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 22, 2018
(Deep Learning, Exercise series 4 - solutions)

Solution 1 (Backpropagation and Computational Graphs):

a) From the chain rule we have that

$$\frac{\partial l}{\partial h_1} = \frac{\partial l}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1} \left(= \frac{\partial l}{\partial h_2}\frac{\partial h_2}{\partial h_1} \right) \tag{1}$$

which can be represented as a computational graph as follows:

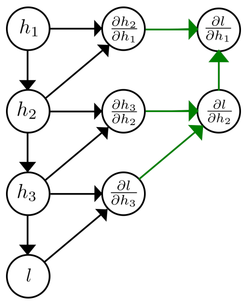

Figure 1: Gradient computations are shown in black, application of the chain rule in green.

b) Using the chain rule, we can derive the following expression for $\frac{\partial l}{\partial x}$ from the given DAG:

$$\frac{\partial l}{\partial x} = \left(\frac{\partial l}{\partial h_2}\frac{\partial h_2}{\partial h_1} + \frac{\partial l}{\partial h_3}\frac{\partial h_3}{\partial h_1} \right) \frac{\partial h_1}{\partial x}$$

We then compute each partial derivative to obtain:

$$\frac{\partial l}{\partial h_2} = h_3 = 2x + y, \quad \frac{\partial l}{\partial h_3} = h_2 = 8x^2, \quad \frac{\partial h_2}{\partial h_1} = 4h_1 = 8x, \quad \frac{\partial h_3}{\partial h_1} = 1, \quad \frac{\partial h_1}{\partial x} = 2$$
$$\frac{\partial l}{\partial x} = 48x^2 + 16xy$$

Solution 2 (Approximate Hessian for feed-forward networks):

(1)

$$\begin{aligned}
\frac{\partial^2 L_n}{\partial w_{ji}^2} &= \frac{\partial}{\partial w_{ji}} \cdot \frac{\partial L_n}{\partial w_{ji}} \\
&= \frac{\partial}{\partial w_{ji}} \left(\frac{\partial L_n}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ji}} \right) && \text{(by the chain rule)} \\
&= \frac{\partial}{\partial w_{ji}} \left(\frac{\partial L_n}{\partial a_j} \cdot z_i \right), && (2)
\end{aligned}$$

where we used the fact that $\frac{\partial a_j}{\partial w_{ji}} = \frac{\partial \sum_i w_{ji} z_i}{\partial w_{ji}} = z_i$ for one specific i. Furthermore, the product rule $\frac{d}{dx}(u \cdot v) = \frac{du}{dx} \cdot v + u \cdot \frac{dv}{dx}$ extends naturally to partial derivatives and thus we have

$$\frac{\partial}{\partial w_{ji}} \left(\frac{\partial L_n}{\partial a_j} \cdot z_i \right) = \frac{\partial}{\partial w_{ji}} \left(\frac{\partial L_n}{\partial a_j} \right) \cdot z_i + \frac{\partial L_n}{\partial a_j} \frac{\partial z_i}{\partial w_{ji}}. \tag{3}$$

Note that the second summand is zero, since the activation of the $i-$th unit does not depend on the weight w_{ji} going from i to j. Thus $\frac{\partial z_i}{\partial w_{ji}} = 0$. Finally,

$$\begin{aligned}
\frac{\partial}{\partial w_{ji}} \left(\frac{\partial L_n}{\partial a_j} \right) \cdot z_i =& \frac{\partial}{\partial a_j} \left(\frac{\partial L_n}{\partial w_{ji}} \right) \cdot z_i && \text{(by Schwarz-Theorem)} \\
=& \frac{\partial}{\partial a_j} \left(\frac{\partial L_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}} \right) \cdot z_i && \text{(by the product rule)} \\
=& \frac{\partial^2 L_n}{\partial a_j^2} \cdot z_i^2. && (4)
\end{aligned}$$

As a result, we get

$$\frac{\partial^2 L_n}{\partial w_{ji}^2} = \frac{\partial^2 L_n}{\partial a_j^2} \cdot z_i^2. \tag{5}$$

Note that the term z_i has already been computed during the forward-pass. Thus, part (2) of this exercise takes a closer look at the first term.

(2)

We first evaluate the first partial derivative, using the fact that a unit j contributes to the loss via the units k that it outputs to i.e.

$$\frac{\partial L_n}{\partial a_j} = \sum_k \frac{\partial L_n}{\partial a_k} \cdot \frac{\partial a_k}{\partial a_j}. \qquad \text{(by the chain rule)} \tag{6}$$

Clearly, for one specific k we have

$$\frac{\partial a_k}{\partial a_j} = \frac{\partial \sum_i w_{ki} z_i}{\partial a_j} = \frac{\partial w_{kj} z_j}{\partial a_j} = h'(a_j) w_{kj} \tag{7}$$

since $z_j = h(a_j)$. Combined, (9) and (10) yield

$$\begin{aligned}
\frac{\partial^2 L_n}{\partial a_j^2} &= \frac{\partial}{\partial a_j} \frac{\partial L_n}{\partial a_j} \\
&= \frac{\partial}{\partial a_j} h'(a_j) \sum_k w_{kj} \frac{\partial L_n}{\partial a_k} \\
&= h''(a_j) \sum_k w_{kj} \frac{\partial L_n}{\partial a_k} + h'(a_j) \sum_k w_{kj} \frac{\partial^2 L_n}{\partial a_k \partial a_j}. && \text{(by the product rule)}
\end{aligned} \tag{8}$$

We can further develop the last term as

$$\begin{aligned}
\frac{\partial^2 L_n}{\partial a_k \partial a_j} = \frac{\partial}{\partial a_k} \cdot \frac{\partial L_n}{\partial a_j} &= \frac{\partial}{\partial a_k} \left(h'(a_j) \sum_{k'} w_{k'j} \frac{\partial L_n}{\partial a_{k'}} \right) \\
&= h'(a_j) \sum_{k'} w_{k'j} \frac{\partial^2 L_n}{\partial a_{k'} \partial a_k},
\end{aligned} \tag{9}$$

which, together with (11), proves the assertion.

(3) If we now neglect off-diagonal elements in the second-derivative terms, we obtain

$$\frac{\partial^2 L_n}{\partial a_j^2} = h'(a_j)^2 \sum_k w_{kj}^2 \frac{\partial^2 L_n}{\partial a_k^2} + h''(a_j) \sum_k w_{kj} \frac{\partial L_n}{\partial a_k}. \tag{10}$$

The computational complexity required to compute this approximate Hessian is $O(W)$, where W is the total number of weights in the network, compared to $O(W^2)$ for the full Hessian.

Solution 3 (Chain-rule and Jacobians in more than 2 Dimensions):

1. Follows by writing the chain rule explicitly for each triplet of indices:

$$\frac{\partial L}{\partial W_{jk}} = \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial W_{jk}}$$

2. $\frac{\partial L}{\partial W} = 2yx^\top$. This results as follows:

$$\frac{\partial L}{\partial y} = 2y$$

$$\frac{\partial y}{\partial W} = T \in \mathbb{R}^{d_1 \times d_1 \times d_2}, \quad T_{i,j,k} = \mathbb{1}_{i=j} x_k$$

   From chain rule: $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \times_{d_1} \frac{\partial y}{\partial W}$ which implies that $\left(\frac{\partial L}{\partial W}\right)_{j,k} = \sum_i \mathbb{1}_{i=j} y_i x_k = y_j x_k$ which is what we wanted.

   For a different perspective on this see also this link.

3. Similar with point 1.

4. $\frac{\partial}{\partial A}(tr(BA)) = \frac{\partial tr(BA)}{\partial BA} \times_{i_1,i_2} \frac{\partial BA}{\partial A} = B^\top$