SCS 3201

Machine Learning and Neural Computing

Kasun Gunawardana

University of Colombo School of Computing

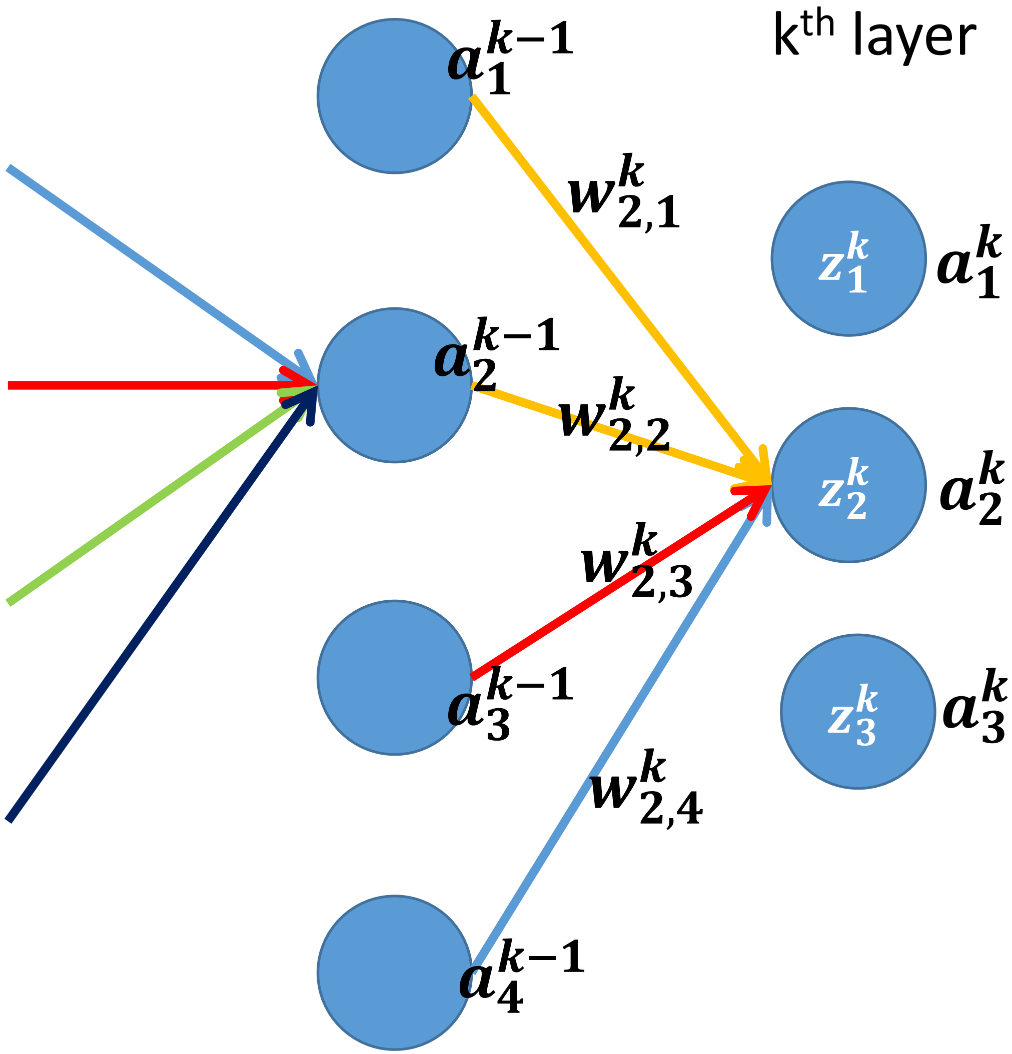
Artificial Neural Networks

- Backpropagation Learning

Multilayer ANN Training

- Weight adjustment
  - Different Layers
  - Error minimization
- Cost
  - Cost contribution – Each Layer / Each Neuron

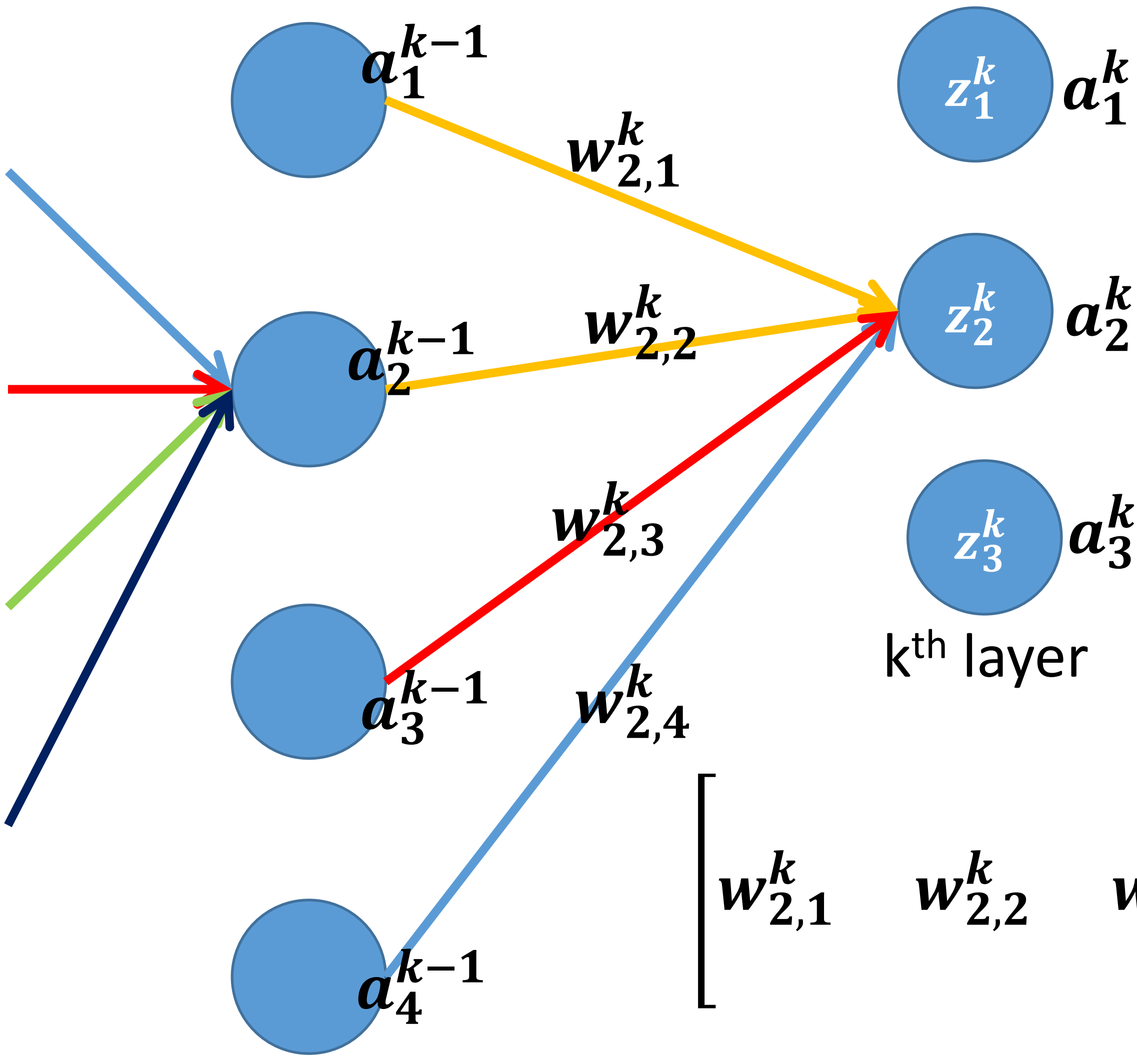
Notation



Notation

- a_i^k - Activation of i^{th} neuron in k^{th} layer.
- $z_i^k = \sum w_i a_i$
- $w_{i,j}^k$ - weight between j^{th} neuron in $(k-1)^{th}$ layer and i^{th} neuron in k^{th} layer.

Notation



$$\begin{bmatrix} w_{2,1}^k & w_{2,2}^k & w_{2,3}^k & w_{2,4}^k \end{bmatrix} \begin{bmatrix} a_1^{k-1} \\ a_2^{k-1} \\ a_3^{k-1} \\ a_4^{k-1} \end{bmatrix} = \begin{bmatrix} z_1^k \\ z_2^k \\ z_3^k \end{bmatrix}$$

Computation

$$\begin{bmatrix} \boldsymbol{w_{1,1}^k} & \boldsymbol{w_{1,2}^k} & \boldsymbol{w_{1,3}^k} & \boldsymbol{w_{1,4}^k} \\ \boldsymbol{w_{2,1}^k} & \boldsymbol{w_{2,2}^k} & \boldsymbol{w_{2,3}^k} & \boldsymbol{w_{2,4}^k} \\ \boldsymbol{w_{3,1}^k} & \boldsymbol{w_{3,2}^k} & \boldsymbol{w_{3,3}^k} & \boldsymbol{w_{3,4}^k} \end{bmatrix} \begin{bmatrix} \boldsymbol{a_1^{k-1}} \\ \boldsymbol{a_2^{k-1}} \\ \boldsymbol{a_3^{k-1}} \\ \boldsymbol{a_4^{k-1}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z_1^k} \\ \boldsymbol{z_2^k} \\ \boldsymbol{z_3^k} \end{bmatrix}$$

Note - $z_i^k = \sum_j w_{i,j}\, a_j^{k-1}$

Forward Propagation of Activation

- $a^1 = x$
- $z^2 = w^2 a^1$
- $a^2 = h(z^2)$
- ……
- $a^K = h(z^K)$
- $a^K = g(x)$

$\boldsymbol{K}$ – Output layer
$\boldsymbol{k}$ – Any layer in general

Activation Function

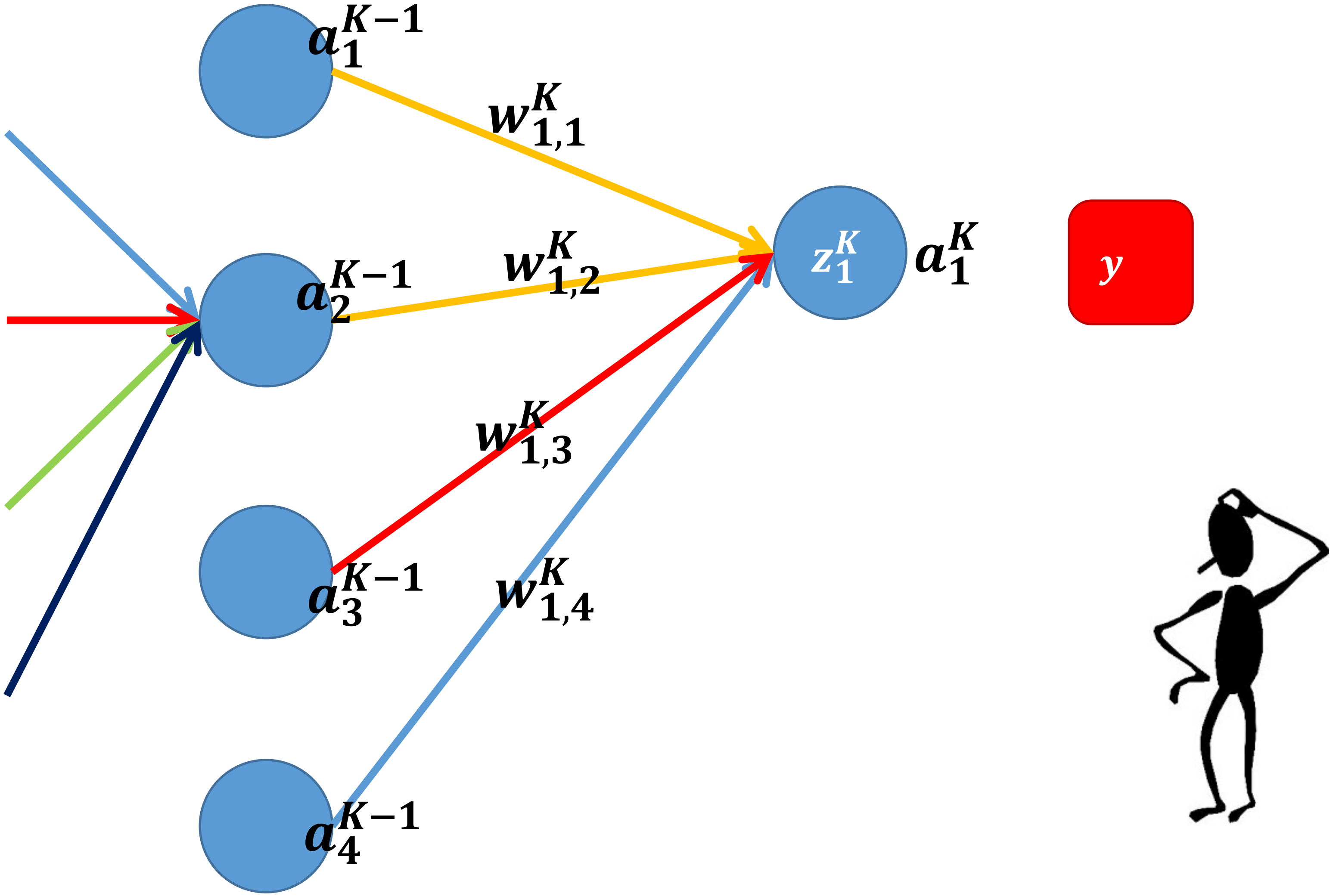
- Activation function

$$a_i^k = h\left(\sum_t w_{i,t}^k \, a_t^{(k-1)} + \, b_i^k\right)$$

- By vectorizing the function $h(\mathrm{x})$,

$$a^k = h\left(w^k a^{(k-1)} + b^k\right)$$

Cost Minimization - Complexity



Cost Minimization

- We know
  - Generated Output (Activation)
  - Expected Output (Label)

For a given input

- Cost can be defined (Differentiable Cost Function)
- Gradient Descent can be employed

Gradient Descent

- Cost Function – $J(w)$ (or $C(w)$ with respect to a single input)
- Minimization $J(w)$ with respect to different $w_{i,j}^{k}$

$$\frac{\partial}{\partial w_{i,j}^{k}} J(w)$$

- $w_{i,j}^{k} \in \mathbb{R}$

Backpropagation - The Challenge

Search a large hypothesis space defined by

- All possible weight values
- For all the neurons
- In all the layers

of the entire network.

Backpropagation

- Error terms are computed backward
- Output layer first
- Error propagates from last layer till second layer
- No error in first layer as it is the input (x)

Overview of the Backpropagation

- Initialize all weights (e.g., to random values)
- Repeat for each (x, y)
- Feed input x and compute activation for each neuron
- Compute error term for each neuron
  - Adjust weights (Update weights backwards)
- Do until converge

Understanding Backpropagation

- Let's consider a neural network with 4 layers and each layer has single neuron.

$$x \xrightarrow{w_2} a^2 \xrightarrow{w_3} a^3 \xrightarrow{w_4} a^4$$

- Cost Function $J(w_2, w_3, w_4)$
- Considering the bias $J(w_2, b_2, w_3, b_3, w_4, b_4)$

Understanding Backpropagation

- Let's consider only the last layer for this instance.
- Activation of last neuron a^4

  (Assume total of K layers, then it is a^K)

$$x \xrightarrow{w_2} a^2 \xrightarrow{w_3} a^3 \xrightarrow{w_4} a^4$$

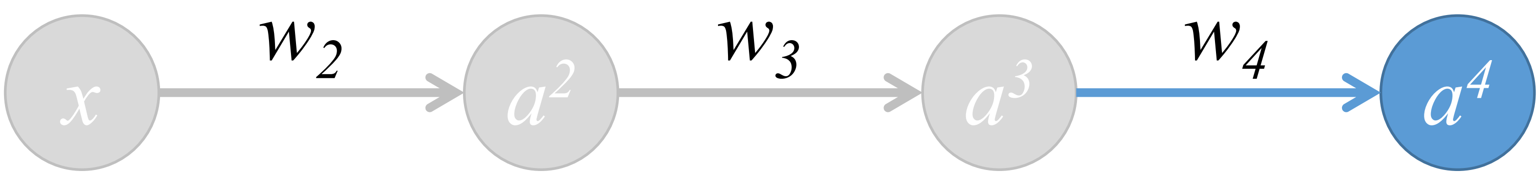
- Let's consider a single input.
- Expected output for last neuron is y.
- Cost Function $J(w_2, w_3, w_4) = Cost(a^K, y)$

Understanding Backpropagation

If $s(x)$ is the activation function,

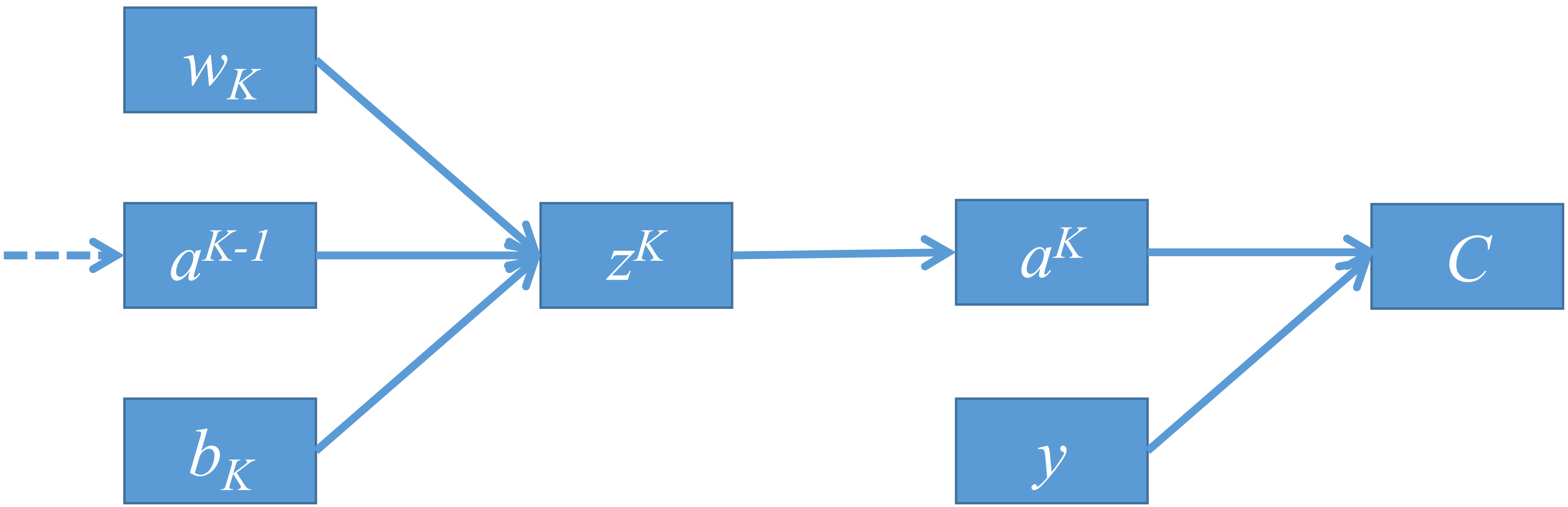
$$a^K = s(w^K a^{K-1} + b^K)$$

$$a^K = s(z^K)$$



- $s(x)$ – is the Sigmoid function.

Understanding Backpropagation



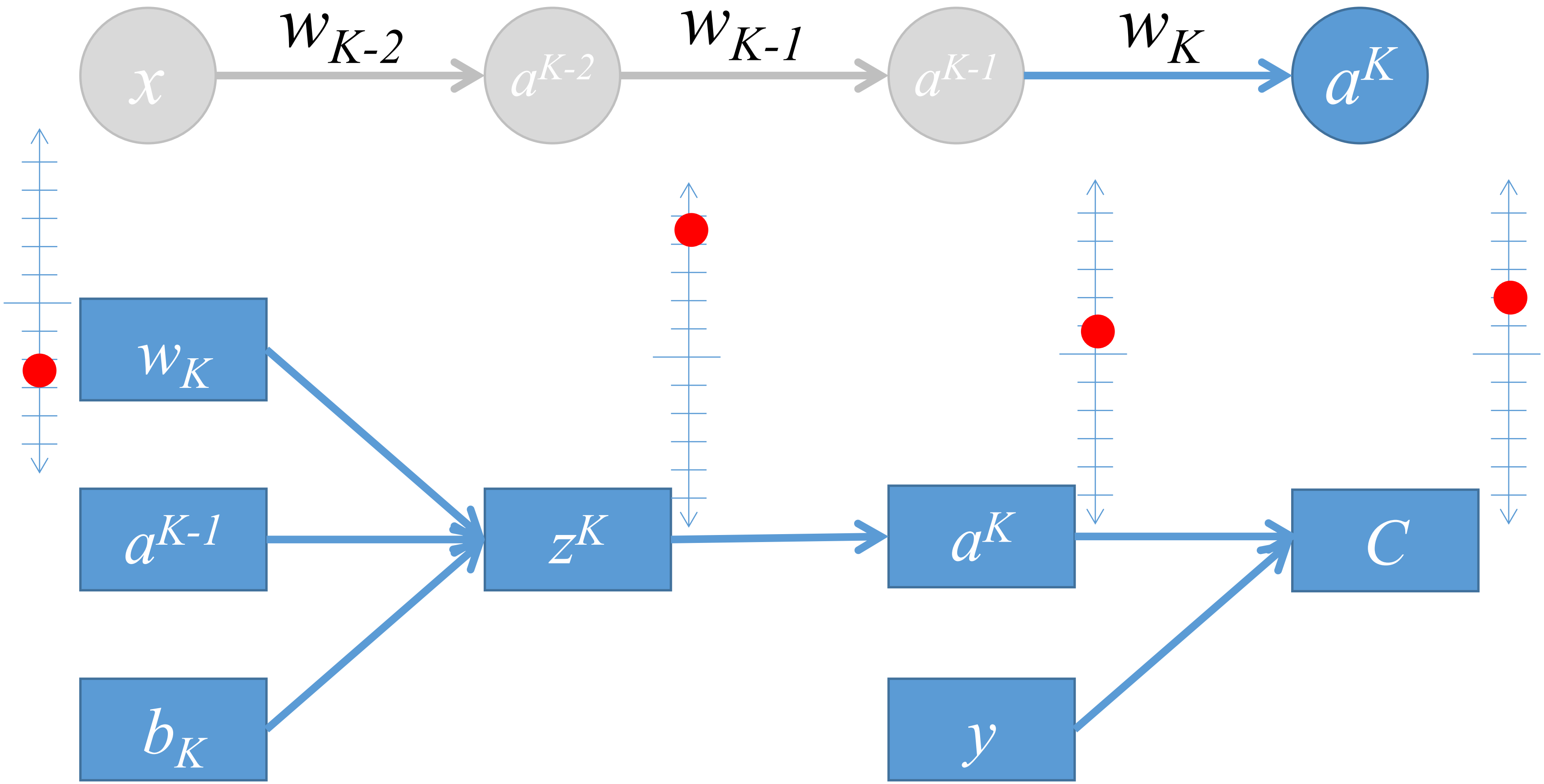
Cost Minimization

- To minimize the cost
  - Change weights (w_k)
  - Change activation of the previous layer (a^{K-1})

- Can we change a^{K-1}
  - To change previous layer activation
    - Change w_{k-1}
    - Change a^{K-2}

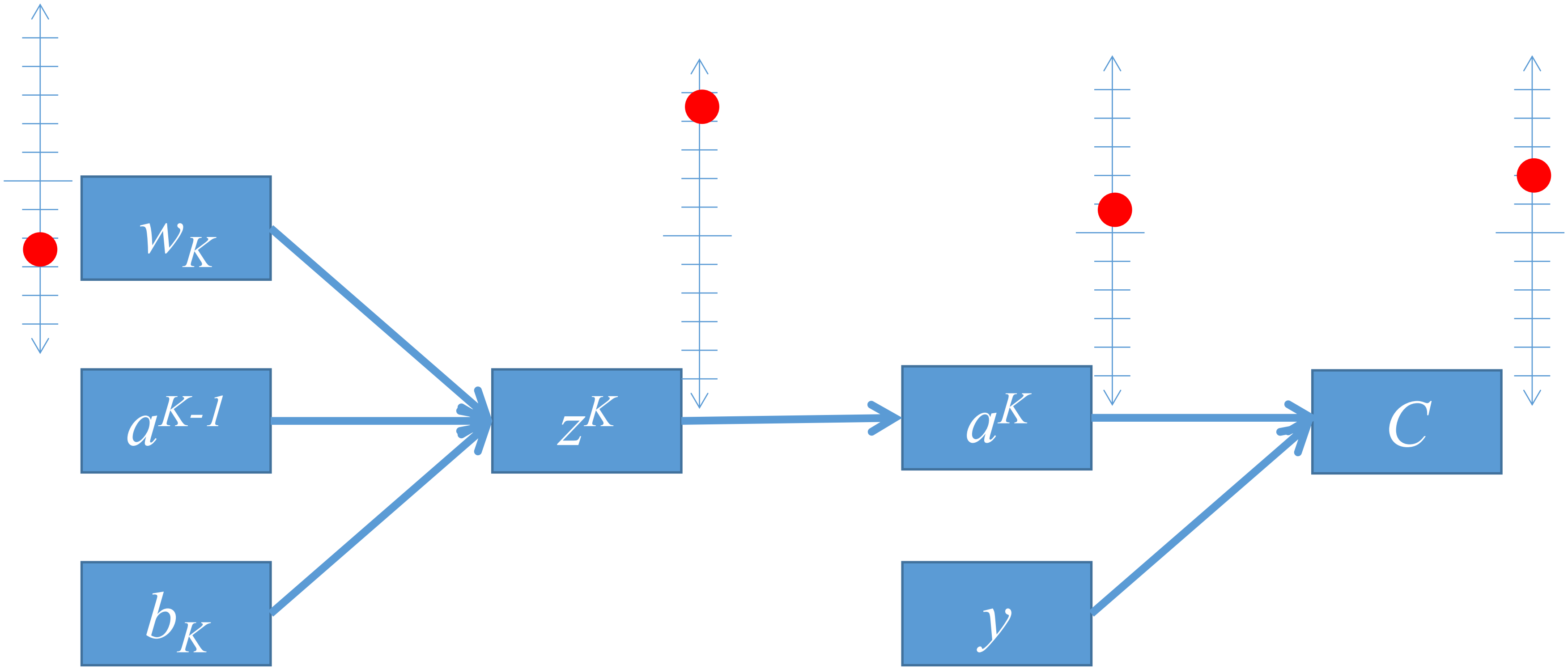
Backward Propagation

Understanding Backpropagation



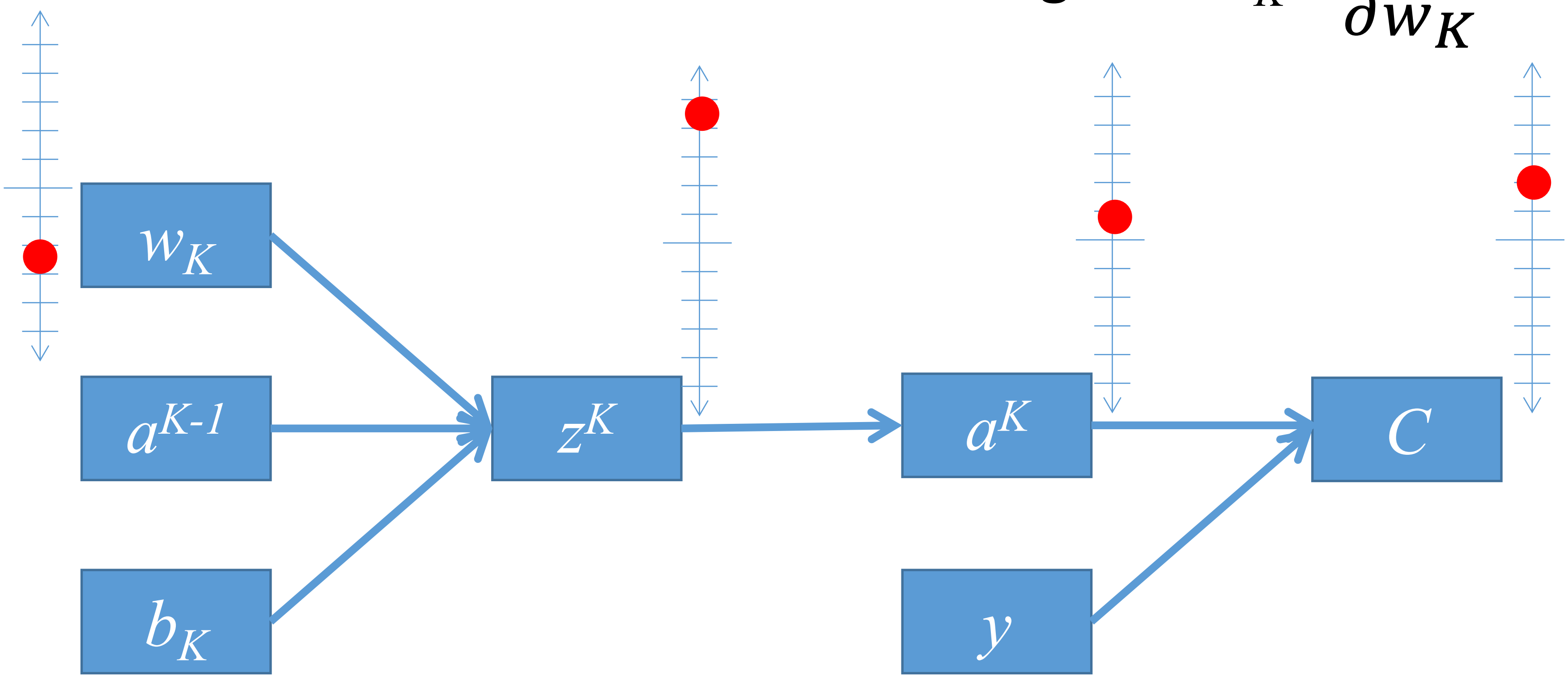
Understanding Backpropagation

- How sensitive our C to small changes in w_K?



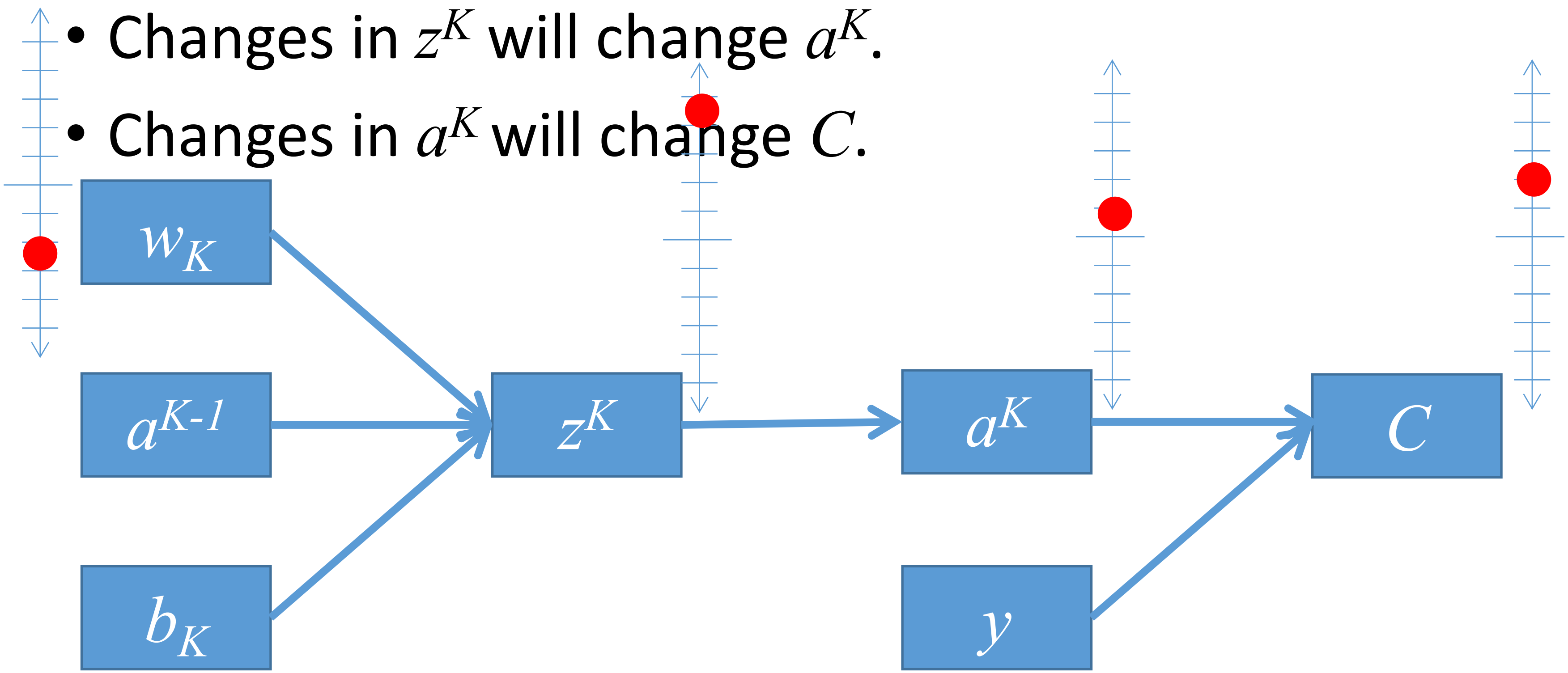
Understanding Backpropagation

- How sensitive our C to small changes in w_K? $\dfrac{\partial C}{\partial w_K}$



Propagation

- Changes in w_K will change z^K.
- Changes in z^K will change a^K.
- Changes in a^K will change C.



Propagation

- How sensitive our C to small changes in w_K? $\frac{\partial C}{\partial w_K}$
- Changes in w_K will change $z^K \Rightarrow \frac{\partial z^K}{\partial w_K}$
- Changes in z^K will change $a^K \Rightarrow \frac{\partial a^K}{\partial z^K}$
- Changes in a^K will change $C \Rightarrow \frac{\partial C}{\partial a^K}$

$$\frac{\partial C}{\partial w_K} = \frac{\partial z^K}{\partial w_K}\frac{\partial a^K}{\partial z^K}\frac{\partial C}{\partial a^K}$$

Chain of derivatives

$$\frac{\partial C}{\partial w_K} = \frac{\partial z^K}{\partial w_K} \frac{\partial a^K}{\partial z^K} \frac{\partial C}{\partial a^K}$$

- Chain Rule

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial w_K}$$

Chain of derivatives

$$\frac{\partial C}{\partial w_K} = \frac{\partial z^K}{\partial w_K} \frac{\partial a^K}{\partial z^K} \frac{\partial C}{\partial a^K}$$

- $\frac{\partial C}{\partial a^K} = \frac{\partial}{\partial a^K} Cost(a^K, y) = C'$
- $\frac{\partial a^K}{\partial z^K} = s'(z^K)$
- $\frac{\partial z^K}{\partial w_K} = a^{K-1}$

$$C(w) = Cost(a^K, y)$$

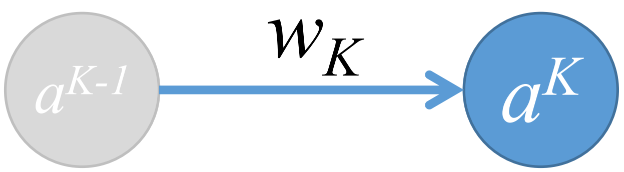
$$a^K = s(z^K)$$

$$z^K = w^K a^{K-1} + b^K$$

Chain of derivatives

$$\frac{\partial C}{\partial w_K} = \frac{\partial z^K}{\partial w_K} \frac{\partial a^K}{\partial z^K} \frac{\partial C}{\partial a^K}$$

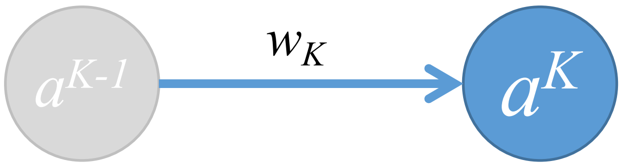
- $\frac{\partial C}{\partial a^K} = \frac{\partial}{\partial a^K} Cost(a^K, y) = C'$
- $\frac{\partial a^K}{\partial z^K} = s'(z^K)$
- $\frac{\partial z^K}{\partial w_K} = a^{K-1}$



Something to Note…

- Change happen to z from a small change in w depends on a^{K-1}
- The ratio that z influence by small change in w completely rely on a^{K-1}

$$\frac{\partial z^K}{\partial w_K} = a^{K-1}$$



"Fire together – Wire together"

Weight Adjustment – Last Layer

- Weight will be adjusted

$$W_K = W_K - \alpha \frac{\partial C}{\partial w_K}$$

Gradient Descent Learning Rule

- Next is the (K-1) layer weight update
  - How sensitive C to small changes in w_{K-1}?

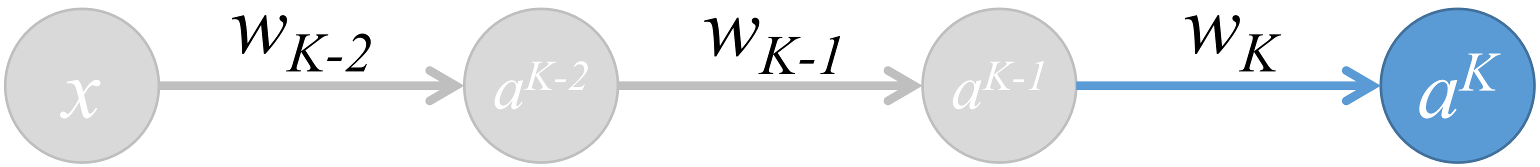
Backpropagation with Hidden Layers

Chain Rule

- The chain rule is a formula to compute the derivative of a composite function.
- Ex. If a variable $\boldsymbol{w}$ depends on the variable $\boldsymbol{x}$, variable $\boldsymbol{x}$ depends on the variable $\boldsymbol{y}$ and variable $\boldsymbol{y}$ depends on the variable $\boldsymbol{z}$.

$$\frac{dw}{dz} = \frac{dw}{dx}\frac{dx}{dy}\frac{dy}{dz} = \frac{dw}{dx}\frac{dx}{dz} = \frac{dw}{dy}\frac{dy}{dz}$$

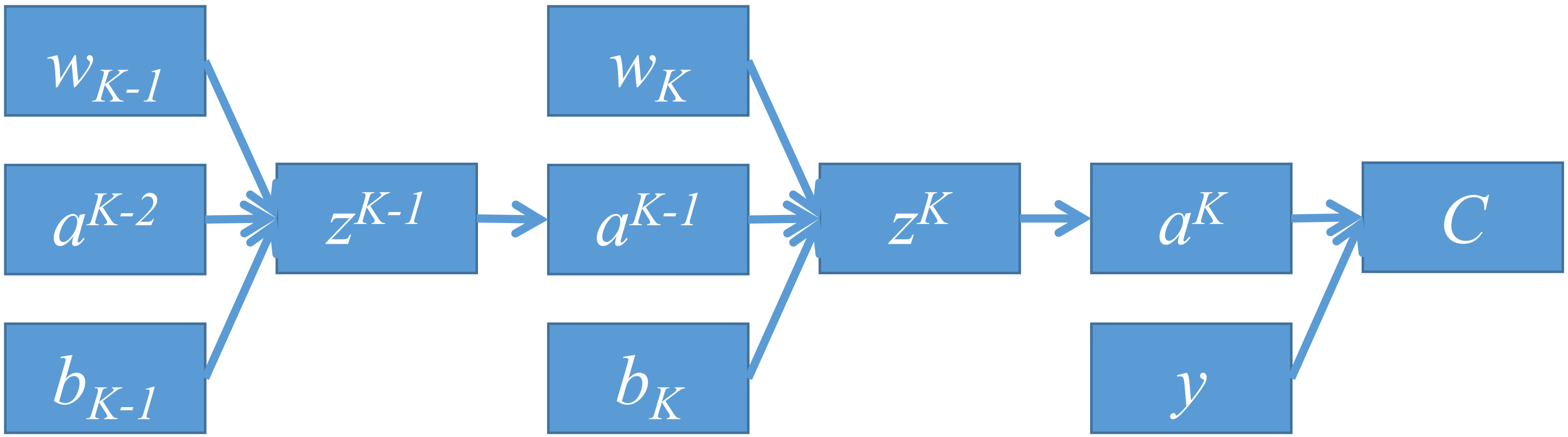
(K-1)th Layer: Weight Adjustment



- How sensitive C to small changes in w_{K-1}?

$$\frac{\partial C}{\partial w_{K-1}}$$

(K-1)th Layer: Weight Adjustment



- How sensitive C to small changes in w_{K-1}?

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial w_{K-1}}$$

(K-2)th Layer: Weight Adjustment

- How sensitive C to small changes in w_{K-1}?

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K}} \frac{\partial a^{K}}{\partial z^{K}} \frac{\partial z^{K}}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial w_{K-2}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial w_K}$$

- Layer K-1

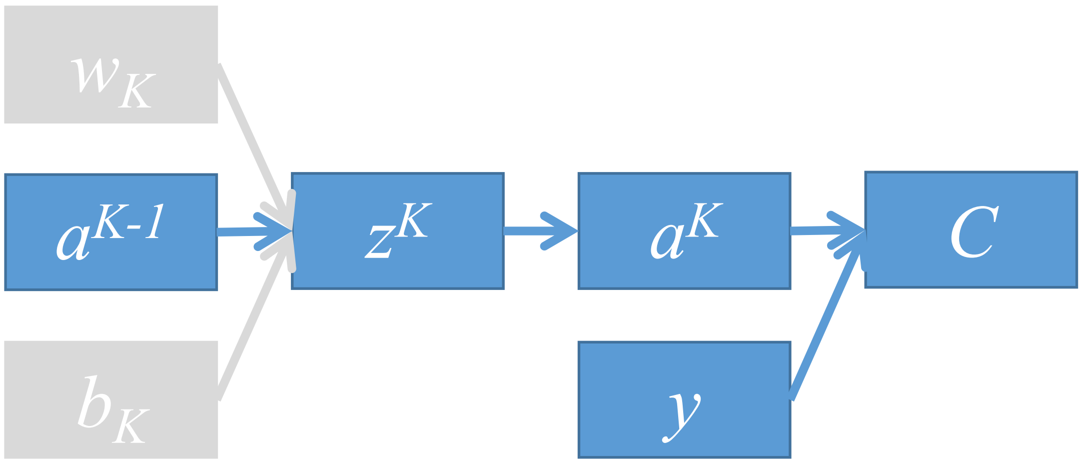
$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial w_{K-1}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial w_{K-2}}$$

Time to Think..!

- How sensitive C to small changes in a^{K-1}?



- $\frac{\partial C}{\partial a^{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial w_K}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial w_{K-1}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial a^{K-1}}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial w_{K-2}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial w_K}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial w_{K-1}}$$

$$\frac{\partial C}{\partial a^{K-1}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial w_{K-2}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial w_K}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K} \frac{\partial a^K}{\partial z^K} \frac{\partial z^K}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial w_{K-1}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^{K-1}}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial w_{K-2}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial w_K}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial w_{K-1}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial w_{K-2}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial a^{K-3}}\frac{\partial a^{K-3}}{\partial z^{K-3}}\frac{\partial z^{K-3}}{\partial w_{K-3}}$$

Similar to this, we can continue for any layer…

Layerwise Weight Adjustment

$$z^k = w_k \times a^{k-1}$$

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\boxed{\frac{\partial z^K}{\partial w_K}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\boxed{\frac{\partial z^{K-1}}{\partial w_{K-1}}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\boxed{\frac{\partial z^{K-2}}{\partial w_{K-2}}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial a^{K-3}}\frac{\partial a^{K-3}}{\partial z^{K-3}}\boxed{\frac{\partial z^{K-3}}{\partial w_{K-3}}}$$

Layerwise Weight Adjustment

$$\boxed{z^k = w_k \times a^{k-1}}$$

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\boxed{\frac{\partial z^K}{\partial w_K}} \rightarrow \boxed{a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\boxed{\frac{\partial z^{K-1}}{\partial w_{K-1}}} \rightarrow \boxed{a^{K-2}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\boxed{\frac{\partial z^{K-2}}{\partial w_{K-2}}} \rightarrow \boxed{a^{K-3}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial a^{K-3}}\frac{\partial a^{K-3}}{\partial z^{K-3}}\boxed{\frac{\partial z^{K-3}}{\partial w_{K-3}}} \rightarrow \boxed{a^{K-4}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\boxed{a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\boxed{a^{K-2}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\boxed{a^{K-3}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial a^{K-3}}\frac{\partial a^{K-3}}{\partial z^{K-3}}\boxed{a^{K-4}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \boxed{\delta^K} \boxed{a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^{K-1}} \boxed{a^{K-2}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^{K-2}} \boxed{a^{K-3}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^{K-3}} \boxed{a^{K-4}}$$

Gradient Descent

- Weight Adjustment Rule

$$w_k = w_k - \alpha \frac{\partial C}{\partial w_k}$$

- Generalized form for any layer

$$w_k = w_k - \alpha \, \delta^k a^{k-1}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \boxed{\frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}}\boxed{a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\boxed{a^{K-2}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}\frac{\partial z^{K-1}}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\boxed{a^{K-3}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}}\frac{\partial z^{K-2}}{\partial a^{K-3}}\frac{\partial a^{K-3}}{\partial z^{K-3}}\boxed{a^{K-4}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \; a^{K-1}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \delta^K \frac{\partial z^K}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} a^{K-2}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial a^{K-3}} \frac{\partial a^{K-3}}{\partial z^{K-3}} a^{K-4}$$

Layerwise Weight Adjustment

$$z^k = w_k \times a^{k-1}$$
$$a^{k-1} = s(z^{k-1})$$

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \; a^{K-1}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \delta^K \frac{\partial z^K}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} a^{K-2}$$

w_k

$s'(z^{K-1})$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial a^{K-3}} \frac{\partial a^{K-3}}{\partial z^{K-3}} a^{K-4}$$

Layerwise Weight Adjustment

$$z^k = w_k \times a^{k-1}$$

$$a^{k-1} = s(z^{k-1})$$

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \; a^{K-1}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \delta^K \; w_K \; s'(z^{K-1}) \; a^{K-2}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \frac{\partial z^{K-1}}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \frac{\partial z^{K-2}}{\partial a^{K-3}} \frac{\partial a^{K-3}}{\partial z^{K-3}} a^{K-4}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \; a^{K-1}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \delta^K \; w_K \; s'(z^{K-1}) \; a^{K-2}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}} \frac{\partial a^{K-1}}{\partial z^{K-1}} \; w_{K-1} \; s'(z^{K-2}) \; a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \; w_{K-2} \; s'(z^{K-3}) \; a^{K-4}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \quad a^{K-1} \qquad\qquad \frac{\partial C}{\partial a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}} \quad a^{K-2}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}} \quad w_{K-1} \quad s'(z^{K-2}) \quad a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}} \quad w_{K-2} \quad s'(z^{K-3}) \quad a^{K-4}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \; a^{K-1} \qquad \frac{\partial C}{\partial a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^K}\frac{\partial a^K}{\partial z^K}\frac{\partial z^K}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}} \; a^{K-2}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\frac{\partial C}{\partial a^{K-1}}\frac{\partial a^{K-1}}{\partial z^{K-1}}} \; w_{K-1} \; s'(z^{K-2}) \; a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}}\frac{\partial a^{K-2}}{\partial z^{K-2}} \; w_{K-2} \; s'(z^{K-3}) \; a^{K-4}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \boxed{\delta^K} \boxed{a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^K} \boxed{w_K} \boxed{s'(z^{K-1})} \boxed{a^{K-2}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^{K-1}} \boxed{w_{K-1}} \boxed{s'(z^{K-2})} \boxed{a^{K-3}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \frac{\partial C}{\partial a^{K-2}} \frac{\partial a^{K-2}}{\partial z^{K-2}} \boxed{w_{K-2}} \boxed{s'(z^{K-3})} \boxed{a^{K-4}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \boxed{\delta^K} \boxed{a^{K-1}}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^K} \boxed{w_K} \boxed{s'(z^{K-1})} \boxed{a^{K-2}}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^{K-1}} \boxed{w_{K-1}} \boxed{s'(z^{K-2})} \boxed{a^{K-3}}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \boxed{\delta^{K-2}} \boxed{w_{K-2}} \boxed{s'(z^{K-3})} \boxed{a^{K-4}}$$

Layerwise Weight Adjustment

- Layer K

$$\frac{\partial C}{\partial w_K} = \delta^K \quad a^{K-1}$$

- Layer K-1

$$\frac{\partial C}{\partial w_{K-1}} = \delta^{K-1} \quad a^{K-2}$$

- Layer K-2

$$\frac{\partial C}{\partial w_{K-1}} = \delta^{K-2} \quad a^{K-3}$$

- Layer K-3

$$\frac{\partial C}{\partial w_{K-1}} = \delta^{K-3} \quad a^{K-4}$$

Gradient Descent

- Weight Adjustment Rule

$$w_k = w_k - \alpha \frac{\partial C}{\partial w_k}$$

- Generalized form for any layer

$$w_k = w_k - \alpha \, \delta^k a^{k-1}$$

$$w_k = w_k - \alpha \, \delta^{k+1} \, w_{k+1} \, s'\left(z^k\right) \, a^{k-1}$$

Generalized Rule for δ

- When k is a hidden layer

$$\delta^{k} = \delta^{k+1} * w_{k+1} * s'(z^{k})$$

Algorithm

SGD

- $repeat\ for\ each\ input\ (x, y)$
- $Compute\ output\ h(x)$
- $for\ each$
  - $output\ layer\ neuron\ i, compute\ its\ error\ term\ \delta$
  $$\delta_i^K = \frac{\partial C}{\partial a_i^K} \frac{\partial a_i^K}{\partial z_i^K}$$
  - $hidden\ neuron, compute\ its\ error\ term$
  $$\delta_i^k = s'(z^k) \sum_{j \in k+1} w_{j,i}^{k+1} \delta_j^{k+1}$$
- $Update\ each\ weight,\quad w_{j,i}^k = w_{j,i}^k - \Delta w_{j,i}^k$

  $where,\quad \Delta w_{j,i}^k = \alpha\ \delta_j^k\ a_i^{k-1}$
- $Do\ Until\ Converge$

Algorithm

SGD

- $repeat\ for\ each\ input\ (x, y)$
- $Compute\ output\ h(x)$
- $for\ each$
  - $output\ layer\ neuron\ i, compute\ its\ error\ term\ \delta$

$$\delta_i^K = \frac{\partial C}{\partial a_i^K}\frac{\partial a_i^K}{\partial z_i^K}$$

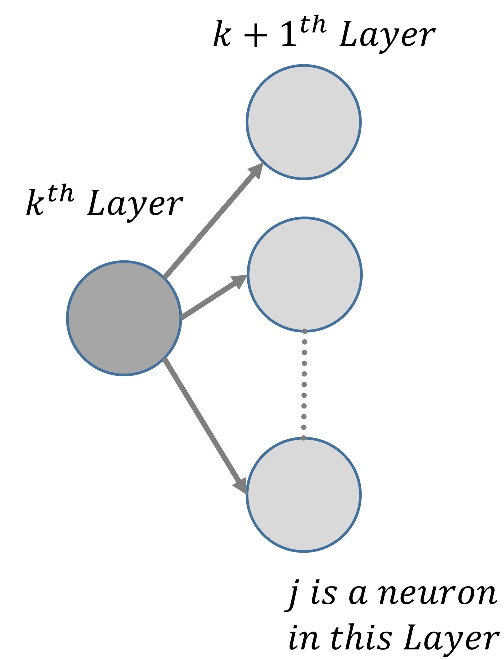
  - $hidden\ neuron, compute\ its\ error\ term$

$$\delta_i^k = s'(z^k) \sum_{j \in k+1} w_{j,i}^{k+1} \delta_j^{k+1}$$

- $Update\ each\ weight,\quad w_{j,i}^k = w_{j,i}^k - \Delta w_{j,i}^k$

$where,\quad \Delta w_{j,i}^k = \alpha\ \delta_j^k\ a_i^{k-1}$

- $Do\ Until\ Converge$



An article that should be read..!

A Step-by-Step Backpropagation Example

- https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/

External Resources

- Watch these videos
  - https://www.youtube.com/watch?v=aircAruvnKk
  - https://www.youtube.com/watch?v=IHZwWFHWa-w
  - https://www.youtube.com/watch?v=Ilg3gGewQ5U
  - https://www.youtube.com/watch?v=tIeHLnjs5U8

Q & A

Thank you..!