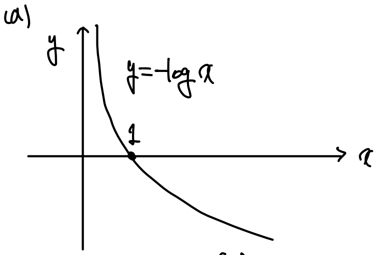
#3.

(a)



This is the graph of $f(x) = -\log x$

$\forall x \in (0,1)$, $0 < f(x) < \infty$ since $f'(x) = -\frac{1}{x} < 0$

f is strictly decreasing, and $f(1) = 0$.

$\therefore f(x) > 0$ for $x \in (0.1)$

Since $0 < \dfrac{\exp(f_y)}{\sum_{j=1}^{k} \exp(f_j)} < 1$, $0 < \ell^{CE}(f, y) < \infty$.

$\blacksquare$

(b) $\ell^{CE}(\lambda e_y, y) = -\log\left(\dfrac{\exp(\lambda)}{\exp(\lambda)}\right) = 0$ for $\forall \lambda$.

$\therefore \ell^{CE}(\lambda e_y, y) \to 0$ as $\lambda \to \infty$.

$\blacksquare$

#4.

Since I is unique at $x \in \mathbb{R}$, $\exists \delta > 0$ s.t. $\forall t \in (x-\delta, x+\delta)$, $\underset{i \in [1,N]}{\operatorname{argmax}} \{f_i(t)\} = I$.

pf) Let's assume that $\forall \delta > 0$, $\exists t \in (x-\delta, x+\delta)$ s.t. $\underset{i \in [1,N]}{\operatorname{argmax}} \{f_i(t)\} = J \neq I$

$f_I(t) < f_J(t)$, and take $\delta \to 0$. Then $t \to x$, and since f_I and f_J are continuous, $\lim_{t \to x} f_I(t) \le \lim_{t \to x} f_J(t)$ means $f_I(x) \le f_J(x)$ which is contradiction to I being the argmax of $\{f_i(x)\}$ i.e. $f_I(x) > f_J(x)$. $\blacksquare$.

Therefore, $\lim_{h \to 0} \dfrac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \dfrac{f_I(x+h) - f_I(x)}{h} = f_I'(x)$.

(Send $h < \delta$).

$\blacksquare$

#5.

(a) ① $z \ge 0$: $\sigma(\sigma(z)) = \sigma(z)$

② $z < 0$: $\sigma(\sigma(z)) = \sigma(0) = 0 = \sigma(z)$ $\blacksquare$

(b) $\sigma'(z) = \dfrac{e^z}{1+e^z}$, $\sigma''(z) = \dfrac{e^z}{(1+e^z)^2} \le 1$

$\forall x, y \in \mathbb{R}$, $|\sigma'(x) - \sigma'(y)| = |\sigma''(c)||x-y| \le |x-y|$ $\Rightarrow$ Lipschitz continuous.

($c \in (x,y)$, by MVT)

Relu: first, the derivative is not well defined, but if we define it as follows,

$\text{Relu}'(x) = \begin{cases} 0 & (x<0) \\ 1 & (x \ge 0) \end{cases}$, this isn't Lipschitz since if we assume it is for $\exists K > 0$,

Let $\varepsilon = \lfloor \frac{1}{4K} \rfloor$, and take $x_1 = -\varepsilon$, $x_2 = \varepsilon$

$\Rightarrow |f(x_1) - f(x_2)| = 1 \le K \cdot 2\varepsilon < 1$ $\to$ contradiction, not Lipschitz

$\blacksquare$

(c) ($\Rightarrow$)

Given $A_1, \cdots, A_L, b_1, \cdots, b_L$, $y_1 \sim y_L$ are the states of the MLP with sigmoid.

$\sigma(z) = \frac{1}{2}\rho(\frac{z}{2}) + \frac{1}{2}$

Set $C_1 = \frac{1}{2}A_1$, $d_1 = \frac{1}{2}b_1$, then $y_1' := \rho(C_1 x + d_1) = 2y_1 - v_1$ $\quad (v_1 = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n_1})$

Set $C_2 = \frac{1}{4}A_2$, $d_2 = \frac{1}{4}A_2 v_1 + \frac{1}{2}b_2$, then $y_2' := \rho(C_2 y_1' + d_2) = 2y_2 - v_2$ $\quad (v_2 = (1, \cdots, 1)^T \in \mathbb{R}^{n_2})$

$\vdots$ $C_i = \frac{1}{4}A_i$, $d_i = \frac{1}{4}A_i v_{i-1} + \frac{1}{2}b_i$, $(2 \leq i \leq L-1)$ $\quad (v_i \in (1, \cdots, 1)^T \in \mathbb{R}^{n_i})$

then $y_{L-1}' := \rho(C_{L-1} y_{L-2}' + d_{L-1}) = 2y_{L-1} - v_{L-1}$.

Set $C_L = \frac{1}{2}A_L$, $d_L = \frac{1}{2}A_L v_{L-1} + b_L$

$\Rightarrow y_L' := \frac{1}{2}A_L(2y_{L-1} - v_{L-1}) + \frac{1}{2}A_L v_{L-1} + b_L = A_L y_{L-1} + b_L = y_L$ $\Rightarrow$ this new MLP with tanh represent the given MLP $\blacksquare$

($\Leftarrow$)

Given $C_1 \sim C_L, d_1 \sim d_L$, and states $y_1 \sim y_L$ with MLP with tanh, $\quad$ (v_i's are defined same as above)

$\rho(z) = 2\sigma(2z) - 1$

Set $A_1 = 2C_1$, $b_1 = 2d_1$, then $y_1' := \sigma(A_1 x + b_1) = \frac{1}{2}(y_1 + v_1)$.

Set $A_2 = 4C_2$, $b_2 = 2d_2 - 2C_2 v_1$, then $y_2' := \sigma(A_2 x + b_2) = \frac{1}{2}(y_2 + v_2)$

$\vdots$ Set $A_i = 4C_i$, $b_i = 2d_i - 2C_i v_{i-1}$, $(2 \leq i \leq L-1)$

then $y_{L-1}' := \frac{1}{2}(y_{L-1} + v_{L-1})$

Set $A_L = 2C_L$, $b_L = d_L - C_L v_{L-1}$, then $y_L' := C_L y_{L-1} + d_L = y_L$ $\Rightarrow$ this new MLP with sigmoid represent the given MLP $\blacksquare$

#6

In the initialization step, if $a_j^0 X_i + b_j^0 < 0$ for $\forall i \in [1, N]$

$$\frac{\partial}{\partial a_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \cdot \frac{\partial}{\partial a_j} f_\theta(X_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \cdot u_j \sigma'(a_j X_i + b_j) \cdot X_i$$

$$\frac{\partial}{\partial b_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \cdot u_j \sigma'(a_j X_i + b_j)$$

Consider the k^{th} iteration of the SGD.

when $k=1 \Rightarrow$ since $a_j^0 X_i + b_j^0 < 0$, $\sigma'(a_j^0 X_i + b_j^0) = 0$, and $\frac{\partial}{\partial a_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial b_j} l(f_\theta(X_i), Y_i) = 0$

$\Rightarrow a_j^1 X_i + b_j^1 < 0$.

Assume $a_j^k X_i + b_j^k < 0 \Rightarrow a_j^{k+1} = a_j^k$, $b_j^{k+1} = b_j^k$ since the gradients are 0 vectors.

$\therefore$ By mathematical induction, $\forall k$, $a_j^k X_i + b_j^k < 0$, hence the gradient vanishes.

#7.

$$\frac{\partial}{\partial a_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \cdot u_j \sigma'(a_j X_i + b_j) \cdot X_i$$

$$\frac{\partial}{\partial b_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \cdot u_j \sigma'(a_j X_i + b_j)$$

But, $\sigma'(a_j^0 X_i + b_j^0) = \alpha$, and $\frac{\partial}{\partial a_j} l(f_\theta(X_i), Y_i) = \frac{\partial}{\partial f_\theta} l(f_\theta(X_i), Y_i) \cdot (\alpha u_j X_i) \neq 0$, and similar for $\frac{\partial l}{\partial b_j}$.

Therefore, there are SGD updates to the parameters a_j and b_j, and the gradient no longer vanishes. ■