Title: Derivation of the back-propagation algorithm
Author: Rui Zhang
Email: ray.rui.zhang@gmail.com

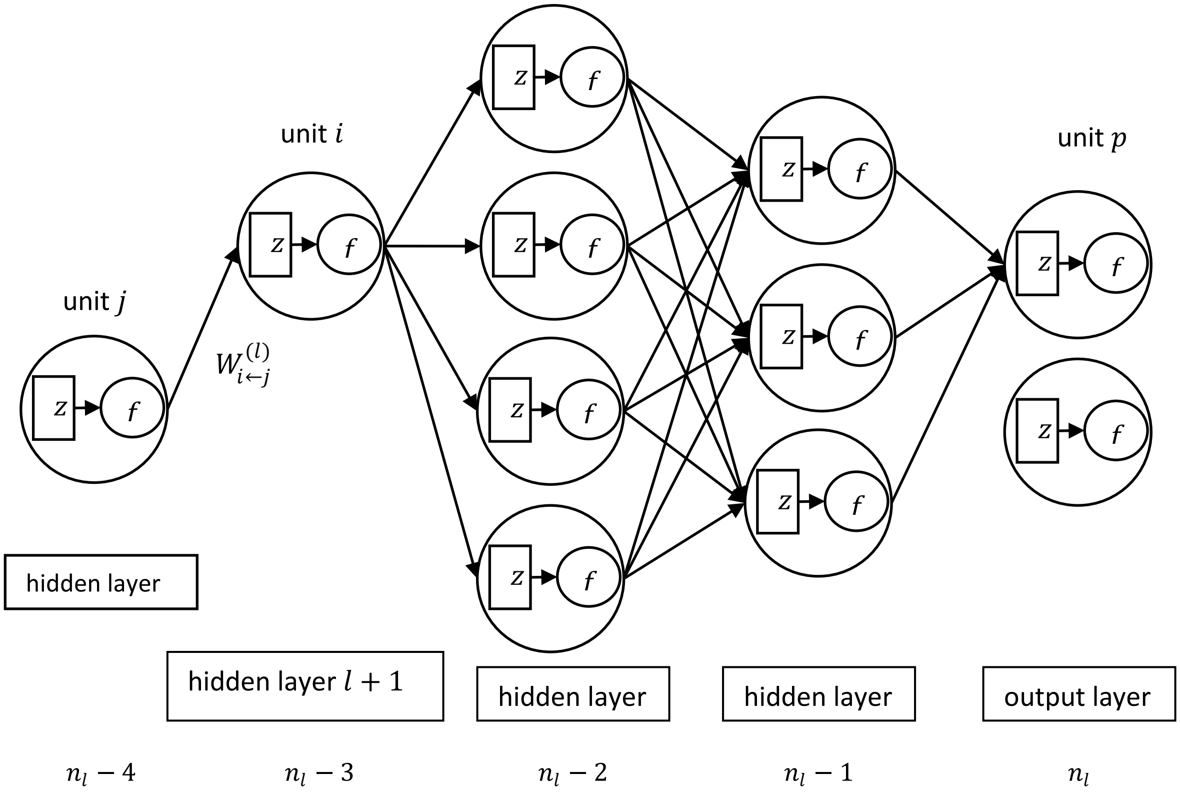

Figure 1 Part of a neural network used to illustrate the gradient of the output with respect to the weight $W_{i\leftarrow j}^{(l)}$

Let use j, i, k, q, p as the indexes for the five layers from left to right, and thus we have

$j = 1,2, \ldots, s_{n_l-4}$, $i = 1,2, \ldots, s_{n_l-3}$, $k = 1,2, \ldots, s_{n_l-2}$, $q = 1,2, \ldots, s_{n_l-1}$ and $p = 1,2, \ldots, s_{n_l}$.

As an example, we will derive the $\partial J(W, b; x, y)/\partial W_{i\leftarrow j}^{(n_l-4)}$, i.e. the gradient of cost function with respect to the weight $W_{i\leftarrow j}^{(n_l-4)}$. Although we use this particular example to show the derivation, it can be easily generalized to the generic form of the gradient with respect to any weight other of the network. The gradients with respect to the biases can also be derived straightforwardly as discussed later. Also note that in the derivation of the gradients, the layers are indexed using the numbers on the bottom line of the figure.

$$\frac{\partial J(W,b;x,y)}{\partial W_{i\leftarrow j}^{(n_l-4)}} = \frac{\partial \left(\frac{1}{2}\|y - h(W,b;x)\|^2\right)}{\partial W_{i\leftarrow j}^{(n_l-4)}} = \frac{\partial \left(\frac{1}{2}\sum_{p=1}^{s_{n_l}} \left(y_p - h_p(W,b;x)\right)^2\right)}{\partial W_{i\leftarrow j}^{(n_l-4)}}$$

$$= -\sum_{p=1}^{s_{n_l}} \left(y_p - h_p(W,b;x)\right) \cdot \frac{\partial h_p(W,b;x)}{\partial W_{i\leftarrow j}^{(n_l-4)}}$$

Title: Derivation of the back-propagation algorithm
Author: Rui Zhang
Email: ray.rui.zhang@gmail.com

Let us first look into the right-hand side part of each term in the above sum.

$$\frac{\partial h_p(W,b;x)}{\partial W_{i\leftarrow j}^{(n_l-4)}} = \frac{\partial a_p^{(n_l)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} = \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \frac{\partial z_p^{(n_l)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} = \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \frac{\partial \left(W_p^{(n_l-1)} \cdot a^{(n_l-1)}\right)}{\partial W_{i\leftarrow j}^{(n_l-4)}}$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \frac{\partial a_q^{(n_l-1)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) = \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \frac{\partial z_q^{(n_l-1)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) \right)$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \frac{\partial a_k^{(n_l-2)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) \right) \right)$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot \frac{\partial z_k^{(n_l-2)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) \right) \right) \right)$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot \left(\sum_{i=1}^{s_{n_l-3}} W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{\partial a_i^{(n_l-3)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) \right) \right) \right) \right)$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{\partial a_i^{(n_l-3)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) \right) \right) \right)$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \frac{\partial z_i^{(n_l-3)}}{\partial W_{i\leftarrow j}^{(n_l-4)}} \right) \right) \right) \right)$$

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \Bigg(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)}$$

$$\cdot \Bigg(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}}$$

$$\cdot \Bigg(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \Bigg(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \frac{\partial \left(W_i^{(n_l-4)} \cdot a^{(n_l-4)}\right)}{\partial W_{i\leftarrow j}^{(n_l-4)}} \Bigg) \Bigg) \Bigg) \Bigg)$$

Title: Derivation of the back-propagation algorithm
Author: Rui Zhang
Email: ray.rui.zhang@gmail.com

$$= \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{k=1}^{s_{n_l-2}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot a_j^{(n_l-4)} \right) \right) \right) \right)$$

$$= \sum_{q=1}^{s_{n_l-1}} \sum_{k=1}^{s_{n_l-2}} \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot W_{p\leftarrow q}^{(n_l-1)} \cdot \frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot W_{q\leftarrow k}^{(n_l-2)} \cdot \frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot a_j^{(n_l-4)}$$

$$= \left(\sum_{q=1}^{s_{n_l-1}} \sum_{k=1}^{s_{n_l-2}} \frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \cdot W_{p\leftarrow q}^{(n_l-1)} \cdot \frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot W_{q\leftarrow k}^{(n_l-2)} \cdot \frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot W_{k\leftarrow i}^{(n_l-3)} \right) \cdot \frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot a_j^{(n_l-4)}$$

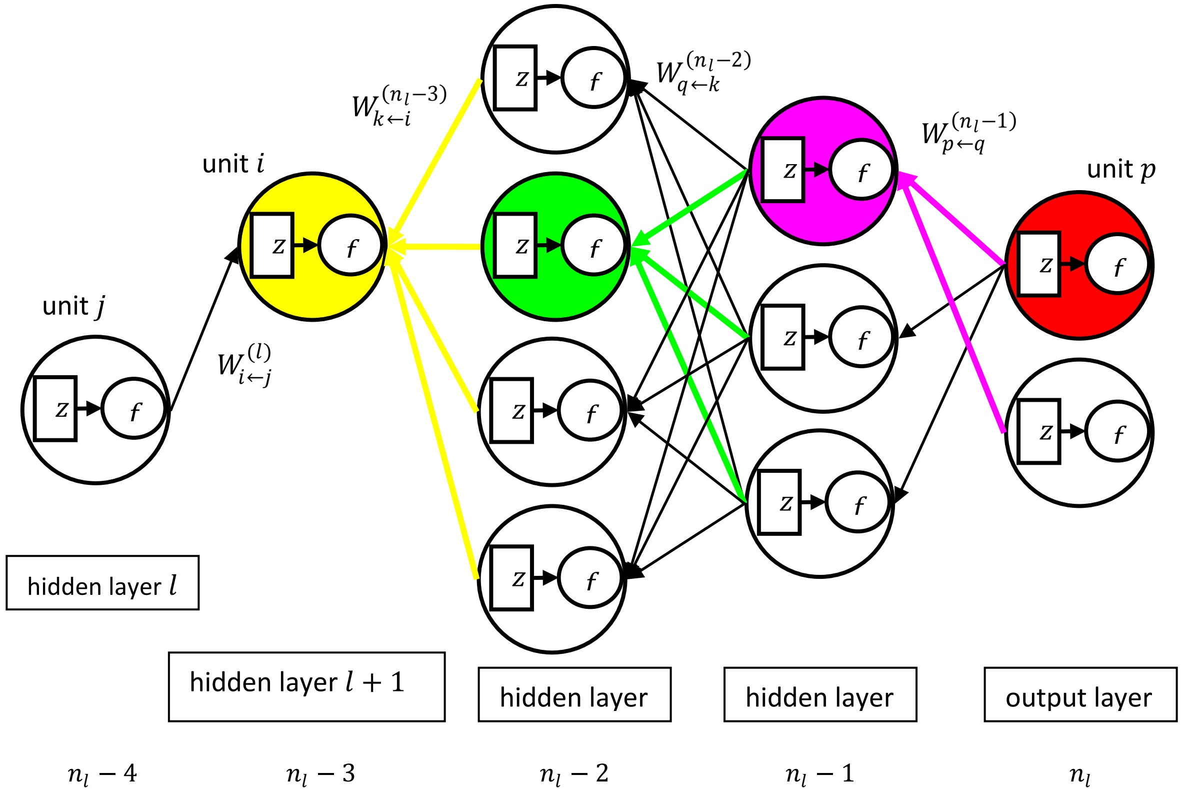

Figure 2 On each layer, the error is propagated to every unit there, but only one of them is shown using a coloured path.

The trick comes into play in this step ...

$$= \left(\frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \sum_{k=1}^{s_{n_l-2}} W_{k\leftarrow i}^{(n_l-3)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot \sum_{q=1}^{s_{n_l-1}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot W_{p\leftarrow q}^{(n_l-1)} \cdot \left(\frac{da_p^{(n_l)}}{dz_p^{(n_l)}} \right) \right) \right) \right) \cdot a_j^{(n_l-4)}$$

Title: Derivation of the back-propagation algorithm
Author: Rui Zhang
Email: ray.rui.zhang@gmail.com

$$= \left(\frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \sum_{k=1}^{s_{n_l-2}} W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{q\leftarrow k}^{(n_l-2)} \cdot \frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(W_{p\leftarrow q}^{(n_l-1)} \cdot \delta_p^{(n_l)}\right)\right)\right) \cdot a_j^{(n_l-4)}$$

$$= \left(\frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \sum_{k=1}^{s_{n_l-2}} W_{k\leftarrow i}^{(n_l-3)} \cdot \frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot \left(\sum_{q=1}^{s_{n_l-1}} W_{q\leftarrow k}^{(n_l-2)} \cdot \delta_q^{(n_l-1)}\right)\right) \cdot a_j^{(n_l-4)}$$

$$= \left(\frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \sum_{k=1}^{s_{n_l-2}} W_{k\leftarrow i}^{(n_l-3)} \delta_k^{(n_l-2)}\right) \cdot a_j^{(n_l-4)}$$

$$= \delta_i^{(n_l-3)} \cdot a_j^{(n_l-4)}$$

Now if we bring back the sum over the units in the output layer, the above result will become

$$\frac{\partial J(W,b;x,y)}{\partial W_{i\leftarrow j}^{(n_l-4)}} = \left(\frac{da_i^{(n_l-3)}}{dz_i^{(n_l-3)}} \cdot \sum_{k=1}^{s_{n_l-2}} W_{k\leftarrow i}^{(n_l-3)} \cdot \left(\frac{da_k^{(n_l-2)}}{dz_k^{(n_l-2)}} \cdot \sum_{q=1}^{s_{n_l-1}} W_{q\leftarrow k}^{(n_l-2)} \cdot \left(\frac{da_q^{(n_l-1)}}{dz_q^{(n_l-1)}} \cdot \left(\sum_{p=1}^{s_{n_l}} W_{p\leftarrow q}^{(n_l-1)} \left(-\frac{da_p^{(n_l)}}{dz_p^{(n_l)}}\left(y_p - h_p(W,b;x)\right)\right)\right)\right)\right)\right) \cdot a_j^{(n_l-4)}$$

The above steps indicates the notion of propagating the error from the output unit to the weight of interest.

Assuming the activations of all the units on all the layers have been computed through a forward pass. The error propagated to each of them (or in other words the error contributed by each of them) can be computed via a backward pass.

1) The error at the output unit *p* is calculated by

$$\delta_p^{(n_l)} = -\left(y - h_p(W,b;x,y)\right)\left(\frac{dh_p(W,b;x,y)}{dz_p^{(n_l)}}\right) == -\left(y - h_p(W,b;x,y)\right)\left(\frac{da_p^{(n_l)}}{dz_p^{(n_l)}}\right)$$

2) The error propagated to each of the units of the network from the output layer until the first hidden layer can be computed by

$$\delta_k^{(l)} = \left(\sum_{q=1}^{s_{l+1}} W_{q \leftarrow k}^{(l)} \cdot \delta_q^{(l+1)} \right) \cdot \frac{da_k^{(l)}}{dz_k^{(l)}}$$

where we kind of abuse the notation of *k* and *q* because they used to be the indexes of the units on the layer $n_l - 1$ and $n_l - 2$. In the above equation, they are the indexes of the units on any two adjacent layers.

3) The gradient of the cost function with respect to a certain weight, say $W_{i \leftarrow j}^{(l)}$, can be computed by

$$\frac{\partial h_p(W, b; x, y)}{\partial W_{i \leftarrow j}^{(l)}} = \delta_i^{(l+1)} \cdot a_j^{(l)}$$

The gradient of the cost function with respect to a certain bias can be derived in the same way.