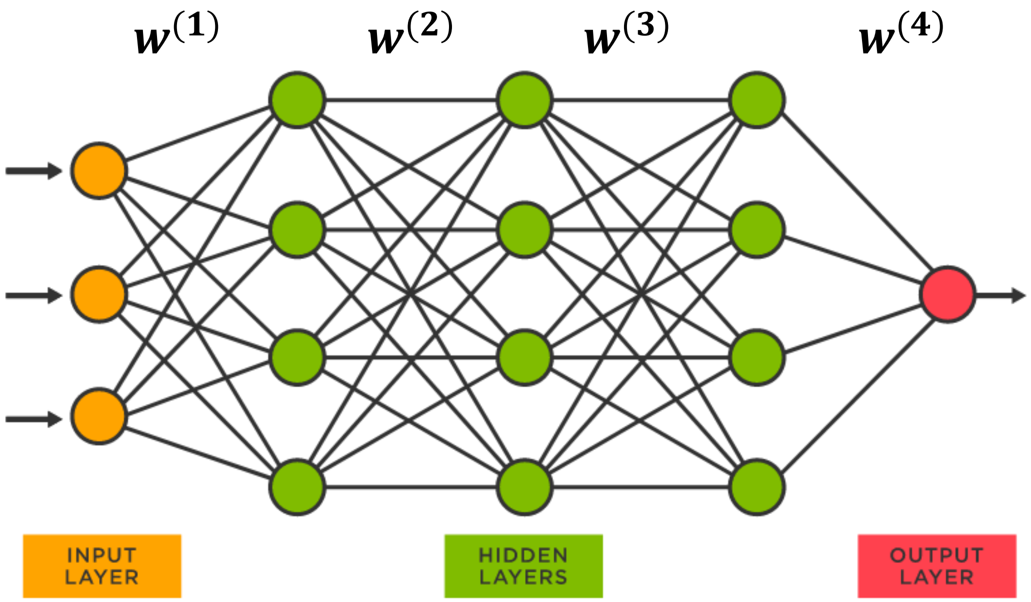
딥러닝을 활용한 **디지털 영상 처리**
Digital Image Processing via Deep Learning

Lecture 5 – Backpropagation

From Last Lecture

$$\boldsymbol{w}^{(z)} = \begin{pmatrix} w_{b1}^{1} & w_{11}^{1} & w_{21}^{1} & \cdots & w_{n1}^{1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{bp}^{1} & w_{1p}^{1} & w_{2p}^{1} & \cdots & w_{np}^{1} \end{pmatrix}$$

$$\boldsymbol{w}^{(1)} = \begin{pmatrix} w_{b1}^{1} & w_{11}^{1} & w_{21}^{1} & w_{31}^{1} \\ w_{b2}^{1} & w_{12}^{1} & w_{22}^{1} & w_{32}^{1} \\ w_{b3}^{1} & w_{13}^{1} & w_{23}^{1} & w_{33}^{1} \\ w_{b4}^{1} & w_{14}^{1} & w_{24}^{1} & w_{34}^{1} \end{pmatrix}$$



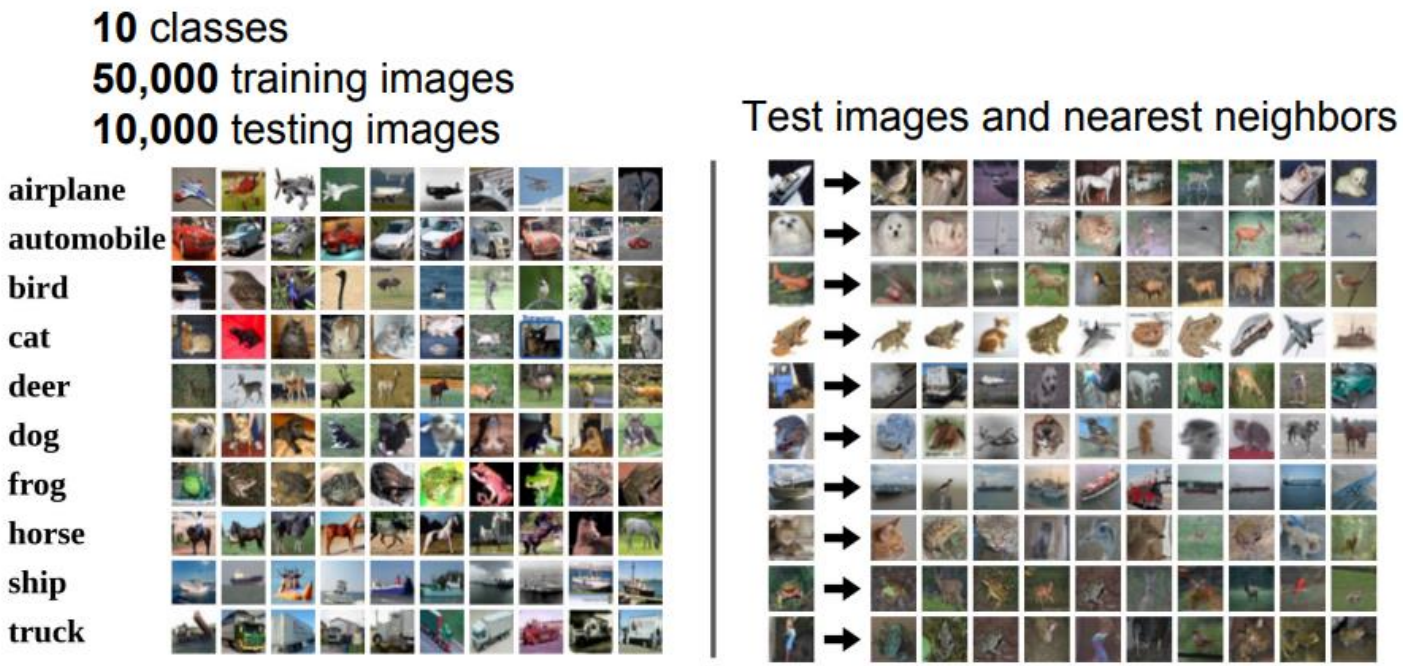
We usually have weight parameters in a scale of millions. 60 in the example on right.

From Last Lecture
Learning from data

We are living in a data era.

We do not set weights by ourselves.

The Neural Network Learn from data.



From Last Lecture
Finding optimal weights – Loss function

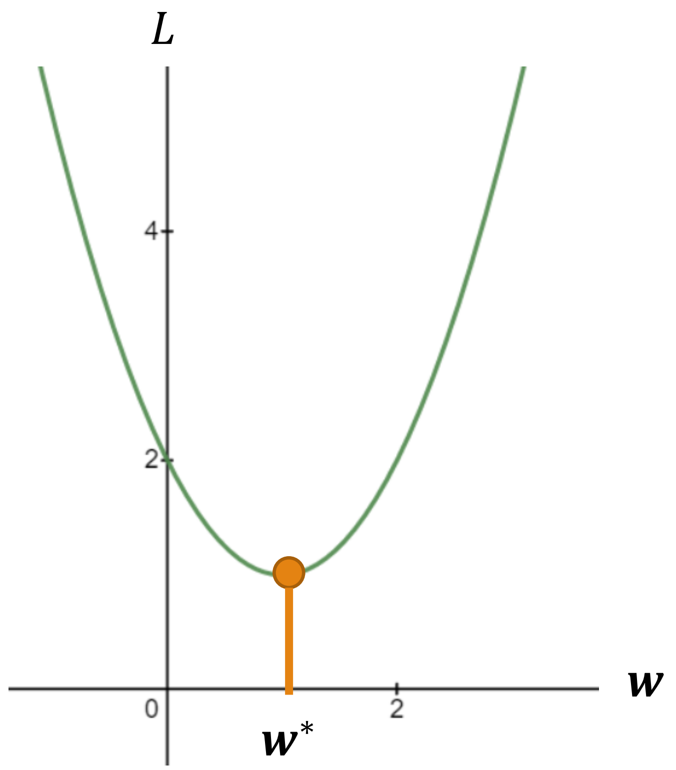
Consider: $L(\boldsymbol{x}, \boldsymbol{w}) = (y - t)^2 = (f(\boldsymbol{x}, \boldsymbol{w}) - t)^2$

Our task is finding weights, $\boldsymbol{w}$, that minimizes the loss, L.

From intuition, we can simply find the optimal weights is finding the turning points of the loss function by differentiating it.

$$\boldsymbol{w}^* = \boldsymbol{argmin_w}(L(\boldsymbol{x}, \boldsymbol{w}))$$

$$0 = \frac{\partial L}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}=\boldsymbol{w}^*}$$



From Last Lecture
Gradient Descent

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$L = w^2 - 2w + 2$

Let's start with $w_0 = 3$

Then, $L = 5$

$$\frac{\partial L}{\partial w} = 2w - 2$$

$$\frac{\partial L}{\partial w}\Big|_{w=3} = 4$$

$$w_1 = 3 - 0.1 * 4 = 2.6$$

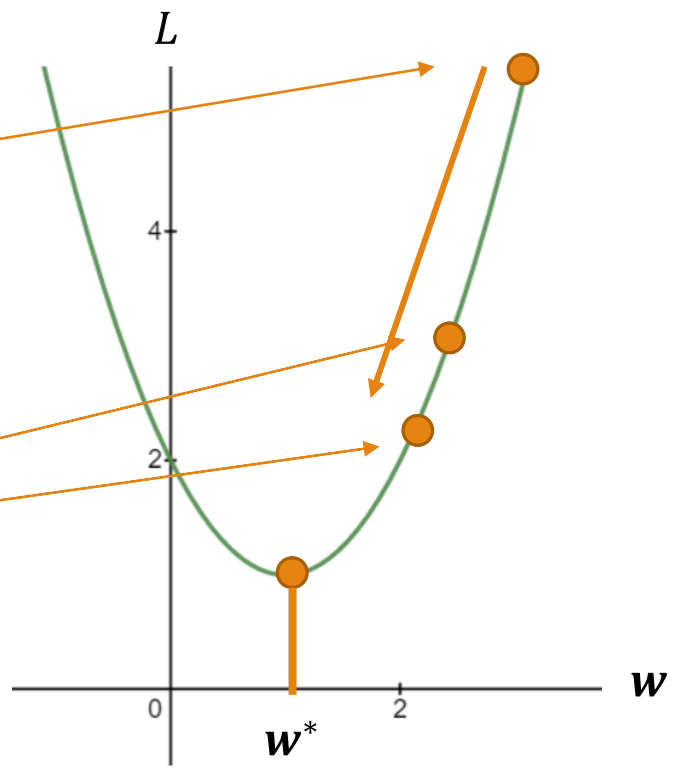
$L = w^2 - 2w + 2$

Repeat with $w_1 = 2.6$

Then, $L = 3.56$

$$\frac{\partial L}{\partial w}\Big|_{w=2.6} = 3.2$$

$$w_2 = 2.6 - 0.1 * 3.2 = 2.28$$



Here, η is called the learning rate. It determines how fast the learning will take place. In the example above, it is set to 0.1

From Last Lecture
Minibatch

Computing Loss with only a single data is very inefficient.

We usually take a batch of data and compute the mean loss.

Take Sum of Squares for Error as an example, where B denotes the batch size.

$$L = \frac{1}{|B|} \sum_{B} \frac{1}{2} \sum_{k} (y_k - t_k)^2$$

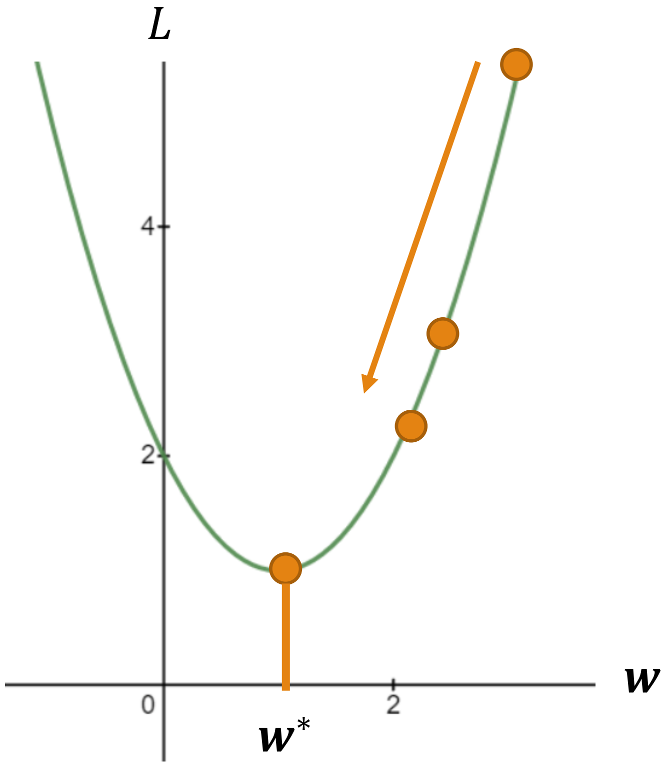
We call this approach a “Minibatch”, due to a small batch size being more efficient in training.

From Last Lecture
Stochastic Gradient Descent

Stochastic Gradient Descent: Applying mean loss of randomly sampled minibatch data and performing gradient descent algorithm.

$$L = \frac{1}{|B|}\sum_{B}\frac{1}{2}\sum_{k}(y_k - t_k)^2$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta\frac{\partial L}{\partial \boldsymbol{w}}$$



From Last Lecture Algorithm

Algorithm: Stochastic Gradient Descent

Input: Training data D ={X,Y}, epoch e, learning rate η, stop threshold τ

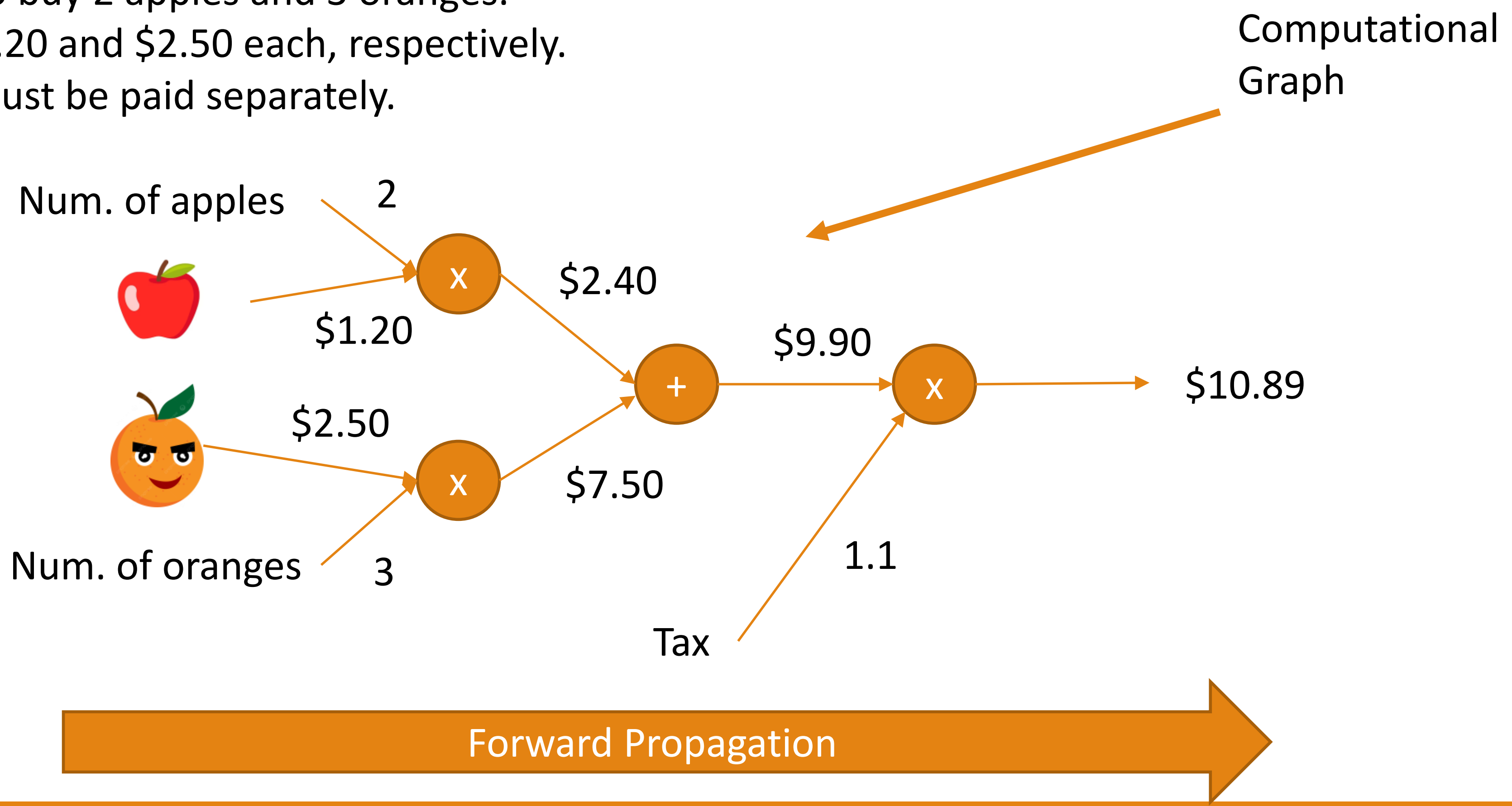
Output: Optimum weight matric, $\boldsymbol{w}^*$

1. **Initialize** $\boldsymbol{w_0}$ with random numbers
2. For i=1,2,3,…,e repeat
    3. **Sample minibatch** data D_M from D
    4. **Compute** $\boldsymbol{y}$ via forward propagation → Forward Propagation
    5. **Compute loss,** $\boldsymbol{L}$
        6. If L is below τ break
    7. **Compute** $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{w}}$ → Backward Propagation
    8. $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \eta \frac{\partial L}{\partial \boldsymbol{w}}$

9. Return $\boldsymbol{w_e}$

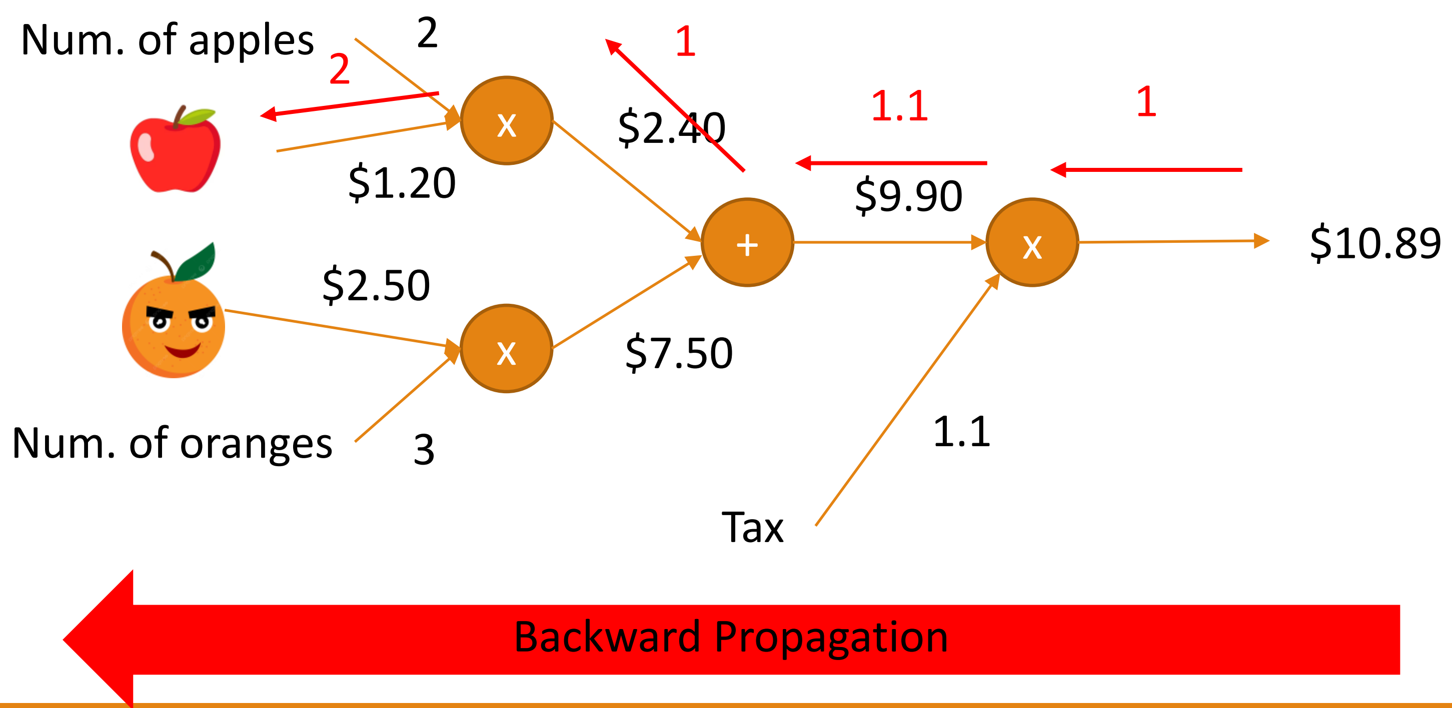
Backward Propagation
A Simple Example

Arin wants to buy 2 apples and 3 oranges.
They cost \$1.20 and \$2.50 each, respectively.
10% of tax must be paid separately.



Backward Propagation
A Simple Example

How sensitive is the overall price to the price of an apple?
Say, if the price of an apple changes to $2.2, how much will the overall price change?
Think of this problem as a differentiation problem!!



Backward Propagation
A Simple Example

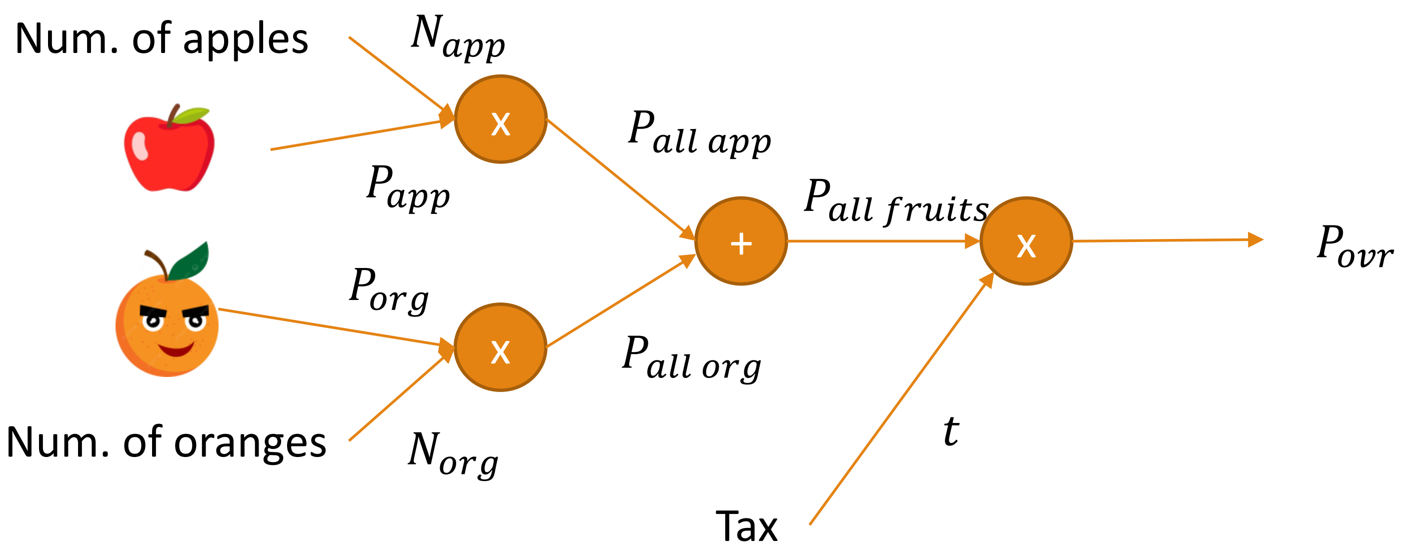
Using Chain rule to back propagate;

$$P_{ovr} = t\left(N_{app}P_{app} + N_{org}P_{org}\right)$$

$$\frac{\partial P_{ovr}}{\partial P_{app}} = tN_{app}$$

$$P_{ovr} = tP_{all\ fruits}$$
$$P_{all\ fruits} = N_{app}P_{app} + N_{org}P_{org}$$

$$\frac{\partial P_{ovr}}{\partial P_{app}} = \frac{\partial P_{ovr}}{\partial P_{all\ fruits}} \frac{\partial P_{all\ fruits}}{\partial P_{app}} = tN_{app}$$

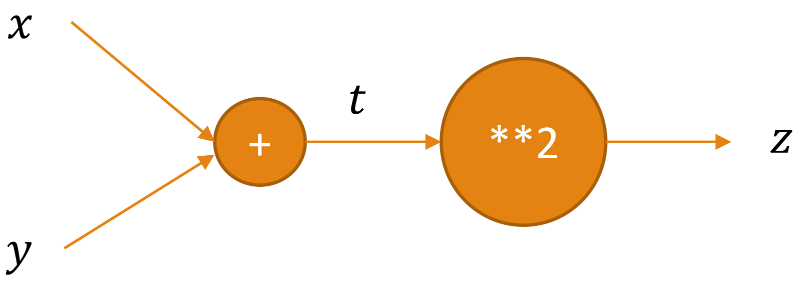


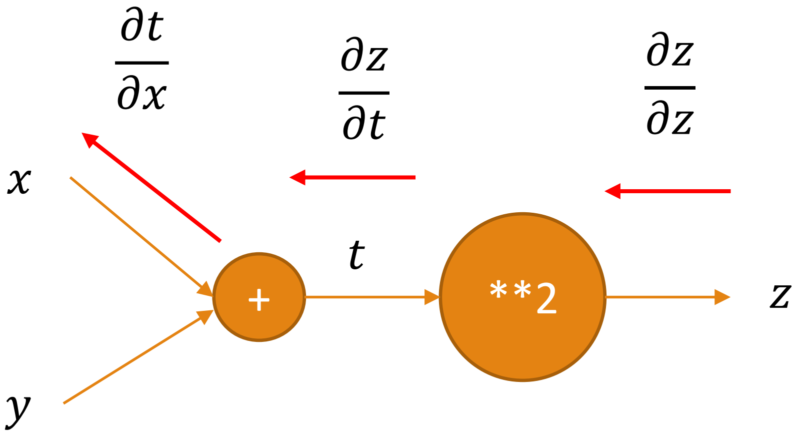
Backward Propagation
A Simple Example 2

Consider;

$$z = t^2$$
$$t = x + y$$

Sketch the Computational Graph of the equation system.
Compute forward and backward propagation of x.





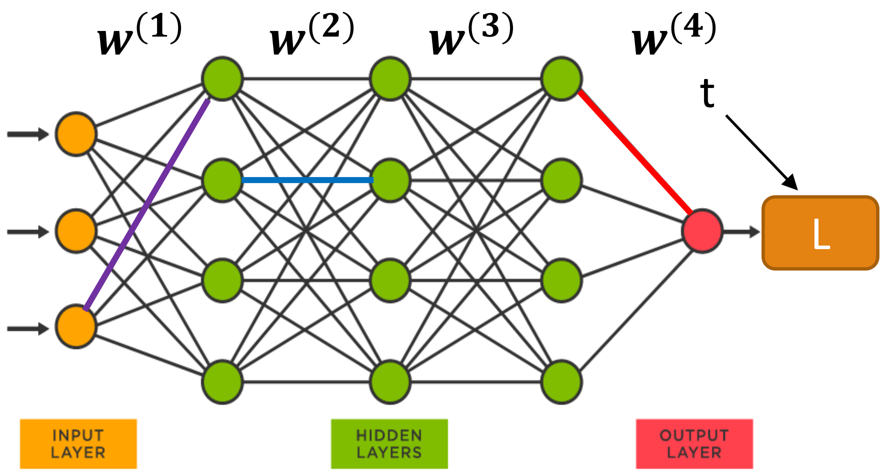
$$\boxed{\frac{\partial z}{\partial x} = \frac{\partial z}{\partial z}\frac{\partial z}{\partial t}\frac{\partial t}{\partial x} = \frac{\partial z}{\partial t}\frac{\partial t}{\partial x} = 2t = 2(x + y)}$$

Intuition

$$L = \frac{1}{|B|}\sum_{B}\frac{1}{2}\sum_{k}(y_k - t_k)^2$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta\frac{\partial L}{\partial \boldsymbol{w}}$$

$\frac{\partial L}{\partial w_{ij}^{z}}$ tells us how sensitive the loss is to a specific weight parameter

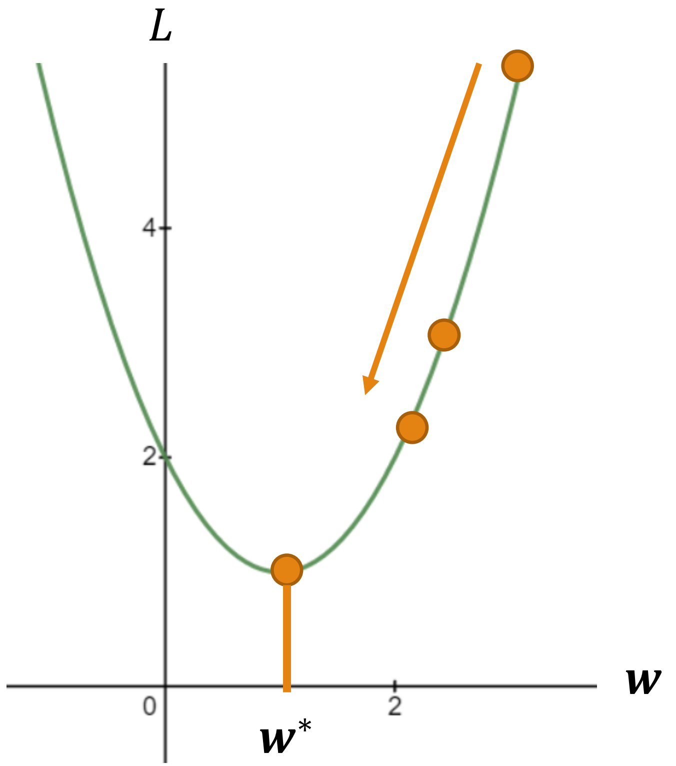


$$\boxed{\frac{\partial L}{\partial w_{31}^{1}} = 1.5}$$

The most sensitive weight!

$$\frac{\partial L}{\partial w_{22}^{2}} = 0.9$$

$$\frac{\partial L}{\partial w_{1y}^{4}} = -0.4$$



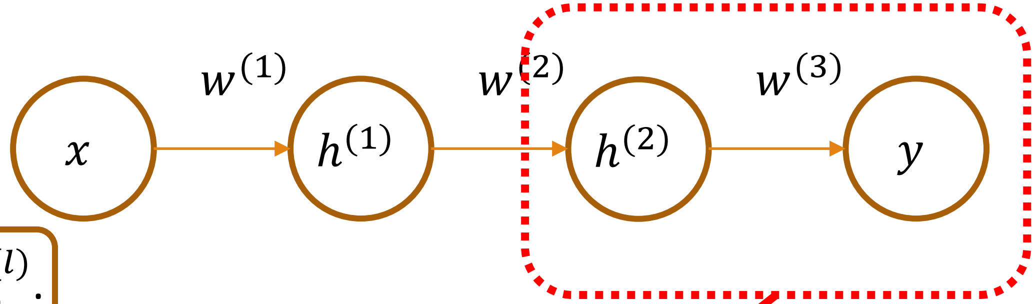
Simple Neural Network Example

Let the activation function, $a(\cdot)$, be identical for all layers.

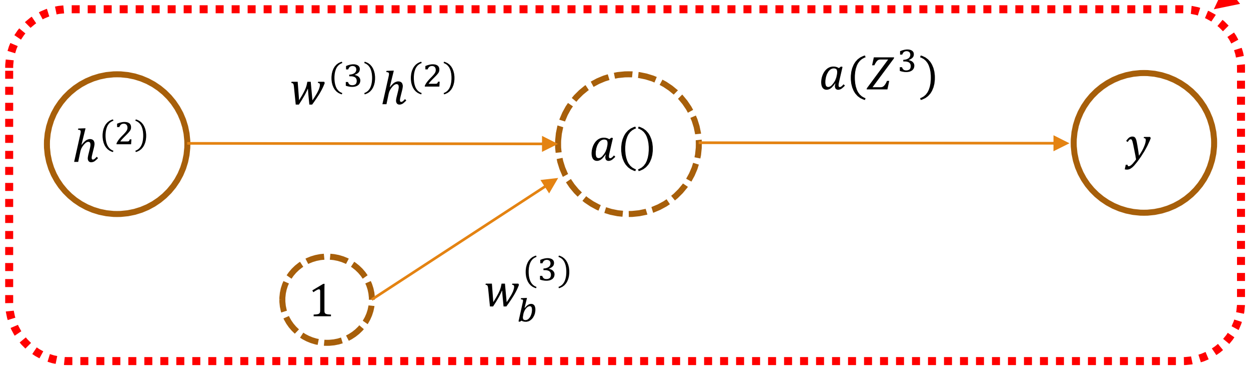
Let the Loss function be $L = \frac{1}{2}(y - t)^2$, where t is the data label.

Find:

1. $\frac{\partial L}{\partial w^{(3)}}$ and $\frac{\partial L}{\partial w^{(1)}}$ wrt $\frac{\partial h^{(l)}}{\partial Z^{(l)}} = a'\left(Z^{(l)}\right)$, where $\boxed{Z^{(l)} = w^{(l)} h^{(l-1)} + w_b^{(l)}.}$



A part of the diagram above in more detail



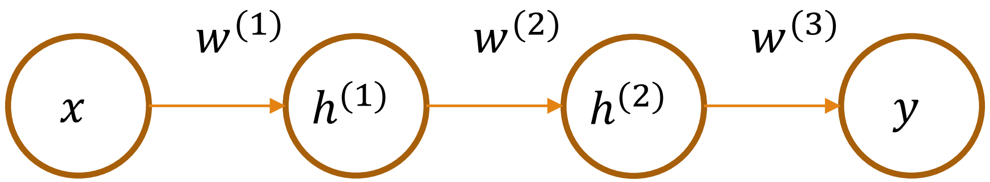
Simple Neural Network Example

Let the activation function, $a(\cdot)$, be identical for all layers.

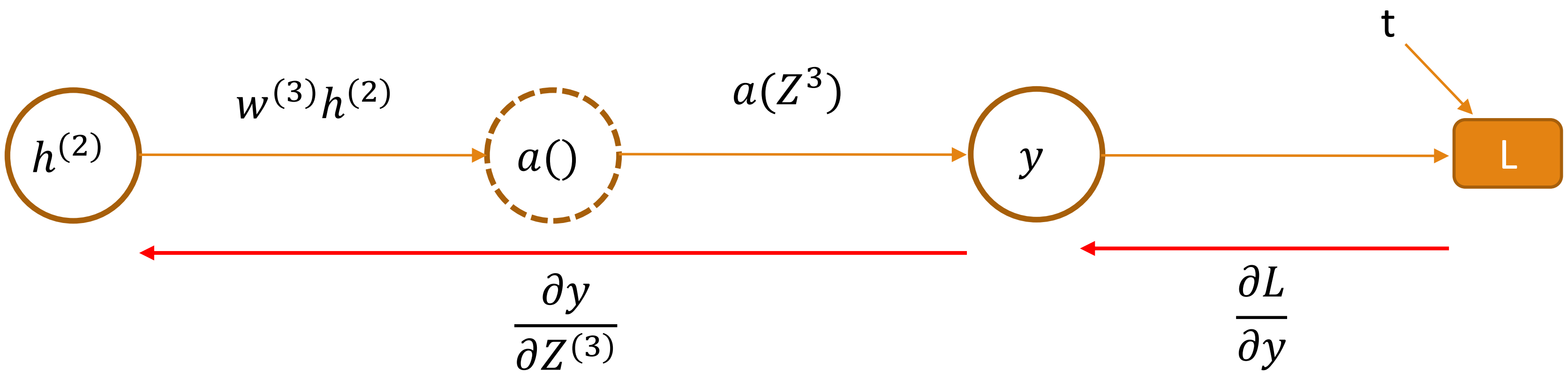
Let the Loss function be $L = \frac{1}{2}(y - t)^2$, where t is the data label.

Find:

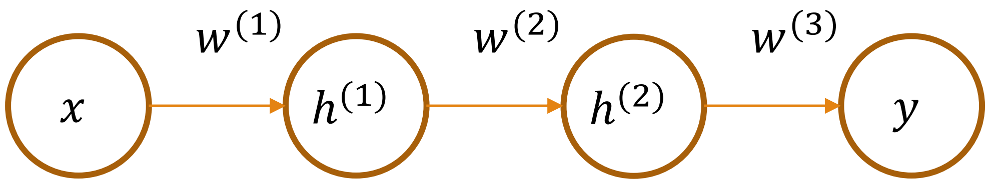
1. $\frac{\partial L}{\partial w^{(3)}}$ and $\frac{\partial L}{\partial w^{(1)}}$ wrt $\frac{\partial h^{(l)}}{\partial Z^{(l)}} = a'\left(Z^{(l)}\right)$, where $Z^{(l)} = w^{(l)} h^{(l-1)} + w_b^{(l)}$.



A part of the diagram above in more detail



Simple Neural Network Example



Let the activation function, $a(\cdot)$, be identical for all layers.

Let the Loss function be $L = \frac{1}{2}(y - t)^2$, where t is the data label.

Find:

1. $\frac{\partial L}{\partial w^{(3)}}$ and $\frac{\partial L}{\partial w^{(1)}}$ wrt $\frac{\partial h^{(l)}}{\partial Z^{(l)}} = a'\left(Z^{(l)}\right)$, where $Z^{(l)} = w^{(l)}h^{(l-1)} + w_b^{(l)}$.

$$y = a\left(w^{(3)}h^{(2)} + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}h^{(1)} + w_b^{(2)}\right) + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}a(w^{(1)}x + w_b^{(1)}) + w_b^{(2)}\right) + w_b^{(3)}\right)$$

$$\frac{\partial L}{\partial y} = (y - t)$$

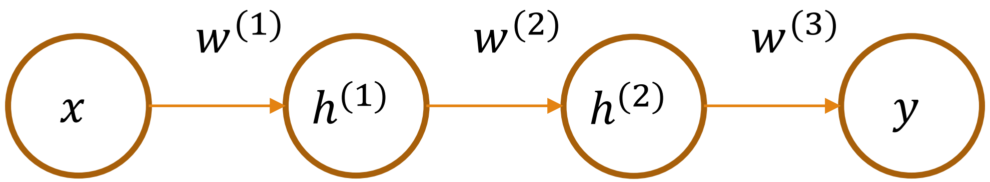
$$\frac{\partial L}{\partial w^{(3)}} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial Z^{(3)}}\frac{\partial Z^{(3)}}{\partial w^{(3)}} = (y - t)\frac{\partial y}{\partial Z^{(3)}}h^{(2)} = (y - t)a'\left(Z^{(3)}\right)h^{(2)}$$

Simple Neural Network Example

Let the activation function, $a(\cdot)$, be identical for all layers.

Let the Loss function be $L = \frac{1}{2}(y - t)^2$, where t is the data label.

Find:



1. $\frac{\partial L}{\partial w^{(3)}}$ and $\frac{\partial L}{\partial w^{(1)}}$ wrt $\frac{\partial h^{(l)}}{\partial Z^{(l)}} = a'\left(Z^{(l)}\right)$, where $Z^{(l)} = w^{(l)}h^{(l-1)} + w_b^{(l)}$.

$$y = a\left(w^{(3)}h^{(2)} + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}h^{(1)} + w_b^{(2)}\right) + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}a(w^{(1)}x + w_b^{(1)}) + w_b^{(2)}\right) + w_b^{(3)}\right)$$

$$\frac{\partial L}{\partial y} = (y - t)$$

$$\frac{\partial L}{\partial w^{(1)}} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial Z^{(3)}}\frac{\partial Z^{(3)}}{\partial h^{(2)}}\frac{\partial h^{(2)}}{\partial Z^{(2)}}\frac{\partial Z^{(2)}}{\partial h^{(1)}}\frac{\partial h^{(1)}}{\partial Z^{(1)}}\frac{\partial Z^{(1)}}{\partial w^{(1)}} = (y - t)a'\left(Z^{(3)}\right)w^{(3)}a'\left(Z^{(2)}\right)w^{(2)}a'\left(Z^{(1)}\right)x$$

From Last Lecture
Activation Functions

Step Function:

$$a(x) = \begin{cases} 1, & x \geq 0 \\ \beta, & x < 0 \end{cases}$$

Where $\beta = 0$ or $\beta = -1$

Logistic Sigmoid:

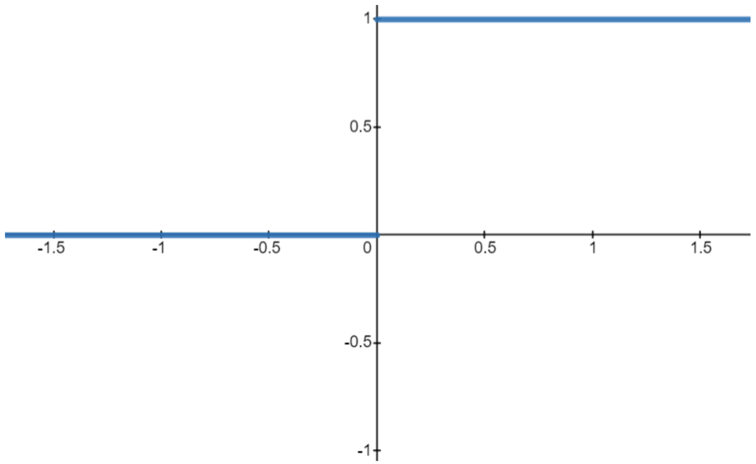
$$a(x) = \frac{1}{1 + e^{-\beta x}}$$

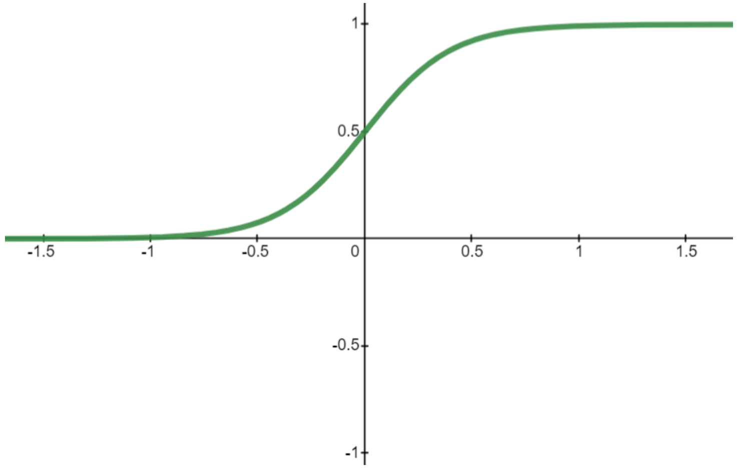
Where $\beta > 0$

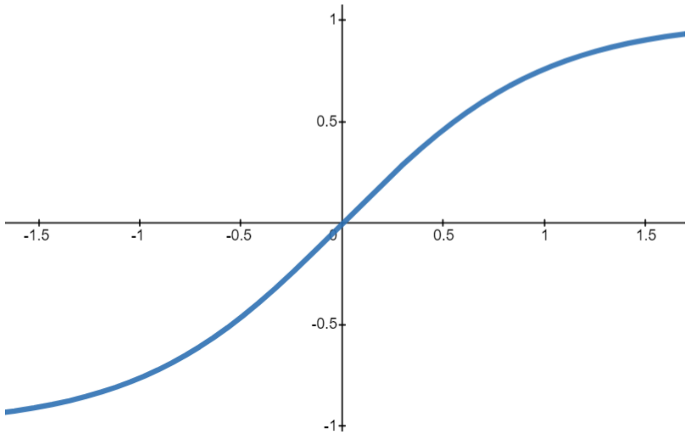
Hyperbolic Tangent:

$$a(x) = \tanh \beta x$$

Where $\beta > 0$







Activation Functions – Gradients

Step Function:

$$a(x) = \begin{cases} 1, & x \geq 0 \\ \beta, & x < 0 \end{cases}$$

Where $\beta = 0$ or $\beta = -1$

Logistic Sigmoid:

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

Where $\beta > 0$

Hyperbolic Tangent:

$$a(x) = \tanh \beta x$$

Where $\beta > 0$

Gradient:

$$a'(x) = \begin{cases} 0, & x \neq 0 \\ inf, & otherwise \end{cases}$$

Gradient:

$$a'(x) = \beta a(x)(1 - a(x))$$

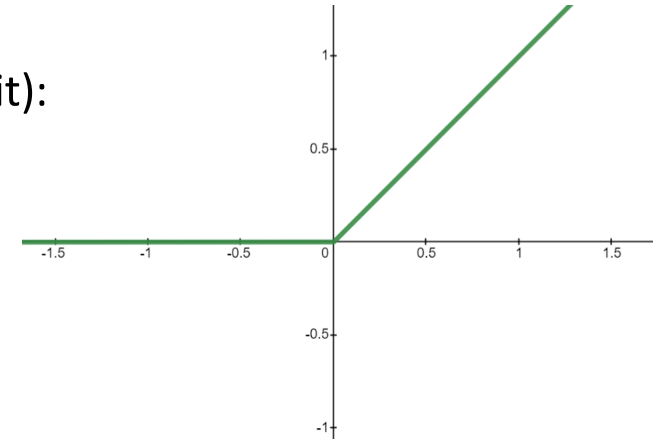
Gradient:

$$a'(x) = \beta \operatorname{sech}^2 \beta x$$

From Last Lecture
Activation Functions

ReLU(Rectified Linear Unit):

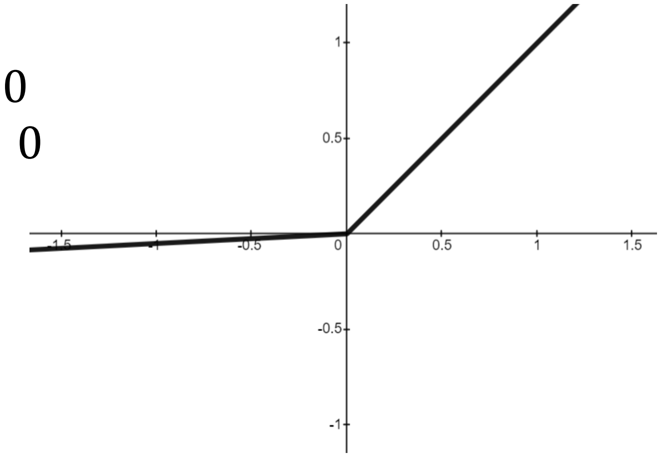
$$a(x) = \max(0, x)$$



Leaky ReLU:

$$a(x) = \begin{cases} x, & x \geq 0 \\ \beta x, & x < 0 \end{cases}$$

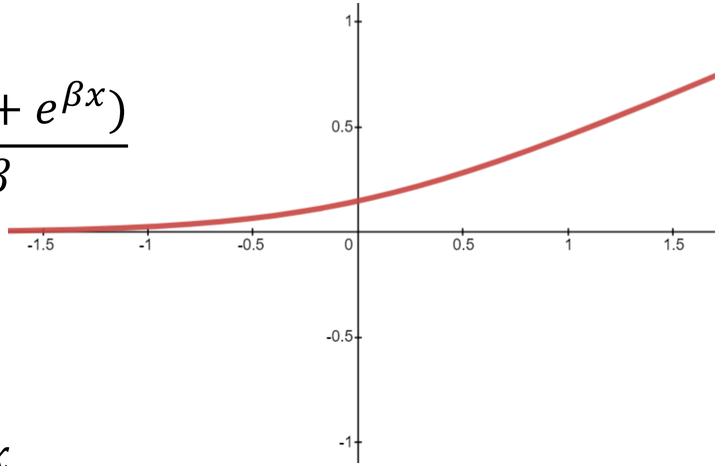
Where $0 < \beta \ll 1$



Softplus:

$$a(x) = \frac{\log(1 + e^{\beta x})}{\beta}$$

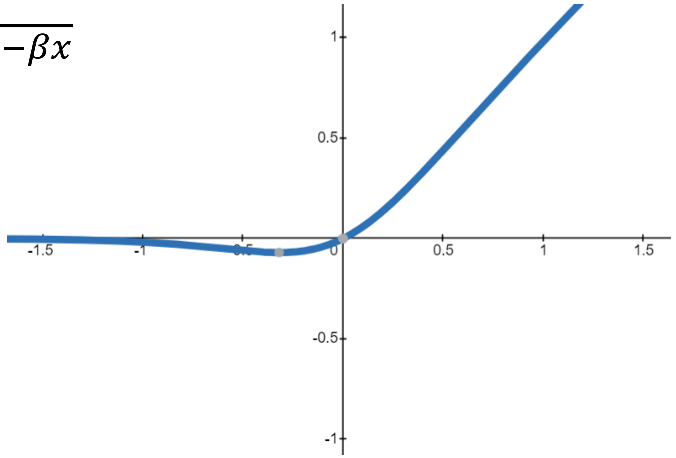
Where $\beta > 0$



Swish:

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

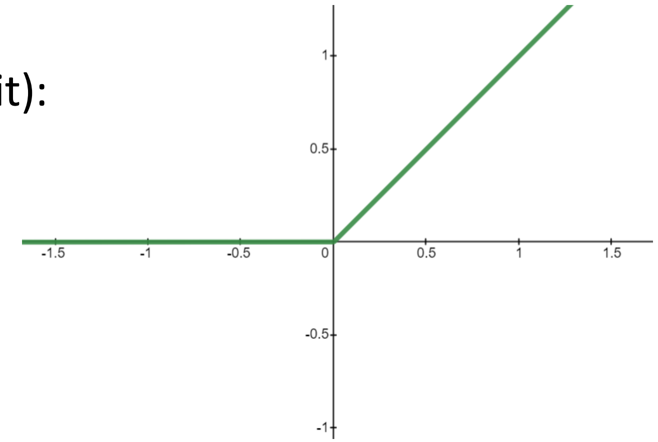
Where $\beta > 0$



Activation Functions – Gradients

ReLU(Rectified Linear Unit):

$$a(x) = \max(0, x)$$



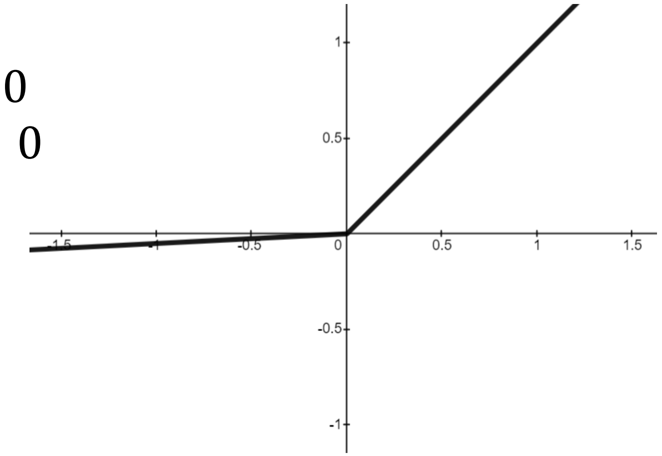
Gradient:

$$a'(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Leaky ReLU:

$$a(x) = \begin{cases} x, & x \geq 0 \\ \beta x, & x < 0 \end{cases}$$

Where $0 < \beta \ll 1$



Gradient:

$$a'(x) = \begin{cases} 1, & x \geq 0 \\ \beta, & x < 0 \end{cases}$$

Activation Functions

Step Function:

$$a(x) = \begin{cases} 1, & x \geq 0 \\ \beta, & x < 0 \end{cases}$$

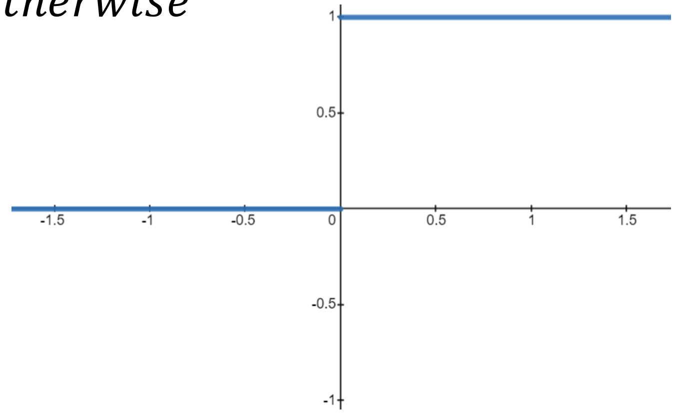
Where $\beta = 0$ or $\beta = -1$

Gradient:

$$a'(x) = \begin{cases} 0, & x \neq 0 \\ inf, & otherwise \end{cases}$$

Is Never Used.

0 Gradient makes the gradient matrix 0.



ReLU(Rectified Linear Unit):

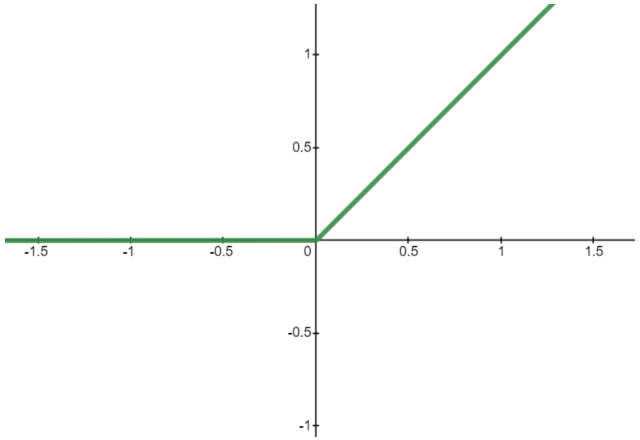
$$a(x) = \max(0, x)$$

Gradient:

$$a'(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Is Widely Used.

Gradient, 1 is very simple and easy for computation.



Simple Neural Network Example

Let the activation function, $a(\cdot)$, be the **ReLU function for all layers**.

Let the Loss function be $L = \frac{1}{2}(y-t)^2$, where t is the data label.

Find:

1. $\frac{\partial L}{\partial w^{(3)}}$ and $\frac{\partial L}{\partial w^{(1)}}$.

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

$$y = a\left(w^{(3)}h^{(2)} + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}h^{(1)} + w_b^{(2)}\right) + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}a(w^{(1)}x + w_b^{(1)}) + w_b^{(2)}\right) + w_b^{(3)}\right)$$

$$\frac{\partial L}{\partial y} = (y-t), \quad a'\left(Z^{(l)}\right) = 1 \text{ assuming positive } Z^{(l)}$$

$$\frac{\partial L}{\partial w^{(3)}} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial Z^{(3)}}\frac{\partial Z^{(3)}}{\partial w^{(3)}} = (y-t)\frac{\partial y}{\partial Z^{(3)}}h^{(2)} = (y-t)a'\left(Z^{(3)}\right)h^{(2)} = \boxed{(y-t)h^{(2)}}$$

Simple Neural Network Example

Let the activation function, $a(\cdot)$, be the **ReLU function for all layers**.

Let the Loss function be $L = \frac{1}{2}(y-t)^2$, where t is the data label.

Find:

1. $\frac{\partial L}{\partial w^{(3)}}$ and $\frac{\partial L}{\partial w^{(1)}}$.

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

$$y = a\left(w^{(3)}h^{(2)} + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}h^{(1)} + w_b^{(2)}\right) + w_b^{(3)}\right) = a\left(w^{(3)}a\left(w^{(2)}a(w^{(1)}x + w_b^{(1)}) + w_b^{(2)}\right) + w_b^{(3)}\right)$$

$$\frac{\partial L}{\partial y} = (y-t), \quad a'\left(Z^{(l)}\right) = 1 \text{ assuming positive } Z^{(l)}$$

$$\frac{\partial L}{\partial w^{(1)}} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial Z^{(3)}}\frac{\partial Z^{(3)}}{\partial h^{(2)}}\frac{\partial h^{(2)}}{\partial Z^{(2)}}\frac{\partial Z^{(2)}}{\partial h^{(1)}}\frac{\partial h^{(1)}}{\partial Z^{(1)}}\frac{\partial Z^{(1)}}{\partial w^{(1)}} = (y-t)a'\left(Z^{(3)}\right)w^{(3)}a'\left(Z^{(2)}\right)w^{(2)}a'\left(Z^{(1)}\right)x = \boxed{(y-t)w^{(3)}w^{(2)}x}$$

Gradient Descent

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$\boldsymbol{w}^{(1)} = \begin{pmatrix} w_b^{(1)} & w^{(1)} \end{pmatrix}$$
$$\boldsymbol{w}^{(2)} = \begin{pmatrix} w_b^{(2)} & w^{(2)} \end{pmatrix}$$
$$\boldsymbol{w}^{(3)} = \begin{pmatrix} w_b^{(3)} & w^{(3)} \end{pmatrix}$$

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} \dfrac{\partial L}{\partial \boldsymbol{w}^{(1)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(2)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(3)}}^T \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix}_n - \eta \begin{pmatrix} \dfrac{\partial L}{\partial \boldsymbol{w}^{(1)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(2)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(3)}}^T \end{pmatrix}$$

Gradient Descent

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$\boldsymbol{w}^{(1)} = \begin{pmatrix} w_b^{(1)} & w^{(1)} \end{pmatrix}$$
$$\boldsymbol{w}^{(2)} = \begin{pmatrix} w_b^{(2)} & w^{(2)} \end{pmatrix}$$
$$\boldsymbol{w}^{(3)} = \begin{pmatrix} w_b^{(3)} & w^{(3)} \end{pmatrix}$$

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\boldsymbol{w} = \begin{pmatrix} w_b^{(1)} \\ w^{(1)} \\ w_b^{(2)} \\ w^{(2)} \\ w_b^{(3)} \\ w^{(3)} \end{pmatrix} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} \partial L/\partial w_b^{(1)} \\ \partial L/\partial w^{(1)} \\ \partial L/\partial w_b^{(2)} \\ \partial L/\partial w^{(2)} \\ \partial L/\partial\, w_b^{(3)} \\ \partial L/\partial w^{(3)} \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix}_{n} - \eta \begin{pmatrix} \dfrac{\partial L}{\partial \boldsymbol{w}^{(1)}}^{T} \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(2)}}^{T} \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(3)}}^{T} \end{pmatrix}$$

Gradient Descent

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

$$\boldsymbol{w}^{(\boldsymbol{1})} = \begin{pmatrix} w_b^{(1)} & w^{(1)} \end{pmatrix}$$
$$\boldsymbol{w}^{(\boldsymbol{2})} = \begin{pmatrix} w_b^{(2)} & w^{(2)} \end{pmatrix}$$
$$\boldsymbol{w}^{(\boldsymbol{3})} = \begin{pmatrix} w_b^{(3)} & w^{(3)} \end{pmatrix}$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\boldsymbol{w} = \begin{pmatrix} w_b^{(1)} \\ w^{(1)} \\ w_b^{(2)} \\ w^{(2)} \\ w_b^{(3)} \\ w^{(3)} \end{pmatrix} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} \partial L/\partial w_b^{(1)} = (y-t)w^{(3)}w^{(2)} \\ \partial L/\partial w^{(1)} = (y-t)w^{(3)}w^{(2)}x \\ \partial L/\partial w_b^{(2)} = (y-t)h^{(2)}w^{(3)} \\ \partial L/\partial w^{(2)} = (y-t)h^{(2)}w^{(3)}h^{(1)} \\ \partial L/\partial\, w_b^{(3)} = (y-t) \\ \partial L/\partial w^{(3)} = (y-t)h^{(2)} \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}^{(\boldsymbol{1})^T} \\ \boldsymbol{w}^{(\boldsymbol{2})^T} \\ \boldsymbol{w}^{(\boldsymbol{3})^T} \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{w}^{(\boldsymbol{1})^T} \\ \boldsymbol{w}^{(\boldsymbol{2})^T} \\ \boldsymbol{w}^{(\boldsymbol{3})^T} \end{pmatrix}_{n} - \eta \begin{pmatrix} \dfrac{\partial L}{\partial \boldsymbol{w}^{(1)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(2)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(3)}}^T \end{pmatrix}$$

Gradient Descent

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$\left.\begin{array}{l} \boldsymbol{w}^{(\mathbf{1})} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \\ \boldsymbol{w}^{(\mathbf{2})} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \\ \boldsymbol{w}^{(\mathbf{3})} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}\right\} \text{IC}$$

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\boldsymbol{w}_0 = \begin{array}{c} \text{IC} \\ \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \end{array} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} (y-t)w^{(3)}w^{(2)} \\ (y-t)w^{(3)}w^{(2)}x \\ (y-t)h^{(2)}w^{(3)} \\ (y-t)h^{(2)}w^{(3)}h^{(1)} \\ (y-t) \\ (y-t)h^{(2)} \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}^{(\mathbf{1})^T} \\ \boldsymbol{w}^{(\mathbf{2})^T} \\ \boldsymbol{w}^{(\mathbf{3})^T} \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{w}^{(\mathbf{1})^T} \\ \boldsymbol{w}^{(\mathbf{2})^T} \\ \boldsymbol{w}^{(\mathbf{3})^T} \end{pmatrix}_{n} - \eta \begin{pmatrix} \frac{\partial L}{\partial \boldsymbol{w}^{(1)}}^T \\ \frac{\partial L}{\partial \boldsymbol{w}^{(2)}}^T \\ \frac{\partial L}{\partial \boldsymbol{w}^{(3)}}^T \end{pmatrix}$$

Gradient Descent

$$D = \begin{pmatrix} x \\ t \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Label, t

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

IC

$$\boldsymbol{w}_0 = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} (y-t)w^{(3)}w^{(2)} \\ (y-t)w^{(3)}w^{(2)}x \\ (y-t)h^{(2)}w^{(3)} \\ (y-t)h^{(2)}w^{(3)}h^{(1)} \\ (y-t) \\ (y-t)h^{(2)} \end{pmatrix}$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{w}^{(1)T} \\ \boldsymbol{w}^{(2)T} \\ \boldsymbol{w}^{(3)T} \end{pmatrix}_{n} - \eta \begin{pmatrix} \dfrac{\partial L}{\partial \boldsymbol{w}^{(1)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(2)}}^T \\ \dfrac{\partial L}{\partial \boldsymbol{w}^{(3)}}^T \end{pmatrix}$$

$$\begin{pmatrix} h_1 \\ h_2 \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \eta \frac{\partial L}{\partial \boldsymbol{w}_0} = \eta \begin{pmatrix} -0.25 \\ -0.25 \\ -0.5 \\ -0.5 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.025 \\ -0.025 \\ -0.05 \\ -0.05 \\ -0.1 \\ -0.1 \end{pmatrix}$$

Gradient Descent

$$D = \binom{x}{t} = \binom{1}{2}$$

Label, t

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

IC

$$\boldsymbol{w}_0 = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} (y-t)w^{(3)}w^{(2)} \\ (y-t)w^{(3)}w^{(2)}x \\ (y-t)h^{(2)}w^{(3)} \\ (y-t)h^{(2)}w^{(3)}h^{(1)} \\ (y-t) \\ (y-t)h^{(2)} \end{pmatrix}$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\begin{pmatrix} \boldsymbol{w}^{(1)^T} \\ \boldsymbol{w}^{(2)^T} \\ \boldsymbol{w}^{(3)^T} \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{w}^{(1)^T} \\ \boldsymbol{w}^{(2)^T} \\ \boldsymbol{w}^{(3)^T} \end{pmatrix}_n - \eta \begin{pmatrix} \frac{\partial L}{\partial \boldsymbol{w}^{(1)}}^T \\ \frac{\partial L}{\partial \boldsymbol{w}^{(2)}}^T \\ \frac{\partial L}{\partial \boldsymbol{w}^{(3)}}^T \end{pmatrix}$$

$$\eta \frac{\partial L}{\partial \boldsymbol{w}_0} = \eta \begin{pmatrix} -0.25 \\ -0.25 \\ -0.5 \\ -0.5 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.025 \\ -0.025 \\ -0.05 \\ -0.05 \\ -0.1 \\ -0.1 \end{pmatrix} \qquad \boldsymbol{w}_1 = \begin{pmatrix} 0.525 \\ 0.525 \\ 0.55 \\ 0.55 \\ 0.6 \\ 0.6 \end{pmatrix}$$

$$\boxed{\begin{pmatrix} h_1 \\ h_2 \\ y \end{pmatrix} = \begin{pmatrix} 1.05 \\ 1.1275 \\ 1.2765 \end{pmatrix}}$$

Gradient Descent

$$D = \begin{pmatrix} x \\ t \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Label, t

$$x \xrightarrow{w^{(1)}} h^{(1)} \xrightarrow{w^{(2)}} h^{(2)} \xrightarrow{w^{(3)}} y$$

IC

$$\boldsymbol{w}_1 = \begin{pmatrix} 0.525 \\ 0.525 \\ 0.55 \\ 0.55 \\ 0.6 \\ 0.6 \end{pmatrix} \qquad \frac{\partial L}{\partial \boldsymbol{w}} = \begin{pmatrix} (y-t)w^{(3)}w^{(2)} \\ (y-t)w^{(3)}w^{(2)}x \\ (y-t)h^{(2)}w^{(3)} \\ (y-t)h^{(2)}w^{(3)}h^{(1)} \\ (y-t) \\ (y-t)h^{(2)} \end{pmatrix}$$

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \eta \frac{\partial L}{\partial \boldsymbol{w}}$$

$$\begin{pmatrix} h_1 \\ h_2 \\ y \end{pmatrix} = \begin{pmatrix} 1.05 \\ 1.1275 \\ 1.2765 \end{pmatrix} \quad \eta \frac{\partial L}{\partial \boldsymbol{w}_0} = \begin{pmatrix} -0.0238755 \\ -0.0238755 \\ -0.0489448 \\ -0.0513920 \\ -0.0723500 \\ -0.0815746 \end{pmatrix} \quad \boldsymbol{w}_2 = \begin{pmatrix} 0.5488755 \\ 0.5488755 \\ 0.5989448 \\ 0.6013920 \\ 0.6723500 \\ 0.6815746 \end{pmatrix} \longrightarrow \boxed{\begin{pmatrix} h_1 \\ h_2 \\ y \end{pmatrix} = \begin{pmatrix} 1.097751 \\ 1.259123 \\ 1.530537 \end{pmatrix}}$$

Softmax Gradient

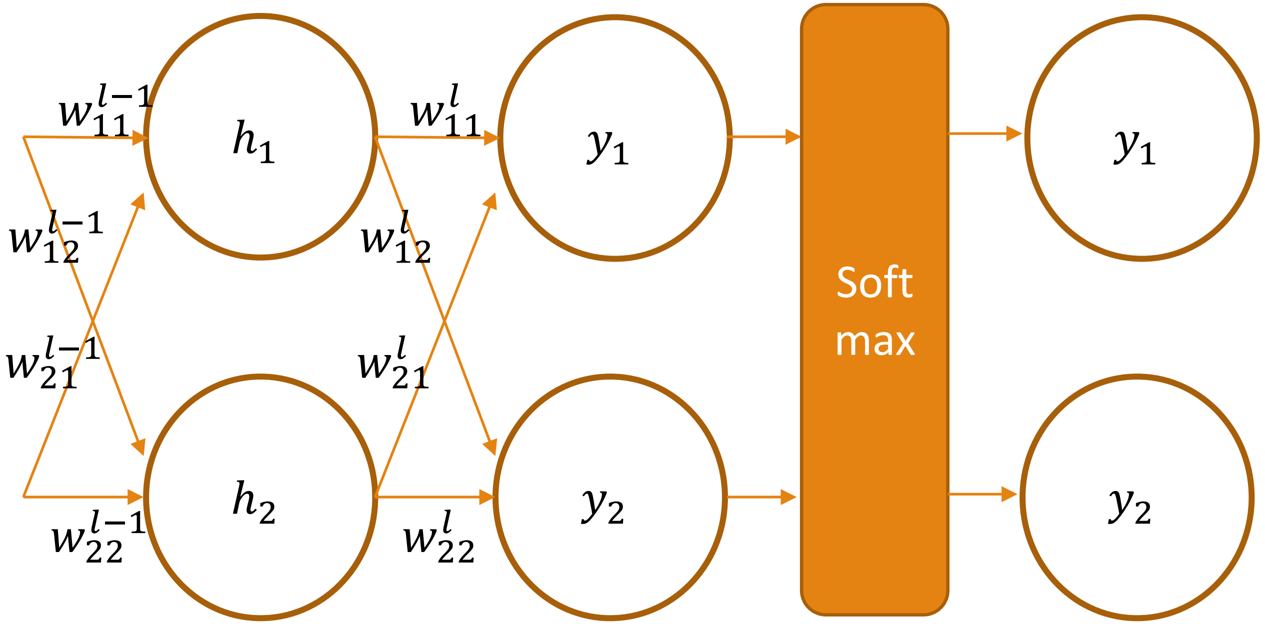
$$\begin{pmatrix} 1.058 \\ 0.013 \\ 0.568 \\ 1.345 \end{pmatrix} \Rightarrow \begin{pmatrix} 0.303 \\ 0.107 \\ 0.186 \\ 0.404 \end{pmatrix}$$

$$s(i, \boldsymbol{x}) = \ softmax(i, \boldsymbol{x}) = \frac{\exp(x_i)}{\sum_i \exp(x_i)}$$

$$\frac{\partial s}{\partial x_i} = \frac{\exp(x_i)\sum_i \exp(x_i) - \exp(x_i)\exp(x_i)}{(\sum_i \exp(x_i))^2}$$

$$= \frac{\exp(x_i)}{\sum_i \exp(x_i)} - \left[\frac{\exp(x_i)}{\sum_i \exp(x_i)}\right]^2$$

$$\boxed{= s(x_i)(1 - s(x_i))}$$



Next Lecture

Programming!!!!