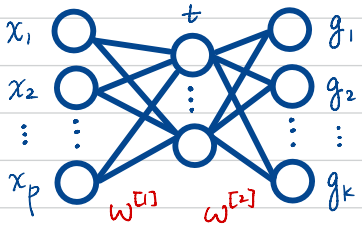
SPINN for Softmax-NN with gMLP. (參考 Feng-Simon 的文章)



$$L = \prod_{i=1}^{n} \prod_{\ell=1}^{k} \left[\frac{\exp(g_\ell(x_i))}{\sum_{j=1}^{k} \exp(g_j(x_i))} \right]^{1\{y_i=\ell\}}$$

$$J = -\log L = -\sum_{i=1}^{n} \sum_{\ell=1}^{k} 1\{y_i=\ell\} \log \frac{\exp(g_\ell(x_i))}{\sum_{j=1}^{k} \exp(g_j(x_i))}$$

$$= -\sum_{i=1}^{n} \sum_{\ell=1}^{k} 1\{y_i=\ell\}\, g_\ell(x_i) + \sum_{i=1}^{n} \log \sum_{j=1}^{k} \exp(g_j(x_i))$$

$$\nabla_{g_{k'}} J = -\sum_{i=1}^{n} 1\{y_i=k'\} + \sum_{i=1}^{n} \frac{\exp(g_{k'}(x_i))}{\sum_{j=1}^{k} \exp(g_j(x_i))}$$

Forward: X^T $p \times n$

$$\underset{t\times n}{\vec{h_t}} = a\left(\underset{t\times p}{w^{[1]}} \underset{p\times n}{X^T}\right)$$

$$\underset{k\times n}{\vec{g}} = \underset{k\times t}{w^{[2]}} \underset{t\times n}{\vec{h_t}}$$

Backprop:

$$\underset{k\times t}{\frac{\partial J}{\partial w^{[1]}}} = \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot \frac{\partial g_\ell}{\partial w^{[1]}} = \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot \begin{bmatrix} — & 0 & — \\ — & \frac{\partial g_\ell}{\partial w_\ell^{[1]}} & — \\ — & \vdots & — \\ — & 0 & — \end{bmatrix} = \underset{k\times k}{\mathrm{diag}\left(\frac{\partial J}{\partial g_\ell}\right)} \cdot \underset{k\times 1}{\mathbf{1}_k} \cdot \underset{1\times t}{\vec{h_t}^T}$$

$$\underset{t\times p}{\frac{\partial J}{\partial w^{[1]}}} = \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot \frac{\partial g_\ell}{\partial \vec{h_t}} \cdot \frac{\partial \vec{h_t}}{\partial (w^{[1]} X^T)} \cdot \frac{\partial (w^{[1]} X^T)}{\partial w^{[1]}}$$

$$= \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot w_\ell^{[2]T} * \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} \cdot X_i^T$$

$$= \underbrace{\begin{bmatrix} | & & | \\ w_1^{[2]T} & \cdots & w_k^{[2]T} \\ | & & | \end{bmatrix}}_{t\times k} \cdot \underbrace{\begin{bmatrix} \nabla_{g_1} J \\ \vdots \\ \nabla_{g_k} J \end{bmatrix}}_{k\times 1} * \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix}}_{\text{w.r.t. } w^{[1]} X^T,\ t\times 1} \cdot \underbrace{X_i^T}_{1\times p}$$

$$= \underset{t\times t}{\mathrm{diag}\left(\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix}\right)} \cdot \underset{t\times k}{w^{[2]T}} \cdot \underset{k\times 1}{\nabla_g J} \cdot \underset{1\times p}{X_i^T}$$

Objective:

$$\min_{w} \quad \mathcal{L}(y, w) + \frac{\lambda_0}{2}\|w^{[i]}\|_2^2 + \sum_{j=1}^{p} P_\gamma\Big(\sum_{t=1}^{I} P_\gamma(|w_{jt}^{[i]}|;\lambda);\lambda\Big).$$

Solution:

(i) Smooth part: $\mathcal{L}(y, w) + \frac{\lambda_0}{2}\|w^{[i]}\|_2^2$

Therefore, $\frac{\partial \mathcal{L}}{\partial w^{[i]}} = \mathrm{diag}\left(\frac{\partial J}{\partial q_k}\right) \cdot \mathbf{1}_K \cdot \tilde{h}_t + \lambda_0 \cdot w_{kt}^{[i]}$.

(ii) Thresholding part:

$$R = \frac{1}{2}\sum_{j=1}^{p}\|w_j^{[i]} - u\|_2^2 + \sum_{j=1}^{p} P_\gamma\Big(\sum_{t=1}^{I} P_\gamma(|w_{jt}^{[i]}|;\lambda);\lambda\Big).$$

$$\frac{\partial R}{\partial w_{jt}^{[i]}} = w_{jt}^{[i]} - u_t + \underbrace{\partial P_\gamma\Big[\sum_{t=1}^{I} P_\gamma(|\tilde{w}_{jt}^{[i]}|;\lambda);\lambda\Big] \cdot \partial P_\gamma(|\tilde{w}_{jt}^{[i]}|;\lambda) \cdot \partial(|w_{jt}^{[i]}|)}_{a}.$$

Let $\frac{\partial R}{\partial w_{jt}^{[i]}} = 0$, we have

$$0 = \begin{cases} w_{jt}^{[i]} - u_t + a, & \text{if } w_{jt}^{[i]} > 0, \\ -u_t + a \cdot [-1, 1], & \text{if } w_{jt}^{[i]} = 0, \\ w_{jt}^{[i]} - u_t - a, & \text{if } w_{jt}^{[i]} < 0. \end{cases}$$

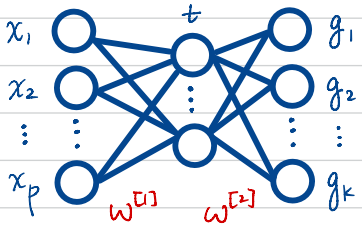
so

$$w_{jt}^{[i]} = \begin{cases} u_t - a, & u_t > a, \\ 0, & -a \le u_t \le a, \\ u_t + a, & u_t < -a. \end{cases}$$

which is $w_{jt}^{[i]} \leftarrow \mathrm{sign}(u_t) \cdot (|u_t| - a)_+$.

Part (i) is embedded in the backprop.
Part (ii) is the proximal operator.

SPINN + Integrative analysis



$$\mathcal{L} = \prod_{i=1}^{n} \prod_{\ell=1}^{k} \left[\frac{\exp(g_\ell(x_i))}{\sum_{j=1}^{k} \exp(g_j(x_i))} \right]^{1\{y_i=\ell\}}$$

$$J = -\log \mathcal{L} = -\sum_{i=1}^{n} \sum_{\ell=1}^{k} 1\{y_i=\ell\} \log \frac{\exp(g_\ell(x_i))}{\sum_{j=1}^{k} \exp(g_j(x_i))}$$

$$= -\sum_{i=1}^{n} \sum_{\ell=1}^{k} 1\{y_i=\ell\}\, g_\ell(x_i) + \sum_{i=1}^{n} \log \sum_{j=1}^{k} \exp(g_j(x_i))$$

$$\nabla_{g_{k'}} J = -\sum_{i=1}^{n} 1\{y_i=k'\} + \sum_{i=1}^{n} \frac{\exp(g_{k'}(x_i))}{\sum_{j=1}^{k} \exp(g_j(x_i))}$$

Forward: X^T $p \times \sum_m n$

$$\vec{h_t} = a(w^{[1]} X^T)$$

$t \times \sum_m n \quad t \times p \quad p \times \sum_m n$

$$\vec{g} = w^{[2]} \vec{h_t}$$

$k \times \sum_m n \quad k \times t \quad t \times \sum_m n$

Backprop:

$$\frac{\partial J}{\partial w^{[1]}} = \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot \frac{\partial g_\ell}{\partial w^{[1]}} = \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot \begin{bmatrix} — & 0 & — \\ — & \frac{\partial g_\ell}{\partial w_\ell^{[1]}} & — \\ — & \vdots & — \\ — & 0 & — \end{bmatrix} = \underset{k \times k}{\text{diag}}\left(\frac{\partial J}{\partial g_\ell}\right) \cdot \underset{k \times 1}{\mathbf{1}_k} \cdot \underset{1 \times t}{\vec{h_t}^T}$$

$t \times p$

$$\frac{\partial J}{\partial w^{[1]}} = \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot \frac{\partial g_\ell}{\partial \vec{h_t}} \cdot \frac{\partial \vec{h_t}}{\partial (w^{[1]} X^T)} \cdot \frac{\partial (w^{[1]} X^T)}{\partial w^{[1]}}$$

$$= \sum_\ell \frac{\partial J}{\partial g_\ell} \cdot w_\ell^{[2]T} * \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} \cdot X_i^T$$

$$= \begin{bmatrix} | & & | \\ w_1^{[2]T} & \cdots & w_k^{[2]T} \\ | & & | \end{bmatrix} \cdot \begin{bmatrix} \nabla_{g_1} J \\ \vdots \\ \nabla_{g_k} J \end{bmatrix} * \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} \cdot X_i^T$$

$t \times k \quad k \times 1$ (w.r.t. $w^{[1]} X^T$, $t \times 1$) $\quad 1 \times p$

$$= \underset{t \times t}{\text{diag}}\left(\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix}\right) \cdot \underset{t \times k}{w^{[2]T}} \cdot \underset{k \times 1}{\nabla_g J} \cdot \underset{1 \times pm}{X_i^T}$$

Objective:

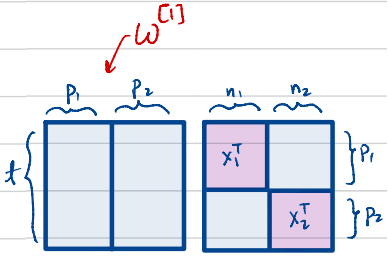
$$\min_{w} \quad \mathcal{L}(y, w) + \frac{\lambda_0}{2}\|w^{[1]}\|_2^2 + \sum_{j=1}^{p} P_\gamma\left[\sum_{m=1}^{M} \Omega_\alpha(w_{jm}^{[1]}); \lambda\right]$$

where $\Omega_\alpha(\theta) = (1-\alpha)\cdot\|\theta\|_1 + \alpha\cdot\|\theta\|_2$.

Solution:

(i) Smooth part: $\mathcal{L}(y, w) + \frac{\lambda_0}{2}\|w^{[1]}\|_2^2$

Therefore, $\frac{\partial \mathcal{L}}{\partial w^{[1]}} = \text{diag}\left(\frac{\partial J}{\partial g_t}\right)\cdot \mathbf{1}_K \cdot \vec{h}_t + \lambda_0 \cdot w_{0t}^{[1]}$.



(ii) Thresholding part:

$$R = \frac{1}{2}\sum_{j=1}^{p}\sum_{m=1}^{M}\|w_{jm}^{[1]} - \theta\|_2^2 + \sum_{j=1}^{p} P_\gamma\left[\sum_{m=1}^{M}\Omega_\alpha(w_{jm}^{[1]}); \lambda\right]$$

$$\frac{\partial R}{\partial w_{jm}^{[1]}} = w_{jm}^{[1]} - \theta + \partial P_\gamma\left[\sum_{m=1}^{M}\Omega_\alpha(\tilde{w}_{jm}^{[1]}); \lambda\right]\cdot \partial\left(\Omega_\alpha(w_{jm}^{[1]})\right)$$

($\partial P_\gamma[\cdots]$: const (var w.r.t. j))

$$= w_{jm}^{[1]} - \theta + \text{const}\cdot\left[(1-\alpha)\cdot\partial\|w_{jm}\|_1 + \alpha\cdot\partial\|w_{jm}\|_2\right].$$

Therefore, $w_j^{[1]} \leftarrow S(w_j^{[1]}, \text{const}\cdot(1-\alpha)\cdot\gamma_k)$, (learning rate)

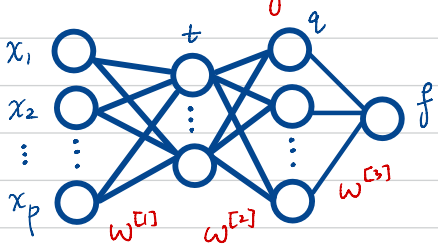
$$w_j^{[1]} \leftarrow \left(1 - \frac{\text{const}\cdot\alpha\cdot\gamma_k}{\|w_j\|_2}\right)_+ w_j^{[1]}$$

for $j = 1, \ldots, p$.

Part (i) is embedded in the backprop.

Part (ii) is the proximal operator.

SPINN + Integrative analysis (Regression, MSE)



$$\mathcal{L} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\nabla_{f_i} \mathcal{L} = -2(y_i - f(x_i)) = 2f(x_i) - 2y_i$$

$$\nabla_{\vec{f}} \mathcal{L} = (\nabla_{f_1} \mathcal{L}, \dots, \nabla_{f_n} \mathcal{L})^T \quad n \times 1$$

Forward: $X^T_{pm \times \sum_m n}$

$$\underset{t \times \sum_m n}{\vec{ht}} = a\left(\underset{t \times pm}{w^{[1]}} \underset{pm \times \sum_m n}{X^T}\right)$$

$$\underset{q \times \sum_m n}{\vec{hq}} = a\left(\underset{q \times t}{w^{[2]}} \underset{t \times \sum_m n}{\vec{ht}}\right)$$

$$\underset{1 \times \sum_m n}{f} = a\left(\underset{1 \times q}{w^{[3]}} \underset{q \times \sum_m n}{\vec{hq}}\right)$$

Integrative

Single

Forward: $X^T_{p \times n}$

$$\underset{t \times n}{\vec{ht}} = a\left(\underset{t \times p}{w^{[1]}} \underset{p \times n}{X^T}\right)$$

$$\underset{q \times n}{\vec{hq}} = a\left(\underset{q \times t}{w^{[2]}} \underset{t \times n}{\vec{ht}}\right)$$

$$\underset{1 \times n}{f} = \underset{1 \times q}{w^{[3]}} \underset{q \times n}{\vec{hq}}$$

Backprop:

$$\underset{1 \times q}{\frac{\partial \mathcal{L}}{\partial w^{[3]}}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i} \cdot \frac{\partial f_i}{\partial (w^{[3]} \vec{hq})} \cdot \frac{\partial (w^{[3]} \vec{hq})}{\partial w^{[3]}} = \underset{1 \times 1}{\frac{\partial \mathcal{L}}{\partial f_i}} \cdot \underset{1 \times 1}{[0/1]} \cdot \underset{1 \times q}{\vec{hq}^T}$$

$$\underset{q \times t}{\frac{\partial \mathcal{L}}{\partial w^{[2]}}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i} \cdot \frac{\partial f_i}{\partial (w^{[3]} \vec{hq})} \cdot \frac{\partial (w^{[3]} \vec{hq})}{\partial \vec{hq}} \cdot \frac{\partial \vec{hq}}{\partial (w^{[2]} \vec{ht})} \cdot \frac{\partial (w^{[2]} \vec{ht})}{\partial w^{[2]}}$$

$$= \sum_i \underset{1 \times 1}{\frac{\partial \mathcal{L}}{\partial f_i}} \cdot \underset{1 \times 1}{[0/1]} \cdot \underset{q \times 1}{w^{[3]T}} \cdot \underbrace{\underset{q \times 1}{\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}}}_{\text{w.r.t. } w^{[3]} \vec{hq}} * \underbrace{\underset{q \times 1}{\vec{ht}^T}}_{\text{w.r.t } w^{[3]} \vec{ht}} \quad \underset{1 \times t}{}$$

$$= \underset{q \times q}{\text{diag}\left(\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}\right)} \cdot \underset{q \times 1}{w^{[3]T}} \cdot \underset{1 \times 1}{\frac{\partial \mathcal{L}}{\partial f_i}} \cdot \underset{1 \times 1}{[0/1]} \cdot \underset{1 \times t}{\vec{ht}^T}$$

$$\underset{t \times p}{\frac{\partial \mathcal{L}}{\partial w^{[1]}}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i} \cdot \frac{\partial f_i}{\partial (w^{[3]} \vec{hq})} \cdot \frac{\partial (w^{[3]} \vec{hq})}{\partial \vec{hq}} \cdot \frac{\partial \vec{hq}}{\partial (w^{[2]} \vec{ht})} \cdot \frac{\partial (w^{[2]} \vec{ht})}{\partial \vec{ht}} \cdot \frac{\partial \vec{ht}}{\partial (w^{[1]} X_i)} \cdot \frac{\partial (w^{[1]} X_i)}{\partial w^{[1]}}$$

$$= \sum_i \underset{1 \times 1}{\frac{\partial \mathcal{L}}{\partial f_i}} \cdot \underset{1 \times 1}{[0/1]} \cdot \underset{q \times 1}{w^{[3]T}} \cdot \underbrace{\underset{q \times 1}{\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}}}_{w^{[3]} \vec{hq}} * \underbrace{\underset{q \times 1}{}}_{w^{[2]} \vec{ht}} \underset{t \times q}{w^{[2]T}} \cdot \underbrace{\underset{t \times 1}{\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}}}_{w^{[1]} X_i} \cdot \underset{1 \times p}{X_i^T}$$

$$= \sum_i \underset{t \times t}{\text{diag}\left(\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}\right)} \cdot \underset{t \times q}{w^{[2]T}} \cdot \underset{q \times q}{\text{diag}\left(\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}\right)} \cdot \underset{q \times 1}{w^{[3]T}} \cdot \underset{1 \times 1}{\frac{\partial \mathcal{L}}{\partial f_i}} \cdot \underset{1 \times 1}{[0/1]} \cdot \underset{1 \times p}{X_i^T}$$