Machine Learning and Physics sheet 05

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

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

$$= \cancel{\frac{1}{2(\cosh(x))^2}}$$

$$= \frac{1}{4\cosh\left(\frac{x}{2}\right)^2}$$

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

we show that $2\sigma(2x) - 1 = \tanh(x)$

$$2\sigma(2x)-1 = 2\left(\frac{1}{1+e^{-2x}}\right)-1$$

$$= \frac{2}{1+e^{-2x}} - \frac{1+e^{-2x}}{1+e^{-2x}}$$

$$= \frac{1-e^{-2x}}{1+e^{-2x}}$$

$$= \frac{1-e^{-2x}}{1+e^{-2x}} \cdot \frac{e^{x}}{e^{x}}$$

$$= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

$$= \tanh(x)$$

c) We want to have $\sigma(w^T x_1 + b) < \frac{1}{2}$ with $x_1 = (1,1)$

$\sigma(w^T x_2 + b) < 1/2$ $\quad x_2 = (2,2)$

$\sigma(w^T x_3 + b) > 1/2$ $\quad x_3 = (2,3)$

$\sigma(w^T (x_4 + b)) > 1/2$ $\quad x_4 = (1,2)$

So $w^T x_1 + b < 0$, etc.

This is the case for $w = \begin{pmatrix}-1\\1\end{pmatrix}$, $b = -1/2$, so

$\sigma\left((-1,1)\begin{pmatrix}1\\1\end{pmatrix} - \frac{1}{2}\right) < 1/2$, since $\sigma(-1/2) < 1/2$

$\sigma\left((-1,1)\begin{pmatrix}2\\2\end{pmatrix} - \frac{1}{2}\right) < 1/2$

$\sigma\left((-2,1)\begin{pmatrix}2\\3\end{pmatrix} - \frac{1}{2}\right) > 1/2$, since $\sigma(1 - 1/2) = \sigma(1/2) > 1/2$

$\sigma\left((-1,1)\begin{pmatrix}1\\2\end{pmatrix} - (1/2)\right) > 1/2$

③

a) constant offset:

$$\mathrm{softmax}(\sigma + c ; \lambda) = \frac{\exp(\lambda(\sigma_k + c_k))}{\sum_{j=1}^{k} \exp(\lambda(\sigma_j + c_j))}$$

$$\overset{c_j = c_k \, \forall j}{=} \frac{e^{\lambda c_k}}{e^{\lambda c_k}} \frac{\exp(\lambda \sigma_k)}{\sum_{j=1}^{k} \exp(\lambda \sigma_j)}$$

$$= \mathrm{softmax}(\sigma, \lambda) \quad \text{for} \quad c = \begin{pmatrix} c \\ c \\ c \\ \vdots \\ c \end{pmatrix} \in \mathbb{R}^k$$

rescaling input:

$$\mathrm{softmax}(c \cdot \sigma ; \lambda) = \frac{\exp(\lambda c \sigma_k)}{\sum_{j=1}^{k} \exp(\lambda(\sigma_j \cdot c))}$$

$$= \frac{\exp(\lambda \sigma_k)^c}{\sum_{j=1}^{k} \exp(\lambda \sigma_j)^c}$$

$$\neq \mathrm{softmax}(\sigma, \lambda)$$

In general the softmax shows identical results only for a constant offset, not for rescaling. We thus follow that σ_1 and σ_2 yield identical results.

d) Derivative after kth component:

$$\frac{\partial}{\partial \sigma_k} \mathrm{lse}(\sigma, \lambda) = \frac{\partial}{\partial \sigma_k} \frac{1}{\lambda} \log\left(\sum_{j=1}^{k} \exp(\lambda \sigma_j)\right)$$

$$= \frac{\lambda}{\lambda} \frac{\exp(\lambda \sigma_k)}{\sum_{j=1}^{k} \exp(\lambda \sigma_j)}$$

$$= \mathrm{softmax}(\sigma, \lambda)_k$$

e) We prove the statement by showing the inequality in both directions.

$$\lim_{\lambda \to \infty} \mathrm{lse}(\sigma, \lambda) = \lim_{\lambda \to \infty} \frac{1}{\lambda} \log\left(\sum_{j=1}^{k} \exp(\lambda \sigma_j)\right) \underset{\text{log monotone}}{\geq} \lim_{\lambda \to \infty} \frac{1}{\lambda} \log\left(e^{\lambda \sigma_{max}}\right) = \sigma_{max} = \max(\sigma)$$

$$\lim_{\lambda \to \infty} \mathrm{lse}(\sigma, \lambda) = \lim_{\lambda \to \infty} \frac{1}{\lambda} \log\left(\sum_{j=1}^{k} \exp(\lambda \sigma_j)\right) \leq \lim_{\lambda \to \infty} \frac{1}{\lambda} \log\left(k \cdot e^{\sigma_{max} \lambda}\right) = \lim_{\lambda \to \infty} \frac{\log(k)}{\lambda} + \sigma_{max} \xrightarrow[\lambda \to \infty]{} \max(\sigma)$$

$\square$