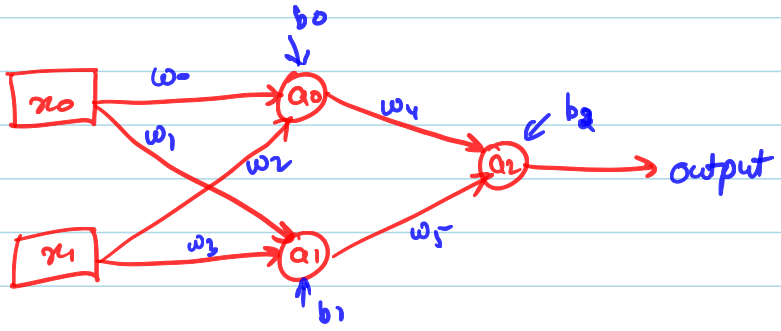
Neural network derivation for backpropagation.

(A)



input: (x_0, x_1)

output: a_2

Error: $L = \frac{1}{2}(y - a_2)^2$ — ①

$\sigma(x) = \frac{1}{1+e^{-x}}$ — ②

Equations:

$z_0 = w_0 x_0 + w_2 x_1 + b_0$ — ③

$a_0 = \sigma(z_0)$ — ④

$z_1 = w_1 x_0 + w_3 x_1 + b_1$ — ⑤

$a_1 = \sigma(z_1)$ — ⑥

$z_2 = w_4 a_0 + w_5 a_1 + b_2$ — ⑦

$a_2 = \sigma(z_2)$ — ⑧

For gradient descent, we want to find

$$\frac{\partial L}{\partial b_i}, \quad \frac{\partial L}{\partial w_j} \quad \forall\, i, j$$

---

① Find $\sigma'(x)$ in terms of $\sigma(x)$.

Ans $\sigma(x) = \frac{1}{1+e^{-x}} \Rightarrow \frac{d\sigma(x)}{dx} = \frac{d}{dx}(1+e^{-x})^{-1} = \frac{-1}{(1+e^{-x})^2}(0 - e^{-x})$

$$= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right) = \sigma(x)\left[\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right]$$

$$= \sigma(x)\left[1 - \sigma(x)\right]$$

② Find all derivatives related to eq ①, ②, ④.

Ans ① $L = \frac{1}{2}(a_2 - y)^2$ where $y \rightarrow$ fixed as per given input, not a variable in the model.

$$\Rightarrow \frac{dL}{da_2} = a_2 - y.$$

② $z_0 = w_0 x_0 + w_2 x_1 + b_0$

$(x_0, x_1) \rightarrow$ input, not a parameter of NN.

$$\Rightarrow \frac{\partial z_0}{\partial w_0} = x_0, \quad \frac{\partial z_0}{\partial w_1} = x_1, \quad \frac{\partial z_0}{\partial b_0} = 1$$

④ $a_0 = \sigma(z_0)$

$\Rightarrow \frac{da_0}{dz_0}, \frac{d}{dz_0}\sigma(z_0) = \sigma(z_0)(1-\sigma(z_0))$

③ Find all partial derivatives: $\frac{\partial L}{\partial b_i}, \frac{\partial L}{\partial w_j}$

Ans $L = \frac{1}{2}(a_2-y)^2 \Rightarrow \frac{dL}{da_2} = a_2-y$

→ $a_2 = \sigma(z_2) \Rightarrow \frac{da_2}{dz_2} = \sigma'(z_2) \Rightarrow \frac{dL}{dz_2} = \sigma'(z_2)(a_2-y)$

→ $z_2 = w_4 a_0 + w_5 a_1 + b_2 \Rightarrow \frac{\partial L}{\partial b_2} = \frac{dL}{dz_2} \times \frac{\partial z_2}{\partial b_2} = \sigma'(z_2)(a_2-y) \times 1$

Similarly, $\frac{\partial L}{\partial w_4} = \frac{dL}{dz_2} \times a_0$, $\frac{\partial L}{\partial w_5} = \frac{dL}{dz_2} a_1$

→ $a_1 = \sigma(z_1) \Rightarrow \frac{\partial L}{\partial z_1} = \frac{dL}{dz_2}\frac{\partial z_2}{\partial z_1}$

$= \frac{dL}{dz_2}\frac{\partial}{\partial z_1}(\cancel{w_4 a_0}^{0} + w_5 a_1 + \cancel{b_2}^{0})$

$= \frac{dL}{dz_2} w_5 \frac{da_1}{dz_1} = w_5 \frac{dL}{dz_2}\sigma'(z_1)$

→ $a_0 = \sigma(z_0) \Rightarrow \frac{\partial L}{\partial z_0} = w_4 \frac{dL}{dz_2}\sigma'(z_0)$

→ $z_1 = w_1 x_0 + w_3 x_1 + b_1 \Rightarrow \frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial b_1} = (a_2-y)\sigma'(z_2) w_5 \sigma'(z_1)$

$\Rightarrow \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_1} = (a_2-y)\sigma'(z_2) w_5 \sigma'(z_1) x_0$

$\Rightarrow \frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_3} = (a_2-y)\sigma'(z_2) w_5 \sigma'(z_1) x_1$

→ $z_0 = w_0 x_0 + w_2 x_1 + b_0$

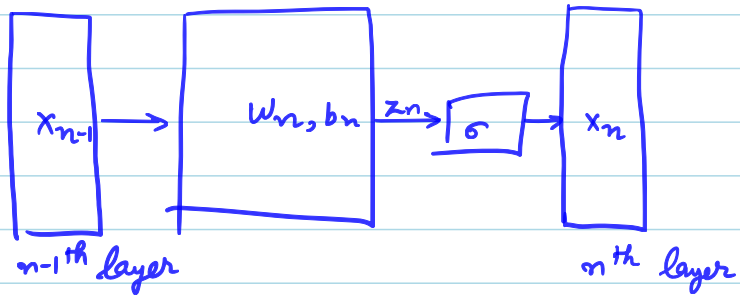
$\Rightarrow \frac{\partial L}{\partial b_0} = \frac{\partial L}{\partial z_0}\frac{\partial z_0}{\partial b_0} = (a_2-y)\sigma'(z_1) w_4 \sigma'(z_0)$

$\Rightarrow \frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial z_0}\frac{\partial z_0}{\partial w_0} = (a_2-y)\sigma'(z_2) w_4 \sigma'(z_0) x_0$

$\Rightarrow \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_0}\frac{\partial z_0}{\partial w_2} = (a_2-y)\sigma'(z_2) w_4 \sigma'(z_0) x_1$

(β) matrix form:-

let the layers are connected as:



Equations:

$x_n = \sigma(z_n)$ where x_n: column vector of size m_n — ①

z_n: column vector of size m_n

σ is applied component wise on z_n.

$z_n = W_n x_{n-1} + b_n$ where b_n: column vector of size m_n — ②

W_n: matrix of size $m_n \times m_{n-1}$

or $x_n, z_n, b_n \in \mathbb{R}^{m_n \times 1}$

$W_n \in \mathbb{R}^{m_n \times m_{n-1}}$

① $\frac{dx_n}{dz_n}$: we first need to define the meaning of vector differentiation.

let $y = f(x)$ $\quad y \in \mathbb{R}^{n \times 1}$, $x \in \mathbb{R}^{m \times 1}$

Then $\frac{dy}{dx} := \begin{bmatrix} \frac{dy_1}{dx_1} & \cdots & \frac{dy_n}{dx_1} \\ \vdots & & \vdots \\ \frac{dy_1}{dx_m} & & \frac{dy_n}{dx_m} \end{bmatrix}_{m \times n}$

Point to remember: if y is a scalar, $\frac{dy}{dx}$ is $\begin{bmatrix} | \\ | \end{bmatrix}$ instead of $[- \; -]$

Coming to $\frac{dx_n}{dz_n}$, $[a_{ij}] = \frac{dx_n}{dz_n}$

$\Rightarrow a_{ij} = \frac{dx_n^{(i)}}{dz_n^{(j)}} = \frac{d\sigma(z_n^{(i)})}{d z_n^{(j)}} = \begin{cases} 0 & i \neq j \\ \sigma'(z_n^{(i)}) & i = j \end{cases}$

$$\Rightarrow \frac{dx_n}{dz_n} = \underset{m_n \times m_n}{\text{diag}}\left[\sigma'(z_n^{(1)}), \sigma'(z_n^{(2)}), \text{---}, \sigma'(z_n^{(m_n)})\right]$$

$$\in \mathbb{R}^{m_n \times m_n} \quad \text{---} \quad (1)$$

(2) $\frac{\partial z_n}{\partial b_n}$: $z_n = w_n x_{n-1} + b_n$

$\Rightarrow \frac{\partial z_n}{\partial b_n}$ should be a matrix of size $m_n \times m_n$.

$$[a_{ij}] = \frac{\partial z_n}{\partial b_n} \Rightarrow a_{ij} = \frac{\partial z_n^{(j)}}{\partial b_n^{(i)}}$$

Now, $z_n^{(j)} = \sum_{k=1}^{m_{n-1}} w_n^{(j,k)} x_{n-1}^{(k)} + b_n^{(j)}$

$$\Rightarrow a_{ij} = \frac{\partial z_n^{(j)}}{\partial b_n^{(i)}} = \frac{\partial}{\partial b_n^{(i)}}\left[\sum_{k=1}^{m_{n-1}} w_n^{(j,k)} x_{n-1}^{(k)} + b_n^{(j)}\right]$$

$$= \frac{\partial b_n^{(j)}}{\partial b_n^{(i)}} = \delta_{ij}$$

$$\Rightarrow a_{ij} = \delta_{ij} \text{ or } \frac{\partial z_n}{\partial b_n} = I_{m_n \times m_n}$$

(3) $\frac{\partial z_n}{\partial x_{n-1}}$: $z_n = w_n x_n + b_n$

Resulting matrix should have a size of $m_{n-1} \times m_n$ (from definition)

$$[a_{ij}] = \frac{\partial z_n}{\partial x_{n-1}} \Rightarrow a_{ij} = \frac{\partial z_n^{(j)}}{\partial x_{n-1}^{(i)}}$$

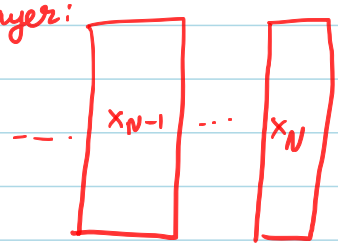
$$= \frac{\partial}{\partial x_{n-1}^{(i)}}\left[\sum_{k=1}^{m_{n-1}} w_n^{(j,k)} x_{n-1}^{(k)} + b_n^{(j)}\right]$$

$$= w_n^{(j,i)}$$

$$\Rightarrow a_{ij} = w_n^{j,i} \text{ or } \frac{\partial z_n}{\partial x_{n-1}} = w_n^T$$

(4) $\frac{\partial z_n}{\partial w_n}$: This will be 3rd order tensor & we don't know / want to go into that. We will go by another approach later.

(c) Last layer:



Loss $L = \frac{1}{2}(x_N - y)^T (x_N - y)$

Now, we will compute all partial derivatives

① $\frac{\partial L}{\partial x_N}$ = column vector of size m_N from definition.

$$[a_i] = \frac{\partial L}{\partial x_N} \Rightarrow a_i = \frac{dL}{dx_N^{(i)}} = \frac{d}{dx_N^{(i)}} \frac{1}{2} \sum_{k=1}^{m_N} (x_N^{(k)} - y^{(k)})^2$$

$$= x_N^{(i)} - y^{(i)}$$

$$\Rightarrow \frac{\partial L}{\partial x_N} = x_N - y \quad \text{— ①}$$

② We have $x_N = \sigma(z_N)$ & know what is $\sigma'(z_N)$

$$\Rightarrow \frac{\partial L}{\partial z_N} = \frac{\partial x_N}{\partial z_N} \frac{\partial L}{\partial x_N} = \underbrace{\underbrace{\sigma'(z_N)}_{m_N \times m_N} \underbrace{(x_N - y)}_{m_N \times 1}}_{m_N \times 1} \quad \text{— ②}$$

Remember that Chain Rule is reversed here!!

③ We have $z_N = W_N x_{N-1} + b_N$.

$$\Rightarrow \frac{\partial L}{\partial b_N} = \frac{\partial z_N}{\partial b_N} \frac{\partial L}{\partial z_N} = I_{m_N \times m_N} \cdot \sigma'(z_N) \frac{\partial L}{\partial x_N}$$

$$\Rightarrow \boxed{\frac{\partial L}{\partial b_N} = \sigma'(z_N) \frac{\partial L}{\partial x_N} \quad \text{or} \quad \frac{\partial L}{\partial b_\ell} = \sigma'(z_\ell) \frac{\partial L}{\partial x_\ell}}$$

$$= \sigma'(z_N)(x_N - y)$$

④ $$\boxed{\frac{\partial L}{\partial x_{N-1}} = \frac{\partial z_N}{\partial x_{N-1}} \frac{\partial L}{\partial z_N} = W_N^T \sigma'(z_N) \frac{\partial L}{\partial x_N}}$$

→ Going back one more layer to achieve generalization,

$$x_{N-1} = \sigma(z_{N-1})$$
$$z_{N-1} = W_{N-1} x_{N-2} + b_{N-1}$$

$$\frac{\partial L}{\partial z_{N-1}} = \frac{\partial x_{N-1}}{\partial z_{N-1}} \frac{\partial L}{\partial x_{N-1}} = \sigma'(z_{N-1}) W_N^T \frac{\partial L}{\partial x_N}$$

$$\frac{\partial L}{\partial x_{N-2}} = \frac{\partial z_{N-1}}{\partial x_{N-2}} \frac{\partial L}{\partial z_{N-1}} = \underbrace{W_{N-1}^T \sigma'(z_{N-1}) W_N^T \sigma'(z_N)}_{\text{pattern.}} \frac{\partial L}{\partial x_N}$$

$\frac{\partial L}{\partial z_{N-1}}$

$$\frac{\partial L}{\partial b_{N-1}} = \frac{\partial z_{N-1}}{\partial b_{N-1}} \cdot \frac{\partial L}{\partial z_{N-1}} = \frac{\partial L}{\partial z_{N-1}}$$

— What about $\frac{\partial L}{\partial W_N}$?

Let $[a_{ij}] = \frac{\partial L}{\partial W_N}$. How it will look like?

$a_{ij} = \frac{\partial L}{\partial W_N^{(i,j)}}$. Effect of $W_N^{(i,j)}$ will be reflected to $x_N^{(i)}$ only

$$\Rightarrow a_{ij} = \frac{\partial L}{\partial x_N^{(i)}} \frac{\partial x_N^{(i)}}{\partial W_N^{(i,j)}}$$ (why order changed? bcz now, this is a real number)

$$= (x_N^{(i)} - y^{(i)}) \cdot \frac{\partial x_N^{(i)}}{\partial z_N^{(i)}} \frac{\partial z_N^{(i)}}{\partial W_N^{(i,j)}}$$

$$= (x_N^{(i)} - y^{(i)}) \sigma'(z_N^{(i)}) \cdot \frac{\partial}{\partial W_N^{(i,j)}} \left(\sum_{k=1}^{m_{N-1}} W_N^{(i,k)} x_{N-1}^{(k)} + b_N^{(i)} \right)$$

$$= \text{———} \cdot (x_{N-1}^{(j)})$$

$$\Rightarrow a_{ij} = \alpha_i \beta_j = \begin{bmatrix} 1 \\ \alpha \\ 1 \end{bmatrix} [\text{— } \beta \text{ —}]$$

$$\Rightarrow \frac{\partial L}{\partial W_N} = \underbrace{\underbrace{[\sigma'(z_N)(x_N - y)]}_{m_N \times 1} \underbrace{x_{N-1}^T}_{1 \times m_{N-1}}}_{m_N \times m_{N-1} \text{ as expected}} = \frac{\partial L}{\partial z_N} x_{N-1}^T$$

Let's say we know the fact that

① if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $vec(A) := \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix}$

② if $Z = ABC$, then $vec(Z) = (C^T \otimes A)\, vec(B)$

can we do something?

For $\frac{\partial L}{\partial W_N}$, we can find $\frac{\partial L}{\partial \omega_N}$ where $\omega_N = vec(W_N)$

and since the result will a vector of same size, we can devectorize?

$Z_N = W_N X_{N-1} + b_N \Rightarrow$ $z_N = vec(W_N x_{N-1}) + b_N$

$= vec(I\, W_N\, X_{N-1}) + b_N$

$= (x_{N-1}^T \otimes I)\, \omega_N + b_N$

$(m_N * m_{N-1}) \times 1$

$W_N \in \mathbb{R}^{m_N \times m_{N-1}} \Rightarrow \omega_N \in \mathbb{R}^{(m_N * m_{N-1}) \times 1}$

$x_{N-1} \in \mathbb{R}^{m_{N-1} \times 1} \Rightarrow x_{N-1}^T \in \mathbb{R}^{1 \times m_{N-1}}$

$I \in \mathbb{R}^{m_N \times m_N} \Rightarrow x_{N-1}^T \otimes I \in \mathbb{R}^{m_N \times (m_N * m_{N-1})}$

$\Rightarrow \frac{\partial L}{\partial \omega_N} \overset{?}{=} \frac{\partial z_N}{\partial \omega_N} \times \frac{\partial x_N}{\partial z_N} \times \frac{\partial L}{\partial x_N}$

$= \underbrace{(x_{N-1}^T \otimes I)^T}_{\omega \times m_N} \underbrace{\sigma'(z_N)}_{m_N \times 1} \underbrace{x_{N-Y}}_{} \rightarrow$ well defined ✓

$= (x_{N-1} \otimes I)\, \sigma'(z_N)\, (x_{N-Y})$

will devec(RHS) $\overset{?}{=} \sigma'(z_N)\, (x_{N-Y})\, x_{N-1}^T$?

Take example: $x_{N-1} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$ $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$\sigma'(z_N)(x_{N-Y}) = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$.

$\Rightarrow \sigma'(z_N)(x_{N-Y})\, x_{N-1}^T \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} [x_0 \; x_1] = \begin{bmatrix} x_0 y_0 & x_1 y_0 \\ x_0 y_1 & x_1 y_1 \\ x_0 y_2 & x_1 y_2 \end{bmatrix}$

& $(x_{N-1} \otimes I) = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} x_0 & 0 & 0 \\ x_1 & 0 & 0 \\ 0 & x_0 & 0 \\ 0 & x_1 & 0 \\ 0 & 0 & x_0 \\ 0 & 0 & x_1 \end{pmatrix}$

Product $= \begin{pmatrix} x_0 & 0 & 0 \\ 0 & x_0 & 0 \\ 0 & 0 & x_0 \\ x_1 & 0 & 0 \\ 0 & x_1 & 0 \\ 0 & 0 & x_1 \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}$ $=$ $\begin{pmatrix} x_0 y_0 \\ x_0 y_1 \\ x_0 y_2 \\ x_1 y_0 \\ x_1 y_1 \\ x_1 y_2 \end{pmatrix}$

⇒ Need to poove for general case:

$$\text{devec}\left((X \otimes I)\, Y\right) = YX^T$$

or

$$(X \otimes I)Y = \text{vec}(YX^T)$$ → This is truth!! So everything matches.

→ From above, if $z = Wx + b$

Then $$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z} x^T$$

$$\frac{\partial L}{\partial x} = W^T \frac{\partial L}{\partial z}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial z}$$

Books referred before these notes:

① Math for deep learning - Ronald T Kneusel

② Kronecker Products & Matrix Calculus: with Applications - Alexander Graham.