Derivations for Backpropagation

Sayantan Auddy

This document contains derivations for backpropagation for a fully connected neural network with 1 hidden layer.

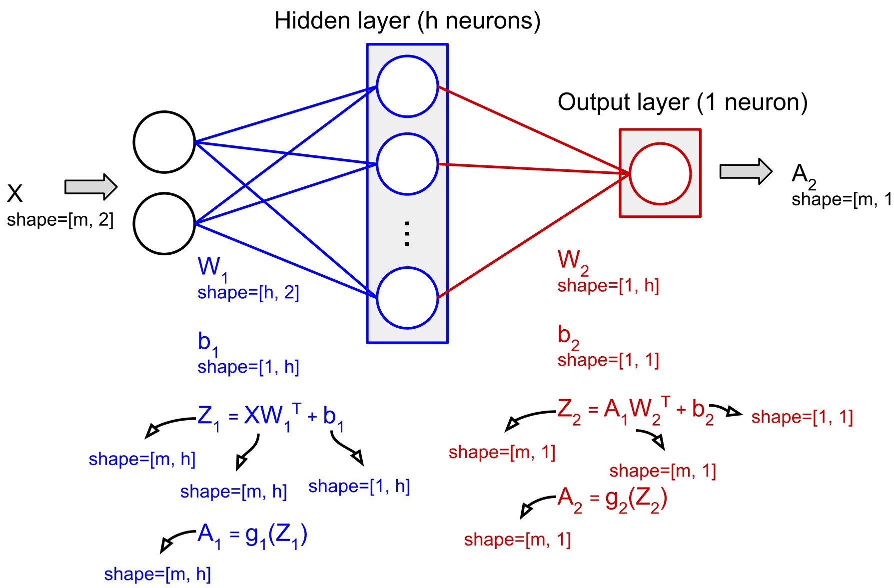

Figure 1: Fully connected neural network with 1 hidden layer. Here, the dimension of each data point is $D = 2$ and there are m data points in a batch. There are h units in the hidden layer.

Let us consider the neural network as a graph. We are passing in data $X \in \mathbb{R}^{m \times D}$ into the network which will perform binary classification. Here m is the number of data points in X and D is the dimension of each data point. Under these assumptions, the forward propagation step computes the following:

$$Z_1 = XW_1^T + b_1 \tag{1}$$
$$A_1 = g_1(Z_1) \text{ (where } g_1 \text{ is the activation function of layer 1)} \tag{2}$$
$$Z_2 = A_1W_2^T + b_2 \tag{3}$$
$$A_2 = g_2(Z_2) \text{ (where } g_2 \text{ is the activation function of the output layer)} \tag{4}$$

The operations of the forward pass are shown as a graph below:

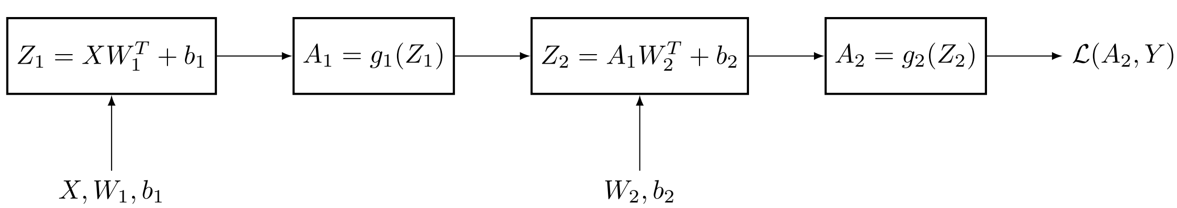

Figure 2: Forward propagation graph.

It makes sense at this point to be aware of the dimensions of the various variables involved. These are specified below:

$$
\begin{aligned}
X &\in \mathbb{R}^{m\times D}\\
W_1 &\in \mathbb{R}^{h\times D}, b_1 \in \mathbb{R}^{1\times h}\\
Z_1 &\in \mathbb{R}^{m\times h}, A_1 \in \mathbb{R}^{m\times h}\\
W_2 &\in \mathbb{R}^{1\times h}, b_2 \in \mathbb{R}^{1\times 1}\\
Z_2 &\in \mathbb{R}^{m\times 1}, A_2 \in \mathbb{R}^{m\times 1}
\end{aligned}
$$

The loss $\mathcal{L}$ is given by

$$\mathcal{L}(A_2, Y) = \frac{1}{m}\sum_{i=1}^{m}\big(-Y^{(i)}\log(A_2^{(i)}) - (1-Y^{(i)})\log(1-A_2^{(i)})\big) \tag{5}$$

We need the final loss to be a scalar value and so we compress the loss by taking the average the term $\big(-Y^{(i)}\log(A_2^{(i)}) - (1-Y^{(i)})\log(1-A_2^{(i)})\big)$ for each data point i. However, for computing the derivatives of the various terms shown in Fig. 2, we will consider the uncompressed loss:

$$\mathcal{L}(A_2, Y) = -Y\log(A_2) - (1-Y)\log(1-A_2) \tag{6}$$

Now, let's derive the gradients for the various elements in the graph shown in Fig. 2. In the code, we use the shorthand notation $\mathrm{d}z_2$ to mean $\dfrac{\partial\mathcal{L}}{\partial z_2}$ and so on. Throughout the following derivations, we will be using the chain rule for differentiation. The forward and backward operations (in red) are summarized together in Fig. 3.

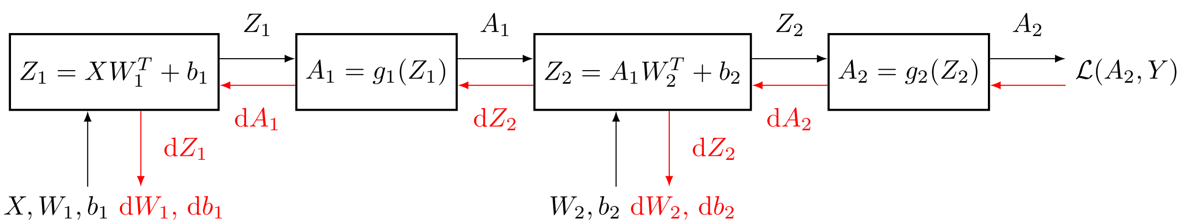

Figure 3: Forward and backward propagation.

- Computing $\mathrm{d}A_2 = \dfrac{\partial\mathcal{L}}{\partial A_2}$ (note that $\mathrm{d}A_2 \in \mathbb{R}^{m\times 1}$)

$$
\begin{aligned}
\mathcal{L}(A_2, Y) &= -Y\log(A_2) - (1-Y)\log(1-A_2)\\
\text{Hence } \frac{\partial\mathcal{L}}{\partial A_2} &= \frac{-Y}{A_2} + \frac{(1-Y)}{(1-A_2)}
\end{aligned} \tag{7}
$$

- Computing $\mathrm{d}Z_2 = \dfrac{\partial\mathcal{L}}{\partial Z_2}$ (note that $\mathrm{d}Z_2 \in \mathbb{R}^{m\times 1}$)

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial Z_2} &= \frac{\partial\mathcal{L}}{\partial A_2}\frac{\partial A_2}{\partial Z_2} = \Big[\frac{-Y}{A_2} + \frac{(1-Y)}{(1-A_2)}\Big]\frac{\partial g_2(Z_2)}{\partial Z_2} = \Big[\frac{-Y}{A_2} + \frac{(1-Y)}{(1-A_2)}\Big]g_2'(Z_2)\\
\Rightarrow \frac{\partial\mathcal{L}}{\partial Z_2} &= \Big[\frac{-Y}{A_2} + \frac{(1-Y)}{(1-A_2)}\Big]g_2(Z_2)\Big(1-g_2(Z_2)\Big) \text{ (since } \sigma'(u) = \sigma(u)(1-\sigma(u)))\\
\Rightarrow \frac{\partial\mathcal{L}}{\partial Z_2} &= \Big[\frac{-Y}{A_2} + \frac{(1-Y)}{(1-A_2)}\Big]A_2(1-A_2) = A_2 - Y
\end{aligned} \tag{8}
$$

- Computing $\mathrm{d}W_2 = \dfrac{\partial \mathcal{L}}{\partial W_2}$ (note that $\mathrm{d}W_2 \in \mathbb{R}^{1\times h}$)

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial Z_2}\frac{\partial Z_2}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial Z_2}A_1 = \mathrm{d}Z_2^T A_1 \text{ (the transpose follows from the shape of the matrices involved)} \tag{9}$$

There is a small caveat in this derivation. When we have m data points in a batch, the derivative $\mathrm{d}W_2$ will contain contributions from each single data point, and these individual contributions are summed up to get the final value of $\mathrm{d}W_2$. To understand this more clearly, let us consider that $m = 3$, $h = 2$, and let

$$\mathrm{d}Z_2 = \begin{pmatrix} z_{11} \\ z_{21} \\ z_{31} \end{pmatrix} A_1 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \tag{10}$$

Thus,

$$\mathrm{d}W_2 = \mathrm{d}Z_2^T A_1 = \begin{pmatrix} z_{11}a_{11} + z_{21}a_{21} + z_{31}a_{31} & z_{11}a_{12} + z_{21}a_{22} + z_{31}a_{32} \end{pmatrix} \tag{11}$$

In general, for m data points in a batch,

$$\mathrm{d}W_2 = \begin{pmatrix} \sum_j^m z_{j1}a_{j1} & \sum_j^m z_{j1}a_{j2} \end{pmatrix} \tag{12}$$

To make the value of $\mathrm{d}W_2$ independent of the batch size m, we divide the sums by m, which gives us

$$\mathrm{d}W_2 = \frac{1}{m}\begin{pmatrix} \sum_j^m z_{j1}a_{j1} & \sum_j^m z_{j1}a_{j2} \end{pmatrix} \tag{13}$$

And so, finally we have

$$\mathrm{d}W_2 = \frac{1}{m}\mathrm{d}Z_2^T A_1 \tag{14}$$

- Computing $\mathrm{d}b_2 = \dfrac{\partial \mathcal{L}}{\partial b_2}$ (note that $\mathrm{d}b_2 \in \mathbb{R}^{1\times 1}$)

$$\frac{\partial \mathcal{L}}{\partial b_2} = \frac{\partial \mathcal{L}}{\partial Z_2}\frac{\partial dZ_2}{\partial db_2} = \frac{\partial \mathcal{L}}{\partial Z_2} = \mathrm{d}Z_2 \tag{15}$$

But this derivation is not yet done. Let us again consider that $m = 3$, $h = 2$, and let

$$\mathrm{d}Z_2 = \begin{pmatrix} z_{11} \\ z_{21} \\ z_{31} \end{pmatrix} \tag{16}$$

So $\mathrm{d}Z_2$ will contain 1 row for each of the m data points. Moreover, we need to make the dimension of $\mathrm{d}b_2$, the same as $b_2 \in \mathbb{R}^{1\times 1}$. To do this, we compute the average of the rows in $\mathrm{d}Z_2$, and so we get

$$\frac{\partial \mathcal{L}}{\partial b_2} = \frac{1}{m}\left(\sum_j^m z_{j1}\right) = \frac{1}{m}\sum_{\text{along rows}} \mathrm{d}Z_2 \tag{17}$$

- Computing $\mathrm{d}A_1 = \dfrac{\partial \mathcal{L}}{\partial A_1}$ (note that $\mathrm{d}A_1 \in \mathbb{R}^{m\times h}$)

$$\frac{\partial \mathcal{L}}{\partial A_1} = \frac{\partial \mathcal{L}}{\partial Z_2}\frac{\partial Z_2}{\partial A_1} = \frac{\partial \mathcal{L}}{\partial Z_2}W_2 = \mathrm{d}Z_2 W_2 \tag{18}$$

- Computing $\mathrm{d}Z_1 = \dfrac{\partial \mathcal{L}}{\partial Z_1}$ (note that $\mathrm{d}Z_1 \in \mathbb{R}^{m\times h}$)

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial Z_1} &= \frac{\partial \mathcal{L}}{\partial A_1}\frac{\partial A_1}{\partial Z_1} = \left(\frac{\partial \mathcal{L}}{\partial Z_2}W_2\right) \odot \left(\frac{\partial g_1(Z_1)}{\partial Z_1}\right) \\ \Rightarrow \frac{\partial \mathcal{L}}{\partial Z_1} &= \left(\frac{\partial \mathcal{L}}{\partial Z_2}W_2\right) \odot (1 - A_1^2) \text{ (since } g_1 = \tanh\text{, and } g_1(u)' = 1 - g_1^2(u)) \\ \Rightarrow & \left(\mathrm{d}Z_2 W_2\right) \odot (1 - A_1^2) \text{ (element-wise multiplication follows from the matrix shapes)} \end{aligned} \tag{19}$$

- Computing $\mathrm{d}W_1 = \frac{\partial \mathcal{L}}{\partial W_1}$ (note that $\mathrm{d}W_1 \in \mathbb{R}^{h \times D}$)

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial Z_1}\frac{\partial Z_1}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial Z_1} X = \mathrm{d}Z_1^T X \text{ (transpose follows from the matrix shapes)} \tag{20}$$

  Applying the same logic we did while computing $\mathrm{d}W_2$, we get

$$\mathrm{d}W_1 = \frac{1}{m}\mathrm{d}Z_1^T X \tag{21}$$

- Computing $\mathrm{d}b_1 = \frac{\partial \mathcal{L}}{\partial b_1}$ (note that $\mathrm{d}b_1 \in \mathbb{R}^{1 \times h}$)

$$\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial Z_1}\frac{\partial Z_1}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial Z_1} = \mathrm{d}Z_1 \tag{22}$$

  Again, using the logic for computing $\mathrm{d}b_2$, we average the rows of $\mathrm{d}Z_1 \in \mathbb{R}^{m \times h}$ to get $\mathrm{d}b_1 \in \mathbb{R}^{1 \times h}$.

$$\mathrm{d}b_1 = \frac{1}{m}\sum_{\text{along rows}} \mathrm{d}Z_1 \tag{23}$$

Thus, the backpropagation equations can be summarized as:

$$\mathrm{d}Z_2 = \frac{\partial \mathcal{L}}{\partial Z_2} = A_2 - Y \tag{24}$$

$$\mathrm{d}W_2 = \frac{\partial \mathcal{L}}{\partial W_2} = \frac{1}{m}\mathrm{d}Z_2^T A_1 \tag{25}$$

$$\mathrm{d}b_2 = \frac{\partial \mathcal{L}}{\partial b_2} = \frac{1}{m}\sum_{\text{along rows}} \mathrm{d}Z_2 \tag{26}$$

$$\mathrm{d}Z_1 = \frac{\partial \mathcal{L}}{\partial Z_1} = \Big(\mathrm{d}Z_2 W_2\Big) \odot (1 - A_1^2) \tag{27}$$

$$\mathrm{d}W_1 = \frac{\partial \mathcal{L}}{\partial W_1} = \frac{1}{m}\mathrm{d}Z_1^T X \tag{28}$$

$$\mathrm{d}b_1 = \frac{\partial \mathcal{L}}{\partial b_1} = \frac{1}{m}\sum_{\text{along rows}} \mathrm{d}Z_1 \tag{29}$$