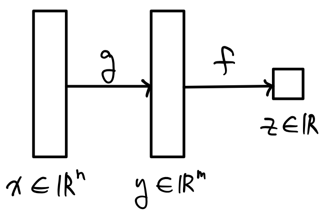
Chain rule and linear algebra

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad : \text{ column vector}$$

$$x = [x_1, x_2, \cdots x_n] \quad : \text{ row vector}$$



m×1
column vector

① See gradient as column vector

$$\underset{n\times 1}{\frac{dz}{dx}} = \underset{n\times m}{\frac{dy}{dx}} \underset{m\times 1}{\frac{dz}{dy}}$$

$$\left\{ \begin{array}{l} \dfrac{dz}{dy} = [dz/dy_1, dz/dy_2, \cdots dz/dy_m]^T \\[2ex] \dfrac{dy}{dx} = \begin{pmatrix} dy_1/dx_1, & dy_2/dx_1, & \cdots, & dy_m/dx_1 \\ dy_1/dx_2, & dy_2/dx_2, & \cdots, & dy_m/dx_2 \\ \vdots & & & \\ dy_1/dx_n, & dy_2/dx_n, & \cdots, & dy_m/dx_n \end{pmatrix} \end{array} \right.$$

e.g.) $\underset{1\times 1}{z} = \underset{1\times m}{C^T} \underset{m\times 1}{y} = C^T \underset{m\times n}{M} \underset{n\times 1}{x}$ $\qquad \underset{n\times 1}{\frac{dz}{dx}} = \frac{dy}{dx}\frac{dz}{dy} = \underset{n\times m}{M^T} \underset{m\times 1}{C}$

transpose

② See gradient as row vector

$$\underset{1\times n}{\frac{dz}{dx}} = \underset{1\times m}{\frac{dz}{dy}} \underset{m\times n}{\frac{dy}{dx}}$$

same

e.g.) $\underset{1\times 1}{z} = \underset{1\times m}{C^T} \underset{m\times 1}{y} = C^T \underset{m\times n}{M} \underset{n\times 1}{x}$ $\qquad \underset{1\times n}{\frac{dz}{dx}} = \underset{1\times m}{\frac{dz}{dy}} \underset{m\times n}{\frac{dy}{dx}} = C^T M$

Backpropagation

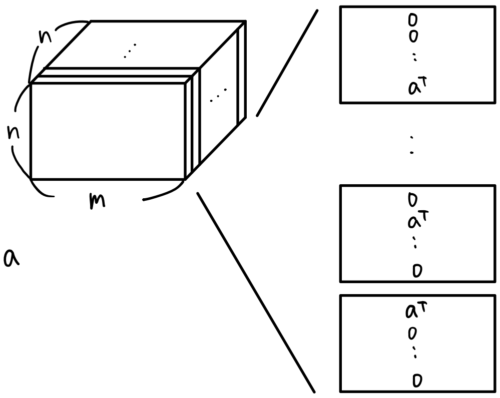
Linear layer

$$\underset{n\times m}{\frac{dL}{dW}} = \underset{n\times m\times n}{\frac{dz}{dW}}\ \underset{n\times 1}{\frac{dL}{dz}} = \sum_{i=1}^{n} \frac{dz_i}{dW}\frac{dL}{dz_i} \qquad \begin{bmatrix} a^T \cdot dL/dz_1 \\ a^T \cdot dL/dz_2 \\ \vdots \\ a^T \cdot dL/dz_n \end{bmatrix}$$

$n \times m \times 1$

$n \times m$

$$\underset{n\times 1}{z} = \underset{n\times m}{W}\ \underset{m\times 1}{a} + \underset{n\times 1}{b} \qquad \frac{dz}{dW} :$$



$$\underset{1\times 1}{z_i} = \underset{1\times m}{W_i}\ \underset{m\times 1}{a} + \underset{1\times 1}{b_i} \qquad \underset{m\times 1}{\frac{dz_i}{dW_i}} = a$$

Sigmoid

$$\underset{n\times 1}{\frac{dL}{dz}} = \underset{n\times n}{\frac{d\sigma(z)}{dz}}\ \underset{n\times 1}{\frac{dL}{d\sigma(z)}}$$

$$\sigma(z) = [\sigma(z_1), \sigma(z_2), \cdots, \sigma(z_n)]^T \qquad \frac{d\sigma(z)}{dz} = \begin{bmatrix} d\sigma(z_1)/dz_1, & d\sigma(z_2)/dz_1, & \cdots, & d\sigma(z_n)/dz_1 \\ d\sigma(z_1)/dz_2, & d\sigma(z_2)/dz_2, & \cdots, & d\sigma(z_n)/dz_2 \\ & \vdots & & \\ d\sigma(z_1)/dz_n, & d\sigma(z_2)/dz_n, & \cdots, & d\sigma(z_n)/dz_n \end{bmatrix}$$

$$= \begin{bmatrix} d\sigma(z_1)/dz_1, & 0, & \cdots, & 0 \\ 0, & d\sigma(z_2)/dz_2, & \cdots, & 0 \\ & \vdots & & \\ 0, & 0, & \cdots, & d\sigma(z_n)/dz_n \end{bmatrix}$$

$$\sigma(z_i) = \frac{1}{1+\exp(-z_i)} \qquad \frac{d\sigma(z_i)}{dz_i} = (1-\sigma(z_i))\sigma(z_i)$$

ReLU

$$f(z) = \max(0, z) \qquad \frac{df(z)}{dz} = \begin{bmatrix} \mathrm{Ind}(z_1 \geq 0), & 0, & \cdots, & 0 \\ 0, & \mathrm{Ind}(z_2 \geq 0), & \cdots, & 0 \\ & \vdots & & \\ 0, & 0, & \cdots, & \mathrm{Ind}(z_n \geq 0) \end{bmatrix}$$

$$f(z_i) = \max(0, z_i) \qquad \frac{df(z_i)}{dz_i} = \mathrm{Ind}(z_i \geq 0)$$