MLPH

Sheet 5.

1. ~~(a)~~ Logistic sigmoid

(a) $\sigma(x) = \frac{1}{1+e^{-x}}$,

$$\sigma'(x) = \frac{-1}{(1+e^{-x})^2} \cdot (-e^{-x}) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x)(1-\sigma(x)).$$

(b) $$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{1}{1+e^{-2x}} - \frac{e^{-2x}}{1+e^{-2x}}$$

$$= \sigma(2x) - (1-\sigma(2x)) = 2\sigma(2x) - 1.$$

(c) choose: $\vec{w} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$, $b = \cancel{0.5}\,1$, then we have:

$$\vec{w}^T\vec{x} + b = \begin{cases} \cancel{0.5}\,1, & \vec{x} = (1,1) \text{ or } (2,2). \\ \cancel{-0.5}\,-1, & \vec{x} = (1,2) \text{ or } (2,3) \end{cases}$$

$$\sigma(\vec{w}^T\vec{x} + b) = \begin{cases} 0.269, & \vec{x} = (1,2) \text{ or } (2,3) \\ 0.731, & \vec{x} = (1,1) \text{ or } (2,2) \end{cases}$$

3. ~~(a)~~ Log-Sum-EXp and soft(arg)max.

(a) on subset $\{\vec{\sigma}^1 = (1,2,3)^T, \vec{\sigma}^2 = (11,12,13)^T\}$.

(i) constant offset b:

$$\text{softmax}(\vec{\sigma}+b;\lambda)_i = \frac{e^{\lambda(\sigma_i+b)}}{\sum_{j=1}^{K} e^{\lambda(\sigma_j+b)}} = \frac{\exp(\lambda\sigma_i)\,\cancel{e^{\lambda b}}}{\sum_{j=1}^{K}\exp(\lambda\sigma_j)\,\cancel{e^{\lambda b}}}$$

$$= \text{softmax}(\vec{\sigma};\lambda)_i. \qquad \text{invariant.}$$

(ii) rescaling factor a:

$$\text{softmax}(a\vec{\sigma};\lambda)_i = \frac{\exp(\lambda a\sigma_i)}{\sum_{j=1}^{K}\exp(\lambda a\sigma_j)} = \text{softmax}(\vec{\sigma};\lambda a)_i$$

$$\neq \text{softmax}(\vec{\sigma},\lambda)_i. \qquad \text{not invariant.}$$

(d) $$\frac{d\,\text{lse}(\vec{\sigma},\lambda)}{d\sigma_i} = \frac{1}{\lambda}\cdot\frac{1}{\sum_{j=1}^{K}\exp(\lambda\sigma_j)}\cdot\lambda\exp(\lambda\sigma_i) = \frac{\exp(\lambda\sigma_i)}{\sum_{j=1}^{K}\exp(\lambda\sigma_j)}$$

$$= \text{softmax}(\vec{\sigma},\lambda)_i.$$

3.

(e). rearrange $\{\sigma_1, \sigma_2 \cdots, \sigma_k\}$ to $\{b_1, b_2 \cdots, b_k\}$,

s.t. $b_1 \geq b_2 \geq \cdots \geq b_k$.

$$S := \sum_{j=1}^{k} e^{\lambda b_j} = e^{\lambda b_1}\left(1 + \sum_{j=2}^{k} e^{\lambda(b_j - b_1)}\right).$$

~~($\forall \lambda > 0$)~~

since $b_j \leq b_1$, or $b_j - b_1 \leq 0$, we have ~~with~~ λ

$\forall \lambda > 0$, $0 \leq e^{\lambda(b_j - b_1)} \leq 1$.

therefore:

$$e^{\lambda b_1} \leq S \leq k e^{\lambda b_1}$$

$$b_1 \leq \frac{1}{\lambda}\log S \leq b_1 + \frac{1}{\lambda}\log k.$$

Because $\lim_{\lambda \to \infty} b_1 = b_1$, $\lim_{\lambda \to \infty} b_1 + \frac{1}{\lambda}\log k = b_1$,

we have (according to Squeeze Theorem)

$$\lim_{\lambda \to \infty} lse(\vec{\sigma}; \lambda) = \lim_{\lambda \to \infty} \frac{1}{\lambda}\log S = b_1 = \max(\vec{\sigma})$$