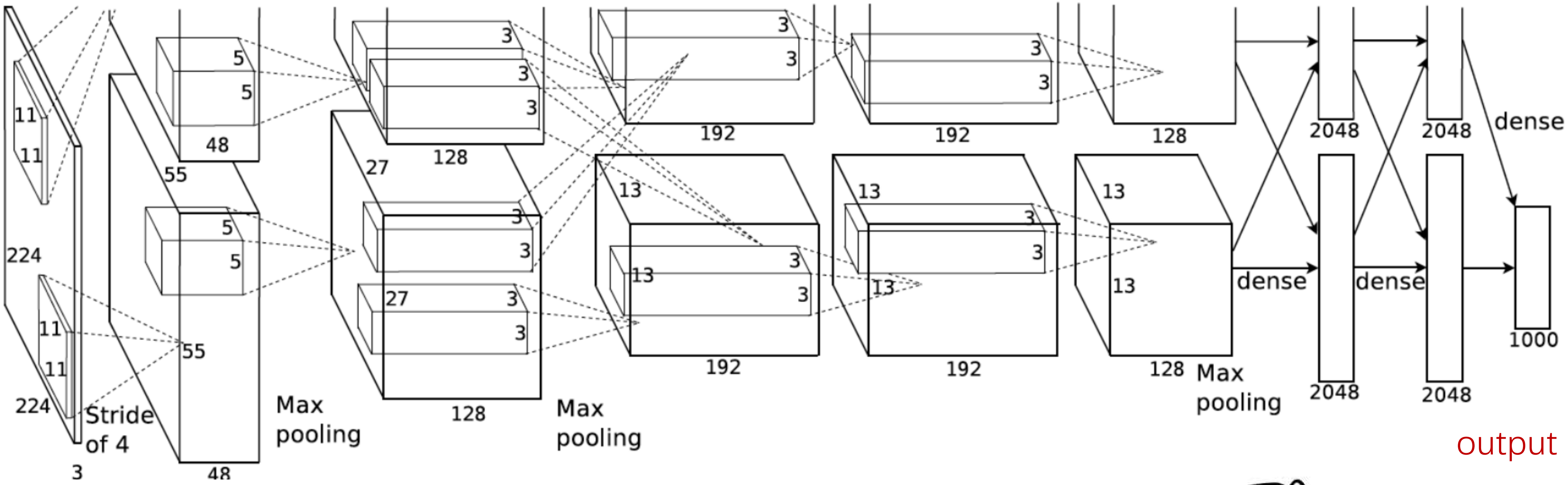
$$L = \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{h}_i, y_i) + \lambda R(\boldsymbol{W^1}) + \lambda R(\boldsymbol{W^2})$$

优化

计算 $\frac{\partial L}{\partial \boldsymbol{W^1}}, \frac{\partial L}{\partial \boldsymbol{W^2}}$ $\quad \frac{\partial L}{\partial W_{ij}^k}, k = 1,2$

$$\nabla_W L = \nabla_W [\frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{h}_i, y_i) + \lambda R(\boldsymbol{W^1}) + \lambda R(\boldsymbol{W^2})]$$

AlexNet



output layer

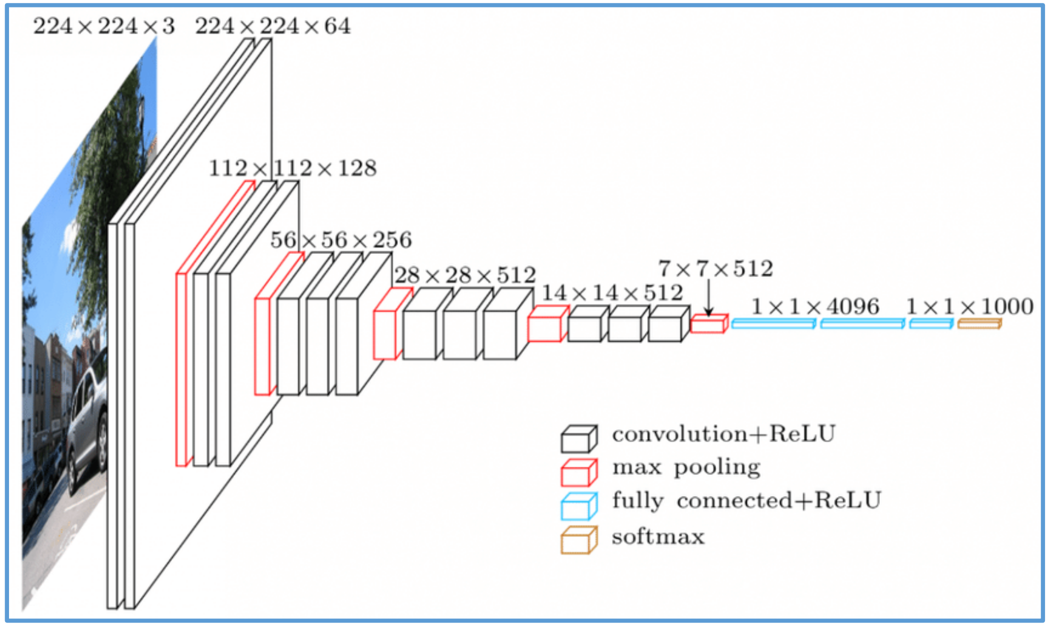
input layer

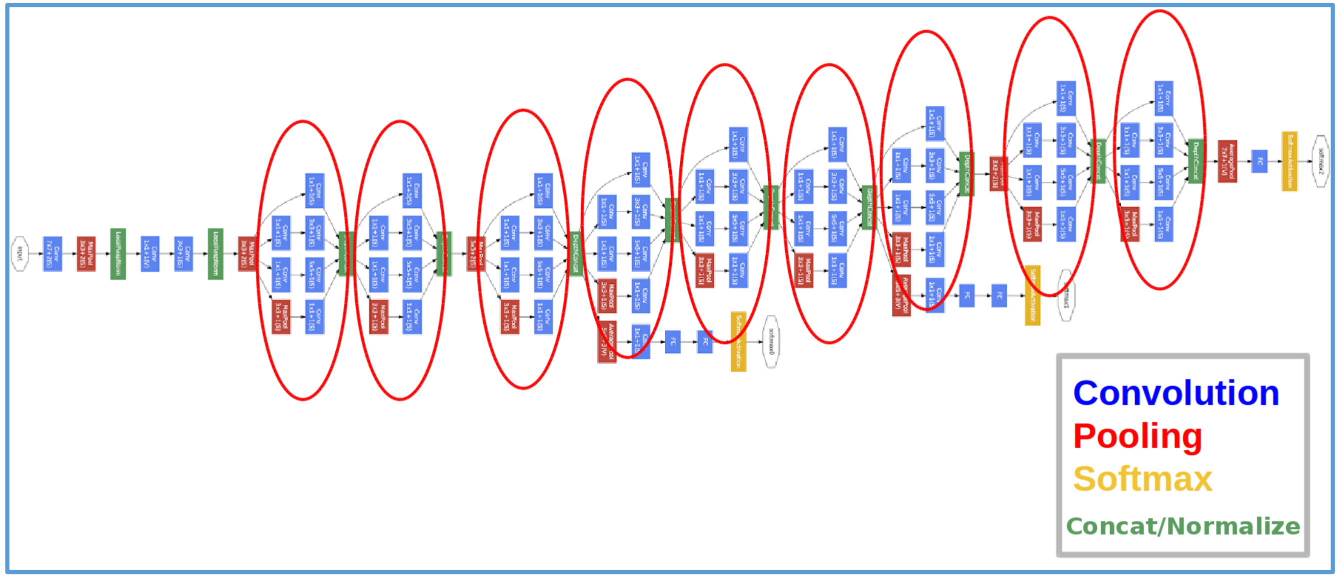
hidden layer

$\nabla_W L$

心好累

VGG, GoogleNet





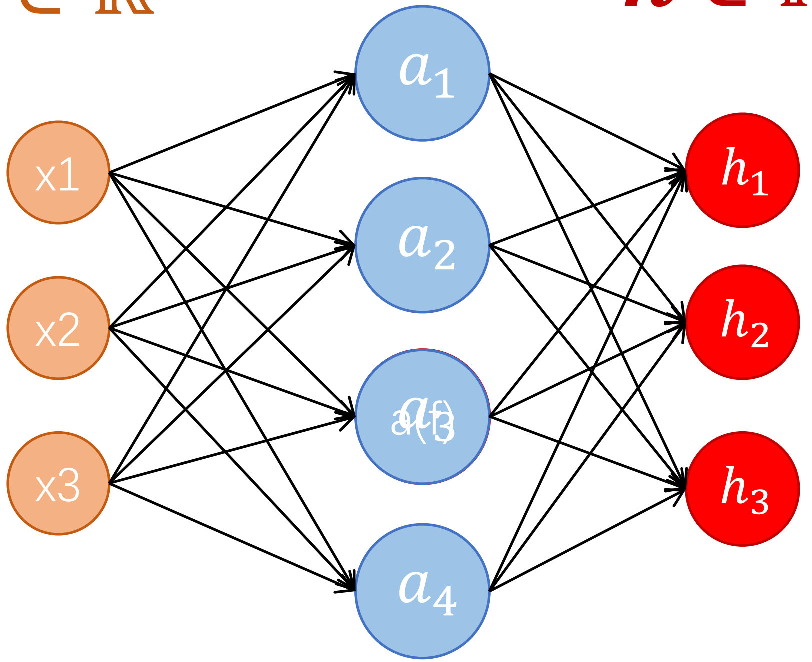
$\nabla_W L$

神经网络

训练集：$(x_i, y_i)_{i=1}^{N}$

前向传递（forward pass）

$\boldsymbol{x} \in \mathbb{R}^D$ $\quad \boldsymbol{a} \in \mathbb{R}^H$ $\quad \boldsymbol{h} \in \mathbb{R}^C$



input layer hidden layer output layer

2-layer neural network

ReLU
$\max(0, x)$

$$\boldsymbol{a} = \max(0,\ f(\boldsymbol{x}; \boldsymbol{W}^{\mathbf{1}})); \qquad \boldsymbol{s} = f(\boldsymbol{a}, \boldsymbol{W}^{\mathbf{2}});$$

Softmax

$$h_k = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Loss func

$$L = \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{h}_i, y_i) + \lambda R(\boldsymbol{W}^{\mathbf{1}}) + \lambda R(\boldsymbol{W}^{\mathbf{2}})$$

反向传递（backward pass）

反向传播

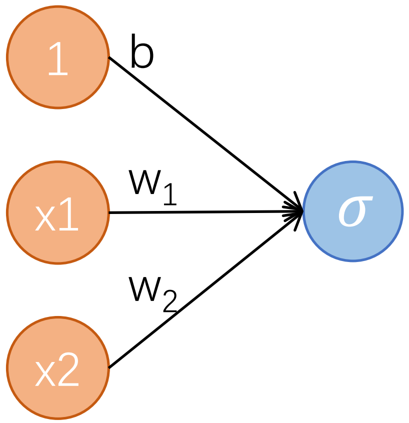
(Backpropagation)

$$\frac{\partial L}{\partial \boldsymbol{W}^{\mathbf{2}}} = \frac{\partial L}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{W}^{\mathbf{2}}}, \qquad \frac{\partial L}{\partial \boldsymbol{W}^{\mathbf{1}}} = \frac{\partial L}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}}\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{W}^{\mathbf{1}}}$$

Chain rule

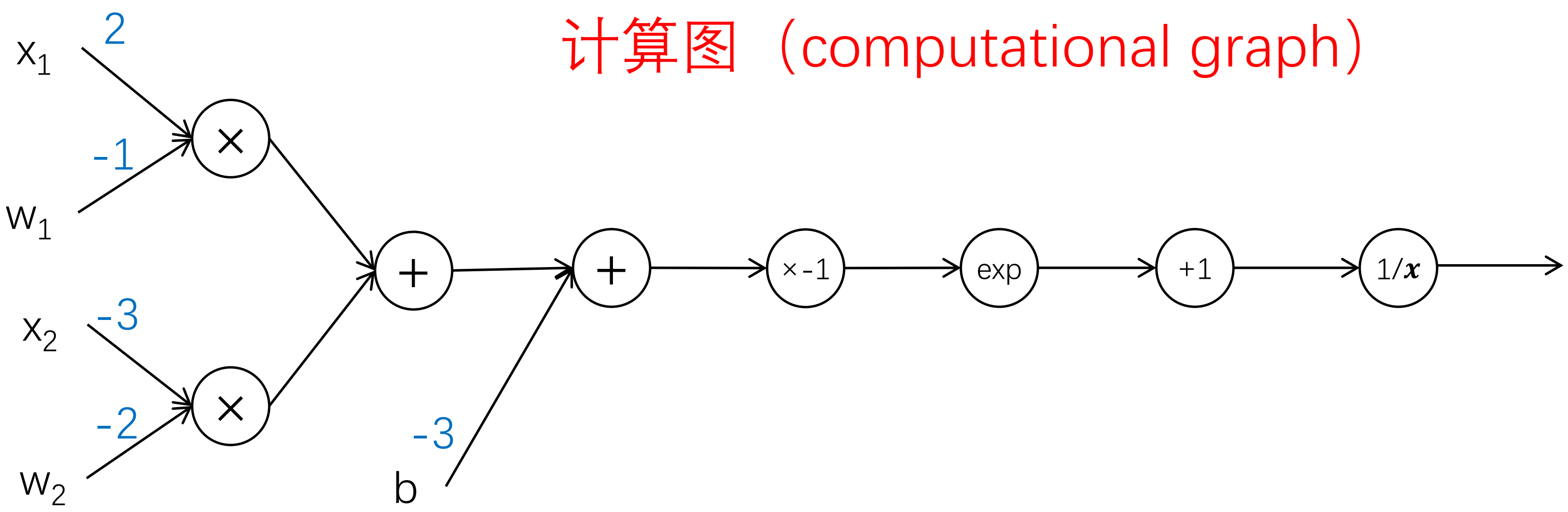
反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$



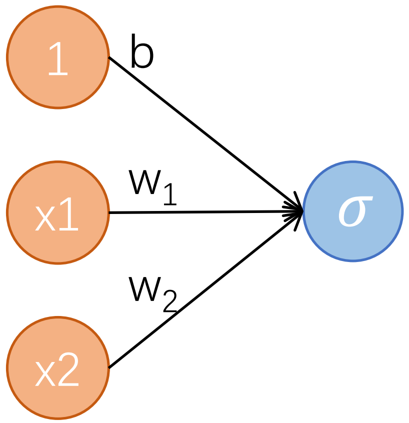
$$\frac{\partial\sigma}{\partial w_1}, \frac{\partial\sigma}{\partial w_2}, \frac{\partial\sigma}{\partial b}, \frac{\partial\sigma}{\partial x_1}, \frac{\partial\sigma}{\partial x_2}$$

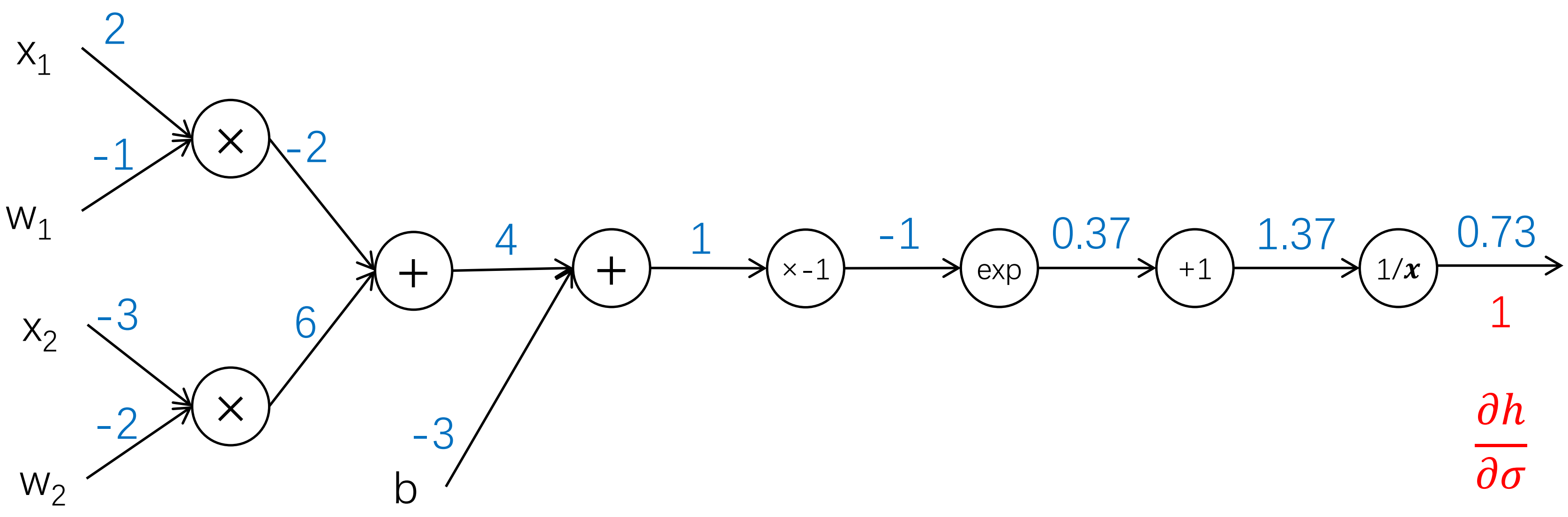
计算图（computational graph）



反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





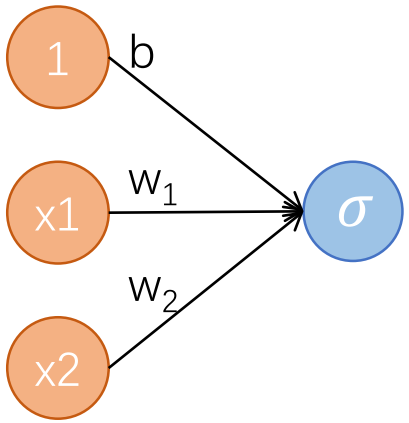
$$\frac{\partial\sigma}{\partial w_1}, \frac{\partial\sigma}{\partial w_2}, \frac{\partial\sigma}{\partial b}, \frac{\partial\sigma}{\partial x_1}, \frac{\partial\sigma}{\partial x_2}$$

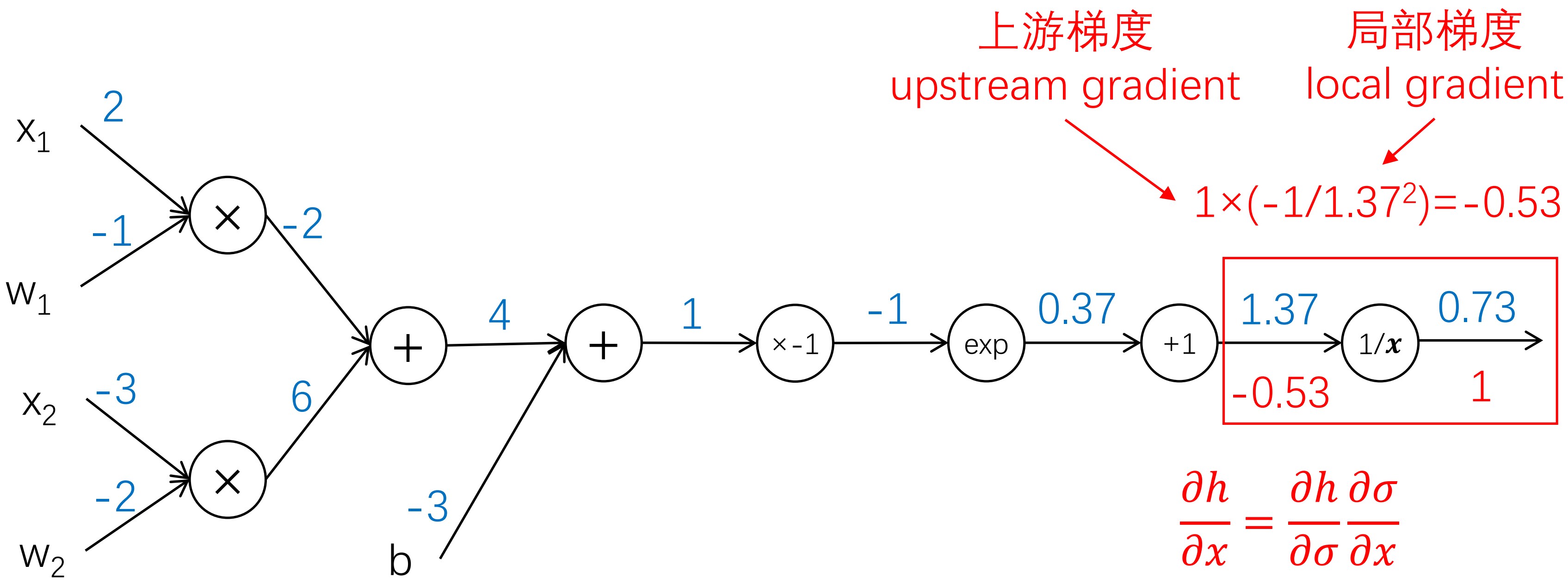
$$f(x) = e^x \longrightarrow \frac{df}{dx} = e^x \qquad f(x) = \frac{1}{x} \longrightarrow \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \longrightarrow \frac{df}{dx} = a \qquad f_c(x) = c + x \longrightarrow \frac{df}{dx} = 1$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial x}$$

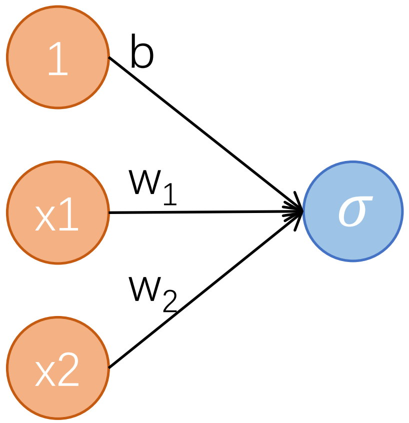
$$\frac{\partial \sigma}{\partial w_1}, \frac{\partial \sigma}{\partial w_2}, \frac{\partial \sigma}{\partial b}, \frac{\partial \sigma}{\partial x_1}, \frac{\partial \sigma}{\partial x_2}$$

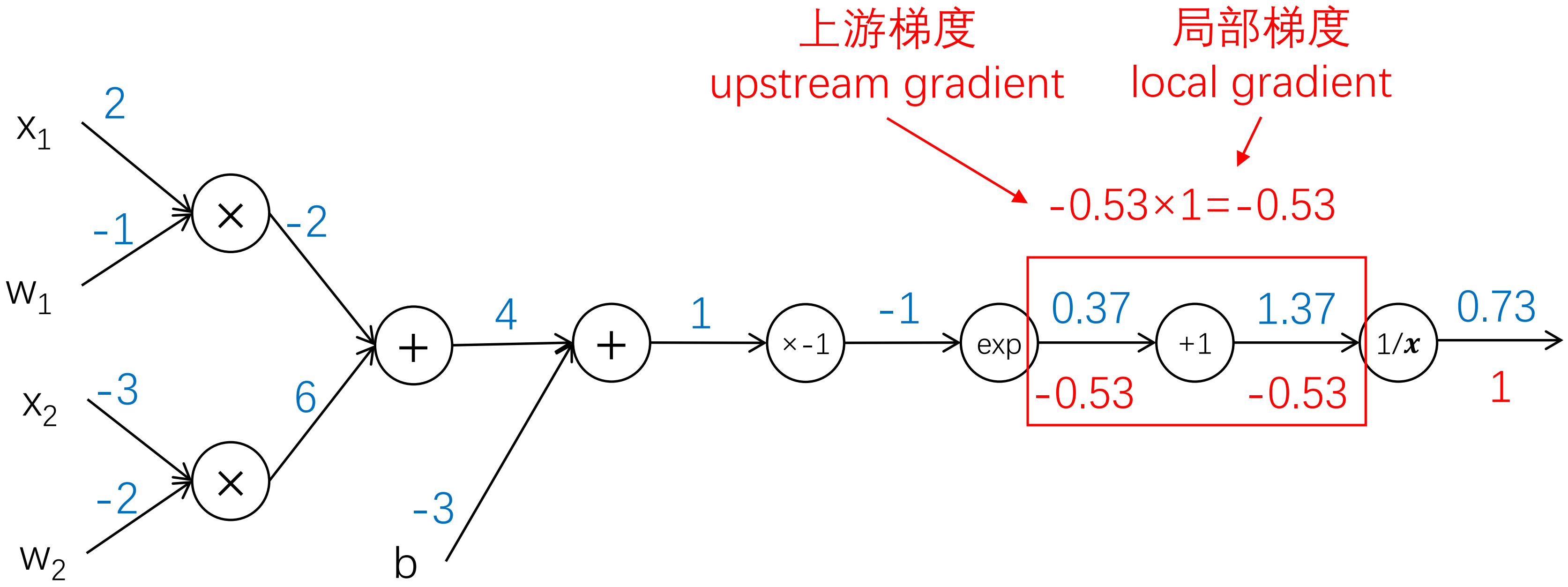
$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad \boxed{f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2}$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





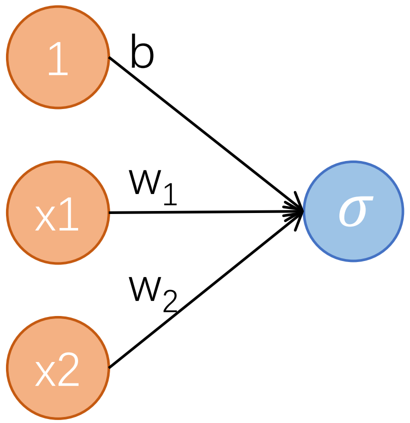
$$\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}$$

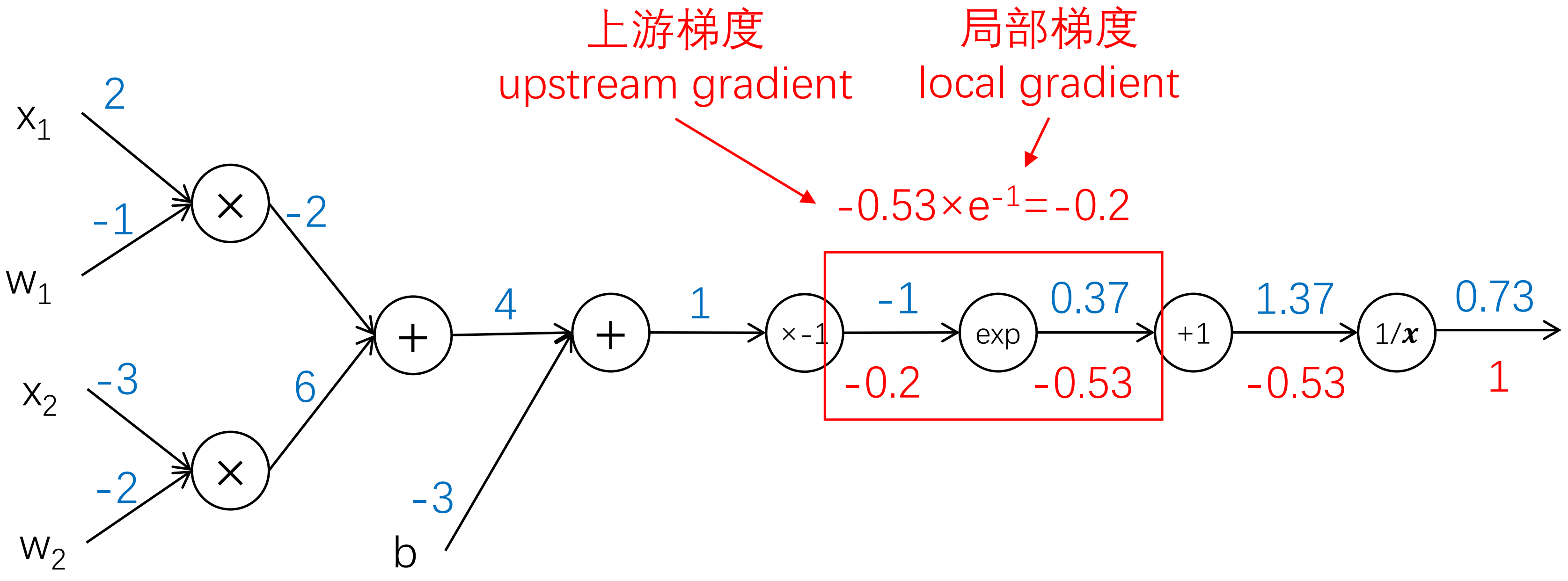
$$f(x)=e^x \rightarrow \frac{df}{dx}=e^x \qquad f(x)=\frac{1}{x} \rightarrow \frac{df}{dx}=-1/x^2$$

$$f_a(x)=ax \rightarrow \frac{df}{dx}=a \qquad \boxed{f_c(x)=c+x \rightarrow \frac{df}{dx}=1}$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W}, \boldsymbol{x}) = \frac{1}{1 + e^{-(w_1x_1 + w_2x_2 + b)}}$$





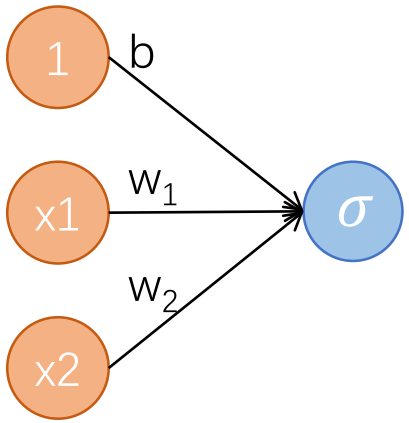
$$\frac{\partial\sigma}{\partial w_1}, \frac{\partial\sigma}{\partial w_2}, \frac{\partial\sigma}{\partial b}, \frac{\partial\sigma}{\partial x_1}, \frac{\partial\sigma}{\partial x_2}$$

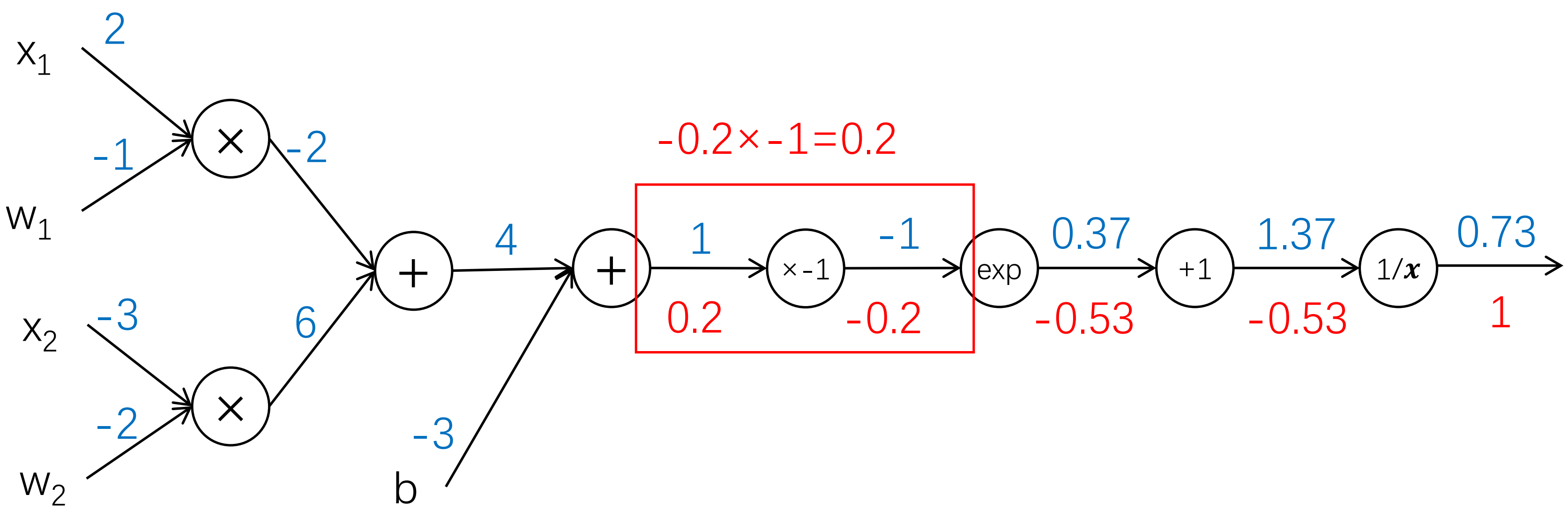
$$\boxed{f(x) = e^x \quad \longrightarrow \quad \frac{df}{dx} = e^x} \qquad f(x) = \frac{1}{x} \quad \longrightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \longrightarrow \quad \frac{df}{dx} = a \qquad f_c(x) = c + x \quad \longrightarrow \quad \frac{df}{dx} = 1$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





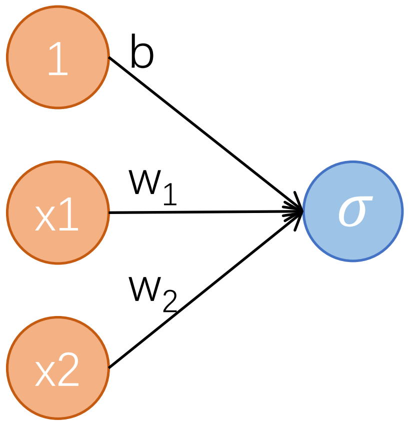
$$\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}$$

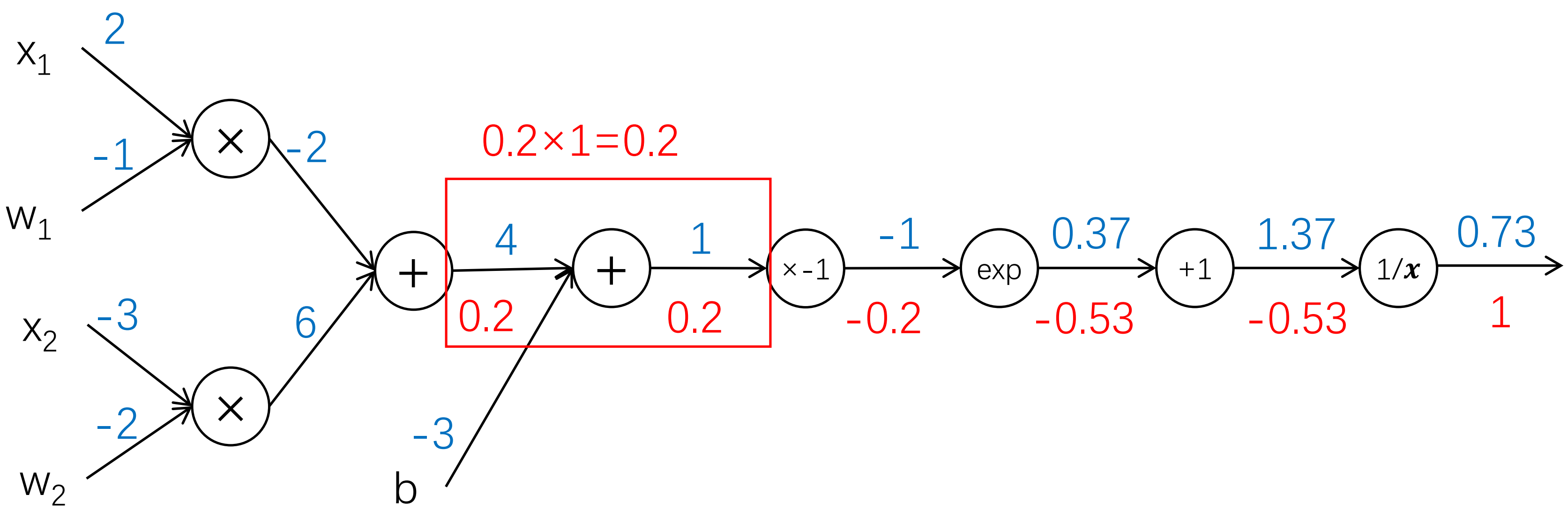
$$f(x)=e^x \quad\rightarrow\quad \frac{df}{dx}=e^x \qquad\qquad f(x)=\frac{1}{x} \quad\rightarrow\quad \frac{df}{dx}=-1/x^2$$

$$f_a(x)=ax \quad\rightarrow\quad \frac{df}{dx}=a \qquad\qquad f_c(x)=c+x \quad\rightarrow\quad \frac{df}{dx}=1$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





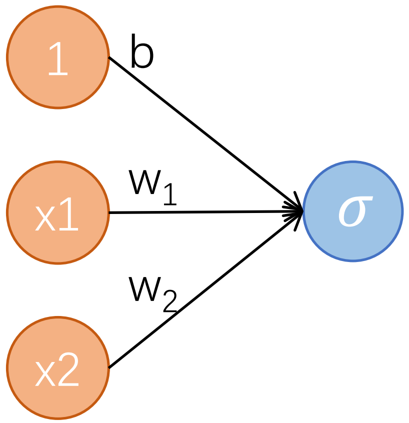
$$\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}$$

$$f(x)=e^x \quad \rightarrow \quad \frac{df}{dx}=e^x \qquad f(x)=\frac{1}{x} \quad \rightarrow \quad \frac{df}{dx}=-1/x^2$$

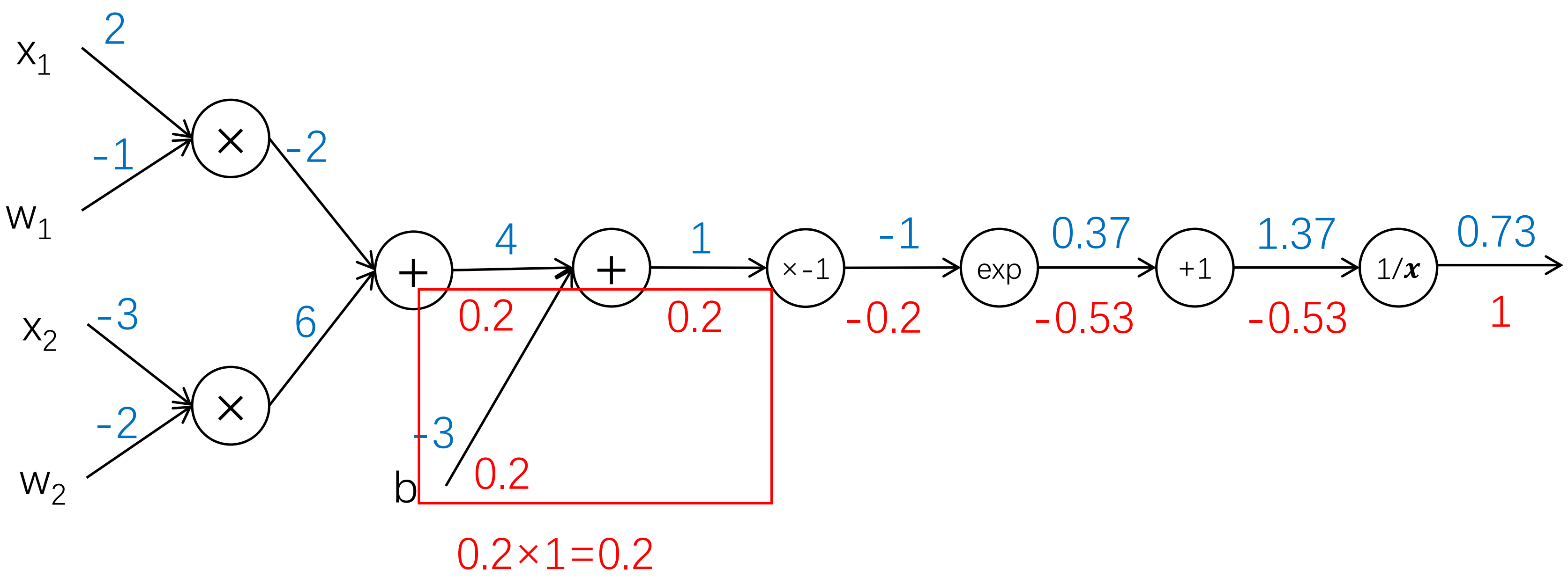
$$f_a(x)=ax \quad \rightarrow \quad \frac{df}{dx}=a \qquad \boxed{f_c(x)=c+x \quad \rightarrow \quad \frac{df}{dx}=1}$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$



$$\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}$$

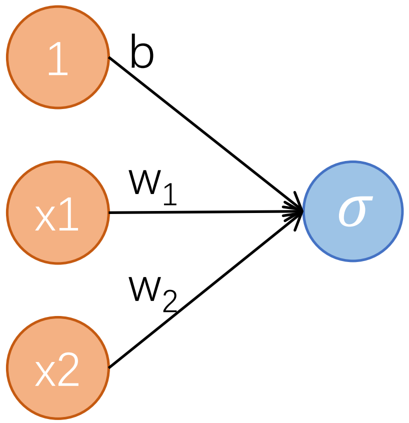


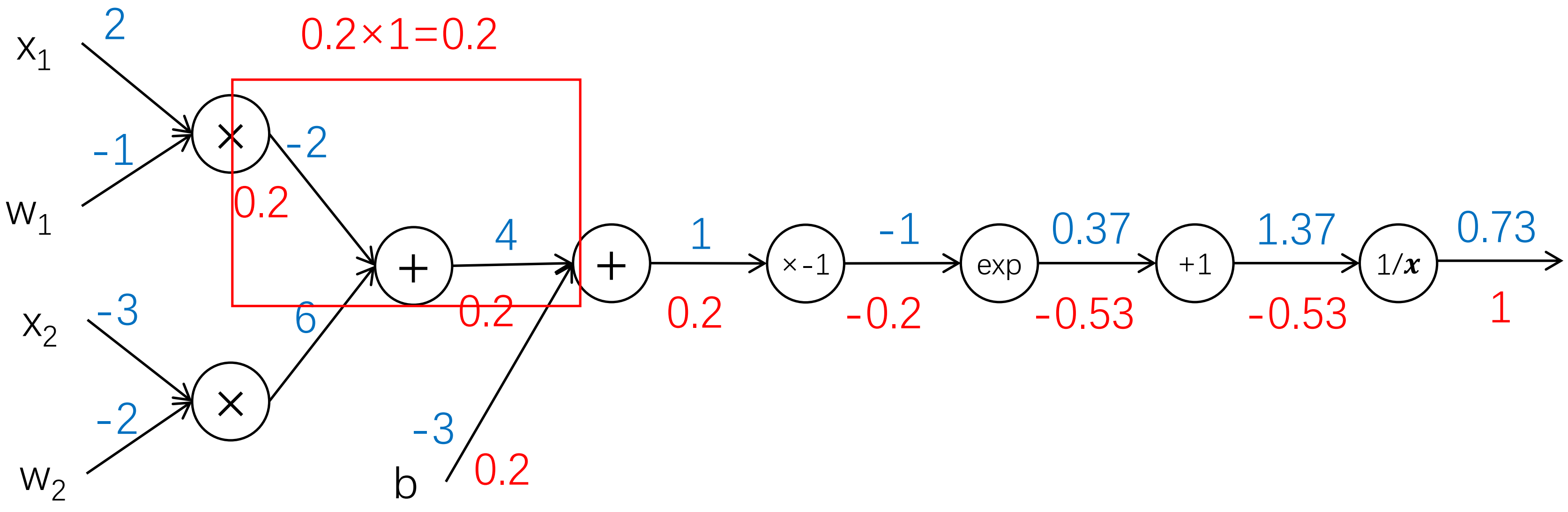
$$f(x)=e^x \rightarrow \frac{df}{dx}=e^x \qquad f(x)=\frac{1}{x} \rightarrow \frac{df}{dx}=-1/x^2$$

$$f_a(x)=ax \rightarrow \frac{df}{dx}=a \qquad \boxed{f_c(x)=c+x \rightarrow \frac{df}{dx}=1}$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





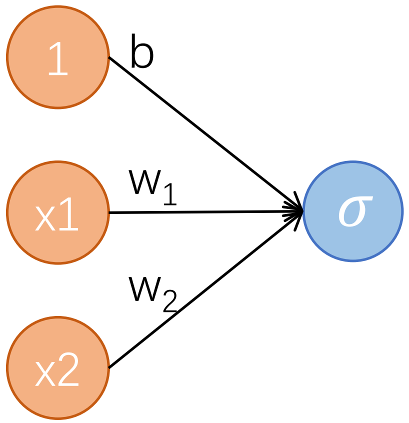
$$\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}$$

$$f(x)=e^x \longrightarrow \frac{df}{dx}=e^x \qquad f(x)=\frac{1}{x} \longrightarrow \frac{df}{dx}=-1/x^2$$

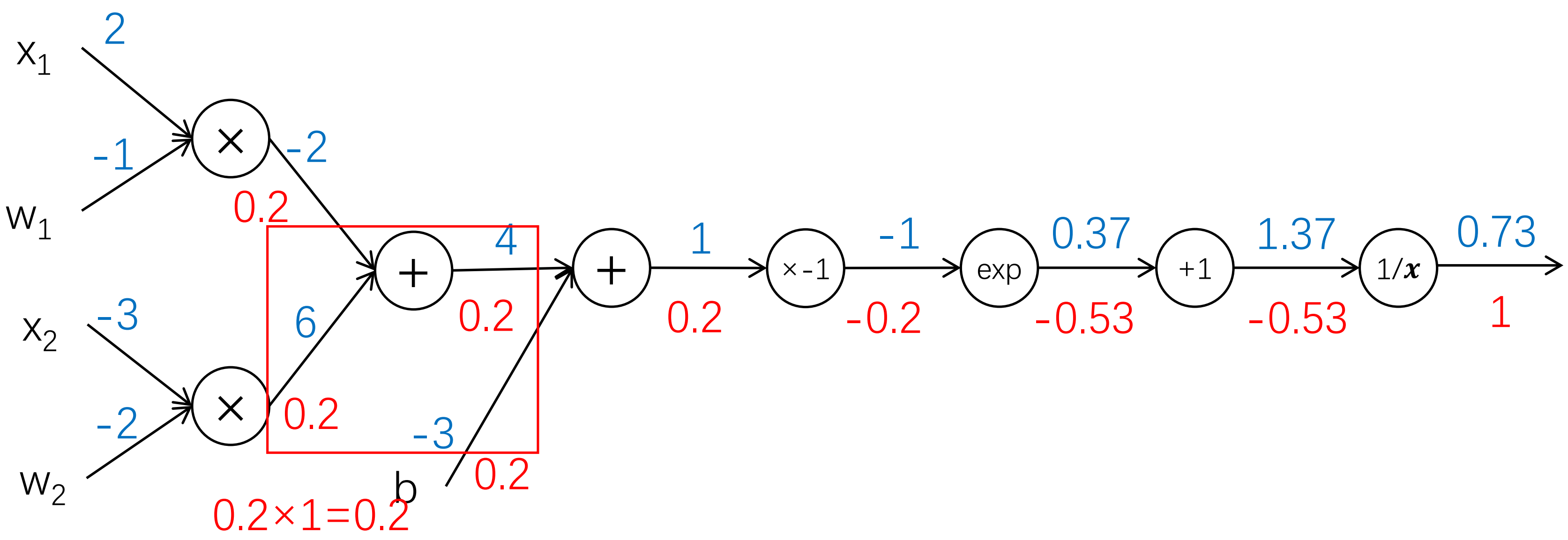
$$f_a(x)=ax \longrightarrow \frac{df}{dx}=a \qquad \boxed{f_c(x)=c+x \longrightarrow \frac{df}{dx}=1}$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$



$$\frac{\partial\sigma}{\partial w_1}, \frac{\partial\sigma}{\partial w_2}, \frac{\partial\sigma}{\partial b}, \frac{\partial\sigma}{\partial x_1}, \frac{\partial\sigma}{\partial x_2}$$

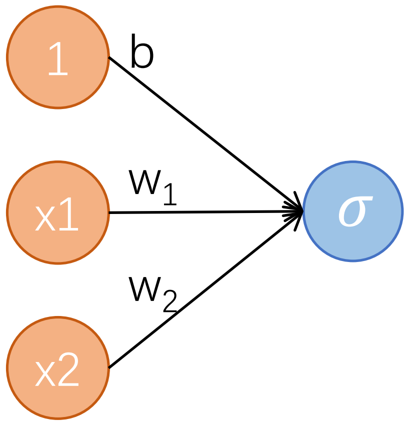


$$f(x) = e^x \longrightarrow \frac{df}{dx} = e^x \qquad f(x) = \frac{1}{x} \longrightarrow \frac{df}{dx} = -1/x^2$$

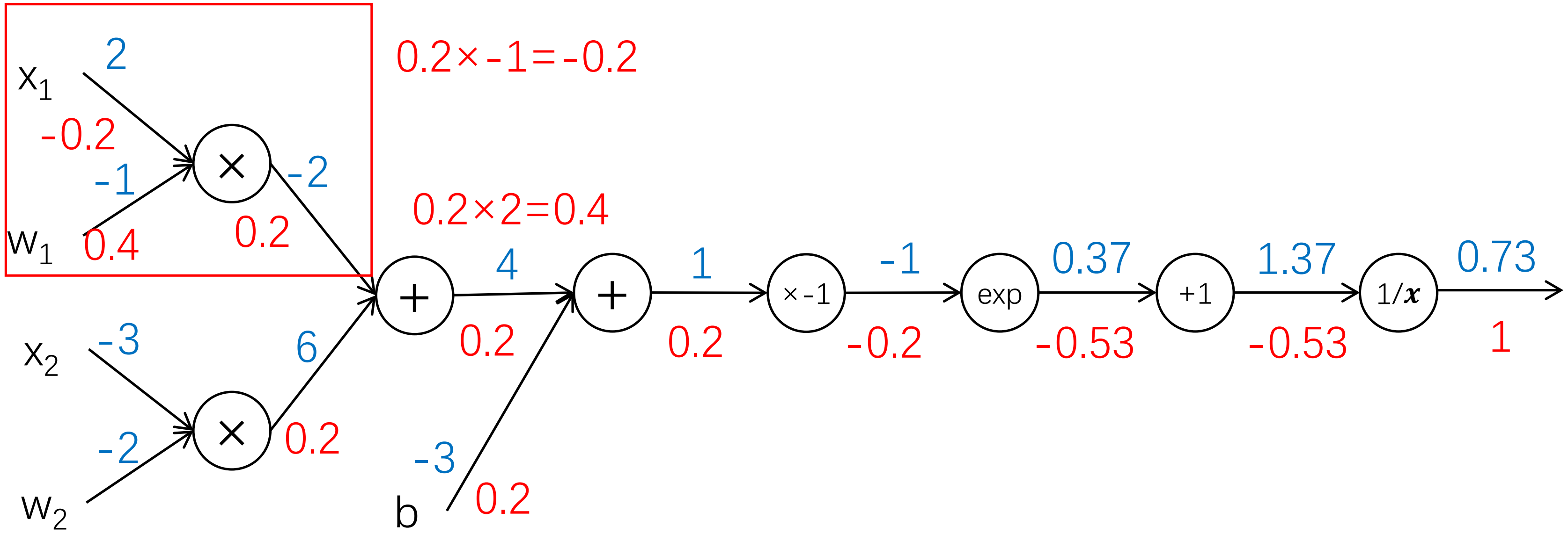
$$f_a(x) = ax \longrightarrow \frac{df}{dx} = a \qquad \boxed{f_c(x) = c + x \longrightarrow \frac{df}{dx} = 1}$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$



$$\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}$$

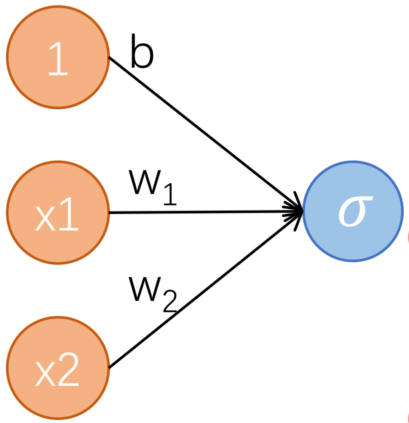


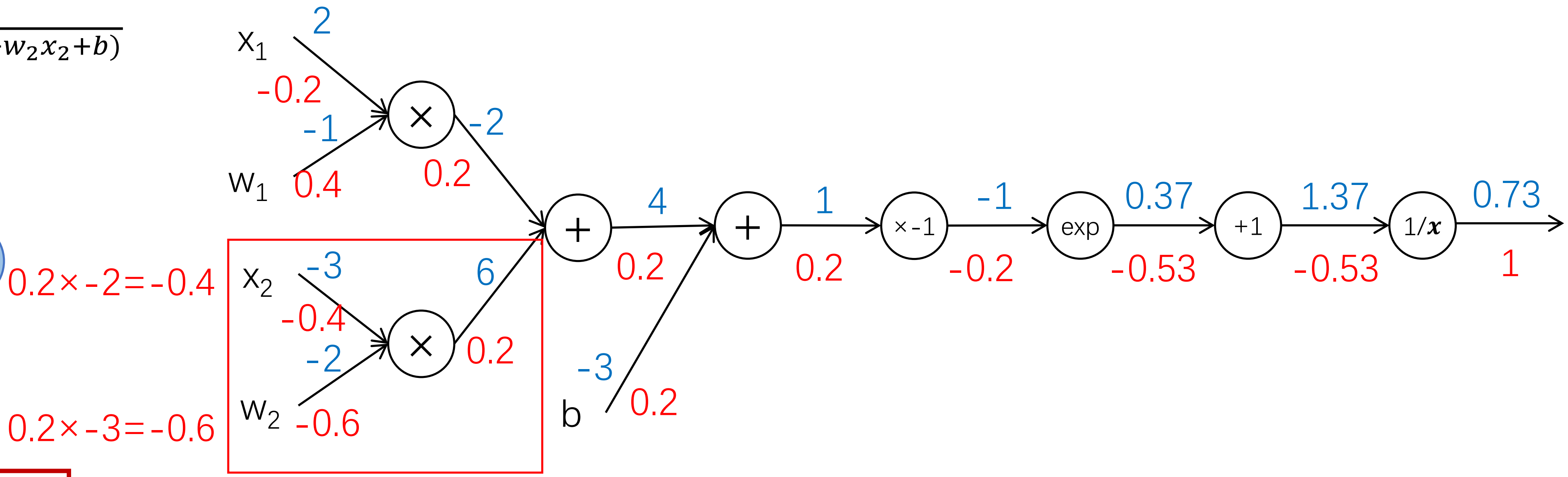
$f(x)=e^x \longrightarrow \frac{df}{dx}=e^x$ $\quad f(x)=\frac{1}{x} \longrightarrow \frac{df}{dx}=-1/x^2$

$f_a(x)=ax \longrightarrow \frac{df}{dx}=a$ $\quad f_c(x)=c+x \longrightarrow \frac{df}{dx}=1$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





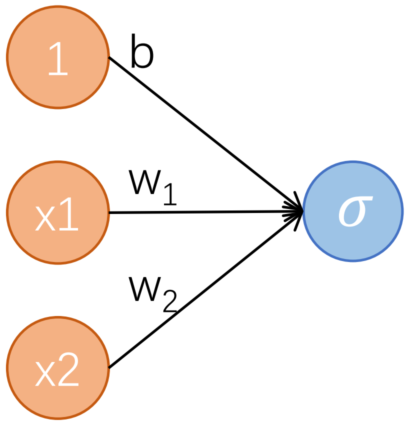
$$\boxed{\frac{\partial\sigma}{\partial w_1}, \frac{\partial\sigma}{\partial w_2}, \frac{\partial\sigma}{\partial b}, \frac{\partial\sigma}{\partial x_1}, \frac{\partial\sigma}{\partial x_2}}$$

$f(x) = e^x \longrightarrow \frac{df}{dx} = e^x$ $\qquad$ $f(x) = \frac{1}{x} \longrightarrow \frac{df}{dx} = -1/x^2$

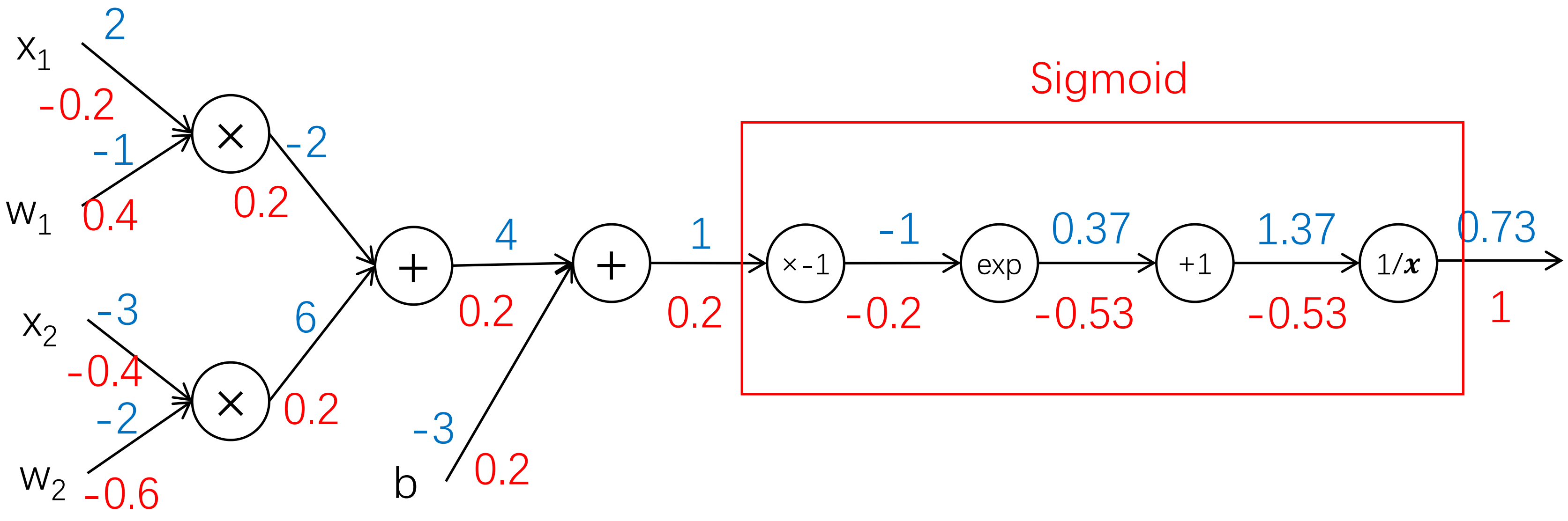
$\boxed{f_a(x) = ax \longrightarrow \frac{df}{dx} = a}$ $\qquad$ $f_c(x) = c + x \longrightarrow \frac{df}{dx} = 1$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$



$$\boxed{\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}}$$

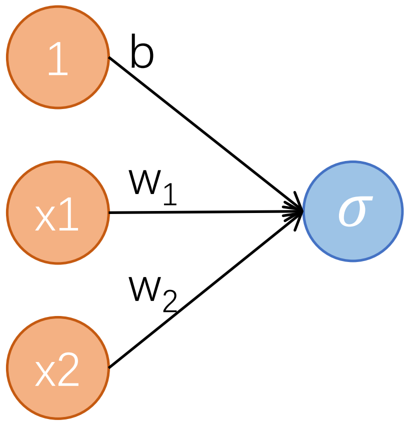


Sigmoid本地梯度

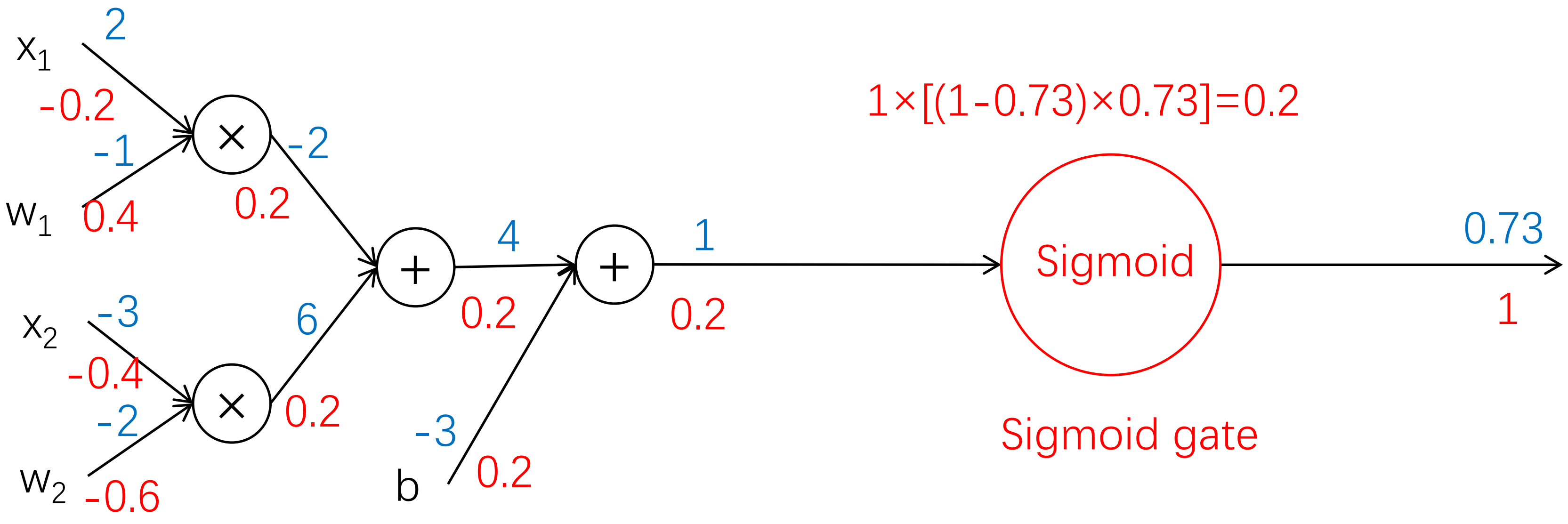
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\sigma(x)$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$



$$\boxed{\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}}$$

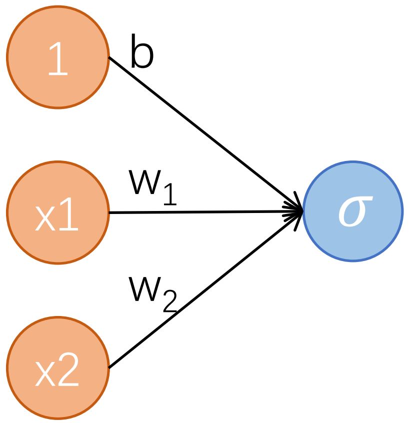


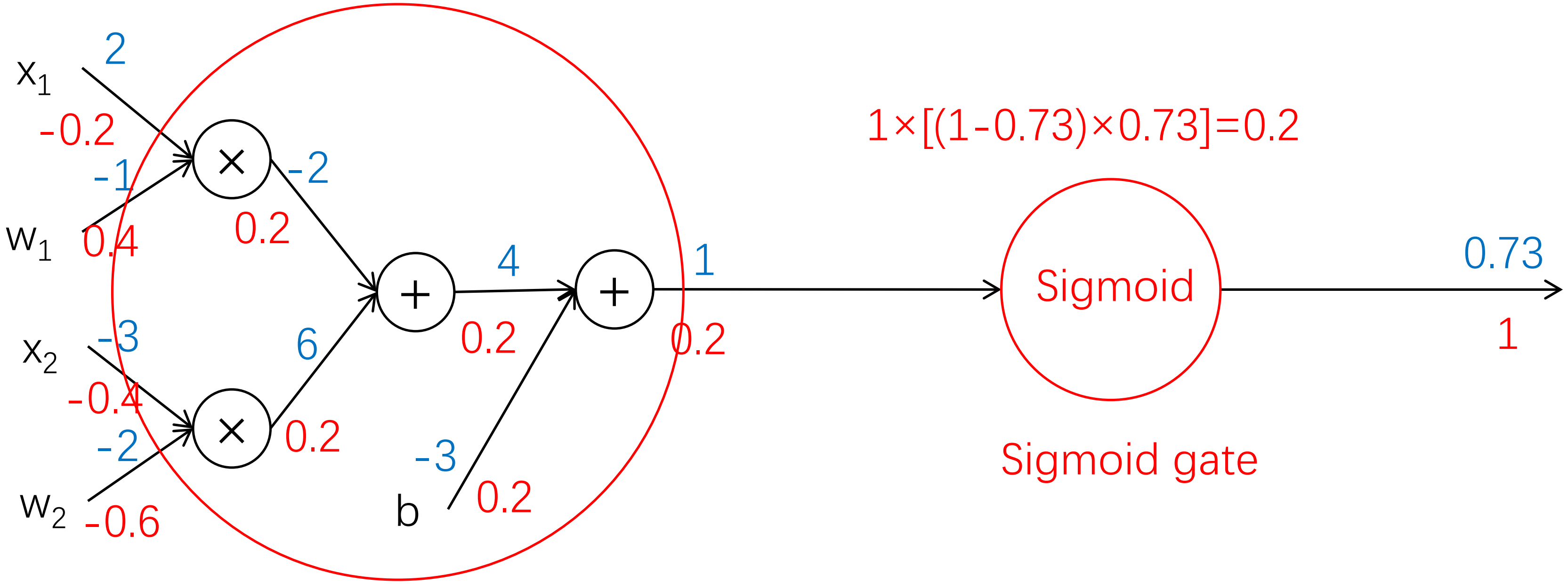
Sigmoid本地梯度

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\sigma(x)$$

反向传播（Backpropagation）

$$\sigma(\boldsymbol{W},\boldsymbol{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+b)}}$$





$$\boxed{\frac{\partial\sigma}{\partial w_1},\frac{\partial\sigma}{\partial w_2},\frac{\partial\sigma}{\partial b},\frac{\partial\sigma}{\partial x_1},\frac{\partial\sigma}{\partial x_2}}$$

$$\boxed{\frac{\partial\sigma}{\partial w_i} = \frac{\partial\sigma}{\partial y}\frac{\partial y}{\partial w_i} \qquad \frac{\partial\sigma}{\partial x_i} = \frac{\partial\sigma}{\partial y}\frac{\partial y}{\partial x_i}}$$

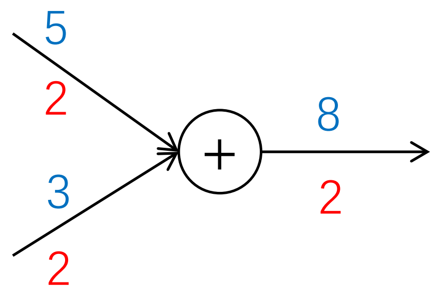
$$y = w_1x_1 + w_2x_2 + b$$

Sigmoid本地梯度

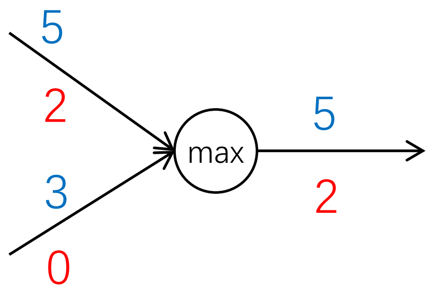
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\sigma(x)$$

梯度流的常见模式

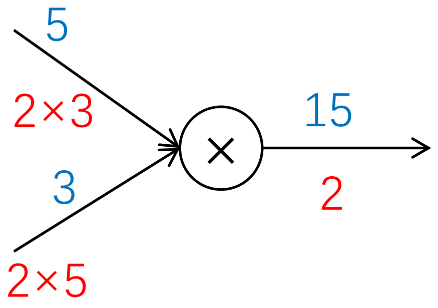
add gate：拷贝上游梯度



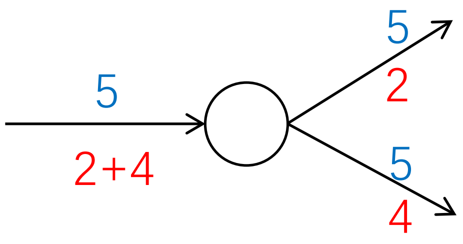
max gate：上游梯度路由给较大变量



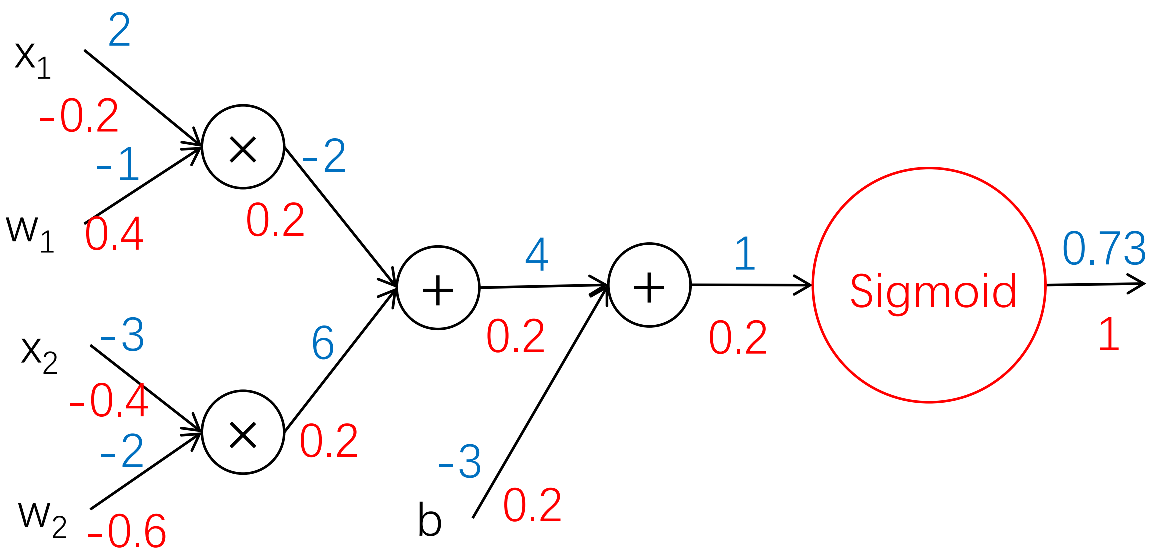
mul gate：上游梯度×互换变量



copy gate：上游梯度相加



代码实现



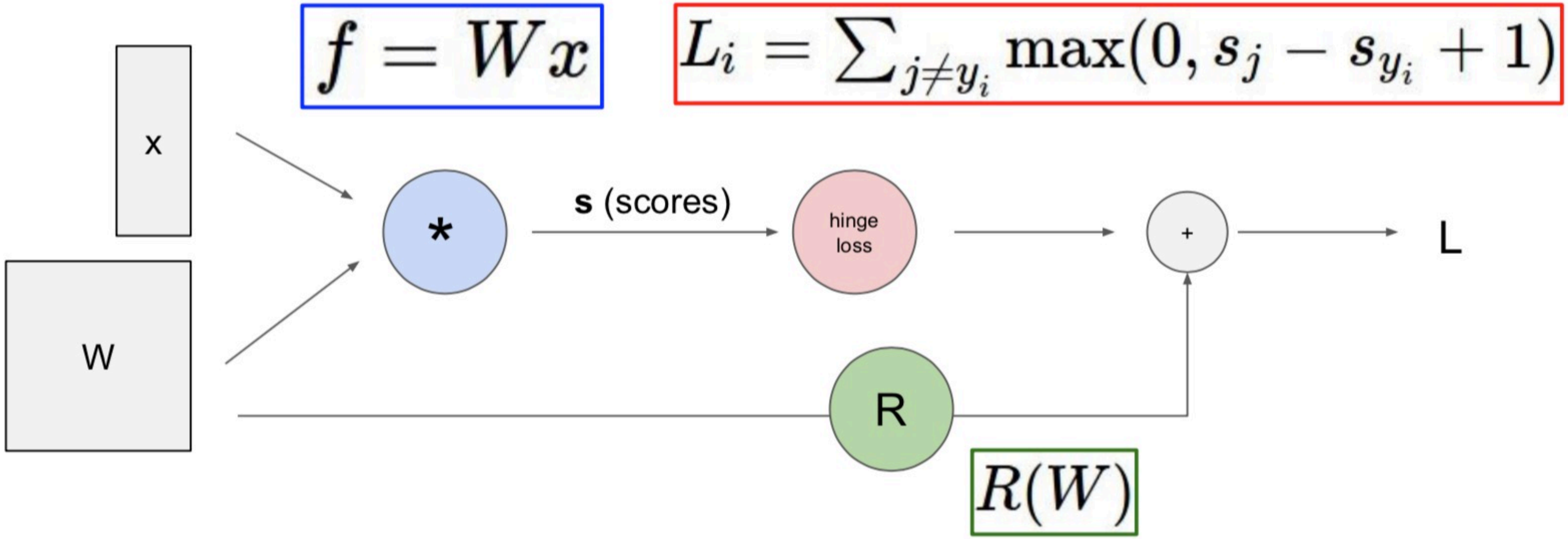

```
def f(w1,x1,w2,x2,b)
    s1 = w1 * x1
    s2 = w2 * x2
    s3 = s1 + s2
    s4 = s3 + b
    L = sigmoid(s4)


    grad_L = 1.0
    grad_s4 = grad_L * (1-L) * L
    grad_b = grad_s4
    grad_s3 = grad_s4
    grad_s1 = grad_s3
    grad_s2 = grad_s3
    grad_w1 = grad_s1 * x1
    grad_x1 = grad_s1 * w1
    grad_w2 = grad_s2 * x2
    grad_x2 = grad_s12 * w2
```

前向传递，计算输出

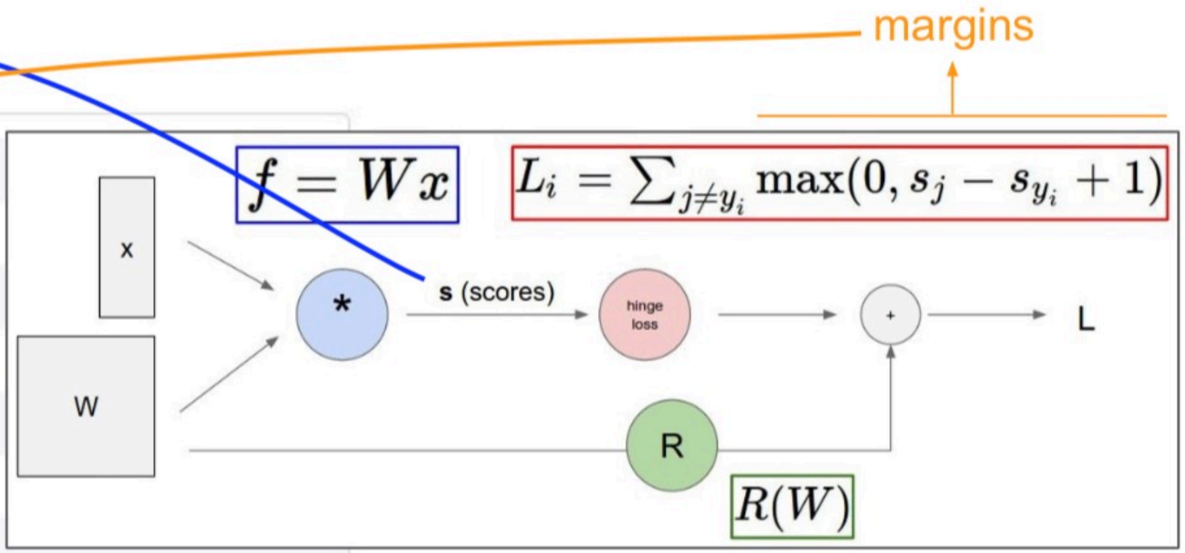
反向传递，计算梯度

SVM分类器计算图+反向传播

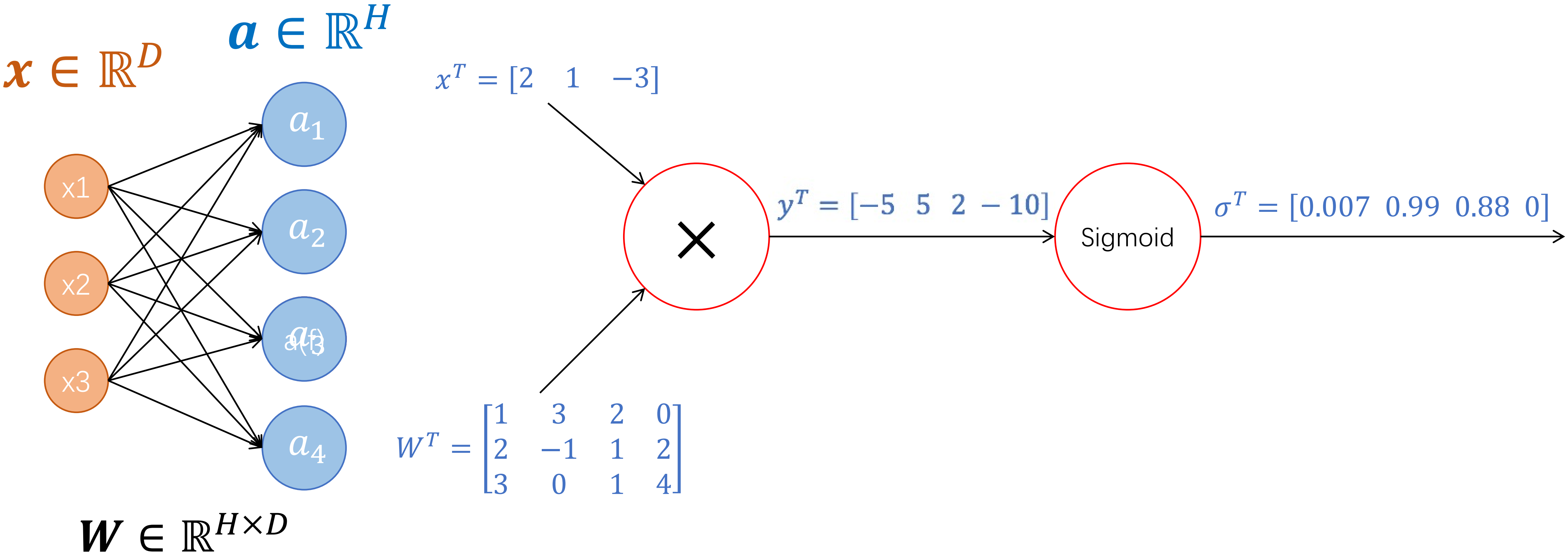


SVM分类器计算图+反向传播实现

```
# receive W (weights), X (data)
# forward pass (we have 8 lines)
scores = #...
margins = #...
data_loss = #...
reg_loss = #...
loss = data_loss + reg_loss
# backward pass (we have 5 lines)
dmargins = # ... (optionally, we go direct to dscores)
dscores = #...
dW = #...
```



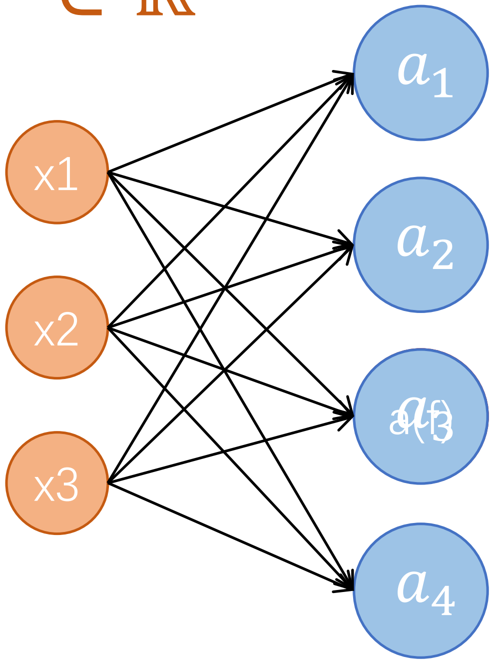
反向传播的矩阵运算



$\boldsymbol{x} \in \mathbb{R}^D$
$\boldsymbol{a} \in \mathbb{R}^H$
x1
x2
x3
a_1
a_2
a_3
a_4
$\boldsymbol{W} \in \mathbb{R}^{H \times D}$
$x^T = [2 \quad 1 \quad -3]$
$W^T = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$
$\times$
$y^T = [-5 \quad 5 \quad 2 \quad -10]$
Sigmoid
$\sigma^T = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

反向传播的矩阵运算

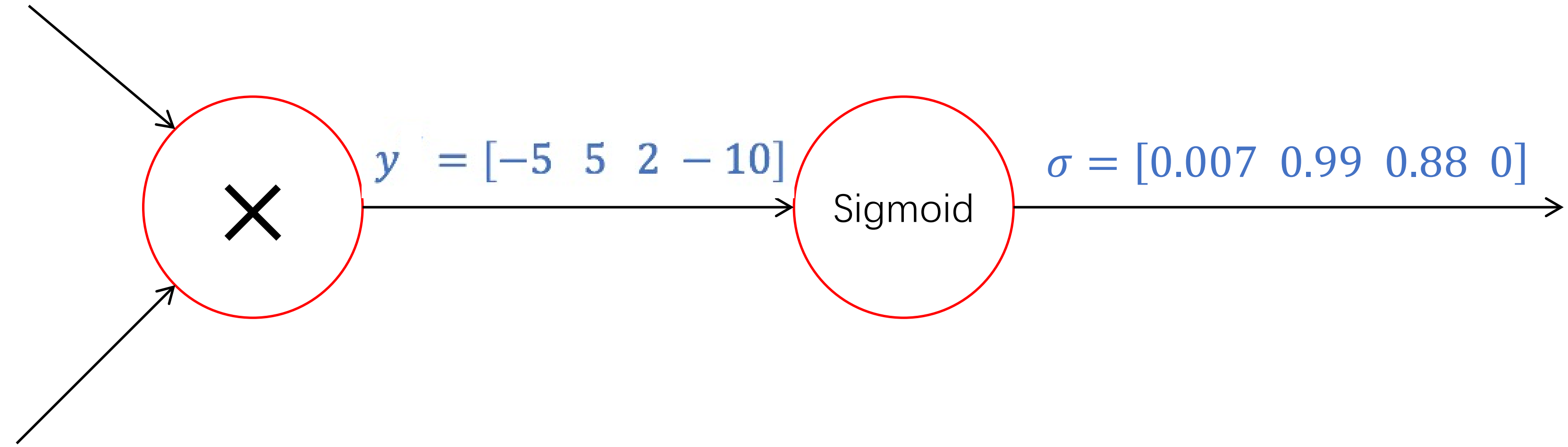
$\boldsymbol{x}^T \in \mathbb{R}^D$ $\boldsymbol{a}^T \in \mathbb{R}^H$



$\boldsymbol{W}^T \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$



$y = [-5 \quad 5 \quad 2 \quad -10]$

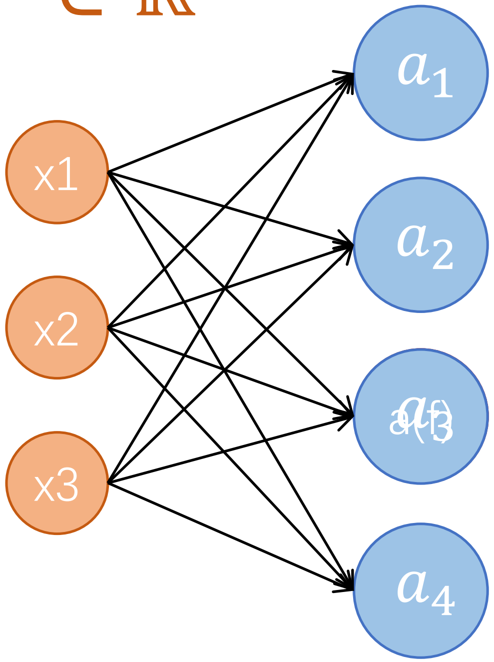
$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

反向传播的矩阵运算

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

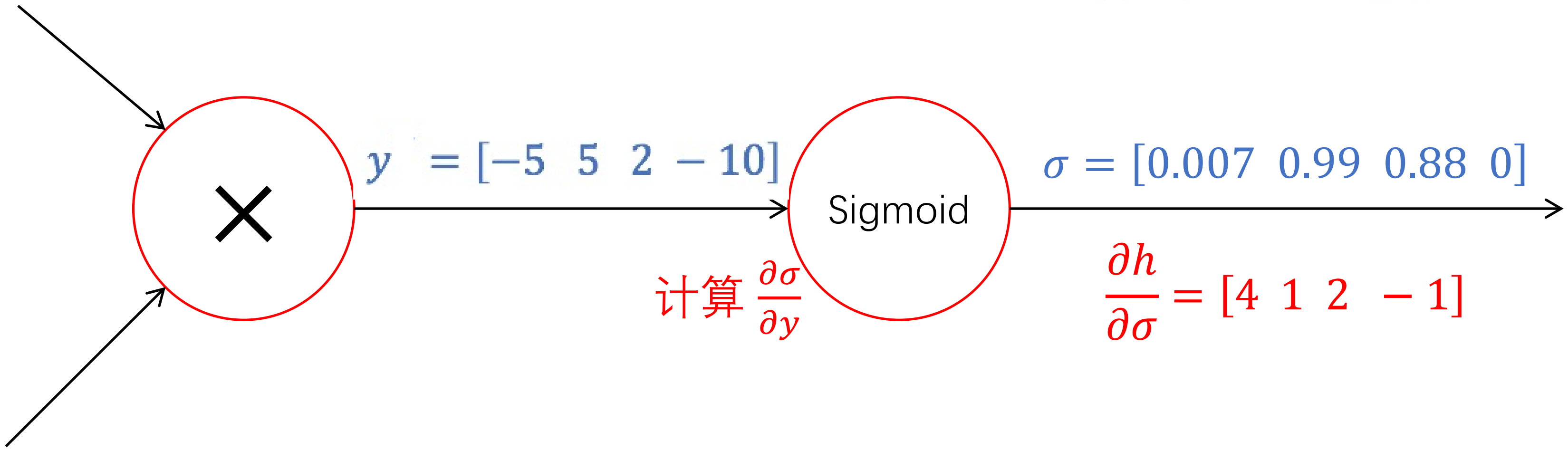
$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$



$y = [-5 \quad 5 \quad 2 \quad -10]$

$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

计算 $\dfrac{\partial \sigma}{\partial y}$

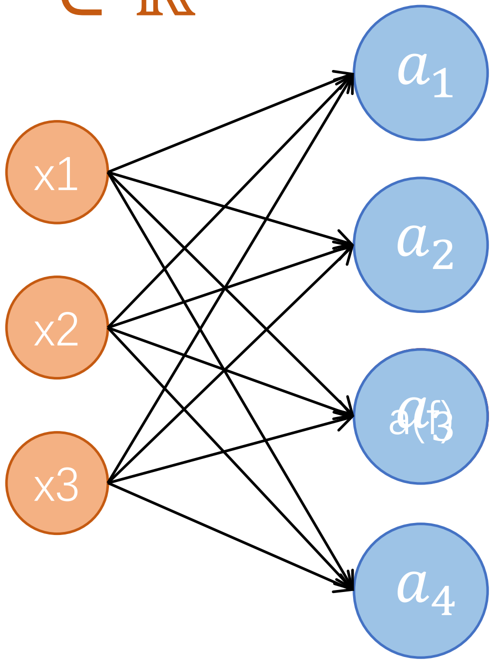
$\dfrac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$

反向传播的矩阵运算

雅可比矩阵 (Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^T \in \mathbb{R}^D$ $\boldsymbol{a}^T \in \mathbb{R}^H$



$\boldsymbol{W}^T \in \mathbb{R}^{H\times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$\times$

$y = [-5 \quad 5 \quad 2 \quad -10]$

Sigmoid

$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

计算 $\frac{\partial \sigma}{\partial y}$

$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$

$$\left(\frac{\partial \sigma}{\partial y}\right)_{n,m} = \left(\frac{\partial \sigma_n}{\partial y_m}\right) \longrightarrow \frac{\partial \sigma}{\partial y} \in \mathbb{R}^{H\times H}$$

假设H=4096，需要至少16777216≈16M内存

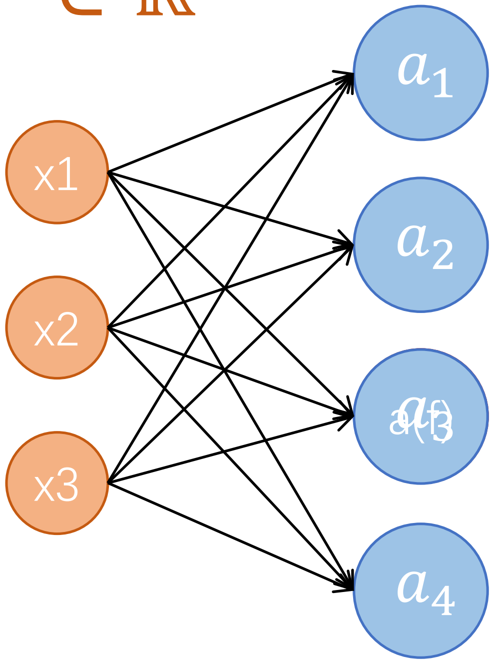
实际上我们不需要显示实现Jacobian Matrix

反向传播的矩阵运算

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

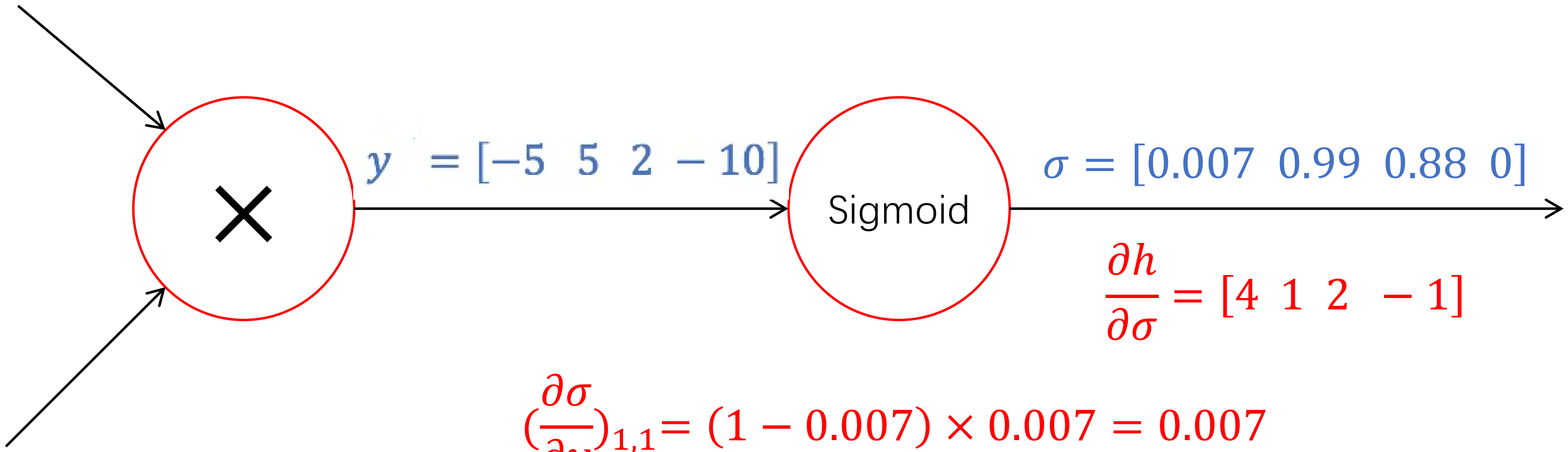
$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$



$y = [-5 \quad 5 \quad 2 \quad -10]$

$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

$$\left(\frac{\partial \sigma}{\partial y}\right)_{1,1} = (1 - 0.007) \times 0.007 = 0.007$$

$$\left(\frac{\partial \sigma}{\partial y}\right)_{1,2} = 0 \qquad \left(\frac{\partial \sigma}{\partial y}\right)_{1,3} = 0 \qquad \left(\frac{\partial \sigma}{\partial y}\right)_{1,4} = 0$$

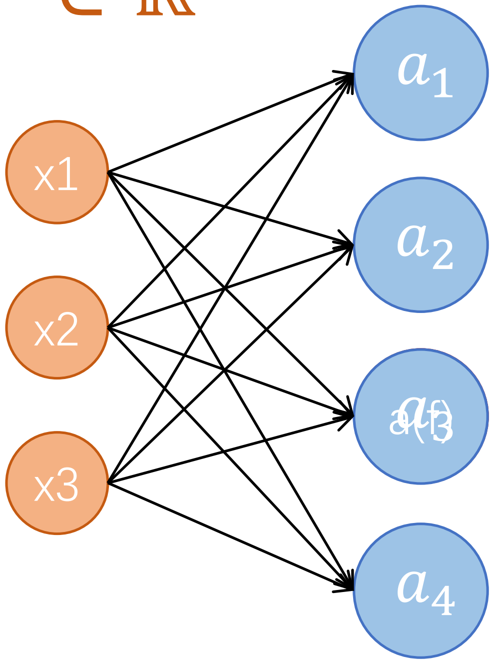
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\sigma(x)$$

反向传播的矩阵运算

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

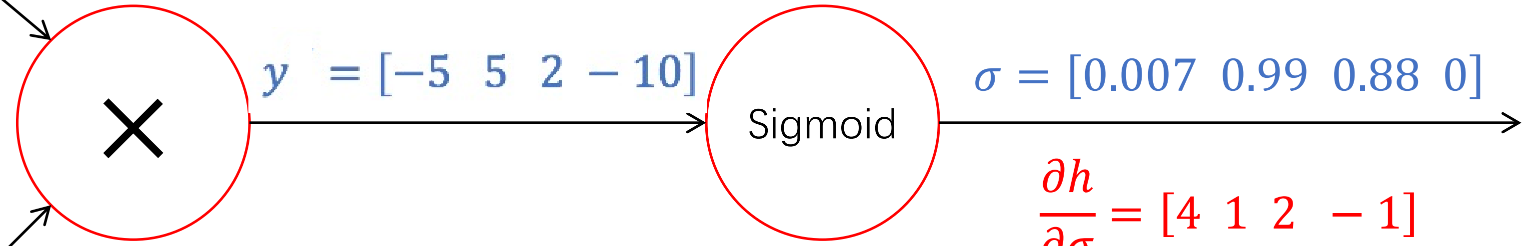
$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\quad \boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$



$y = [-5 \;\; 5 \;\; 2 \;\; -10]$

$\sigma = [0.007 \;\; 0.99 \;\; 0.88 \;\; 0]$

$$\frac{\partial h}{\partial \sigma} = [4 \;\; 1 \;\; 2 \;\; -1]$$

$$\left(\frac{\partial \sigma}{\partial y}\right)_{1,1} = (1 - 0.007) \times 0.007 = 0.007$$

$\left(\frac{\partial \sigma}{\partial y}\right)_{1,2} = 0$ $\quad$ y_i只影响σ_i $\quad$ $\left(\frac{\partial \sigma}{\partial y}\right)_{1,4} = 0$

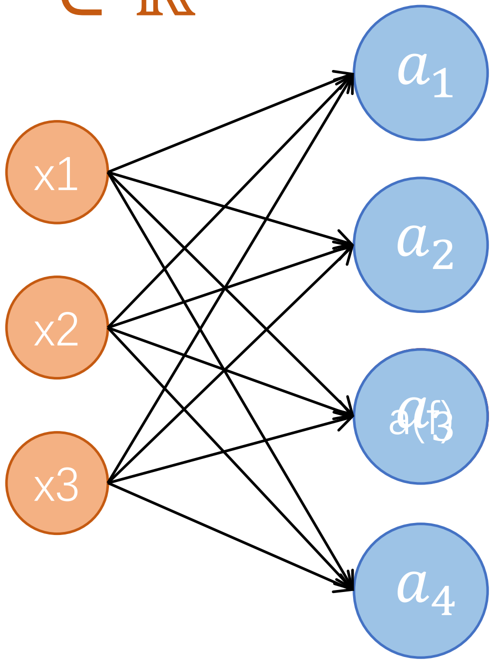
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\sigma(x)$$

反向传播的矩阵运算

雅可比矩阵 (Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^{D}$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^{H}$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

×

$y = [-5 \quad 5 \quad 2 \quad -10]$

Sigmoid

$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

$$\frac{\partial \sigma}{\partial y} = \begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

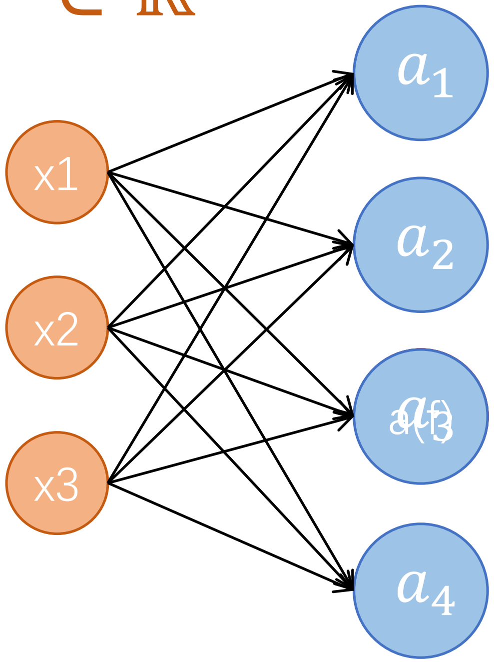
$$\frac{\partial h}{\partial y_i} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y_i} = \sum_j \frac{\partial h}{\partial \sigma_j}\frac{\partial \sigma_j}{\partial y_i}$$

反向传播的矩阵运算

雅可比矩阵 (Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\qquad$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$

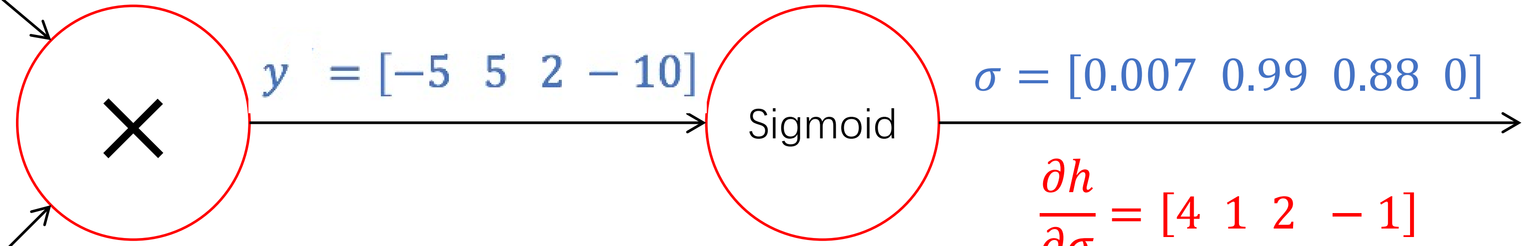


$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$y = [-5 \quad 5 \quad 2 \quad -10]$



$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

$$\frac{\partial \sigma}{\partial y} = \begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial h}{\partial y_i} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y_i} = \sum_j \frac{\partial h}{\partial \sigma_j}\frac{\partial \sigma_j}{\partial y_i}$$

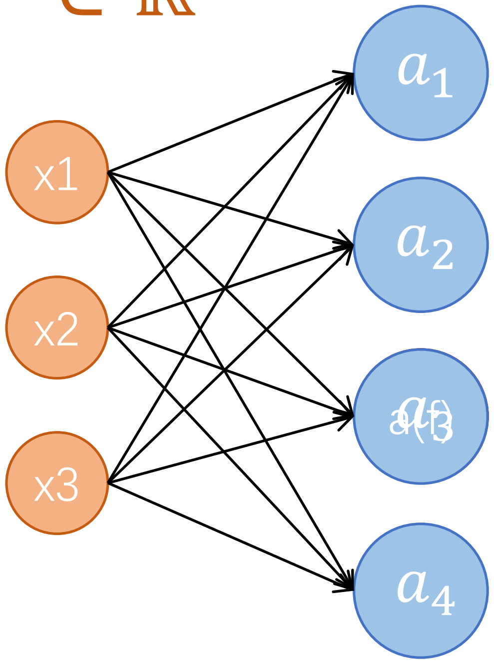
$$\frac{\partial h}{\partial y} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y} = [4 \quad 1 \quad 2 \quad -1]\begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

反向传播的矩阵运算

雅可比矩阵 (Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\quad$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$\times$ $\quad y = [-5 \quad 5 \quad 2 \quad -10]$ $\quad$ Sigmoid $\quad \sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

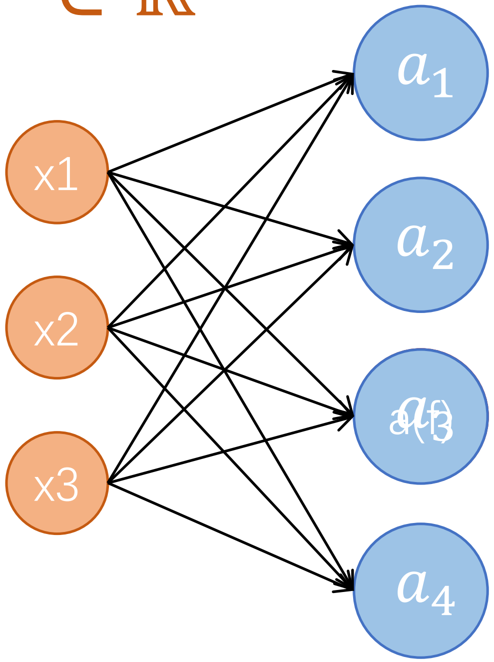
$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0] \qquad \frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

$$\frac{\partial \sigma}{\partial y} = \begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \frac{\partial h}{\partial y_i} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y_i} = \sum_j \frac{\partial h}{\partial \sigma_j}\frac{\partial \sigma_j}{\partial y_i}$$

$$\frac{\partial h}{\partial y} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y} = [4 \quad 1 \quad 2 \quad -1] \begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

反向传播的矩阵运算

$x^T \in \mathbb{R}^D$ $\quad a^T \in \mathbb{R}^H$



$W^T \in \mathbb{R}^{H\times D}$

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$x = [2 \quad 1 \quad -3]$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$\times$

$y = [-5 \quad 5 \quad 2 \quad -10]$

Sigmoid

$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

$$\frac{\partial \sigma}{\partial y} = \begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial h}{\partial y_i} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y_i} = \sum_j \frac{\partial h}{\partial \sigma_j}\frac{\partial \sigma_j}{\partial y_i}$$

$$\frac{\partial h}{\partial y} = \frac{\partial h}{\partial \sigma}\frac{\partial \sigma}{\partial y} = [4 \quad 1 \quad 2 \quad -1] \begin{bmatrix} 0.007 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

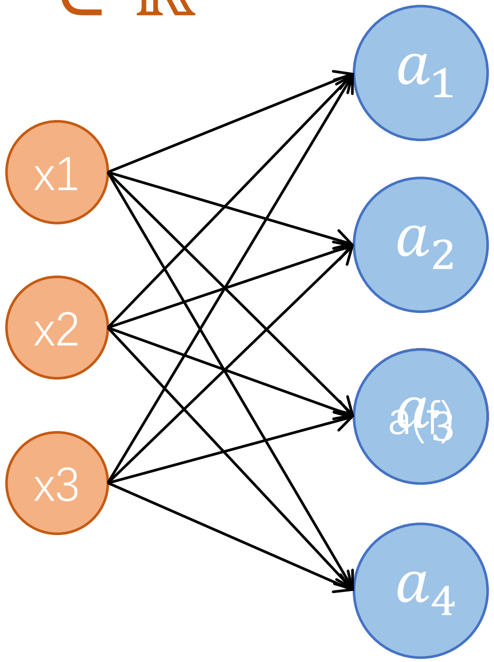
只需要进行pairwise的梯度计算

反向传播的矩阵运算

雅可比矩阵 (Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H\times D}$

$$x = [2 \quad 1 \quad -3]$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$$y^T = \begin{bmatrix} x_1w_{11} + x_2w_{21} + x_3w_{31} \\ x_1w_{12} + x_2w_{22} + x_3w_{32} \\ x_1w_{13} + x_2w_{23} + x_3w_{33} \\ x_1w_{14} + x_2w_{24} + x_3w_{34} \end{bmatrix}$$

×

$$y = [-5 \quad 5 \quad 2 \quad -10]$$

$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

Sigmoid

$$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$$

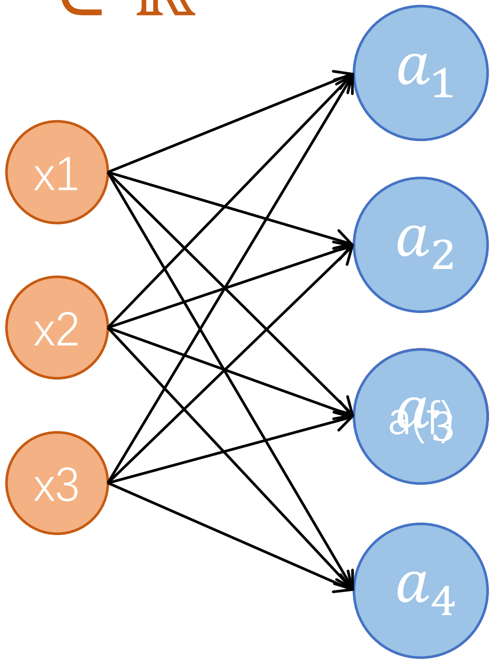
$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

反向传播的矩阵运算

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\quad \boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H\times D}$

$$x = \begin{bmatrix} 2 & 1 & -3 \end{bmatrix}$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$\times$

$$y^T = \begin{bmatrix} x_1w_{11} + x_2w_{21} + x_3w_{31} \\ x_1w_{12} + x_2w_{22} + x_3w_{32} \\ x_1w_{13} + x_2w_{23} + x_3w_{33} \\ x_1w_{14} + x_2w_{24} + x_3w_{34} \end{bmatrix}$$

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \end{bmatrix}$$

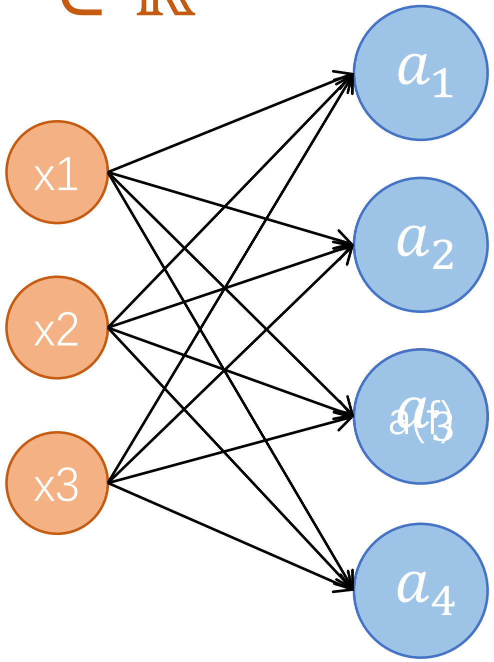
$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \end{bmatrix}$$

$$\left(\frac{\partial y}{\partial x}\right)_{n,m} = \left(\frac{\partial y_n}{\partial x_m}\right) \longrightarrow \frac{\partial y}{\partial x} \in \mathbb{R}^{H\times D}$$

和$\boldsymbol{W}^{\boldsymbol{T}}$形状一样!

反向传播的矩阵运算

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^{D}$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^{H}$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H\times D}$

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$$x = [2 \quad 1 \quad -3]$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$\times$

$$y^T = \begin{bmatrix} x_1w_{11} + x_2w_{21} + x_3w_{31} \\ x_1w_{12} + x_2w_{22} + x_3w_{32} \\ x_1w_{13} + x_2w_{23} + x_3w_{33} \\ x_1w_{14} + x_2w_{24} + x_3w_{34} \end{bmatrix}$$

$$\boldsymbol{y} = [-5 \quad 5 \quad 2 \quad -10]$$

Sigmoid

$$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$$

$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

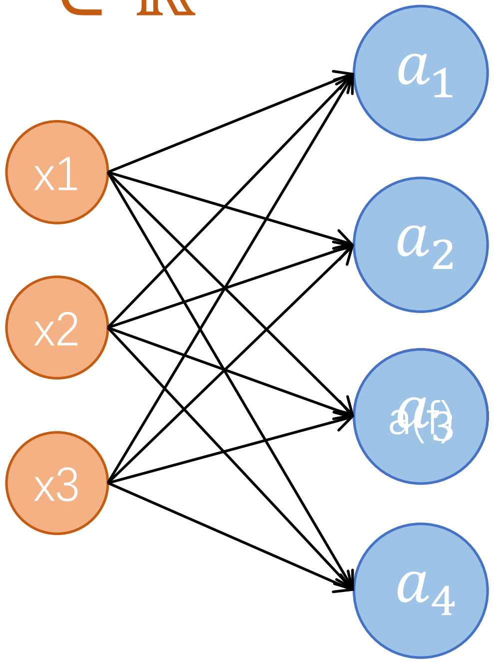
$$\left(\frac{\partial y}{\partial x}\right)_{n,m} = \left(\frac{\partial y_n}{\partial x_m}\right) \longrightarrow \frac{\partial y}{\partial x} \in \mathbb{R}^{H\times D}$$

和$\boldsymbol{W}^{\boldsymbol{T}}$形状一样!

$$\frac{\partial y}{\partial x} = \begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \\ w_{14} & w_{24} & w_{34} \end{bmatrix} = \boldsymbol{W}^{\boldsymbol{T}}\ !$$

反向传播的矩阵运算

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H \times D}$

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \left[\frac{\partial \mathbf{f}}{\partial x_1} \cdots \frac{\partial \mathbf{f}}{\partial x_n}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$$x = [2 \quad 1 \quad -3]$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$$y^T = \begin{bmatrix} x_1 w_{11} + x_2 w_{21} + x_3 w_{31} \\ x_1 w_{12} + x_2 w_{22} + x_3 w_{32} \\ x_1 w_{13} + x_2 w_{23} + x_3 w_{33} \\ x_1 w_{14} + x_2 w_{24} + x_3 w_{34} \end{bmatrix}$$

$\times$

$$y = [-5 \quad 5 \quad 2 \quad -10]$$

Sigmoid

$$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$$

$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

$$\left(\frac{\partial y}{\partial x}\right)_{n,m} = \left(\frac{\partial y_n}{\partial x_m}\right) \longrightarrow \frac{\partial y}{\partial x} \in \mathbb{R}^{H \times D}$$

和$\boldsymbol{W}^{\boldsymbol{T}}$形状一样!

$$\frac{\partial y}{\partial x} = \begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \\ w_{14} & w_{24} & w_{34} \end{bmatrix} = \boldsymbol{W}^{\boldsymbol{T}}\ !$$

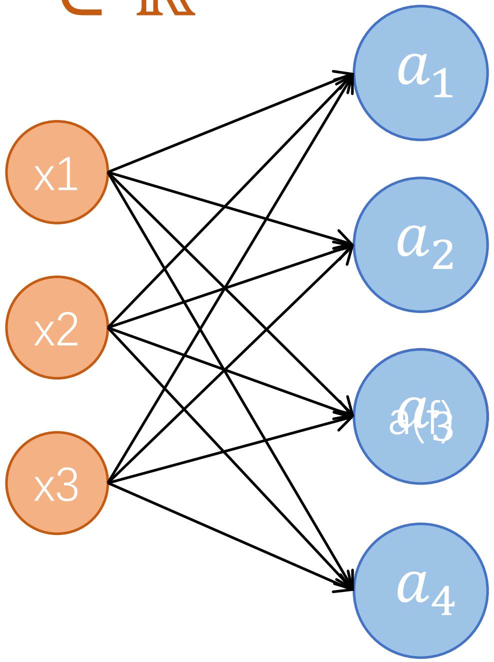
$$\frac{\partial h}{\partial x_i} = \frac{\partial h}{\partial y}\frac{\partial y}{\partial x_i} = \sum_j \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial x_i}$$

$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial y}\frac{\partial y}{\partial x} = \frac{\partial h}{\partial y} W^T$$

反向传播的矩阵运算

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^{D}$

$\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^{H}$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H\times D}$

$x = [2 \quad 1 \quad -3]$

$\frac{\partial h}{\partial x} = [0.50 \quad 0.27 \quad 0.30]$

和$\boldsymbol{x}$的形状一样

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

×

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$$y^T = \begin{bmatrix} x_1 w_{11} + x_2 w_{21} + x_3 w_{31} \\ x_1 w_{12} + x_2 w_{22} + x_3 w_{32} \\ x_1 w_{13} + x_2 w_{23} + x_3 w_{33} \\ x_1 w_{14} + x_2 w_{24} + x_3 w_{34} \end{bmatrix}$$

$y = [-5 \;\; 5 \;\; 2 \;\; -10]$

Sigmoid

$\sigma = [0.007 \;\; 0.99 \;\; 0.88 \;\; 0]$

$\frac{\partial h}{\partial y} = [0.028 \;\; 0.01 \;\; 0.22 \;\; 0]$

$\frac{\partial h}{\partial \sigma} = [4 \;\; 1 \;\; 2 \;\; -1]$

$$\left(\frac{\partial y}{\partial x}\right)_{n,m} = \left(\frac{\partial y_n}{\partial x_m}\right) \longrightarrow \frac{\partial y}{\partial x} \in \mathbb{R}^{H\times D}$$

和$\boldsymbol{W}^{\boldsymbol{T}}$形状一样!

$$\frac{\partial y}{\partial x} = \begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \\ w_{14} & w_{24} & w_{34} \end{bmatrix} = \boldsymbol{W}^{\boldsymbol{T}}\ !$$

$$\frac{\partial h}{\partial x_i} = \frac{\partial h}{\partial y}\frac{\partial y}{\partial x_i} = \sum_j \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial x_i}$$

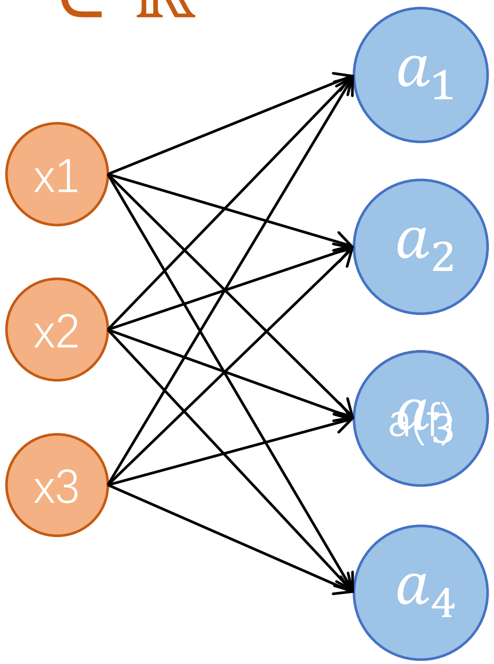
$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial y}\frac{\partial y}{\partial x} = \frac{\partial h}{\partial y} W^T$$

反向传播的矩阵运算

雅可比矩阵 (Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^{D}$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^{H}$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H\times D}$

$$x = [2 \quad 1 \quad -3]$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$$\frac{\partial h}{\partial x} = [0.50 \quad 0.27 \quad 0.30]$$

和$\boldsymbol{x}$的形状一样

$\times$

$$y^T = \begin{bmatrix} x_1 w_{11} + x_2 w_{21} + x_3 w_{31} \\ x_1 w_{12} + x_2 w_{22} + x_3 w_{32} \\ x_1 w_{13} + x_2 w_{23} + x_3 w_{33} \\ x_1 w_{14} + x_2 w_{24} + x_3 w_{34} \end{bmatrix}$$

$$y = [-5 \quad 5 \quad 2 \quad -10]$$

$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

Sigmoid

$$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

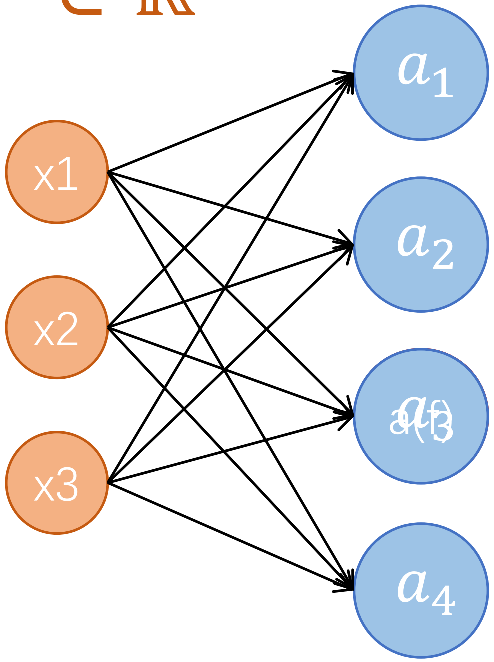
$$\left(\frac{\partial y}{\partial W}\right)_{n,(ij)} = \left(\frac{\partial y_n}{\partial w_{ij}}\right) \longrightarrow \frac{\partial y}{\partial W} \in \mathbb{R}^{H\times D\times H}$$

反向传播的矩阵运算

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^{\boldsymbol{T}} \in \mathbb{R}^D$ $\boldsymbol{a}^{\boldsymbol{T}} \in \mathbb{R}^H$



$\boldsymbol{W}^{\boldsymbol{T}} \in \mathbb{R}^{H\times D}$

$x = [2 \quad 1 \quad -3]$

$\frac{\partial h}{\partial x} = [0.50 \quad 0.27 \quad 0.30]$

和$\boldsymbol{x}$的形状一样

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$$y^T = \begin{bmatrix} x_1w_{11} + x_2w_{21} + x_3w_{31} \\ x_1w_{12} + x_2w_{22} + x_3w_{32} \\ x_1w_{13} + x_2w_{23} + x_3w_{33} \\ x_1w_{14} + x_2w_{24} + x_3w_{34} \end{bmatrix}$$

$\times$

$y = [-5 \quad 5 \quad 2 \quad -10]$

$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$

Sigmoid

$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$

$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$

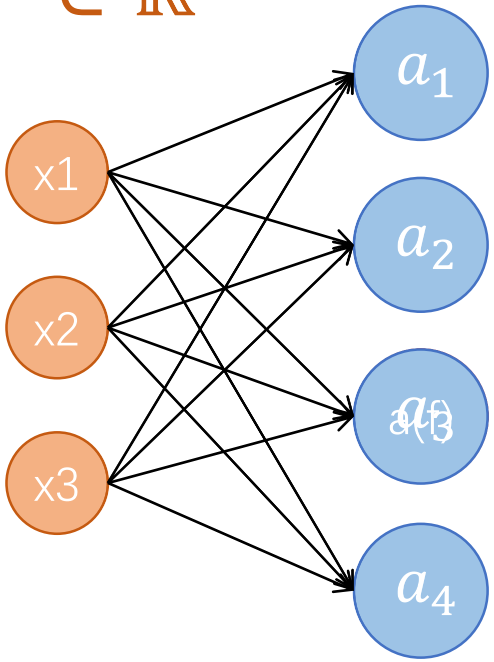
$$\left(\frac{\partial y}{\partial W}\right)_{n,(ij)} = \left(\frac{\partial y_n}{\partial w_{ij}}\right) \longrightarrow \frac{\partial y}{\partial W} \in \mathbb{R}^{H\times D\times H}$$

假设H=4096， D=3072, 需要至少51539607552≈5G内存!!

w_{ij}只影响y_j！！！

反向传播的矩阵运算

$\boldsymbol{x}^T \in \mathbb{R}^D$ $\quad \boldsymbol{a}^T \in \mathbb{R}^H$



$\boldsymbol{W}^T \in \mathbb{R}^{H\times D}$

$$x = [2 \quad 1 \quad -3]$$

$$\frac{\partial h}{\partial x} = [0.50 \quad 0.27 \quad 0.30]$$

和$\boldsymbol{x}$的形状一样

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

×

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$$y^T = \begin{bmatrix} x_1 w_{11} + x_2 w_{21} + x_3 w_{31} \\ x_1 w_{12} + x_2 w_{22} + x_3 w_{32} \\ x_1 w_{13} + x_2 w_{23} + x_3 w_{33} \\ x_1 w_{14} + x_2 w_{24} + x_3 w_{34} \end{bmatrix}$$

$$\boldsymbol{y} = [-5 \;\; 5 \;\; 2 \;\; -10]$$

$$\frac{\partial h}{\partial y} = [0.028 \;\; 0.01 \;\; 0.22 \;\; 0]$$

Sigmoid

$$\sigma = [0.007 \;\; 0.99 \;\; 0.88 \;\; 0]$$

$$\frac{\partial h}{\partial \sigma} = [4 \;\; 1 \;\; 2 \;\; -1]$$

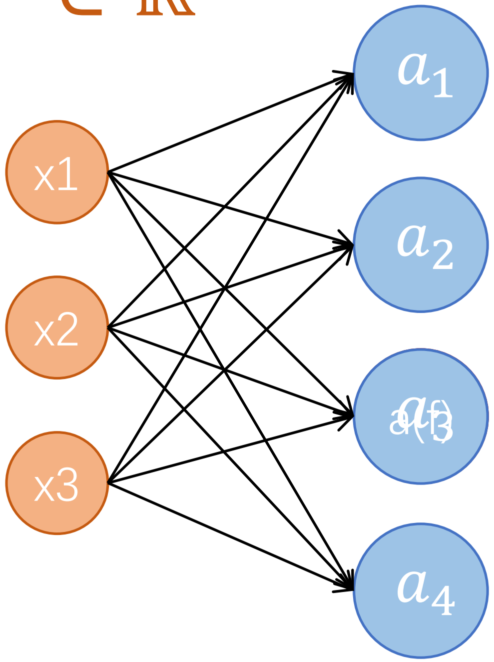
$$\frac{\partial h}{\partial w_{ij}} = \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial w_{ij}} = \frac{\partial h}{\partial y_j} x_i = x_i \frac{\partial h}{\partial y_j}$$

反向传播的矩阵运算

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$\boldsymbol{x}^T \in \mathbb{R}^D$ $\quad$ $\boldsymbol{a}^T \in \mathbb{R}^H$



$\boldsymbol{W}^T \in \mathbb{R}^{H\times D}$

$$x = [2 \quad 1 \quad -3]$$

$$\frac{\partial h}{\partial x} = [0.50 \quad 0.27 \quad 0.30]$$

和$\boldsymbol{x}$的形状一样

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$$y^T = \begin{bmatrix} x_1w_{11} + x_2w_{21} + x_3w_{31} \\ x_1w_{12} + x_2w_{22} + x_3w_{32} \\ x_1w_{13} + x_2w_{23} + x_3w_{33} \\ x_1w_{14} + x_2w_{24} + x_3w_{34} \end{bmatrix}$$

$\times$ $\xrightarrow{\boldsymbol{y} = [-5 \;\; 5 \;\; 2 \;\; -10]}$ Sigmoid $\xrightarrow{\sigma = [0.007 \;\; 0.99 \;\; 0.88 \;\; 0]}$

$$\frac{\partial h}{\partial y} = [0.028 \;\; 0.01 \;\; 0.22 \;\; 0]$$

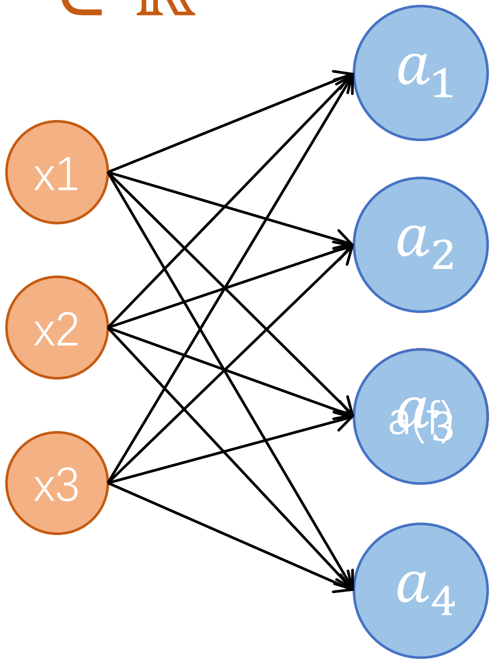
$$\frac{\partial h}{\partial \sigma} = [4 \;\; 1 \;\; 2 \;\; -1]$$

$$\frac{\partial h}{\partial w_{ij}} = \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial w_{ij}} = \frac{\partial h}{\partial y_j}x_i = x_i\frac{\partial h}{\partial y_j}$$

$$\frac{\partial h}{\partial W} = \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial w_{ij}} = x^T\frac{\partial h}{\partial y} = \begin{matrix} 2 \\ 1 \\ -3 \end{matrix} [0.028 \;\; 0.01 \;\; 0.22 \;\; 0]$$

反向传播的矩阵运算

$x^T \in \mathbb{R}^D$ $\quad a^T \in \mathbb{R}^H$



$W^T \in \mathbb{R}^{H\times D}$

雅可比矩阵
(Jacobian matrix)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

$$x = [2 \quad 1 \quad -3]$$

$$\frac{\partial h}{\partial x} = [0.50 \quad 0.27 \quad 0.30]$$

和$\boldsymbol{x}$的形状一样

$$y^T = \begin{bmatrix} x_1 w_{11} + x_2 w_{21} + x_3 w_{31} \\ x_1 w_{12} + x_2 w_{22} + x_3 w_{32} \\ x_1 w_{13} + x_2 w_{23} + x_3 w_{33} \\ x_1 w_{14} + x_2 w_{24} + x_3 w_{34} \end{bmatrix}$$

$$\times$$

$$y = [-5 \quad 5 \quad 2 \quad -10]$$

$$\frac{\partial h}{\partial y} = [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

Sigmoid

$$\sigma = [0.007 \quad 0.99 \quad 0.88 \quad 0]$$

$$\frac{\partial h}{\partial \sigma} = [4 \quad 1 \quad 2 \quad -1]$$

和$\boldsymbol{W}$的形状一样

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix}$$

$$\frac{\partial h}{\partial w_{ij}} = \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial w_{ij}} = \frac{\partial h}{\partial y_j} x_i = x_i \frac{\partial h}{\partial y_j}$$

$$\frac{\partial h}{\partial W} = \begin{bmatrix} 0.056 & 0.02 & 0.44 & 0 \\ 0.028 & 0.01 & 0.22 & 0 \\ -0.084 & -0.03 & -0.66 & 0 \end{bmatrix}$$

$$\frac{\partial h}{\partial W} = \frac{\partial h}{\partial y_j}\frac{\partial y_j}{\partial w_{ij}} = x^T \frac{\partial h}{\partial y} = \begin{matrix} 2 \\ 1 \\ -3 \end{matrix} [0.028 \quad 0.01 \quad 0.22 \quad 0]$$

反向传播的矩阵运算： minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

×

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

$$\frac{\partial h}{\partial y}$$

y_{ij}只影响σ_{ij}, 只需pairwise计算

反向传播的矩阵运算：minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

×

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

x_{ij}只影响y的第i行

反向传播的矩阵运算：minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

×

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

$$\frac{\partial h}{\partial x_{ij}} = \sum_k \frac{\partial h}{\partial y_{ik}} \frac{\partial y_{ik}}{\partial x_{ij}} = \sum_k \frac{\partial h}{\partial y_{ik}} w_{jk} \Rightarrow \frac{\partial h}{\partial x} = \frac{\partial h}{\partial y} W^T$$

x_{ij}只影响y的第i行

反向传播的矩阵运算： minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$\frac{\partial y}{\partial x}$$

$$\frac{\partial h}{\partial x} = \begin{bmatrix} 0.50 & 0.27 & 0.30 \\ -0.21 & -0.24 & -0.17 \end{bmatrix}$$

$\times$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

$$\frac{\partial h}{\partial x_{ij}} = \sum_k \frac{\partial h}{\partial y_{ik}} \frac{\partial y_{ik}}{\partial x_{ij}} = \sum_k \frac{\partial h}{\partial y_{ik}} w_{jk} \Rightarrow \frac{\partial h}{\partial x} = \frac{\partial h}{\partial y} W^T$$

x_{ij}只影响y的第i行

反向传播的矩阵运算：minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

$$\frac{\partial h}{\partial x} = \begin{bmatrix} 0.50 & 0.27 & 0.30 \\ -0.21 & -0.24 & -0.17 \end{bmatrix}$$

$\times$

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

w_{ij}只影响y的第j列

反向传播的矩阵运算：minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

$$\frac{\partial h}{\partial x} = \begin{bmatrix} 0.50 & 0.27 & 0.30 \\ -0.21 & -0.24 & -0.17 \end{bmatrix}$$

$\times$

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

$$\frac{\partial h}{\partial w_{ij}} = \sum_k \frac{\partial h}{\partial y_{kj}} \frac{\partial y_{kj}}{\partial w_{ij}} = \sum_k \frac{\partial h}{\partial y_{kj}} x_{ki} = \sum_k x_{ki} \frac{\partial h}{\partial y_{kj}} \Rightarrow \frac{\partial h}{\partial W} = X^T \frac{\partial h}{\partial y}$$

w_{ij}只影响y的第j列

反向传播的矩阵运算：minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$\frac{\partial h}{\partial x} = \begin{bmatrix} 0.50 & 0.27 & 0.30 \\ -0.21 & -0.24 & -0.17 \end{bmatrix}$$

×

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

$$\frac{\partial h}{\partial W} = \begin{bmatrix} 0.046 & -0.04 & 0.64 & 0 \\ 0.038 & 0.07 & 0.02 & 0 \\ -0.064 & 0.09 & -1.06 & 0 \end{bmatrix}$$

$$\frac{\partial h}{\partial w_{ij}} = \sum_k \frac{\partial h}{\partial y_{kj}} \frac{\partial y_{kj}}{\partial w_{ij}} = \sum_k \frac{\partial h}{\partial y_{kj}} x_{ki} = \sum_k x_{ki} \frac{\partial h}{\partial y_{kj}} \Rightarrow \frac{\partial h}{\partial W} = X^T \frac{\partial h}{\partial y}$$

w_{ij}只影响y的第j列

反向传播的矩阵运算：minibatch

$$x = \begin{bmatrix} 2 & 1 & -3 \\ -1 & 1 & 2 \end{bmatrix} \in N \times D$$

$$\frac{\partial h}{\partial x} = \begin{bmatrix} 0.50 & 0.27 & 0.30 \\ -0.21 & -0.24 & -0.17 \end{bmatrix}$$

$\times$

$$y = \begin{bmatrix} -5 & 5 & 2 & -10 \\ 7 & -4 & 1 & 10 \end{bmatrix}$$

$$\frac{\partial h}{\partial y} = \begin{bmatrix} 0.028 & 0.01 & 0.22 & 0 \\ 0.01 & 0.06 & -0.2 & 0 \end{bmatrix}$$

Sigmoid

$$\sigma = \begin{bmatrix} 0.007 & 0.99 & 0.88 & 0 \\ 0.99 & 0.02 & 0.73 & 1 \end{bmatrix}$$

$$\frac{\partial h}{\partial \sigma} = \begin{bmatrix} 4 & 1 & 2 & -1 \\ 1 & 3 & -1 & 2 \end{bmatrix}$$

$$W = \begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & -1 & 1 & 2 \\ 3 & 0 & 1 & 4 \end{bmatrix} \in D \times M$$

$$\frac{\partial h}{\partial W} = \begin{bmatrix} 0.046 & -0.04 & 0.64 & 0 \\ 0.038 & 0.07 & 0.02 & 0 \\ -0.064 & 0.09 & -1.06 & 0 \end{bmatrix}$$

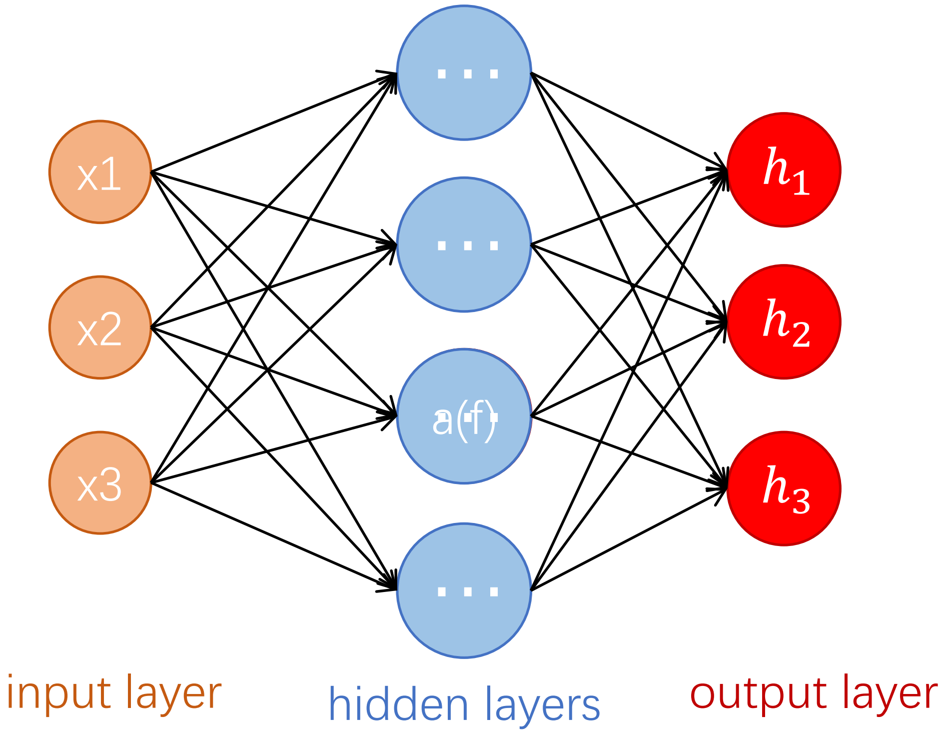
$\frac{\partial h}{\partial w_{ij}} =$

矩阵的梯度和矩

状的形状一样

w_{ij}只影响y的第j列

多层神经网络反向传播



Chain rule

$$\frac{\partial L}{\partial \boldsymbol{W}^{l}} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{W}^{l}}$$

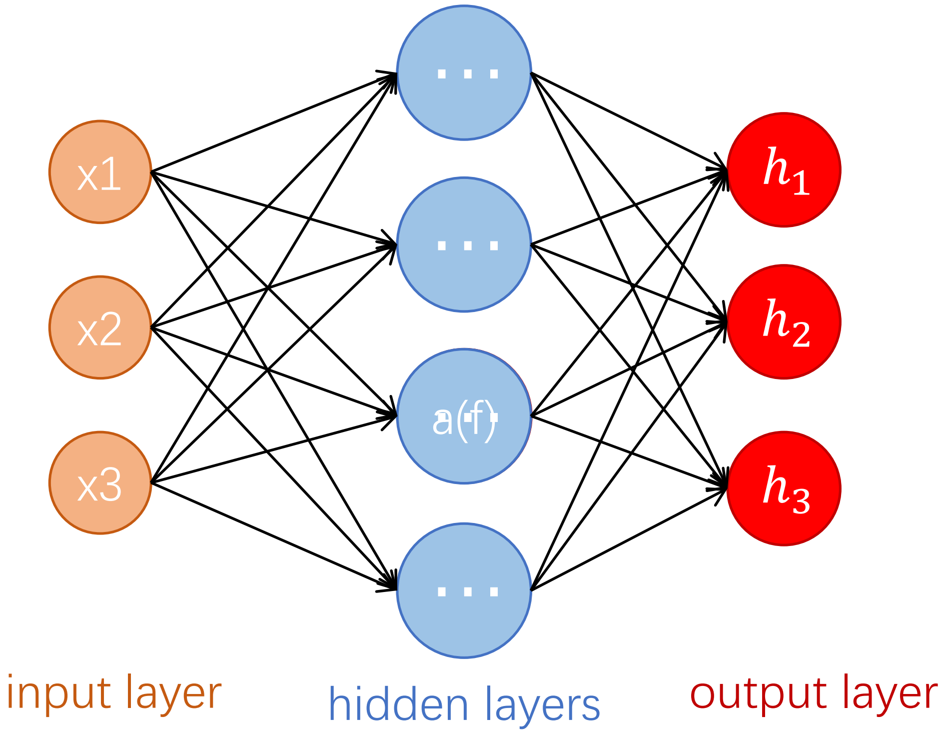
$$\frac{\partial L}{\partial \boldsymbol{W}^{l-1}} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}^{l}} \frac{\partial \boldsymbol{a}^{l}}{\partial \boldsymbol{W}^{l-1}}$$

$$\frac{\partial L}{\partial \boldsymbol{W}^{l-2}} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}^{l}} \frac{\partial \boldsymbol{a}^{l}}{\partial \boldsymbol{a}^{l-1}} \frac{\partial \boldsymbol{a}^{l-1}}{\partial \boldsymbol{W}^{l-2}}$$

$$\vdots$$

$$\frac{\partial L}{\partial \boldsymbol{W}^{1}} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}^{l}} \frac{\partial \boldsymbol{a}^{l}}{\partial \boldsymbol{a}^{l-1}} \frac{\partial \boldsymbol{a}^{l-1}}{\partial \boldsymbol{a}^{l-2}} \cdots \cdots \frac{\partial \boldsymbol{a}^{2}}{\partial \boldsymbol{W}^{1}}$$

多层神经网络反向传播



Chain rule

$$\frac{\partial L}{\partial \boldsymbol{W}^l} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{W}^l}$$

$$\frac{\partial L}{\partial \boldsymbol{W}^{l-1}} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}^l} \frac{\partial \boldsymbol{a}^l}{\partial \boldsymbol{W}^{l-1}}$$

$$\frac{\partial L}{\partial \boldsymbol{W}^{l-2}} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}^l} \frac{\partial \boldsymbol{a}^l}{\partial \boldsymbol{a}^{l-1}} \frac{\partial \boldsymbol{a}^{l-1}}{\partial \boldsymbol{W}^{l-2}}$$

$$\vdots$$

不要忘记每一层的偏置项权重b，以及正则项！

$$\frac{\partial L}{\partial \boldsymbol{W}^1} = \frac{\partial L}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}^l} \frac{\partial \boldsymbol{a}^l}{\partial \boldsymbol{a}^{l-1}} \frac{\partial \boldsymbol{a}^{l-1}}{\partial \boldsymbol{a}^{l-2}} \cdots\cdots \frac{\partial \boldsymbol{a}^2}{\partial \boldsymbol{W}^1}$$

Assignment1: 两层神经网络实现

✓loss()函数：返回loss和梯度

✓train()函数：梯度下降优化，调整超参数

✓Predict()函数：测试集分类

```python
# Compute the forward pass
scores = None
#############################################################################
# TODO: Perform the forward pass, computing the class scores for the input. #
# Store the result in the scores variable, which should be an array of      #
# shape (N, C).                                                             #
#############################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

pass

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

# If the targets are not given then jump out, we're done
if y is None:
    return scores

# Compute the loss
loss = None
#############################################################################
# TODO: Finish the forward pass, and compute the loss. This should include  #
# both the data loss and L2 regularization for W1 and W2. Store the result  #
# in the variable loss, which should be a scalar. Use the Softmax           #
# classifier loss.                                                          #
#############################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

pass

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

# Backward pass: compute gradients
grads = {}
#############################################################################
# TODO: Compute the backward pass, computing the derivatives of the weights #
# and biases. Store the results in the grads dictionary. For example,       #
# grads['W1'] should store the gradient on W1, and be a matrix of same size #
#############################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

pass

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

return loss, grads
```

小结

- 神经网络
  - ✓激活函数
  - ✓神经网络的层次
  - ✓前馈计算

- 优化
  - ✓计算图
  - ✓反向传播
  - ✓链式规则

L05

- Convolutional Neural Networks