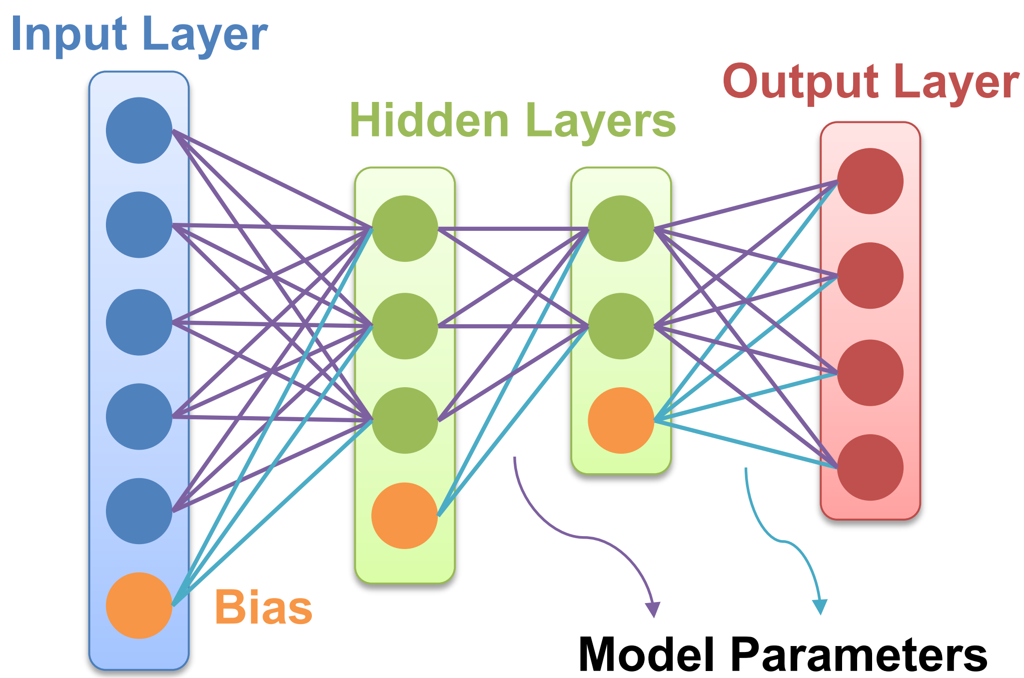
Backpropagation

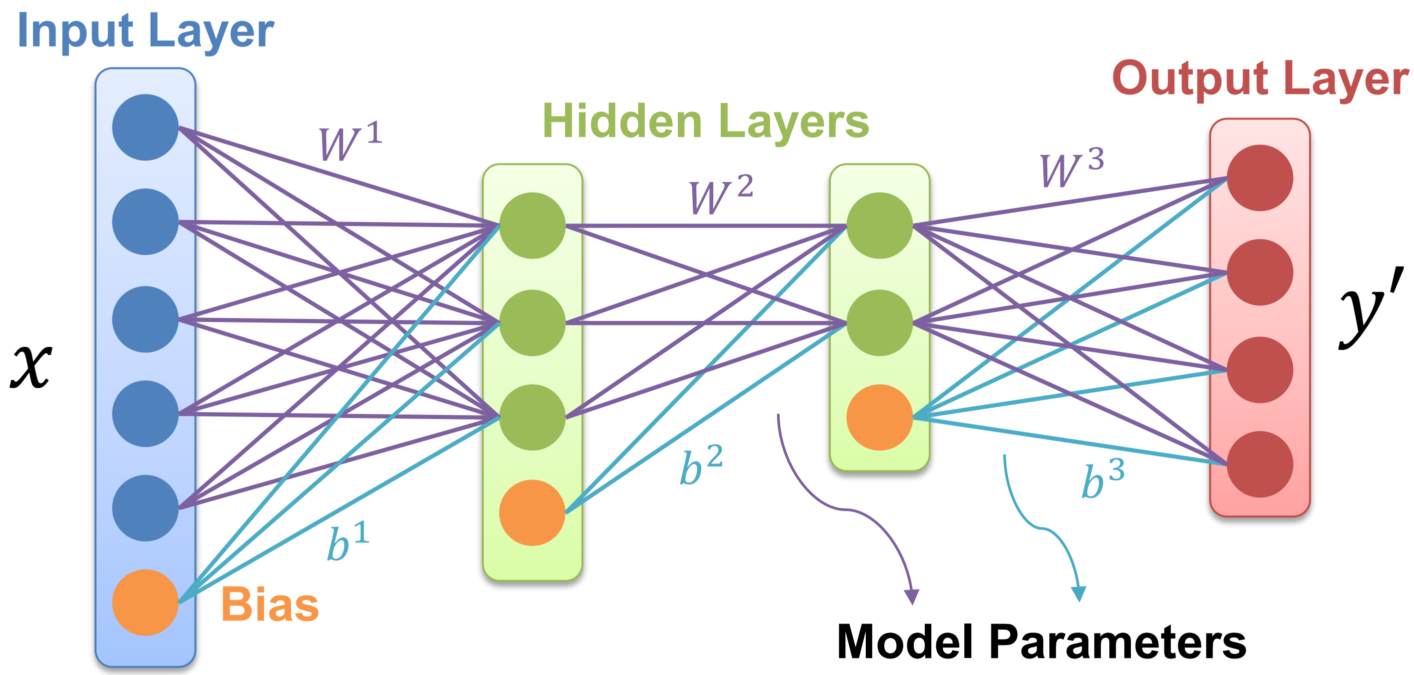
Kuan-Yu Chen (陳冠宇)

2018/03/15 @ TR-409, NTUST

Fully-Connected Feed-Forward

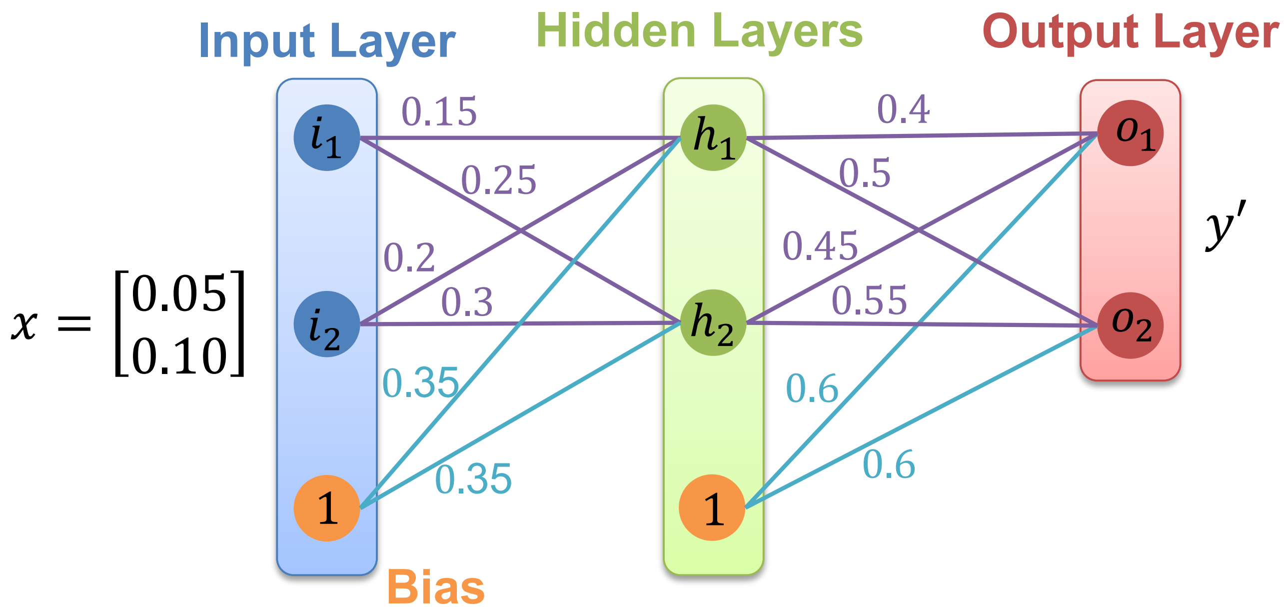


Forward Propagation – 1

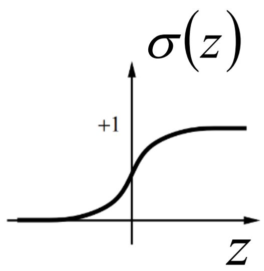


$$y' = \sigma(W^3\sigma(W^2\sigma(W^1x + b^1) + b^2) + b^3\ )$$

Forward Propagation – 2



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



$$y' = \sigma(W^2 \sigma(W^1 x + b^1) + b^2)$$

$$h_1 = \sigma(0.15 \times 0.05 + 0.2 \times 0.1 + 0.35) = 0.593$$

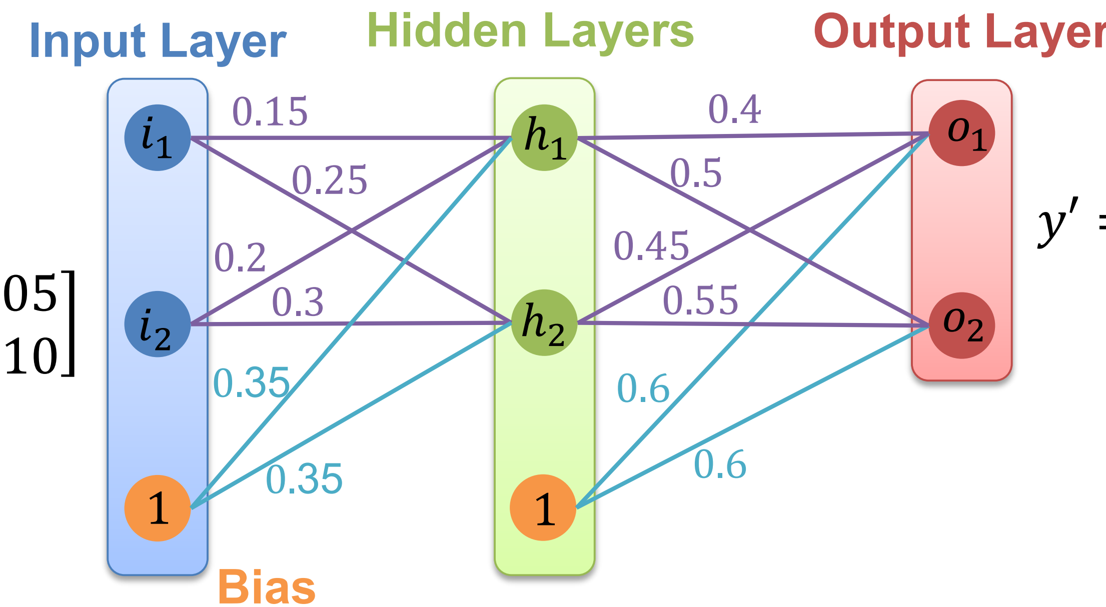
$$h_2 = \sigma(0.25 \times 0.05 + 0.3 \times 0.1 + 0.35) = 0.597$$

$$o_1 = \sigma(0.4 \times 0.593 + 0.45 \times 0.597 + 0.6) = 0.751$$

$$o_2 = \sigma(0.5 \times 0.593 + 0.55 \times 0.597 + 0.6) = 0.773$$

Mean Squared Error

$$x = \begin{bmatrix} 0.05 \\ 0.10 \end{bmatrix}$$



$$y' = \begin{bmatrix} 0.751 \\ 0.773 \end{bmatrix} \quad y = \begin{bmatrix} 0.01 \\ 0.99 \end{bmatrix}$$

$$MSE = \frac{1}{N \times D} \sum_{n=1}^{N} (y_n - y'_n)^2$$

number of sample

size of output

$$MSE = \frac{1}{2} \sum_{n=1}^{N} (y_n - y'_n)^2$$

$$MSE = \frac{1}{N} \sum_{n=1}^{N} (y_n - y'_n)^2$$

$$MSE = \frac{1}{2}\left((0.01 - 0.751)^2 + (0.99 - 0.773)^2\right) = 0.275$$

Gradient Descent – 1

- Gradient descent is based on the observation that if the multi-variable function $f_\theta(\cdot)$ is defined and differentiable in a neighborhood of a point x, then $f_\theta(x)$ decreases fastest if one goes from x in the direction of the negative gradient of $f_\theta(x)$

$$f = MSE = \frac{1}{N \times D} \sum_{n=1}^{N} (y_n - y_n')^2$$

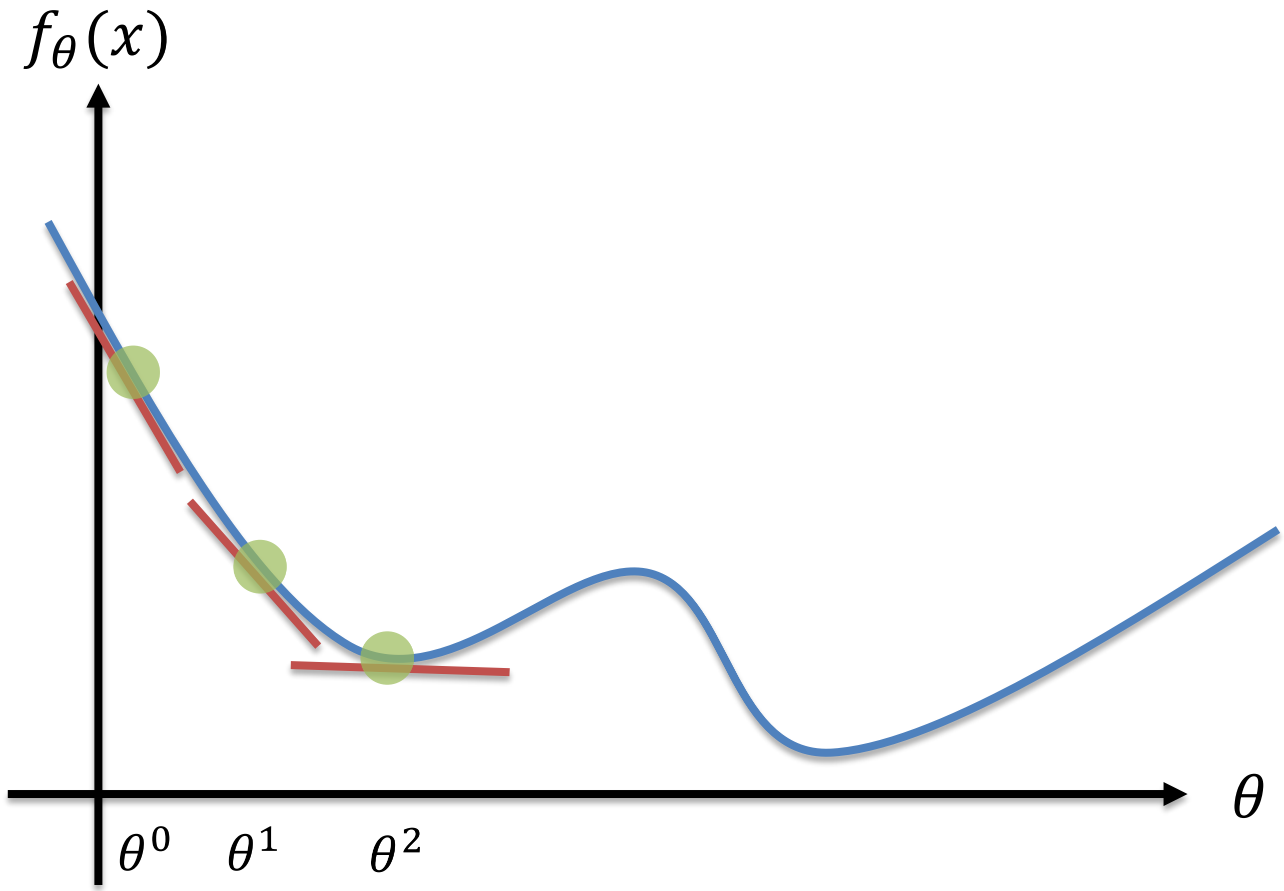
$$\theta = \{W^3, W^2, W^1, b^1, b^2, b^3\}$$

$$y' = \sigma(W^3 \sigma(W^2 \sigma(W^1 x + b^1) + b^2) + b^3)$$

$$\theta^{i+1} = \theta^i - \underbrace{\eta}_{\text{step size}} \frac{\partial f}{\partial \theta}$$

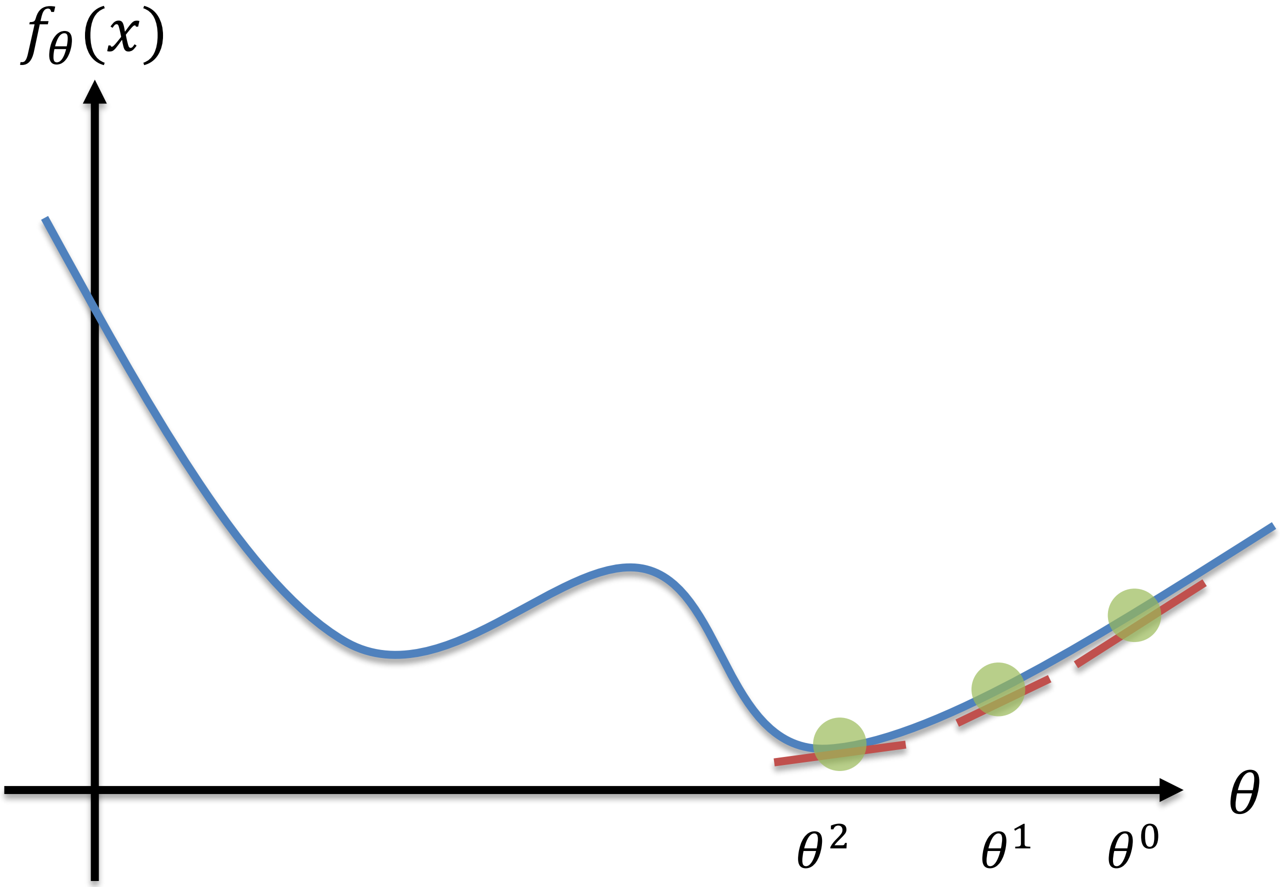
Gradient Descent – 2

$$\theta^{i+1} = \theta^{i} - \eta \frac{\partial f}{\partial \theta}$$

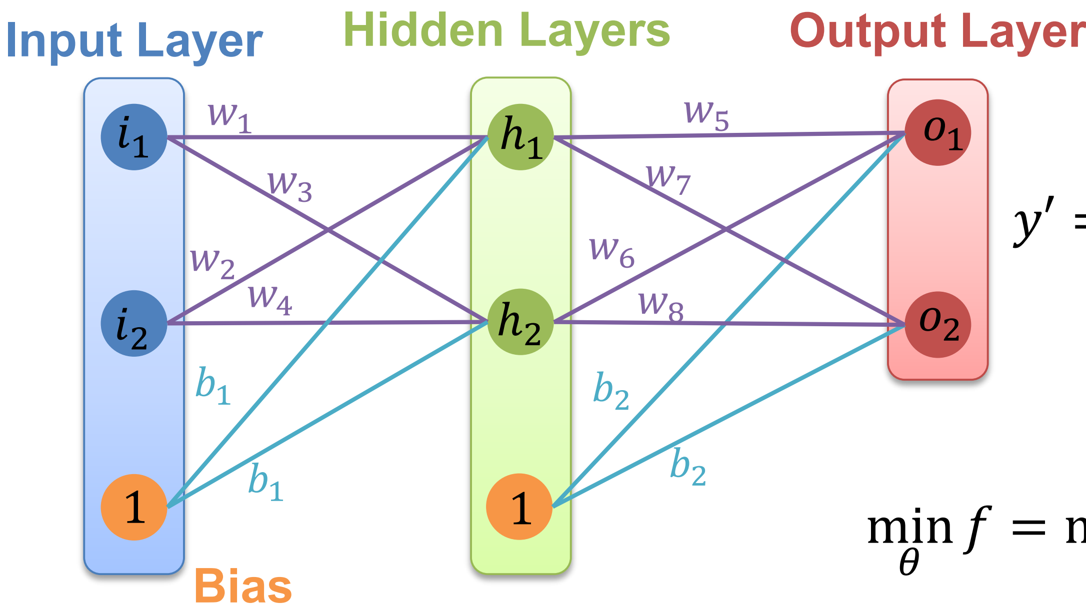


Gradient Descent – 3

$$\theta^{i+1} = \theta^{i} - \eta \frac{\partial f}{\partial \theta}$$



Update the Model Parameters! – 1



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} 0.01 \\ 0.99 \end{bmatrix}$$

$$\min_\theta f = \min_\theta \frac{1}{N \times D} \sum_{n=1}^{N} (y_n - y_n')^2$$

$$h_1 = \sigma(w_1 \times i_1 + w_2 \times i_2 + b_1) = \sigma\left(net_{h_1}\right)$$

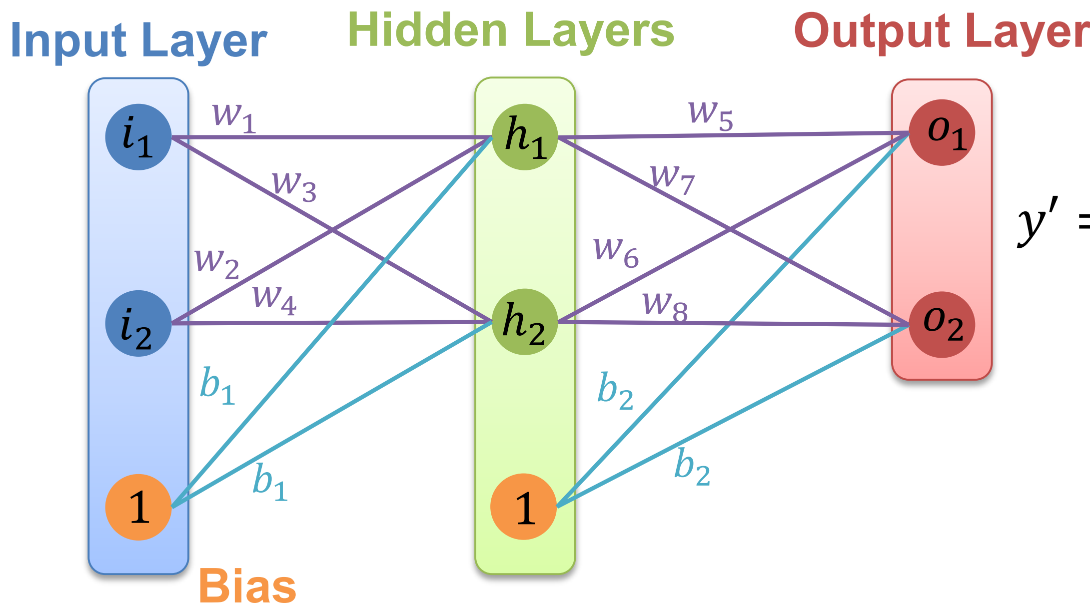
$$h_2 = \sigma(w_3 \times i_1 + w_4 \times i_2 + b_1) = \sigma\left(net_{h_2}\right)$$

$$o_1 = \sigma(w_5 \times h_1 + w_6 \times h_2 + b_2) = \sigma\left(net_{o_1}\right)$$

$$o_2 = \sigma(w_7 \times h_1 + w_8 \times h_2 + b_2) = \sigma\left(net_{o_2}\right)$$

Update the Model Parameters! – 2

$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$



$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\min_\theta f = \min_\theta \frac{1}{N \times D} \sum_{n=1}^{N} (y_n - y'_n)^2$$

$$h_1 = \sigma(w_1 \times i_1 + w_2 \times i_2 + b_1) = \sigma(net_{h_1})$$

$$h_2 = \sigma(w_3 \times i_1 + w_4 \times i_2 + b_1) = \sigma(net_{h_2})$$

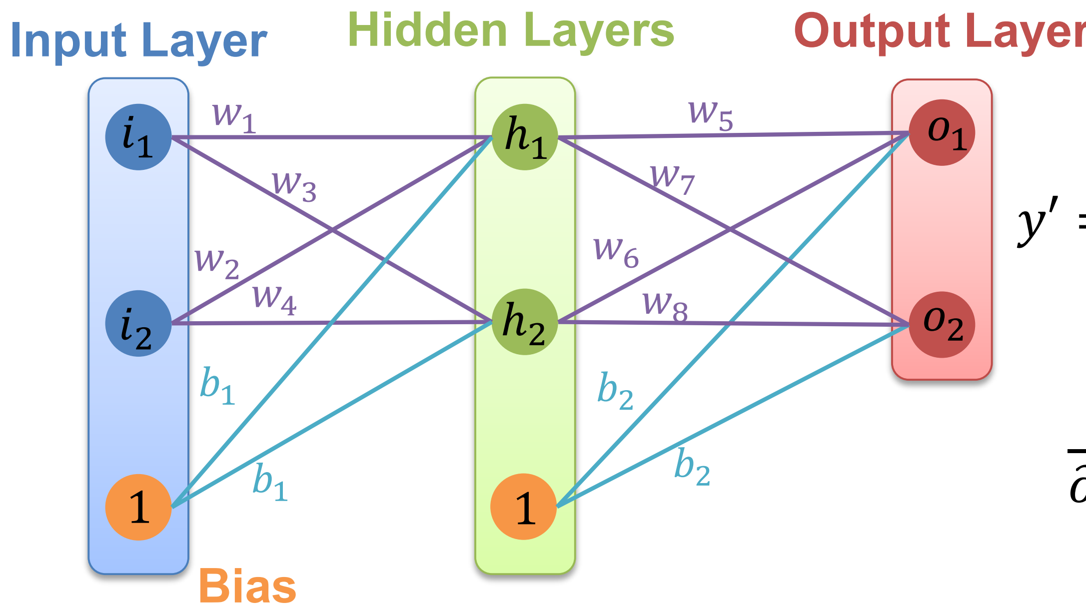
$$o_1 = \sigma(w_5 \times h_1 + w_6 \times h_2 + b_2) = \sigma(net_{o_1})$$

$$o_2 = \sigma(w_7 \times h_1 + w_8 \times h_2 + b_2) = \sigma(net_{o_2})$$

$$w_5^{new} = w_5^{old} - \eta \frac{\partial f}{\partial w_5}$$

$$\frac{\partial f}{\partial w_5} = \frac{\partial f}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial w_5}$$

Update the Model Parameters! – 3



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

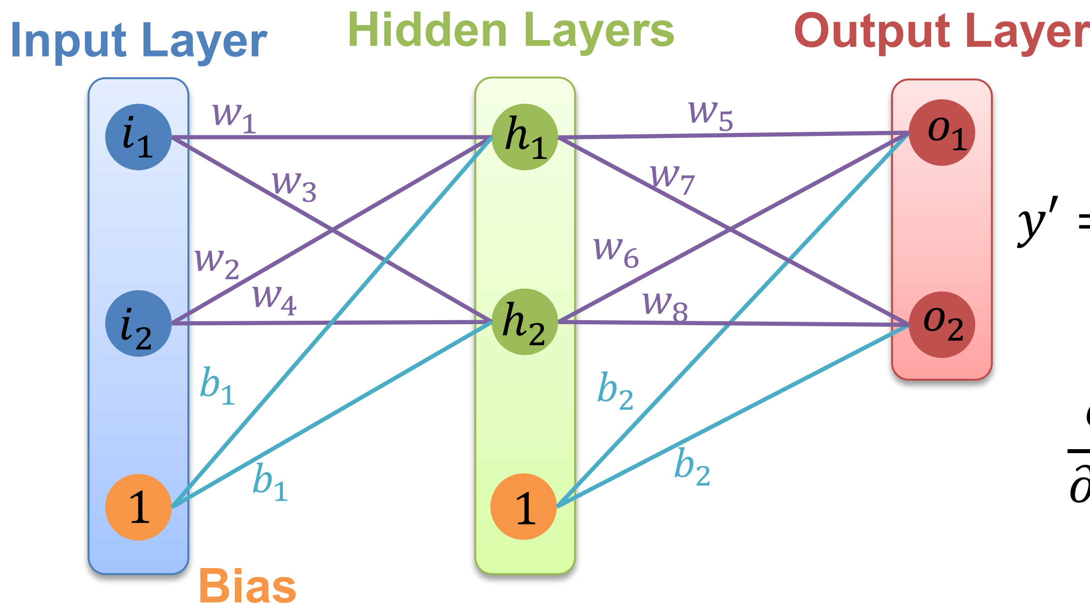
$$\frac{\partial f}{\partial w_5} = \frac{\partial f}{\partial o_1}\frac{\partial o_1}{\partial net_{o_1}}\frac{\partial net_{o_1}}{\partial w_5}$$

$$f = \frac{1}{N \times D}\sum_{n=1}^{N}(y_n - y_n')^2 = \frac{1}{2}\left((t_1 - o_1)^2 + (t_2 - o_2)^2\right)$$

$$= \frac{1}{2} \times (t_1 - o_1)^2 + \frac{1}{2} \times (t_2 - o_2)^2$$

$$\frac{\partial f}{\partial o_1} = \frac{1}{2} \times 2 \times (t_1 - o_1) \times (-1)$$

Update the Model Parameters! – 4



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

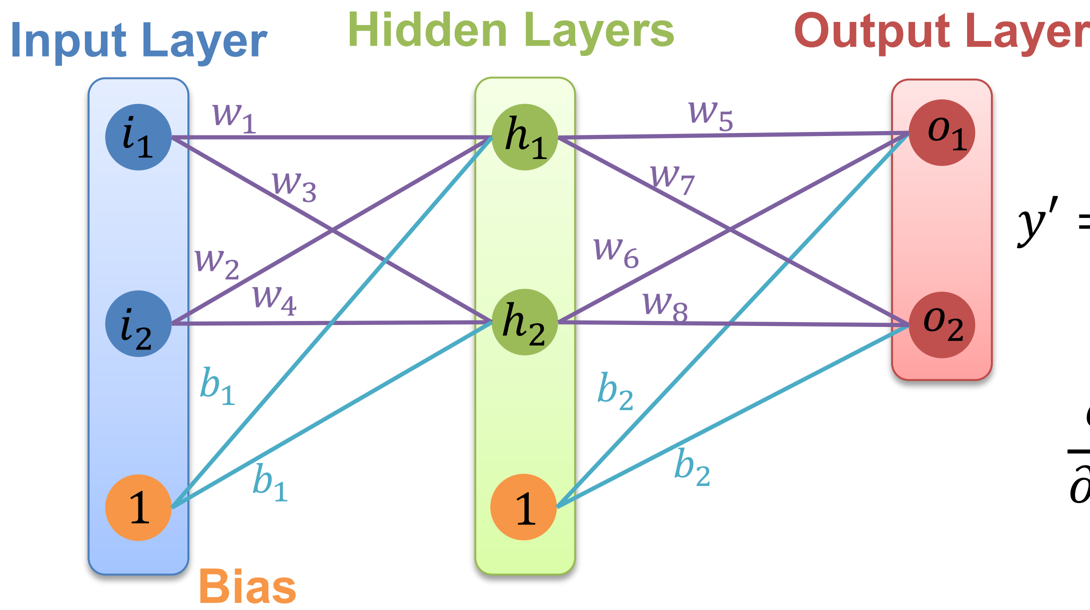
$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_5} = \frac{\partial f}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial w_5}$$

$$o_1 = \sigma(w_5 \times h_1 + w_6 \times h_2 + b_2) = \sigma(net_{o_1}) = \frac{1}{1 + e^{-net_{o_1}}}$$

$$\frac{\partial o_1}{\partial net_{o_1}} = o_1 \times (1 - o_1)$$

Update the Model Parameters! – 5



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

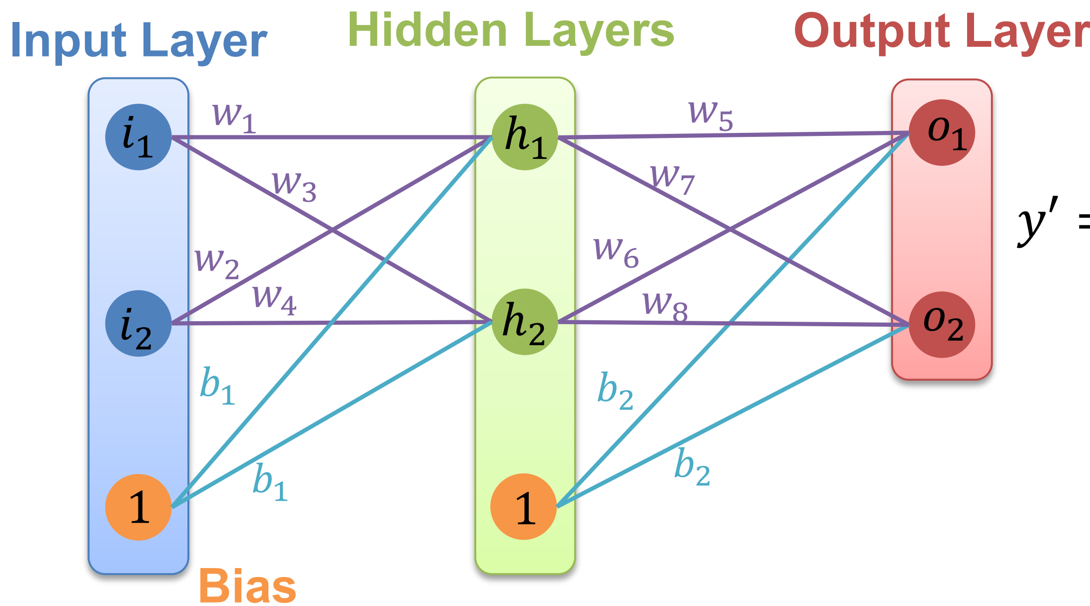
$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_5} = \frac{\partial f}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial w_5}$$

$$net_{o_1} = w_5 \times h_1 + w_6 \times h_2 + b_2$$

$$\frac{\partial net_{o_1}}{\partial w_5} = h_1$$

Update the Model Parameters! – 6



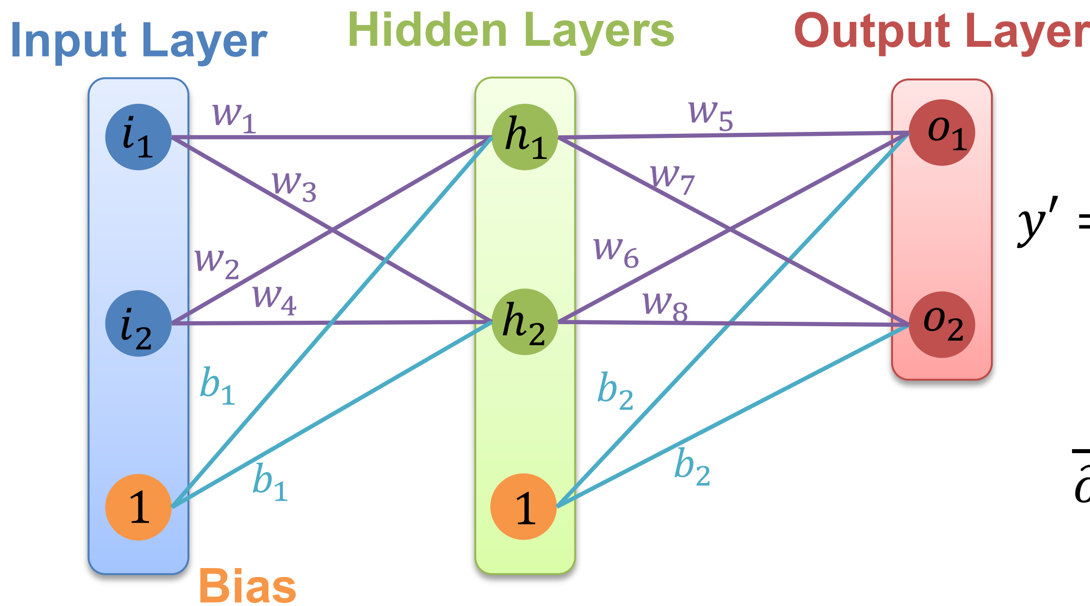
$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \qquad y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_5} = \frac{\partial f}{\partial o_1}\frac{\partial o_1}{\partial net_{o_1}}\frac{\partial net_{o_1}}{\partial w_5} = \left(\frac{1}{2} \times 2 \times (t_1 - o_1) \times (-1)\ \right)(o_1 \times (1 - o_1)\ )h_1$$

$$w_5^{new} = w_5^{old} - \eta \frac{\partial f}{\partial w_5}$$

Update the Model Parameters! – 7

$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$



$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_1}$$

$$o_1 = \sigma(w_5 \times h_1 + w_6 \times h_2 + b_2) = \sigma\left(net_{o_1}\right)$$

$$o_2 = \sigma(w_7 \times h_1 + w_8 \times h_2 + b_2) = \sigma\left(net_{o_2}\right)$$

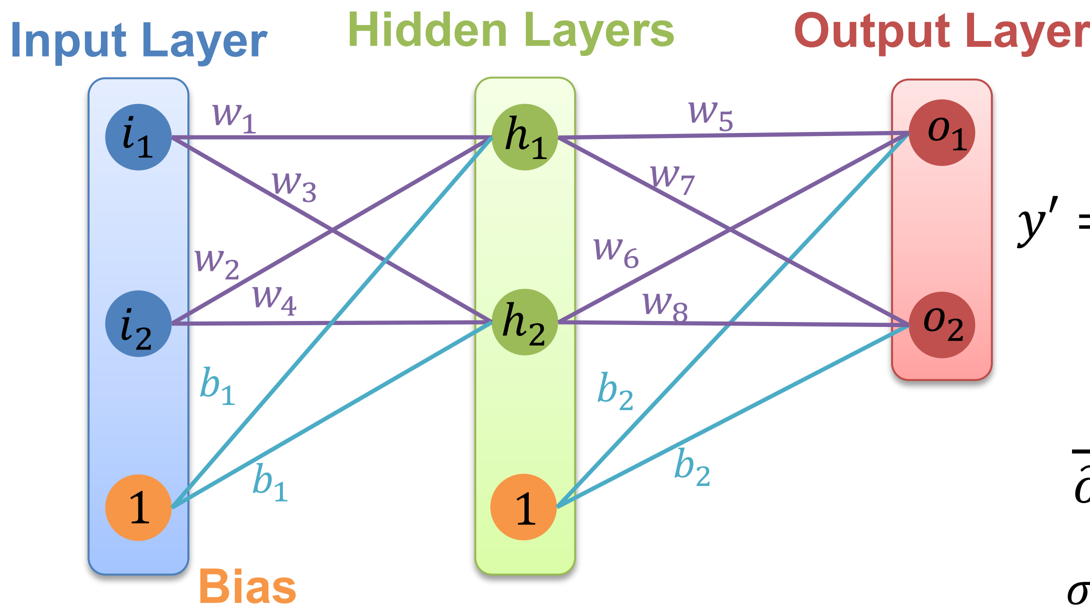
$$f = \frac{1}{2} \times (t_1 - o_1)^2 + \frac{1}{2} \times (t_2 - o_2)^2$$

$$\frac{\partial f}{\partial h_1} = \frac{\partial \frac{1}{2} \times (t_1 - o_1)^2}{\partial h_1} + \frac{\partial \frac{1}{2} \times (t_2 - o_2)^2}{\partial h_1}$$

$$= \frac{\partial \frac{1}{2} \times (t_1 - o_1)^2}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial h_1} + \frac{\partial \frac{1}{2} \times (t_2 - o_2)^2}{\partial o_2} \frac{\partial o_2}{\partial net_{o_2}} \frac{\partial net_{o_2}}{\partial h_1}$$

$$= \left(\frac{1}{2} \times 2 \times (t_1 - o_1) \times (-1)\right)(o_1 \times (1 - o_1))(w_5) + \left(\frac{1}{2} \times 2 \times (t_2 - o_2) \times (-1)\right)(o_2 \times (1 - o_2))(w_7)$$

Update the Model Parameters! – 8



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_1}$$

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

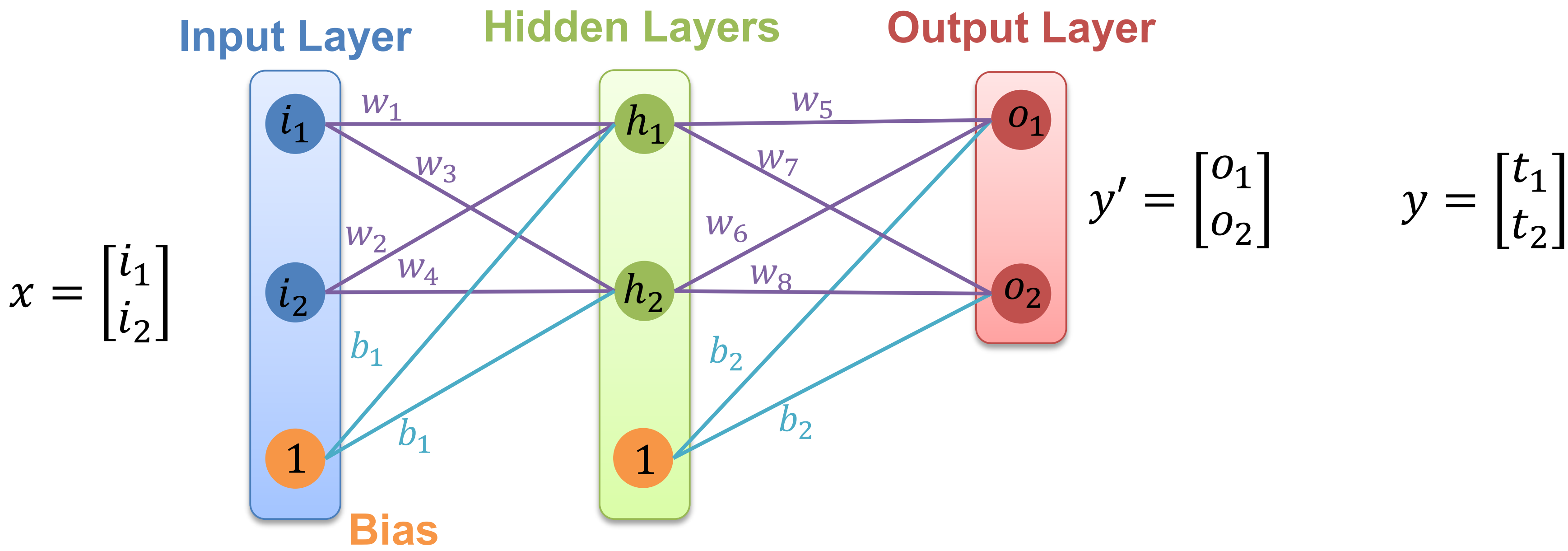
$$f = \frac{1}{2} \times (t_1 - o_1)^2 + \frac{1}{2} \times (t_2 - o_2)^2$$

$$h_1 = \sigma(w_1 \times i_1 + w_2 \times i_2 + b_1) = \sigma\left(net_{h_1}\right)$$

$$\frac{\partial h_1}{\partial net_{h_1}} = h_1 \times (1 - h_1)$$

$$\frac{\partial net_{h_1}}{\partial w_1} = i_1$$

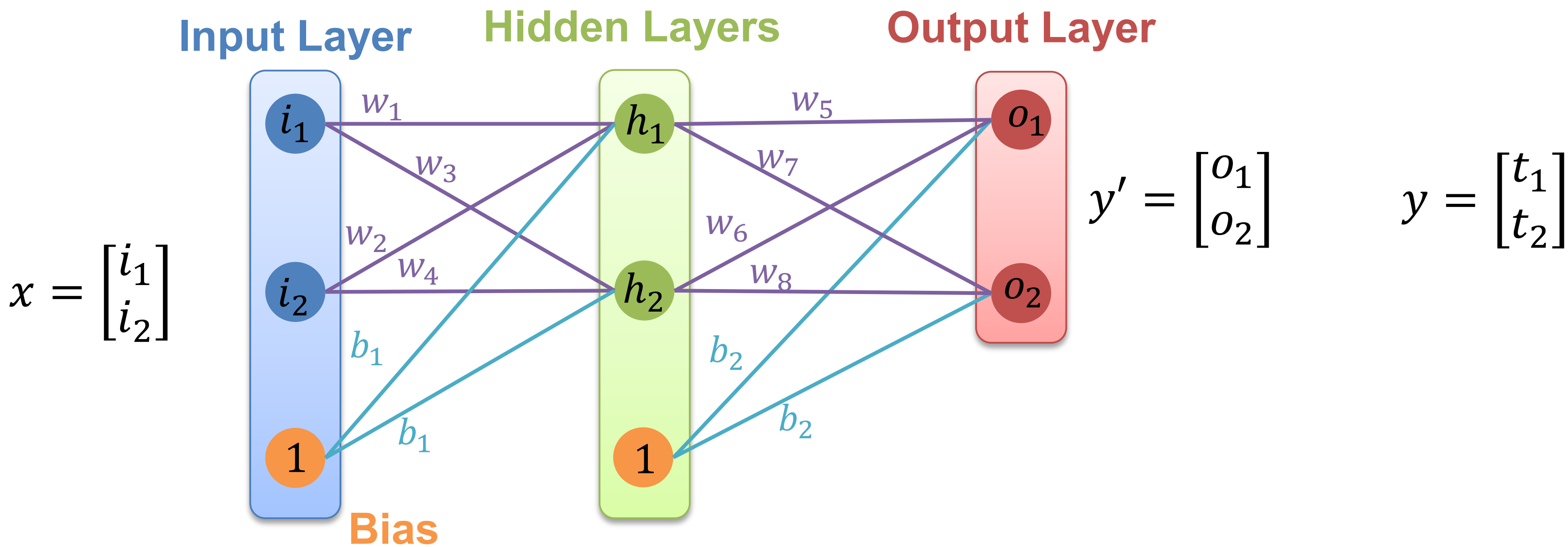
Update the Model Parameters! – 9



$$\frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_1}$$

$$= \left(\frac{\partial \frac{1}{2} \times (t_1 - o_1)^2}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial h_1} + \frac{\partial \frac{1}{2} \times (t_2 - o_2)^2}{\partial o_2} \frac{\partial o_2}{\partial net_{o_2}} \frac{\partial net_{o_2}}{\partial h_1} \right) \frac{\partial h_1}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_1}$$

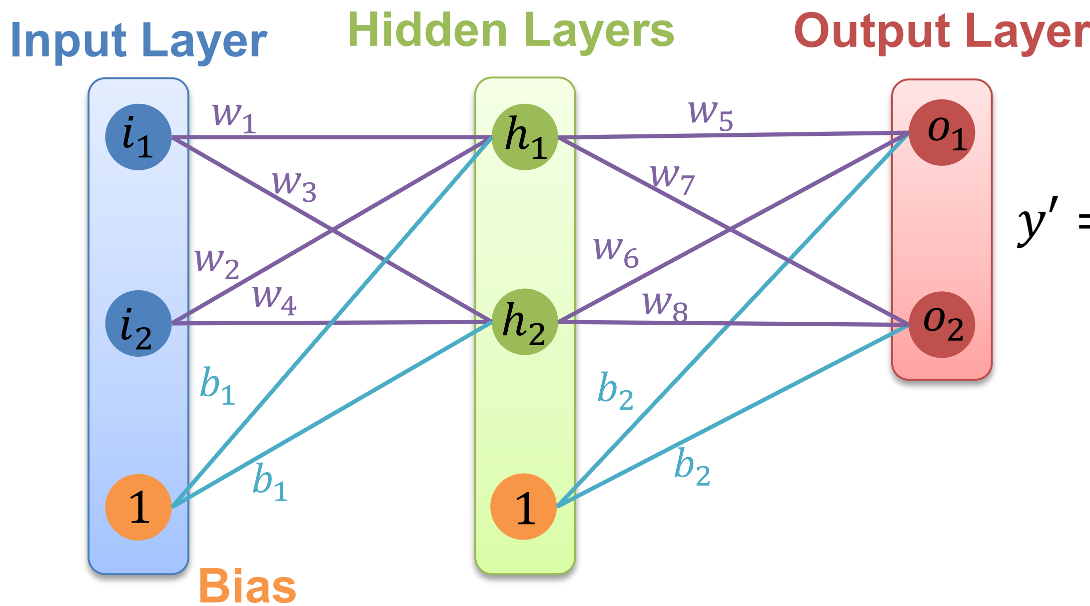
Update the Model Parameters! – 10



$$\frac{\partial f}{\partial w_1} = \boxed{\frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial net_{h_1}}} \frac{\partial net_{h_1}}{\partial w_1}$$

$$= \boxed{\left(\frac{\partial \frac{1}{2} \times (t_1 - o_1)^2}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial h_1} + \frac{\partial \frac{1}{2} \times (t_2 - o_2)^2}{\partial o_2} \frac{\partial o_2}{\partial net_{o_2}} \frac{\partial net_{o_2}}{\partial h_1} \right) \frac{\partial h_1}{\partial net_{h_1}}} \frac{\partial net_{h_1}}{\partial w_1}$$

Update the Model Parameters! – 11

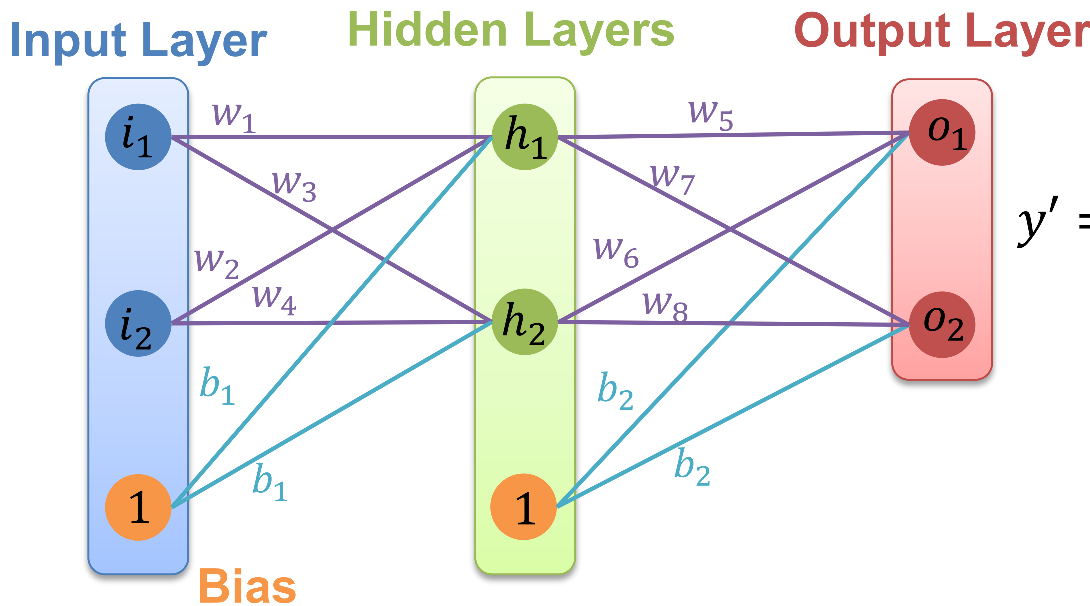


$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \qquad y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_1} = \boxed{\frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial net_{h_1}}} \frac{\partial net_{h_1}}{\partial w_1}$$

$$= \boxed{\left(\boxed{\frac{\partial \frac{1}{2} \times (t_1 - o_1)^2}{\partial o_1} \frac{\partial o_1}{\partial net_{o_1}}} \frac{\partial net_{o_1}}{\partial h_1} + \boxed{\frac{\partial \frac{1}{2} \times (t_2 - o_2)^2}{\partial o_2} \frac{\partial o_2}{\partial net_{o_2}}} \frac{\partial net_{o_2}}{\partial h_1} \right) \frac{\partial h_1}{\partial net_{h_1}}} \frac{\partial net_{h_1}}{\partial w_1}$$

Update the Model Parameters! – 12



$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$ $\quad y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix}$ $\quad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$

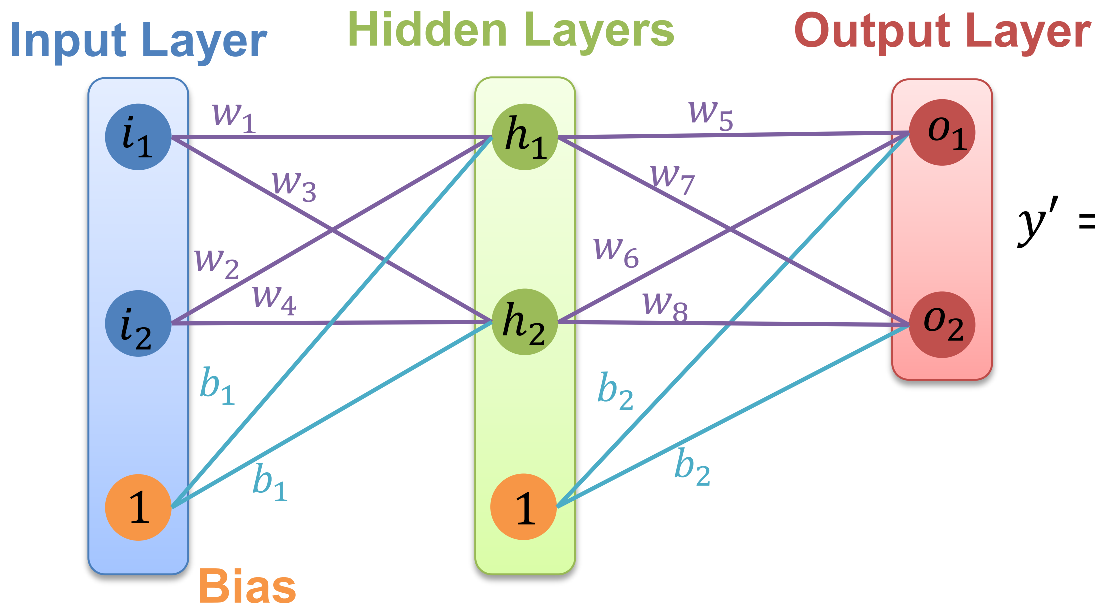
$$\frac{\partial f}{\partial a_n}\frac{\partial a_n}{\partial net_{a_n}} = \delta_{a_n}$$

$$\frac{\partial f}{\partial w_1} = \boxed{\frac{\partial f}{\partial h_1}\frac{\partial h_1}{\partial net_{h_1}}}\frac{\partial net_{h_1}}{\partial w_1} = \delta_{h_1}\frac{\partial net_{h_1}}{\partial w_1}$$

$$= \boxed{\left(\boxed{\frac{\partial \frac{1}{2}\times(t_1-o_1)^2}{\partial o_1}\frac{\partial o_1}{\partial net_{o_1}}}\frac{\partial net_{o_1}}{\partial h_1} + \boxed{\frac{\partial \frac{1}{2}\times(t_2-o_2)^2}{\partial o_2}\frac{\partial o_2}{\partial net_{o_2}}}\frac{\partial net_{o_2}}{\partial h_1}\right)\frac{\partial h_1}{\partial net_{h_1}}}\frac{\partial net_{h_1}}{\partial w_1}$$

$$= \left(\delta_{o_1}\frac{\partial net_{o_1}}{\partial h_1} + \delta_{o_2}\frac{\partial net_{o_2}}{\partial h_1}\right)\frac{\partial h_1}{\partial net_{h_1}}\frac{\partial net_{h_1}}{\partial w_1}$$

Update the Model Parameters! – 13



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

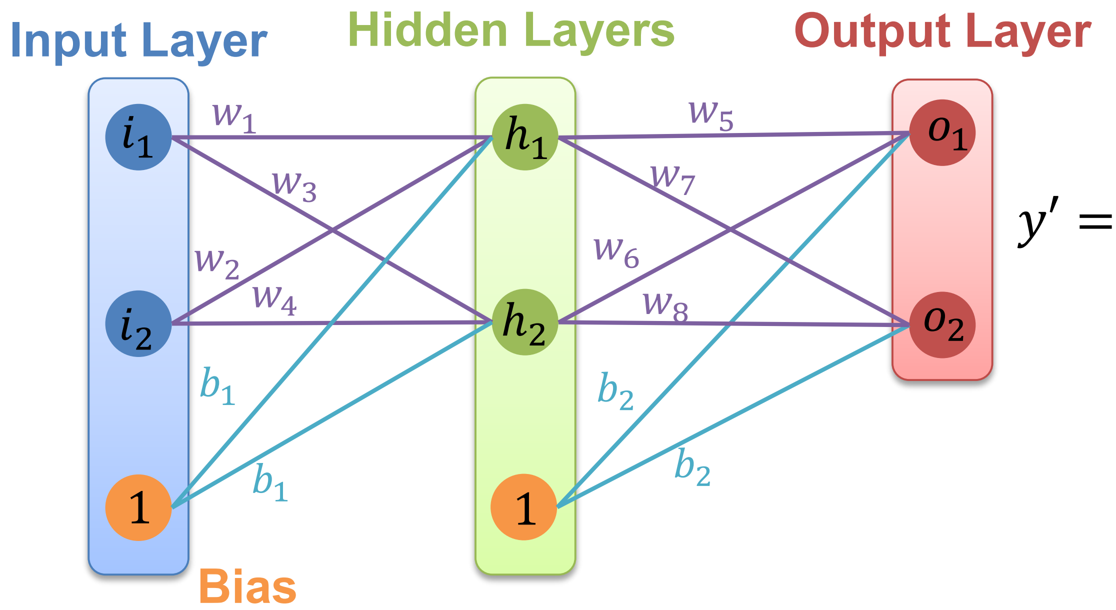
$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial a_n}\frac{\partial a_n}{\partial net_{a_n}} = \delta_{a_n}$$

$$\frac{\partial f}{\partial w_5} = \frac{\partial f}{\partial o_1}\frac{\partial o_1}{\partial net_{o_1}}\frac{\partial net_{o_1}}{\partial w_5} = \left(\frac{1}{2} \times 2 \times (t_1 - o_1) \times (-1)\right)(o_1 \times (1 - o_1))h_1$$

$$= \delta_{o_1}\frac{\partial net_{o_1}}{\partial w_5}$$

Update the Model Parameters! – 14



$$x = \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

$$y' = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} \qquad y = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\frac{\partial f}{\partial a_n} \frac{\partial a_n}{\partial net_{a_n}} = \delta_{a_n}$$

$$\frac{\partial f}{\partial w_5} = \delta_{o_1} \frac{\partial net_{o_1}}{\partial w_5} = \delta_{o_1} h_1$$

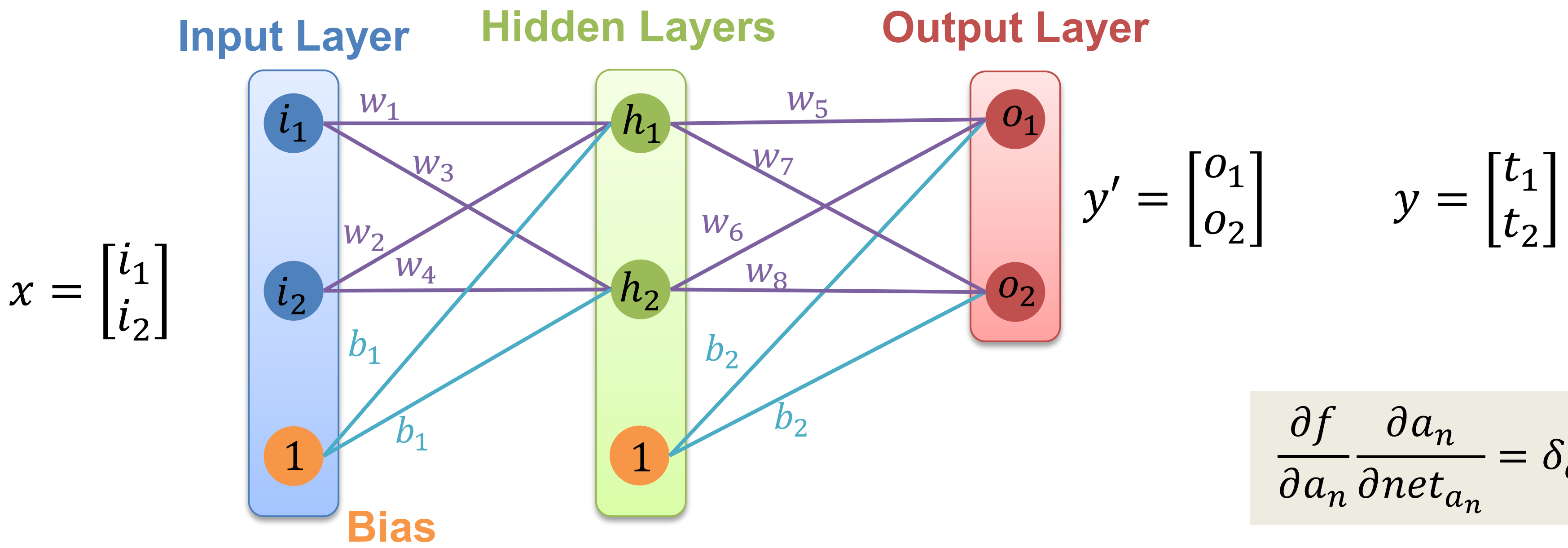
$$\frac{\partial f}{\partial w_1} = \delta_{h_1} \frac{\partial net_{h_1}}{\partial w_1} = \delta_{h_1} i_1$$

$$\frac{\partial f}{\partial w_6} = \delta_{o_1} \frac{\partial net_{o_1}}{\partial w_6} = \delta_{o_1} h_2$$

$$\frac{\partial f}{\partial w_4} = \delta_{h_2} \frac{\partial net_{h_2}}{\partial w_4} = \delta_{h_2} i_2$$

$$\vdots$$

Backpropagation – 1

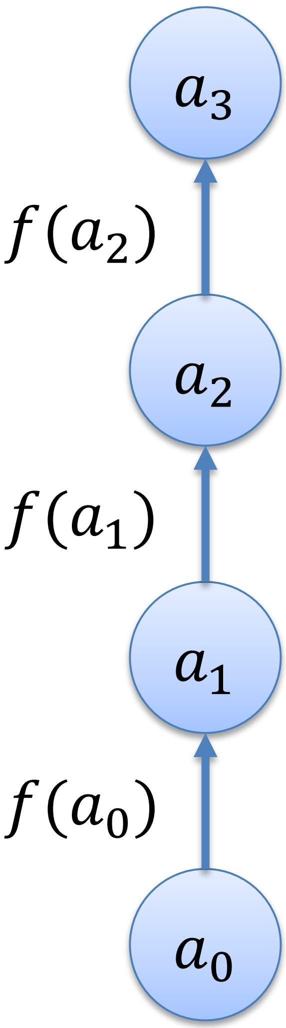


$$\frac{\partial f}{\partial a_n}\frac{\partial a_n}{\partial net_{a_n}} = \delta_{a_n}$$

$$\frac{\partial f}{\partial w_1} = \delta_{h_1}\frac{\partial net_{h_1}}{\partial w_1} = \delta_{h_1} i_1 = \left(\delta_{o_1}\frac{\partial net_{o_1}}{\partial h_1} + \delta_{o_2}\frac{\partial net_{o_2}}{\partial h_1}\right)\frac{\partial h_1}{\partial net_{h_1}} i_1$$

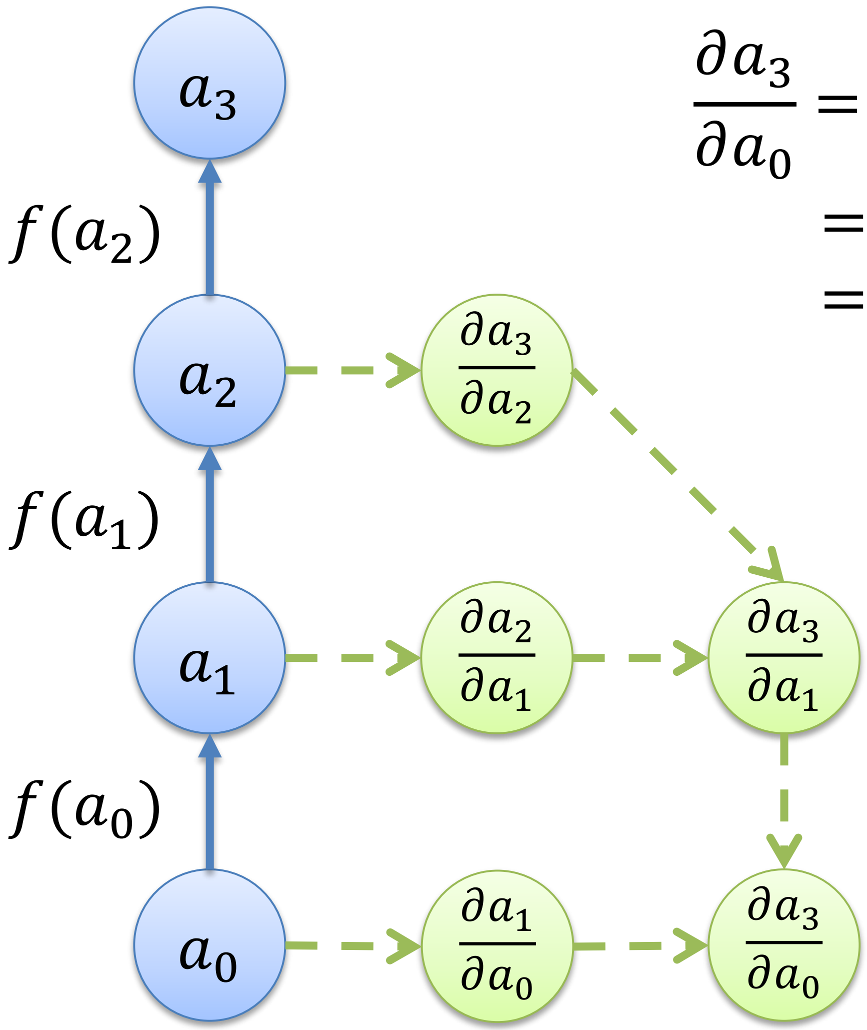
It is a recursive equation!

Backpropagation – 2



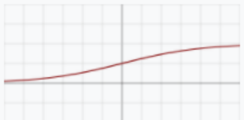
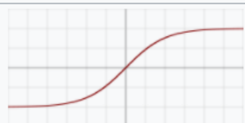
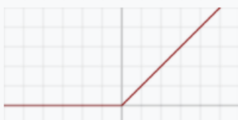
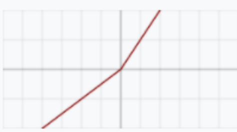
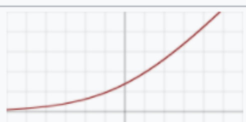
$$
\begin{aligned}
\frac{\partial a_3}{\partial a_0} &= \frac{\partial a_3}{\partial a_2}\frac{\partial a_2}{\partial a_1}\frac{\partial a_1}{\partial a_0} \\
&= f'(a_2)f'(a_1)f'(a_0) \\
&= f'(f(f(a_0))\,)f'(f(a_0)\,)f'(a_0)
\end{aligned}
$$

Backpropagation – 3



$$\frac{\partial a_3}{\partial a_0} = \frac{\partial a_3}{\partial a_2}\frac{\partial a_2}{\partial a_1}\frac{\partial a_1}{\partial a_0}$$
$$= f'(a_2)f'(a_1)f'(a_0)$$
$$= f'(f(f(a_0))\ )f'(f(a_0)\ )f'(a_0)$$

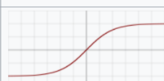
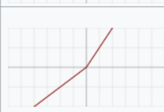
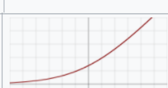
Activation Functions – 1

Name	Plot	Equation	Derivative (with respect to *x*)	Range
Logistic (a.k.a. Sigmoid or Soft step)		$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$ [1]	$f'(x) = f(x)(1-f(x))$	$(0,1)$
TanH		$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$	$f'(x) = 1 - f(x)^2$	$(-1,1)$
Rectified linear unit (ReLU)[10]		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$[0,\infty)$
Leaky rectified linear unit (Leaky ReLU)[11]		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty,\infty)$
SoftPlus[18]		$f(x) = \ln(1+e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$	$(0,\infty)$

Name	Equation	Derivatives	Range
Softmax	$f_i(\vec{x}) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}$ for $i = 1, \ldots, J$	$\frac{\partial f_i(\vec{x})}{\partial x_j} = f_i(\vec{x})(\delta_{ij} - f_j(\vec{x}))$ [6]	$(0,1)$
Maxout[23]	$f(\vec{x}) = \max_i x_i$	$\frac{\partial f}{\partial x_j} = \begin{cases} 1 & \text{for } j = \underset{i}{\operatorname{argmax}}\, x_i \\ 0 & \text{for } j \neq \underset{i}{\operatorname{argmax}}\, x_i \end{cases}$	$(-\infty,\infty)$

Activation Functions – 2

$$\frac{\partial f}{\partial a_n}\frac{\partial a_n}{\partial net_{a_n}} = \delta_{a_n} \qquad a_n = \sigma(net_{a_n}) \qquad \frac{\partial a_n}{\partial net_{a_n}} = \sigma'(net_{a_n})$$

Name	Plot	Equation	Derivative (with respect to *x*)	Range
Logistic (a.k.a. Sigmoid or Soft step)		$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$ [1]	$f'(x) = f(x)(1 - f(x))$	$(0,1)$
TanH		$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$	$f'(x) = 1 - f(x)^2$	$(-1,1)$
Rectified linear unit (ReLU)[10]		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \ge 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \ge 0 \end{cases}$	$[0,\infty)$
Leaky rectified linear unit (Leaky ReLU)[11]		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \ge 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \ge 0 \end{cases}$	$(-\infty,\infty)$
SoftPlus[18]		$f(x) = \ln(1 + e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$	$(0,\infty)$

Name	Equation	Derivatives	Range
Softmax	$f_i(\vec{x}) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}$ for i = 1, …, J	$\frac{\partial f_i(\vec{x})}{\partial x_j} = f_i(\vec{x})(\delta_{ij} - f_j(\vec{x}))$ [6]	$(0,1)$
Maxout[23]	$f(\vec{x}) = \max_i x_i$	$\frac{\partial f}{\partial x_j} = \begin{cases} 1 & \text{for } j = \underset{i}{\operatorname{argmax}}\, x_i \\ 0 & \text{for } j \ne \underset{i}{\operatorname{argmax}}\, x_i \end{cases}$	$(-\infty,\infty)$

Loss Functions

$f(x) = y'$

- Squared Loss

$$L_{SL} = \sum_{n=1}^{N} (y_n - y_n')^2$$

- Cross-entropy Loss
  - Usually paired with softmax

$$L_{CE} = \sum_{n=1}^{N} -y_n \log(y_n')$$

- Logistic Loss
- Hinge Loss
- Absolute Loss

Squared Loss & Cross Entropy Loss

$$y'_1 = \begin{bmatrix} 0.4 \\ 0.5 \\ 0.1 \end{bmatrix} \qquad y'_2 = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix} \qquad y = \begin{bmatrix} 0.3 \\ 0.5 \\ 0.2 \end{bmatrix}$$

- Squared Loss

$$L_{SL} = \sum_{n=1}^{N} (y_n - y'_n)^2$$

$$L_{SL}(y, y'_1) = (0.3 - 0.4)^2 + (0.5 - 0.5)^2 + (0.2 - 0.1)^2 = 0.01 + 0.01 = 0.02$$

$$L_{SL}(y, y'_2) = (0.3 - 0.2)^2 + (0.5 - 0.5)^2 + (0.2 - 0.3)^2 = 0.01 + 0.01 = 0.02$$

- Cross Entropy Loss
  - **Preserving the order?**

$$L_{CE} = \sum_{n=1}^{N} -y_n \log(y'_n)$$

$$L_{CE}(y, y'_1) = \big(-0.3\log(0.4)\big) + \big(-0.5\log(0.5)\big) + \big(-0.2\log(0.1)\big) = 1.08198$$

$$L_{CE}(y, y'_2) = \big(-0.3\log(0.2)\big) + \big(-0.5\log(0.5)\big) + \big(-0.2\log(0.3)\big) = 1.0702$$

Mini-Batch

Given a set of trining samples $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$

epoch

for 1 to M

  Select a subset of samples without replacement

    Do forward propogation

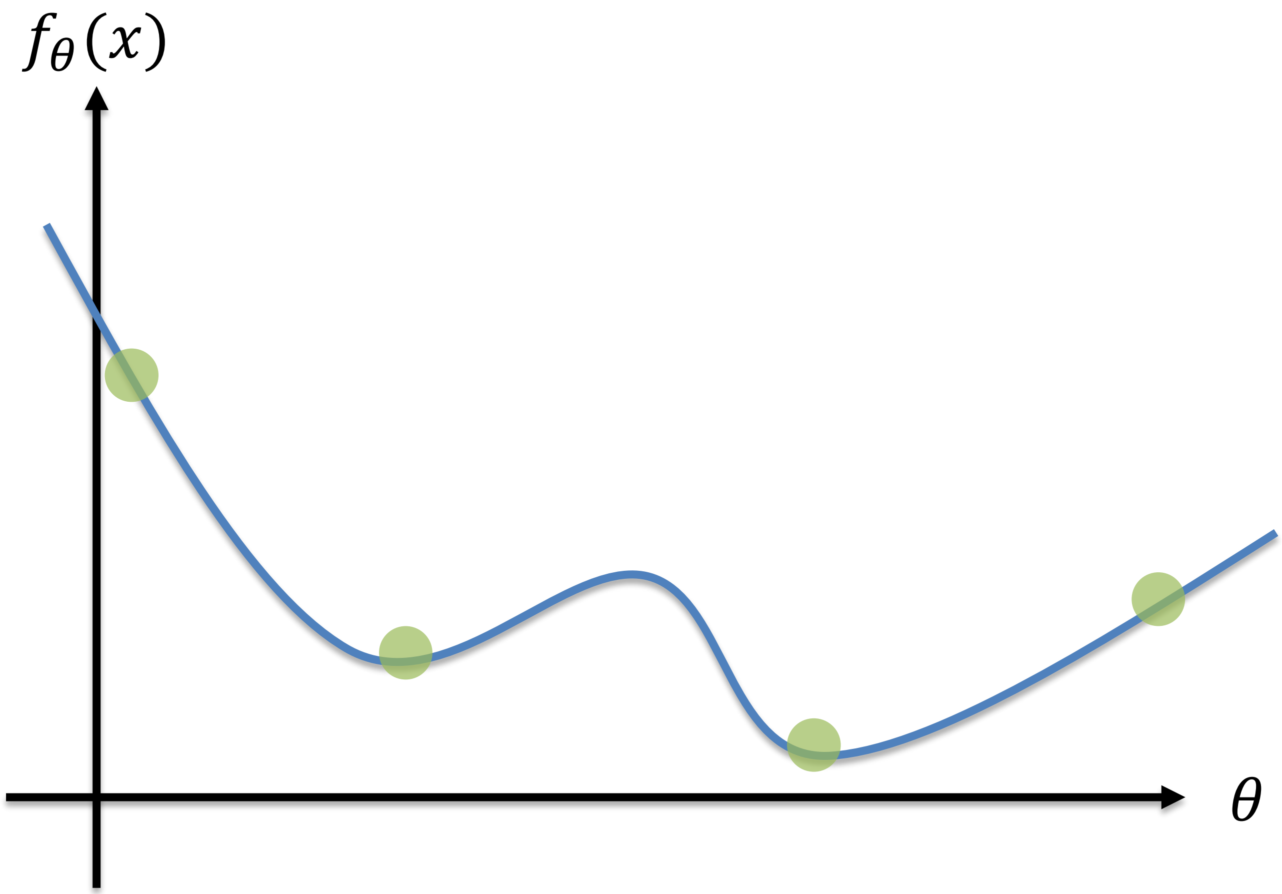
    Calculate errors

    Update model parameters

mini-batch

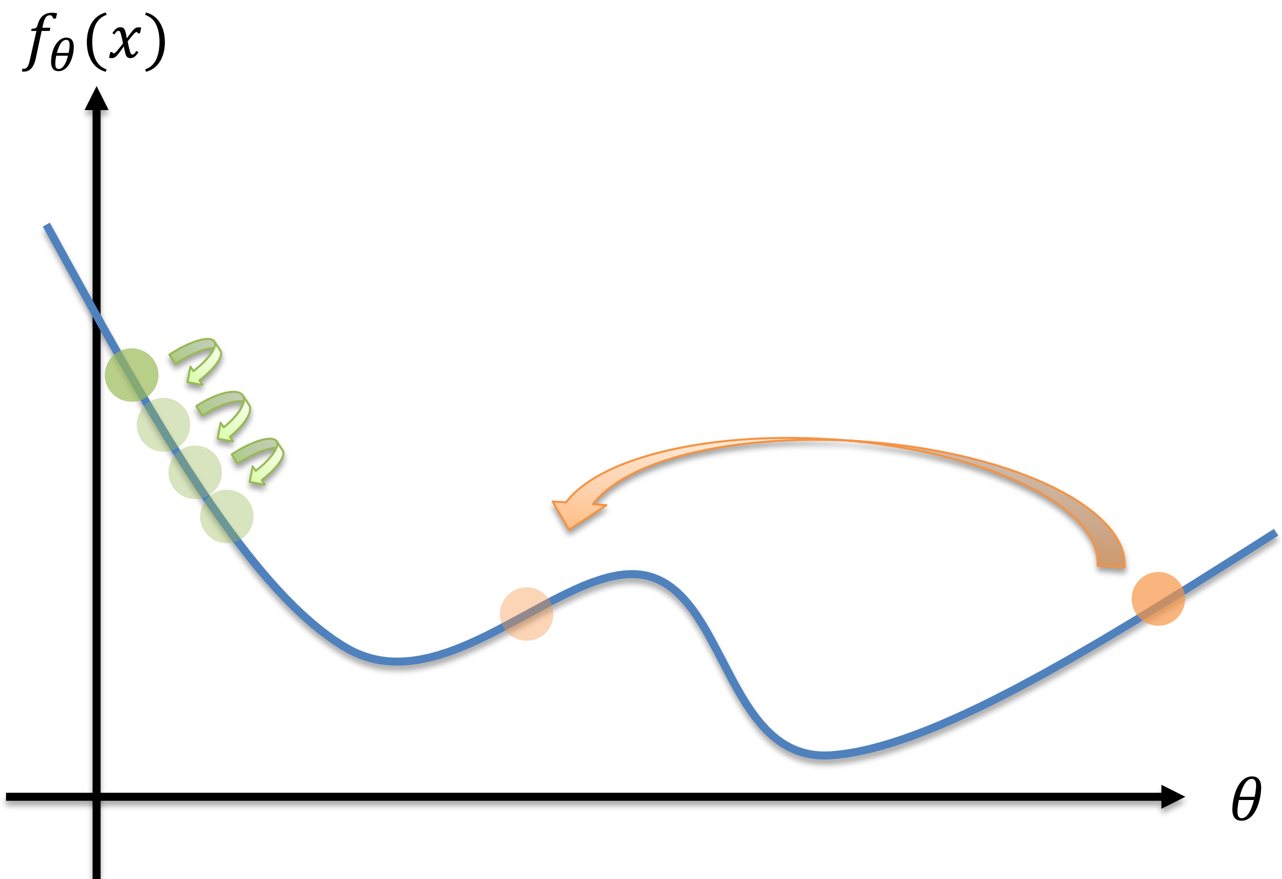
Initialization & Step Size – 1

- Initialization is the beginning



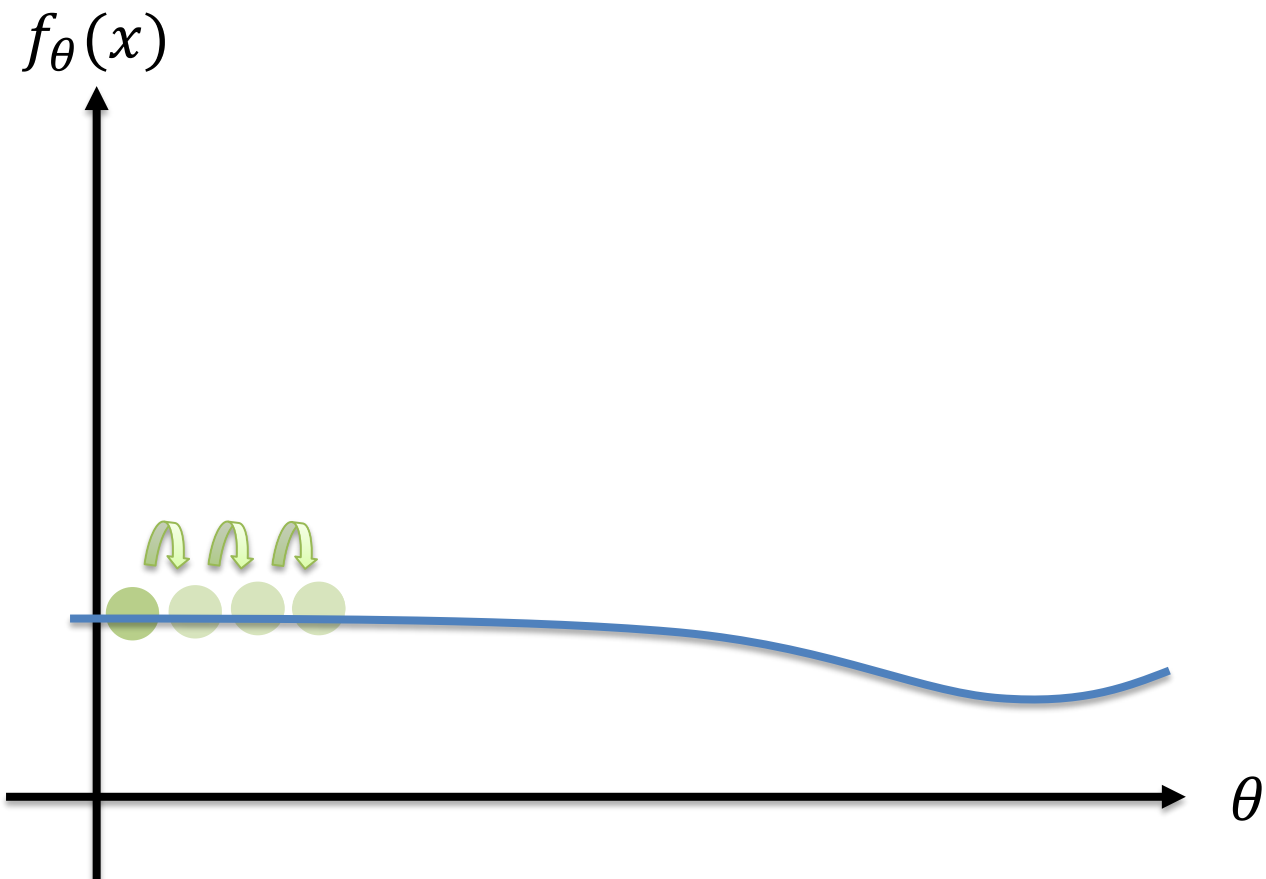
Initialization & Step Size – 2

- Small step size: slow convergence
- Large step size: hard to converge



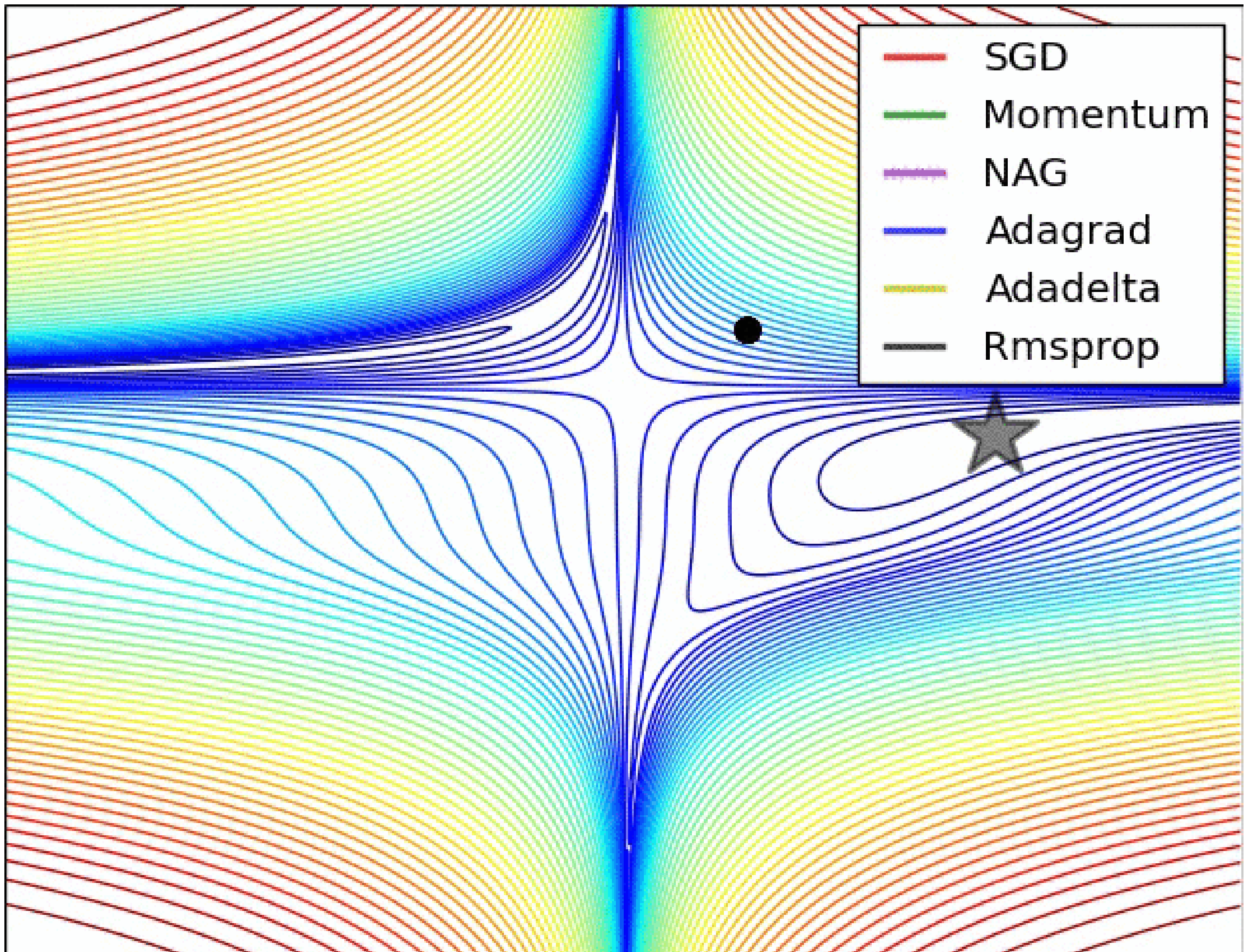
Initialization & Step Size – 3

- Small step size: slow convergence
- Large step size: hard to converge



Optimizers

- SGD
- RMSprop
- Adagrad
- Adadelta
- Adam



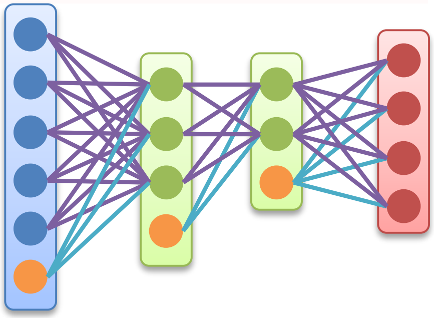
Hard to Learn but Easy to Do

```python
from keras.layers import Input, Dense
from keras.models import Model

# This returns a tensor
inputs = Input(shape=(784,))

# a layer instance is callable on a tensor, and returns a tensor
x = Dense(64, activation='relu')(inputs)
x = Dense(64, activation='relu')(x)
predictions = Dense(10, activation='softmax')(x)

# This creates a model that includes
# the Input layer and three Dense layers
model = Model(inputs=inputs, outputs=predictions)
model.compile(optimizer='rmsprop',
              loss='categorical_crossentropy',
              metrics=['accuracy'])
model.fit(data, labels)  # starts training
```



Questions?

kychen@mail.ntust.edu.tw