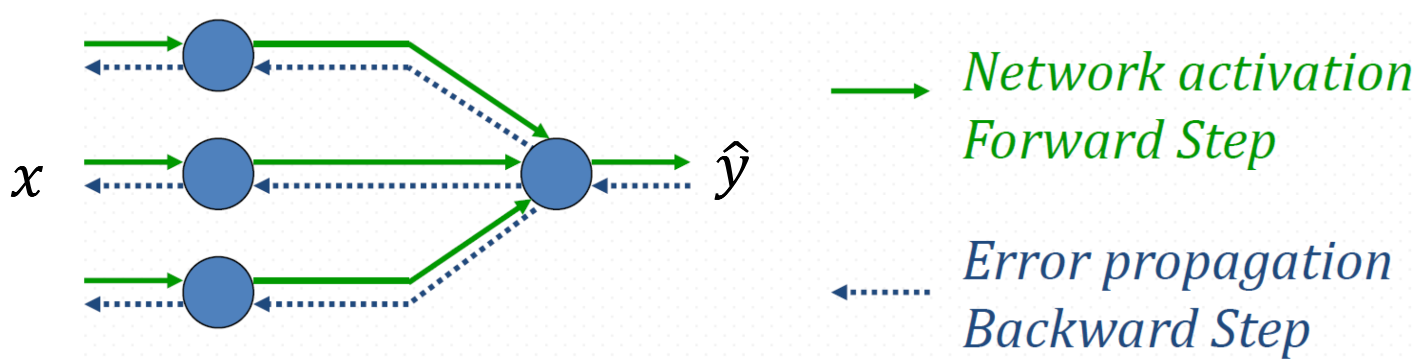
Machine Learning

Backpropagation

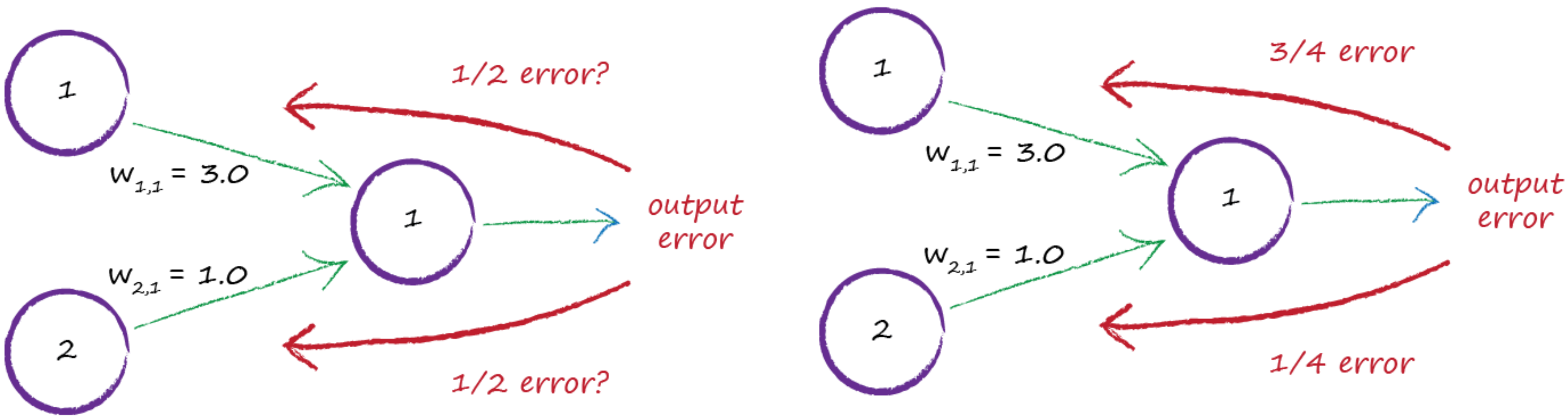
김선녕(ksycafe@gmail.com)

역전파(Backpropagation)

- 1986년 데이비드 럼멜하트(David Rumelhart), 제프리 힌튼(Geoffrey Hinton), 로날드 윌리암스(Ronald Williams) 발표
  - Learning representations by back-propagating errors
- 실제 값과 모델이 계산한 결과 값(예측 값)의 차이(오차 값)를 역으로 전파해 $weight$값을 갱신하는 알고리즘(학습)
  - 효율적인 기법으로 Gradient를 자동으로 계산하는 경사 하강법
  - 정방향, 역방향 각 한번 통과하는 것만으로 모든 모델 파라미터에 대한 네트워크 오차의 Gradient를 계산
  - <u>x에서 시작해서 $\hat{y}$까지 미치는 영향(미분 값)을 알아야 $weight$값을 조정할 수 있다</u>



오차분배방법

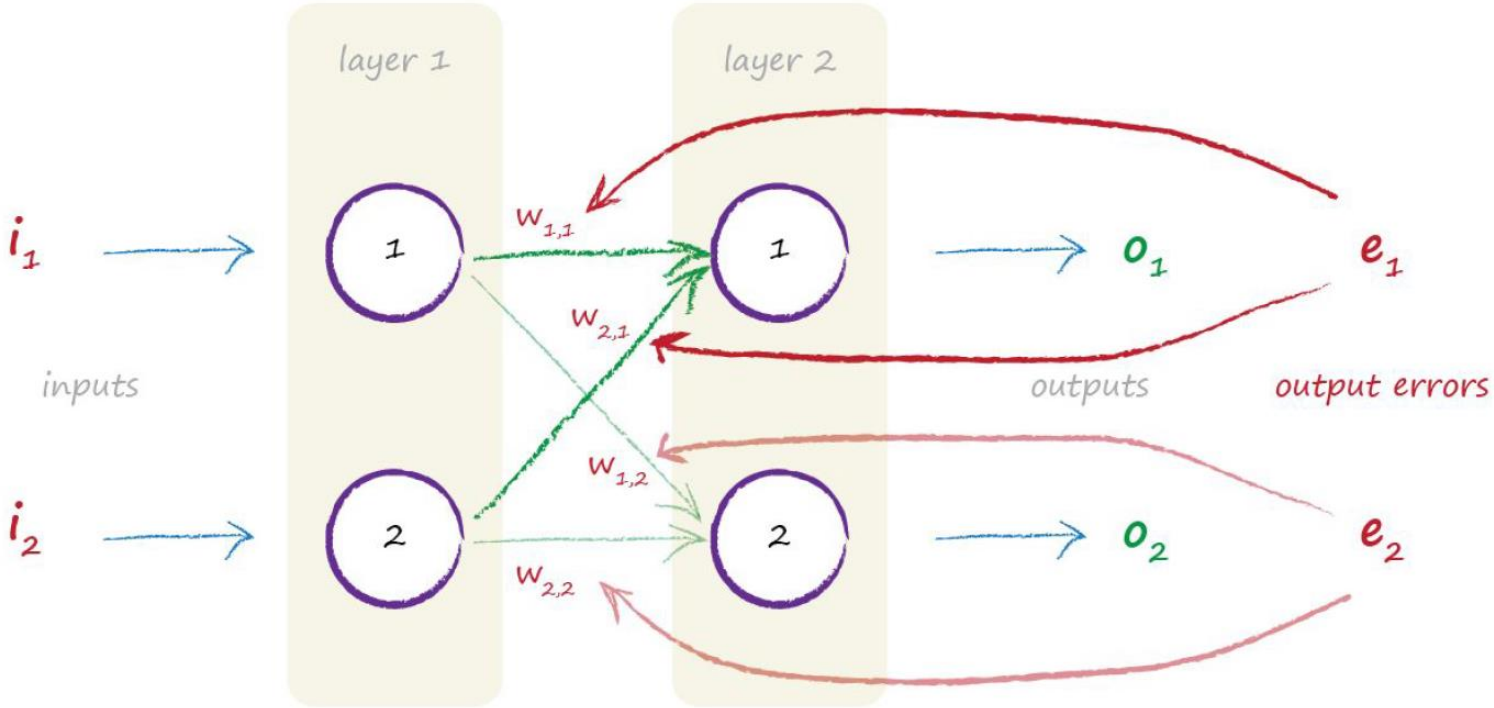


$$e_{hidden} = \begin{pmatrix} \frac{w_{11}}{w_{11}+w_{21}+\cdots} & \frac{w_{12}}{w_{12}+w_{22}+\cdots} & \cdots \\ \frac{w_{21}}{w_{11}+w_{21}+\cdots} & \frac{w_{22}}{w_{12}+w_{22}+\cdots} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \cdot e_{output}$$

ref : Make your own neural network - Triq Rashid

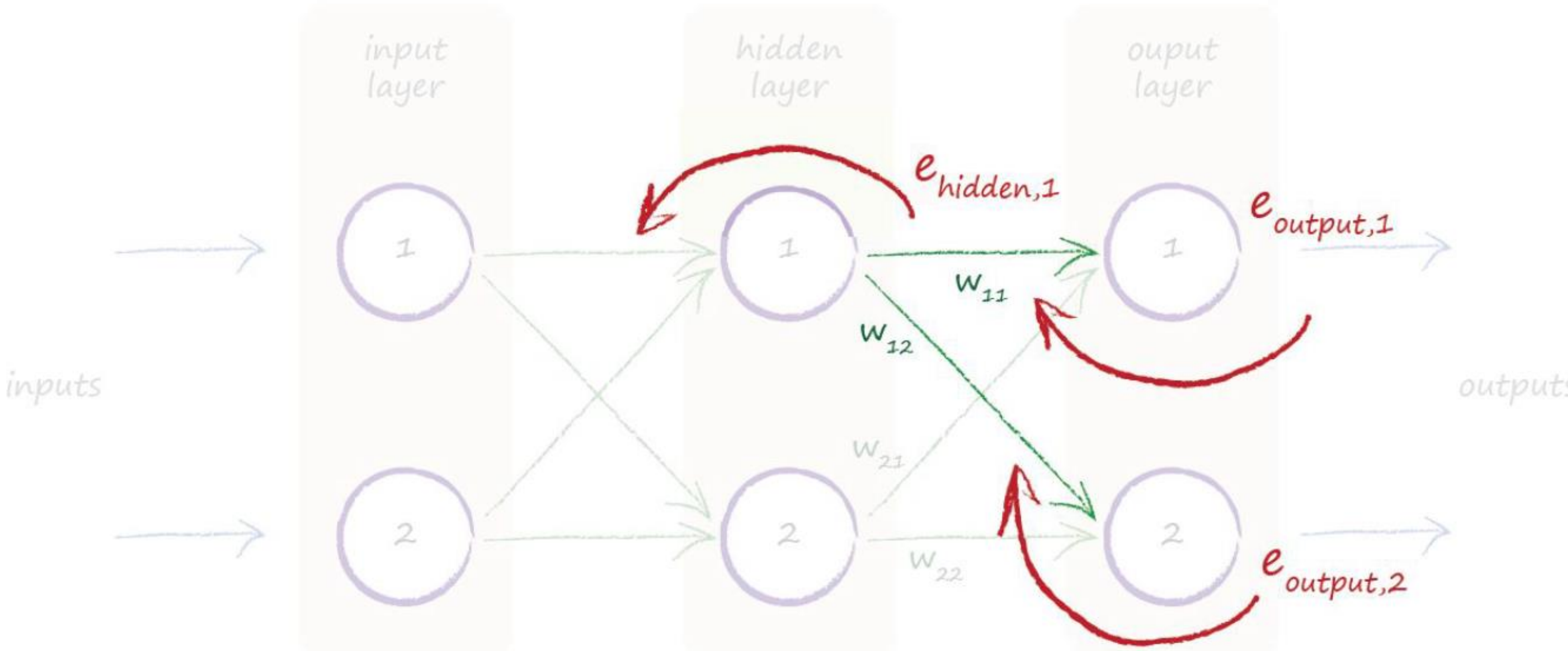
- 실제 값이 y_1이면 $e_1 = y_1 - o_1$
- 오차 e_1은 나뉘어 전달될 때 작은 가중치를 가지는 연결 노드보다 큰 가중치를 가지는 연결 노드에 더 많이 전달(가중치 비례에 따라서)

$$w_{1,1} = \frac{w_{11}}{w_{11} + w_{21}}, \qquad w_{2,1} = \frac{w_{21}}{w_{11} + w_{21}}$$

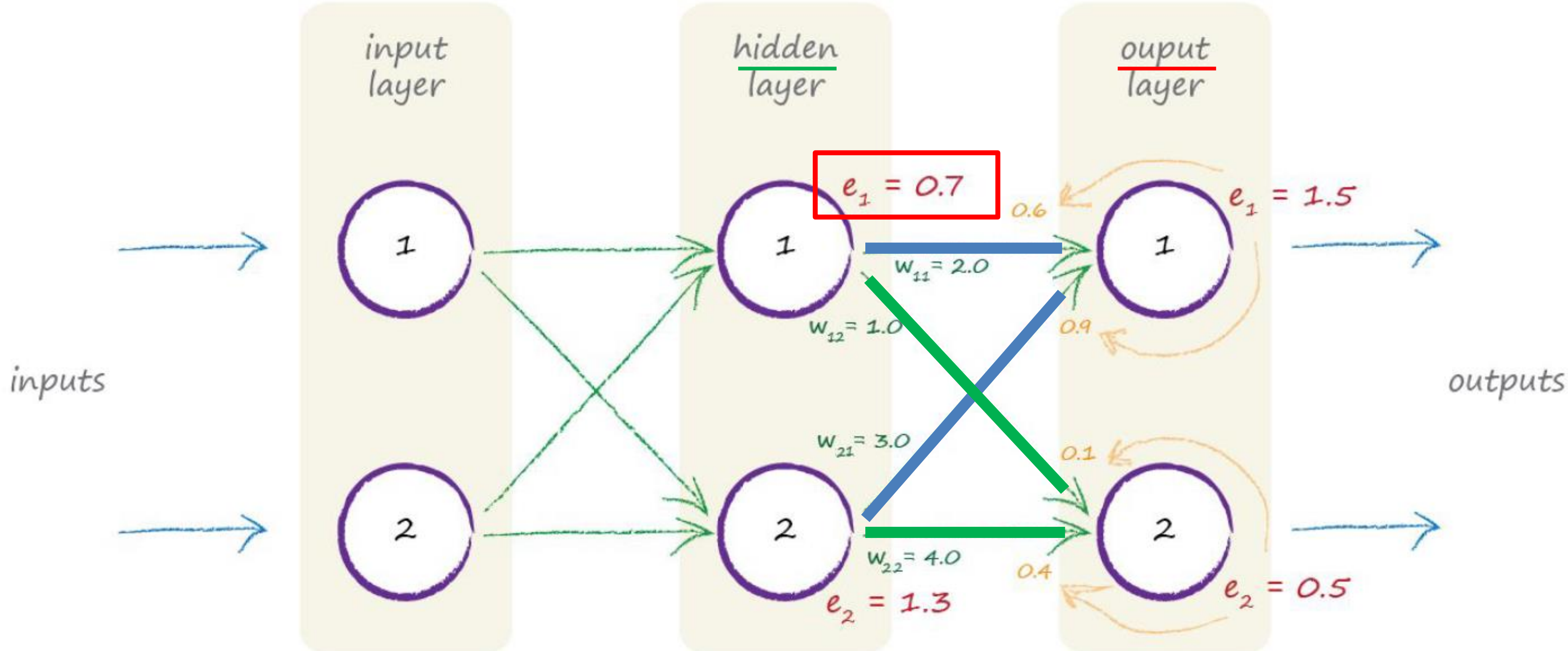


ref : Make your own neural network - Triq Rashid

다중 계층에서의 오차의 역전파



ref : Make your own neural network - Triq Rashid

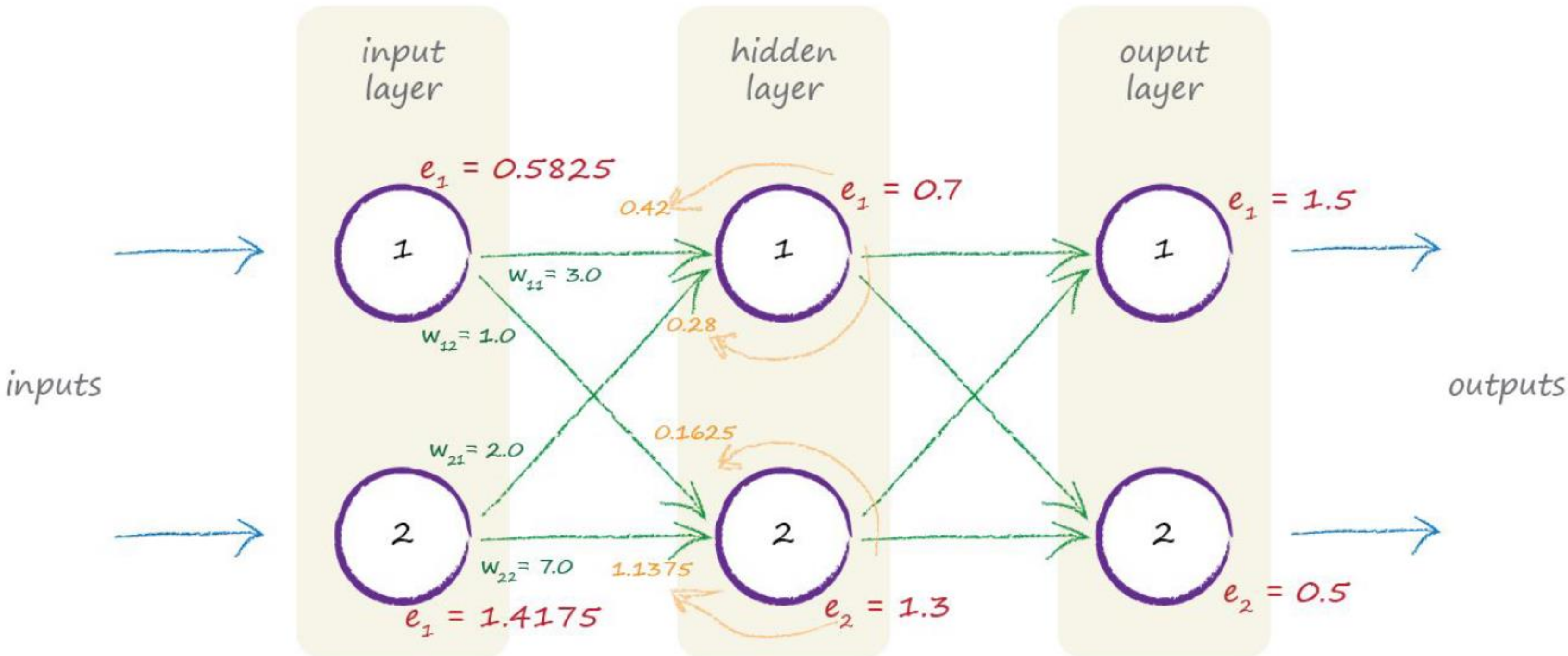


- $e_{hidden,\,node1}$ = 연결노드 w_{11}, w_{12}로 나뉘어 전달되는 오차의 합

$$= e_{output,\ node1} * \frac{w_{11}}{w_{11}+w_{21}} + e_{output,\,node2} * \frac{w_{12}}{w_{12}+w_{22}}$$

$$= 1.5 * \frac{2.0}{2.0+3.0} + 0.5 * \frac{1.0}{1.0+4.0} = \mathbf{0.6 + 0.1 = 0.7}$$

- $e_{hidden,\,node2} = \mathbf{0.9 + 0.4 = 1.3}$



- $e_{input,\, node1} = 0.42 + 0.1625 = 0.5825$
- $e_{input,\, node2} = 0.28 + 1.1375 = 1.4175$

행렬곱을 이용한 오차의 역전파

$$error_{output} = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

$$error_{hidden} = \begin{pmatrix} \frac{w_{11}}{w_{11}+w_{21}} & \frac{w_{12}}{w_{12}+w_{22}} \\ \frac{w_{21}}{w_{21}+w_{11}} & \frac{w_{22}}{w_{22}+w_{12}} \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

※ 가중치가 크면 클수록 더 많은 출력오차 발생.분수에서 분모는 일종의 정규화 인자(normalizing factor).
오류의 크기만 잃을 뿐.

- 전치행렬(transpose matrix)

$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{pmatrix}^T$$

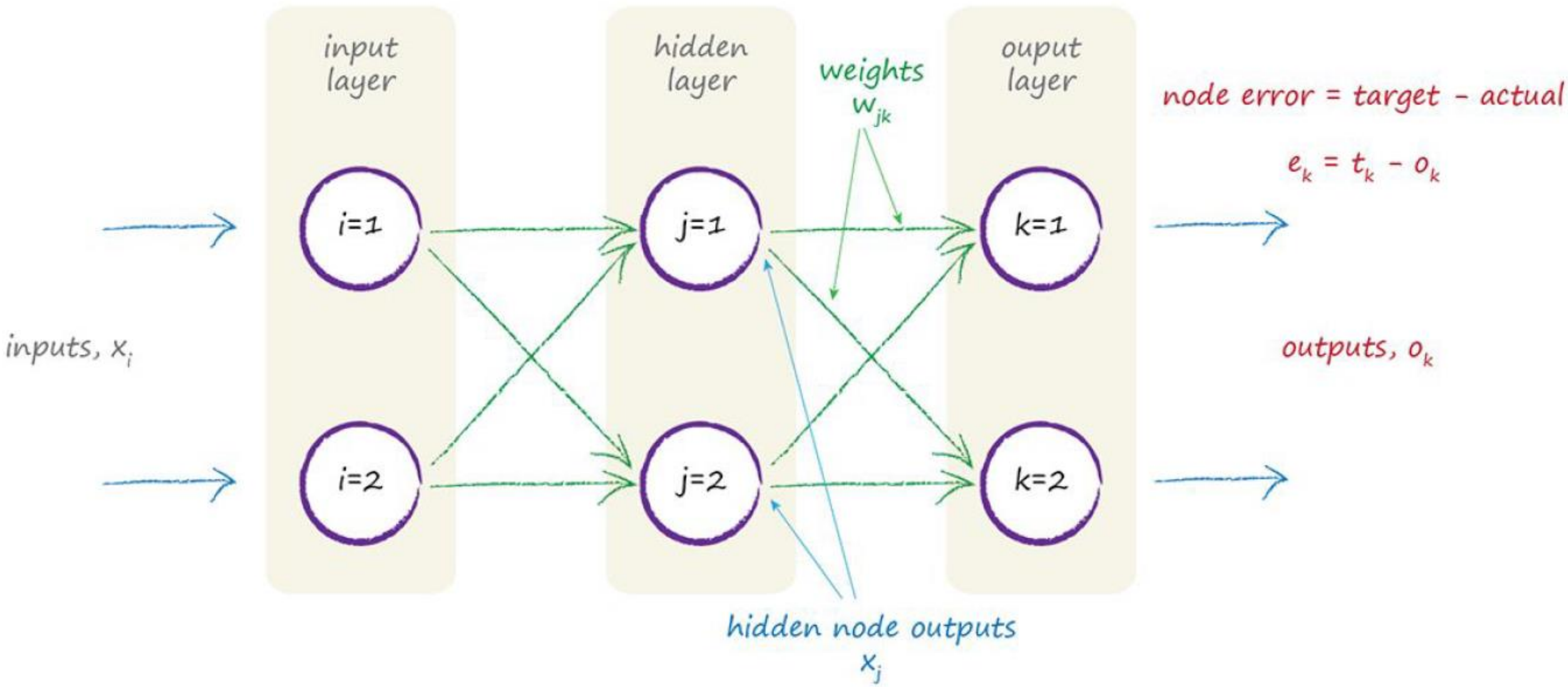
- 오차역전파의 행렬 표현

$$error_{hidden} = \boldsymbol{W^T}_{hidden\_output} \cdot error_{output}$$

가중치 조정

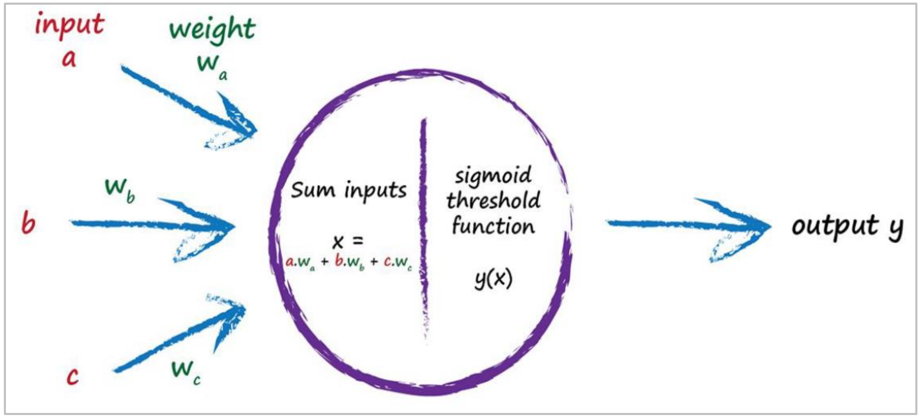
가중치 w_{jk}의 값이 변화함에 따라 오차 E의 값이 얼마만큼 변하는지 표현

$$\frac{\delta E}{\delta w_{jk}} = \frac{\delta}{\delta w_{jk}}(y_n - \hat{y}_n)^2 = \frac{\delta}{\delta w_{jk}}(t_n - o_n)^2$$



가중치 W_{jk} 조정 수식

$$
\begin{aligned}
\frac{\delta E}{\delta w_{jk}} &= \frac{\delta}{\delta w_{jk}}(t_n - o_n)^2 \\
&= \frac{\delta E}{\delta o_k} \cdot \frac{\delta o_k}{\delta w_{jk}} \\
&= -2(t_k - o_k) \cdot \frac{\delta o_k}{\delta w_{jk}} \\
&= -2(t_k - o_k) \cdot \frac{\delta}{\delta w_{jk}} \sigma\left(\sum_j w_{jk} \cdot o_j\right) \\
&= -2(t_k - o_k) \cdot \sigma\left(\sum_j w_{jk} \cdot o_j\right)\left(1 - \sigma\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot \frac{\delta}{\delta w_{jk}}\left(\sum_j w_{jk} \cdot o_j\right) \\
&= -2(t_k - o_k) \cdot \sigma\left(\sum_j w_{jk} \cdot o_j\right)\left(1 - \sigma\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot o_j
\end{aligned}
$$



$$
\frac{\boldsymbol{\delta E}}{\boldsymbol{\delta w_{jk}}} = -(\boldsymbol{t_k} - \boldsymbol{o_k}) \cdot \boldsymbol{\sigma}\left(\sum_j w_{jk} \cdot o_j\right)\left(1 - \sigma\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot \boldsymbol{o_j}
$$

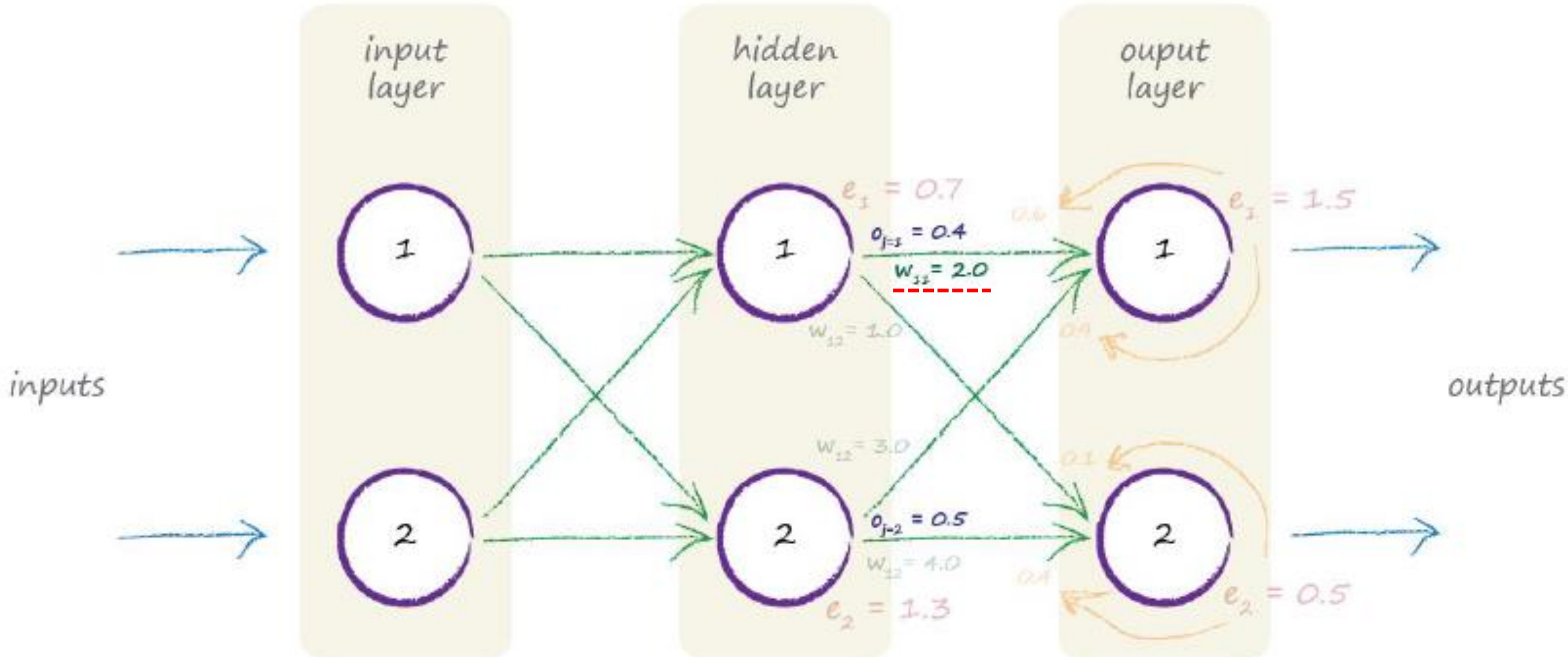
[$sigmoid\ function(\sigma)$미분]

$$
\frac{\delta}{\delta x}\sigma(x) = \sigma(x)(1 - \sigma(x))
$$

가중치 업데이트 예제(1/2)

$o_{j=1} = 0.4,\ o_{j=2} = 0.5$ 으로 가정하자

hidden layer와 output layer사이의 가중치 w_{11}을 **업데이트**해야 한다. **현재 2.0**



가중치 업데이트 예제(2/2)

$$\frac{\delta \boldsymbol{E}}{\delta \boldsymbol{w_{jk}}} = -(\boldsymbol{t_k} - \boldsymbol{o_k}) \cdot \boldsymbol{\sigma}\left(\sum_j w_{jk} \cdot o_j\right)\left(1 - \sigma\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot \boldsymbol{o_j}$$

- 첫 번째 항 $(t_k - o_k)$는 오차 : $e_1 = 1.5$
- sigmoid함수 내의 합

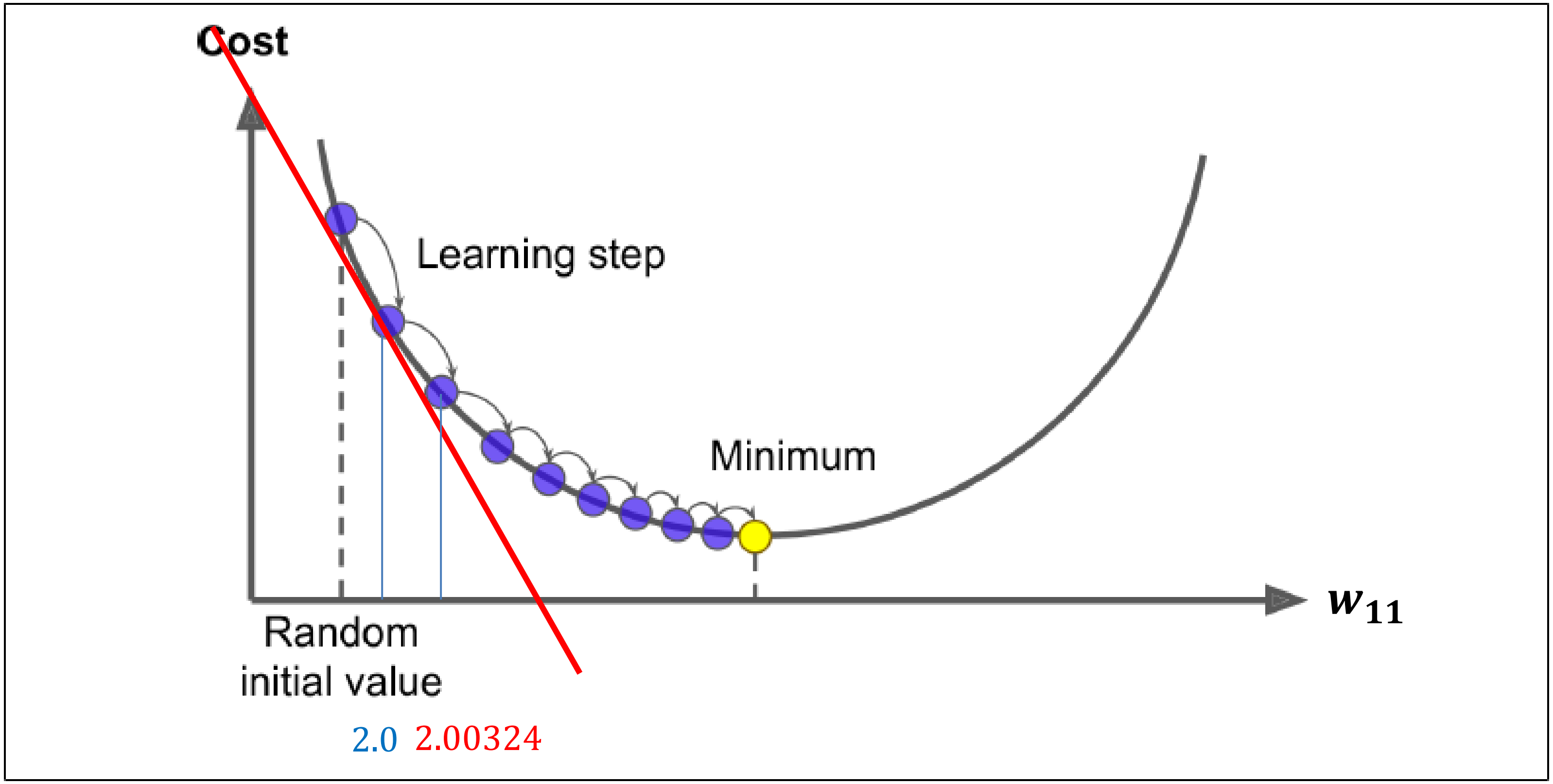
  $\sum_j w_{jk} \cdot o_j = (2.0 * 0.4)\ +\ (4.0 * 0.5)\ =\ 2.8$
- sigmoid함수 값

  $\sigma(2.8) = \frac{1}{1 + e^{-2.8}} = 0.943$

  $\sigma(2.8)(1 - \sigma(2.8)) = 0.943 * (1 - 0.943) = 0.054$
- 마지막 항 $o_j = 0.4$
- 변화량

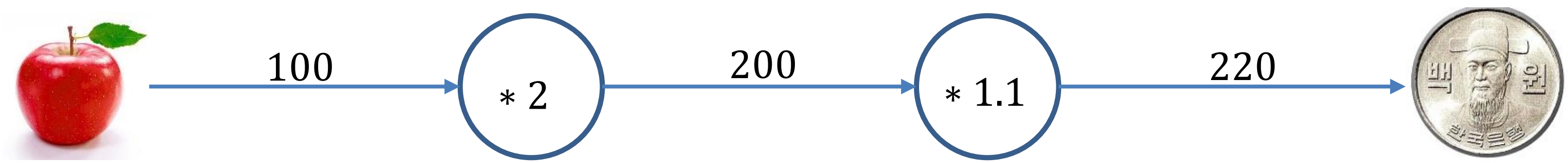
  $\frac{\delta E}{\delta w_{jk}} = -(1.5 * 0.054 * 0.4) = -0.0324$

  $if\ learning\ rate\ =\ 0.1\ then\ (0.1 * -0.0324) = \boldsymbol{-0.00324}$
- $\boldsymbol{w_{11} = 2.0 - (-0.00324) = 2.00324}$

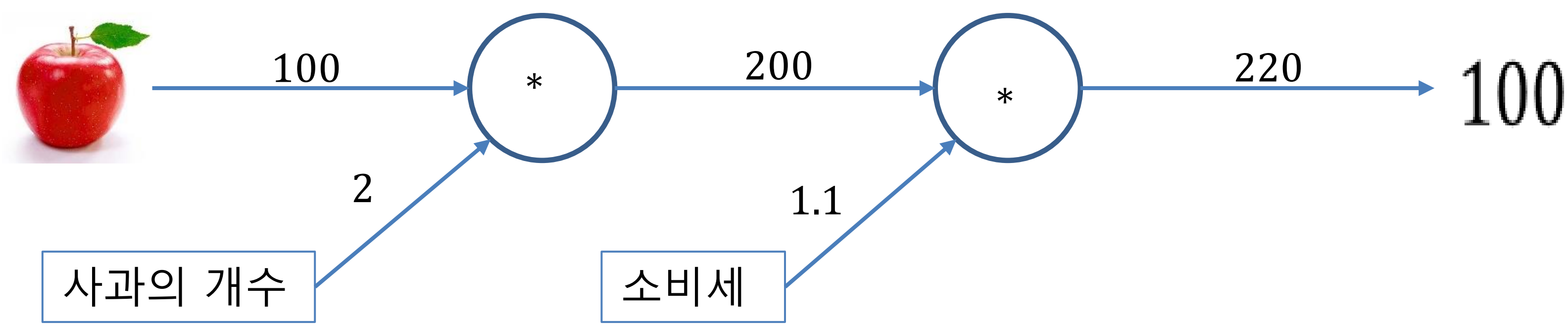

Cost
Learning step
Minimum
w_{11}
Random
initial value
2.0
2.00324

계산그래프(computation graph)

- 1개에 100원인 사과를 2개 산다면 지불금액은? 단 소비세 10% 부과

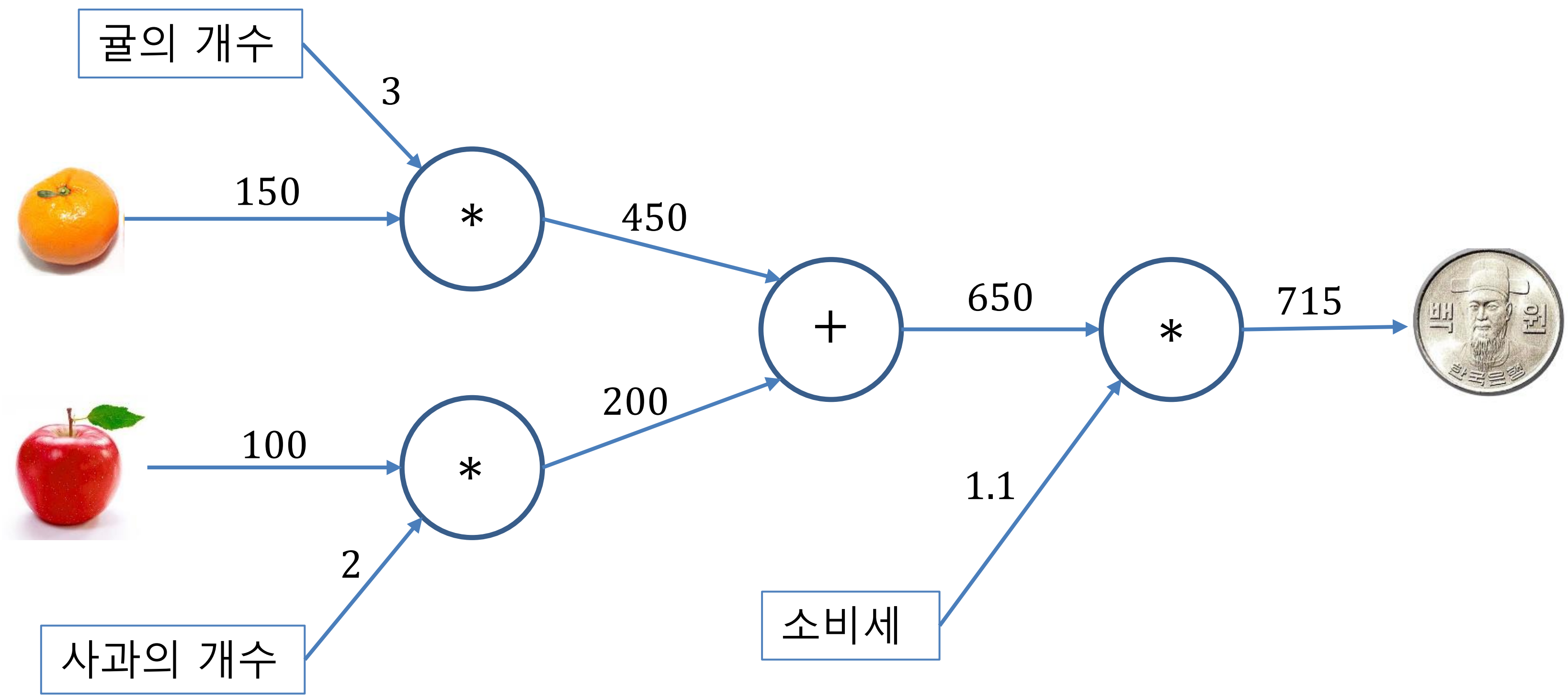


- 연산자 분리

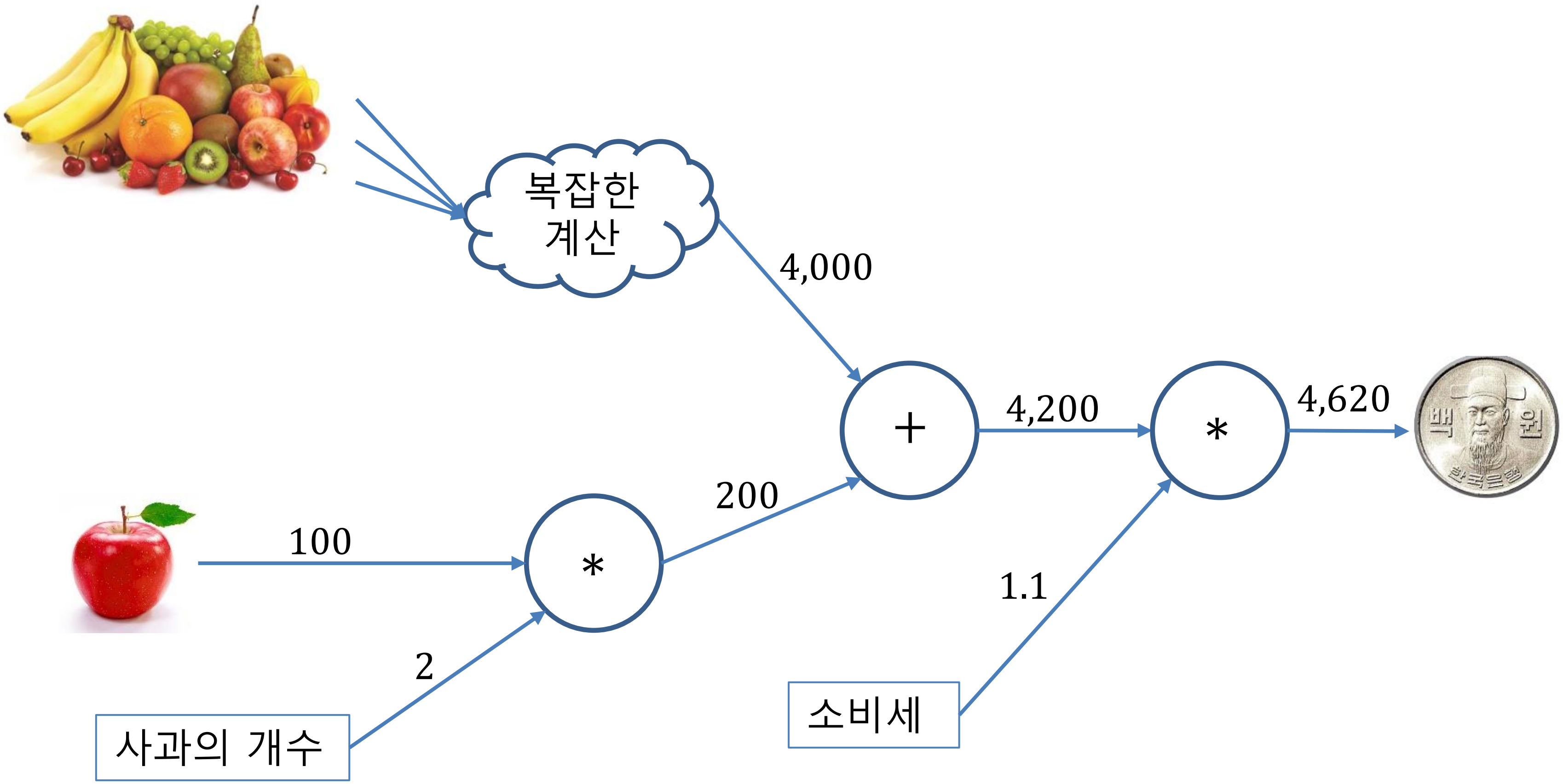


ref :밑바닥부터 시작하는 딥러닝, 한빛미디어

- 사과 2개, 귤 3개에 산다면? 단, 사과 1개에 100원, 귤 1개에 150원



ref :밑바닥부터 시작하는 딥러닝, 한빛미디어

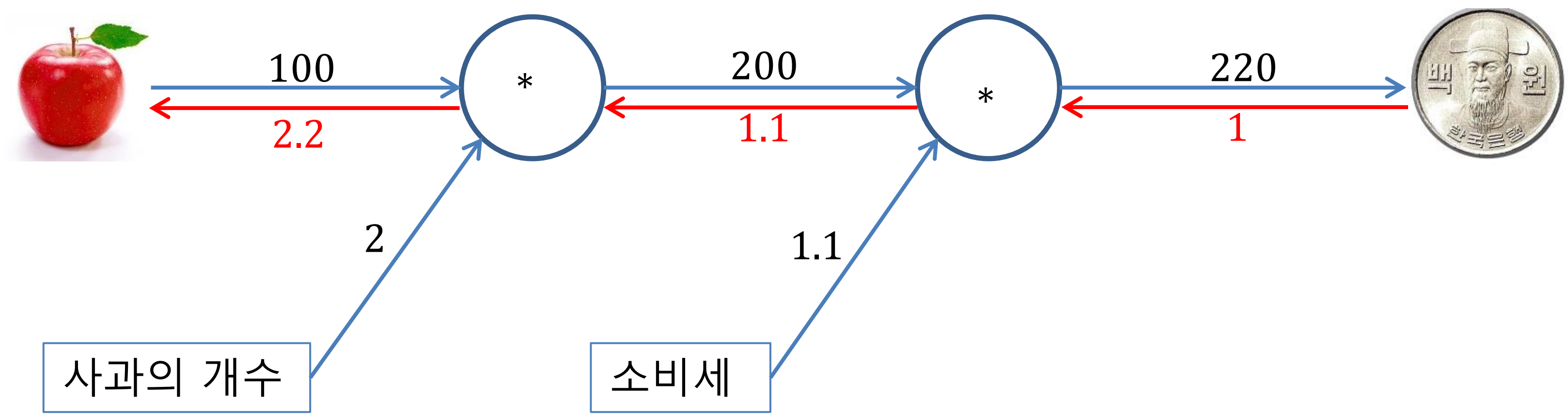


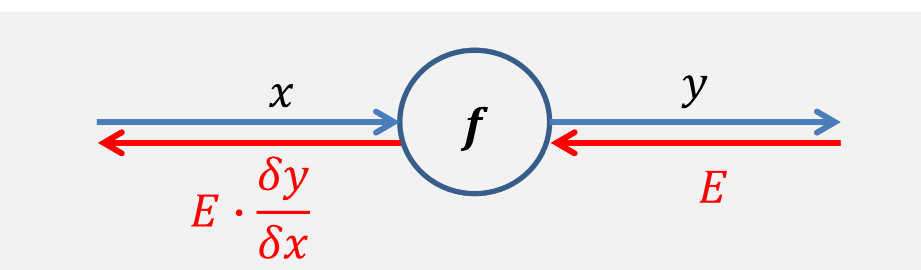
ref :밑바닥부터 시작하는 딥러닝, 한빛미디어

사과 가격이 오르면 최종금액에 어떤 영향을 끼치는가?

사과값 x, 지불금액 E 이라 하면 : $\dfrac{\delta E}{\delta x}$

즉, 미분값을 전달 : 사과가 1원 오르면 금액은 2.2원 오른다는 뜻

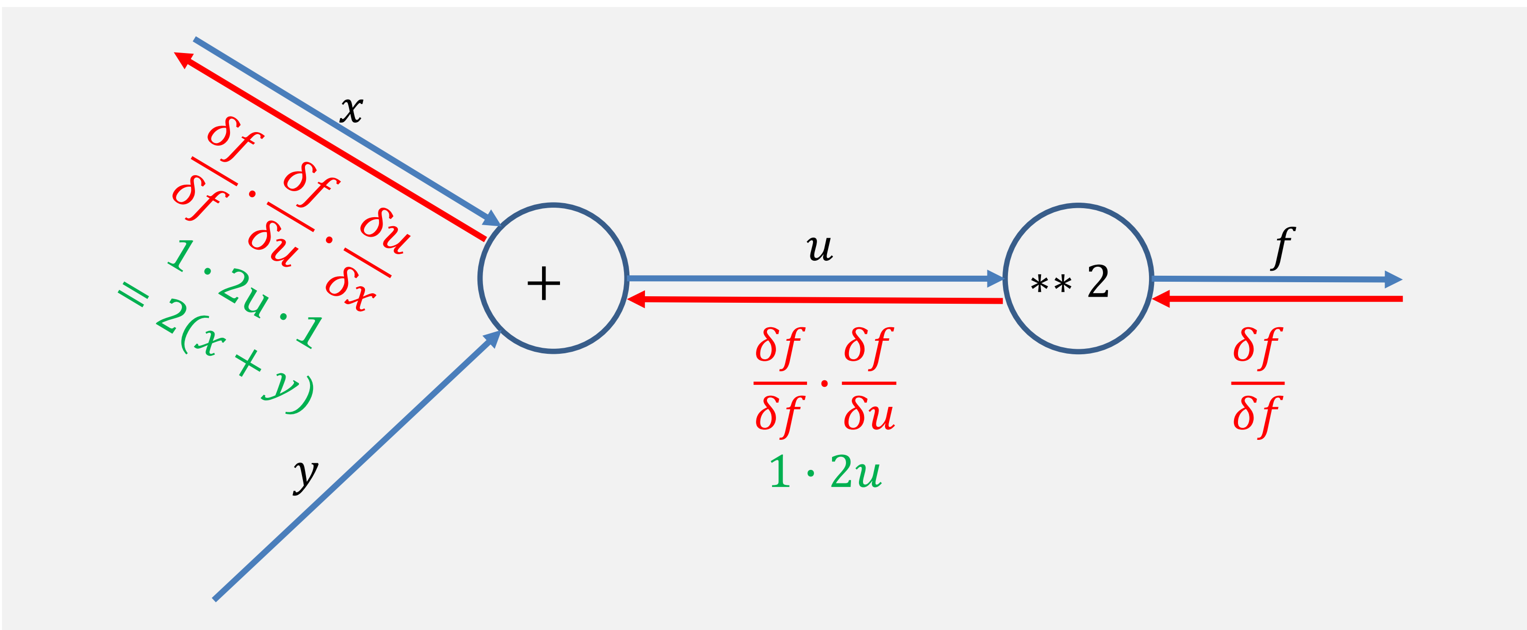




연쇄법칙(chain rule)과 계산그래프

$$f = (x+y)^2$$
$$f = u^2$$
$$u = x + y$$

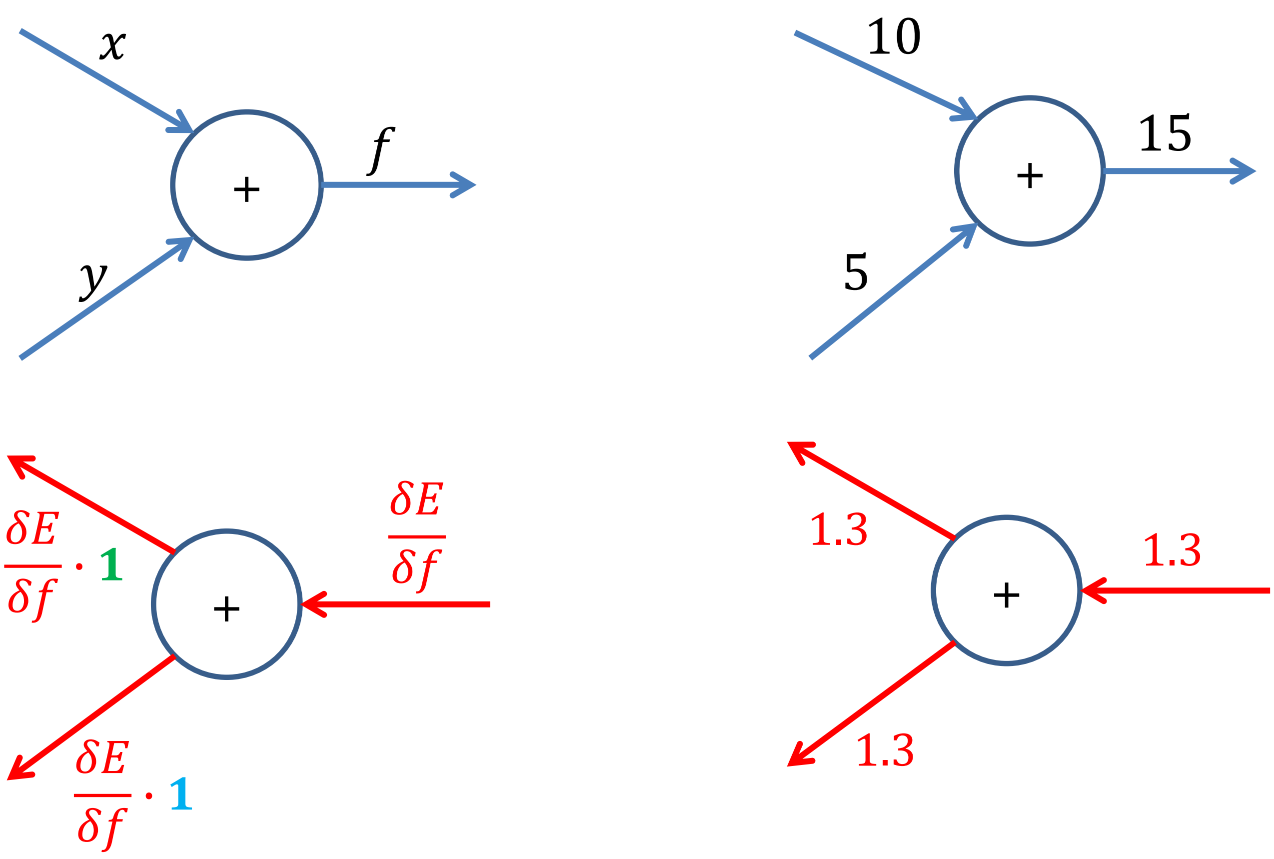
$$\frac{\delta f}{\delta x} = \frac{\delta f}{\delta u}\cdot\frac{\delta u}{\delta x} = 2u \cdot 1 = 2(x+y)$$



덧셈 노드의 역전파

$$f = x + y$$

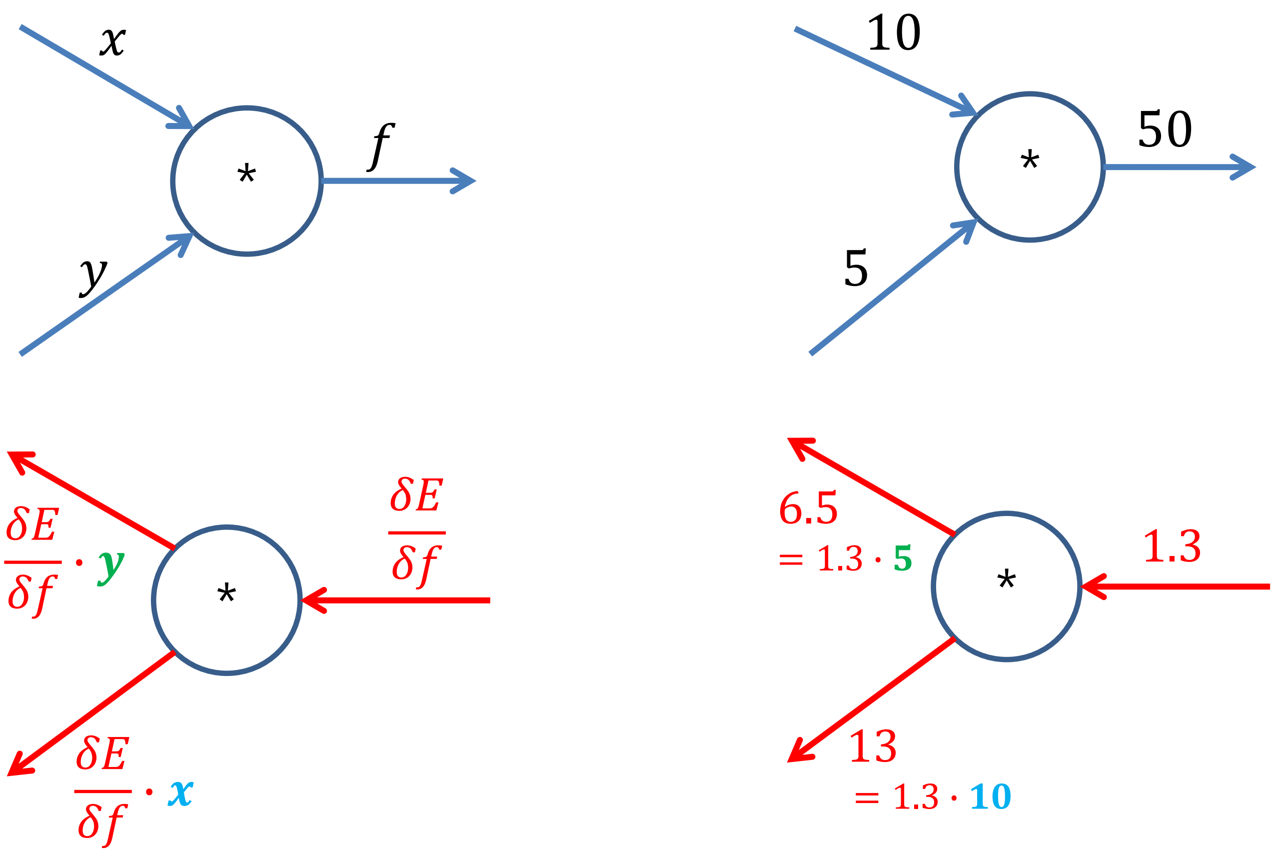
$$\frac{\delta f}{\delta x} = 1, \qquad \frac{\delta f}{\delta y} = 1$$



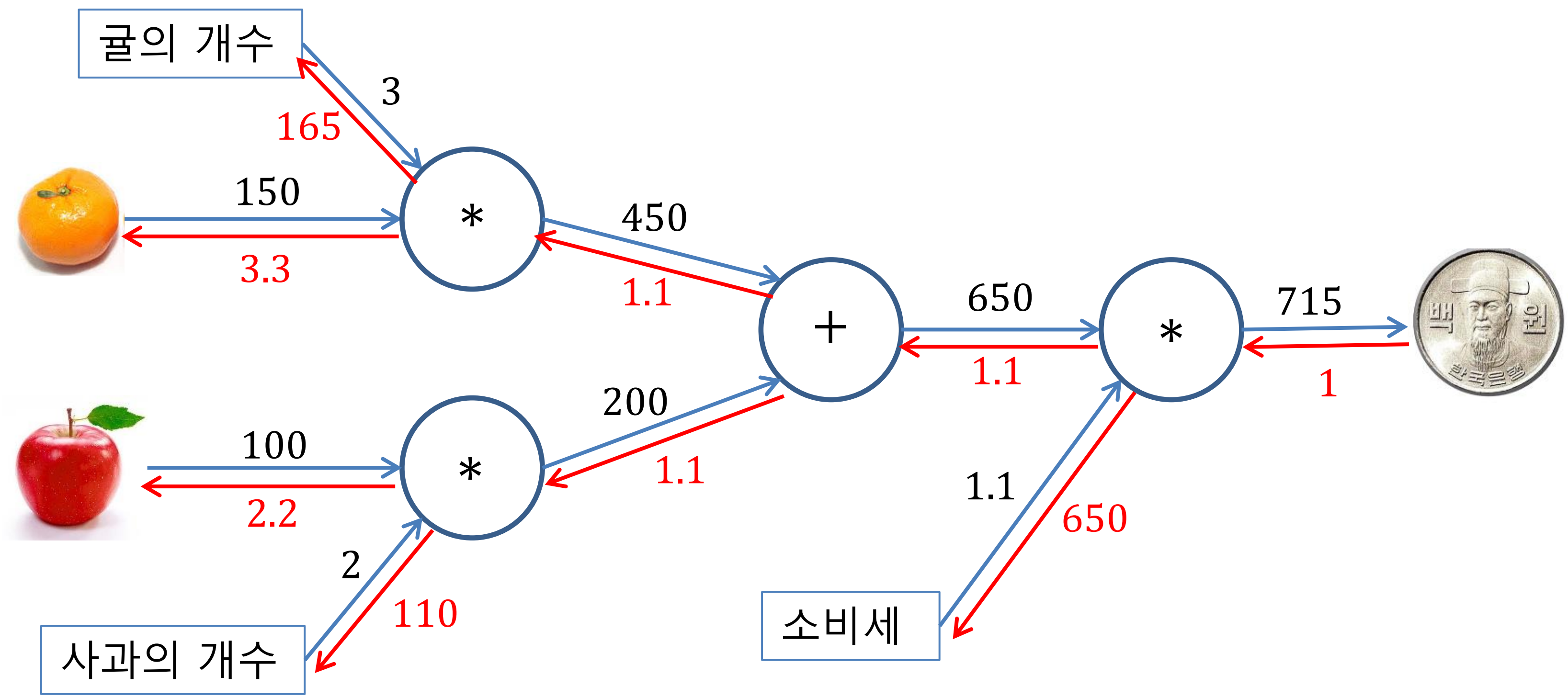
곱셈 노드의 역전파

$$f = x * y$$

$$\frac{\delta f}{\delta x} = \boldsymbol{y}, \qquad \frac{\delta f}{\delta y} = \boldsymbol{x}$$



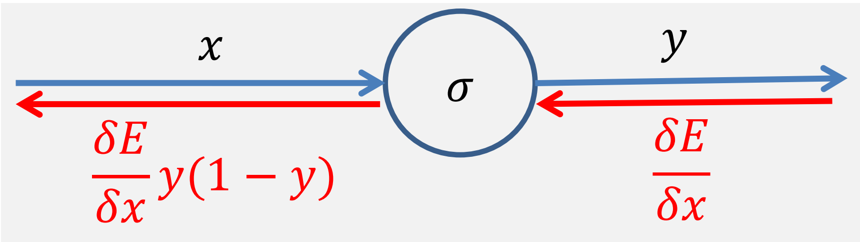
- 사과 예



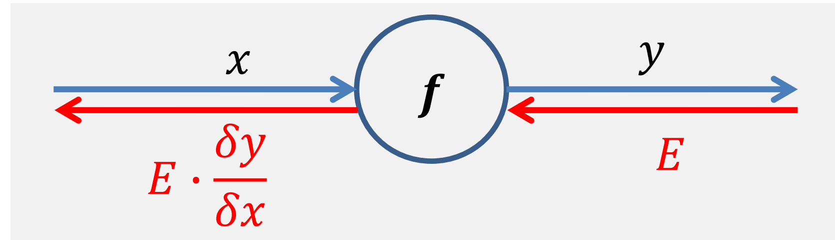
Ref :밑바닥부터 시작하는 딥러닝, 한빛미디어

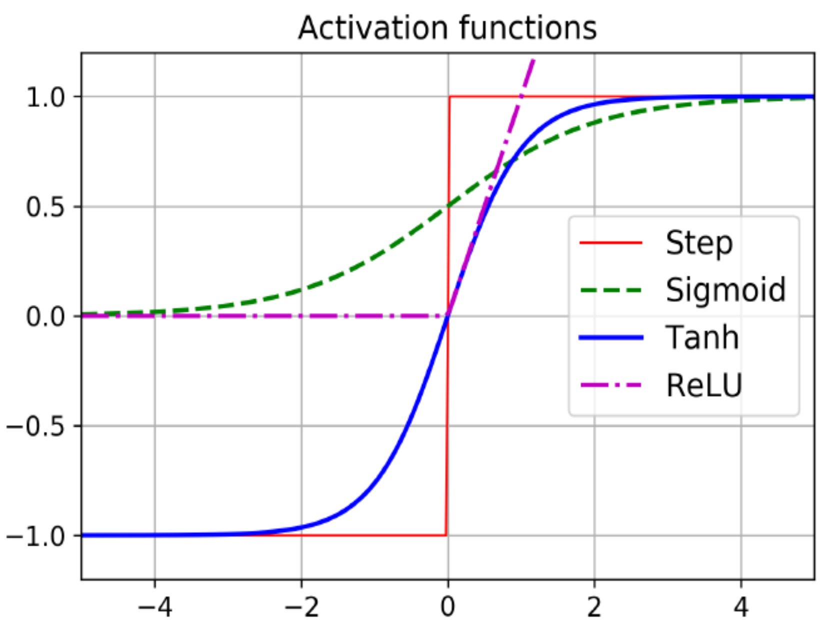
활성화 함수

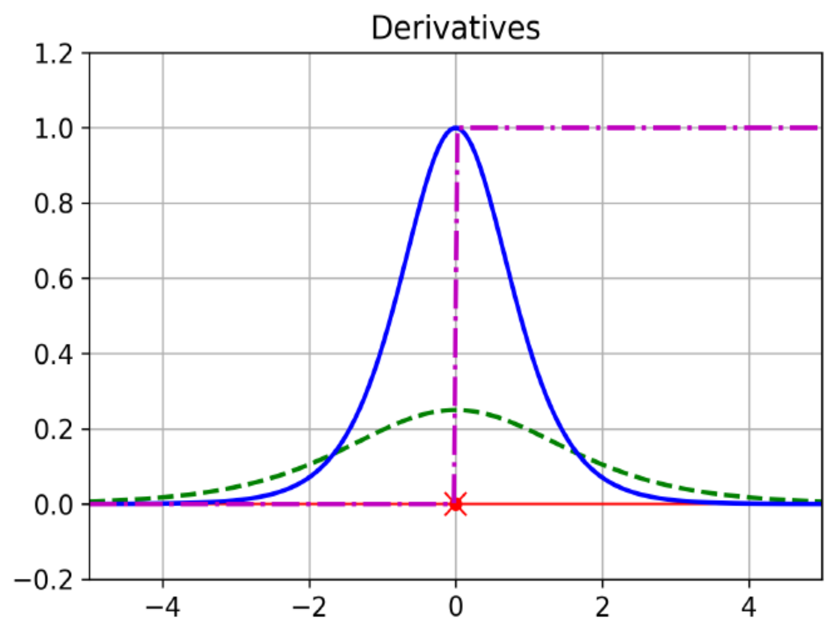
Sigmoid의 계산그래프



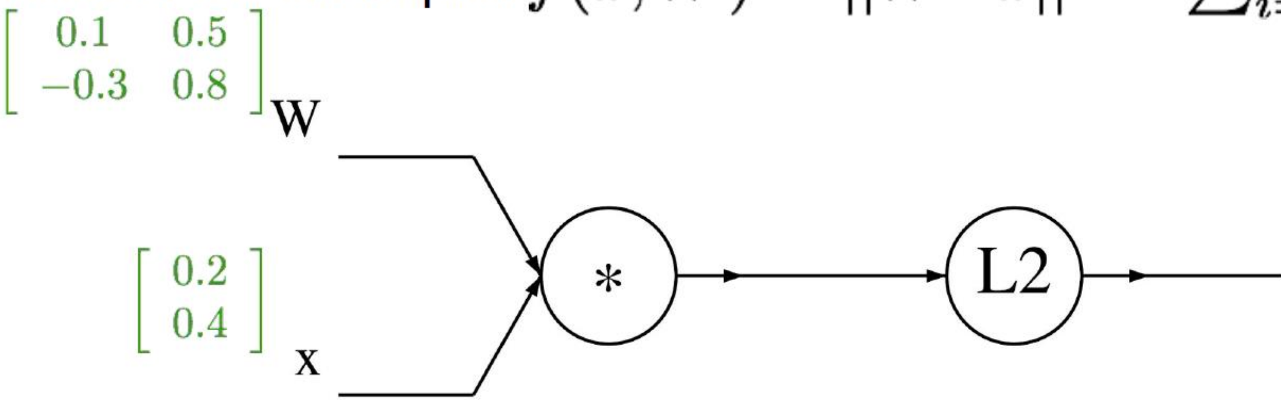
역전파 계산그래프







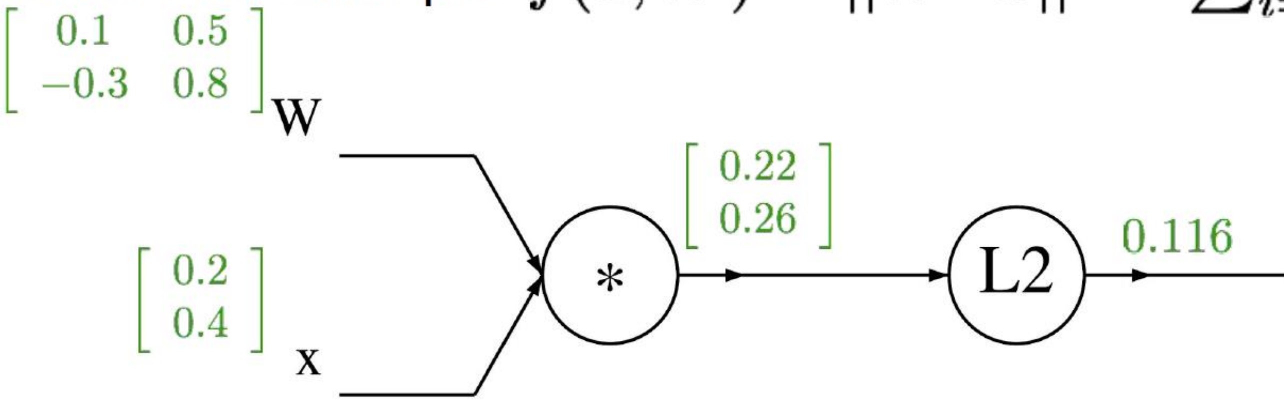
A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$

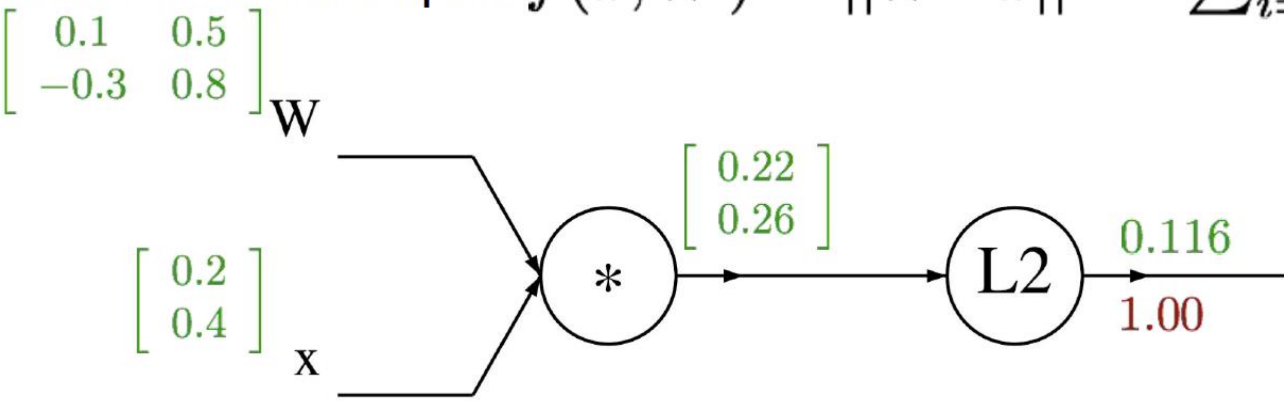


$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix} \cdot \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



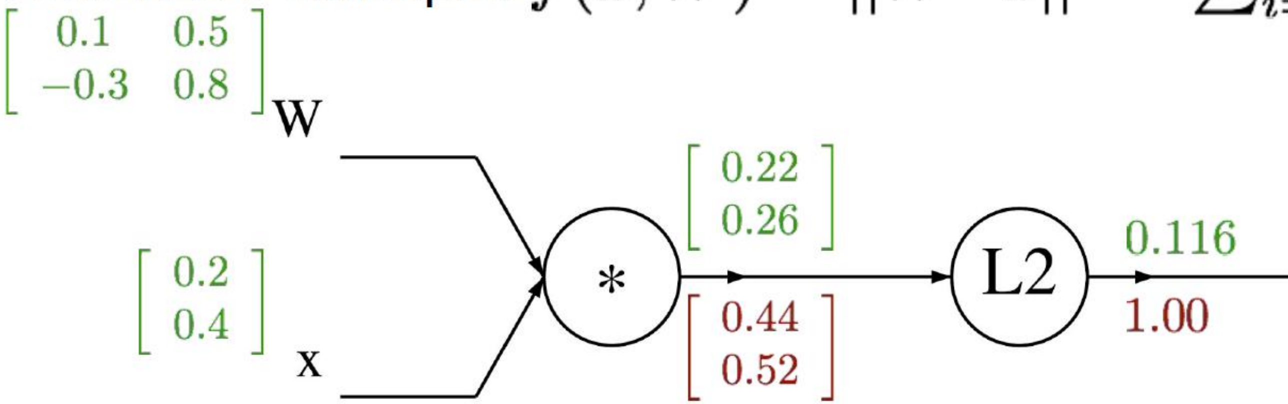
$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\boxed{\nabla_q f = 2q}$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



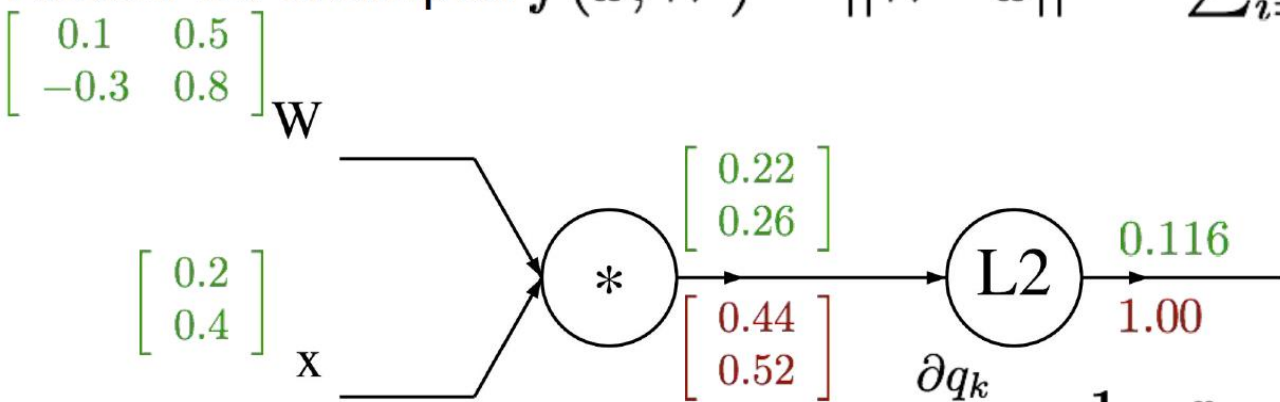
$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\boxed{\nabla_q f = 2q}$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$

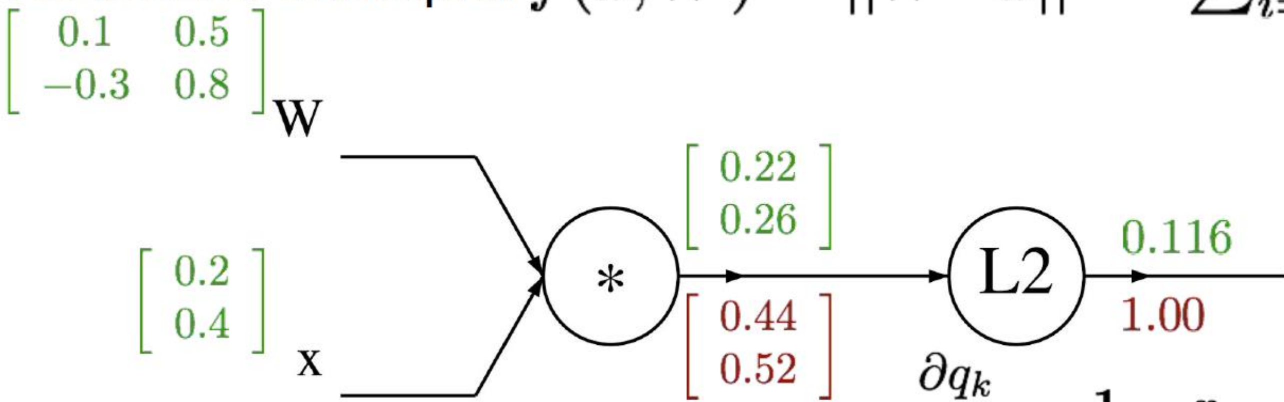


$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



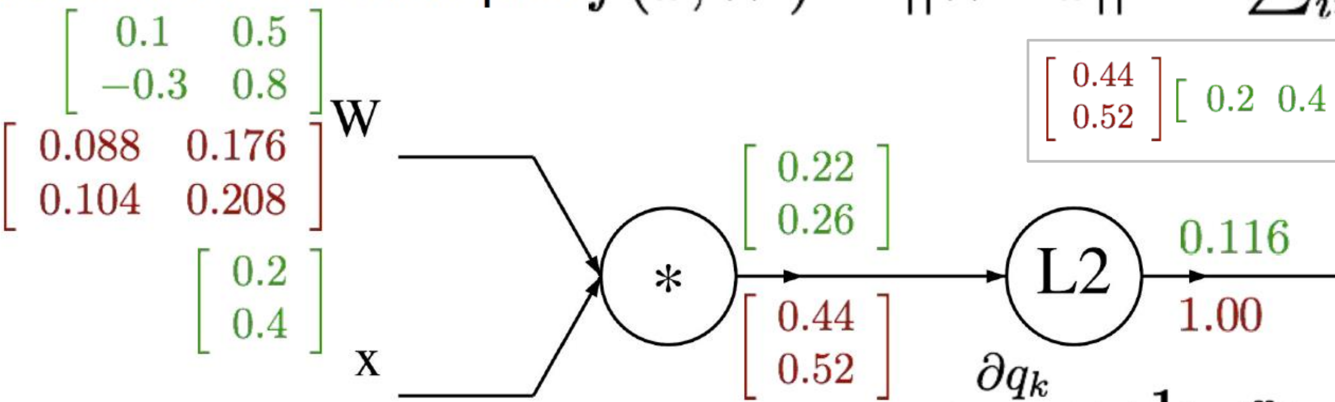
$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$ → q에 대하여 W미분

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\begin{aligned} \frac{\partial f}{\partial W_{i,j}} &= \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}} \\ &= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j) \\ &= 2q_i x_j \end{aligned}$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix} \begin{bmatrix} 0.2 & 0.4 \end{bmatrix} \quad \begin{matrix} 0.44 * 0.2 & 0.44 * 0.4 \\ 0.52 * 0.2 & 0.52 * 0.4 \end{matrix}$$

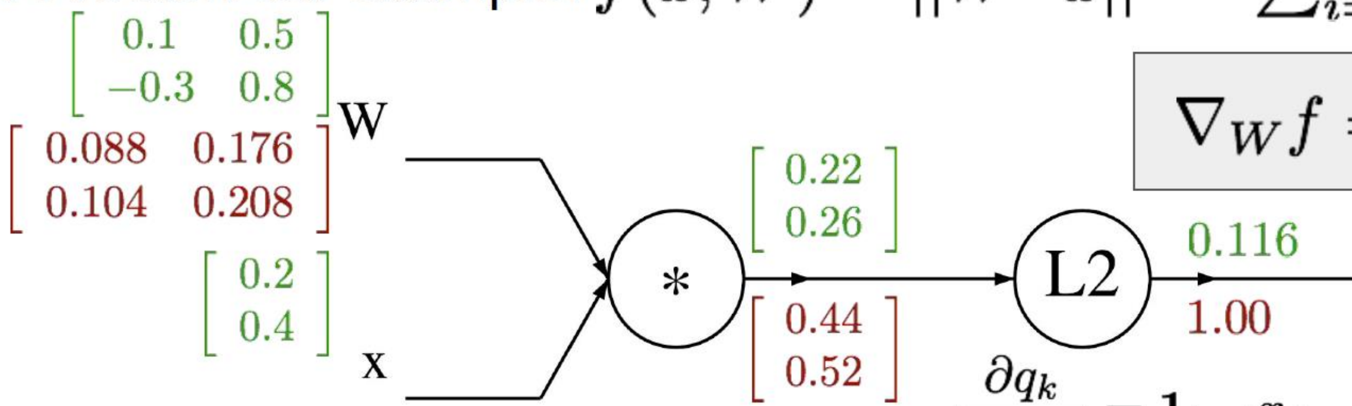
$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\begin{aligned} \frac{\partial f}{\partial W_{i,j}} &= \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}} \\ &= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j) \\ &= 2q_i x_j \end{aligned}$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



$$\boxed{\nabla_W f = 2q \cdot x^T}$$

$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$

$$\frac{\partial f}{\partial W_{i,j}} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}}$$

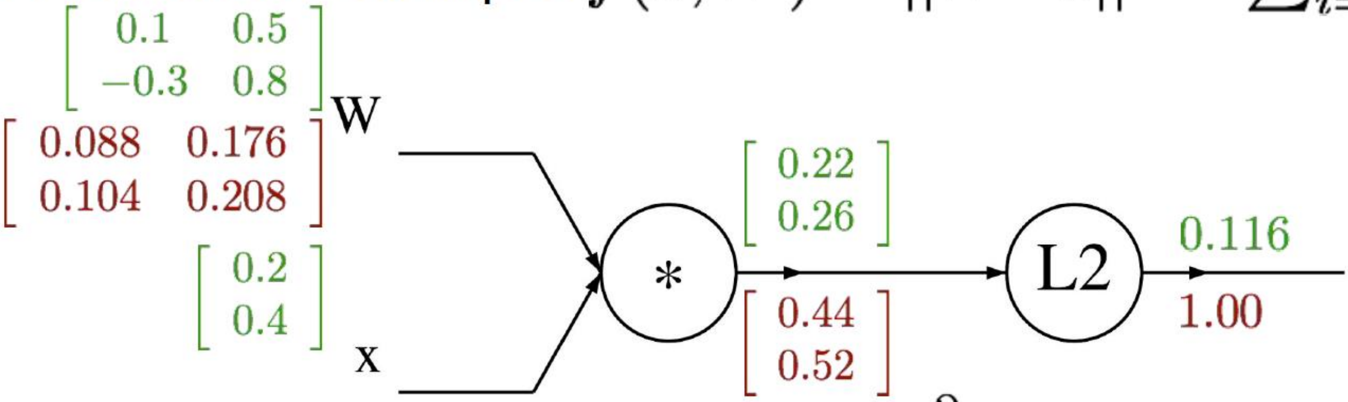
$$= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j)$$

$$= 2q_i x_j$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

A vectorized example: $f(x,W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$

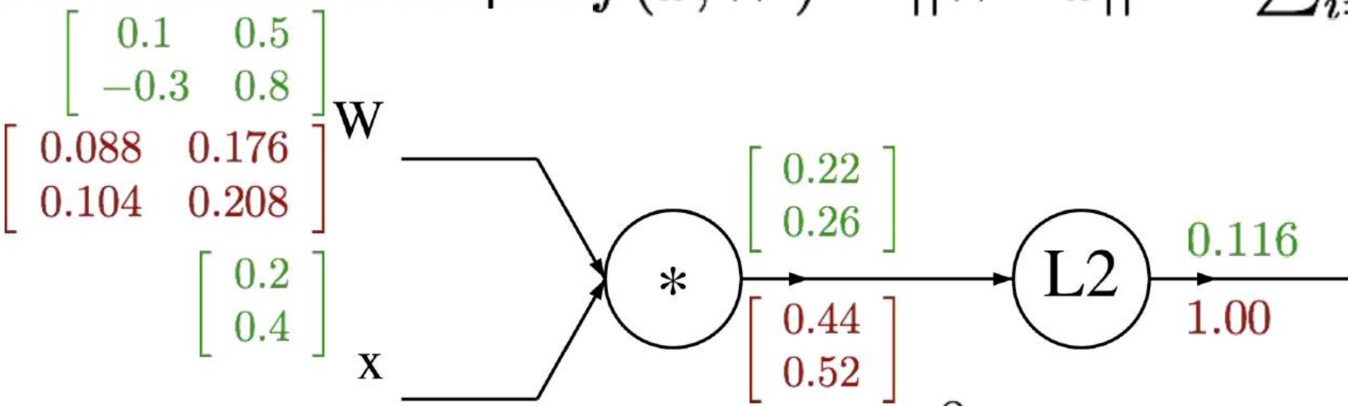


$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$



$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

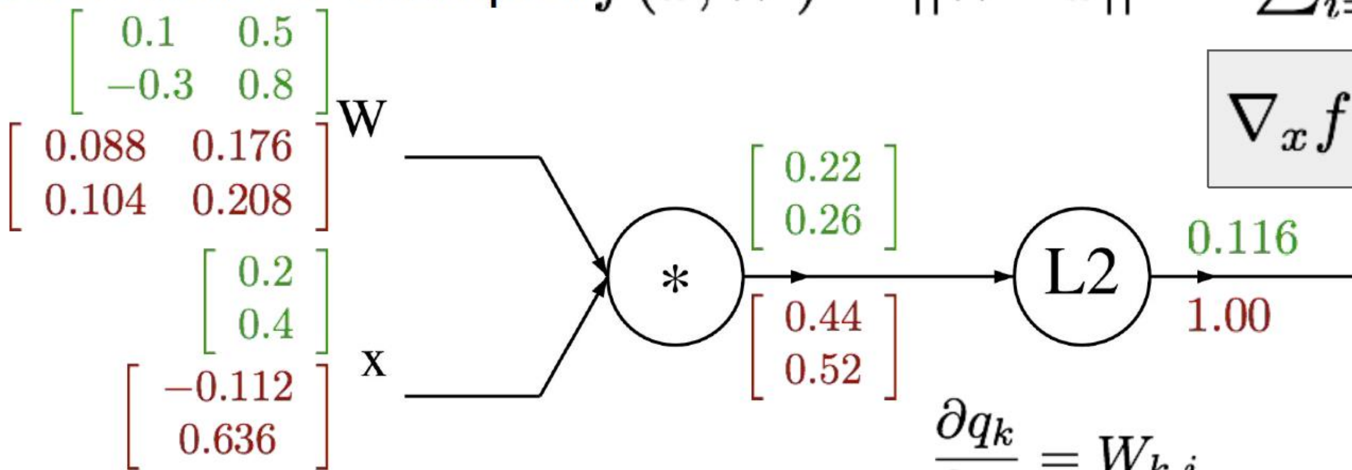
$$\frac{\partial f}{\partial x_i} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i}$$

$$= \sum_k 2 q_k W_{k,i}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$

$$\boxed{\nabla_x f = 2W^T \cdot q}$$



$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$\frac{\partial f}{\partial x_i} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i}$$

$$= \sum_k 2 q_k W_{k,i}$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

Intermediate **Variables**

(forward propagation)

$$h_1 = xW_1 + b_1$$

$$z_1 = \sigma(h_1)$$

$$z_2 = z_1 W_2 + b_2$$

$$Loss = (z_2 - y)^2$$

Intermediate **Gradients**

(backward propagation)

$$\frac{\partial h_1}{\partial x} = W_1^T$$

$$\frac{\partial z_1}{\partial h_1} = \sigma'(h_1) = z_1 \circ (1 - z_1)$$

$$\frac{\partial z_2}{\partial z_1} = W_2^\top$$

$$\frac{\partial Loss}{\partial z_2} = 2(z_2 - y)$$