#3 $\ell^{CE}(f,y) = -\log\left(\dfrac{\exp(f_y)}{\sum_{j=1}^{k}\exp(f_j)}\right)$, $f \in \mathbb{R}^k$, $y \in \{1,\dots,k\}$

(a) $\dfrac{\exp(f_y)}{\sum_{j=1}^{k}\exp(f_j)} = \dfrac{\exp(f_y)}{\exp(f_1)+\dots+\exp(f_y)+\dots+\exp(f_k)}$

Since $\exp(f_i) > 0$ for all $i \in \{1,\dots,k\}$,

$0 < \dfrac{\exp(f_y)}{\sum\exp(f_j)} < 1$

Also, we know that $-\infty < \log(x) < 0$ when $x \in (0,1)$

Therefore, $-\infty < \log\left(\dfrac{\exp(f_y)}{\sum\exp(f_j)}\right) < 0$ and thus, $0 < \ell^{CE}(f,y) < \infty$

(b) $\ell^{CE}(\lambda e_y, y) = -\log\left(\dfrac{\exp(\lambda)}{(k-1)+\exp(\lambda)}\right) = \log\left(\dfrac{e^\lambda + k-1}{e^\lambda}\right) = \log\left(1+\dfrac{k-1}{e^\lambda}\right) \xrightarrow{\lambda\to\infty} \log(1) = 0!$

#4 Let $f(x) = \max\{f_1(x),\dots,f_k(x)\}$. $f_i(x)$: diff. univariate fn.

pf) for a given x, maximizing index $I = \arg\max_i \{f_i(x)\}$ is unique.

Therefore, $f_I(x) > f_i(x)$ holds for all $i \neq I$

Since $\forall i$, $f_i(x)$ is differentiable, $\forall i$ $f_i(x)$ is continuous. $\therefore$ $\forall i \neq I$ $f_I(x) - f_i(x)$ is continuous.

By the definition of continuity, $f_I(x) - f_i(x) > 0$, $\exists \delta > 0$. s.t if $x-\delta < t < x+\delta$, $f_I(t) - f_i(t) > 0$ holds.

This means that $\forall t \in (x-\delta, x+\delta)$, $f_I(t) = f(t)$.

Therefore for $|h| < |\delta|$, $\lim_{h\to 0}\dfrac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\dfrac{f_I(x+h)-f_I(x)}{h}$ holds.

i.e. $f'(x) = f_I'(x)$! ㊊

#5 (a) $\sigma(z) = \max\{0, z\}$, $\sigma(\sigma(z)) = \sigma(z)$

pf) $\sigma(\sigma(z)) = \max\{0, \sigma(z)\} = \max\{0, \max\{0, z\}\} = \max\{0, 0, z\} = \max\{0, z\} = \sigma(z)$

(b) Softplus $\sigma(z) = \log(1+e^z)$, $\sigma'(z) = \dfrac{e^z}{1+e^z}$

Let's show that $\sigma(z)$ has Lipschitz continuous derivative

$$|\sigma'(x) - \sigma'(y)| = \left| \frac{e^x}{1+e^x} - \frac{e^y}{1+e^y} \right| = \left| \frac{e^x + e^x e^y - e^y - e^x e^y}{(1+e^x)(1+e^y)} \right| = \left| \frac{e^x - e^y}{(1+e^x)(1+e^y)} \right|$$

$$= \left| \frac{e^c}{(1+e^x)(1+e^y)} \right| |x-y| \quad \exists c,\ c \in (\min(x,y), \max(x,y)) \text{ by Mean Value Theorem}$$

Since $c < \max(x,y)$, $e^c < 1+e^x$ or $e^c < 1+e^y$ holds.

$\therefore |\sigma'(x) - \sigma'(y)| = \left| \dfrac{e^c}{(1+e^x)(1+e^y)} \right| |x-y| < |x-y|$ $\therefore$ Softplus has Lipschitz continuous derivative.

On the other hand, Consider ReLU, $\sigma(z) = \max(0, z)$

$$\frac{|\sigma'(z) - \sigma'(-z)|}{|z-(-z)|} = \frac{\pm 1}{2|z|} \to \infty \text{ as } z \to 0$$

$\therefore \forall L, \exists \delta > 0$ s.t if $0 < |z| < \delta$, $|\sigma'(z) - \sigma'(-z)| > L|z-(-z)|$. $\therefore$ ReLU does not have Lipschitz continuous derivative!

(c) $$\rho(z) = \frac{1-e^{-2z}}{1+e^{-2z}} = \frac{e^{2z}-1}{e^{2z}+1} = \frac{2e^{2z} - (1+e^{2z})}{1+e^{2z}} = 2\frac{1}{1+e^{-2z}} - 1 = 2\sigma(2z) - 1.$$

$L > 1$, $A_1 \ldots, A_L$, $b_1 \ldots b_L$ 이 주어져서, $y_L = A_L y_{L-1} + b_L, \ldots, y_1 = \sigma(A_1 x + b_1)$ 이라 하자.

① $$y_L = A_L \sigma(A_{L-1} y_{L-2} + b_{L-1}) + b_L = C_L \rho(C_{L-1} y_{L-2} + d_{L-1}) + d_L$$

$$= C_L\left(2\sigma\left(\boxed{2C_{L-1} y_{L-2} + 2d_{L-1}}\right)\right) + d_L - C_L\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

$\Rightarrow$ $\boxed{C_L = \frac{1}{2}A_L,\ d_L = b_L + \frac{1}{2}A_L\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}}$

$$A_{L-1}\sigma(A_{L-2} y_{L-3} + b_{L-2}) + b_{L-1} = 2C_{L-1}\rho(C_{L-2} y_{L-3} + d_{L-2}) + 2d_{L-1}$$

$$= 2C_{L-1}\left(2\sigma(2C_{L-2} y_{L-3} + 2d_{L-2}) - \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\right) + 2d_{L-1}$$

$$= 4C_{L-1}\sigma\left(\boxed{2C_{L-2} y_{L-3} + 2d_{L-2}}\right) + 2d_{L-1} - 2C_{L-1}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

$\Rightarrow$ $C_{L-1} = \frac{1}{4}A_{L-1}$, $d_{L-1} = \frac{1}{2}\left(b_{L-1} + \frac{1}{2}A_{L-1}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\right)$

$= \frac{1}{2}b_{L-1} + \frac{1}{4}A_{L-1}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$

같은 꼴의 반복!

② $\boxed{C_i = \frac{1}{4}A_i,\ d_i = \frac{1}{2}b_i + \frac{1}{4}A_i\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}}$ (1로 구성된 $\mathbb{R}^i$ vector) $(1 < i < L)$

③ $\boxed{C_1 = \frac{1}{2}A_1,\ d_1 = \frac{1}{2}b_1}$

임의의 A_i, b_i 에 대하여, 위와 같이 C_i, d_i 를 잡으면 두 MLP는 equivalent 하다!

pf) <i> L=2 인 경우

$y_2 = A_2 y_1 + b_2$ $\qquad$ $\bar{y}_2 = C_2 \bar{y}_1 + d_2 = \frac{1}{2} A_2 \bar{y}_1 + b_2 + \frac{1}{2} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$y_1 = \sigma(A_1 x + b_1)$ $\qquad$ $\bar{y}_1 = \rho(C_1 x + d_1) = \rho\left(\frac{1}{2} A_1 + \frac{1}{2} b_1\right)$

$$\bar{y}_2 = \frac{1}{2} A_2 \left(2\sigma(A_1 x + b_1) - \cancel{\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} \right) + b_2 + \cancel{\frac{1}{2} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}}$$

$$= A_2 \sigma(A_1 x + b_1) + b_2 = y_2.$$

<ii> L=3 인 경우

$y_3 = A_2 y_2 + b_3$ $\qquad$ $\bar{y}_3 = \frac{1}{2} A_3 \bar{y}_2 + b_3 + \frac{1}{2} A_3 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$y_2 = \sigma(A_2 y_1 + b_2)$ $\qquad$ $\bar{y}_2 = \rho\left(\frac{1}{4} A_2 \bar{y}_1 + \frac{1}{2} b_2 + \frac{1}{4} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right)$

$y_1 = \sigma(A_1 x + b_1)$ $\qquad$ $\bar{y}_1 = \rho\left(\frac{1}{2} A_1 x + \frac{1}{2} b_1\right) = 2\sigma(A_1 x + b_1) - \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$$\bar{y}_2 = \rho\left(\frac{2}{4} A_2 \sigma(A_1 x + b_1) - \cancel{\frac{1}{4} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} + \frac{1}{2} b_2 + \cancel{\frac{1}{4} A_2 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} \right)$$

$$= 2\sigma\left(A_2 \sigma(A_1 x + b_1) + b_2 \right) - \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

$$\bar{y}_3 = A_3 \sigma\left(A_2 \sigma(A_1 x + b_1) + b_2 \right) + b_3 = y_3 \text{ !}$$

<iii> L 일때 성립 가정.

즉 $y_L = A_L y_{L-1} + b_L = \bar{y}_L = \frac{1}{2} A_L \bar{y}_{L-1} + b_L + \frac{1}{2} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$ 인 상황이라 하자. ····(*)

C_L은 $\frac{1}{4} A_L$로, d_L은 $\frac{1}{2} b_L + \frac{1}{4} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$로 바꿀때 $y_{L+1} = \bar{y}_{L+1}$ 임을 보임으로서 L+1 일때 성립함을 보이자.

$$\cancel{A_{L+1}} \sigma(A_L y_{L-1} + b_L) + \cancel{b_{L+1}} = \frac{1}{2} A_{L+1} \rho\left(\frac{1}{4} A_L \bar{y}_{L-1} + \frac{1}{2} b_L + \frac{1}{4} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \right) + b_{L+1} + \frac{1}{2} A_{L+1} \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

$$= \cancel{A_{L+1}} \sigma\left(\frac{1}{2} A_L \bar{y}_{L-1} + b_L + \frac{1}{2} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \right) + \cancel{b_{L+1}}$$

$\Leftrightarrow A_L y_{L-1} + b_L = \frac{1}{2} A_L \bar{y}_{L-1} + b_L + \frac{1}{2} A_L \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$ ··· (*)와 동일한 식! 즉 성립한다!

즉 <i>, <ii>, <iii>, M.I 에 의하여 전 page에 주는대로 C_l, d_l를 잡으면 동일한 MLP를 얻게 된다는 사실을 엄밀히 증명하였다.

#6. $a_j^0 X_i + b_j^0 < 0 \quad \forall i \in \{1, \dots, N\}$. $\quad \sigma$: ReLU

We want to show that $a_j^k X_i + b_j^k < 0$ for all i and k (to show that jth ReLU output remains dead)

$$\frac{\partial}{\partial a_j} l(f_\theta(x_i), Y_i) = \frac{\partial l}{\partial f}\frac{\partial f}{\partial a_j} = \frac{\partial l}{\partial f}\frac{\partial u^T \sigma(a x_i + b)}{\partial a_j} = \frac{\partial l}{\partial f} u_j \cdot \sigma'(a_j x_i + b_j)\, x_i$$

$$= \begin{cases} 0 & \text{if } a_j X_i + b_j < 0 \\ \frac{\partial l}{\partial f} u_j X_i & \text{if } a_j X_i + b_j > 0 \end{cases}$$

$$\frac{\partial}{\partial b_j} l(f_\theta(x_i), Y_i) = \frac{\partial l}{\partial f}\frac{\partial f}{\partial b_j} = \frac{\partial l}{\partial f}\frac{\partial u^T \sigma(a x_i + b)}{\partial b_j} = \frac{\partial l}{\partial f} u_j \sigma'(a_j X_i + b_j)$$

$$= \begin{cases} 0 & \text{if } a_j X_i + b_j < 0 \\ \frac{\partial l}{\partial f} u_j & \text{if } a_j X_i + b_j > 0 \end{cases}$$

If we think about SGD, $\theta_{k+1} = \theta_k - lr \times \frac{\partial l}{\partial \theta_k}$.

Since <i> $a_j^0 X_i + b_j^0 < 0$ ($\forall i$),

<ii> $a_j^k X_i + b_j^k < 0$. Then, $\frac{\partial}{\partial a_j} l = 0$ and $\frac{\partial l}{\partial b_j} = 0$ for all i.

Thus, a_j^{k+1}, b_j^{k+1} remains unchanged and thus $a_j^{k+1} X_i + b_j^{k+1} < 0$ for all i.

By Mathematical induction, <i>, <ii>, we proved that $\forall i, k$, $a_j^k X_i + b_j^k < 0$ holds.

The jth ReLU output is dead throughout training.

#7 If we use Leaky ReLU, $\sigma(z) = \begin{cases} z & z>0 \\ \alpha z & z\le 0 \end{cases}$

$$\frac{\partial l(f_\theta(x_i), Y_i)}{\partial a_j} = \begin{cases} \left(\frac{\partial l}{\partial f} u_j x_i\right)\alpha & (a_j x_i + b_j < 0) \\ \left(\frac{\partial l}{\partial f} u_j x_i\right) & (a_j x_i + b_j > 0) \end{cases} \qquad \frac{\partial l(f_\theta(x_i), Y_i)}{\partial b_j} = \begin{cases} \left(\frac{\partial l}{\partial f} u_j\right)\alpha & (a_j x_i + b_j < 0) \\ \frac{\partial l}{\partial f} u_j & (a_j x_i + b_j > 0) \end{cases}$$

$\alpha \neq 0$ 이라면, 비록 $a_j^0 x_i + b_j^0 < 0$ 더라도

$\frac{\partial l}{\partial a_j} = \left(\frac{\partial l}{\partial f} u_j x_i\right)\alpha$: not identically zero

$\frac{\partial l}{\partial b_j} = \left(\frac{\partial l}{\partial f} u_j\right)\alpha$: not identically zero 이므로,

SGD 에 의해 $a_j^{k+1} = a_j^k - \frac{\partial l}{\partial a_j} \times \text{learning\_rate}$

$b_j^{k+1} = b_j^k - \frac{\partial l}{\partial b_j} \times \text{learning\_rate}$ 와 같이 update 되고,

#6 번 문제와 같이 gradient 가 "exactly vanish" 하지는 않는다!