Back Propagation

Alon Filler, Nadav Menirav

August 2025

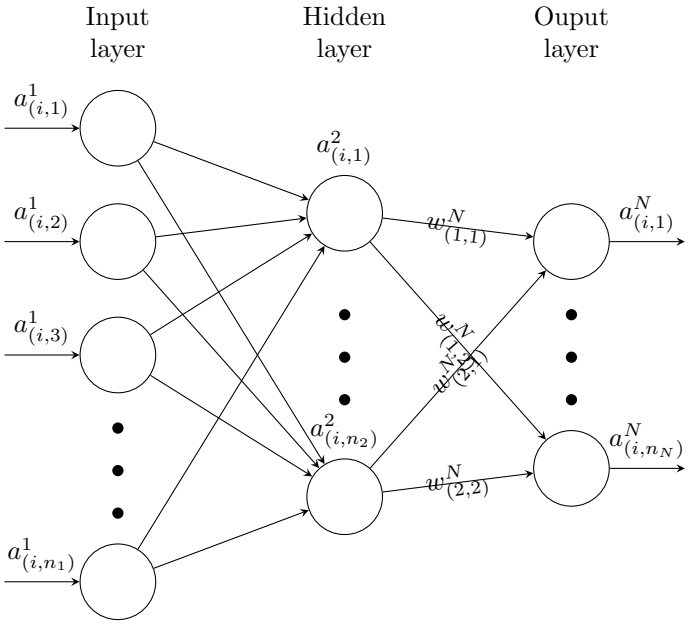
Abstract

Back Propagation is one of many algorithms used to determine the gradient of a neural network's cost function. Its approach differs from other algorithms, in that it computes the derivatives for the last layer, and uses the results to compute the derivatives of earlier layers recursively. In this paper, we are going to derive the formula for the derivatives of each of the parameters (weights and biases) in a general neural network.

Notations

Layers	$= N$
Neurons per layer	$= n_l$
Training data count	$= D$
Entry in training data	$= i \in \{1, \dots, D\}$
Neuron in layer l	$= j \in \{1, \dots, n_l\}$
Activation of neuron before non-linearity	$= \sigma^{-1}(a^l_{(i,j)}) = \left(\sum_{k=1}^{n_{l-1}} w^l_{(k,j)} a^{l-1}_{(i,k)}\right) + b^l_j$
Activation of neuron	$= a^l_{(i,j)} = \sigma\left(\left(\sum_{k=1}^{n_{l-1}} w^l_{(k,j)} a^{l-1}_{(i,k)}\right) + b^l_j\right)$
Input	$= a^1_{(i,j)}$
Output	$= a^N_{(i,j)}$

For the ease of the reader, here is an illustration of the notations for the i input:



Calculating Gradient

Deriving the cost

$$
\begin{aligned}
cost(X) &= \frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}(prediction_{(i,j)} - expected_{(i,j)})^2\right) \\
&= \frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}(a^N_{(i,j)} - expected_{(i,j)})^2\right)
\end{aligned}
$$

$$
\begin{aligned}
\partial\left(cost(X)\right)_{b_m^l} &= \partial\left(\frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}(a_{(i,j)}^N - expected_{(i,j)})^2\right)\right)_{b_m^l} \\
&= \frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}\partial\left((a_{(i,j)}^N - expected_{(i,j)})^2\right)_{b_m^l}\right) \\
&= \frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}2(a_{(i,j)}^N - expected_{(i,j)})\partial\left(a_{(i,j)}^N - expected_{(i,j)}\right)_{b_m^l}\right) \\
&= \frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}2(a_{(i,j)}^N - expected_{(i,j)})\partial\left(a_{(i,j)}^N\right)_{b_m^l}\right)
\end{aligned}
$$

The partial derivative was calculated below.

$$
\partial\left(cost(X)\right)_{w_{(m,t)}^l} = \frac{1}{D}\sum_{i=1}^{D}\left(\sum_{j=1}^{n_N}2(a_{(i,j)}^N - expected_{(i,j)})\partial\left(a_{(i,j)}^N\right)_{w_{(m,t)}^l}\right)
$$

This partial derivative was also calculated below.

Sub derivatives of the cost

$$\partial\left(a^N_{(i,j)}\right)_{a^{N-1}_{(i,m)}} = \partial\left(\left(\sum_{k=1}^{n_{N-1}} w^N_{(k,j)} a^{N-1}_{(i,k)}\right) + b^N_j\right)_{a^{N-1}_{(i,m)}} = w^N_{(m,j)}$$

$$\partial\left(a^N_{(i,j)}\right)_{w^N_{(m,j)}} = \partial\left(\left(\sum_{k=1}^{n_{N-1}} w^N_{(k,j)} a^{N-1}_{(i,k)}\right) + b^N_j\right)_{w^N_{(m,j)}} = a^{N-1}_{(i,m)}$$

$$\partial\left(a^N_{(i,j)}\right)_{b^N_j} = \partial\left(\left(\sum_{k=1}^{n_{N-1}} w^N_{(k,j)} a^{N-1}_{(i,k)}\right) + b^N_j\right)_{b^N_j} = 1$$

$$\frac{\partial a^l_{(i,t)}}{\partial w^l_{(m,t)}} = \partial\left(a^l_{(i,t)}\right)_{w^l_{(m,t)}} = a^l_{(i,t)} \cdot \left(1 - a^l_{(i,t)}\right) \cdot a^{l-1}_{(i,m)}$$

$$\begin{aligned}
\frac{\partial a^{l+1}_{(i,t)}}{\partial b^l_m} &= \partial\left(a^{l+1}_{(i,t)}\right)_{b^l_m} = \partial\sigma\left(\left(\sum_{k=1}^{n_l} w^{l+1}_{(k,t)} a^l_{(i,k)}\right) + b^{l+1}_t\right)_{b^l_m} \\
&= / : \partial\left(\sigma\left(f(x)\right)\right)_{x_i} = \sigma(f(x))\left(1 - \sigma(f(x))\,\partial\left(f(x)\right)_{x_i}\right. \\
&= a^{l+1}_{(i,t)} \cdot \left(1 - a^{l+1}_{(i,t)}\right) \cdot \partial\left(\sum_{k=1}^{n_l} w^{l+1}_{(k,t)} a^l_{(i,k)}\right)_{b^l_m} \\
&= a^{l+1}_{(i,t)} \cdot \left(1 - a^{l+1}_{(i,t)}\right) \cdot w^{l+1}_{(m,t)} \cdot \partial\left(a^l_{(i,m)}\right)_{b^l_m} \\
&= a^{l+1}_{(i,t)} \cdot \left(1 - a^{l+1}_{(i,t)}\right) \cdot w^{l+1}_{(m,t)} \cdot a^l_{(i,m)} \cdot \left(1 - a^l_{(i,m)}\right) \partial\left(\left(\sum_{k=1}^{n_{l-1}} w^l_{(k,m)} a^{l-1}_{(i,k)}\right) + b^l_m\right)_{b^l_m} \\
&= a^{l+1}_{(i,t)} \cdot \left(1 - a^{l+1}_{(i,t)}\right) \cdot w^{l+1}_{(m,t)} \cdot a^l_{(i,m)} \cdot \left(1 - a^l_{(i,m)}\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial a^{l+1}_{(i,k)}}{\partial a^l_{(i,t)}} &= \partial\left(a^{l+1}_{(i,k)}\right)_{\partial a^l_{(i,t)}} = a^{l+1}_{(i,k)} \cdot (1 - a^{l+1}_{(i,k)}) \cdot \partial\left(\sigma^{-1}\left(a^{l+1}_{(i,k)}\right)\right)_{a^l_{(i,t)}} \\
&= a^{l+1}_{(i,k)} \cdot (1 - a^{l+1}_{(i,k)}) \cdot \partial\left(\left(\sum_{j=1}^{n_l} w^{l+1}_{(j,k)} a^l_{(i,j)}\right) + b^{l+1}_k\right)_{a^l_{(i,t)}} \\
&= a^{l+1}_{(i,k)} \cdot (1 - a^{l+1}_{(i,k)}) \cdot \partial\left(w^{l+1}_{(t,k)} a^l_{(i,t)}\right)_{a^l_{(i,t)}} \\
&= a^{l+1}_{(i,k)} \cdot (1 - a^{l+1}_{(i,k)}) \cdot w^{l+1}_{(t,k)}
\end{aligned}$$

$$\frac{\partial a^N_{(i,j)}}{\partial a^{l+1}_{(i,t)}} = \partial\left(a^N_{(i,j)}\right)_{a^{l+1}_{(i,t)}} = \sum_{k=1}^{n_{l+2}} \frac{\partial a^N_{(i,j)}}{\partial a^{l+2}_{(i,k)}} \cdot \frac{\partial a^{l+2}_{(i,k)}}{\partial a^{l+1}_{(i,t)}}$$

$$\partial\left(a^N_{(i,j)}\right)_{b^l_m} = \sum_{t=1}^{n_{l+1}} \frac{\partial a^N_{(i,j)}}{\partial a^{l+1}_{(i,t)}} \cdot \frac{\partial a^{l+1}_{(i,t)}}{\partial b^l_m}$$

$$\partial\left(a^N_{(i,j)}\right)_{w^l_{(m,t)}} = \frac{\partial a^N_{(i,j)}}{\partial a^l_{(i,t)}} \cdot \frac{\partial a^l_{(i,t)}}{\partial w^l_{(m,t)}}$$