Ch 5. Back Propagation

- Loss function’s gradient for weight matrix was calculated by numerical differential.

- Backpropagation (BP) is introduced to efficiently calculate Loss function’s gradient for weight

- Computation Graph of BP
  - https://karpathy.medium.com/yes-you-should-understand-backprop-e2f06eab496b
  - http://cs231n.stanford.edu/slides/2019/cs231n_2019_lecture04.pdf

Numerical Derivative vs. Chain Rule + Derivative Formula

$$\mathbb{R} \xrightarrow{f} \mathbb{R} \xrightarrow{g} \mathbb{R} \xrightarrow{h} \mathbb{R}$$

Numerical derivative : $\dfrac{h(g(f(x+\varepsilon))) - h(g(f(x)))}{\varepsilon}$

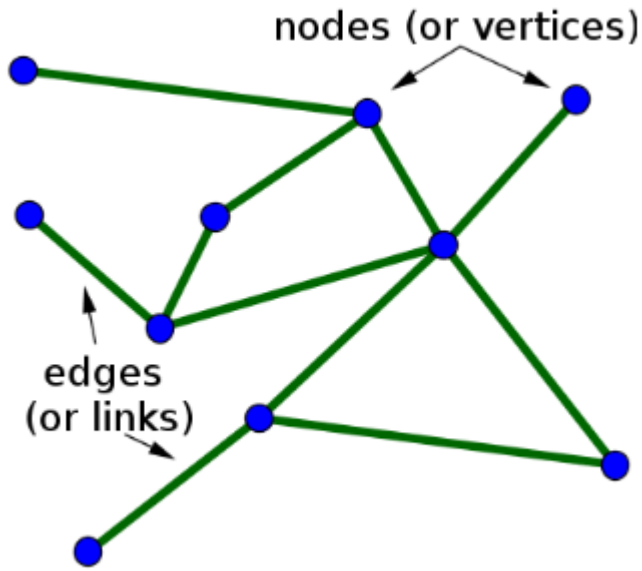
Chain rule + Derivative formula : $h'(g(f(x)))g'(f(x))f'(x)$

e.g., $f(x) = 2x+1,\ g(y) = y^2,\ h(z) = \sin(z)$

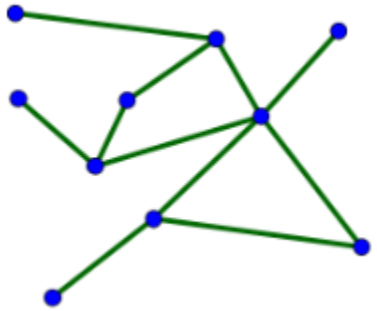
Numerical derivative : $\dfrac{\sin\left((2(x+\varepsilon)+1)^2\right) - \sin\left((2x+1)^2\right)}{\varepsilon}$

Chain rule + Derivative formula : $\cos\left((2x+1)^2\right) \times 2(2x+1) \times 2$

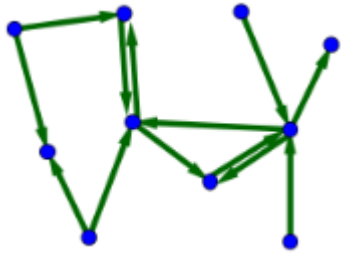
Graph

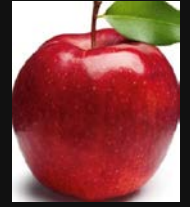


undirected graph :

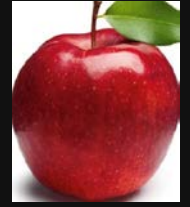


directed graph :



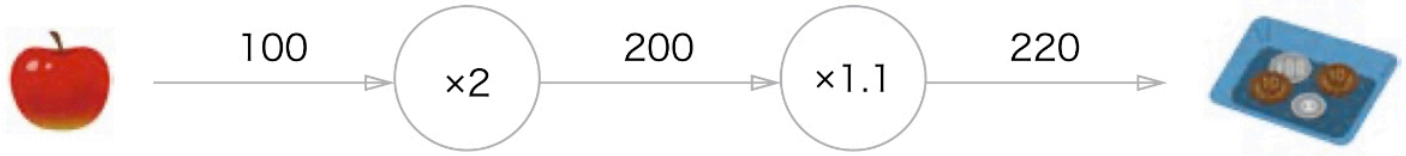
Q) Hyunbin bought  . One  costs ₩100.

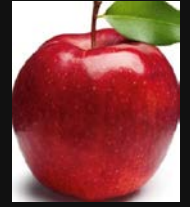
Consumption tax is 10%. How much should he pay for them?

Q) Hyunbin bought  . One  costs ₩100.

Consumption tax is 10%. How much should he pay for them?

그림 5-1 계산 그래프로 풀어본 문제 1의 답



Q) Hyunbin bought  . One  costs ₩100.

Consumption tax is 10%. How much should he pay for them?

그림 5-1 계산 그래프로 풀어본 문제 1의 답

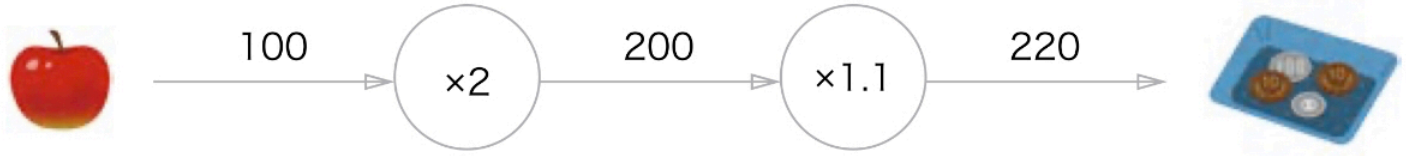
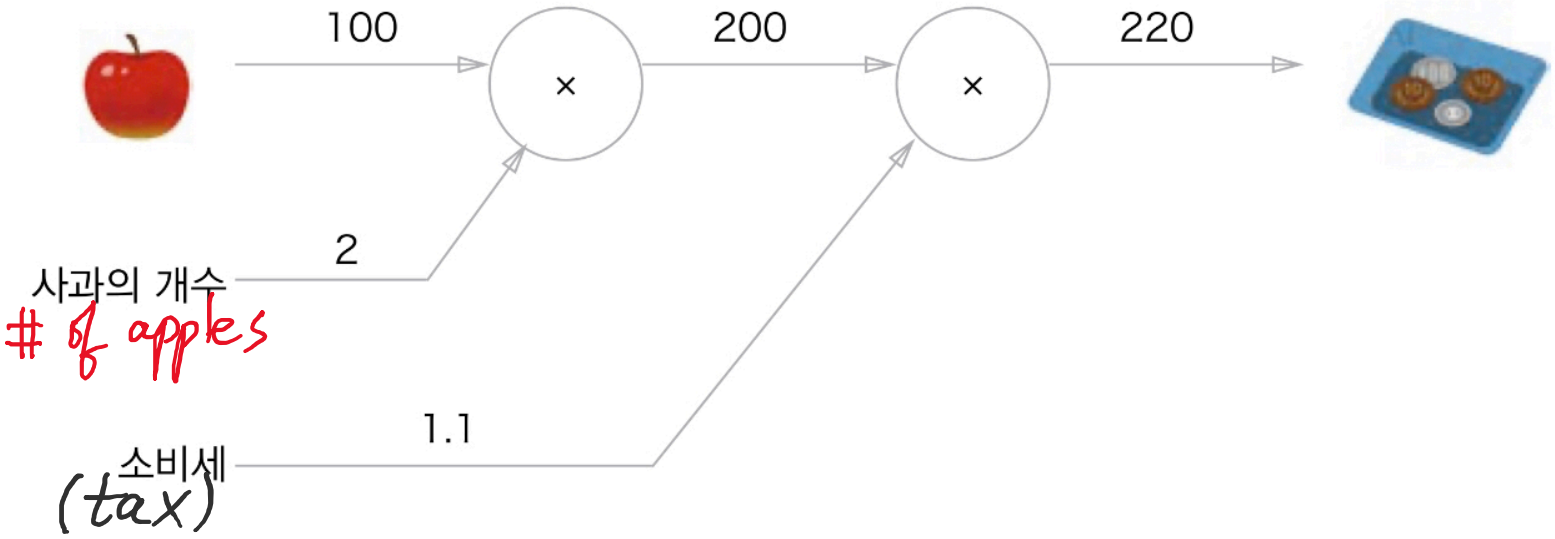


그림 5-2 계산 그래프로 풀어본 문제 1의 답 : '사과의 개수'와 '소비세'를 변수로 취급해 원 밖에 표기



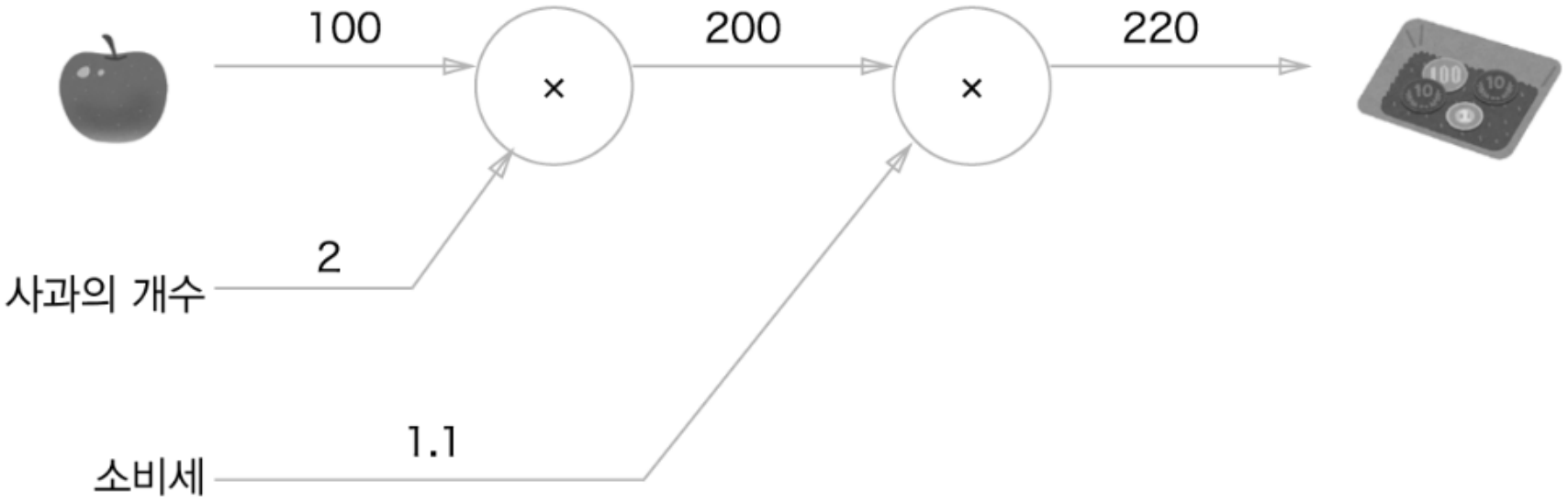
Computation Graph: Forward Propagation

Forward propagation flow:

1. **Build a computation graph**
2. **Compute from left to right**

Apple's price : 100, # of apples : 2, Tax : 10%

Total price : $100 \times 2 \times 1.1 = 220$

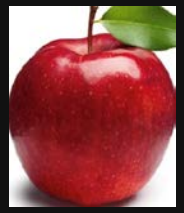


Q) Hyunbin bought  .

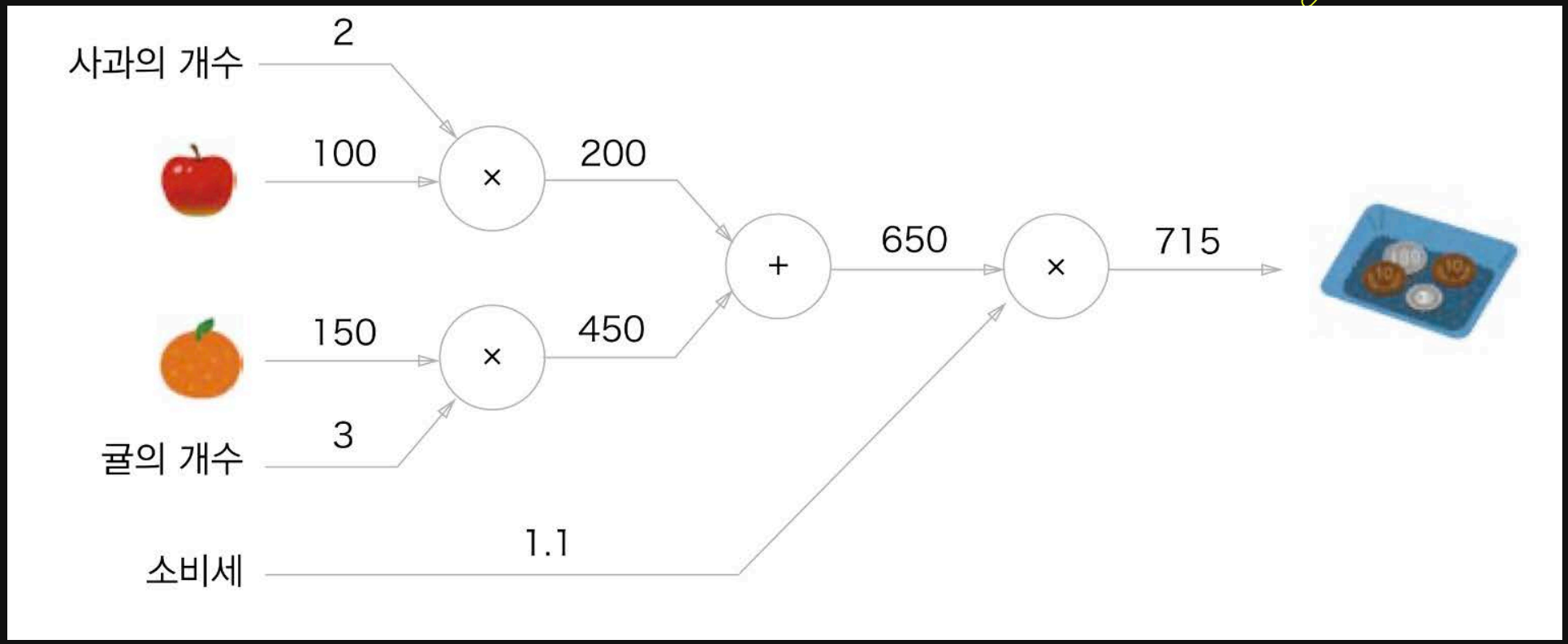
One  costs ₩100, one  costs ₩150.

Consumption tax is 10%. How much should he pay for them?

Q) Hyunbin bought  .

One  costs ₩100, one  costs ₩150.

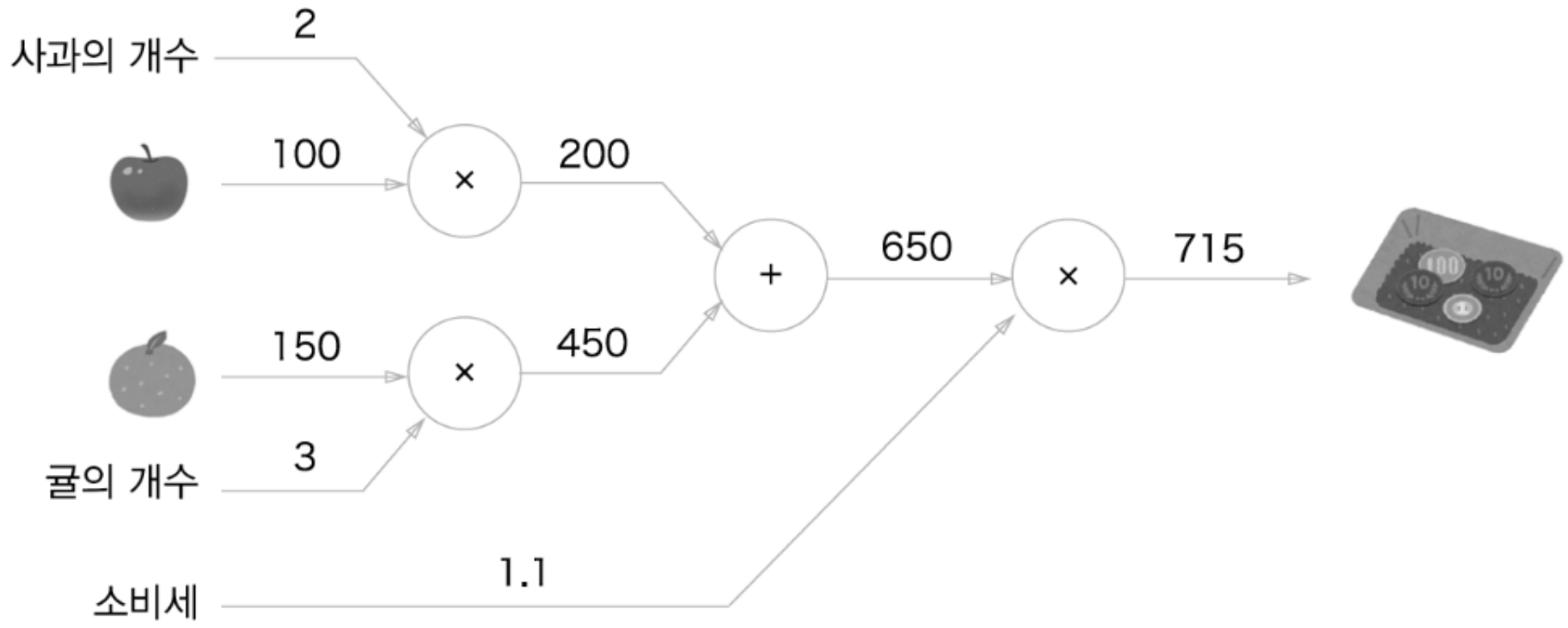
Consumption tax is 10%. How much should he pay for them?

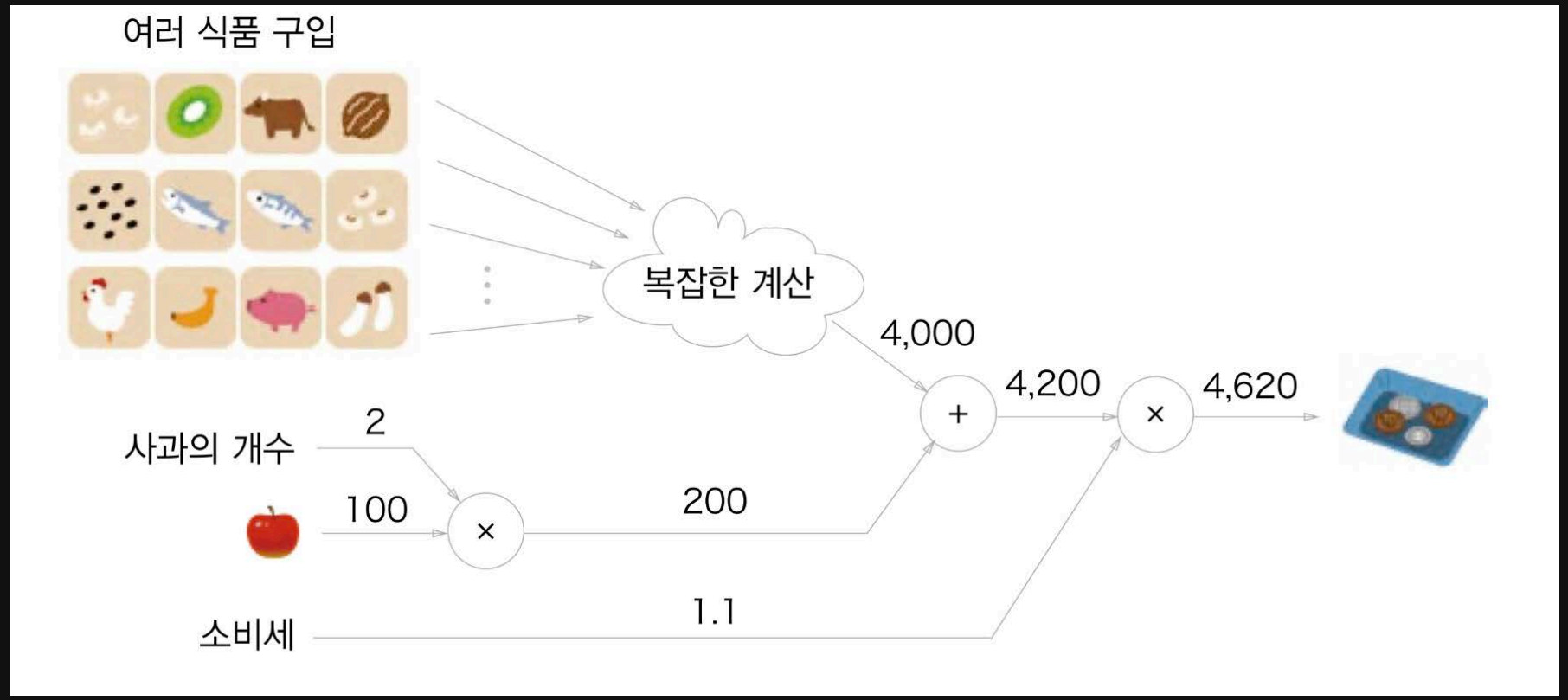


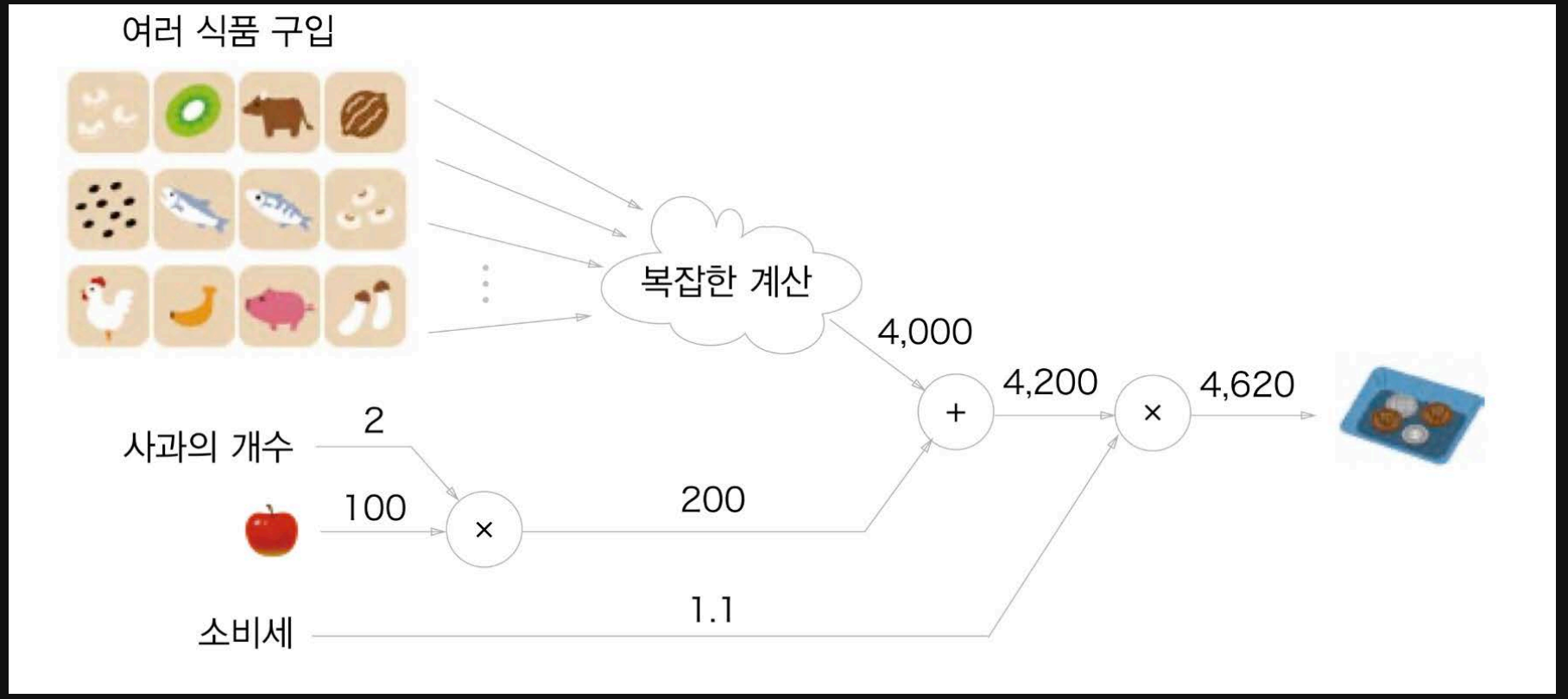
Computation Graph: Forward Propagation

Apple price : 100, # of apples : 2, orange price : 150, # of oranges : 3, tax : 10%

Total price : $(100 \times 2 + 150 \times 3) \times 1.1 = 715$




여러 식품 구입
복잡한 계산
4,000
+
4,200
×
4,620
사과의 개수
2
100
×
200
소비세
1.1



The important thing is that the computation on a certain node is very local. It does not care how the 4,000 is computed. At the plus (+) node, two inputs are 4,000 and 200. Its result is just 4,200. That's it.

At a node, just care for the input values.

Computation Graph: Back Propagation

Back propagation flow:

1. **Starting from the value of 1**
2. **It flows from right to left.**
3. **At multiplication nodes, another input values (black) are multiplied with the back-propagated (red) input.**

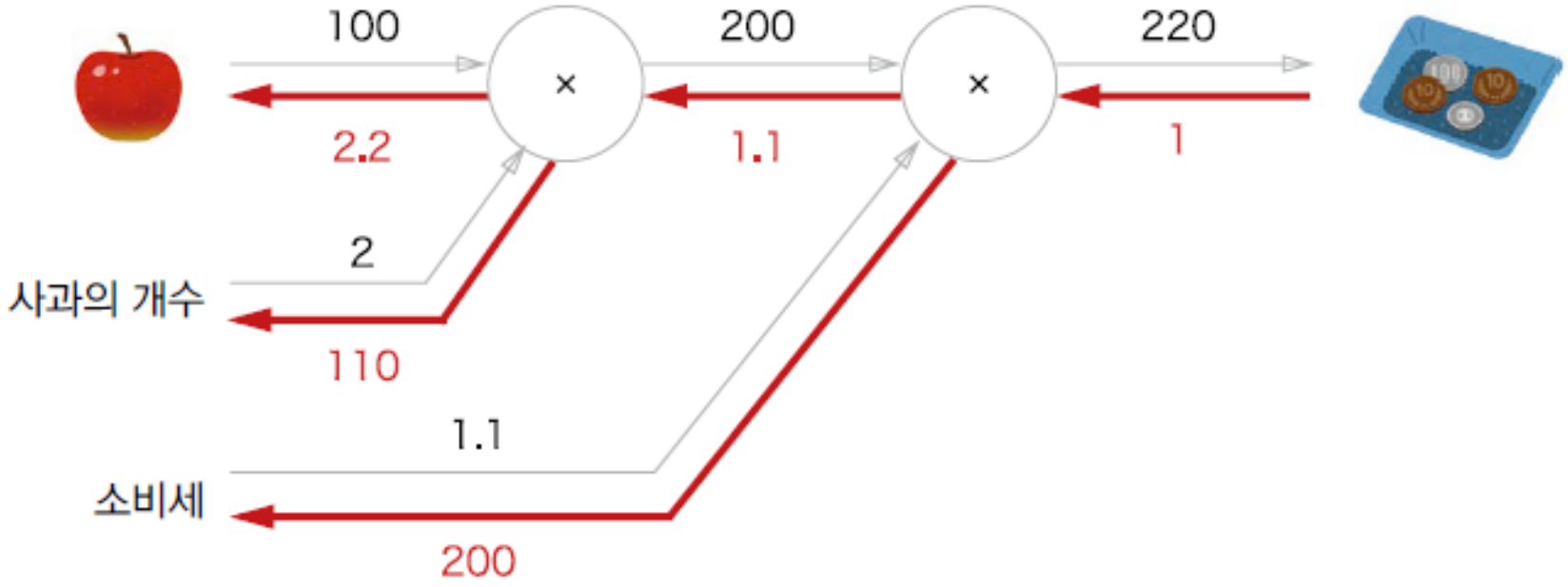
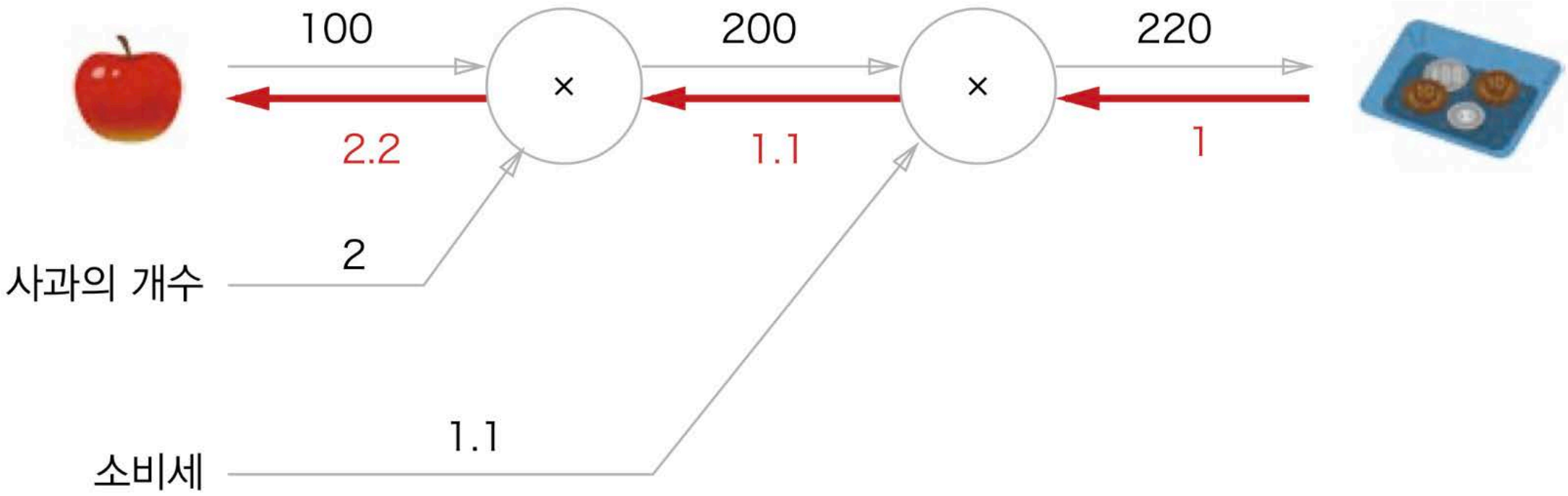


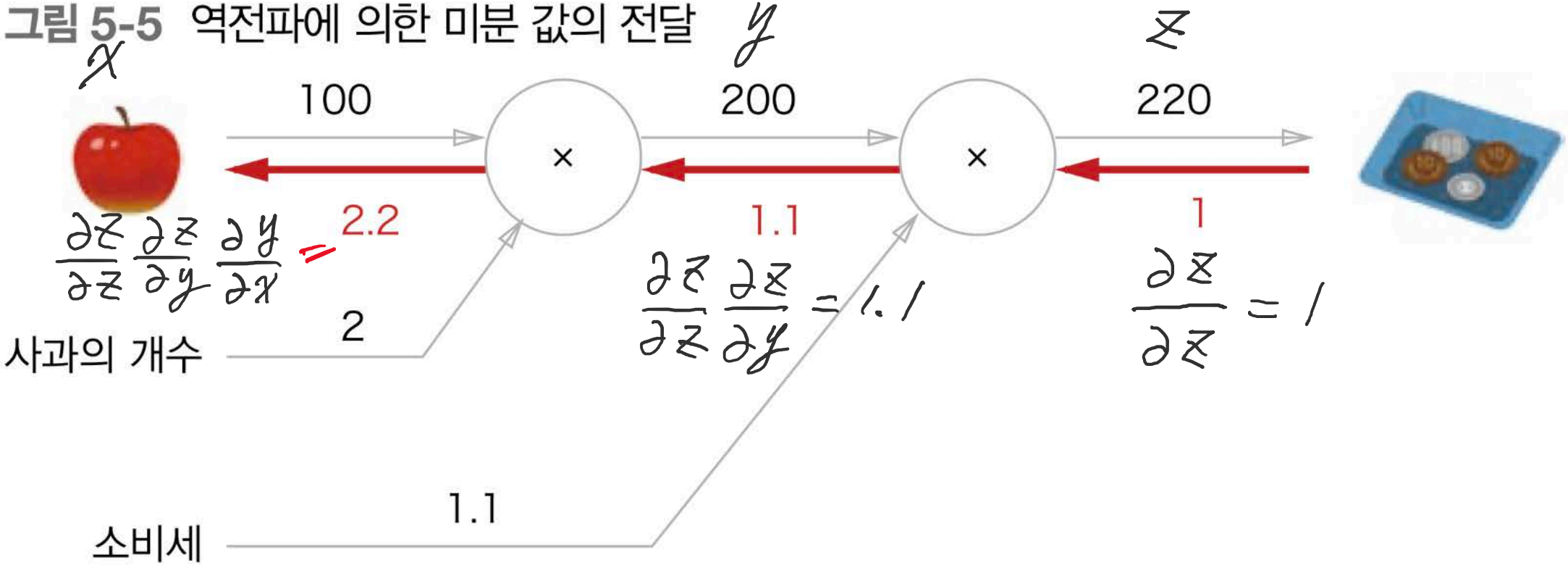
그림 5-5 역전파에 의한 미분 값의 전달



$y = 2x$

$z = 1.1y$

그림 5-5 역전파에 의한 미분 값의 전달

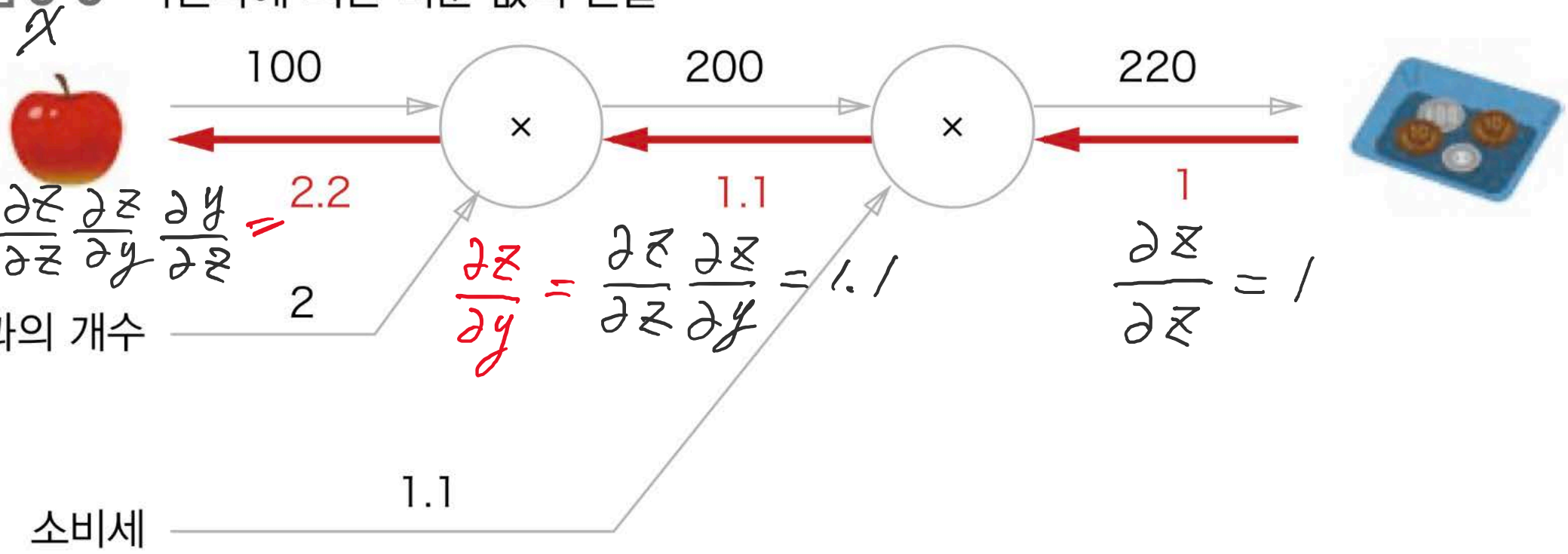


$y = 2x$

$z = 1.1y \Rightarrow \frac{\partial z}{\partial y} = 1.1$

$= 1.1 \cdot 2x \Rightarrow \frac{\partial z}{\partial x} = 2.2$

그림 5-5 역전파에 의한 미분 값의 전달

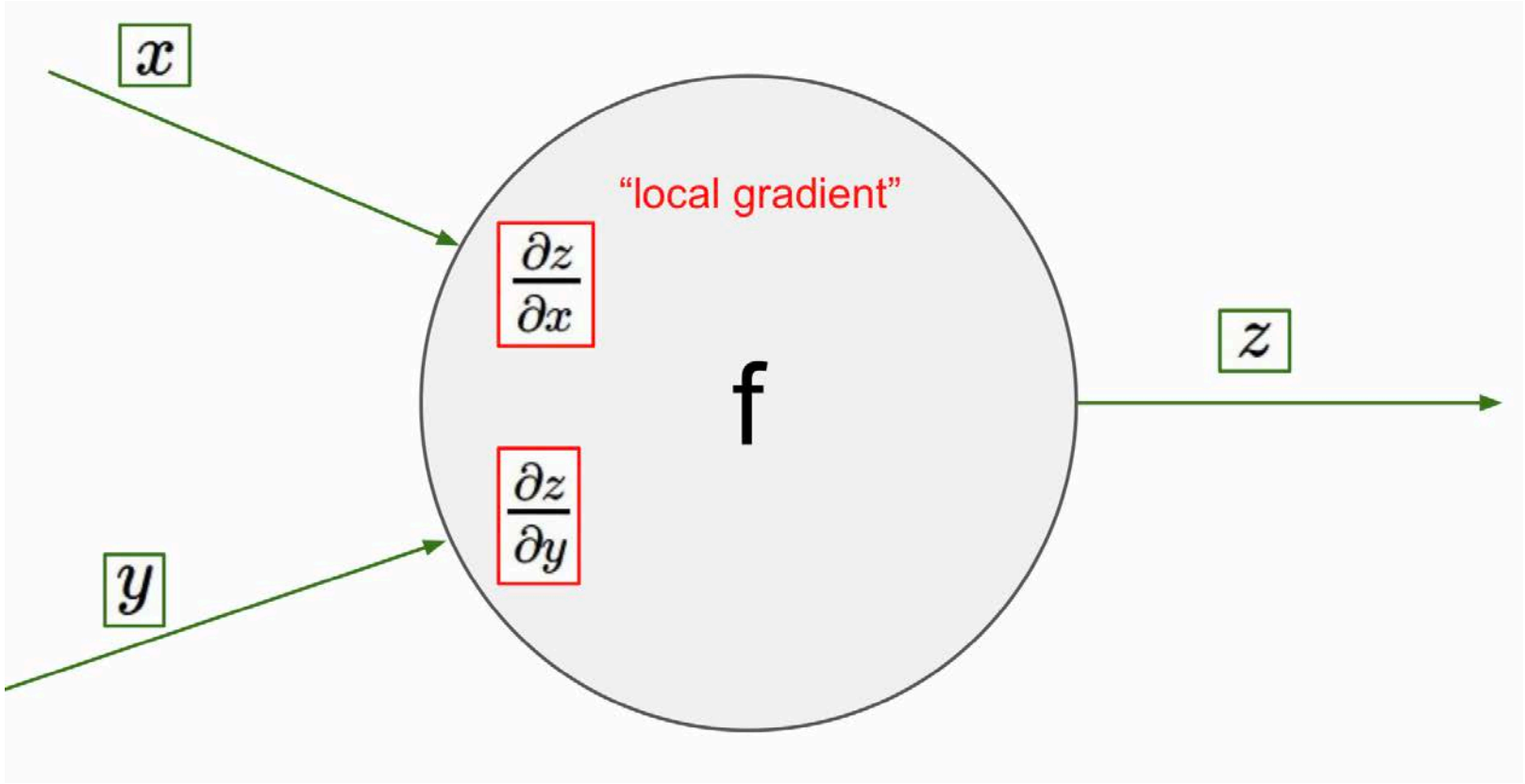


$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial z}\frac{\partial z}{\partial y}\frac{\partial y}{\partial x} = 2.2$

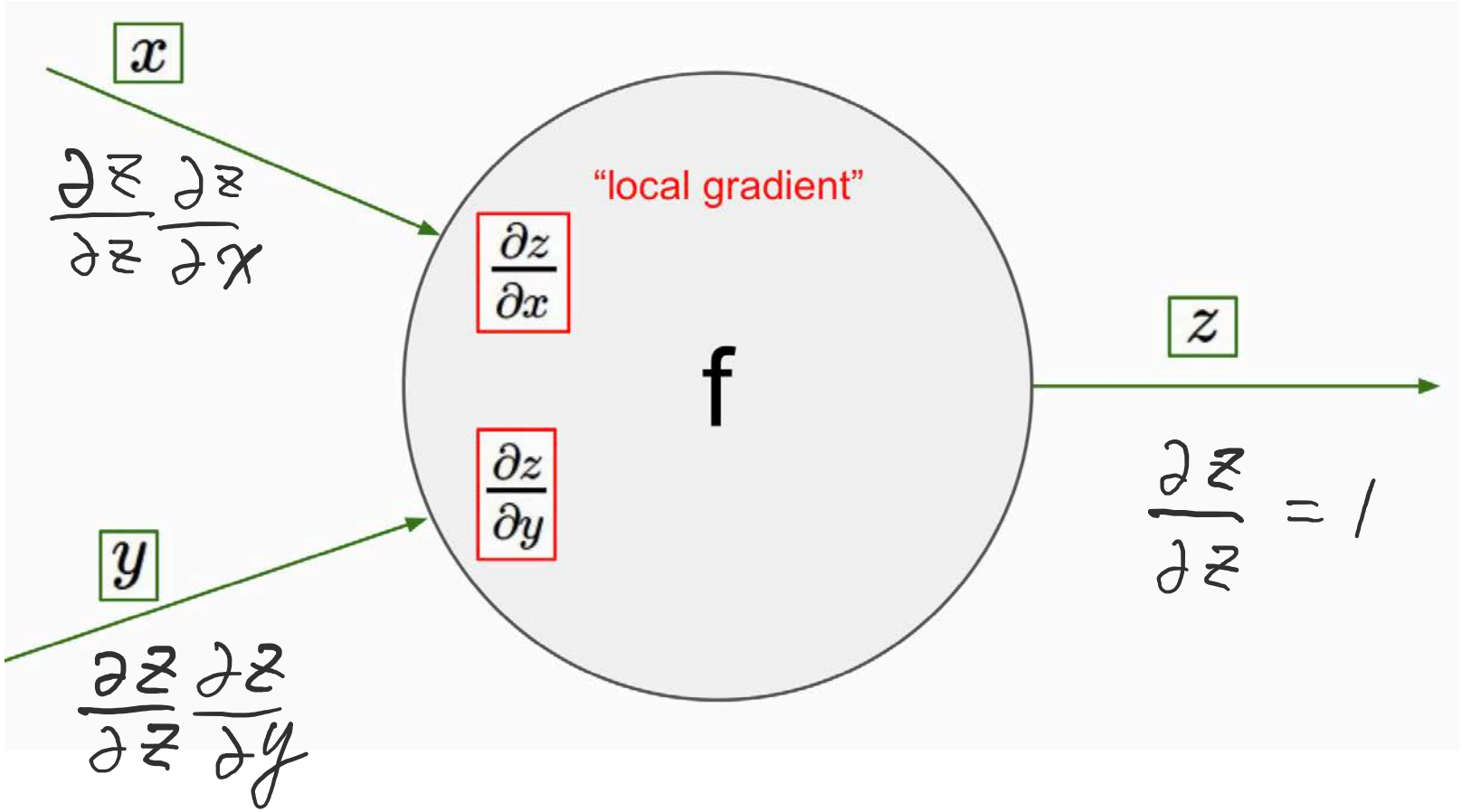
$\frac{\partial z}{\partial y} = \frac{\partial z}{\partial z}\frac{\partial z}{\partial y} = 1.1$

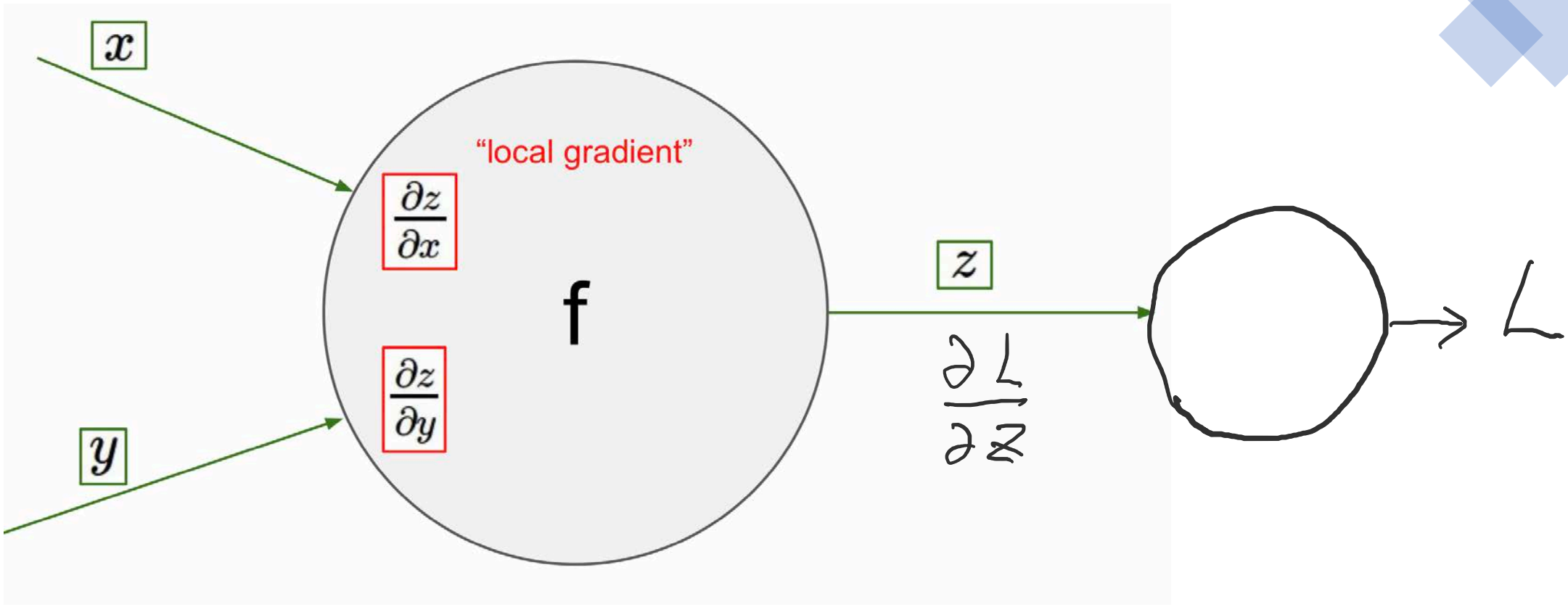
$\frac{\partial z}{\partial z} = 1$

$$z = f(x, y)$$


x
y
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
f
z

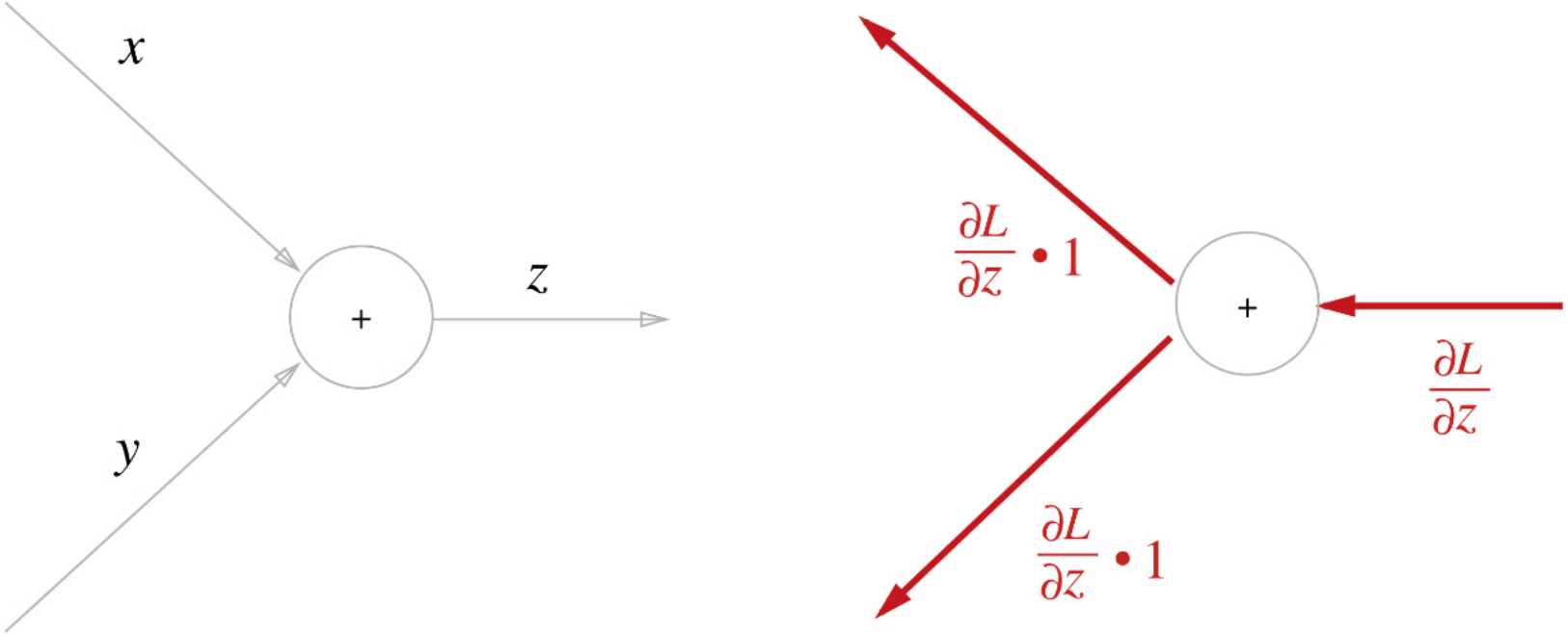
$$z = f(x, y)$$


x
y
z
f
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
$\frac{\partial z}{\partial z} \frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial z} \frac{\partial z}{\partial y}$
$\frac{\partial z}{\partial z} = 1$

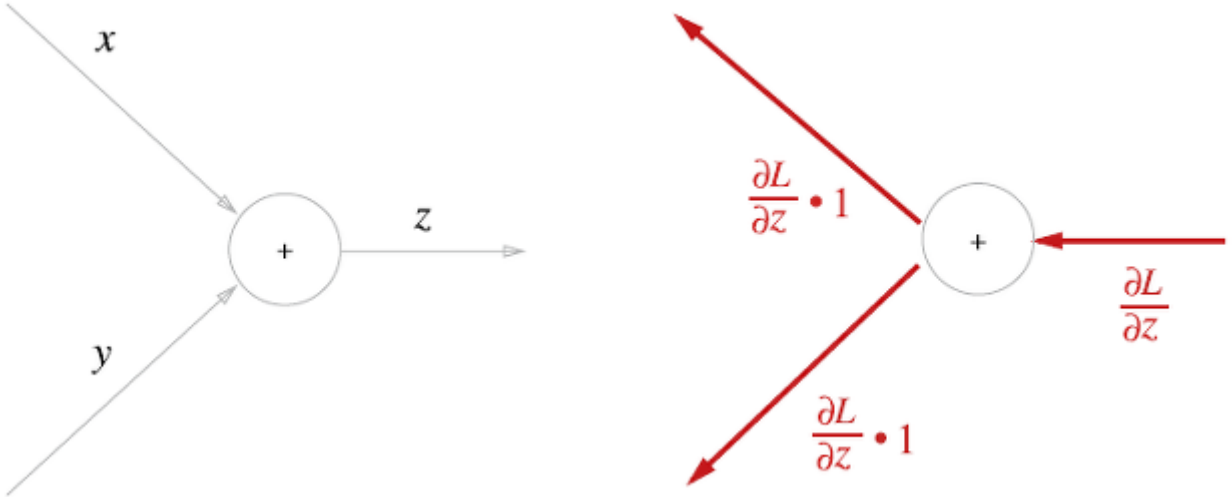

x
y
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
f
z
$\frac{\partial L}{\partial z}$
L

Addition Node

그림 5-9 덧셈 노드의 역전파 : 왼쪽이 순전파, 오른쪽이 역전파다. 덧셈 노드의 역전파는 입력 값을 그대로 흘려보낸다.



Back Propagation of Addition Node

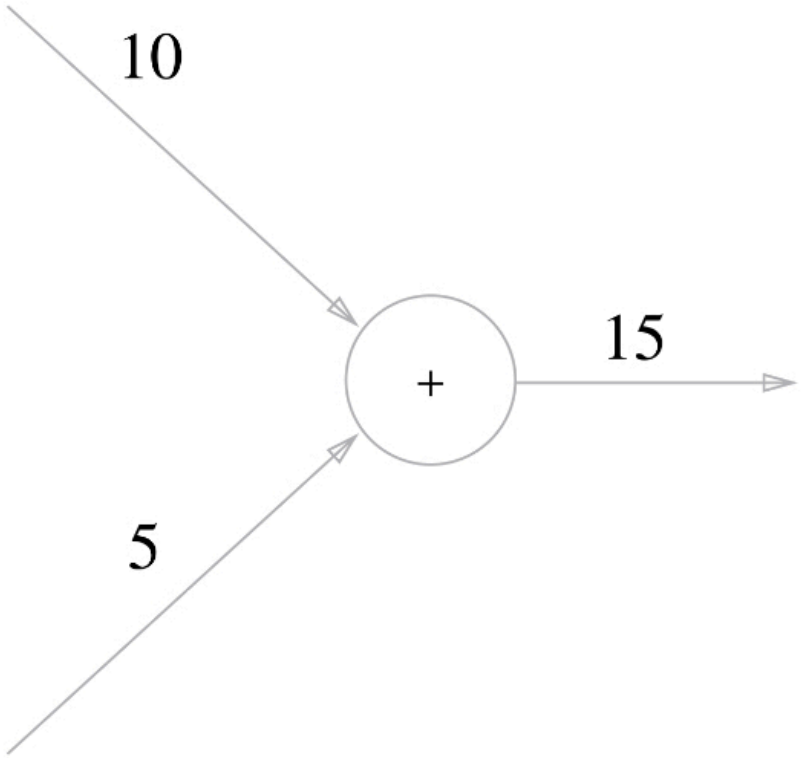


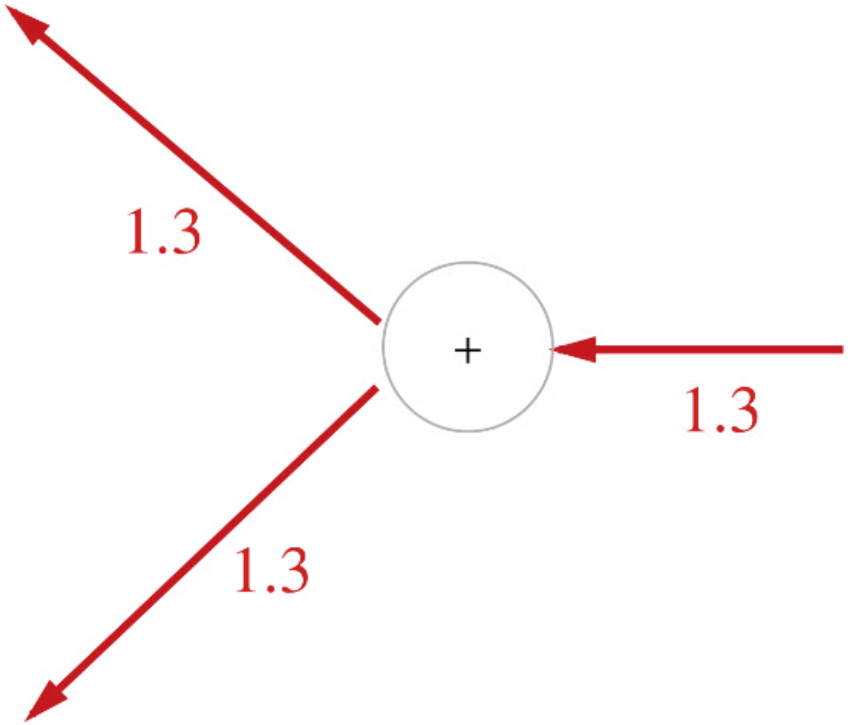
$$L(z) = L(x + y)$$

$$\frac{\partial}{\partial x} L(x + y) = L'(x + y) \times 1$$

$$\frac{\partial}{\partial y} L(x + y) = L'(x + y) \times 1$$

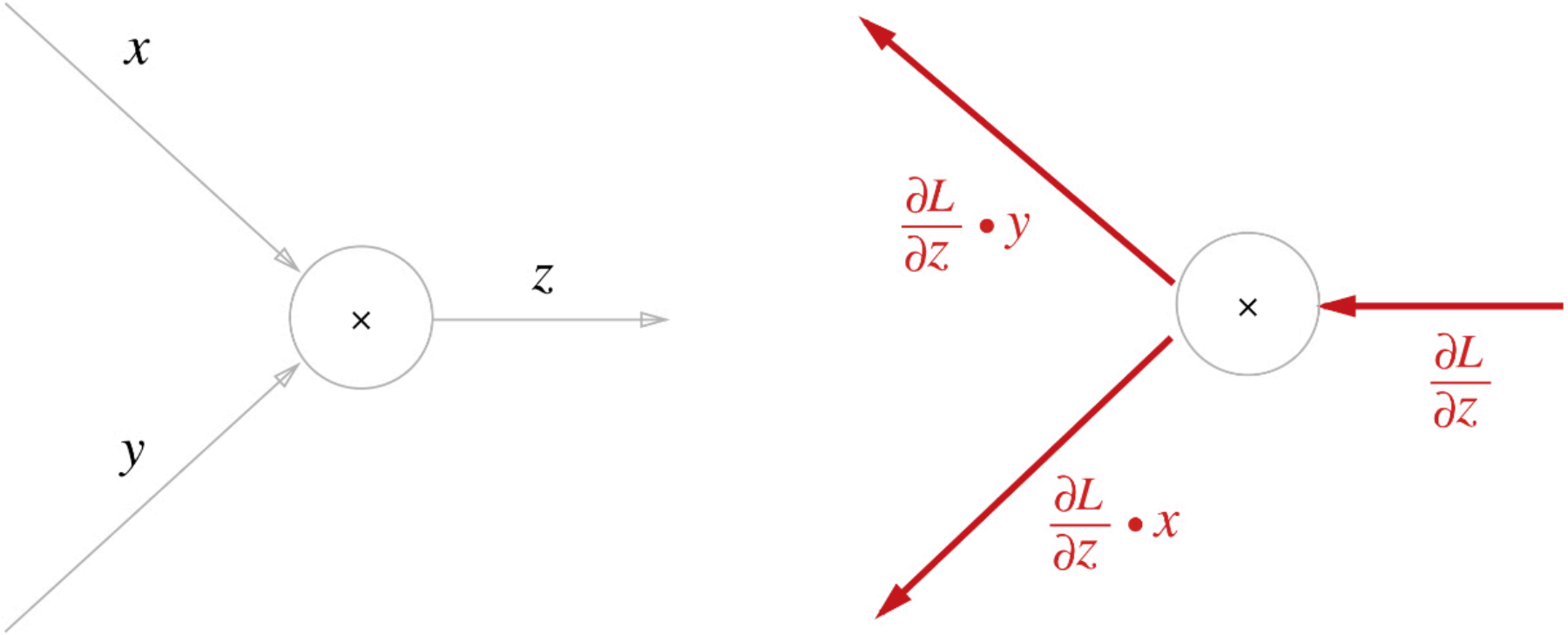
Addition Node



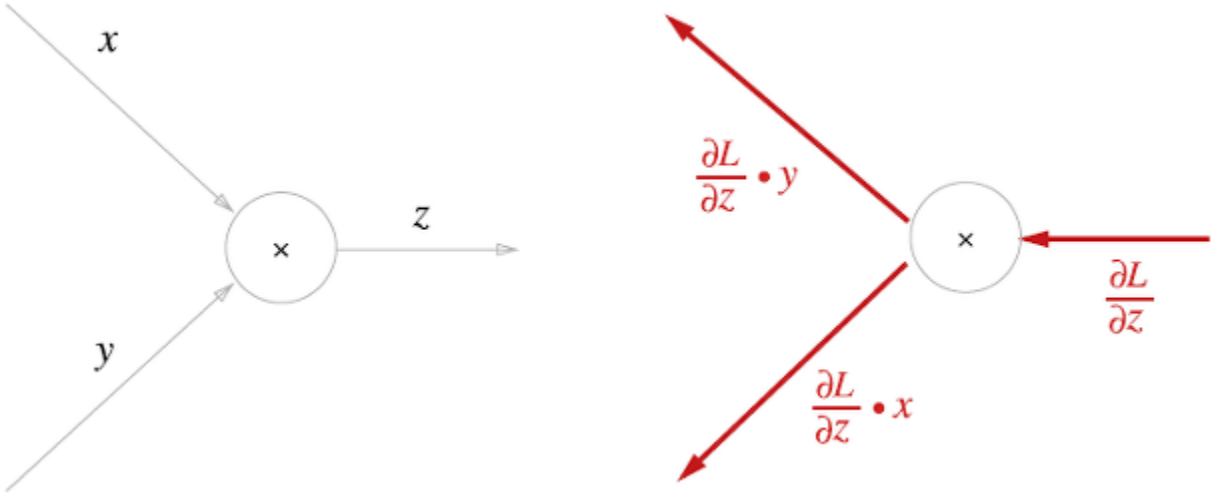


Multiplication Node

그림 5-12 곱셈 노드의 역전파 : 왼쪽이 순전파, 오른쪽이 역전파다.



Back Propagation of Multiplication Node

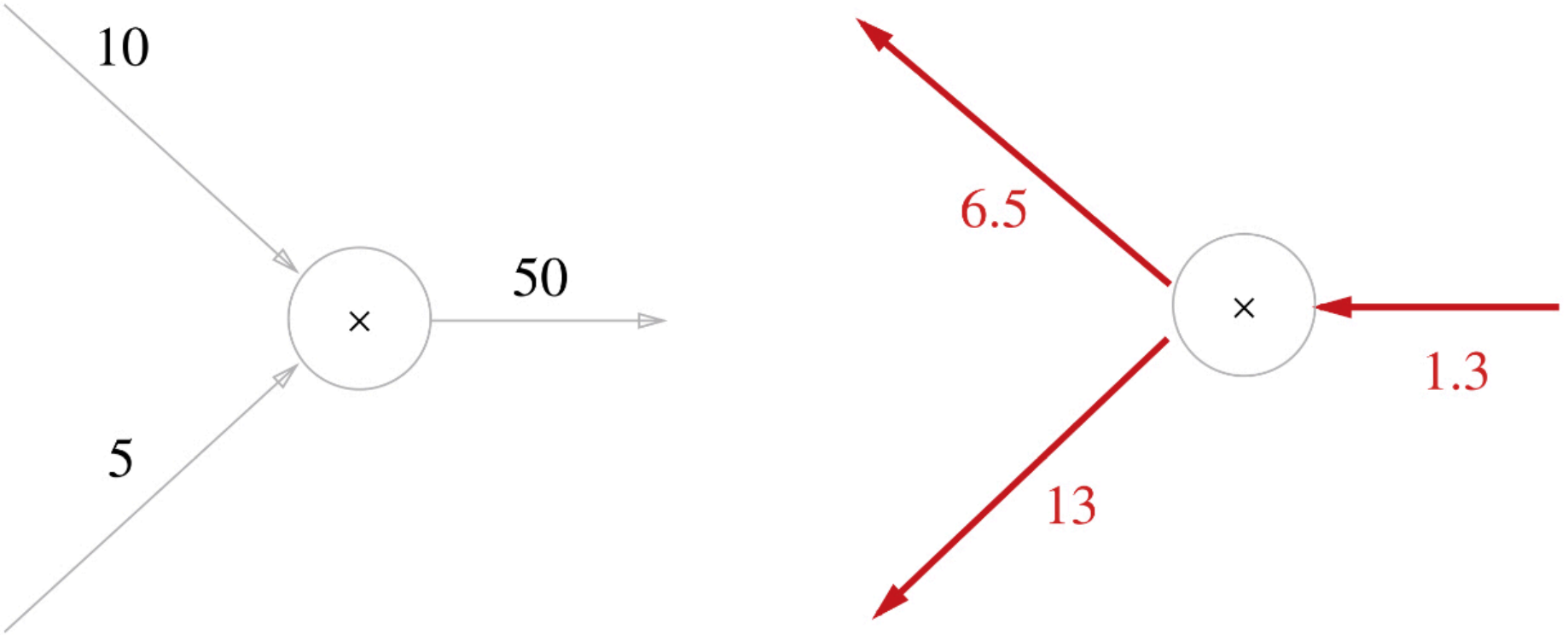


$$L(z) = L(xy)$$

$$\frac{\partial}{\partial x} L(xy) = L'(xy) \times y$$

$$\frac{\partial}{\partial y} L(xy) = L'(xy) \times x$$

Multiplication Node



BP and Chain Rule

Apple price x # of apples , m

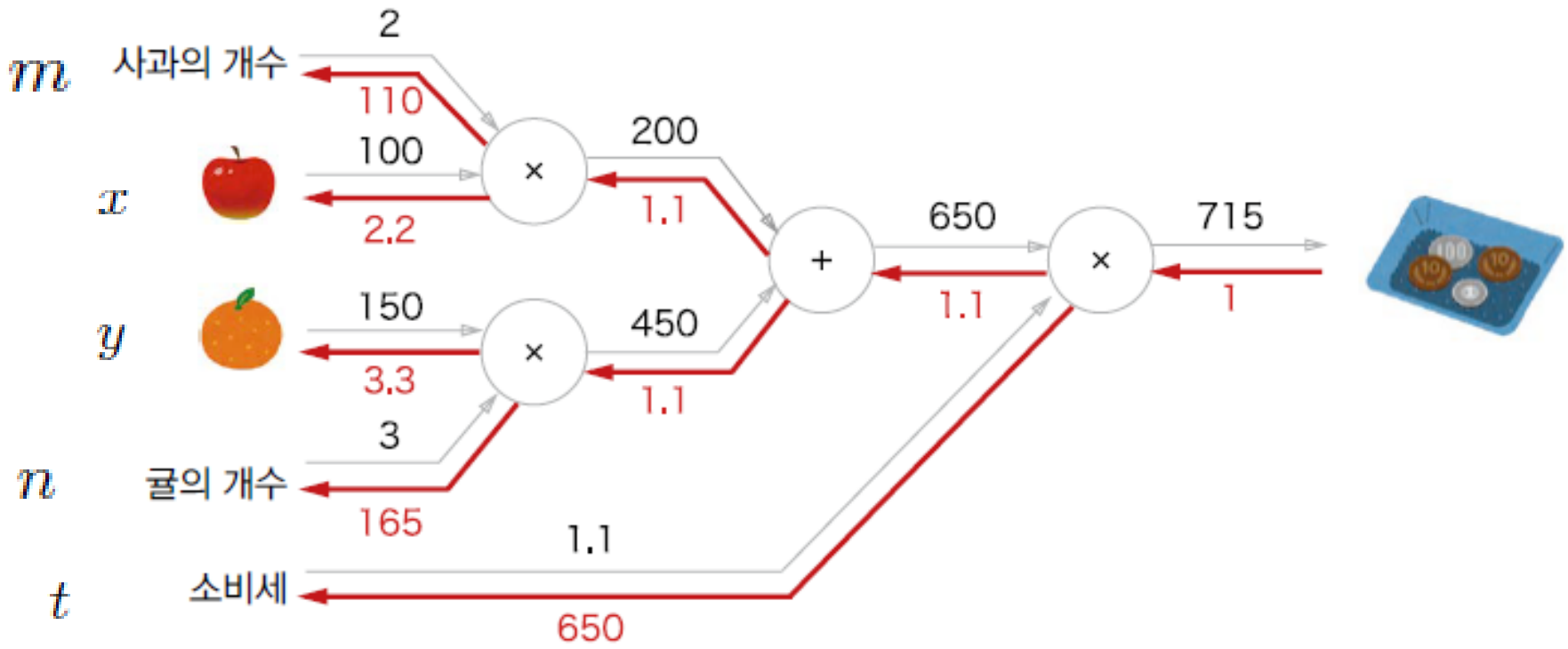
Orange price : , y of oranges , t n t

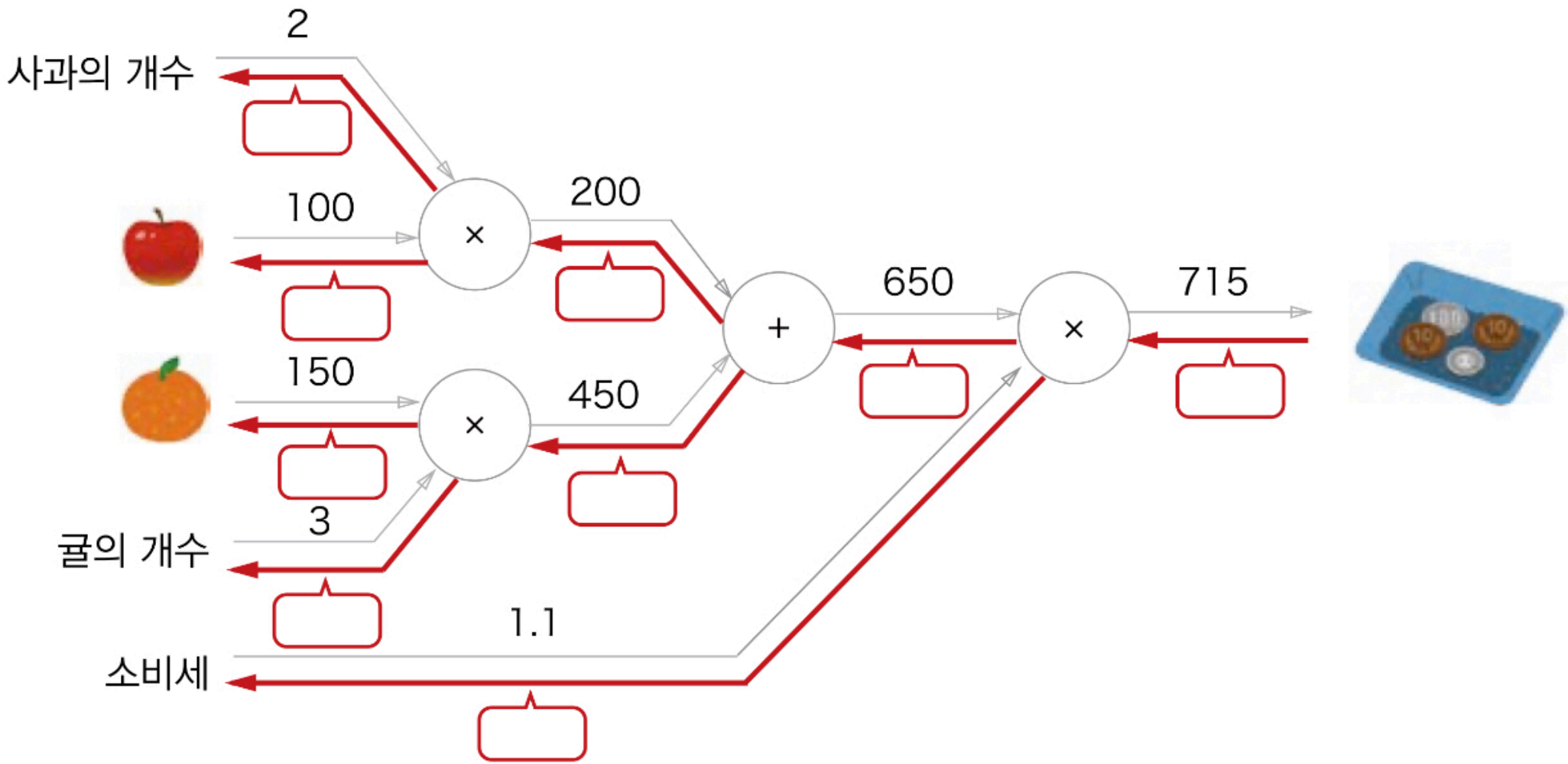
Total price : $f(x, m, y, n, t) = (xm + yn)t$

z= $f_1(x) = mx$, w=$f_2(z) = z + yn$, L= $f_3(w) = wt$

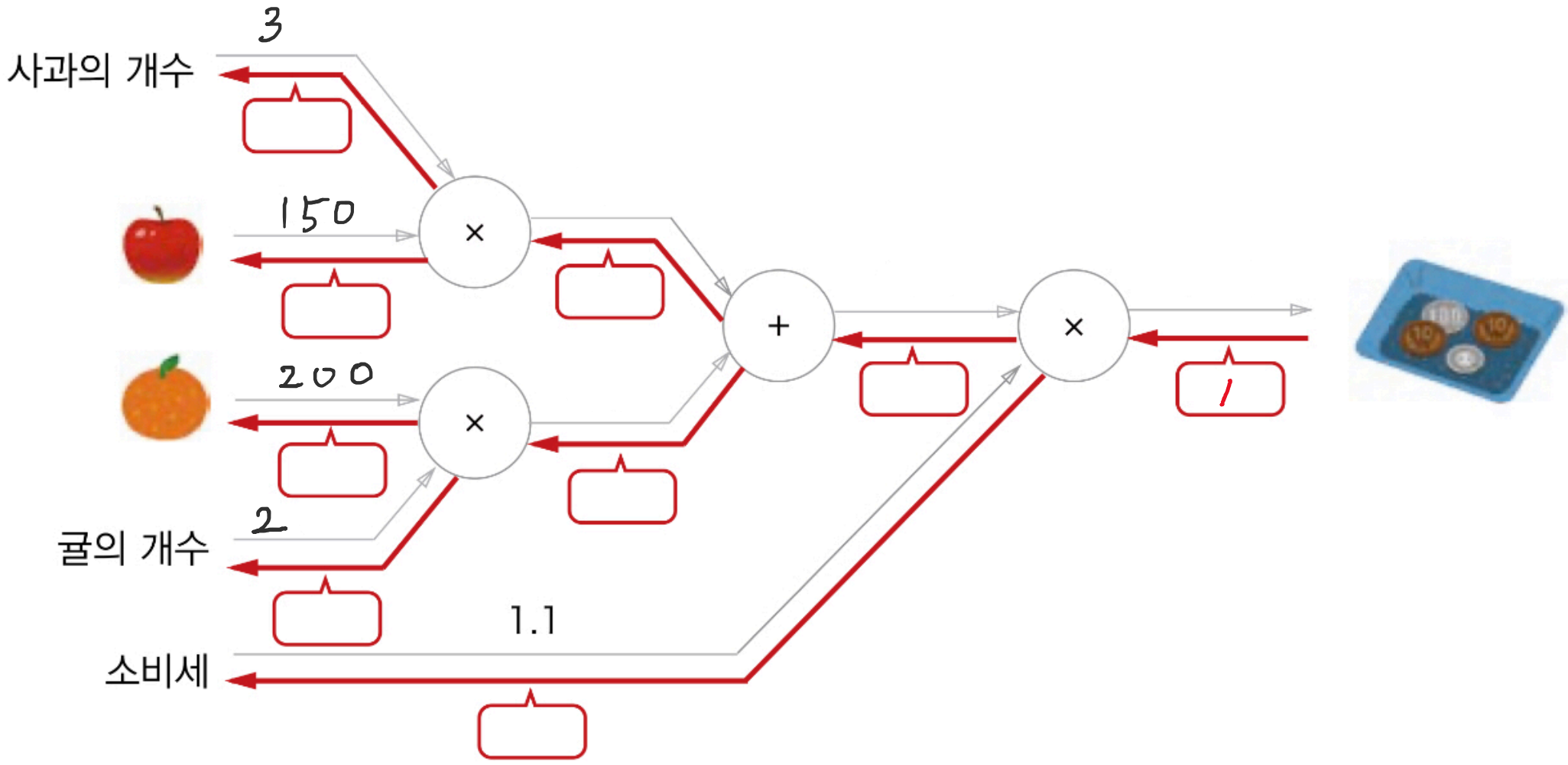
$$f_3 \circ f_2 \circ f_1(x) = (xm + yn)t$$

$$\frac{\partial L}{\partial x} = (f_3 \circ f_2 \circ f_1)'(x) = f_3'(f_2(f_1(x))) \times f_2'(f_1(x)) \times f_1'(x) = t \times 1 \times m$$



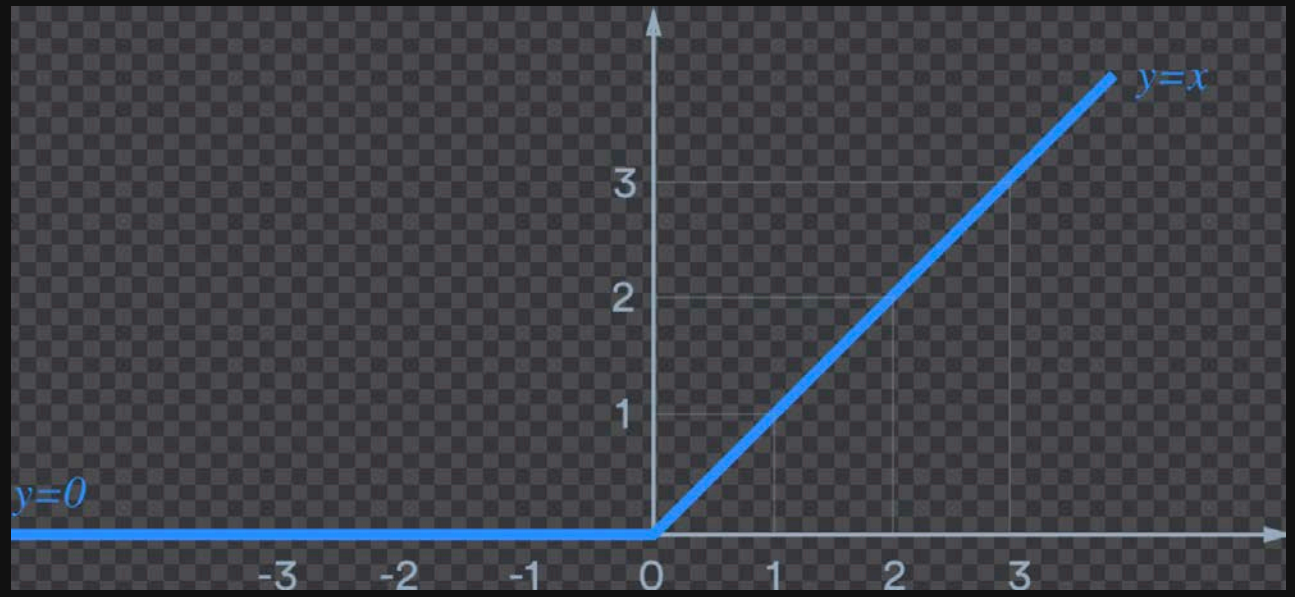

2
사과의 개수
100
200
×
150
×
450
+
650
×
715
3
귤의 개수
1.1
소비세

Quiz : Computation Graph: Back Propagation



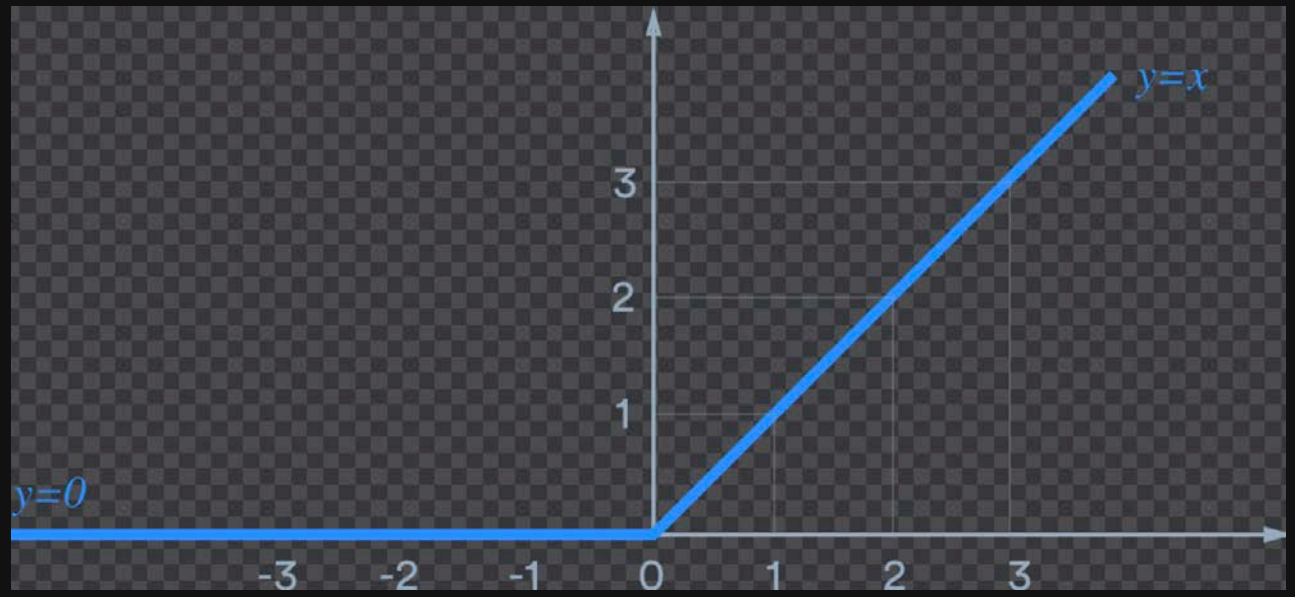
5.5 Implementing Activation Function Layer
(활성화 함수 계층 구현)

- ReLU (Rectifier linear unit)



Implementing Activation Function Layer

- **ReLU**

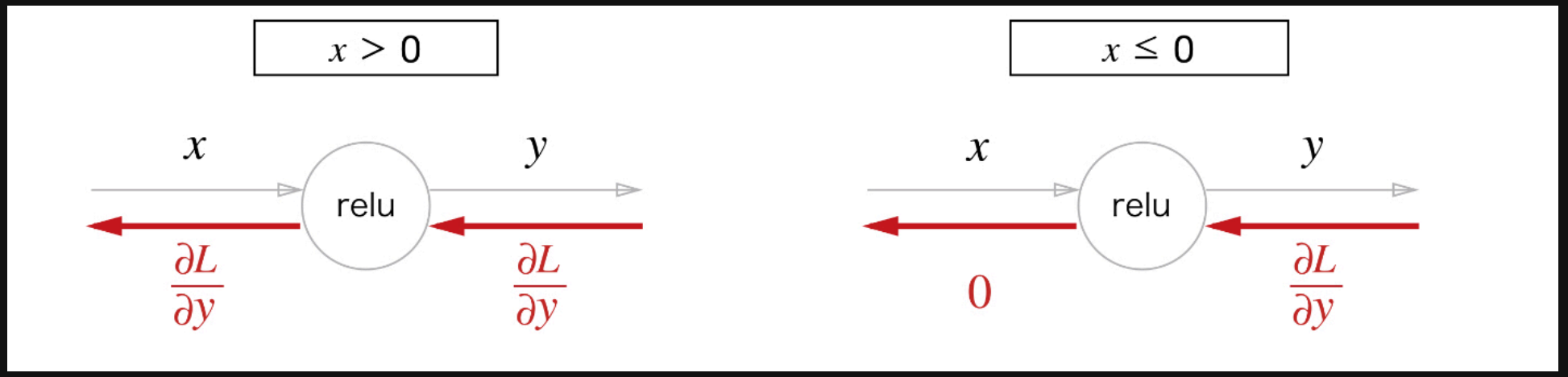


$$y = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$\frac{\partial y}{\partial x} = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Implementing Activation Function Layer

- **ReLU**



$\frac{\partial y}{\partial x} = 1$

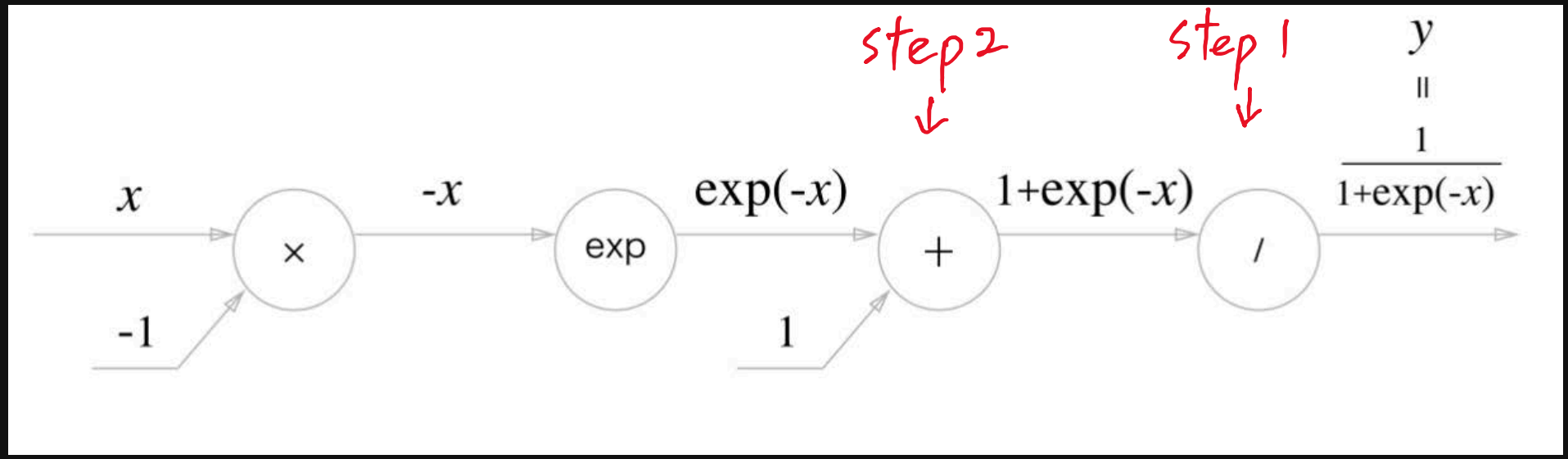
$\frac{\partial y}{\partial x} = 0$

5.5.2 Sigmoid Function Layer

$$y = \frac{1}{1 + \exp(-x)}$$

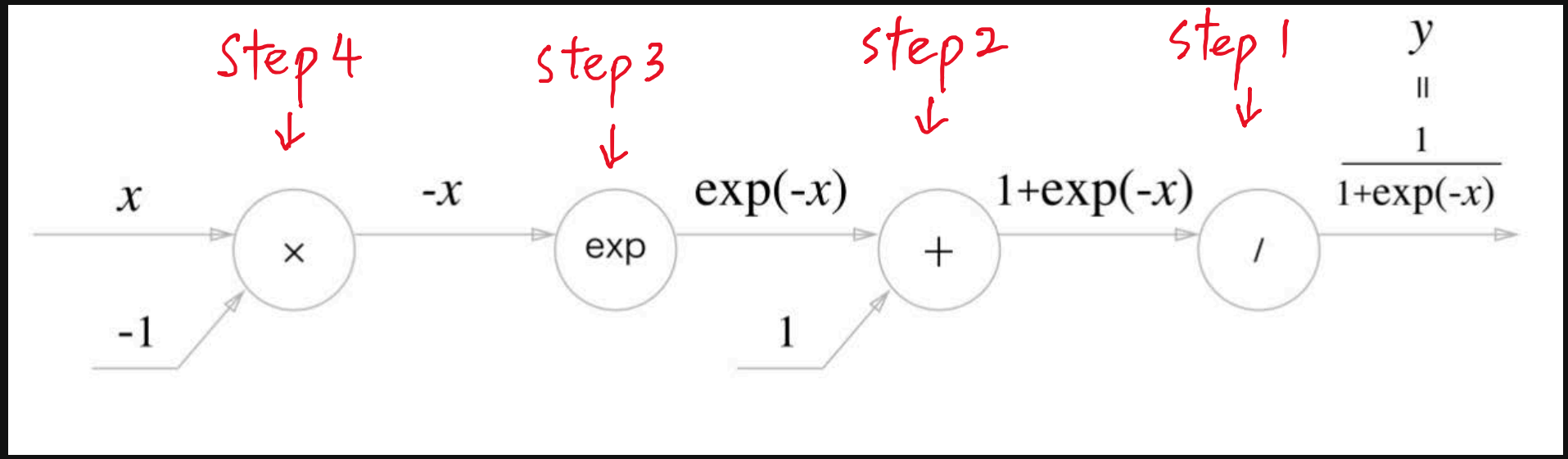
5.5.2 Sigmoid Layer

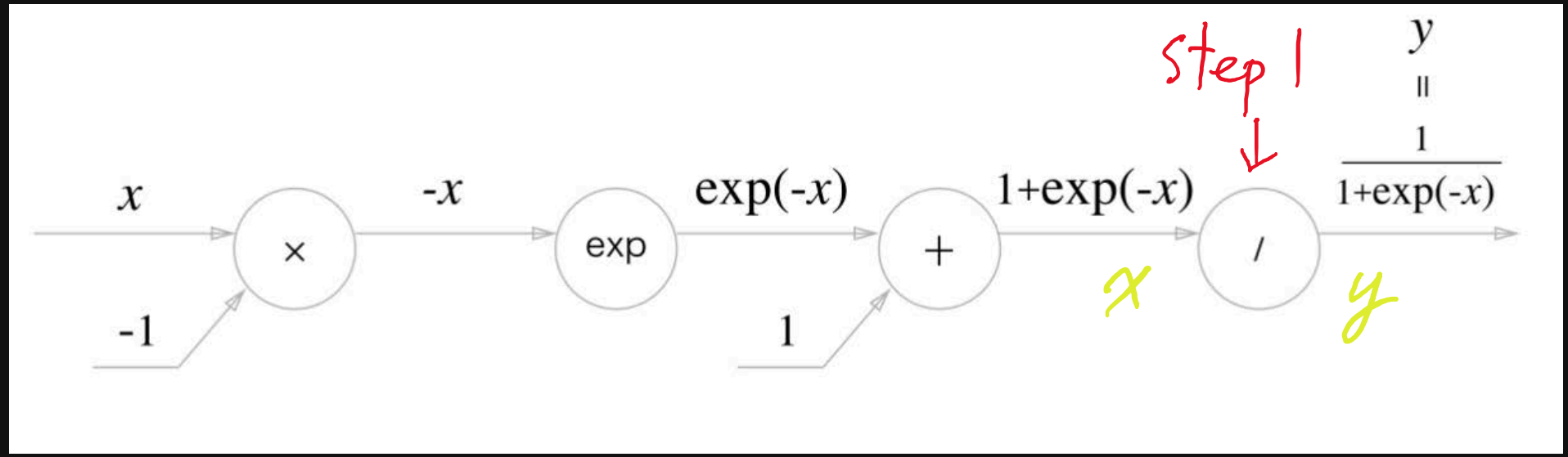
$$y = \frac{1}{1 + \exp(-x)}$$



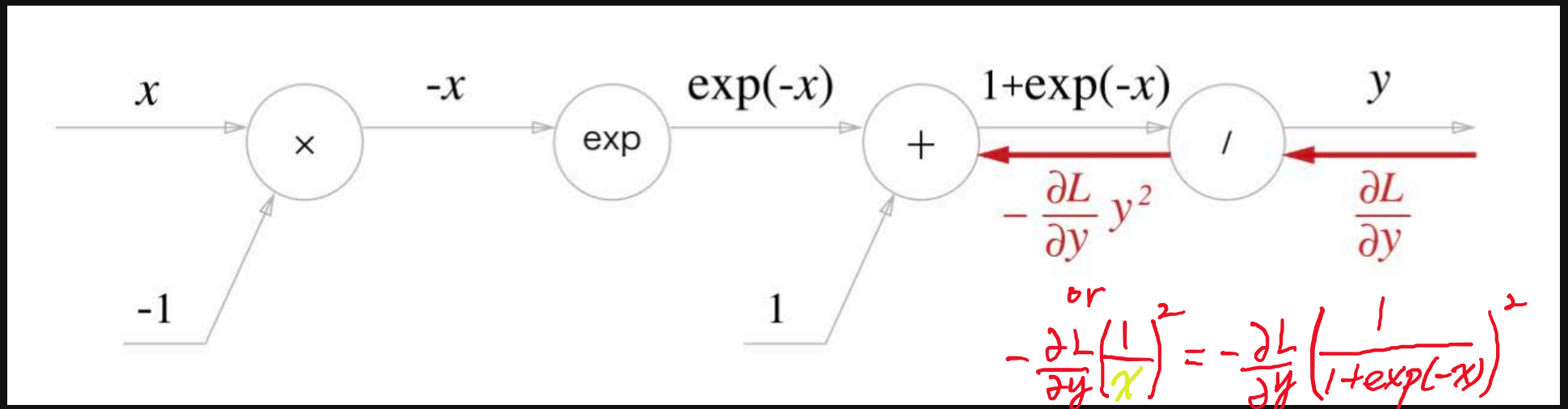
5.5.2 Sigmoid Layer

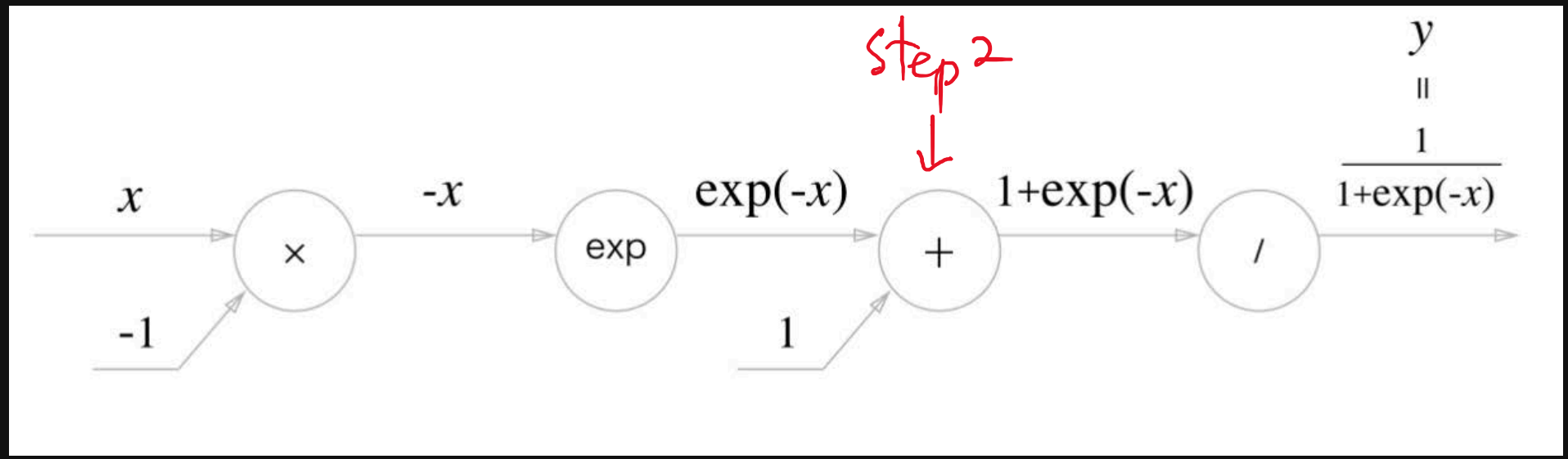
$$y = \frac{1}{1 + \exp(-x)}$$

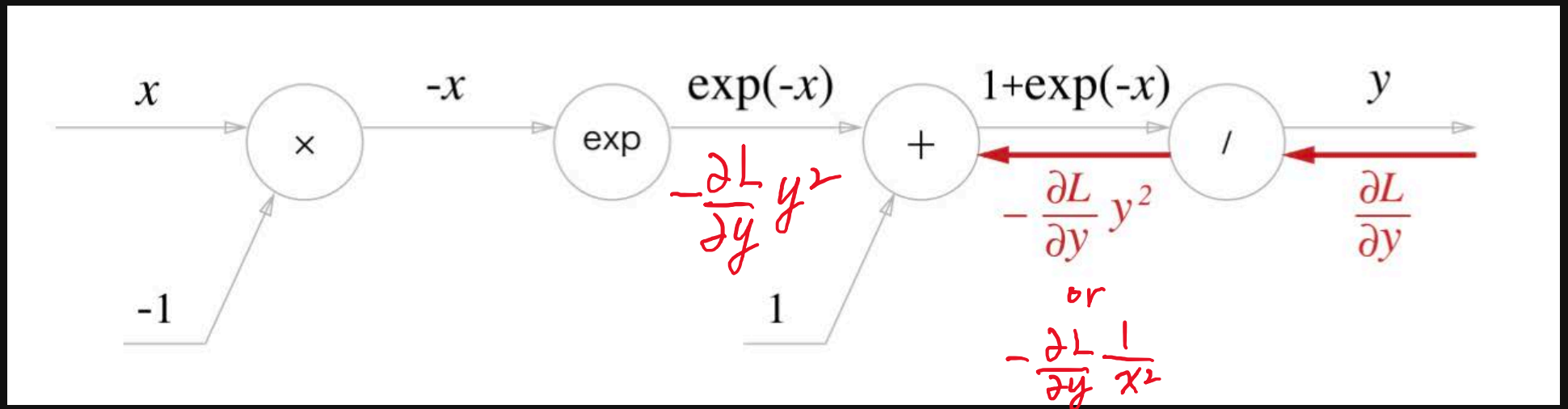


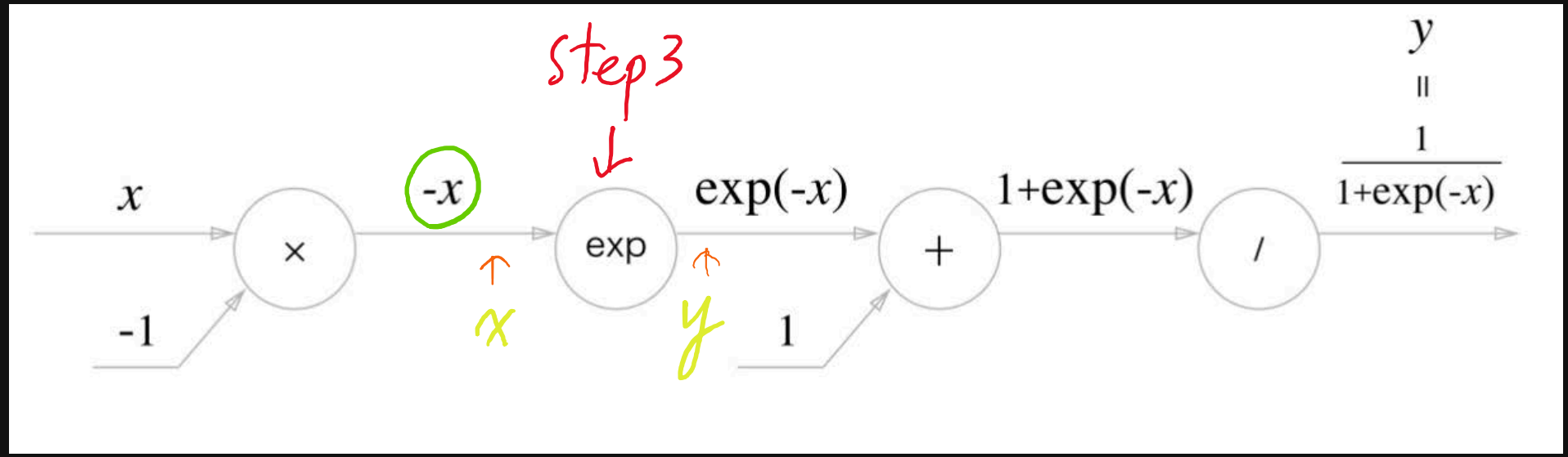


Step 1 : $y = \frac{1}{x}$, $\frac{\partial y}{\partial x} = -\frac{1}{x^2} = -y^2$

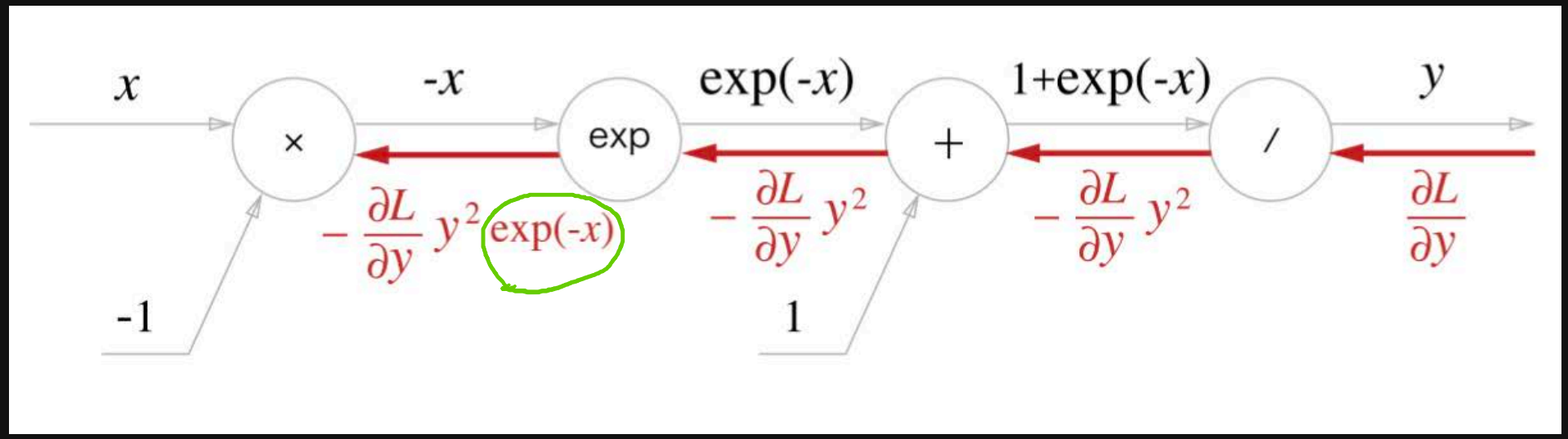


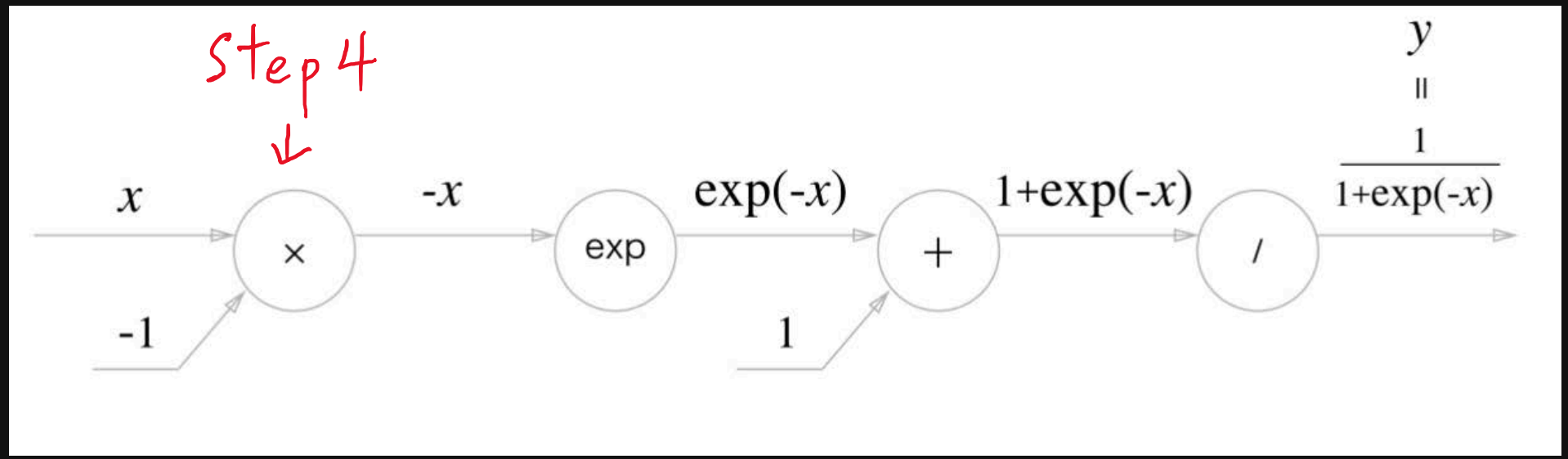

step 2
x
-x
exp(-x)
1+exp(-x)
y
=
1
1+exp(-x)
×
exp
+
/
-1
1

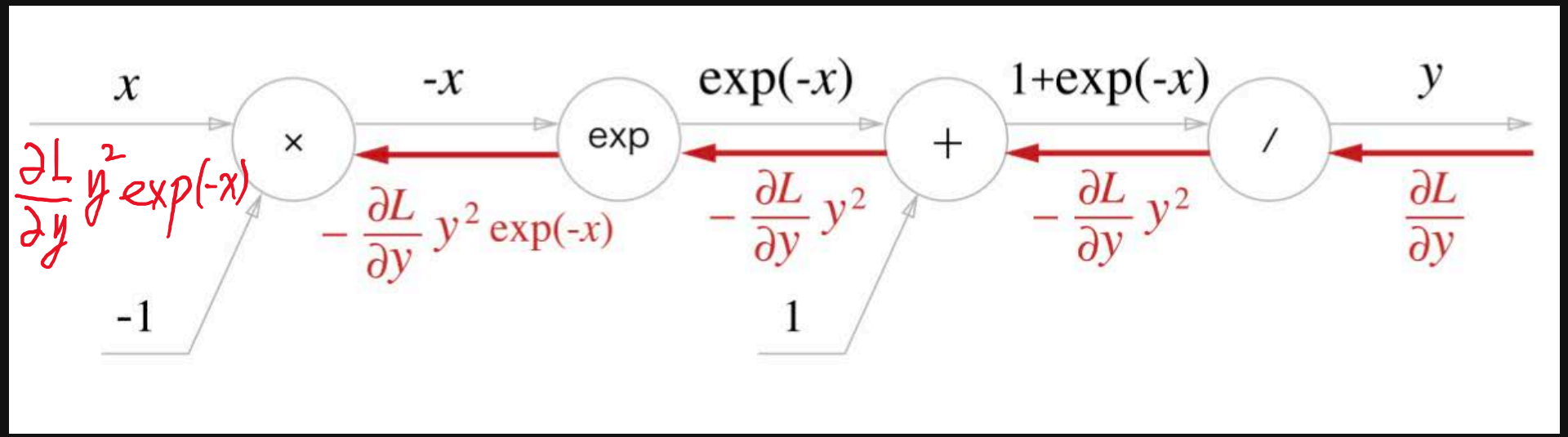

x
-x
exp(-x)
1+exp(-x)
y
×
exp
+
/
-1
1
$-\frac{\partial L}{\partial y}y^2$
$-\frac{\partial L}{\partial y}y^2$
$\frac{\partial L}{\partial y}$
or
$-\frac{\partial L}{\partial y}\frac{1}{x^2}$

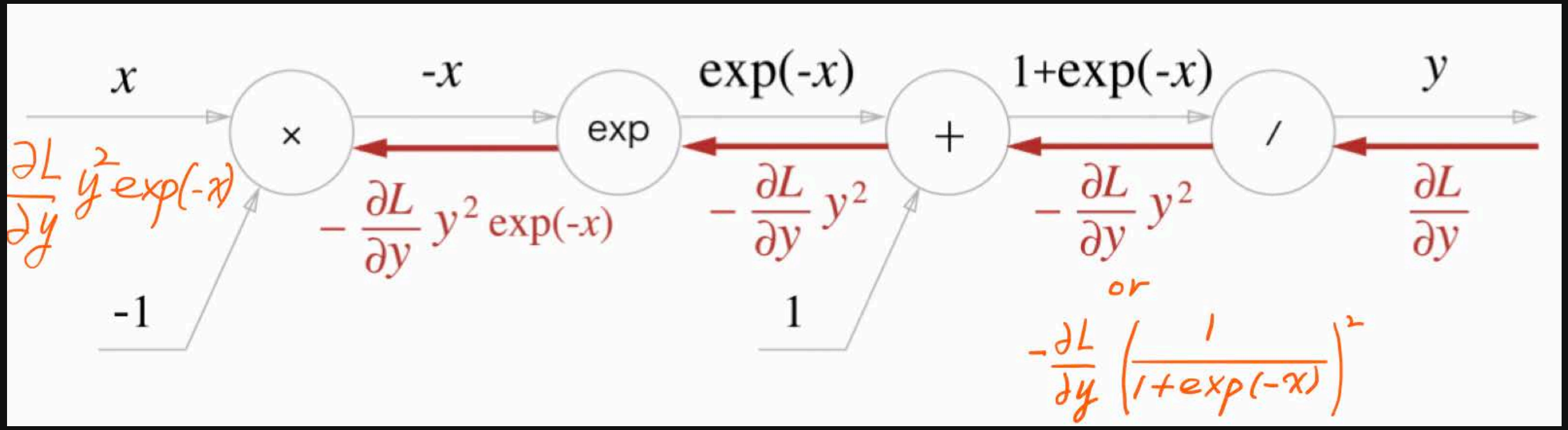


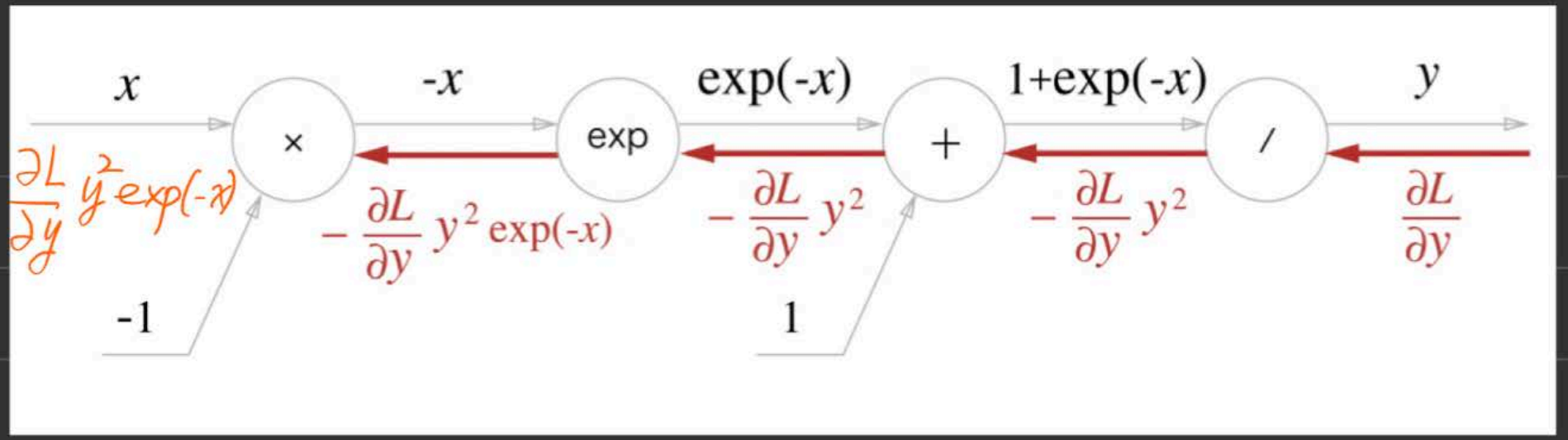
$y = \exp(x)$, $\frac{\partial y}{\partial x} = \exp(x)$, $-x$ 가 입력 되므로 $\frac{\partial y}{\partial x} = \exp(-x)$




Step 4
x
-1
×
-x
exp
exp(-x)
1
+
1+exp(-x)
/
y
=
1
1+exp(-x)


x
-1
×
-x
exp
exp(-x)
1
+
1+exp(-x)
/
y
$\frac{\partial L}{\partial y} y^2 \exp(-x)$
$-\frac{\partial L}{\partial y} y^2 \exp(-x)$
$-\frac{\partial L}{\partial y} y^2$
$-\frac{\partial L}{\partial y} y^2$
$\frac{\partial L}{\partial y}$


x
-x
exp(-x)
1+exp(-x)
y
×
exp
+
/
-1
1
$\frac{\partial L}{\partial y} y^2 \exp(-x)$
$-\frac{\partial L}{\partial y} y^2 \exp(-x)$
$-\frac{\partial L}{\partial y} y^2$
$-\frac{\partial L}{\partial y} y^2$
$\frac{\partial L}{\partial y}$
or
$-\frac{\partial L}{\partial y}\left(\frac{1}{1+\exp(-x)}\right)^2$


x
-x
exp(-x)
1+exp(-x)
y
×
exp
+
/
-1
1
$\frac{\partial L}{\partial y} y^2 \exp(-x)$
$-\frac{\partial L}{\partial y} y^2 \exp(-x)$
$-\frac{\partial L}{\partial y} y^2$
$-\frac{\partial L}{\partial y} y^2$
$\frac{\partial L}{\partial y}$


$\frac{\partial L}{\partial y} y^2 \exp(-x)$
Sigmoid
$\frac{\partial L}{\partial y}$

$$\frac{\partial L}{\partial y} \left(\frac{1}{1+\exp(-x)} \right)^2 \exp(-x)$$

$$= \frac{\partial L}{\partial y} \frac{1}{1+\exp(-x)} \frac{\exp(-x)}{1+\exp(-x)}$$

$$= \frac{\partial L}{\partial y} y (1-y)$$

$$\frac{\partial L}{\partial y} \left(\frac{1}{1+\exp(-x)} \right)^2 \exp(-x)$$

$$= \frac{\partial L}{\partial y} \frac{1}{1+\exp(-x)} \frac{\exp(-x)}{1+\exp(-x)}$$

$$= \frac{\partial L}{\partial y} y (1-y)$$

Sigmoid back propagation can be expressed by the only forward propagation output, y

5.6 Affine / Softmax Layer

Graph of a Affine Layer

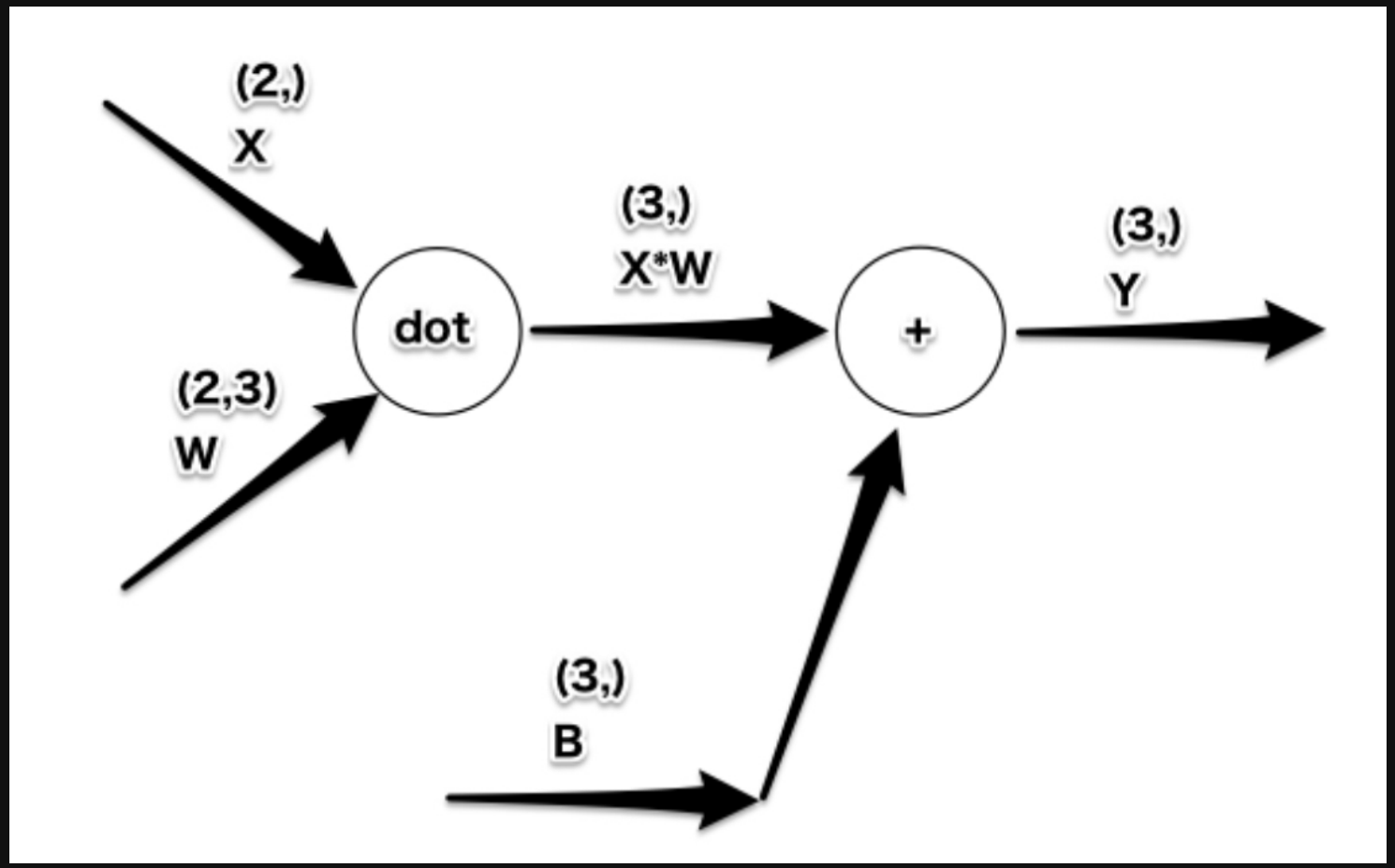
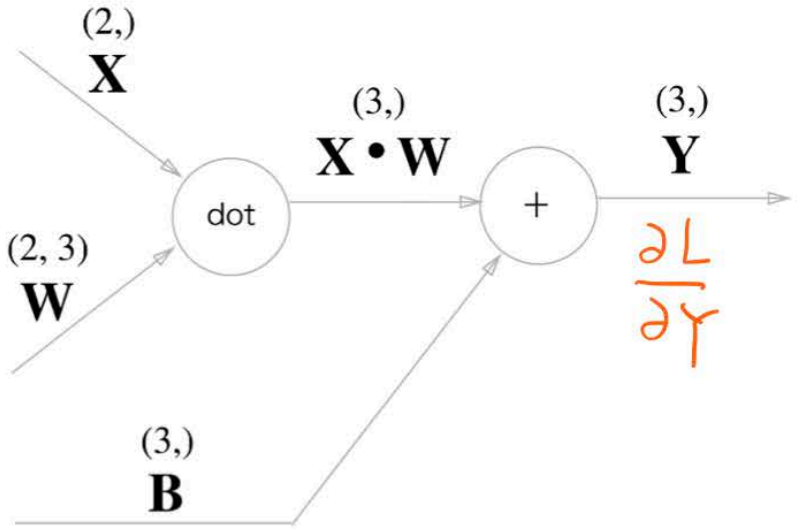


그림 5-24 Affine 계층의 계산 그래프 : 변수가 행렬임에 주의. 각 변수의 형상을 변수명 위에 표기했다.

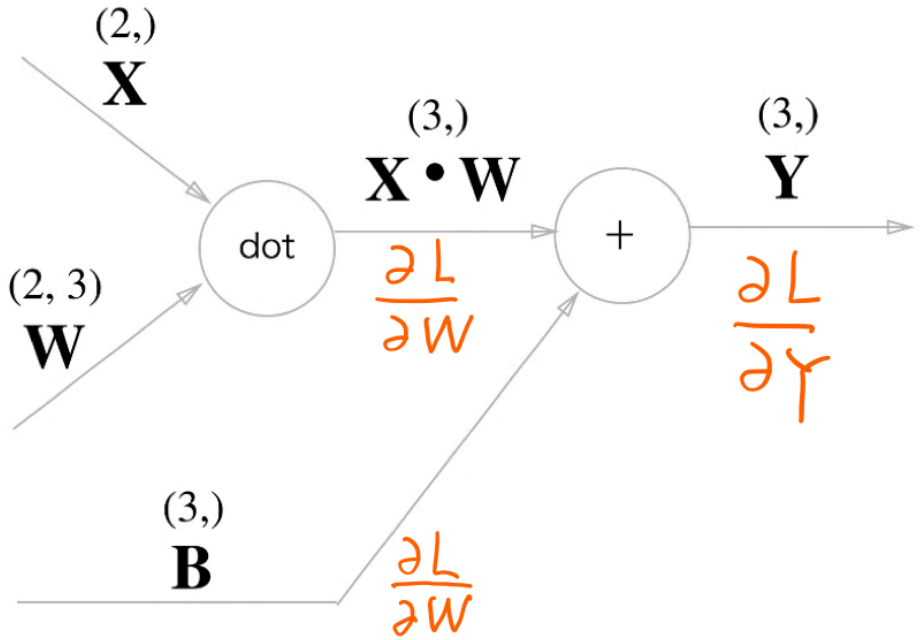


$$\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$$

$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$$

$$X = [x_1, x_2], \quad W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

그림 5-24 Affine 계층의 계산 그래프 : 변수가 행렬임에 주의. 각 변수의 형상을 변수명 위에 표기했다.

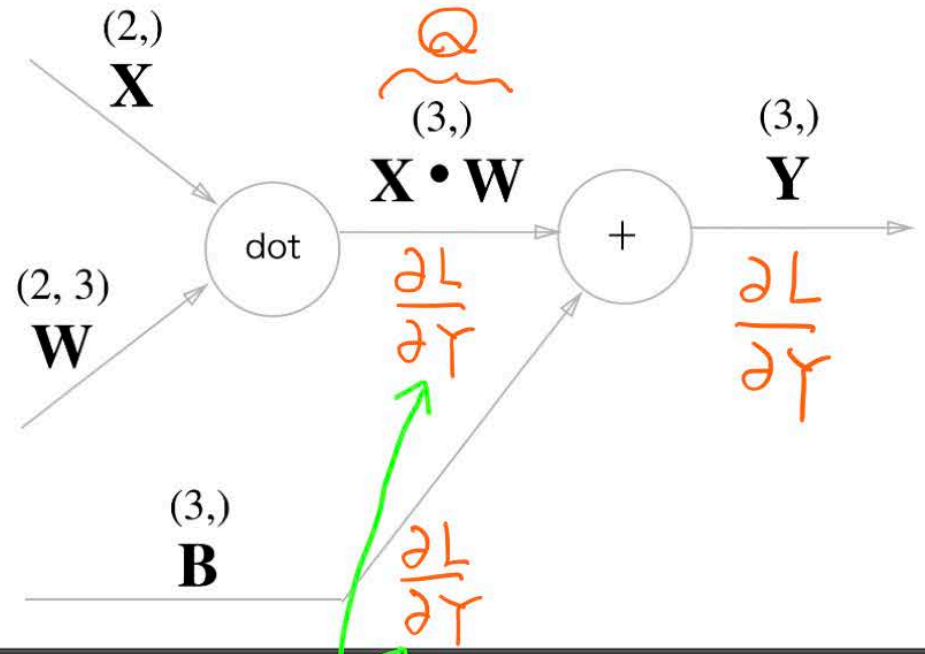


$$\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$$

$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$$

$$X = [x_1, x_2], \quad W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

그림 5-24 Affine 계층의 계산 그래프 : 변수가 행렬임에 주의. 각 변수의 형상을 변수명 위에 표기했다.

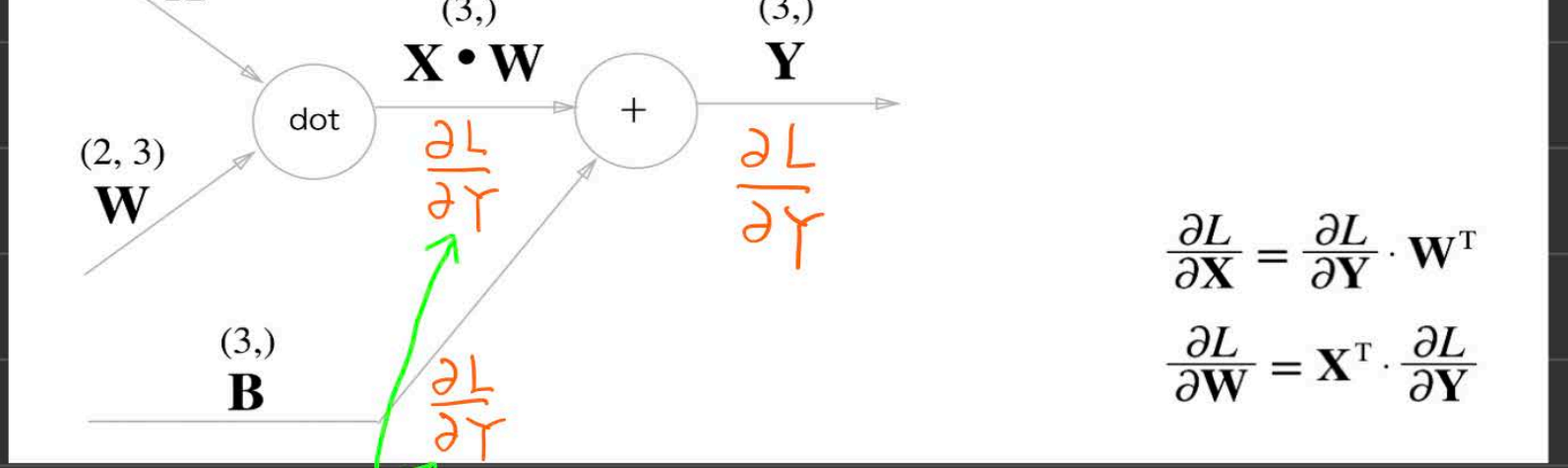


$$\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$$

$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$$

$X = [x_1, x_2]$, $W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$

$\frac{\partial L}{\partial Y} \frac{\partial Y}{\partial Q} = \frac{\partial L}{\partial Y}$



$$\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$$

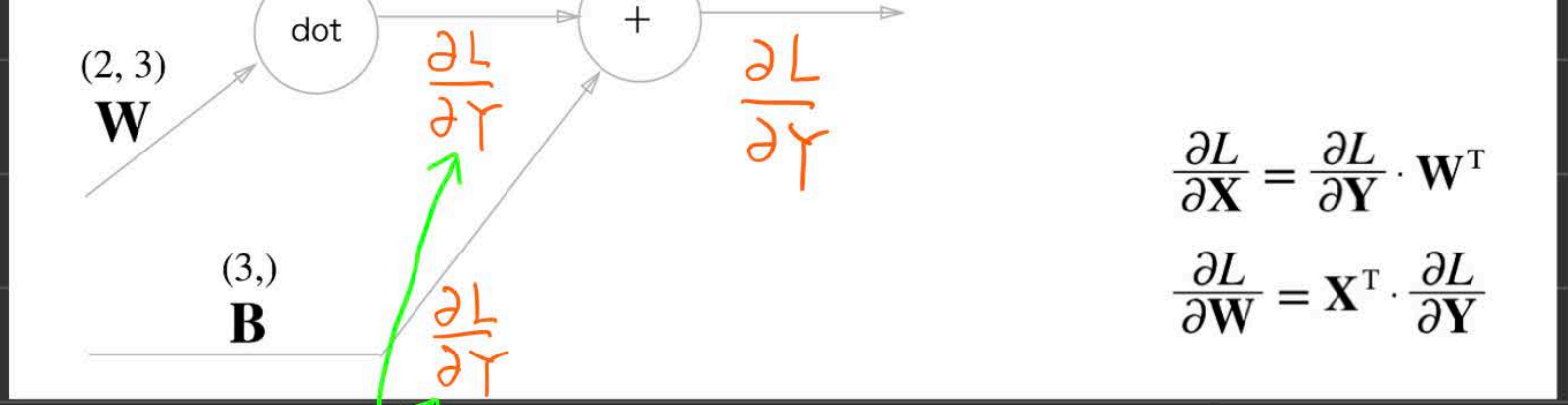
$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$$

$X = [x_1, x_2]$, $W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$

$\frac{\partial L}{\partial Y} \underbrace{\frac{\partial Y}{\partial Q}}_{1} = \underbrace{\frac{\partial L}{\partial Y}}$

$\frac{\partial L}{\partial X}$ 의 size는 $\overset{(2,)}{X}$ 와 같아야 한다. $3 \neq 2$ ($(3,)$ vs $(2,3)$)

그리고 Multiplication node 이므로 $\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \cdot W$ ($\frac{\partial L}{\partial X}$: $(2,)$)

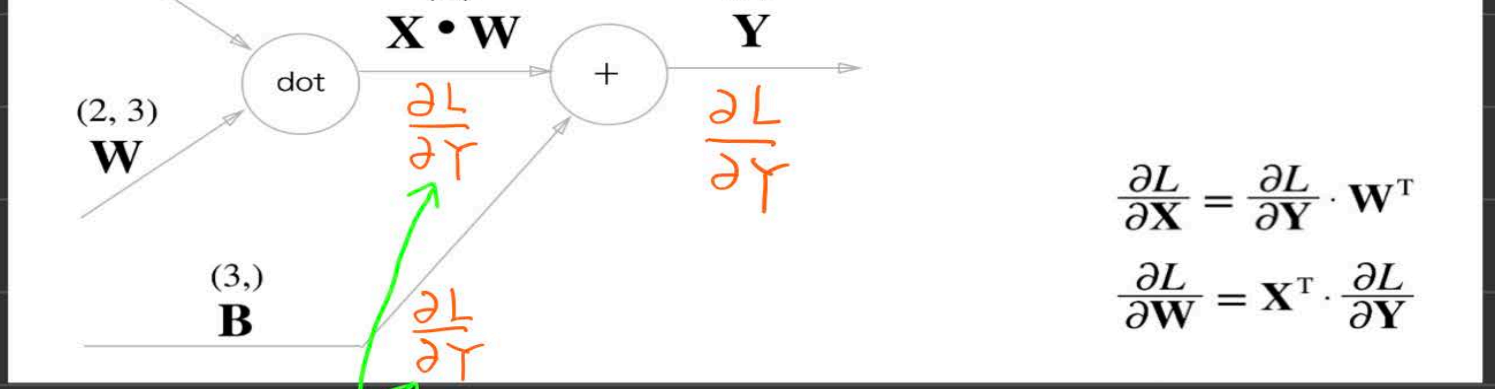


$X = [x_1, x_2]$, $W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$

$$\frac{\partial L}{\partial Y} \overbrace{\frac{\partial Y}{\partial Q}}^{1} = \overbrace{\frac{\partial L}{\partial Y}}$$

$\frac{\partial L}{\partial X}$ 의 size는 $X^{(2,)}$ 와 같아야 한다.

그리고 Multiplication node 이므로 $\frac{\partial L}{\partial X}^{(2,)} = \frac{\partial L}{\partial Y}^{(3,)} \cdot W^{(2,3)} = \frac{\partial L}{\partial Y} W^{T}$ $(3,)\ (3,2)$



$$X = [x_1, x_2], \quad W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

$$\frac{\partial L}{\partial Y}\overbrace{\frac{\partial Y}{\partial Q}}^{1} = \frac{\partial L}{\partial Y}$$

$\frac{\partial L}{\partial X}$ 의 size는 $X^{(2,)}$ 와 같아야 한다.

그리고 Multiplication node 이므로 $\frac{\partial L}{\partial X}^{(2,)} = \frac{\partial L}{\partial Y}^{(3,)} \cdot W^{(2,3)} = \frac{\partial L}{\partial Y}^{(3,)} W^{T(3,2)}$

$$\frac{\partial L}{\partial W}^{(2,3)} = X^{T(2\times1)} \frac{\partial L}{\partial Y}^{(3,)}$$

$$X = [x_1, x_2], \quad W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

$$\frac{\partial L}{\partial Y} \overbrace{\frac{\partial Y}{\partial Q}}^{1} = \frac{\partial L}{\partial Y}$$

$\frac{\partial L}{\partial X}$ 의 size는 $X^{(2,)}$ 와 같아야 한다.

그리고 Multiplication node 이므로 $\frac{\partial L}{\partial X}^{(2,)} = \frac{\partial L}{\partial Y}^{(3,)} \cdot W^{(2,3)} = \frac{\partial L}{\partial Y} W^T$

$$\frac{\partial L}{\partial W}^{(2,3)} = X^T \frac{\partial L}{\partial Y}^{(3,)}$$

(2×1)

where $X^T = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $W^T = \begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix}$

1 $\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$

(2,) (3,) (3, 2)

2 $\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$

(2, 3) (2, 1) (1, 3)

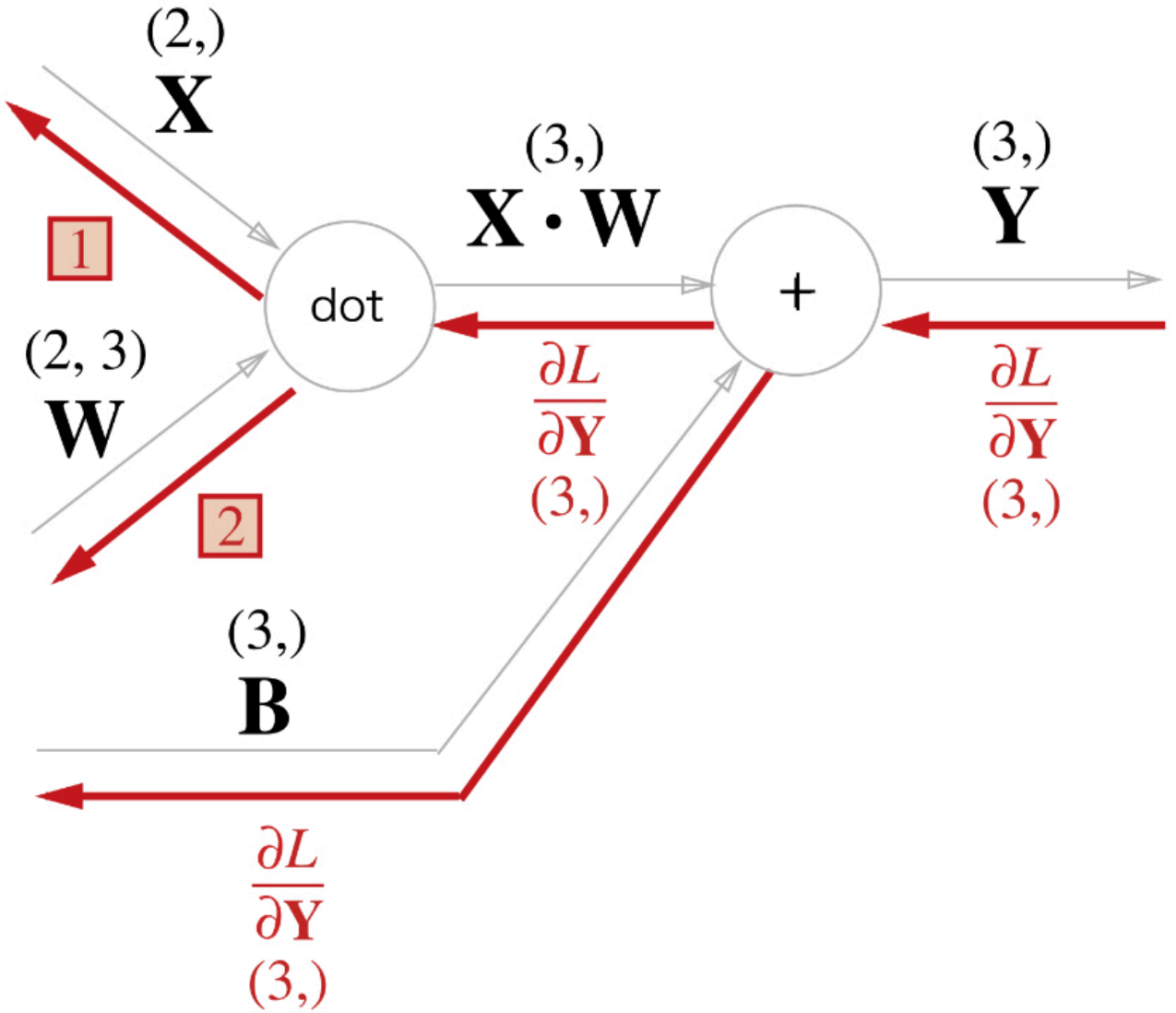
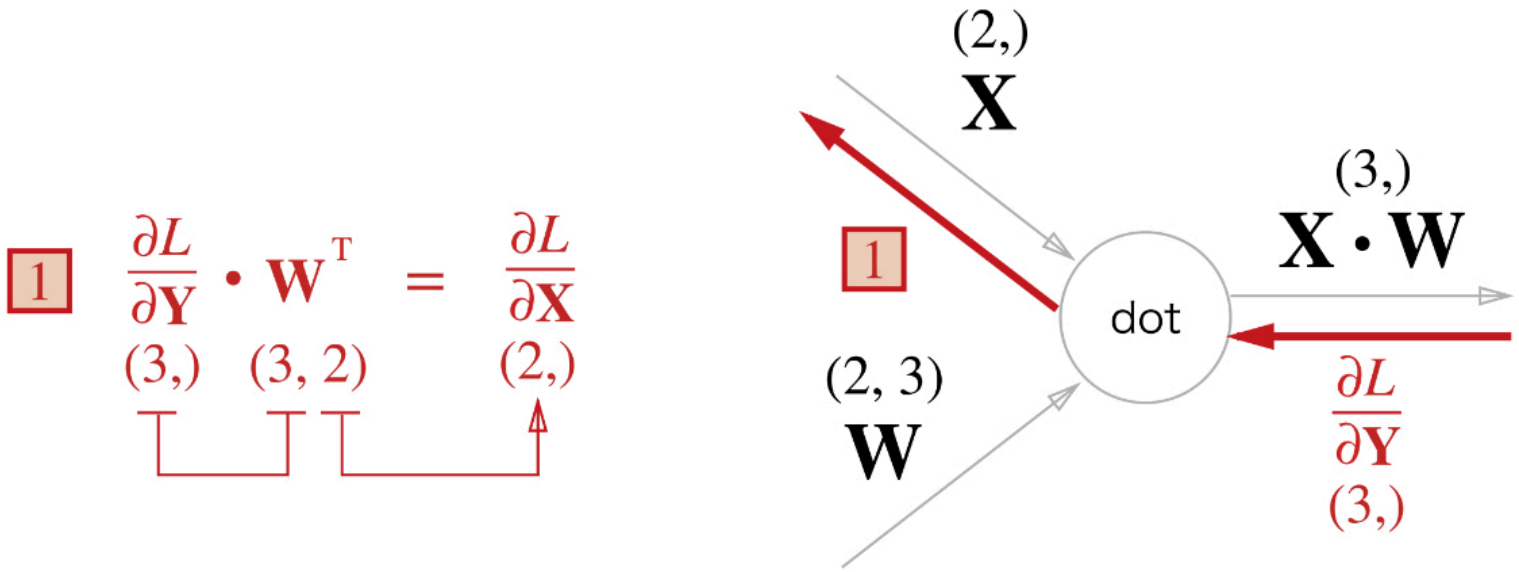


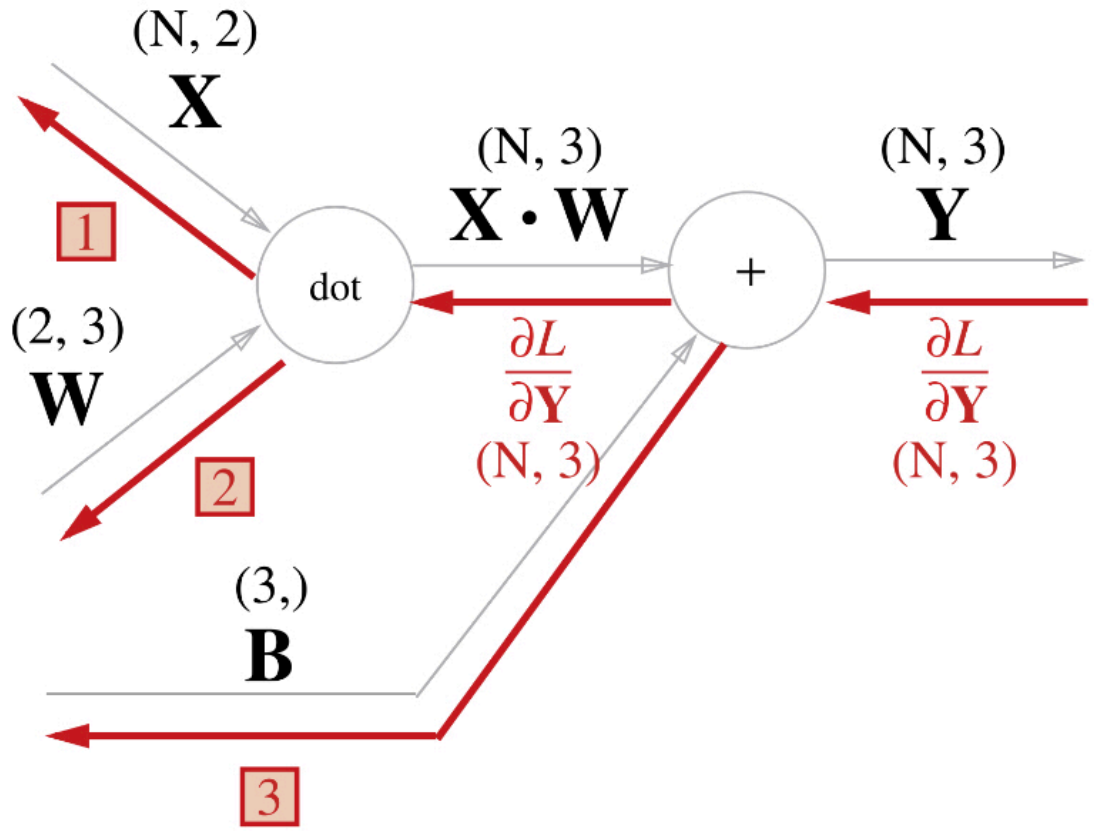
그림 5-26 행렬 곱('dot' 노드)의 역전파는 행렬의 대응하는 차원의 원소 수가 일치하도록 곱을 조립하여 구할 수 있다.



$$\mathbf{X} = (x_0, x_1, \cdots, x_n)$$

$$\frac{\partial L}{\partial \mathbf{X}} = \left(\frac{\partial L}{\partial x_0}, \frac{\partial L}{\partial x_1}, \cdots, \frac{\partial L}{\partial x_n} \right)$$

5.6.2 Affine Layer for Batch



1 $\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$

(N, 2) (N, 3) (3, 2)

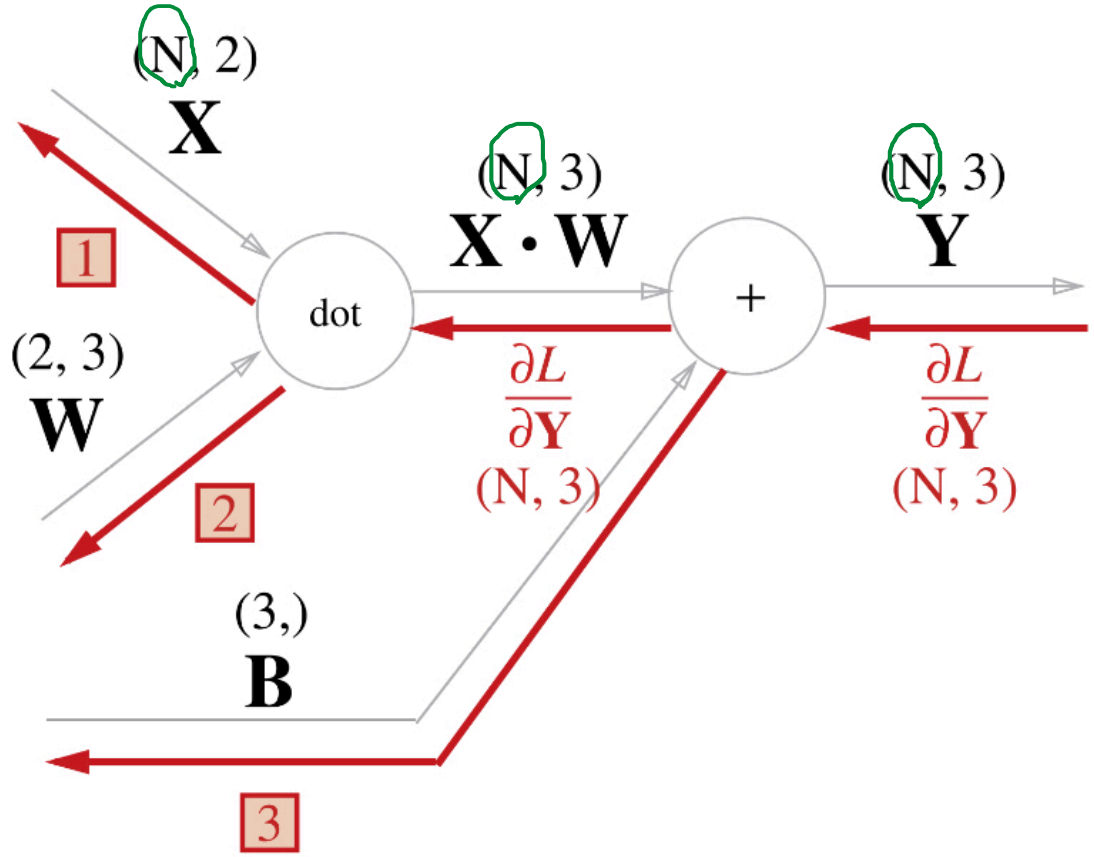
2 $\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$

(2, 3) (2, N) (N, 3)

3 $\frac{\partial L}{\partial \mathbf{B}} = \frac{\partial L}{\partial \mathbf{Y}}$ 의 첫 번째 축(0축, 열방향)의 합

(3) (N, 3)

5.6.2 Affine Layer for Batch



1 $\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$

(N, 2) (N, 3) (3, 2)

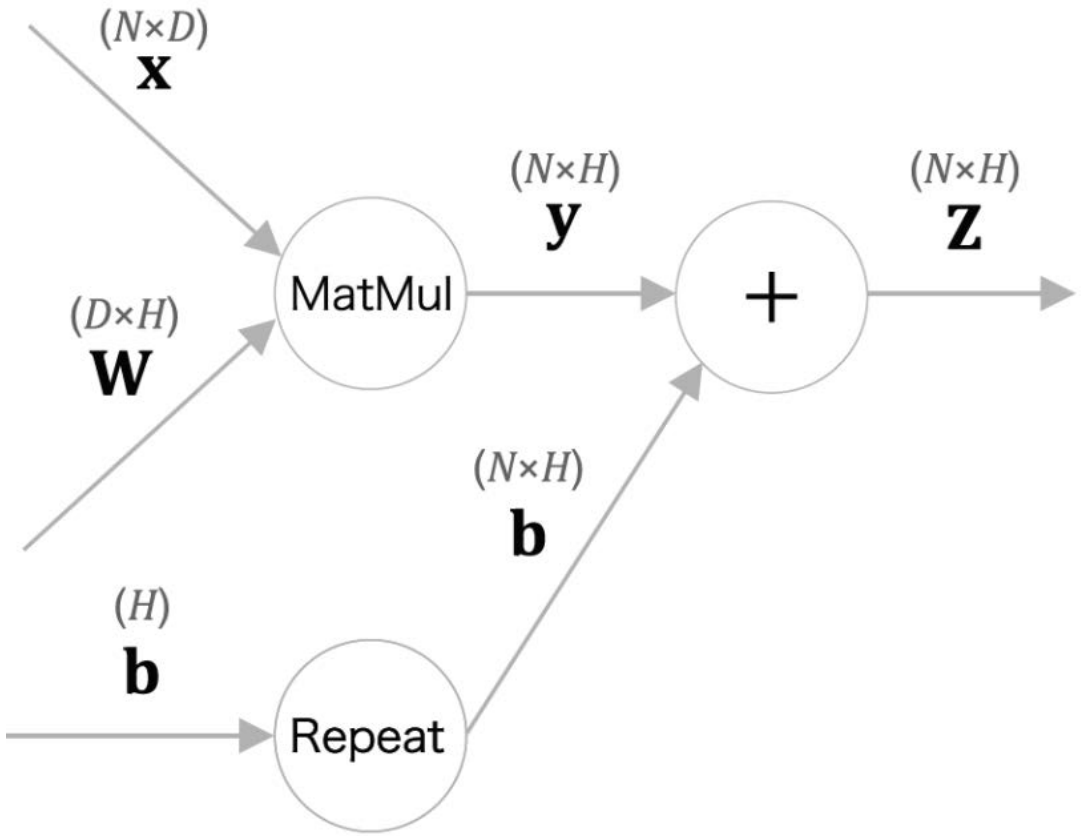
2 $\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$

(2, 3) (2, N) (N, 3)

3 $\frac{\partial L}{\partial \mathbf{B}} = \frac{\partial L}{\partial \mathbf{Y}}$ 의 첫 번째 축(0축, 열방향)의 합

(3) (N, 3)

Affine layer's Computation Graph : Batch Process



Sigmoid function's Derivative

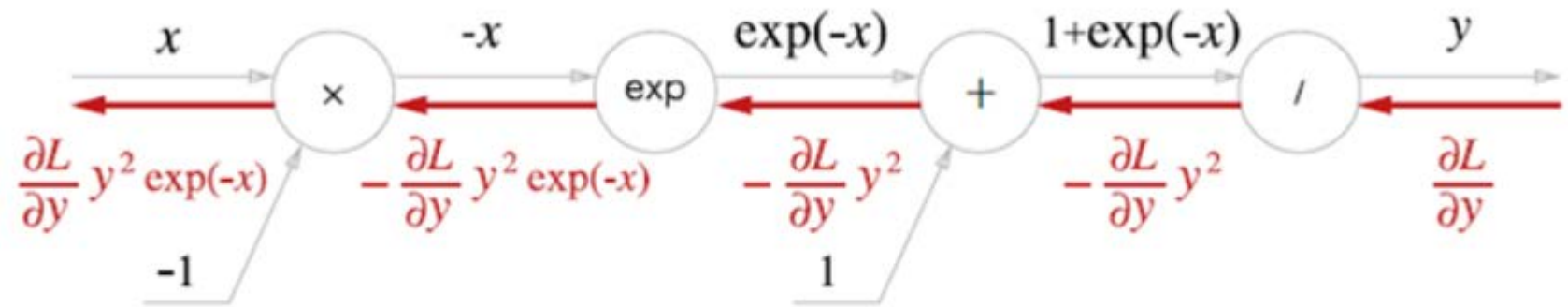
$$y = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1}$$

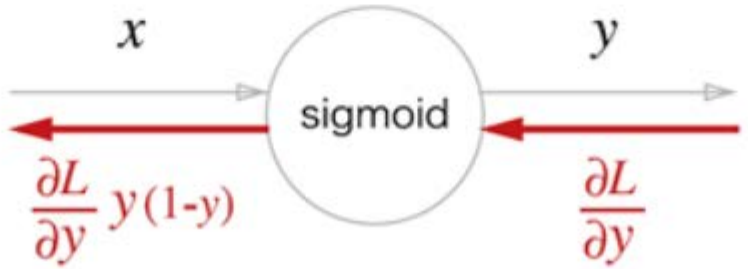
미분

$$\begin{aligned}
y' &= -(1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1+e^{-x}-1}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} \\
&= y - y^2 \\
&= y(1-y)
\end{aligned}$$

Sigmoid function‘s Computation Graph

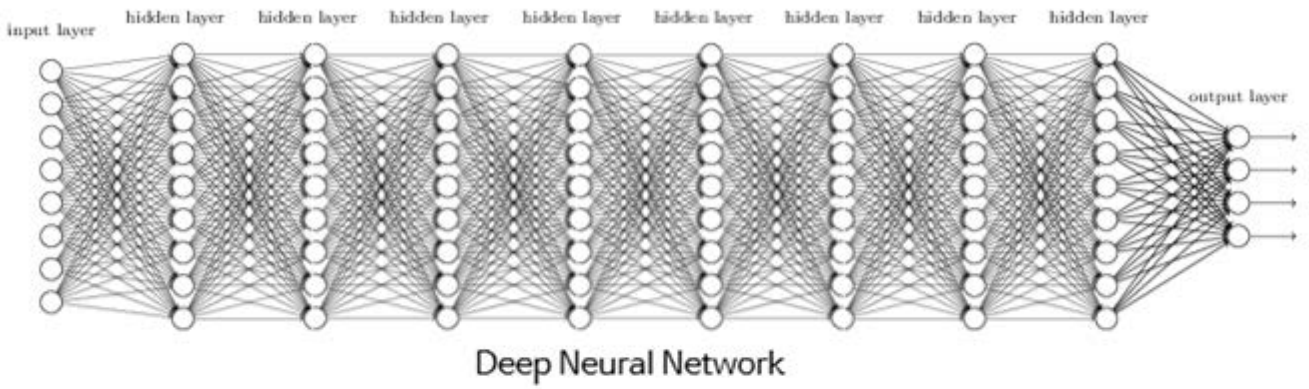
$$y = \frac{1}{1 + \exp(-x)}$$



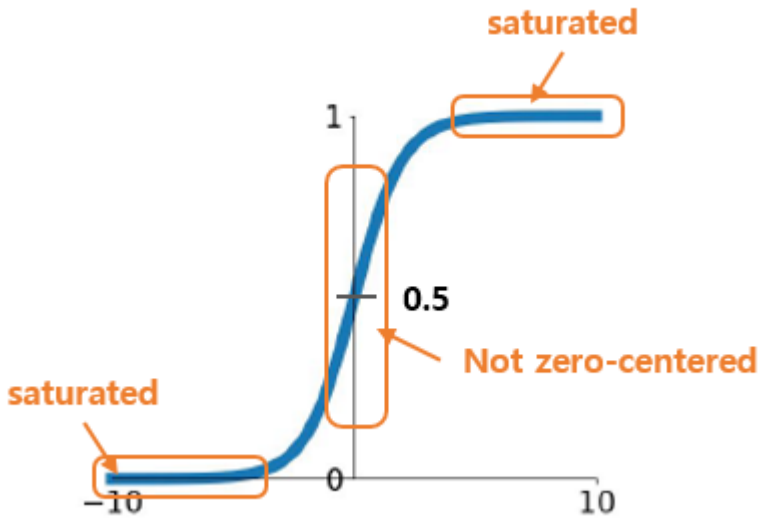


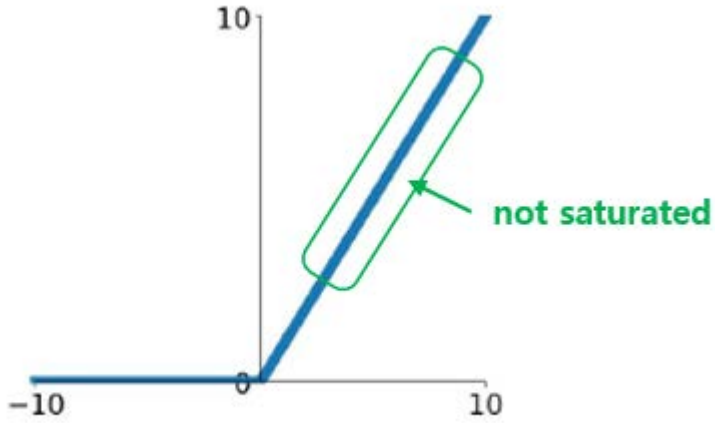
$$\begin{aligned}
\frac{\partial L}{\partial y} y^2 exp(-x) &= \frac{\partial L}{\partial y} \frac{1}{(1 + exp(-x))^2} \exp(-x) \\
&= \frac{\partial L}{\partial y} \frac{1}{1 + exp(-x)} \frac{exp(-x)}{1 + exp(-x)} \\
&= \frac{\partial L}{\partial y} y(1 - y)
\end{aligned}$$

Vanishing Gradient



Sigmoid vs ReLu

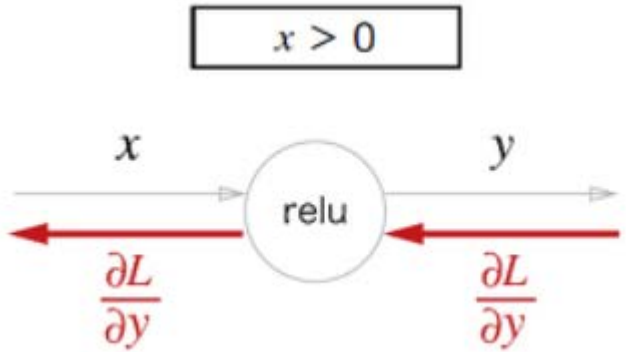


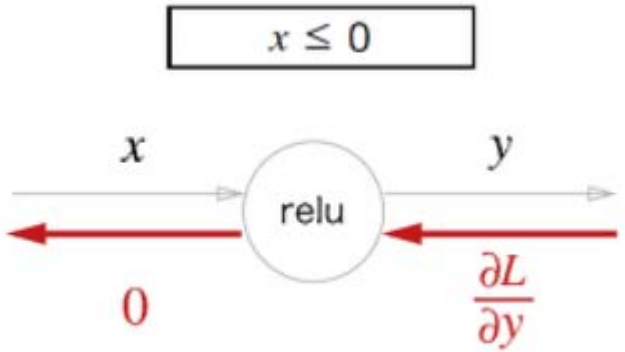


Relu function's Computation Graph

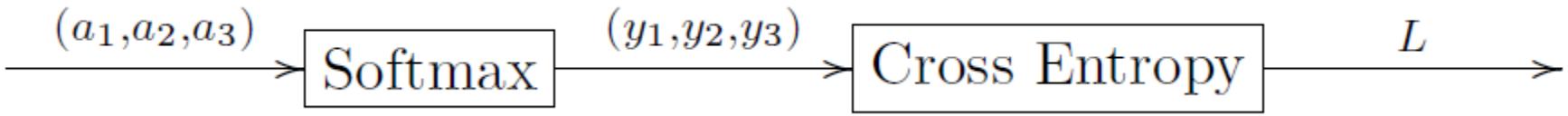
$$y = \begin{cases} x & (x > 0) \\ 0 & (x \leq 0) \end{cases}$$

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & (x > 0) \\ 0 & (x \leq 0) \end{cases}$$





Softmax-with-Loss Function's Derivative



$$
\begin{aligned}
L(a_1, a_2, a_3) &= -t_1 \log y_1 - t_2 \log y_2 - t_3 \log y_3 \\
&= -t_1 \log \frac{e^{a_1}}{e^{a_1} + e^{a_2} + e^{a_3}} - t_2 \log \frac{e^{a_2}}{e^{a_1} + e^{a_2} + e^{a_3}} - t_3 \log \frac{e^{a_3}}{e^{a_1} + e^{a_2} + e^{a_3}} \\
&= -t_1 \log e^{a_1} - t_2 \log e^{a_2} - t_3 \log e^{a_3} + (t_1 + t_2 + t_3) \log(e^{a_1} + e^{a_2} + e^{a_3}) \\
&= -t_1 a_1 - t_2 a_2 - t_3 a_3 + \log(e^{a_1} + e^{a_2} + e^{a_3})
\end{aligned}
$$

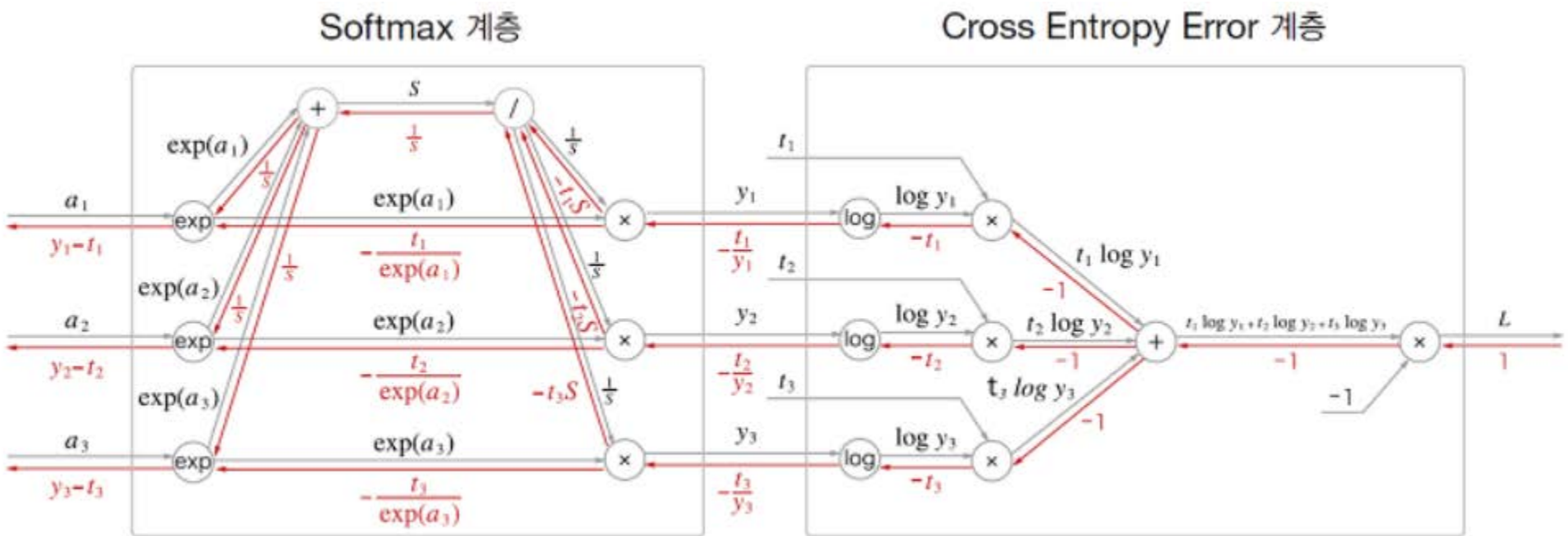
$$
\frac{\partial L}{\partial a_i} = -t_i + \frac{e^{a_i}}{e^{a_1} + e^{a_2} + e^{a_3}} = -t_i + y_i
$$

$$
\frac{\partial L}{\partial a} = (y_1 - t_1, y_2 - t_2, y_3 - t_3) = y - t
$$

Softmax-with-Loss layer's Computation Graph

$$\left(\frac{e^{a_1}}{e^{a_1} + e^{a_2} + e^{a_3}}, \frac{e^{a_2}}{e^{a_1} + e^{a_2} + e^{a_3}}, \frac{e^{a_3}}{e^{a_1} + e^{a_2} + e^{a_3}} \right)$$

$$-t_1 \log y_1 - t_2 \log y_2 - t_3 \log y_3$$

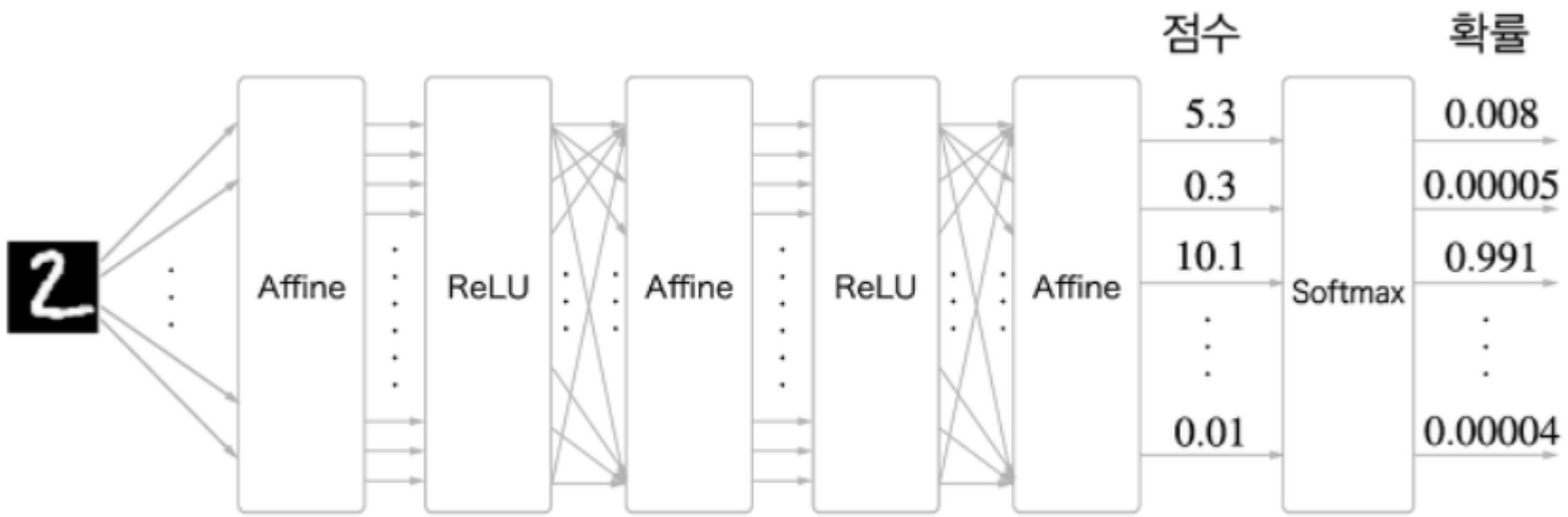


점수(score) vs 확률분포

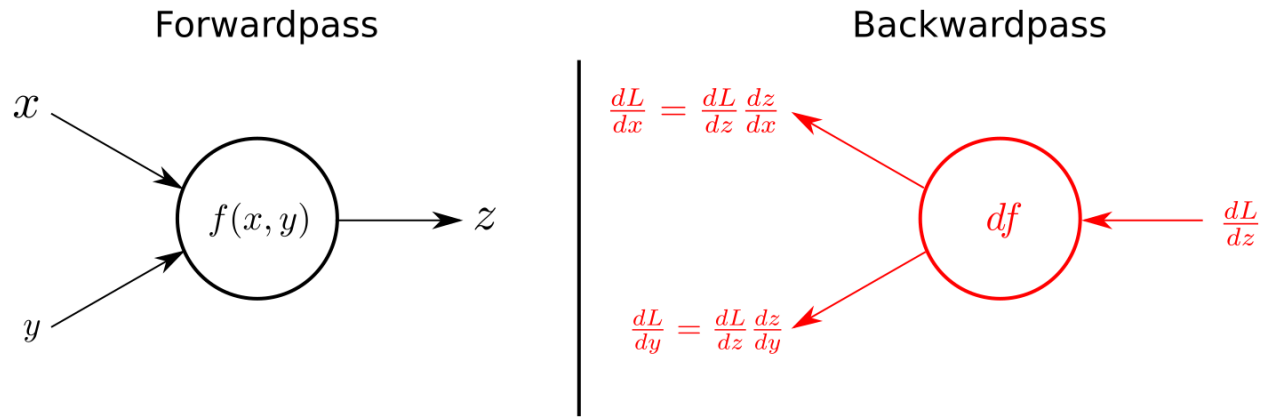
학습중일때는 확률분포사이의 거리를 재기 위해 softmax층과 loss층이 필요하다.

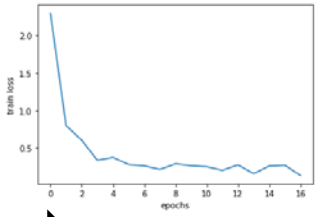
학습이 끝나고 추론을 할 때는 softmax층과 loss층은 계산비용의 낭비일 뿐이다.

최대의 점수가 나오는 항만 알면 된다.



Learning through BP (Chain rule)

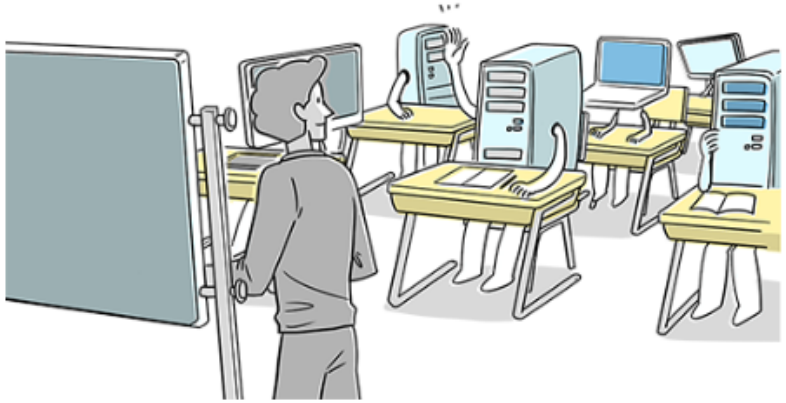


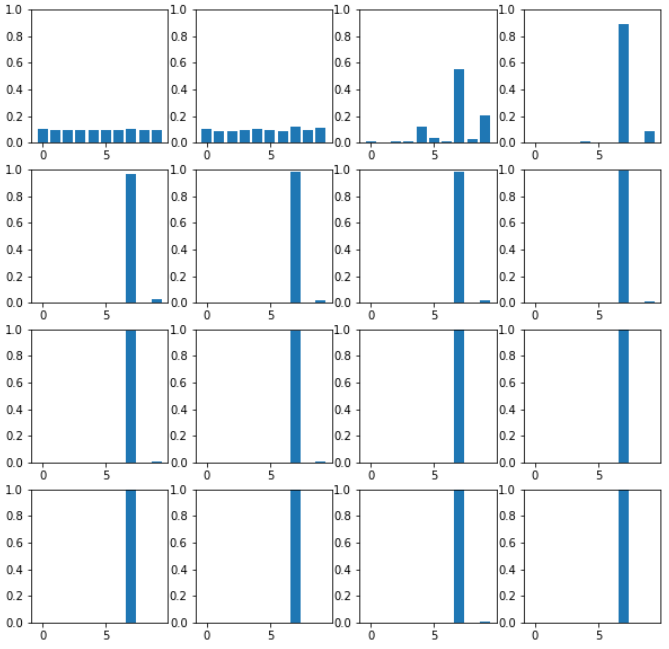


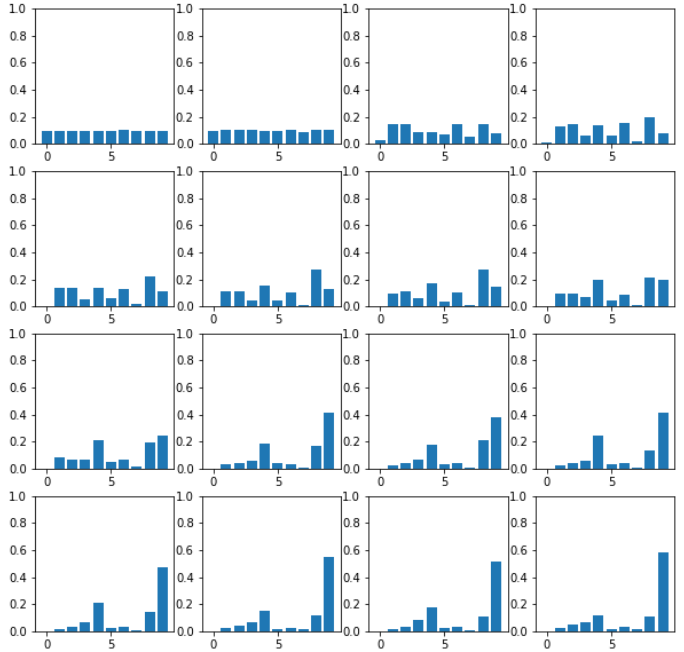
며칠이 몇초로 줄어들었어요!

97% accuracy

Machine Learning

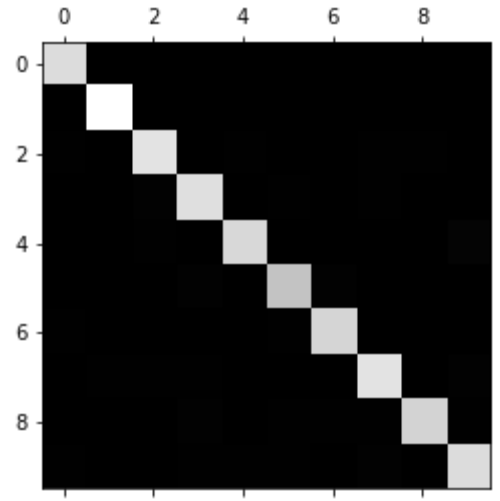




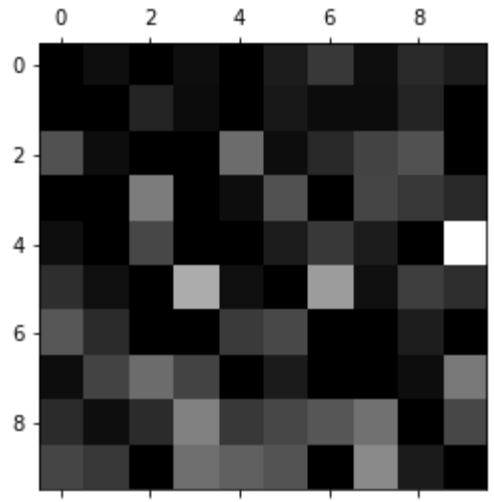


Confusion Matrix

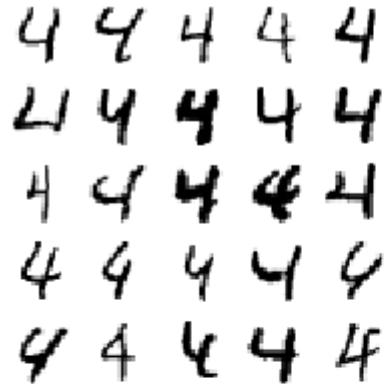
```
[[ 966    1    0    1    0    2    4    1    3    2]
 [   0 1124    3    1    0    2    1    1    3    0]
 [   6    1 1002    0    8    1    3    5    6    0]
 [   0    0    9  982    1    6    0    5    4    3]
 [   1    0    5    0  950    2    4    2    0   18]
 [   3    1    0   11    1  858   10    1    4    3]
 [   6    3    0    0    4    5  938    0    2    0]
 [   1    5    8    5    0    2    0  997    1    9]
 [   3    1    3    9    4    5    6    8  930    5]
 [   5    4    0    8    7    6    0   10    2  967]]
```



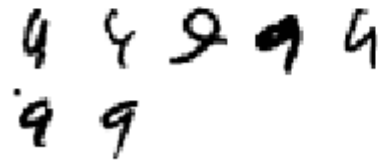
row : label
column : hand-written numbers that the machine predicts
element value : how many times it predicts



4를 4로예측



4를 9로 예측



9를 4로예측

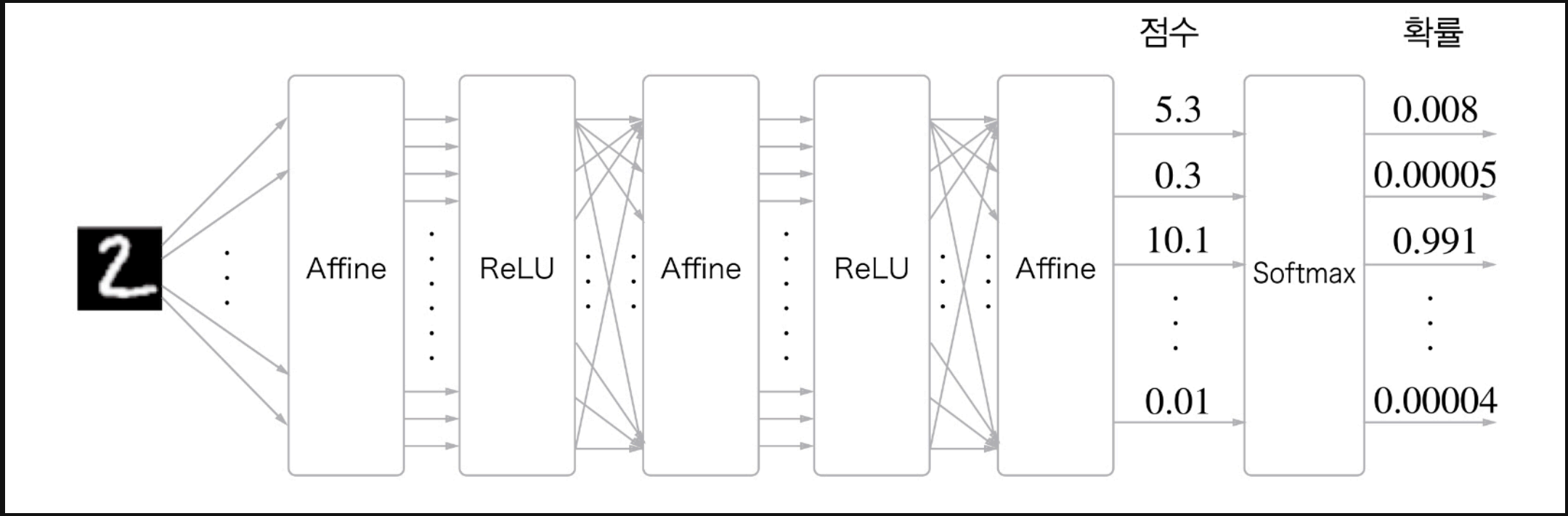
9를 9로 예측

Image Reference

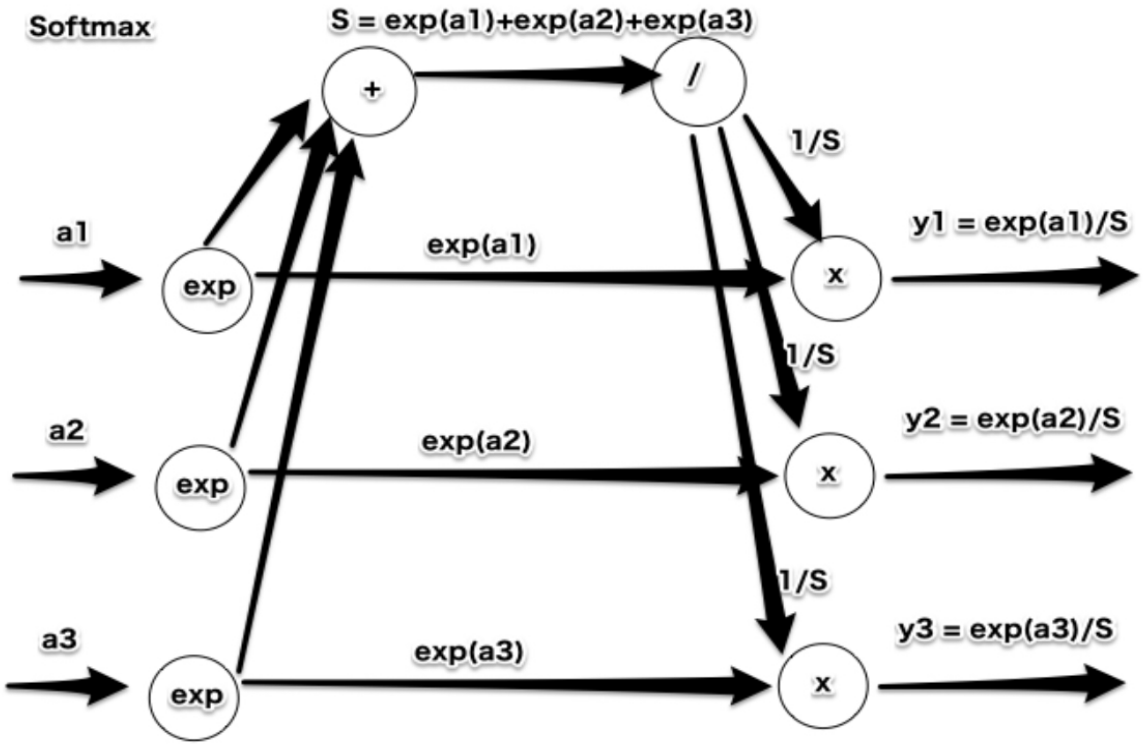
p.6~p.14, p.18, p.21, p.23 : https://excelsior-cjh.tistory.com/171

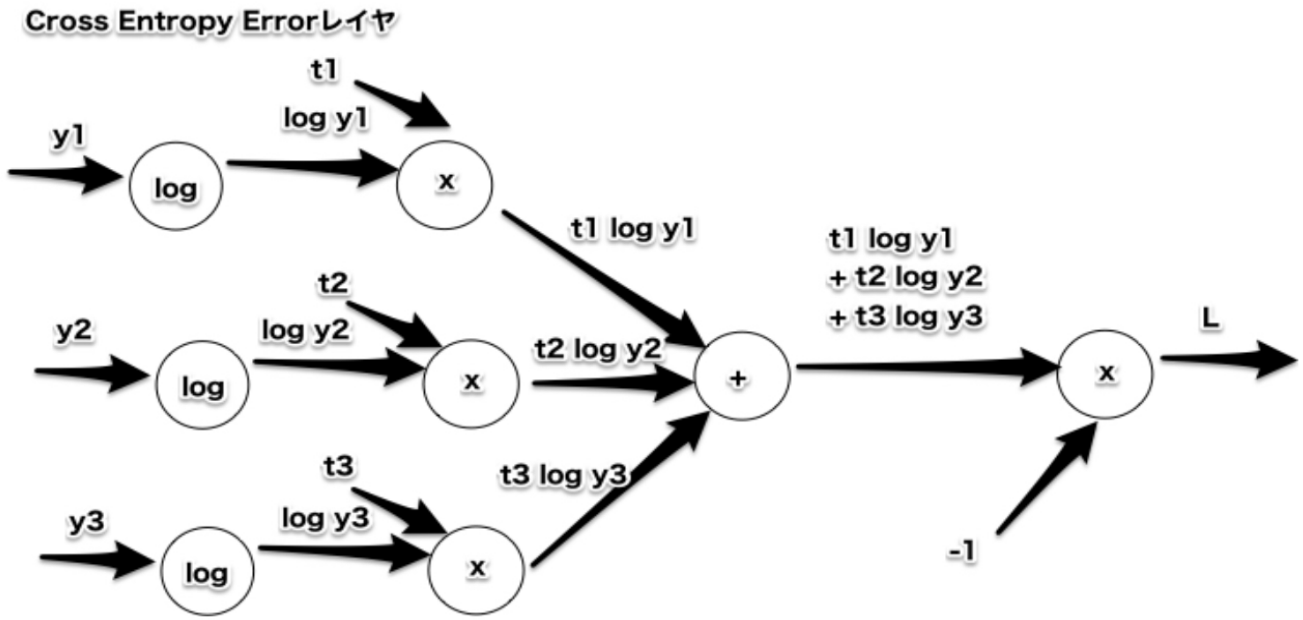
5.6.3 Softmax-with-Loss Layer

Softmax Layer normalize the 10 input values producing their probabilities.

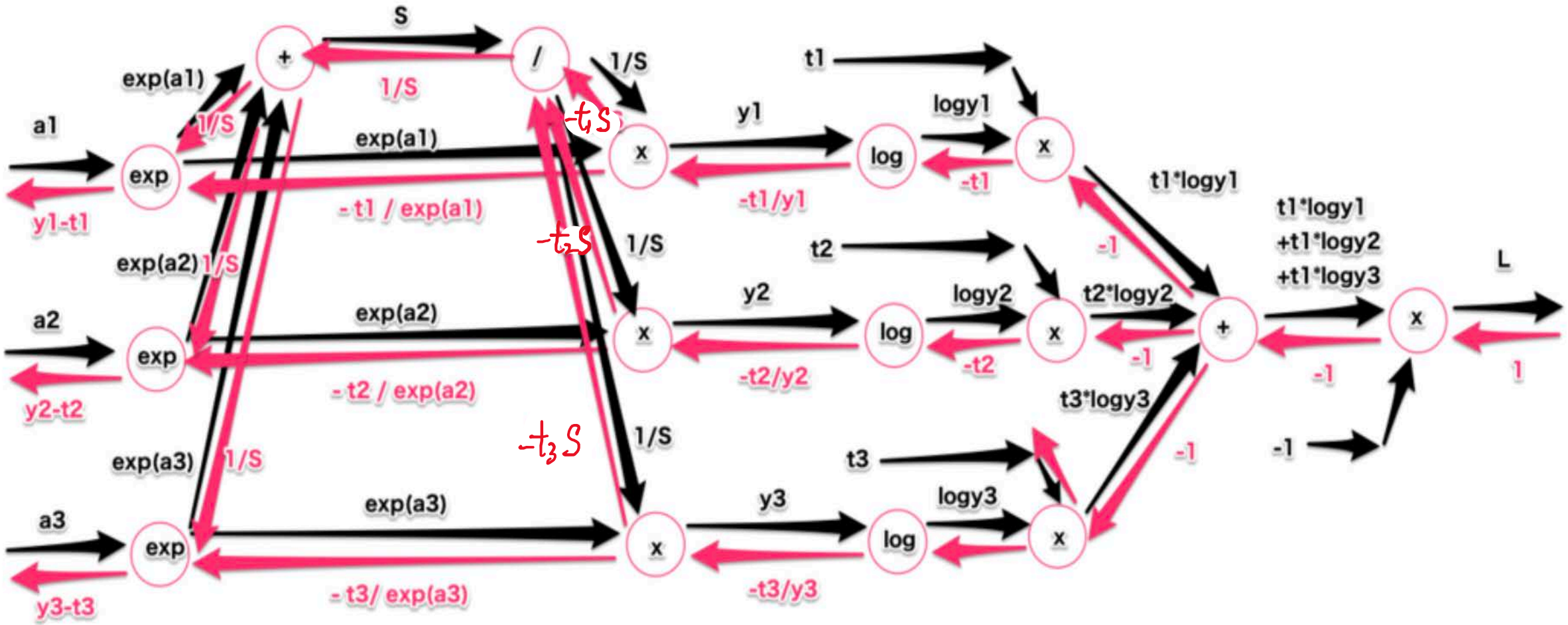


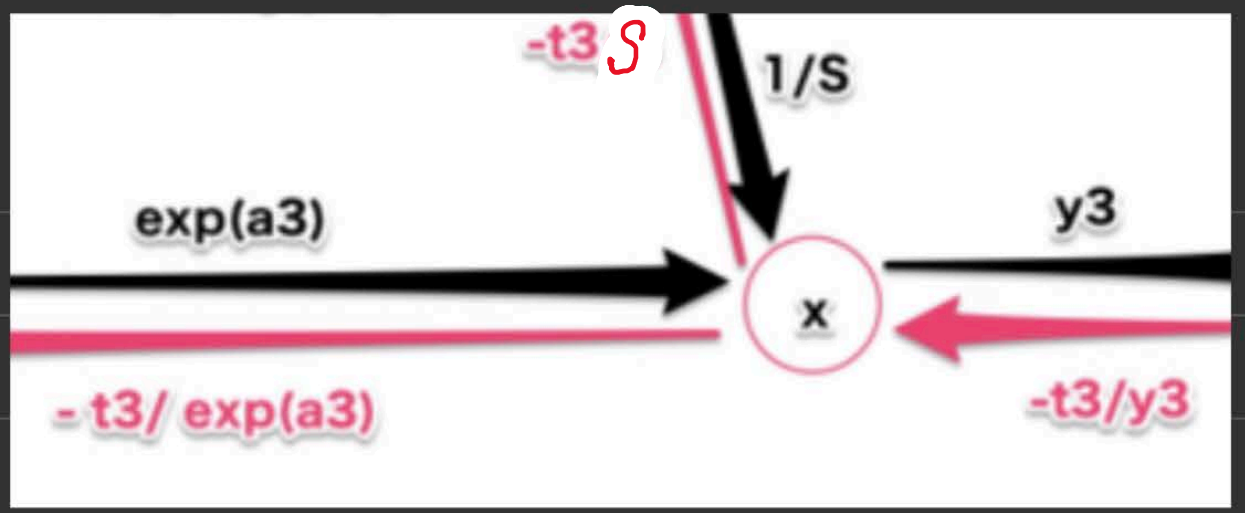
5.6.3 Softmax-with-Loss Layer





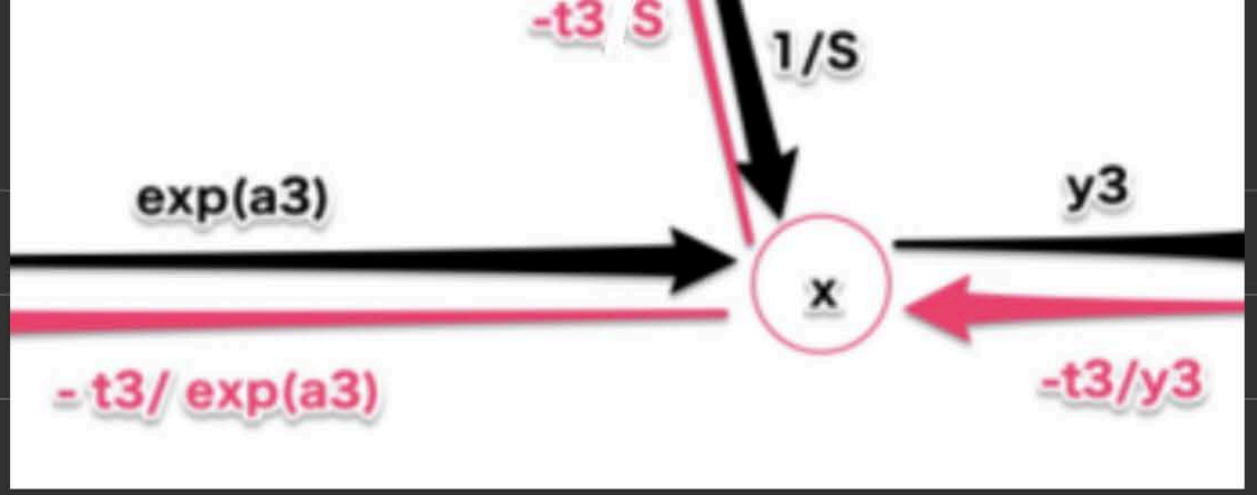
Back propagation of Softmax-with-Loss Layer





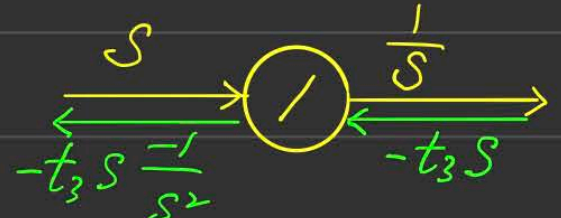
$$y_3 = \frac{\exp(a_3)}{\exp(a_1)+\exp(a_2)+\exp(a_3)}, \quad \frac{1}{S} = \frac{1}{\exp(a_1)+\exp(a_2)+\exp(a_3)}$$

$$-\frac{t_3}{y_3} \times \frac{1}{S} = \frac{-t_3}{\exp(a_3)}$$

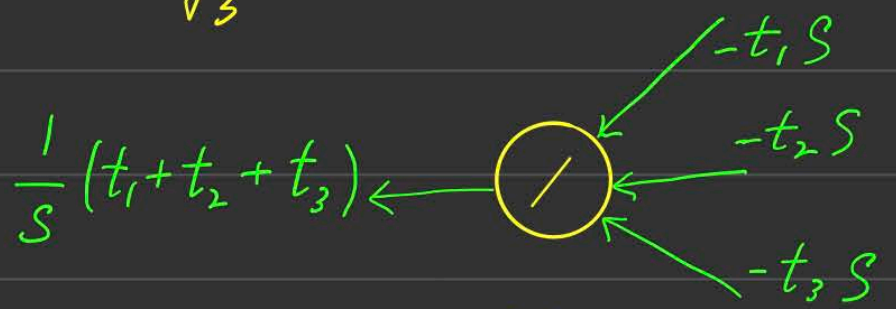


$$y_3 = \frac{\exp(a_3)}{\exp(a_1)+\exp(a_2)+\exp(a_3)}, \quad \frac{1}{S} = \frac{1}{\exp(a_1)+\exp(a_2)+\exp(a_3)}$$

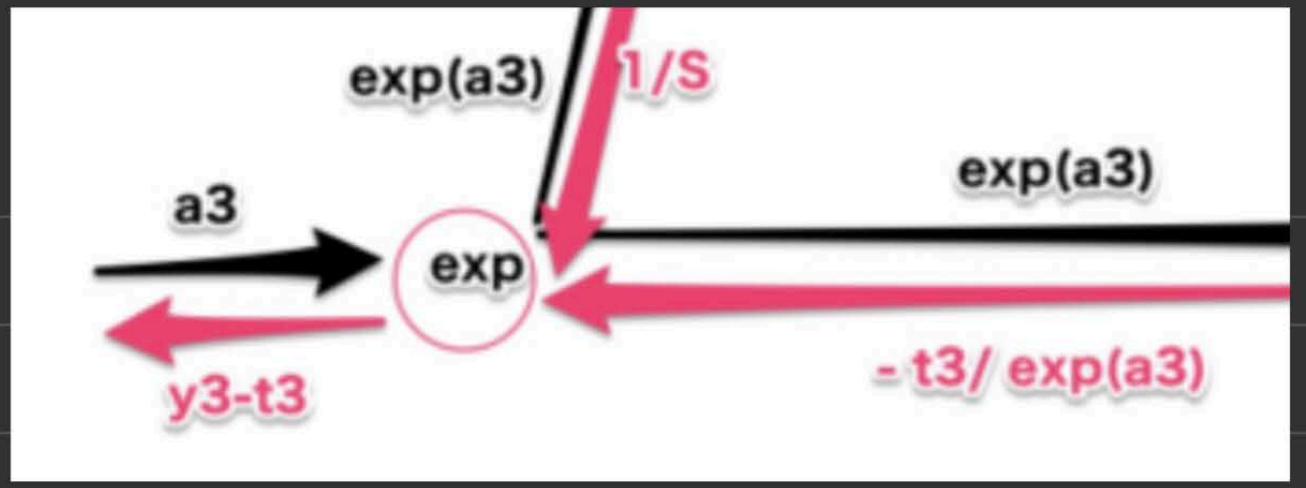
$$-\frac{t_3}{y_3} \times \frac{1}{S} = \frac{-t_3}{\exp(a_3)}, \quad -\frac{t_3}{y_3} \times \exp(a_3) = -t_3 S$$

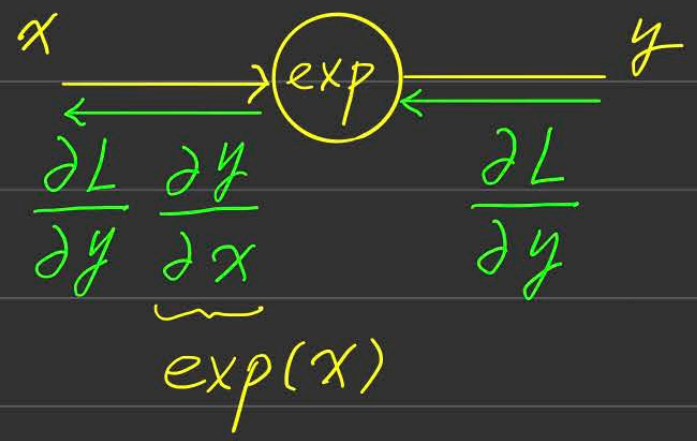


$= \frac{t_3}{S}$



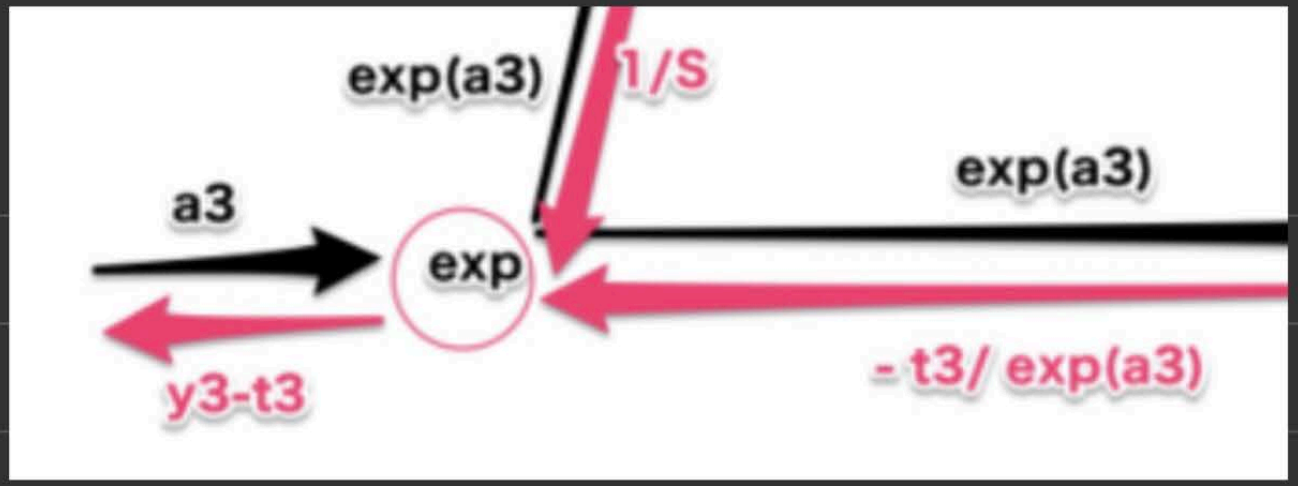
* t_{k} 은 정답 레이블 (target label) 로
정답에 해당하는 원소만 1이고 나머지는 0. $(t_1+t_2+t_3 = 1)$





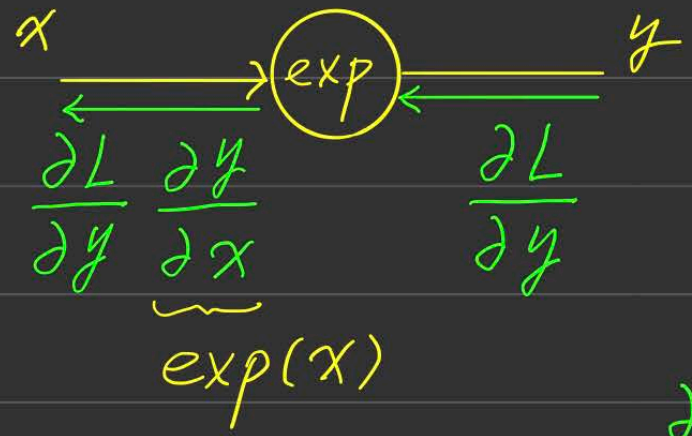
$$y = \exp(x)$$

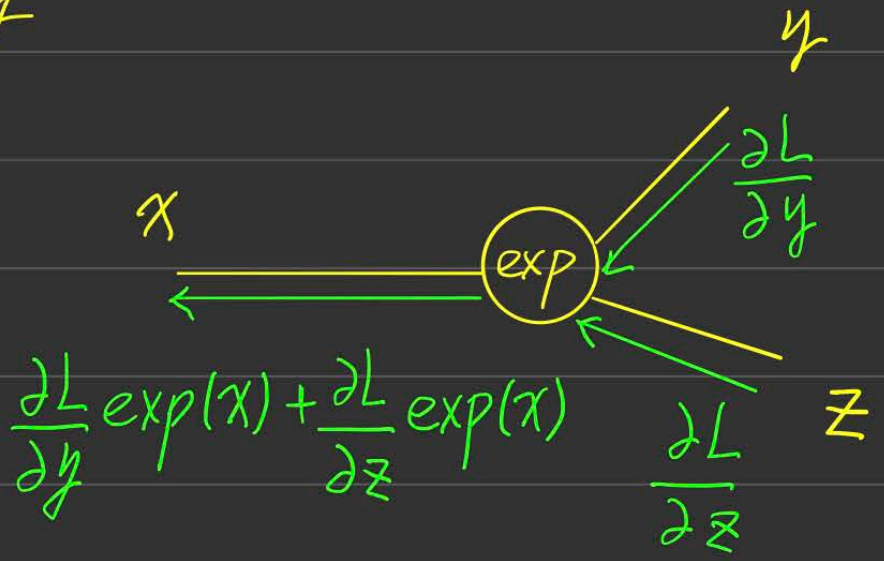
$$\frac{\partial y}{\partial x} = \exp(x)$$

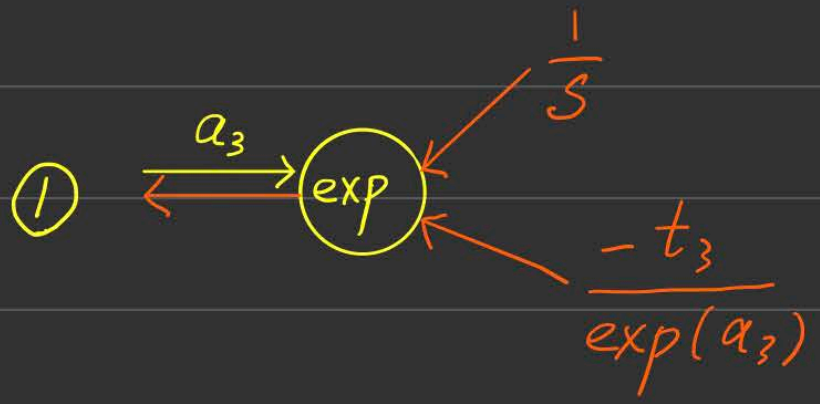


$$y = \exp(x)$$

$$\frac{\partial y}{\partial x} = \exp(x)$$

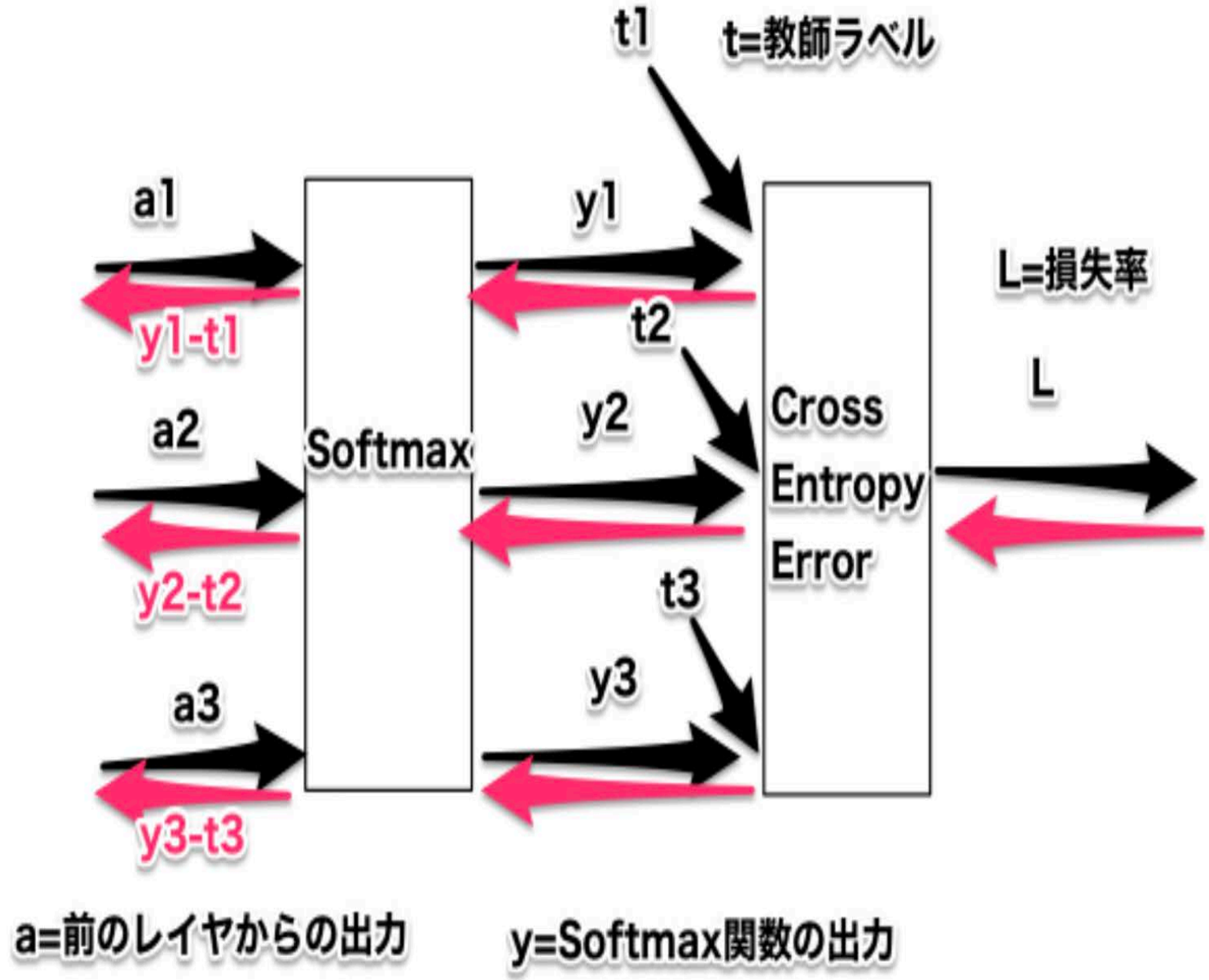




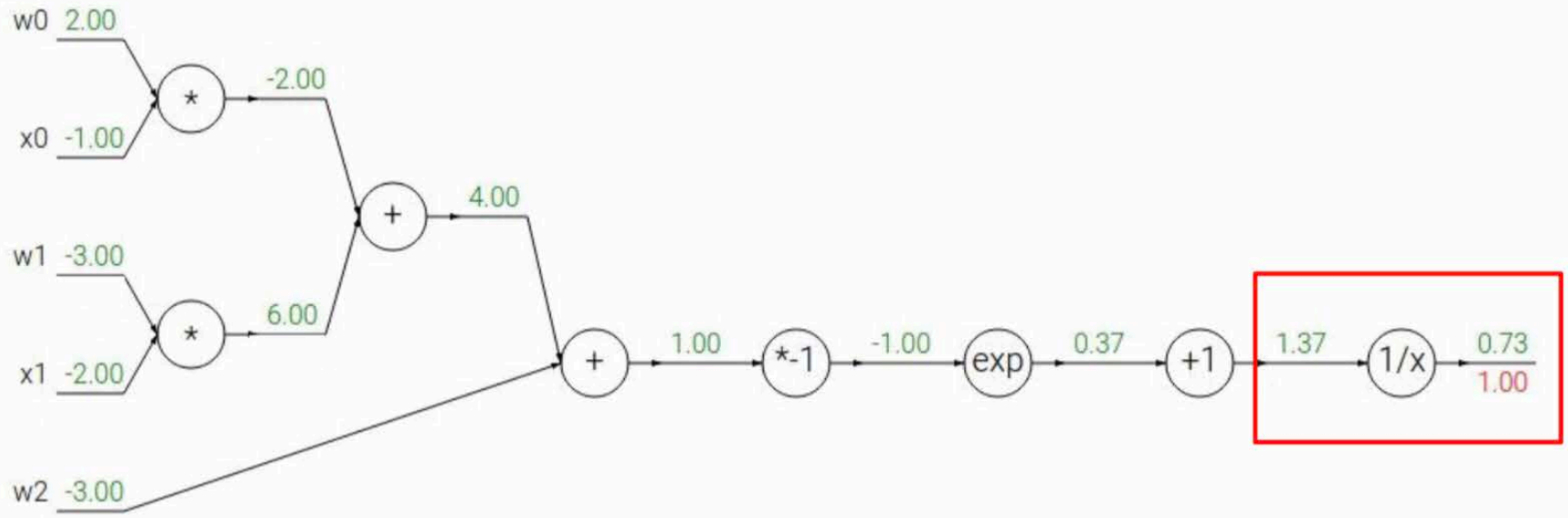


$$① = -\frac{t_3}{\exp(a_3)} \cdot \exp(a_3) + \frac{1}{S} \cdot \exp(a_3) = y_3 - t_3$$

Softmax-with-Loss Layer



Another example: $$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

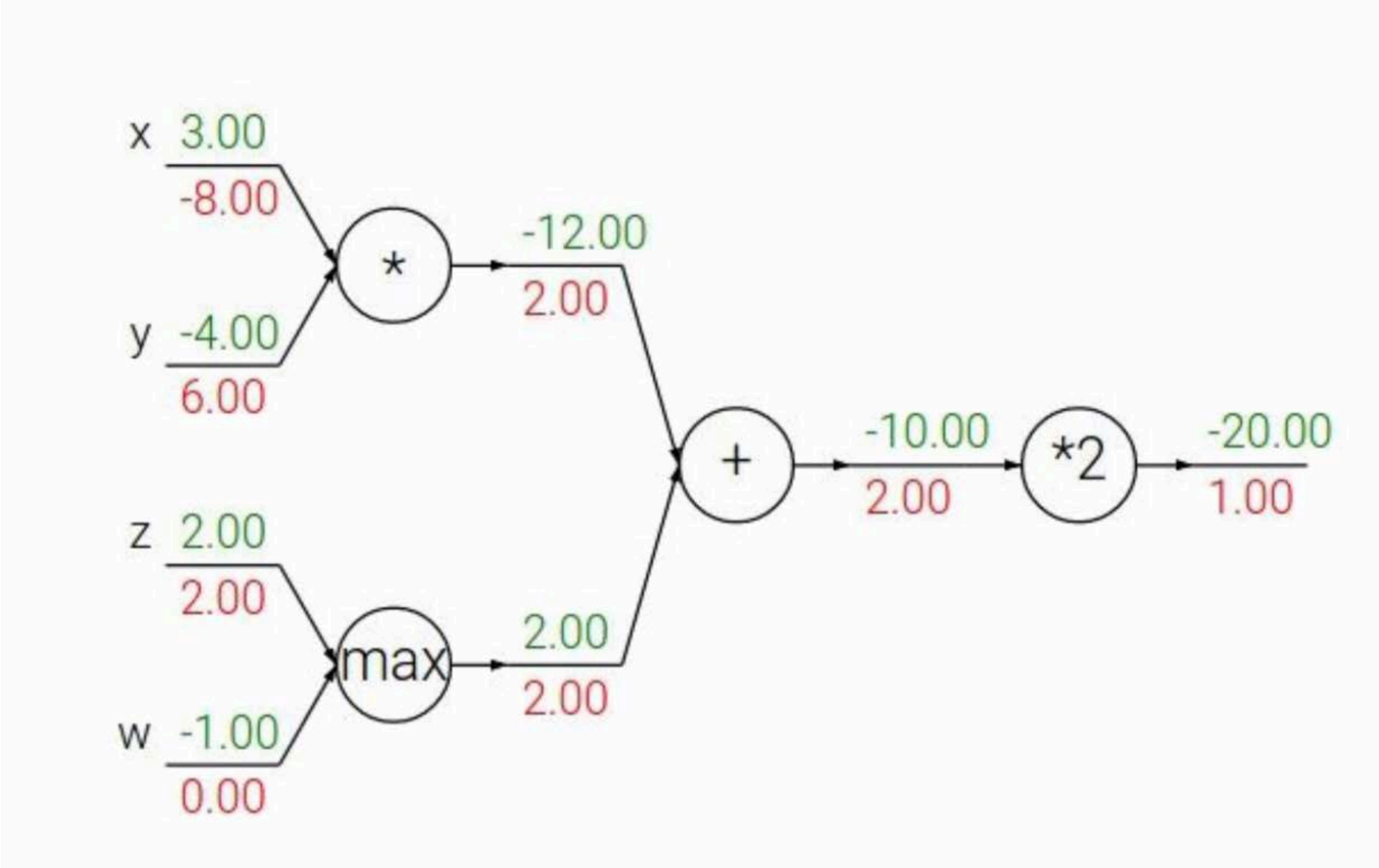


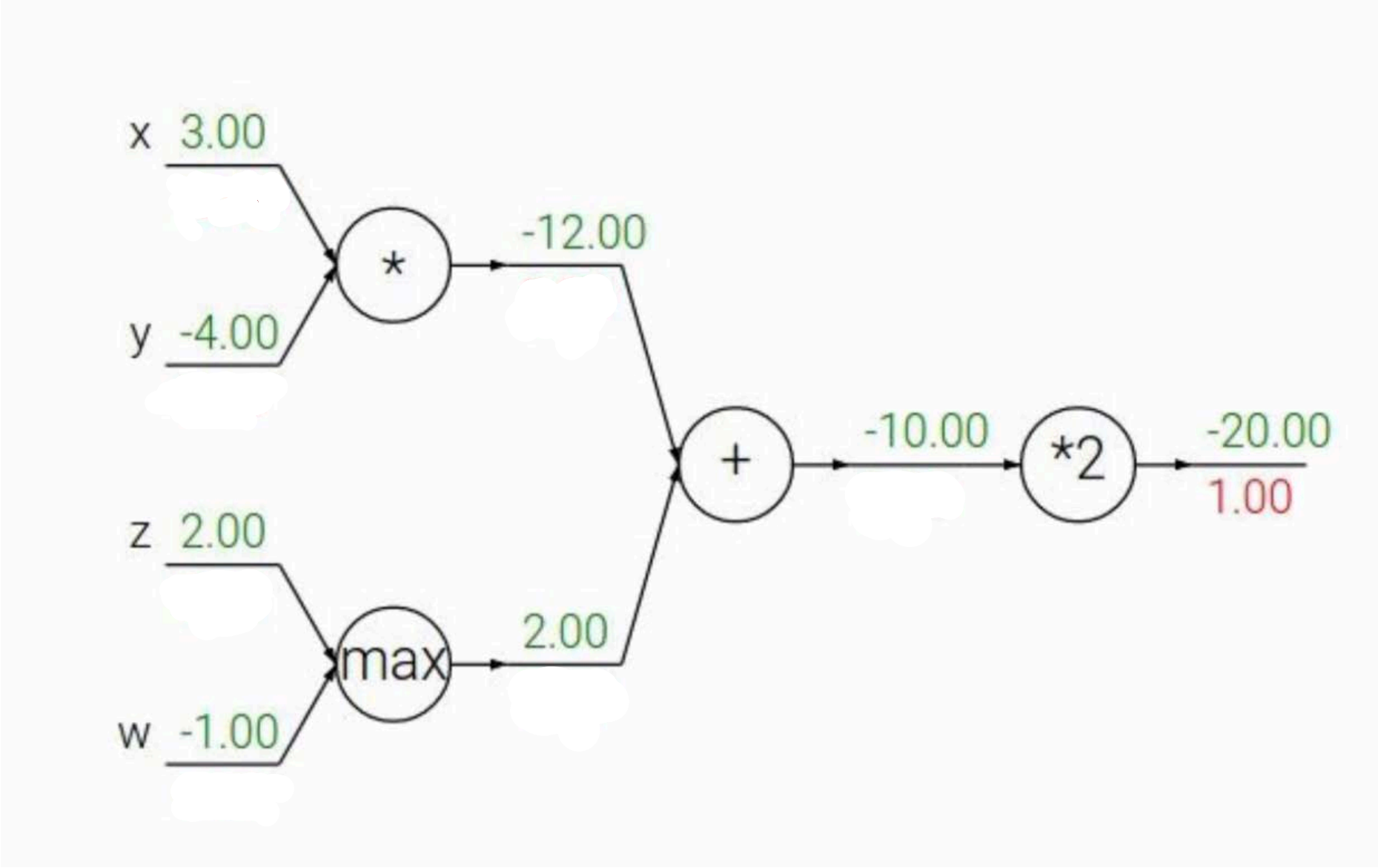
$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

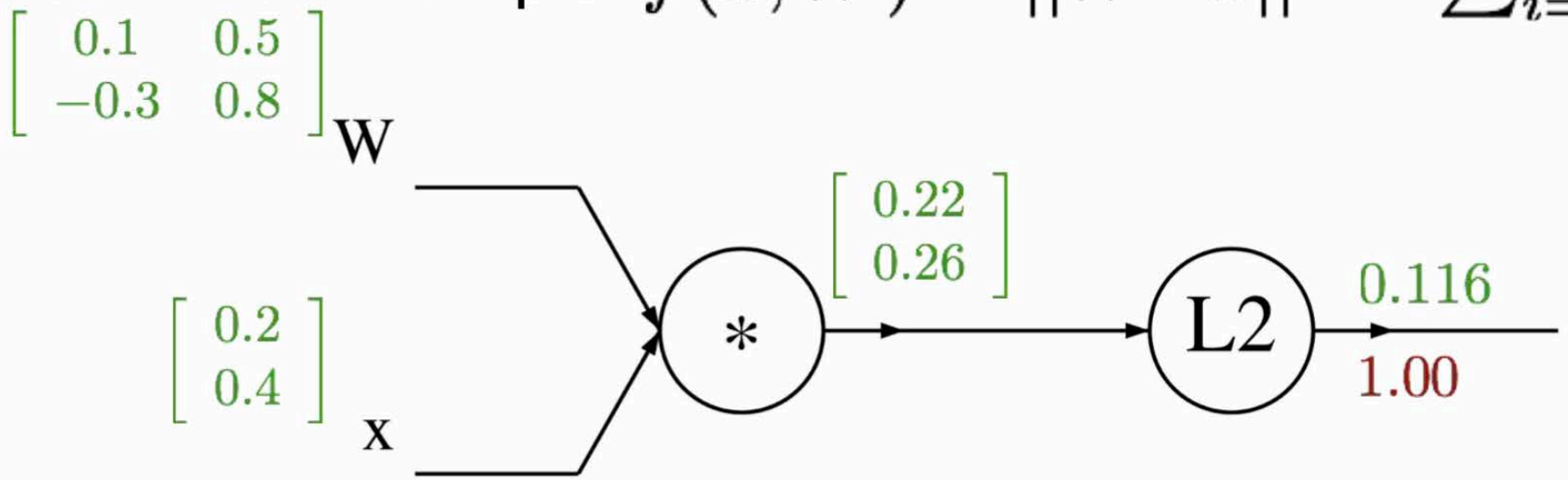
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

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


x
3.00
-8.00
y
-4.00
6.00
*
-12.00
2.00
z
2.00
2.00
w
-1.00
0.00
max
2.00
2.00
+
-10.00
2.00
*2
-20.00
1.00


x 3.00
y -4.00
*
-12.00
z 2.00
w -1.00
max
2.00
+
-10.00
*2
-20.00
1.00

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



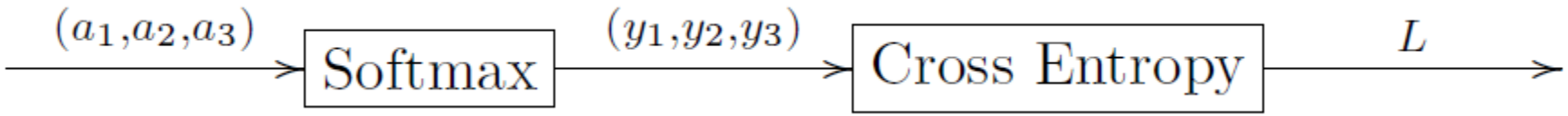
$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

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

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

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

Softmax-with-Loss Function's Derivative



$$
\begin{aligned}
L(a_1, a_2, a_3) &= -t_1 \log y_1 - t_2 \log y_2 - t_3 \log y_3 \\
&= -t_1 \log \frac{e^{a_1}}{e^{a_1} + e^{a_2} + e^{a_3}} - t_2 \log \frac{e^{a_2}}{e^{a_1} + e^{a_2} + e^{a_3}} - t_3 \log \frac{e^{a_3}}{e^{a_1} + e^{a_2} + e^{a_3}} \\
&= -t_1 \log e^{a_1} - t_2 \log e^{a_2} - t_3 \log e^{a_3} + (t_1 + t_2 + t_3) \log(e^{a_1} + e^{a_2} + e^{a_3}) \\
&= -t_1 a_1 - t_2 a_2 - t_3 a_3 + \log(e^{a_1} + e^{a_2} + e^{a_3})
\end{aligned}
$$

$$
\frac{\partial L}{\partial a_i} = -t_i + \frac{e^{a_i}}{e^{a_1} + e^{a_2} + e^{a_3}} = -t_i + y_i
$$

$$
\frac{\partial L}{\partial a} = (y_1 - t_1, y_2 - t_2, y_3 - t_3) = y - t
$$