3.

(a) $0 < \exp(f_y) < \sum_{j=1}^{K} \exp(f_j) < \infty$

$\Rightarrow \quad 0 < \dfrac{\exp(f_y)}{\sum_{j=1}^{K} \exp(f_j)} < 1$

$\Rightarrow \quad 0 < -\log\left(\dfrac{\exp(f_y)}{\sum_{j=1}^{K} \exp(f_j)}\right) < \infty$

(b) $\ell^{CE}(\lambda e_y, y) = -\log\left(\dfrac{\exp(\lambda)}{\exp(\lambda) + (K-1)\exp(0)}\right)$

$= -\log \dfrac{e^{\lambda}}{e^{\lambda} + (K-1)} \quad \xrightarrow[\text{as } \lambda \to \infty]{} \quad 0$

Since $\dfrac{e^{\lambda}}{e^{\lambda} + (K-1)} \xrightarrow[\text{as } \lambda \to \infty]{} 1$

□

4. Let $x \in \mathbb{R}$ be given and $I = \operatorname{argmax}_i \{f_i(x)\}$ is unique.

Since $\mathbb{R}$ is compact,

$\{y \in \mathbb{R} : f_I(y) > f_n(y), \; n = \{1, 2, \cdots K\} \setminus \{I\}\}$ is open.

$\therefore \quad \exists \delta > 0$ s.t.

$\forall y \in \mathbb{R}, \; \|x - y\| < \delta \quad \Rightarrow \quad f_I(y) > f_n(y), \quad n = \{1, 2, \cdots K\} \setminus \{I\}$

$\Leftrightarrow \quad f(y) = f_I(y)$

$\therefore \forall y \in \mathbb{R}, \; \|x - y\| < \delta \quad \Rightarrow \quad f'(y) = f'_I(y)$

Hence, $f'(x) = f'_I(x)$.

□

5.

(a) i) The case $z \geq 0$

$$\sigma(\sigma(z)) = \sigma(z) = z$$

ii) The case $z < 0$

$$\sigma(\sigma(z)) = \sigma(0) = 0 = \sigma(z)$$

$\therefore \forall z \in \mathbb{R}, \quad \sigma(\sigma(z)) = \sigma(z)$

(b) $\sigma(z) = \log(1 + e^z)$

$$\sigma'(z) = \frac{e^z}{1 + e^z}$$

$$\sigma''(z) = \frac{e^z}{(1+e^z)^2}$$

$\forall x, y \in \mathbb{R} \quad \exists t \in (x, y) \quad s.t.$

$x \neq y$

$$\frac{\sigma'(x) - \sigma'(y)}{x - y} = \sigma''(t) = \frac{e^t}{(1+e^t)^2} < 1$$

$$\therefore |\sigma'(x) - \sigma'(y)| < |x - y|$$

$\therefore \sigma(z)$ has Lipschitz continuous derivatives.

On the other hand, ReLU does not

since ReLU is non-differentiable at $z = 0$.

(c) $\rho(z) = 2\sigma(2z) - \underline{1}$

$$\therefore y_1 = \rho(C_1 x + d_1) = 2\sigma(\underbrace{2C_1}_{C_1'} x + d_1) - \underline{1} = 2\sigma(C_1' x + d_1) - \underline{1}$$

$$y_2 = \rho(C_2 y_1 + d_2) = 2\sigma(\underbrace{2C_2}_{C_2'} \underbrace{\sigma(C_1' x + d_1)}_{y_1'} + \underbrace{(d_2 - C_2 \underline{1})}_{d_2'}) - \underline{1}$$

$$= 2\sigma(C_2' y_1' + d_2') - \underline{1}$$

$$\vdots$$

$$y_{L-1} = 2\sigma(C_{L-1}' y_{L-2}' + d_{L-1}') - \underline{1}$$

$$y_L = C_L y_{L-1} + d_L = 2C_L \sigma(C_{L-1}' y_{L-2}' + d_{L-1}') + d_L - C_L \underline{1} = C_L' \sigma(y_{L-1}') + d_L'$$

By initializing $\begin{cases} C_l' = 2C_l = A_l \in \mathbb{R}^{n_l \times n_{l-1}} \\ d_l' = d_l - C_l \underline{1} = b_l \in \mathbb{R}^{n_l} \end{cases} \quad (l = 1, \cdots, L)$,

y_l' be equivalent with y_l in MLPs built with sigmoid activations.

$(l = 1, \cdots, L)$

□

6. $x \in \mathbb{R} \setminus \{0\} \Rightarrow \begin{cases} \nabla_b f_\theta(x) = \sigma'(ax+b) \odot u \\ \nabla_a f_\theta(x) = (\sigma'(ax+b) \odot u)\, x \end{cases}$

$a_j^k x_i + b_j^k < 0 \Rightarrow \sigma'(a_j^k x_i + b_j^k) = 0$

$\Rightarrow \begin{cases} \nabla_{b_j} f_\theta(x) = 0 \\ \nabla_{a_j} f_\theta(x) = 0 \end{cases}$

$\Rightarrow \begin{cases} \nabla_{b_j} \ell(f_\theta(x), y_i) = \nabla_{b_j} f_\theta(x)\ \nabla_{f_\theta(x)} \ell(f_\theta(x), y_i) = 0 \\ \nabla_{a_j} \ell(f_\theta(x), y_i) = \nabla_{a_j} f_\theta(x)\ \nabla_{f_\theta(x)} \ell(f_\theta(x), y_i) = 0 \end{cases}$

$\therefore$ If $a_j^0 x_i + b_j^0 < 0$ for all i at initialization,

then a_j^k, b_j^k will never update with SGD and

$a_j^k x_i + b_j^k$ remains negative for all $k = 1, 2, \cdots$; i,

which result in $\sigma(a_j^k x_i + b_j^k) = 0$ throughout the training.

□

7. $x \in \mathbb{R} \setminus \{0\} \Rightarrow$

$$\begin{cases} \nabla_u f_\theta(x) = \sigma(ax+b) \\ \nabla_b f_\theta(x) = \sigma'(ax+b) \odot u \\ \nabla_a f_\theta(x) = (\sigma'(ax+b) \odot u)\, x \end{cases}$$

$a_j^{\,k} x_i + b_j^{\,k} < 0 \Rightarrow$

$$\begin{cases} \sigma(a_j^{\,k} x_i + b_j^{\,k}) = 0.01(a_j^{\,k} x_i + b_j^{\,k}) \\ \sigma'(a_j^{\,k} x_i + b_j^{\,k}) = 0.01 \end{cases}$$

$\Rightarrow$

$$\begin{cases} \nabla_{u_j} f_\theta(x) = 0.01(a_j^{\,k} x + b_j^{\,k}) \\ \nabla_{b_j} f_\theta(x) = 0.01\, u_j \\ \nabla_{a_j} f_\theta(x) = 0.01\, u_j x \end{cases}$$

$\Rightarrow$

$$\nabla_{b_j} \ell(f_\theta(x), y_i) = \nabla_{b_j} f_\theta(x)\ \nabla_{f_\theta(x)} \ell(f_\theta(x), y_i)$$

$$\nabla_{a_j} \ell(f_\theta(x), y_i) = \nabla_{a_j} f_\theta(x)\ \nabla_{f_\theta(x)} \ell(f_\theta(x), y_i)$$

both are not identically zero

with proper loss function.

$\therefore$ a_j, b_j can be updated with SGD,

hence $\sigma(a_j^{\,k} x_i + b_j^{\,k})$ can be nonzero with some i, k.

□