Exercises
Deep Learning
Fall 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Fernando Pérez-Cruz
Web http://www.da.inf.ethz.ch/teaching/2018/DeepLearning/

Series Monday, Oct 8, 2018
(Deep Learning, Exercise series 2 - solutions)

Solution 1 (Activation functions):

$$\begin{aligned}
\nabla s(x) &= -\frac{1}{(1+e^{-x})^2} \cdot \nabla(1+e^{-x}) \\
&= -\frac{-e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}}\left(\frac{e^{-x}}{1+e^{-x}}\right) \\
&= \frac{1}{1+e^{-x}}\left(\frac{(1+e^{-x})-1}{1+e^{-x}}\right) \\
&= \frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right)
\end{aligned}$$

$$\begin{aligned}
\nabla_x \mathsf{ReLU}(x) &= \nabla_x \max(0,x) = \begin{cases} 1 & x > 0 \\ 0 & \text{else} \end{cases} \\
\nabla_x \mathsf{tanh}(x) &= \nabla_x \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - \mathsf{tanh}^2(x) \\
\nabla_{x_j} \mathsf{softmax}(x)_i &= \nabla_{x_j} \frac{e^{x_i}}{\sum_k e^{x_k}} = \begin{cases} \mathsf{softmax}(x)_i(1-\mathsf{softmax}(x)_i) & i = j \\ -\mathsf{softmax}(x)_i \mathsf{softmax}(x)_j & i \neq j \end{cases} \\
\nabla_{x_j} \mathsf{tanh}(\mathsf{softmax}(x)_i) &= \frac{\partial}{\partial y}\mathsf{tanh}(y)\nabla_{x_j}\mathsf{softmax}(x)_i \quad \text{where } y = \mathsf{softmax}(x)_i
\end{aligned}$$

Solution 2 (Cross-entropy):

$$\begin{aligned}
\log \mathcal{L}(\boldsymbol{w}) &= \log(\prod_{i=1}^{n} \Pr(x_i)^{y_i}(1-\Pr(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} y_i \log \Pr(x_i) + (1-y_i)\log(1-\Pr(x_i)) \\
&= -H(\boldsymbol{w}) \qquad (1)
\end{aligned}$$

Solution 3 (Finite differences):

(1) Using Taylor expansion, we get

$$f(w_i + \epsilon) = f(w_i) + \epsilon \nabla f(w_i) + O(\epsilon^2). \qquad (2)$$

Re-organizing the terms and using $\frac{O(\epsilon^2)}{\epsilon} = O(\epsilon)$ yields

$$\nabla f(w_i) = \frac{f(w_i+\epsilon) - f(w_i)}{\epsilon} + O(\epsilon). \qquad (3)$$

(2) For the second equation, we again use a Taylor expansion of $f(w_i+\epsilon)$ and $f(w_i-\epsilon)$ around w_i, but this time up to the third-order, i.e.

$$f(w_i+\epsilon) = f(w_i) + \epsilon\nabla f(w_i) + \frac{1}{2}\epsilon\nabla^2 f(w_i)\epsilon + O(\epsilon^3), \tag{4}$$

and

$$f(w_i-\epsilon) = f(w_i) - \epsilon\nabla f(w_i) + \frac{1}{2}\epsilon\nabla^2 f(w_i)\epsilon + O(\epsilon^3) \tag{5}$$

Subtracting Eq. 5 from Eq. 4 yields

$$\nabla f(w_i) = \frac{f(w_i+\epsilon) - f(w_i-\epsilon)}{2\epsilon} + O(\epsilon^2) \tag{6}$$

Solution 4 (Deep linear networks):

We simply expand the composition of the 2 functions, i.e.

$$\begin{aligned}(g_1 \circ g_2)(\boldsymbol{x}) &= \boldsymbol{W}_1(\boldsymbol{W}_2\boldsymbol{x} + \boldsymbol{b}_2) + \boldsymbol{b}_1 \\ &= \boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{x} + \boldsymbol{W}_1\boldsymbol{b}_2 + \boldsymbol{b}_1 \\ &= \boldsymbol{W}_3\boldsymbol{x} + \boldsymbol{b}_3 \in \mathcal{G},\end{aligned}$$

where $\boldsymbol{W}_3 := \boldsymbol{W}_1\boldsymbol{W}_2 \in \mathbb{R}^{k\times d}$.