$k \in \{0,1,2,\dots,m\}$ $k=0$: input layer, $k=m$: output layer $\quad X \in \mathbb{R}^{N\times M}$ N: instances, M: features

for $1 \le k \le m$

$[W^k]_{ij} = W^k_{ij}$ $\quad W^k = \begin{bmatrix} b^k_1 & b^k_2 & \cdots & b^k_{r_k} \\ W^k_{11} & W^k_{12} & \cdots & W^k_{1r_k} \\ \vdots & \vdots & & \vdots \\ W^k_{r_{k-1}1} & W^k_{r_{k-1}2} & \cdots & W^k_{r_{k-1}r_k} \end{bmatrix}$ $\quad \Theta^k = \begin{bmatrix} 1 & g(a^k_{11}) & g(a^k_{12}) & \cdots & g(a^k_{1r_k}) \\ 1 & g(a^k_{21}) & g(a^k_{22}) & \cdots & g(a^k_{2r_k}) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & g(a^k_{N1}) & g(a^k_{N2}) & \cdots & g(a^k_{Nr_k}) \end{bmatrix}$

$i \in \{1,\dots,r_{k-1}\}$ where $r_0 = N$

$j \in \{1,\dots,r_k\}$ $r_m = 1$

$A^k = \Theta^{k-1} W^k \Rightarrow A^k = \begin{bmatrix} a^k_{11} & a^k_{12} & \cdots & a^k_{1r_k} \\ a^k_{21} & a^k_{22} & \cdots & a^k_{2r_k} \\ \vdots & \vdots & & \vdots \\ a^k_{N1} & a^k_{N2} & \cdots & a^k_{Nr_k} \end{bmatrix} \Rightarrow \Theta^k = [\vec{1}, g(A^k)] \in \mathbb{R}^{N\times(r_k+1)}$

$\Theta^0 = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1M} \\ 1 & X_{21} & X_{22} & \cdots & X_{2M} \\ \vdots & & & & \\ 1 & X_{N1} & X_{N2} & \cdots & X_{NM} \end{bmatrix}$

$W^k \in \mathbb{R}^{(r_{k-1}+1)\times r_k}$

$\Theta^{k-1} \in \mathbb{R}^{N\times(r_{k-1}+1)}$

$A^k \in \mathbb{R}^{N\times r_k}$

Denote $W^k = \begin{bmatrix} \vec{b}^k \\ W^k_* \end{bmatrix}$ and $\Theta^k = [\vec{1}, \Theta^k_*] \Rightarrow \Theta^k_* = g(A^k)$

Define $\Delta^m = \hat{Y} - Y \Rightarrow \Delta^m = \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}$ $\quad \hat{Y} = g(A^m)$ where $A^m \in \mathbb{R}^{N\times 1}$

for $0 \le k \le m-1$ $\quad$ filled unitary matrix $N \{ [1 \cdots 1] \}$ (r_k columns)

$$\Delta^k = \Theta^k_* \odot (\mathbb{1}_{(N\times r_k)} - \Theta^k_*) \odot [\Delta^{k+1} (W^{k+1}_*)^T]$$ where $A \odot_{ij} B = a_{ij} b_{ij}$ (Hadamard product)

make dimensional transforms: $\Theta^{k-1} \mapsto T_1(\Theta^{k-1})$; $\Delta^k \mapsto T_2(\Delta^k)$

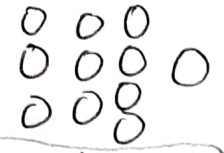
where $T_1(\Theta^{k-1}) \in \mathbb{R}^{N\times[(r_{k-1}+1)\times 1]}$ and $T_2(\Delta^k) \in \mathbb{R}^{N\times 1\times r_k}$

Define $\partial\mathcal{E}^k = T_1(\Theta^{k-1}) \odot_i T_2(\Delta^k) \Rightarrow \partial\mathcal{E}^k \in \mathbb{R}^{N\times(r_{k-1}+1)\times r_k}$

$\Rightarrow \partial\mathcal{E}^k = [[\partial\mathcal{E}^k]_1, [\partial\mathcal{E}^k]_2, \dots, [\partial\mathcal{E}^k]_N]$; $D\mathcal{E}^k = \frac{1}{N}\sum_{d=1}^{N} [\partial\mathcal{E}^k]_d$

where $[\partial\mathcal{E}^k]_d \in \mathbb{R}^{(r_{k-1}+1)\times r_k}$ and $D\mathcal{E}^k \in \mathbb{R}^{(r_{k-1}+1)\times r_k}$

iterate til $\hat{W}^k = W^k - \alpha D\mathcal{E}^k$ converges



given: $[X, Y]$
$X \in \mathbb{R}^{N \times M}$

set: W^K for $1 \le K \le m$
$r_0 = N$, $r_m = 1$
decide r_K for $1 \le K \le m-1$
W^K for $1 \le K \le m$, $W^K \in \mathbb{R}^{(r_{K-1}+1) \times r_K}$

K in range $(1, m)$

algorithm

$K \in \{0, 1, 2, 3\}$

$\Theta^0 = [\vec{1}, X]$

- for $1 \le K \le m-1$:
  $A^K = \Theta^{K-1} W^K$
  $\Theta^K_* = \sigma(A^K)$
  $\Theta^K = [\vec{1}, \Theta^K_*]$

$A^m = \Theta^{m-1} W^m$
$\Delta^m = \hat{Y} - Y$

- for $m-1 \ge K \ge 1$:
  $\Delta^K = \Theta^K_* \odot (I_{(N \times r_K)} - \Theta^K_*) \odot [\Delta^{K+1} (W^{K+1}_*)^T]$

- for $1 \le K \le m-1$:
  $\Theta^{K-1} \mapsto T_1(\Theta^{K-1})$
  $\Delta^K \mapsto T_2(\Delta^K)$

$$\partial \mathcal{E}^K = T_1(\Theta^{K-1}) \odot_i T_2(\Delta^K)$$
$$D\mathcal{E}^K = \frac{1}{N} \sum_{d=1}^{N} [\partial \mathcal{E}^K]_d$$
$$\hat{W}^K = W^K - \alpha D\mathcal{E}^K$$