2019009261_최가온 🧑🏻‍💻

Problem 1

Represent the derivative of the following scalar functions with respect to $\mathbf{X} \in \mathbb{R}^{D\times D}$.

(a) $f(\mathbf{X}) = tr(\mathbf{X}^2)$. Here, $tr(A)$ is the trace of a square matrix A.

(b) $g(\mathbf{X}) = tr(\mathbf{X}^3)$.

(c) $h(\mathbf{X}) = tr(\mathbf{X}^k)$ for $k \in \mathbb{N}$.

(a) $f(X) = tr(X^2) = \sum_{i=1}^{D}\sum_{j=1}^{D} X_{ij}X_{ji}$

$$\frac{df(X)}{dX} = \delta_{kl}^{ij} X_{ji} + X_{ij}\boxed{\delta_{kl}^{ji}}$$

$l=i$ && $k=j$인 항만 살아남음

$$= X_{lk} + X_{lk} = 2X_{lk} = 2(X_{kl})^T = 2X^T$$

(b) $g(X) = tr(X^3) = \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} X_{ij}X_{jk}X_{ki}$

$$\frac{dg(X)}{dX} = \delta_{lm}^{ij}\sum_{k=1}^{D} X_{jk}X_{ki} + \delta_{lm}^{jk}\sum_{i=1}^{D} X_{ij}X_{ki} + \delta_{lm}^{ki}\sum_{i=1}^{D} X_{ij}X_{jk}$$

$$= \sum_{k=1}^{D} X_{mk}X_{kl} + \sum_{i=1}^{D} X_{il}X_{mi} + \sum_{j=1}^{D} X_{mj}X_{jl}$$

$$= \sum_{k=1}^{D}(X_{lk})^T\cdot(X_{km})^T + \sum_{i=1}^{D}(X_{li})^T\cdot(X_{im})^T + \sum_{j=1}^{D}(X_{lj})^T\cdot(X_{jm})^T$$

$$= (X^2)^T + (X^2)^T + (X^2)^T = 3(X^2)^T$$

(c) $h(X) = tr(X^k)$

$$tr[X^k] = \sum_{a_1=1}^{D}\sum_{a_2=1}^{D}\cdots\sum_{a_k=1}^{D}(\overbrace{X_{a_1a_2}X_{a_2a_3}X_{a_3a_4}\cdots X_{a_ka_1}}^{k개})$$

$$\frac{dh(X)}{dX} = \underbrace{\sum_{a_3=1}^{D}\cdots\sum_{a_k=1}^{D} X^{k-1} + \sum_{a_1=1}^{D}\sum_{a_4=1}^{D}\cdots\sum_{a_k=1}^{D} X^{k-1} + \cdots}_{k개}$$

$$= k(X^{k-1})^T$$

2019009261_최가온 🧑🏻‍💻

Problem 2

In order to alleviate overfitting in logistic regression, regularization technique can be used with the L_2-norm of the weight parameter, $||\mathbf{w}||^2 = \sum_{i=1}^{D} w_i^2$. When we are given $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ and $y_1, \ldots, y_N \in \{0, 1\}$ for training set, we want to derive the update rule for $\mathbf{w} \in \mathbb{R}^D$ in order to minimize the following loss function $L(\mathbf{w})$ having L_2-norm,

$$L(\mathbf{w}) = \sum_{i=1}^{N} \Big(-y_i \ln f(\mathbf{x}_i; \mathbf{w}) - (1 - y_i) \ln(1 - f(\mathbf{x}_i; \mathbf{w})) \Big) + \boxed{||\mathbf{w}||^2.} \quad (1)$$

L2-norm

$$f(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})} \quad (2)$$

(1) Derive the update rule for $\mathbf{w}$ when we use gradient descent.

(2) Discuss the effect of L_2-norm regularization.

1. $L(w) = \sum_{i=1}^{N} \Big(-y_i \ln(f(x_i; w)) - (1 - y_i) \ln(1 - f(x_i; w)) \Big) + ||w||^2$

$= -y \ln f - (1-y) \ln(1-f) + ww^T$

$\left[\frac{d}{dw}(w^T w)\right]_i = \frac{d}{dw_i}\left(\sum w_i^2\right) = \frac{d}{dw_i}\left(w_1^2 + w_2^2 + \cdots + w_i^2 + \cdots + w_D^2\right)$

$= \frac{d}{dw_i}(w_i^2) = 2w_i \qquad \therefore \frac{d}{dw}(w^T w) = 2w$ — A

$\frac{df}{dw} = \frac{-\exp(-w^T x)(-x)}{(1 + \exp(-w^T x))^2} = \frac{x \cdot \exp(-w^T x)}{(1 + \exp(-w^T x))^2}$

$= \frac{1}{1 + \exp(-w^T x)} \cdot \frac{\exp(-w^T x)}{1 + \exp(-w^T x)} \cdot x$

$= f(1-f) \cdot x$ — B

2019009261_최가온 👨🏻‍💻

$$\frac{d}{dw}L(w) = -y \cdot \frac{1}{f} \cdot \frac{df}{dw} - (1-y) \cdot \frac{1}{1-f}\left(-\frac{df}{dw}\right) + 2w - ⓒ$$

Ⓐ와 Ⓑ를 ⓒ에 대입하면,

$$\frac{d}{dw}L(w) = \sum_{i=1}^{N}\Big[-y_i \cdot \frac{1}{\cancel{f(x_i;w)}} \cdot \cancel{f(x_i;w)} \cdot (1-f(x_i;w)) \cdot x_i$$

$$+ (1-y_i) \cdot \frac{1}{\cancel{1-f(x_i;w)}} \cdot f(x_i;w)\,\cancel{(f(x_i;w)-1)} \cdot x_i\Big] + 2w$$

$$= \sum_{i=1}^{N}\left(-y_i(1-f(x_i;w)) \cdot x_i + (1-y_i) \cdot f(x_i;w) \cdot x_i\right) + 2w$$

$$\therefore\ w_{t+1} = w_t - \eta \cdot \frac{dL}{dw}$$

$$= w_t - \eta \cdot \left[\sum_{i=1}^{N}(-y(1-f)x + (1-y)f \cdot x) + 2w\right]$$

2019009261_최가온 👨🏻‍💻

(2) 개별적으로 regularization을 한다는 것의 의미는 weight가 작아지도록 학습을 하게 만드는 것을 의미한다.

L_1 loss $= \sum |w|$

L_2 loss $= \sqrt{\sum |w|^2}$

ex) $a = \{0.3, -0.3, 0.4\}$, $b = \{0.5, -0.5, 0\}$

i) L_1 loss $\sum |a| = 0.3 + 0.3 + 0.4 = 1$

$\sum |b| = 0.5 + 0.5 + 0 = 1$

ii) L_2 loss : $\sum |a|^2 = 0.09 + 0.09 + 0.16 = 0.34$

$\sum |b|^2 = 0.25 + 0.25 + 0 = 0.50$

반대로, weight가 너무 큰 값을 가지면 model이 학습한 data를 외운다. (→ 일반화성능 저하)

L2 norm regularization은 모델의 학습에서 local minimum 의 영향을 덜 받게 (특이점의 영향↓) 학습을 하여 일반화 성능을 높인다고 할 수 있다.

2019009261_최가온

Problem 3

Consider the following function,

$$f(\mathbf{x};\mathbf{w}) = \frac{1-\exp\left(-\mathbf{w}^T\mathbf{x}\right)}{1+\exp\left(-\mathbf{w}^T\mathbf{x}\right)}, \tag{3}$$

having the shape in Fig. 1.

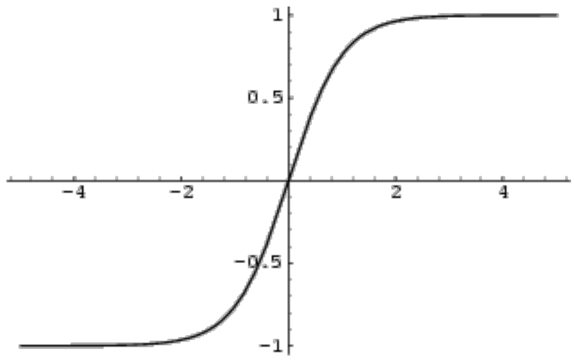


Figure 1: $f(\mathbf{x};\mathbf{w})$

Find the gradient descent update rule for $\mathbf{w}$ to minimize the loss

$$L = \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i;\mathbf{w}) - y_i)^2. \tag{4}$$

Here, we use N number of $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1, 1\}$ for $i \in \{1, \ldots, N\}$ which are given in advance.

$$f(x;w) = \frac{1-\exp(-w^T x)}{1+\exp(-w^T x)}$$

$$L = \frac{1}{2}\sum_{i=1}^{N}\left(f(x_i;w) - y_i\right)^2$$

$$\frac{dL}{dw} = \sum_{i=1}^{N}\left(f(x_i;w) - y_i\right)\cdot\frac{df(x_i;w)}{dw} \quad — \; ①$$

$$\frac{df(x_i;w)}{dw} = \frac{d}{dw}\left(\frac{1-\exp(-w^T x)}{1+\exp(-w^T x)}\right)$$

$$= \frac{\{-\exp(-w^T x)\cdot(-x)\}\{1+\exp(-w^T x)\} - \{1-\exp(-w^T x)\}\cdot\{\exp(-w^T x)\cdot(-x)\}}{\{1+\exp(-w^T x)\}^2}$$

$$\left\{\frac{f(x)}{g(x)}\right\}' = \frac{f'(x)g(x) - f(x)g'(x)}{\{g(x)\}^2}$$

2019009261_최가온 👨🏻‍💻

$$= \frac{x\cdot\exp(-w^Tx) + x\cdot\{\exp(-w^Tx)\}^2 + x\cdot\exp(-w^Tx) - x\{\exp(-w^Tx)\}^2}{\{1+\exp(-w^Tx)\}^2}$$

$$= \frac{2x\cdot\exp(-w^Tx)}{\{1+\exp(-w^Tx)\}^2} \quad — ②$$

② → ① 을 하면,

$$\frac{dL}{dw} = \left[\sum_{i=1}^{N}\left\{\frac{1-\exp(-w^Tx_i)}{1+\exp(-w^Tx_i)} - y_i\right\}\right]\left[\frac{2x\cdot\exp(-w^Tx)}{\{1+\exp(-w^Tx)\}^2}\right]$$

$$= \sum_{i=1}^{N}(f-y_i)\{2x_i\cdot\sigma(1-\sigma)\}$$

$$\therefore w_{t+1} = w_t - \eta\cdot\frac{dL}{dw}$$

$$= w_t - \eta\cdot\left[\sum_{i=1}^{N}(f-y_i)\{2x_i\cdot\sigma(1-\sigma)\}\right]$$

$$\left(\text{단, } \sigma = \frac{1}{1+\exp(-w^Tx)}\right)$$

$$f = \frac{1-\exp(-w^Tx)}{1+\exp(-w^Tx)}$$

Problem 4

For a given data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, derive the closed-form solution for $\mathbf{w}$ that maximizes the following probability $\boldsymbol{P}$.

$$\boldsymbol{P} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - f(\mathbf{x}_i; \mathbf{w}))^2\right), \tag{5}$$

with $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x}$ for $\mathbf{x}, \mathbf{w} \in \mathbb{R}^D$.

exponential 부분이 포함된 항이 $\prod$을 통해 곱의 합으로 전개되는 형태의 식이다. 이에 따라, 지수를 좀 더 간단하게 다루기 위해 양변에 로그를 취하는 방식을 고려하였다.

이때 로그함수는 monotonically increasing의 특성을 지니고 있다. 따라서, 임의의 함수 $f(x)$에 대하여

$$\boxed{\underset{x}{\operatorname{argmin}} f(x) = \underset{x}{\operatorname{argmin}} \ln[f(x)]}$$

가 성립한다.

$$\ln P = -\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \left(\frac{1}{2\sigma^2} (y_i - f(x_i; w))^2 \right)$$

위의 식에서, $\frac{1}{\sqrt{2\pi\sigma^2}}$ 부분을 시그마 앞으로 넘기고, y_i의 index를 생략하여 y로 표현한다.

$$\ln P = -\frac{1}{\sqrt{2\pi\sigma^2}} \left(\frac{1}{2\sigma^2} (y - w^T x)^2 \right)$$

2019009261_최가온 👨🏻‍💻

$$\frac{d\ln P}{dw} = -\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{2\sigma^2} \cdot \frac{d}{dw}(y-w^Tx)^2$$

$$= -\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{2\sigma^2} \cdot 2(y-w^Tx) \cdot \left[\frac{d}{dw}(y-w^Tx)\right]$$

$$= -\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{2\sigma^2} \cdot 2(y-w^Tx) \cdot (-x)$$

$$= \underbrace{\boxed{\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{2\sigma^2}}}_{\text{constant}} \cdot 2(y-w^Tx) \cdot x = 0$$

$$(y-w^Tx)x = 0 \rightarrow w^Tx^2 = yx$$

$$w^Txx = yx \rightarrow w^T = yxx^{-1}x^{-1} = yIx^{-1} = yx^{-1}$$

$$\therefore w^T = yx^{-1}$$

$$w = (yx^{-1})^T = (x^{-1})^T y^T$$

이때, $x \in \mathbb{R}^D$: D by 1 $\quad w \in \mathbb{R}^D$: D by 1

$$\therefore w^Tx : (1 \text{ by } D) \times (D \text{ by } 1) = \text{scalar}$$

$$\therefore \boxed{w = (x^{-1})^T y^T}$$

2019009261_최가온 👨🏻‍💻

Problem 5

The Frobenius dot product $\langle \mathbf{A}, \mathbf{B} \rangle$ is defined as

$$\langle \mathbf{A}, \mathbf{B} \rangle = tr(\mathbf{A}^\top \mathbf{B}), \tag{6}$$

for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ using the scalar function $tr(\cdot)$ for the trace of a matrix,

$$tr(\mathbf{M}) = \sum_i^N M_{ii}, \quad \text{for } \mathbf{M} \in \mathbb{R}^{N \times N}. \tag{7}$$

Find the derivative of $\langle \mathbf{A}, \mathbf{B} \rangle$ with respect to $\mathbf{A}$ using the definition of the matrix derivative:

$$\left[\frac{d}{d\mathbf{A}} \langle \mathbf{A}, \mathbf{B} \rangle \right]_{ij} = \frac{\partial}{\partial A_{ij}} \langle \mathbf{A}, \mathbf{B} \rangle. \tag{8}$$

$<A, B> = tr(A^T B)$: Frobenius dot product

$$\left[\frac{d}{dA} <A, B> \right]_{ij} = \frac{\partial}{\partial A_{ij}} <A, B> = \frac{\partial}{\partial A_{ij}} tr(A^T B)$$

각각의 부분들을 element-wise notation으로 바꾼다.

$$[AB]_{pq} = \sum_{k=1}^{D} A_{pk} B_{kq}$$

$$[A^T B]_{pq} = \sum_{k=1}^{D} A_{kp} B_{kq}$$

$$tr[A^T B] = \sum_{l=1}^{D} \sum_{k=1}^{D} A_{kl} B_{kl}$$

$$\left[\frac{d}{dA} <A, B> \right]_{pq} = \frac{\partial}{\partial A_{pq}} <A, B>$$

$$= \frac{\partial}{\partial A_{pq}} \sum_{l=1}^{D} \sum_{k=1}^{D} a_{kl} b_{kl} = b_{pq}$$

$$\therefore \frac{d}{dA} <A, B> = B$$

2019009261_최가온 🧑🏻‍💻

아래의 상황을 가정

$$A : \begin{array}{|c|c|} \hline a_{11} & a_{12} \\ \hline a_{21} & a_{22} \\ \hline a_{31} & a_{32} \\ \hline \end{array} \qquad B : \begin{array}{|c|c|} \hline b_{11} & b_{12} \\ \hline b_{21} & b_{22} \\ \hline b_{31} & b_{32} \\ \hline \end{array}$$

$$\langle A, B \rangle = tr(A^T B)$$

$$A^T B = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

: $a_{11}b_{11} + a_{21}b_{21} + a_{31}b_{31}$

: $a_{12}b_{12} + a_{22}b_{22} + a_{32}b_{32}$

) 이 둘을 더하면 $tr(A^T B) = \langle A, B \rangle$

따라서, Frobenius inner product는 matrix A, B의 element-wise multiplication의 합이라고 정성적으로 해석할 수 있다.

만약 특수한 상황으로 $\langle A, A \rangle = tr(A^T A)$를 가정하면 이 값은 Frobenius norm으로 각각의 element를 제곱한 것의 합을 의미한다.

Problem 6

We are given a scalar function $f(\mathbf{X}) = \mathbf{w}^\top \mathbf{X} \mathbf{w}$ for $\mathbf{w} \in \mathbb{R}^D, \mathbf{X} \in \mathbb{R}^{D\times D}$. Find the derivative of $f(\mathbf{X})$ with respect to the vector $\mathbf{w}$. Use the definition of the vector derivative $\left[\frac{df}{d\mathbf{w}}\right]_k = \frac{\partial f}{\partial w_k}$, where w_k is the k-th element of $\mathbf{w}$.

$$f(X) = w^T X w \qquad w \in R^D, \quad X \in R^{D\times D}$$

$\hookrightarrow (1 \text{ by } D) \times (D \text{ by } D) \times (D \text{ by } 1)$

$= 1 \text{ by } 1 \Rightarrow$ a scalar

element-wise notation을 이용하여 표현하면 아래와 같다.

$$f(X) = \begin{bmatrix} w_1 & w_2 & \cdots & w_D \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{D1} & x_{D2} & & x_{DD} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{D} w_i x_{i1} & \sum_{i=1}^{D} w_i x_{i2} & \cdots & \sum_{i=1}^{D} w_i x_{iD} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

$$= w_1 \sum_{i=1}^{D} w_i x_{i1} + w_2 \sum_{i=1}^{D} w_i x_{i2} + \cdots + w_D \sum_{i=1}^{D} w_i x_{iD}$$

$$= \sum_{i=1}^{D} w_1 w_i x_{i1} + \sum_{i=1}^{D} w_2 w_i x_{i2} + \cdots + \sum_{i=1}^{D} w_D w_i x_{iD}$$

$$= \sum_{j=1}^{D} \sum_{i=1}^{D} w_j x_{ij} w_i$$

2019009261_최가온 👨🏻‍💻

$$\left[\frac{df}{dw}\right]_k = \frac{\partial f}{\partial w_k} = \delta^i_k \sum_{j=1}^{D} w_j x_{ij} + \delta^j_k \sum_{i=1}^{D} w_i x_{ij}$$

(미분 시, 해당 element 외에는 모두 소거됨)

(δ^a_b는 $a=b$인 항만 남게 함)

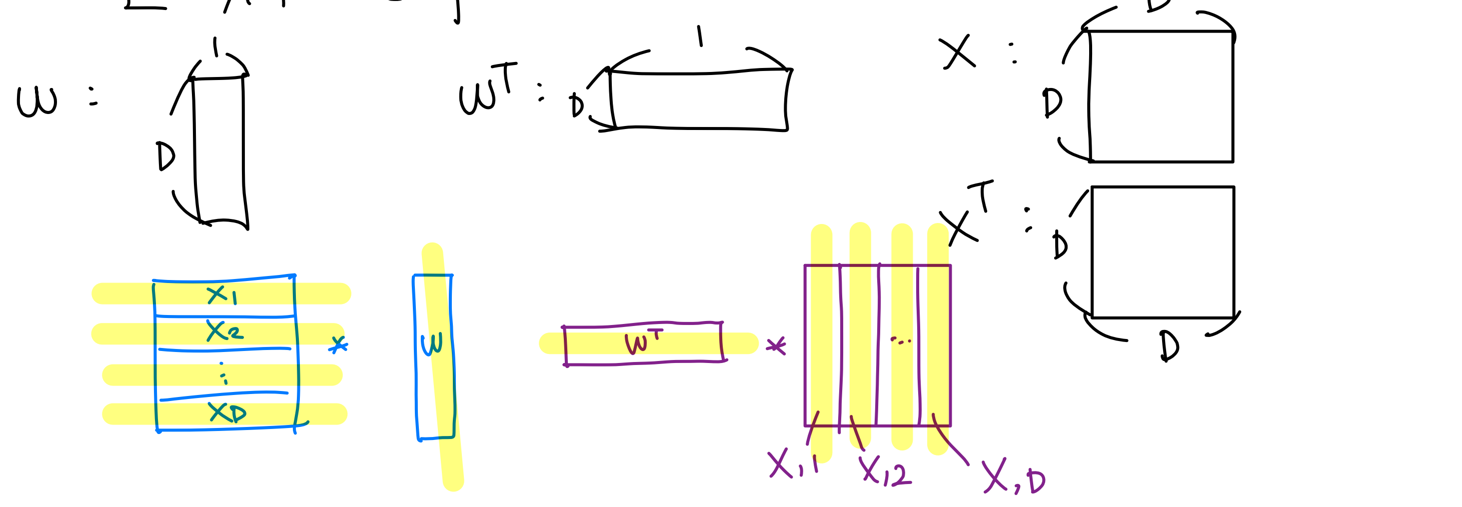
($\delta^a_b = 1$ if $a=b$)

$$= \sum_{j=1}^{D} x_{kj} w_j + \sum_{i=1}^{D} w_i x_{ik}$$

$$\frac{df}{dw} = \begin{bmatrix} \sum_{j=1}^{D} x_{1j} w_j + \sum_{i=1}^{D} w_i x_{i1} \\ \sum_{j=1}^{D} x_{2j} w_j + \sum_{i=1}^{D} w_i x_{i2} \\ \vdots \\ \sum_{j=1}^{D} x_{Dj} w_j + \sum_{i=1}^{D} w_i x_{iD} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{D} x_{1j} w_j \\ \sum_{j=1}^{D} x_{2j} w_j \\ \vdots \\ \sum_{j=1}^{D} x_{Dj} w_j \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{D} w_i x_{i1} \\ \sum_{i=1}^{D} w_i x_{i2} \\ \vdots \\ \sum_{i=1}^{D} w_i x_{iD} \end{bmatrix}$$

$$= ①\begin{bmatrix} w^T x_1 \\ w^T x_2 \\ \vdots \\ w^T x_D \end{bmatrix} + ②\begin{bmatrix} w x_{,1} \\ w x_{,2} \\ \vdots \\ w x_{,D} \end{bmatrix} = Xw + w^T X$$

↑ 위의 shape, element가 만들어지는 과정을 시각화



2019009261_최가온 👨🏻‍💻

Problem 7

For a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$, Suppose that we have a two-layer neural network with residual connections shown in Figure 2. Each component is given as follows:

$$L(W) = \frac{1}{2}\sum_{i=1}^{N}(g(\mathbf{x}_i) - y_i)^2 \tag{9}$$

$$g(\mathbf{x}) = \sigma\left(\sum_{d=1}^{D} w_{2,d} \cdot h_d(\mathbf{x})\right) \tag{10}$$

$$h_i(\mathbf{x}) = x_i + \sigma\left(\sum_{m=1}^{M} w_{1,i,m} \cdot z_m(\mathbf{x})\right) \tag{11}$$

$$z_i(\mathbf{x}) = \sigma\left(\sum_{d=1}^{D} w_{0,i,d} \cdot x_d\right) \tag{12}$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{13}$$

Here, the residual connection from each x_i to the h_i node in the second layer is represented in Eq. (11). The weight $w_{i,j,k}$ connects the j-th node of i-th layer to the k-th node of $(i+1)$-th layer.

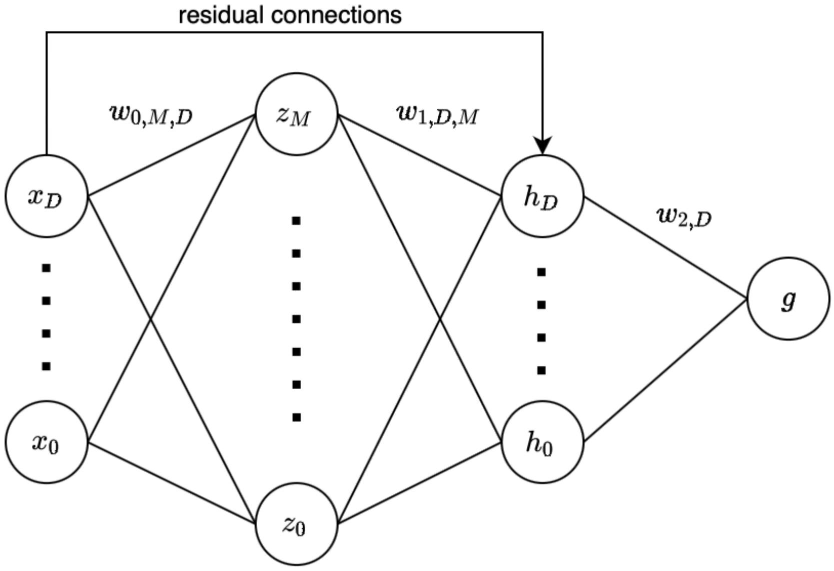


Figure 2: A two-layer network

(a) Calculate derivative of L with respect to h_j.

(b) Calculate derivative of L with respect to $w_{1,d,m}$.

먼저 각각을 matrix format으로 바꾸고

(1), (2)를 풀 것이다.

2019009261_최가온 👨🏻‍💻

$$L(W) = \frac{1}{2}\sum_{i=1}^{N}\left(g(X_i) - y_i\right)^2$$

$$y_i \in \{0, 1\} \qquad g(X_i) \in \mathbb{R}. \rightarrow \text{scalar}$$

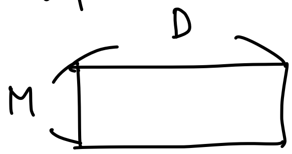
$$g(X) = \sigma\left(\sum_{d=1}^{D} W_{2,d} \cdot h_d(X)\right)$$

$$W_2 \in \mathbb{R}^D : \begin{pmatrix} W_{2,1} \\ \vdots \\ W_{2,D} \end{pmatrix} \qquad h \in \mathbb{R}^D : \begin{pmatrix} h_1(X) \\ h_2(X) \\ \vdots \\ h_D(X) \end{pmatrix}$$

$$\boxed{\therefore\ g(X) = \sigma\left(W_2^T h(X)\right)}$$

$$h_i(X) = X_i + \sigma\left(\sum_{m=1}^{M} W_{1,i,m} \cdot Z_m(X)\right)$$

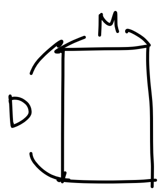
$$X \in \mathbb{R}^D \qquad W_1 \in \mathbb{R}^{M \times D} \qquad Z(X) \in \mathbb{R}^M : \begin{pmatrix} Z_1(X) \\ Z_2(X) \\ \vdots \\ Z_M(X) \end{pmatrix}$$



$$\boxed{\therefore\ h(X) = X + \sigma\left(W_1^T Z\right)}$$

$$Z_i(X) = \sigma\left(\sum_{d=1}^{D} W_{0,i,d} \cdot X_d\right)$$

$$X \in \mathbb{R}^D \qquad W_0 \in \mathbb{R}^{D \times M}$$



$$\boxed{\therefore\ Z(X) = \sigma\left(W_0^T X\right)}$$

2019009261_최가온 🧑🏻‍💻

(1) $$\left[\frac{\partial L}{\partial h}\right]_j = \frac{\partial L}{\partial g} \cdot \left[\frac{\partial g}{\partial h}\right]_j$$

$$= \left[\frac{1}{2} \cdot 2\left(g(x) - y\right)\right] \cdot \sigma\left(w_2^T h(x)\right) \cdot \left(1 - \sigma\left(w_2^T h(x)\right)\right) \cdot \underbrace{w_{2,j}}_{\text{scalar}}$$

$$= \boxed{\left(g(x_j) - y_j\right) \cdot \sigma\left(w_{2,j}\, h(x_j)\right) \cdot \left(1 - \sigma\left(w_{2,j} \cdot h(x_j)\right)\right) \cdot w_{2,j}}$$

(2) $$\left[\frac{\partial L}{\partial w_1}\right]_{d,m} = \boxed{\left[\frac{\partial L}{\partial g} \cdot \frac{\partial g}{\partial h}\right]} \cdot \left[\frac{\partial h}{\partial w_1}\right]_{d,m}$$

└ Delta rule을 고려.

$$= \left(g(x_d) - y_d\right) \cdot \sigma\left(w_{2,d} \cdot h(x_d)\right) \cdot \left(1 - \sigma\left(w_{2,d} \cdot h(x_d)\right)\right) \cdot w_{2,d}$$
$$\cdot\, \sigma\left(w_{1,d,m} \cdot z_d\right) \cdot \left(1 - \sigma\left(w_{1,d,m} \cdot z_d\right)\right) \cdot z_d$$

하위 layer에서 derivative를 구할 때에는 상위 layer에서 구한 derivative 식을 이용하여 recursive한 형태의 도출이 가능하다는 것이 delta rule의 핵심이다.