NN BP Algorithm

$\boxed{\text{Model:}}$ $f_{t+1}(\underline{x}) = \sigma(W_t f_t(\underline{x}) + b_t)$

$\boxed{\text{when } t=0}$ $f_1(\underline{x}) = \sigma(W_0 \underline{x} + b_0)$

$\boxed{\text{when } t=T-1}$ $f_T(\underline{x}) = \sigma(W_{T-1} f_{T-1}(\underline{x}) + b_{T-1})$

$\boxed{\text{Vector-form BP}}$ $\Phi(\underline{\theta}) = \text{Loss}(f_T(\underline{x}), y)$

In vector-BP algo

① $\dfrac{\partial \Phi}{\partial w_{ij}^{T-1}} = \underbrace{\dfrac{\partial \Phi}{\partial f_i^T} \cdot \dfrac{\partial f_i^T}{\partial h_i^T}}_{\delta_i^T} \cdot \underbrace{\dfrac{\partial h_i^T}{\partial w_{ij}^{T-1}}}_{f_j^{T-1}(\underline{x})}$

② $\dfrac{\partial \Phi}{\partial w_{ij}^{T-2}} = \sum_s \boxed{\dfrac{\partial \Phi}{\partial h_s^T} \cdot \dfrac{\partial h_s^T}{\partial f_i^{T-1}}} \cdot \dfrac{\partial f_i^{T-1}}{\partial h_i^{T-1}} \cdot \dfrac{\partial h_i^{T-1}}{\partial w_{ij}^{T-2}}$

$= \underbrace{\left(\sum_s \delta_s^T \cdot W_{si}^{T-1}\right) \cdot \sigma'(h_i^{T-1})}_{\delta_i^{T-1}} \cdot \underbrace{f_j^{T-2}(\underline{x})}_{f_j^{T-2}(\underline{x})}$

$\boxed{\dfrac{\partial \Phi}{\partial f_i^{T-1}} = \dfrac{\partial \Phi}{\partial f_i^T} \cdot \dfrac{\partial f_i^T}{\partial f_i^{T-1}}}$

our interest is: $\nabla_{W_t} \Phi(\underline{\theta}^k)$ for $t=0,1,\dots,T-1$

① $\nabla_{W_{T-1}} \Phi(\underline{\theta}^k) = \nabla_{W_{T-1}} f_T \underbrace{\nabla_{f_T} \Phi(\underline{\theta}^k)}_{P_T}$

② $\nabla_{W_{T-2}} \Phi(\underline{\theta}^k) = \nabla_{W_{T-2}} f_{T-1} \underbrace{\nabla_{f_{T-1}} f_T \; \nabla_{f_T} \Phi(\underline{\theta}^k)}_{P_{T-1}}$

$\boxed{P_{T-1} = \nabla_{f_{T-1}} f_T \; P_T}$ this term will be very complicated

where $f_T = \sigma(W_{T-1} f_{T-1} + b_{T-1})$

ⓚ we have $P_T \to P_{T-1} \to \cdots P_k$ $= \nabla_{f_{T-1}} u \cdot \nabla_u f_T = W_{T-1}^T \cdot \sigma'(W_{T-1} f_{T-1} + b_{T-1})$

a) then $P_{k-1} = \nabla_{f_{k-1}} f_k \, P_k$

b) $\nabla_{W_{k-2}} \Phi(\underline{\theta}) = \nabla_{W_{k-2}} f_{k-1} \nabla_{f_{k-1}} \Phi(\underline{\theta})$

$= \nabla_{W_{k-2}} f_{k-1} \cdot P_{k-1}$

where $f_{k-1} = \sigma(W_{k-2} \cdot f_{k-2} + b_{k-2})$

c) then we have:

$$\begin{array}{ccccccc} P_T & \to \cdots \to & P_k & \to & \boxed{P_{k-1}} \\ & & \downarrow & & \downarrow \\ & & \nabla_{W_{k-1}} \Phi & & \nabla_{W_{k-2}} \Phi \end{array}$$

$\boxed{\text{Scaler- form BP}}$

$$\boxed{W_k} \Rightarrow W^{k}_{d_{k+1}, d_k} \Rightarrow f_i^{k+1}(\underline{x}) = 6\left(\sum_{j=1}^{d_k} W^{k}_{i,j} f_j^{k}(\underline{x}) + b^k\right)$$

$\boxed{k = 0, 1, 2, \dots, T-1}$ $\qquad i = 1, 2, \dots, d_{k+1}$

$\boxed{\text{when } k=0}$ $\quad f_i^1(\underline{x}) = 6\left(\sum_{j=1}^{d_0 \; d} W^0_{i,j} \boxed{f_j^0(\underline{x})} + b^0\right)$ $\quad x_j$

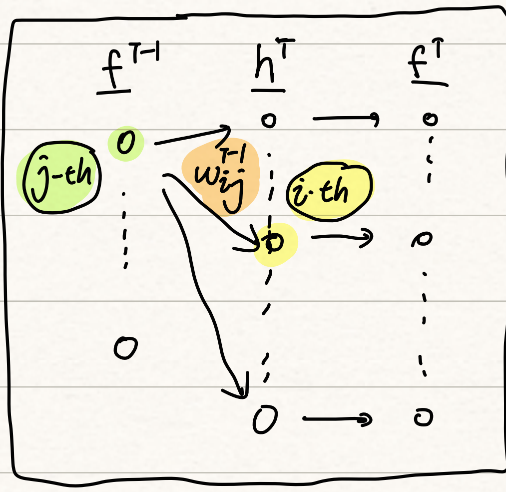
$\boxed{\text{when } k=T-1}$ $\quad f_i^T(\underline{x}) = 6\left(\sum_{j=1}^{d_{T-1}} W^{T-1}_{i,j} f_j^{T-1}(\underline{x}) + b^{T-1}\right)$

$$\Phi(\underline{\theta}) = \text{Loss}\left(f^T(\underline{x}), y\right)$$

① $$\frac{\partial \Phi}{\partial W_{ij}^{T-1}} = \frac{\partial \Phi}{\partial f_i^T} \cdot \frac{\partial f_i^T}{\partial h_i^T} \cdot \frac{\partial h_i^T}{\partial w_{ij}^{T-1}}$$

$$h_i^T := \sum_{j=1}^{d_{T-1}} W^{T-1}_{i,j} f_j^{T-1}(\underline{x}) + b^{T-1}$$

$$= \underbrace{\frac{\partial \Phi}{\partial f_i^T} \cdot 6'(h_i^T)}_{\frac{\partial \Phi}{\partial h_i^T}} \cdot f_j^{T-1}(\underline{x})$$



② $$\frac{\partial \Phi}{\partial W_{ij}^{T-2}} = \sum_{s=1}^{d_T} \frac{\partial \Phi}{\partial h_s^T} \cdot \frac{\partial h_s^T}{\partial w_{ij}^{T-2}}$$

$$= \sum_{s=1}^{d_T} \frac{\partial \Phi}{\partial h_s^T} \cdot \frac{\partial h_s^T}{\partial f_i^{T-1}} \cdot \frac{\partial f_i^{T-1}}{\partial W_{ij}^{T-2}}$$

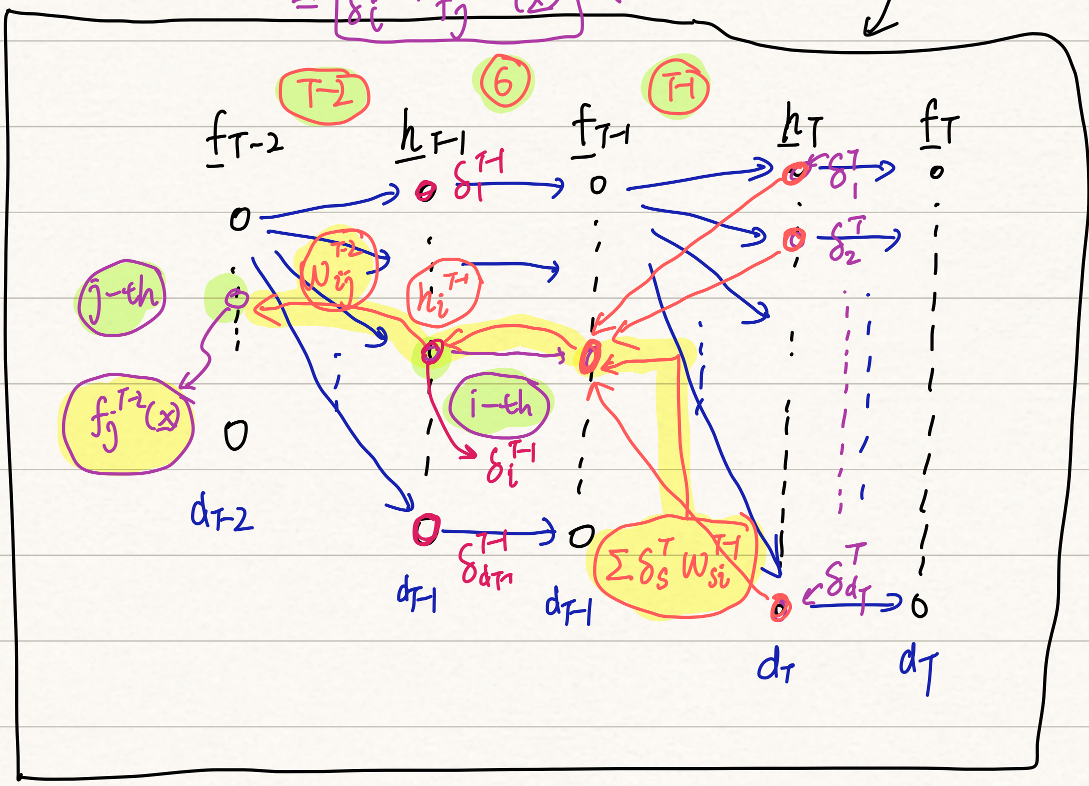
$$= \left(\sum_{s=1}^{d_T} \frac{\partial \Phi}{\partial h_s^T} \cdot \frac{\partial h_s^T}{\partial f_i^{T-1}}\right) 6'(h_i^{T-1}) \cdot f_j^{T-2}(\underline{x})$$

$$f_i^{T-1}(\underline{x}) = 6\left[\sum_{j=1}^{d_{T-2}} w_{ij}^{T-2} f_j^{T-2}(\underline{x}) + b^{T-2}\right]$$

$$h_s^T = \sum_{j=1}^{d_{T-1}} w_{sj}^{T-1} \cdot f_j^{T-1}(x) + b^{T-1}$$

$$= \left(\sum_{s=1}^{d_T} \frac{\partial \Phi}{\partial h_s^T} \cdot w_{si}^{T-1} \right) \sigma'(h_i^{T-1}) \, f_j^{T-2}(x)$$

$$= \boxed{\delta_i^{T-1} \cdot f_j^{T-2}(x)} \; ☆$$



Algo

initialization

$$\Rightarrow \boxed{\delta_j^T := \frac{\partial \Phi}{\partial f_j^T} \cdot \sigma'(h_j^T)}$$

$$\Downarrow$$

Update of error $\;\; \delta_i^k = \left[\sum_{s=1}^{d_{k+1}} \delta_s^{k+1} \cdot w_{si}^k \right] * \sigma'(h_i^k)$

update of gradient $\Rightarrow \boxed{\frac{\partial \Phi}{\partial w_{ij}^{\ell}} = \delta_i^{\ell+1} \cdot f_j^{\ell}(x)}$