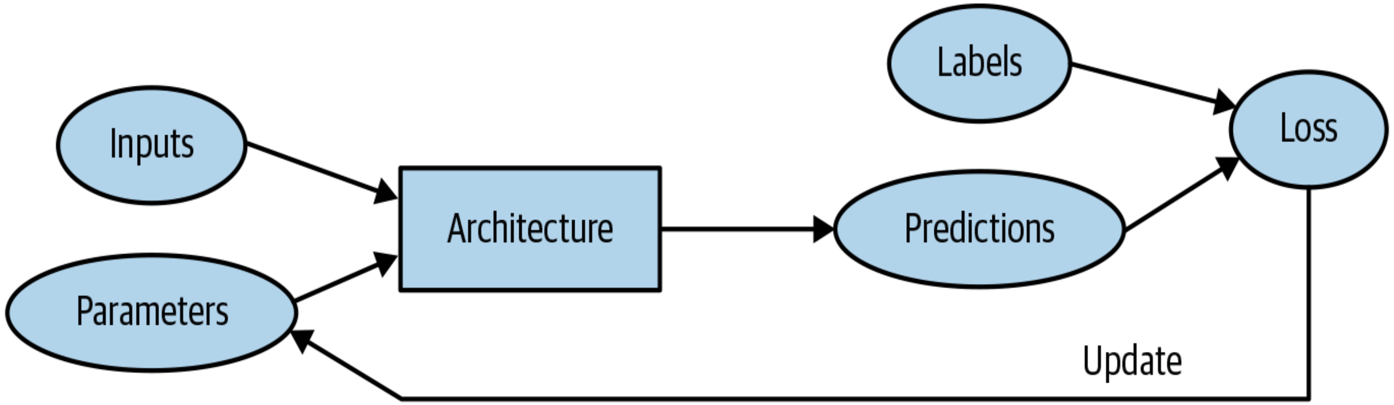
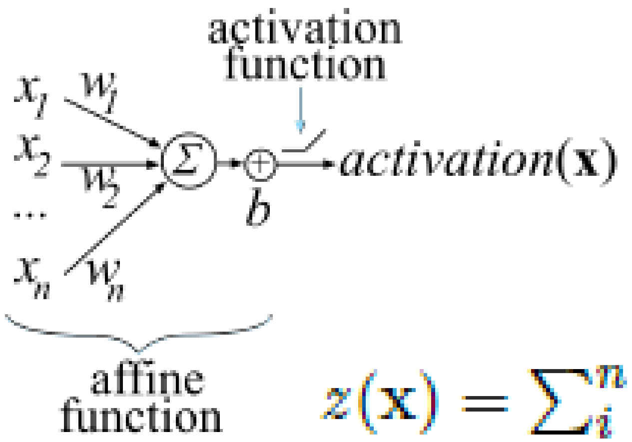
AI 프로그래밍 9

융합학과 권오영
oykwon@koreatech.ac.kr

Matrix Calculus

❖ https://arxiv.org/pdf/1802.01528.pdf





(mean squared error):

$$\frac{1}{N}\sum_{\mathbf{x}}(target(\mathbf{x}) - activation(\mathbf{x}))^2 = \frac{1}{N}\sum_{\mathbf{x}}(target(\mathbf{x}) - max(0, \sum_{i}^{|x|} w_i x_i + b))^2$$

$$z(\mathbf{x}) = \sum_i^n w_i x_i + b = \mathbf{w} \cdot \mathbf{x} + b.$$

미분 정리

Rule	$f(x)$	Scalar derivative notation with respect to x	Example
Constant	c	0	$\frac{d}{dx}99 = 0$
Multiplication by constant	cf	$c\frac{df}{dx}$	$\frac{d}{dx}3x = 3$
Power Rule	x^n	nx^{n-1}	$\frac{d}{dx}x^3 = 3x^2$
Sum Rule	$f + g$	$\frac{df}{dx} + \frac{dg}{dx}$	$\frac{d}{dx}(x^2 + 3x) = 2x + 3$
Difference Rule	$f - g$	$\frac{df}{dx} - \frac{dg}{dx}$	$\frac{d}{dx}(x^2 - 3x) = 2x - 3$
Product Rule	fg	$f\frac{dg}{dx} + \frac{df}{dx}g$	$\frac{d}{dx}x^2x = x^2 + x2x = 3x^2$
Chain Rule	$f(g(x))$	$\frac{df(u)}{du}\frac{du}{dx}$, let $u = g(x)$	$\frac{d}{dx}ln(x^2) = \frac{1}{x^2}2x = \frac{2}{x}$

$$\frac{d}{dx}9(x + x^2) = 9\frac{d}{dx}(x + x^2) = 9(\frac{d}{dx}x + \frac{d}{dx}x^2) = 9(1 + 2x) = 9 + 18x$$

미분 정리

❖ $f(x0 + dx) = f(x0) + f'(x0)dx$

```python
def f(x):
    return 9*(x+x*x)

def df(x):
    return 9+18*x

print(f(1))

print('f(1+0.01) = ', f(1+0.01), 'dy = ', df(1)*0.01)
print('f(1+0.001) = ', f(1+0.001), 'dy = ', df(1)*0.001)
print('f(1+0.0001) = ', f(1+0.0001), 'dy = ', df(1)*0.0001)
```

Shell

```
18
f(1+0.01) =  18.2709 dy =  0.27
f(1+0.001) =  18.027008999999993 dy =  0.027
f(1+0.0001) =  18.00270009 dy =  0.0027
```

편미분

❖ 변수가 여럿일 때 각 변수별로 미분

$$f(x,y) = 3x^2y.$$

$$\frac{\partial}{\partial x}3yx^2 = 3y\frac{\partial}{\partial x}x^2 = 3y2x = 6yx.$$

$$\frac{\partial}{\partial y}3x^2y = 3x^2\frac{\partial}{\partial y}y = 3x^2\frac{\partial y}{\partial y} = 3x^2 \times 1 = 3x^2.$$

❖ gradient of f(x, y)

$$\nabla f(x,y) = [\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}] = [6yx, 3x^2]$$

단위 벡터를 곱하면 scalar값이 됨

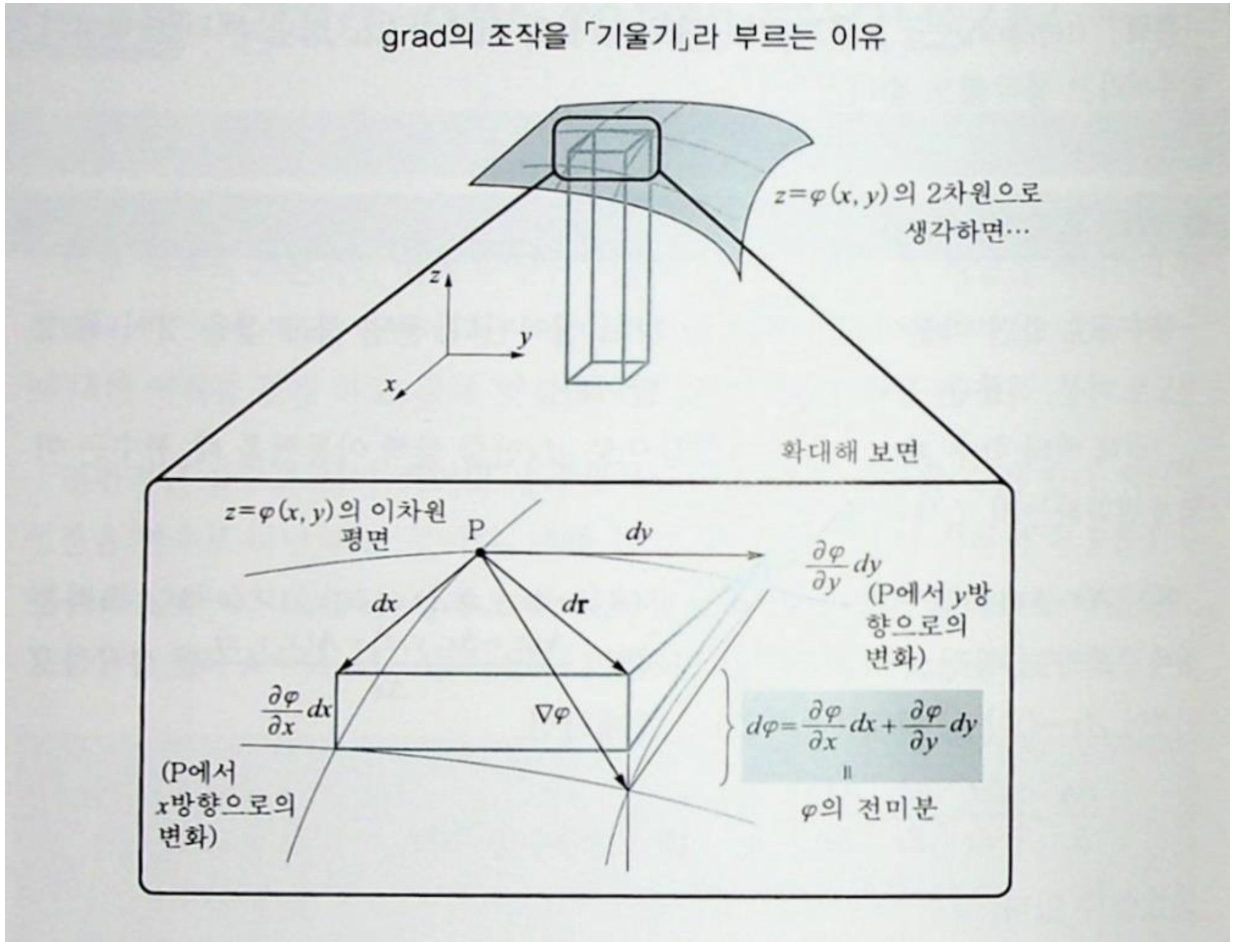
```python
def f(x,y):
    return 3*x*x*y

def dfx(x, y):
    return 6*y*x

def dfy(x, y):
    return 3*x*x

print(f(1,1))

print('f(1+0.01,1) = ', f(1+0.01, 1), 'xdf = ', dfx(1,1)*0.01)
print('f(1+0.001, 1) = ', f(1+0.001, 1), 'xdf = ', dfx(1,1)*0.001)
print('f(1+0.0001, 1) = ', f(1+0.0001, 1), 'xdf = ', dfx(1,1)*0.0001)

print('f(1, 1+0.01) = ', f(1, 1+0.01), 'ydf = ', dfy(1,1)*0.01)
print('f(1, 1+0.001) = ', f(1, 1+0.001), 'ydf = ', dfy(1,1)*0.001)
print('f(1, 1+0.0001) = ', f(1, 1+0.0001), 'ydf = ', dfy(1,1)*0.0001)
```

Shell

```
3
f(1+0.01,1) =  3.0603000000000002 xdf =  0.06
f(1+0.001, 1) =  3.0060029999999993 xdf =  0.006
f(1+0.0001, 1) =  3.00060003 xdf =  0.0006000000000000001
f(1, 1+0.01) =  3.0300000000000002 ydf =  0.03
f(1, 1+0.001) =  3.0029999999999997 ydf =  0.003
f(1, 1+0.0001) =  3.0003 ydf =  0.00030000000000000003
```

Gradient

❖ 변화율과 그 방향을 보여줌



KOREATECH 한국기술교육대학교

matrix calculus

❖ Jacobian matrix: 다변수 벡터 함수의 편미분함수 행렬
(미시 영역에서 비선형적 변화를 선형변환으로 근사)

❖ m개의 함수, n개의 변수를 갖는 벡터 (m개의 gradient 벡터)

$$\begin{aligned} y_1 &= f_1(\mathbf{x}) \\ y_2 &= f_2(\mathbf{x}) \\ &\vdots \\ y_m &= f_m(\mathbf{x}) \end{aligned} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \nabla f_1(\mathbf{x}) \\ \nabla f_2(\mathbf{x}) \\ \cdots \\ \nabla f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}} f_1(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}} f_2(\mathbf{x}) \\ \cdots \\ \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \frac{\partial}{\partial x_1} f_2(\mathbf{x}) & \frac{\partial}{\partial x_2} f_2(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_2(\mathbf{x}) \\ & & \cdots & \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \frac{\partial}{\partial x_2} f_m(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix}$$

matrix calculus

❖ Jacobian의 모습

	scalar x	vector $\mathbf{x}$
scalar f	$\frac{\partial f}{\partial x}$	$\frac{\partial f}{\partial \mathbf{x}}$
vector $\mathbf{f}$	$\frac{\partial \mathbf{f}}{\partial x}$	$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$

chain rule

❖ single-variable chain rule -> divide and conquer

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

$$y = f(g(x)) = sin(x^2):$$

$$\begin{aligned} u &= x^2 && \text{(relative to definition } f(g(x)), g(x) = x^2) \\ y &= sin(u) && (y = f(u) = sin(u)) \end{aligned}$$

$$\begin{aligned} \frac{du}{dx} &= 2x \\ \frac{dy}{du} &= cos(u) \end{aligned}$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = cos(u)2x$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = cos(x^2)2x = 2xcos(x^2)$$

$$\begin{array}{c} y = sin \\ \uparrow \\ u = square \\ \uparrow \\ x \end{array} \quad \begin{array}{l} \left.\vphantom{\frac{a}{b}}\right\} \frac{dy}{du} \\ \left.\vphantom{\frac{a}{b}}\right\} \frac{du}{dx} \end{array}$$

single-variable chain rule

❖ Forward and backward differentiation

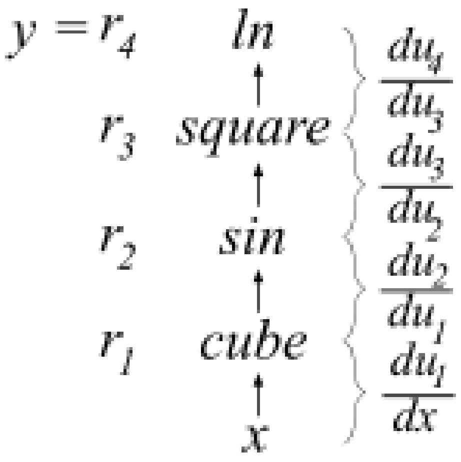
Forward differentiation from x to y

$$\frac{dy}{dx} = \frac{du}{dx}\frac{dy}{du}$$

Backward differentiation from y to x

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

$$y = f(x) = ln(sin(x^3)^2):$$



1. Introduce intermediate variables.

$$\begin{aligned} u_1 &= f_1(x) &&= x^3 \\ u_2 &= f_2(u_1) &&= sin(u_1) \\ u_3 &= f_3(u_2) &&= u_2^2 \\ u_4 &= f_4(u_3) &&= ln(u_3)(y = u_4) \end{aligned}$$

2. Compute derivatives.

$$\begin{aligned} \frac{d}{u_x}u_1 &= \frac{d}{x}x^3 &&= 3x^2 \\ \frac{d}{u_1}u_2 &= \frac{d}{u_1}sin(u_1) &&= cos(u_1) \\ \frac{d}{u_2}u_3 &= \frac{d}{u_2}u_2^2 &&= 2u_2 \\ \frac{d}{u_3}u_4 &= \frac{d}{u_3}ln(u_3) &&= \frac{1}{u_3} \end{aligned}$$

3. Combine four intermediate values.

$$\frac{dy}{dx} = \frac{du_4}{dx} = \frac{du_4}{du_3}\frac{du_3}{du_2}\frac{du_2}{du_1}\frac{du_1}{dx} = \frac{1}{u_3}2u_2cos(u_1)3x^2 = \frac{6u_2x^2cos(u_1)}{u_3}$$

4. Substitute.

$$\frac{dy}{dx} = \frac{6sin(u_1)x^2cos(x^3)}{u_2^2} = \frac{6sin(x^3)x^2cos(x^3)}{sin(u_1)^2} = \frac{6sin(x^3)x^2cos(x^3)}{sin(x^3)^2} = \frac{6x^2cos(x^3)}{sin(x^3)}$$

single-variable total-derivative chain rule

❖ 일반적인 수식으로 확장 예) $y = f(x) = x + x^2$

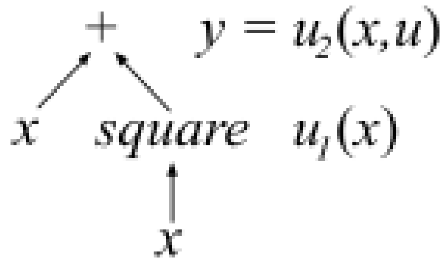
$$\frac{dy}{dx} = \frac{d}{dx}x + \frac{d}{dx}x^2 = 1 + 2x$$

$$\begin{aligned} u_1(x) &= x^2 \\ u_2(x, u_1) &= x + u_1 \qquad (y = f(x) = u_2(x, u_1)) \end{aligned}$$

Let's try it anyway to see what happens. If we pretend that $\frac{du_2}{du_1} = 0 + 1 = 1$ and $\frac{du_1}{dx} = 2x$, then $\frac{dy}{dx} = \frac{du_2}{dx} = \frac{du_2}{du_1}\frac{du_1}{dx} = 2x$ instead of the right answer $1 + 2x$.

$$\begin{aligned} \frac{\partial u_1(x)}{\partial x} &= 2x && (\text{same as } \frac{du_1(x)}{dx}) \\ \frac{\partial u_2(x,u_1)}{\partial u_1} &= \frac{\partial}{\partial u_1}(x + u_1) = 0 + 1 = 1 \\ \frac{\partial u_2(x,u_1)}{\partial x} &\neq \frac{\partial}{\partial x}(x + u_1) = 1 + 0 = 1 && (\text{something's not quite right here!}) \end{aligned}$$

❖ x에 대한 편미분을 할 때 라는 변수는 x에 영향을 받지 않아야 한다.
(u1이 x에 영향을 받음 -> 위 편미분의 가정 위배)



single-variable total-derivative chain rule

❖ total derivative $y = f(x) = x + x^2$

$$\frac{dy}{dx} = \frac{\partial f(x)}{\partial x} = \frac{\partial u_2(x, u_1)}{\partial x} = \frac{\partial u_2}{\partial x}\frac{\partial x}{\partial x} + \frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x} = \frac{\partial u_2}{\partial x} + \frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x}$$

$$\frac{dy}{dx} = \frac{\partial f(x)}{\partial x} = \frac{\partial u_2}{\partial x} + \frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x} = 1 + 1 \times 2x = 1 + 2x$$

$$\frac{\partial f(x, u_1, \ldots, u_n)}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x} + \frac{\partial f}{\partial u_2}\frac{\partial u_2}{\partial x} + \ldots + \frac{\partial f}{\partial u_n}\frac{\partial u_n}{\partial x} = \frac{\partial f}{\partial x} + \sum_{i=1}^{n}\frac{\partial f}{\partial u_i}\frac{\partial u_i}{\partial x}$$

single-variable total-derivative chain rule

- total derivative $f(x) = sin(x + x^2)$

$$\begin{aligned} u_1(x) &= x^2 \\ u_2(x, u_1) &= x + u_1 \\ u_3(u_2) &= sin(u_2) \qquad (y = f(x) = u_3(u_2)) \end{aligned}$$

and partials:

$$\begin{aligned} \frac{\partial u_1}{\partial x} &= 2x \\ \frac{\partial u_2}{\partial x} &= \frac{\partial x}{\partial x} + \frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x} = 1 + 1 \times 2x = 1 + 2x \\ \frac{\partial f(x)}{\partial x} &= \frac{\partial u_3}{\partial x} + \frac{\partial u_3}{\partial u_2}\frac{\partial u_2}{\partial x} = 0 + cos(u_2)\frac{\partial u_2}{\partial x} = cos(x + x^2)(1 + 2x) \end{aligned}$$

- 정리

$$\frac{\partial f(x, u_1, \ldots, u_n)}{\partial x} = \frac{\partial f}{\partial x} + \sum_{i=1}^{n} \frac{\partial f}{\partial u_i}\frac{\partial u_i}{\partial x} \quad \rightarrow \quad \frac{\partial f(u_1, \ldots, u_{n+1})}{\partial x} = \sum_{i=1}^{n+1} \frac{\partial f}{\partial u_i}\frac{\partial u_i}{\partial x}$$

$u_{n+1} = x.$

KOREATECH 한국기술교육대학교

vector chain rule

❖ vector function $\mathbf{y} = \mathbf{f}(x)$;

$$\begin{bmatrix} y_1(x) \\ y_2(x) \end{bmatrix} = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} ln(x^2) \\ sin(3x) \end{bmatrix}$$

$\mathbf{y} = \mathbf{f}(\mathbf{g}(x))$:

$$\begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} x^2 \\ 3x \end{bmatrix}$$

$$\begin{bmatrix} f_1(\mathbf{g}) \\ f_2(\mathbf{g}) \end{bmatrix} = \begin{bmatrix} ln(g_1) \\ sin(g_2) \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial f_1(\mathbf{g})}{\partial x} \\ \frac{\partial f_2(\mathbf{g})}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_1}{\partial g_2}\frac{\partial g_2}{\partial x} \\ \frac{\partial f_2}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_2}{\partial g_2}\frac{\partial g_2}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{1}{g_1}2x + 0 \\ 0 + cos(g_2)3 \end{bmatrix} = \begin{bmatrix} \frac{2x}{x^2} \\ 3cos(3x) \end{bmatrix} = \begin{bmatrix} \frac{2}{x} \\ 3cos(3x) \end{bmatrix}$$

vector chain rule

❖ vector function

$$\begin{bmatrix} \frac{\partial f_1}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_1}{\partial g_2}\frac{\partial g_2}{\partial x} \\ \frac{\partial f_2}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_2}{\partial g_2}\frac{\partial g_2}{\partial x} \end{bmatrix} \qquad \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \frac{\partial f_1}{\partial g_2} \\ \frac{\partial f_2}{\partial g_1} & \frac{\partial f_2}{\partial g_2} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial x} \\ \frac{\partial g_2}{\partial x} \end{bmatrix} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial x}$$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial x} = \begin{bmatrix} \frac{1}{g_1} & 0 \\ 0 & cos(g_2) \end{bmatrix} \begin{bmatrix} 2x \\ 3 \end{bmatrix} = \begin{bmatrix} \frac{1}{g_1}2x + 0 \\ 0 + cos(g_2)3 \end{bmatrix} = \begin{bmatrix} \frac{2}{x} \\ 3cos(3x) \end{bmatrix}$$

$$\frac{\partial}{\partial x}\mathbf{f}(\mathbf{g}(x)) = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial x} \qquad \frac{d}{dx}f(g(x)) = \frac{df}{dg}\frac{dg}{dx}$$

vector chain rule

❖ vector function

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{g}(\mathbf{x})) = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}$$

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{g}(\mathbf{x})) = \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \frac{\partial f_1}{\partial g_2} & \cdots & \frac{\partial f_1}{\partial g_k} \\ \frac{\partial f_2}{\partial g_1} & \frac{\partial f_2}{\partial g_2} & \cdots & \frac{\partial f_2}{\partial g_k} \\ & & \cdots & \\ \frac{\partial f_m}{\partial g_1} & \frac{\partial f_m}{\partial g_2} & \cdots & \frac{\partial f_m}{\partial g_k} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ & & \cdots & \\ \frac{\partial g_k}{\partial x_1} & \frac{\partial g_k}{\partial x_2} & \cdots & \frac{\partial g_k}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{g}(\mathbf{x})) = diag(\frac{\partial f_i}{\partial g_i})diag(\frac{\partial g_i}{\partial x_i}) = diag(\frac{\partial f_i}{\partial g_i}\frac{\partial g_i}{\partial x_i})$$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial x} = \begin{bmatrix} \frac{1}{g_1} & 0 \\ 0 & cos(g_2) \end{bmatrix} \begin{bmatrix} 2x \\ 3 \end{bmatrix}$$

KOREATECH 한국기술교육대학교

vector chain rule

❖ vector function $\mathbf{y} = \mathbf{f}(x)$,

$$\begin{bmatrix} y_1(x) \\ y_2(x) \end{bmatrix} = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} ln(x^2) \\ sin(3x) \end{bmatrix}$$

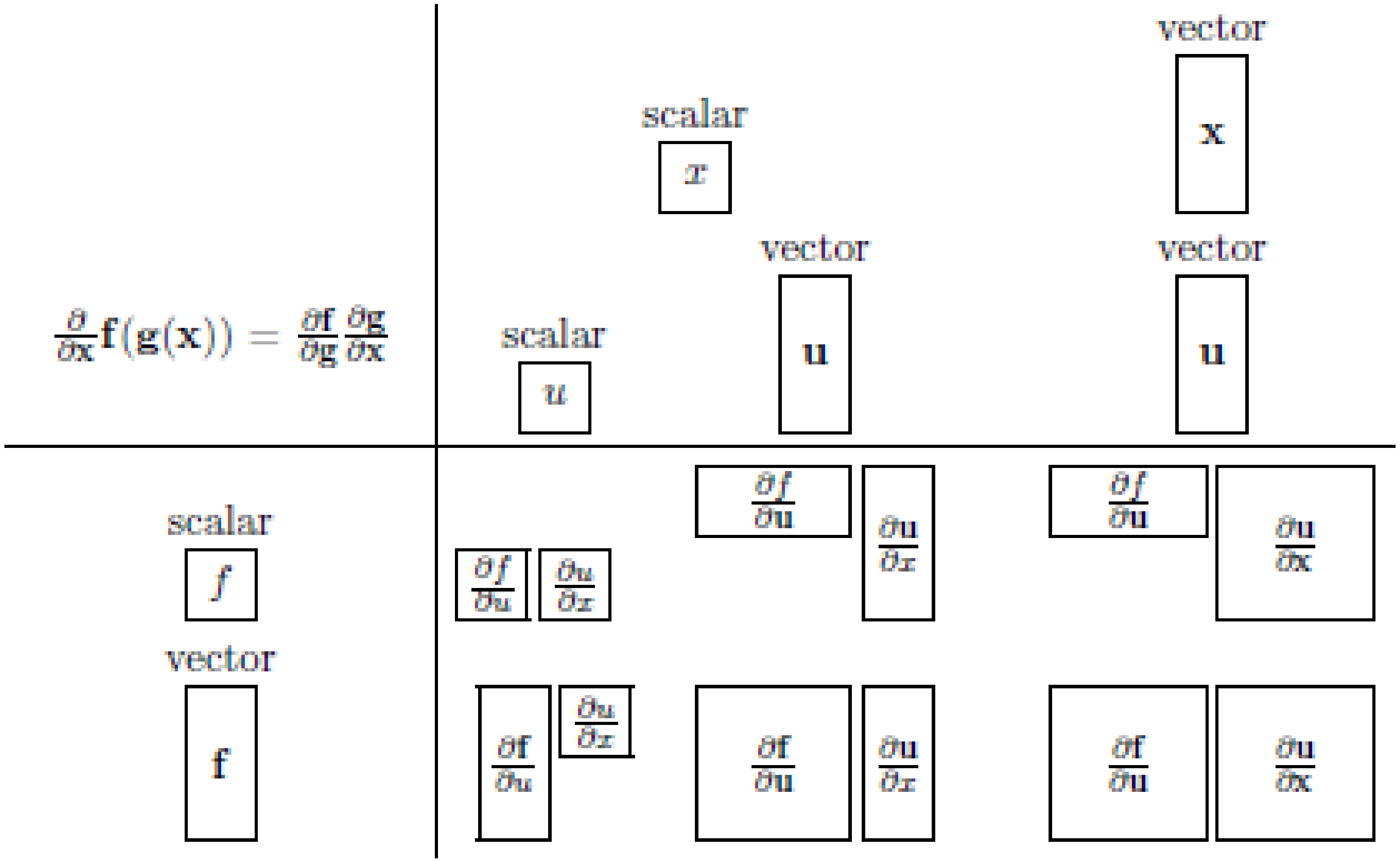
$\mathbf{y} = \mathbf{f}(\mathbf{g}(x))$:

$$\begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} x^2 \\ 3x \end{bmatrix}$$

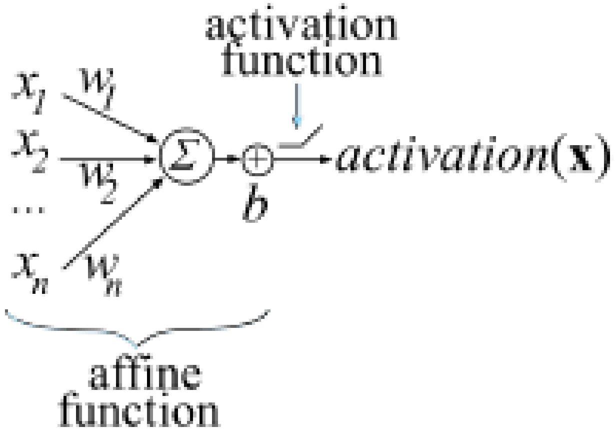
$$\begin{bmatrix} f_1(\mathbf{g}) \\ f_2(\mathbf{g}) \end{bmatrix} = \begin{bmatrix} ln(g_1) \\ sin(g_2) \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial f_1(\mathbf{g})}{\partial x} \\ \frac{\partial f_2(\mathbf{g})}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_1}{\partial g_2}\frac{\partial g_2}{\partial x} \\ \frac{\partial f_2}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_2}{\partial g_2}\frac{\partial g_2}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{1}{g_1}2x + 0 \\ 0 + cos(g_2)3 \end{bmatrix} = \begin{bmatrix} \frac{2x}{x^2} \\ 3cos(3x) \end{bmatrix} = \begin{bmatrix} \frac{2}{x} \\ 3cos(3x) \end{bmatrix}$$

vector chain rule



The gradient of neuron activation



$$activation(\mathbf{x}) = max(0, \mathbf{w} \cdot \mathbf{x} + b)$$

$$z(\mathbf{x}) = \sum_i^n w_i x_i + b = \mathbf{w} \cdot \mathbf{x} + b.$$

w와 b 값을 조정

$$d(u^T v) = du^T v + u^T dv = v^T du + u^T dv$$

$$d(x^T x) = dx^T x + x^T dx = x^T dx + x^T dx$$
$$= (2x)^T dx$$

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w} \cdot \mathbf{x} + b) \text{ and } \frac{\partial}{\partial b}(\mathbf{w} \cdot \mathbf{x} + b).$$

$$\mathbf{w} \cdot \mathbf{x} = \mathbf{w}^T \mathbf{x}.$$

The gradient of neuron activation

$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}\cdot\mathbf{x}+b)$ and $\frac{\partial}{\partial b}(\mathbf{w}\cdot\mathbf{x}+b)$. $\qquad$ $\mathbf{w}\cdot\mathbf{x}=\mathbf{w}^T\mathbf{x}$.

$$
\begin{aligned}
y &= \mathbf{w}\cdot\mathbf{x} &&= \sum_i^n (w_i x_i) \\
\frac{\partial y}{\partial w_j} &= \frac{\partial}{\partial w_j}\sum_i (w_i x_i) &&= \sum_i \frac{\partial}{\partial w_j}(w_i x_i) &&= \frac{\partial}{\partial w_j}(w_j x_j) &&= x_j
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial y}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}}\mathbf{w}\cdot\mathbf{x}+\frac{\partial}{\partial \mathbf{w}}b &&= \mathbf{x}^T+\vec{0}^T &&= \mathbf{x}^T \\
\frac{\partial y}{\partial b} &= \frac{\partial}{\partial b}\mathbf{w}\cdot\mathbf{x}+\frac{\partial}{\partial b}b &&= 0+1 &&= 1
\end{aligned}
$$

The gradient of neuron activation

$$z(\mathbf{w}, b, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

$$activation(z) = max(0, z) \qquad \frac{\partial activation}{\partial \mathbf{w}} = \frac{\partial activation}{\partial z}\frac{\partial z}{\partial \mathbf{w}}$$

$$\frac{\partial activation}{\partial \mathbf{w}} = \begin{cases} 0\frac{\partial z}{\partial \mathbf{w}} = \vec{0}^T & z \leq 0 \\ 1\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial z}{\partial \mathbf{w}} = \mathbf{x}^T & z > 0 \quad (\text{we computed } \frac{\partial z}{\partial \mathbf{w}} = \mathbf{x}^T \text{ previously}) \end{cases}$$

and then substitute $z = \mathbf{w} \cdot \mathbf{x} + b$ back in:

$$\frac{\partial activation}{\partial \mathbf{w}} = \begin{cases} \vec{0}^T & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ \mathbf{x}^T & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

$$\frac{\partial activation}{\partial b} = \begin{cases} 0\frac{\partial z}{\partial b} = 0 & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ 1\frac{\partial z}{\partial b} = 1 & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

The gradient of neural network loss function

multiple vector inputs (multiple images) $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$

scalar targets (분류결과)

$$\mathbf{y} = [target(\mathbf{x}_1), target(\mathbf{x}_2), \ldots, target(\mathbf{x}_N)]^T$$

비용함수 -> 원하는 결과(target)과 신경만의 결과(activation)의 차이

$$C(\mathbf{w}, b, X, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - activation(\mathbf{x}_i))^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - max(0, \mathbf{w}\cdot\mathbf{x}_i + b))^2$$

비용함수가 최소가 되게 w와 b를 조정해야됨 -> 미분을 수행 -> gradient를 구해서 w를 조정하는 과정

$$\begin{aligned} u(\mathbf{w}, b, \mathbf{x}) &= max(0, \mathbf{w}\cdot\mathbf{x} + b) \\ v(y, u) &= y - u \\ C(v) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} v^2 \end{aligned}$$

KOREATECH 한국기술교육대학교

The gradient with respect to the weights

neuron의 결과

$$
\begin{aligned}
u(\mathbf{w}, b, \mathbf{x}) &= max(0, \mathbf{w}\cdot\mathbf{x} + b) \\
v(y, u) &= y - u \\
C(v) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} v^2
\end{aligned}
$$

$$
\frac{\partial}{\partial \mathbf{w}} u(\mathbf{w}, b, \mathbf{x}) = \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x} + b \le 0 \\ \mathbf{x}^T & \mathbf{w}\cdot\mathbf{x} + b > 0 \end{cases}
$$

$$
\frac{\partial v(y,u)}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}(y - u) = \vec{0}^T - \frac{\partial u}{\partial \mathbf{w}} = -\frac{\partial u}{\partial \mathbf{w}} = \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x} + b \le 0 \\ -\mathbf{x}^T & \mathbf{w}\cdot\mathbf{x} + b > 0 \end{cases}
$$

$$
\begin{aligned}
\frac{\partial C(v)}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} v^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \mathbf{w}} v^2 \qquad = \frac{1}{N} \sum_{i=1}^{N} 2v \frac{\partial v}{\partial \mathbf{w}} \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial v^2}{\partial v} \frac{\partial v}{\partial \mathbf{w}} \qquad = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} 2v\vec{0}^T = \vec{0}^T & \mathbf{w}\cdot\mathbf{x}_i + b \le 0 \\ -2v\mathbf{x}^T & \mathbf{w}\cdot\mathbf{x}_i + b > 0 \end{cases}
\end{aligned}
$$

The gradient with respect to the weights

neuron의 결과

$$\begin{aligned} u(\mathbf{w}, b, \mathbf{x}) &= max(0, \mathbf{w}\cdot\mathbf{x}+b) \\ v(y, u) &= y - u \\ C(v) &= \tfrac{1}{N}\sum_{i=1}^{N} v^2 \end{aligned}$$

$$\frac{\partial}{\partial \mathbf{w}} u(\mathbf{w}, b, \mathbf{x}) = \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}+b \le 0 \\ \mathbf{x}^T & \mathbf{w}\cdot\mathbf{x}+b > 0 \end{cases}$$

$$\frac{\partial v(y,u)}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}(y-u) = \vec{0}^T - \frac{\partial u}{\partial \mathbf{w}} = -\frac{\partial u}{\partial \mathbf{w}} = \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}+b \le 0 \\ -\mathbf{x}^T & \mathbf{w}\cdot\mathbf{x}+b > 0 \end{cases}$$

$$\begin{aligned} \frac{\partial C(v)}{\partial \mathbf{w}} &= \frac{1}{N}\sum_{i=1}^{N} \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}_i+b \le 0 \\ -2(y_i-u)\mathbf{x}_i^T & \mathbf{w}\cdot\mathbf{x}_i+b > 0 \end{cases} \\ &= \frac{1}{N}\sum_{i=1}^{N} \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}_i+b \le 0 \\ -2(y_i-max(0,\mathbf{w}\cdot\mathbf{x}_i+b))\mathbf{x}_i^T & \mathbf{w}\cdot\mathbf{x}_i+b > 0 \end{cases} \quad = \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}_i+b \le 0 \\ \frac{-2}{N}\sum_{i=1}^{N}(y_i-(\mathbf{w}\cdot\mathbf{x}_i+b))\mathbf{x}_i^T & \mathbf{w}\cdot\mathbf{x}_i+b > 0 \end{cases} \\ &= \frac{1}{N}\sum_{i=1}^{N} \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}_i+b \le 0 \\ -2(y_i-(\mathbf{w}\cdot\mathbf{x}_i+b))\mathbf{x}_i^T & \mathbf{w}\cdot\mathbf{x}_i+b > 0 \end{cases} \quad = \begin{cases} \vec{0}^T & \mathbf{w}\cdot\mathbf{x}_i+b \le 0 \\ \frac{2}{N}\sum_{i=1}^{N}(\mathbf{w}\cdot\mathbf{x}_i+b-y_i)\mathbf{x}_i^T & \mathbf{w}\cdot\mathbf{x}_i+b > 0 \end{cases} \end{aligned}$$

The gradient with respect to the weights

$$\frac{\partial C}{\partial \mathbf{w}} = \frac{2}{N}\sum_{i=1}^{N} e_i \mathbf{x}_i^T \quad \text{(for the nonzero activation case)}$$

If the error is 0, then the gradient is zero and we have arrived at the minimum loss.
If error is some small positive difference, the gradient is a small step in the directionof x.
If error is large, the gradient is a large step in that direction.
If error is negative, the gradient is reversed, meaning the highest cost is in the negative direction.

결과값과 올바른값의 차이를 최소화가 되도록 w 값을 재조정
학습율(이타,eta) -> 경험적으로 조정

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial C}{\partial \mathbf{w}}$$

The gradient with respect to the bias

$$\frac{\partial u}{\partial b} = \begin{cases} 0 & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ 1 & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

$$\begin{aligned} u(\mathbf{w}, b, \mathbf{x}) &= max(0, \mathbf{w} \cdot \mathbf{x} + b) \\ v(y, u) &= y - u \\ C(v) &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} v^2 \end{aligned}$$

$$\frac{\partial v(y,u)}{\partial b} = \frac{\partial}{\partial b}(y - u) = 0 - \frac{\partial u}{\partial b} = -\frac{\partial u}{\partial b} = \begin{cases} 0 & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ -1 & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

$$\begin{aligned} \frac{\partial C(v)}{\partial b} &= \frac{\partial}{\partial b} \frac{1}{N} \sum_{i=1}^{N} v^2 \quad = \begin{cases} 0 & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \\ \frac{2}{N} \sum_{i=1}^{N} (\mathbf{w} \cdot \mathbf{x}_i + b - y_i) & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases} \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial b} v^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial v^2}{\partial v} \frac{\partial v}{\partial b} \end{aligned}$$

$$b_{t+1} = b_t - \eta \frac{\partial C}{\partial b}$$

KOREATECH 한국기술교육대학교