# NN BP Algorithm

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

when $t=0$: $f_1(x) = \sigma(W_0 x + b_0)$

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

---

**Vector-form BP**

$\Phi(\theta) = \text{Loss}(f_T(x), y)$

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

In vector-BP algo

① $\frac{\partial \Phi}{\partial w_{ij}^{T-1}} = \frac{\partial \Phi}{\partial f_i^T} \cdot \frac{\partial f_i^T}{\partial h_i^T} \cdot \frac{\partial h_i^T}{\partial w_{ij}^{T-1}}$, with $\delta_i^T$ and $f_j^{T-1}(x)$

② $\frac{\partial \Phi}{\partial w_{ij}^{T-2}} = \sum_s \left[\frac{\partial \Phi}{\partial h_s^T} \cdot \frac{\partial h_s^T}{\partial f_i^{T-1}}\right] \frac{\partial f_i^{T-1}}{\partial h_i^{T-1}} \cdot \frac{\partial h_i^{T-1}}{\partial w_{ij}^{T-2}}$

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

with $\delta_i^{T-1}$ and $f_j^{T-2}(x)$

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

① $\nabla_{W_{T-1}} \Phi(\theta^k) = \nabla_{W_{T-1}} f_T \; \nabla_{f_T} \Phi(\theta^k)$, with $\nabla_{f_T} \Phi(\theta^k) = P_T$

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

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

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

③ 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) $\nabla_{W_{k-2}} \Phi(\theta) = \nabla_{W_{k-2}} f_{k-1} \; \nabla_{f_{k-1}} \Phi(\theta)$

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

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

c) then we have:

$P_T \to \cdots \to P_k \to P_{k-1}$

$\downarrow$ $\downarrow$

$\nabla_{W_{k-1}} \Phi$ $\nabla_{W_{k-2}} \Phi$

## Scaler-form BP

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

$k = 0, 1, 2, \ldots, T-1$ $\quad i = 1, 2, \ldots, d_{k+1}$

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

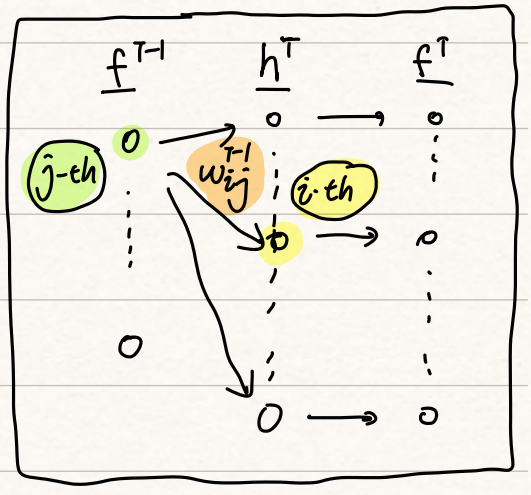
when $k=T-1$ $\quad f_i^T(\underline{x}) = \sigma\left(\sum_{j=1}^{d_{T-1}} w^{T-1}_{i,j} f_j^{T-1}(\underline{x}) + b^{T-1}\right)$

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

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

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

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




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

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

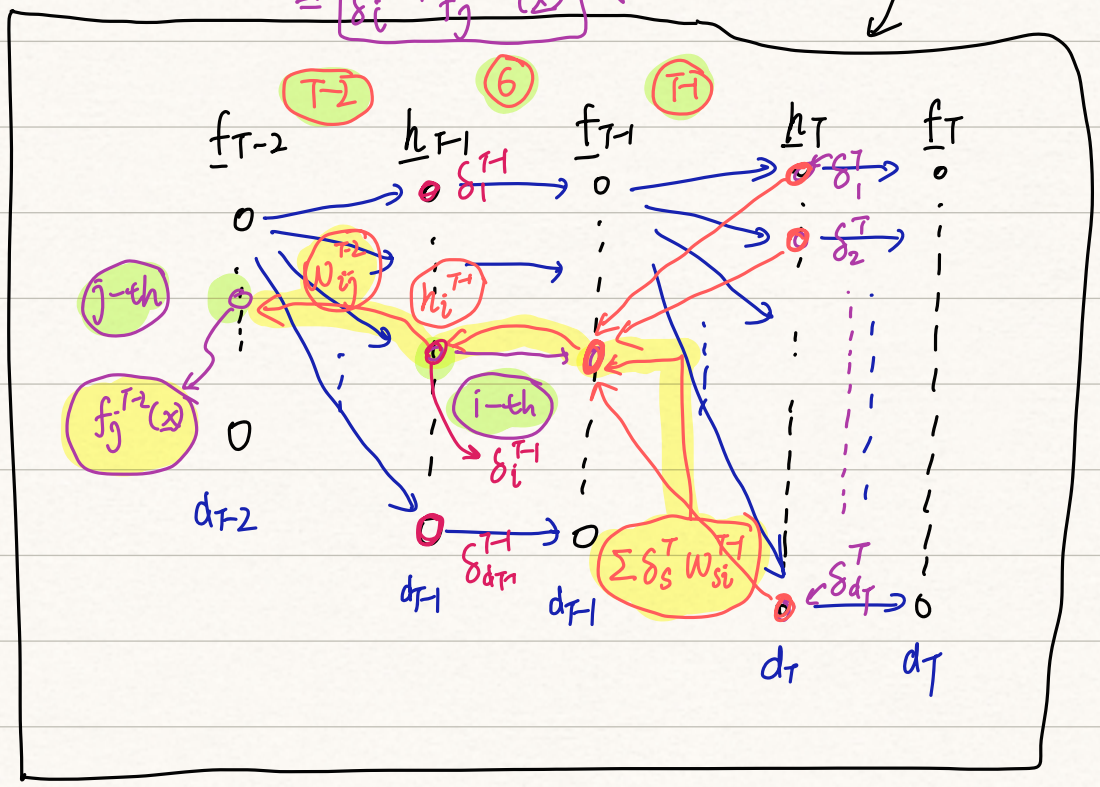
$= \left(\sum_{s=1}^{d_T} \dfrac{\partial \Phi}{\partial h_s^T} \cdot \dfrac{\partial h_s^T}{\partial f_i^{T-1}}\right) \sigma'(h_i^{T-1}) \cdot f_j^{T-2}(\underline{x})$

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

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

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

$$= \delta_i^{T-1} \cdot f_j^{T-2}(x) \quad \star$$




**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** $\Rightarrow \boxed{\frac{\partial \Phi}{\partial w_{ij}^l} = \delta_i^{l+1} \cdot f_j^l(x)}$