Neural Network Mathematics

by Fraser Sabine

version 1.0

Table of contents

1 Notation . . . 2

2 Cost/Loss Function . . . 3

2.1 Squared error . . . 3
2.2 Cross entropy . . . 3

3 Activation Function . . . 4

3.1 Sigmoid . . . 4
3.2 Softmax . . . 4

4 Matrix weights . . . 5

5 Backpropigation . . . 6

5.1 Sigmoid and squared error . . . 6
5.2 Softmax and cross entropy . . . 7

Appendix A Cross entropy derivation . . . 9

Appendix B Softmax derivation . . . 11

1 Notation

Where n is the number of nodes in the neural network

Where L denotes the final layer in the neural network and $L - \ldots$ denotes previous layers

Where $\boldsymbol{z}^L$ is the calculated output $\boldsymbol{z}^L = \boldsymbol{a}^{L-1} W^L + B^L$

Where $\boldsymbol{a}^L$ is the activation function applied to $\boldsymbol{z}^L$. Note $\boldsymbol{a}^{L-H}$ where H is the number of layers, is the input layer of the neural network.

Where W^L is the weights matrix

Where $\boldsymbol{b}^L$ is the biases matrix

Where $\boldsymbol{t}$ is the target output given as a training example

Where c is the loss/cost function

Note $\oslash$ is the Hadamard division or the element-wise division of a vector or matrix

Note $\circ$ is the Hadamard product or the element-wise multiplication of a vector or matrix

Note $\cdot$ is the dot product of a vector or matrix

$$W^L = \begin{pmatrix} w_{00}^L & w_{01}^L & \cdots & w_{0n}^L \\ w_{10}^L & w_{11}^L & \cdots & w_{1n}^L \\ \vdots & \vdots & \ddots & \vdots \\ w_{n0}^L & w_{n1}^L & \cdots & w_{\mathrm{nn}}^L \end{pmatrix} \quad \boldsymbol{b}^L = \begin{pmatrix} b_0^L \\ b_1^L \\ \vdots \\ b_n^L \end{pmatrix}$$

$$\boldsymbol{z}^L = W^L \cdot \boldsymbol{a}^{L-1} + \boldsymbol{b}^L = \begin{pmatrix} w_{00}^L a_0^{L-1} + w_{01}^L a_1^{L-1} + \cdots + w_{0n}^L a_n^{L-1} + b_0^L \\ w_{10}^L a_1^{L-1} + w_{11}^L a_1^{L-1} + \cdots + w_{1n}^L a_n^{L-1} + b_1^L \\ \vdots \\ w_{n0}^L a_n^{L-1} + w_{n1}^L a_n^{L-1} + \cdots + w_{\mathrm{nn}}^L a_n^{L-1} + b_n^L \end{pmatrix} = \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_n \end{pmatrix}$$

$$z_j^L = \sum_{k=0}^{n} w_{\mathrm{jk}}^L a_k^{L-1} + b_j^L$$

where f() is the activation function e.g. sigmoid, tanh, relu etc

$$\boldsymbol{a}^L = \begin{pmatrix} f(z_0) \\ f(z_1) \\ \vdots \\ f(z_n) \end{pmatrix} = \begin{pmatrix} a_0^L \\ a_1^L \\ \vdots \\ a_n^L \end{pmatrix}$$

2 Loss/Cost Function

The loss function is generally for a single training example whereas the cost function is usually the mean of the loss function over a batch (or mini-batch) of training examples. Doing gradient decent with batches (or mini-batches) is more computationally efficient as backpropigation is calculated only as many times as there are batches. The loss/cost function needs to produce a scalar value to make gradient decent possible.

2.1 Squared error

Note 1: this is very similar to the mean squared error function however the $\frac{1}{n}$ is omitted as it is quicker to compute without it and doesn't meaningfully affect the gradient descent.

Note 2: $\frac{1}{2}$ is multiplied as a constant to the general squared error function to remove the constant from the derivative.

Element form

$$c = = \frac{1}{2}\sum_{j=0}^{n}(t_j - a_j^L)^2 = \frac{1}{2}((\boldsymbol{t}_0 - \boldsymbol{a}_0^L)^2 + (\boldsymbol{t}_1 - \boldsymbol{a}_1^L)^2 + \cdots + (\boldsymbol{t}_n - \boldsymbol{a}_n^L)^2)$$

$$\frac{\mathrm{d}c}{\mathrm{d}a_j^L} = t_j - a_j^L$$

Matrix Form

$$c = \frac{1}{2}((\boldsymbol{t} - \boldsymbol{a}^L) \circ (\boldsymbol{t} - \boldsymbol{a}^L))$$

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} = \boldsymbol{t} - \boldsymbol{a}^L = \begin{pmatrix} t_0 - a_0^L \\ t_1 - a_1^L \\ \vdots \\ t_n - a_n^L \end{pmatrix}$$

2.2 Cross entropy

Element form

$$c = -\sum_{j=0}^{n} t_j \log_b(a_j^L) = -(t_0 \log_b(a_0^L) + t_1 \log_b(a_1^L) + \cdots + t_n \log_b(a_n^L))$$

$$\frac{\mathrm{d}c}{\mathrm{d}a_j^L} = -\frac{1}{\ln(b)}\frac{t_j}{a_j^L}$$

Matrix form

$$c = -\boldsymbol{t}^T \cdot \log_b(\boldsymbol{a}^L)$$

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} = -\frac{1}{\ln(b)}(\boldsymbol{t} \oslash \boldsymbol{a}^L) = -\frac{1}{\ln(b)}\begin{pmatrix} \frac{t_0}{a_0^L} \\ \frac{t_1}{a_1^L} \\ \vdots \\ \frac{t_n}{a_n^L} \end{pmatrix}$$

Please see appendix A for full cross entropy derivation.

3 Activation Function

Note that the backpropigation is meaningfully different if the activation function is a vector or scalar function.

3.1 Sigmoid

Element form

$$a_j^L = \sigma(z_j^L) = \frac{1}{1 + e^{-z_j{}^L}}$$

$$\frac{\mathrm{d}a_j^L}{\mathrm{d}z_j^L} = \frac{\mathrm{d}\sigma(z^L)}{\mathrm{d}z^L} = \sigma(z_j^L)\,(1 - \sigma(z_j^L)) = a_j^L\,(1 - a_j^L)$$

Matrix Form

$$\boldsymbol{a}^L = \sigma(\boldsymbol{z}^L) = \frac{1}{1 + e^{-\boldsymbol{z}^L}}$$

$$\frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L} = \frac{\mathrm{d}\sigma(\boldsymbol{z}^L)}{\mathrm{d}\boldsymbol{z}^L} = \sigma(\boldsymbol{z}^L) \circ (1 - \sigma(\boldsymbol{z}^L)) = \boldsymbol{a}^L \circ (1 - \boldsymbol{a}^L)$$

3.2 Softmax

Element form

$$a_j^L = \frac{e^{z_j^L}}{\sum_{i=0}^{n} e^{z_i^L}}$$

$$\text{where } \delta \text{ is the Kronecker delta: } \delta_{jk} \begin{cases} 1 \text{ if } j = k \\ 0 \text{ if } j \neq k \end{cases}$$

$$\frac{\mathrm{d}a_j^L}{\mathrm{d}z_k^L} = a_j^L\,(\delta_{jk} - a_k^L)$$

Matrix form

$$\boldsymbol{a}^L = S(\boldsymbol{z}^L) = \frac{e^{\boldsymbol{z}^L}}{\sum_{i=0}^{n} e^{z_i^L}} = \frac{e^{\boldsymbol{z}^L}}{e^{\boldsymbol{z}_0^L} + e^{\boldsymbol{z}_1^L} + \cdots + e^{\boldsymbol{z}_n^L}}$$

$$\frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L} = \frac{\mathrm{d}S(\boldsymbol{z}^L)}{\mathrm{d}\boldsymbol{z}^L} = \begin{pmatrix} a_0^L\,(1 - a_0^L) & -a_0^L\,a_1^L & \cdots & -a_0^L\,a_n^L \\ -a_1^L\,a_0^L & a_1^L\,(1 - a_1^L) & \cdots & -a_1^L\,a_n^L \\ \vdots & \vdots & \ddots & \vdots \\ -a_n^L\,a_0^L & -a_n^L\,a_1^L & \cdots & a_n^L\,(1 - a_n^L) \end{pmatrix}$$

Please see appendix B for full softmax derivation.

4 Matrix weights

Element form

$$z_j^L = a_k^{L-1} w_{\mathrm{jk}}^L + b_j^L$$
$$\frac{\mathrm{d}z_j^L}{\mathrm{d}w_{jk}^L} = a_k^{L-1}$$
$$\frac{\mathrm{d}z_j^L}{\mathrm{d}b_j^L} = 1$$
$$\frac{\mathrm{d}z_j^L}{\mathrm{d}a_k^{L-1}} = w_{jk}^L$$

Matrix form

$$\boldsymbol{z}^L = W^L \cdot \boldsymbol{a}^{L-1} + \boldsymbol{b}^L$$
$$= \begin{pmatrix} w_{00} & w_{01} & \cdots & w_{0n} \\ w_{10} & w_{11} & \cdots & w_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n0} & w_{n1} & \cdots & w_{\mathrm{nn}} \end{pmatrix} \begin{pmatrix} a_0^{L-1} \\ a_1^{L-1} \\ \vdots \\ a_n^{L-1} \end{pmatrix} + \begin{pmatrix} b_0^L \\ b_1^L \\ \vdots \\ b_n^L \end{pmatrix}$$
$$\frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}W^L} = \boldsymbol{a}^{L-1}$$
$$\frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}\boldsymbol{b}^L} = 1$$
$$\frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}\boldsymbol{a}^{L-1}} = W^L$$

5 Backpropigation

5.1 Sigmoid and squared error

5.1.1 Element form

Layer L

$$\frac{\mathrm{d}c}{\mathrm{d}w_{\mathrm{jk}}^{L}} = \frac{\mathrm{d}c}{\mathrm{d}a_j^L}\frac{\mathrm{d}a_j^L}{\mathrm{d}z_j^L}\frac{\mathrm{d}z_j^L}{\mathrm{d}w_{\mathrm{jk}}^L}$$

$$\frac{\mathrm{d}c}{\mathrm{d}a_j^L} = t_j - a_j^L$$

$$\frac{\mathrm{d}a_j^L}{\mathrm{d}z_j^L} = a_j^L\,(1 - a_j^L)$$

$$\frac{\mathrm{d}c}{\mathrm{d}z_j^L} = \frac{\mathrm{d}c}{\mathrm{d}a_j^L}\frac{\mathrm{d}a_j^L}{\mathrm{d}z_j^L} = (t_j - a_j^L)\,a_j^L\,(1 - a_j^L) = (t_j - a_j^L)\,(1 - a_j^L)\,a_j^L$$

$$\frac{\mathrm{d}z_j^L}{\mathrm{d}w_{\mathrm{jk}}^L} = a_k^{L-1}$$

$$\frac{\mathrm{d}c}{\mathrm{d}w_{\mathrm{jk}}^L} = \frac{\mathrm{d}c}{\mathrm{d}z_j^L}\frac{\mathrm{d}z_j^L}{\mathrm{d}w_{\mathrm{jk}}^L} = (t_j - a_j^L)\,(1 - a_j^L)\,a_j^L\,a_k^{L-1}$$

Layer $L-1$

$$\frac{\mathrm{d}c}{\mathrm{d}w_{kq}^{L-1}} = \left(\sum_{j=0}^{n}\frac{\mathrm{d}c}{\mathrm{d}z_j^L}\frac{\mathrm{d}z_j^L}{\mathrm{d}a_k^{L-1}}\right)\frac{\mathrm{d}a_k^{L-1}}{\mathrm{d}z_k^{L-1}}\frac{\mathrm{d}z_k^{L-1}}{\mathrm{d}w_{kq}^{L-1}}$$

$$\frac{\mathrm{d}z_j^L}{\mathrm{d}a_k^{L-1}} = w_{jk}^L$$

$$\frac{\mathrm{d}c}{\mathrm{d}a_k^{L-1}} = \sum_{j=0}^{n}\frac{\mathrm{d}c}{\mathrm{d}z_j^L}\frac{\mathrm{d}z_j^L}{\mathrm{d}a_k^{L-1}} = \sum_{j=0}^{n}(t_j - a_j^L)\,(1 - a_j^L)\,a_j^L\,w_{jk}^L$$

$$\frac{\mathrm{d}a_k^{L-1}}{\mathrm{d}z_k^{L-1}} = a_k^{L-1}\,(1 - a_k^{L-1})$$

$$\frac{\mathrm{d}z_k^{L-1}}{\mathrm{d}w_{kq}^{L-1}} = a_q^{L-1}$$

$$\frac{\mathrm{d}c}{\mathrm{d}w_{kq}^{L-1}} = \frac{\mathrm{d}c}{\mathrm{d}a_k^{L-1}}\frac{\mathrm{d}a_k^{L-1}}{\mathrm{d}z_k^{L-1}}\frac{\mathrm{d}z_k^{L-1}}{\mathrm{d}w_{kq}^{L-1}} = \left(\sum_{j=0}^{n}(t_j - a_j^L)\,(1 - a_j^L)\,a_j^L\,w_{jk}^L\right)(a_k^{L-1}\,(1 - a_k^{L-1}))\,(a_q^{L-1})$$

5.1.2 Matrix form

Layer L

$$\frac{\mathrm{d}c}{\mathrm{d}W^L} = \left(\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} \circ \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L}\right) \cdot \frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}W^L}^T$$

$$\left(\begin{pmatrix} t_0 - a_0^L \\ t_1 - a_1^L \\ \vdots \\ t_n - a_n^L \end{pmatrix} \circ \begin{pmatrix} a_0^L\,(1 - a_0^L) \\ a_1^L\,(1 - a_1^L) \\ \vdots \\ a_n^L\,(1 - a_n^L) \end{pmatrix}\right) \cdot \begin{pmatrix} a_0^{L-1} \\ a_1^{L-1} \\ \vdots \\ a_n^{L-1} \end{pmatrix}^T$$

$$= \begin{pmatrix} (t_0 - a_0^L)\,(1 - a_0^L)\,a_0^L\,a_0^{L-1} & (t_0 - a_0^L)\,(1 - a_0^L)\,a_0^L\,a_1^{L-1} & \cdots & (t_0 - a_0^L)\,(1 - a_0^L)\,a_0^L\,a_n^{L-1} \\ (t_1 - a_1^L)\,(1 - a_1^L)\,a_1^L\,a_0^{L-1} & (t_1 - a_1^L)\,(1 - a_1^L)\,a_1^L\,a_1^{L-1} & \cdots & (t_1 - a_1^L)\,(1 - a_1^L)\,a_1^L\,a_n^{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ (t_n - a_n^L)\,(1 - a_n^L)\,a_n^L\,a_0^{L-1} & (t_n - a_n^L)\,(1 - a_n^L)\,a_n^L\,a_1^{L-1} & \cdots & (t_n - a_n^L)\,(1 - a_n^L)\,a_n^L\,a_n^{L-1} \end{pmatrix}$$

Layer $L-1$

$$\frac{\mathrm{d}c}{\mathrm{d}W^{L-1}} = \left(\left(\frac{\mathrm{d}c}{\mathrm{d}z^L}^T \cdot \frac{\mathrm{d}z^L}{\mathrm{d}a^{L-1}}\right) \circ \frac{\mathrm{d}a^{L-1}}{\mathrm{d}z^{L-1}}\right) \cdot \frac{\mathrm{d}z^{L-1}}{\mathrm{d}w^{L-1}}$$

$$\frac{\mathrm{d}z^L}{\mathrm{d}a^{L-1}} = W^L$$

$$\frac{\mathrm{d}a^{L-1}}{\mathrm{d}z^{L-1}} = \begin{pmatrix} a_0^{L-1}\,(1 - a_0^{L-1}) \\ a_1^{L-1}\,(1 - a_1^{L-1}) \\ \vdots \\ a_n^{L-1}\,(1 - a_n^{L-1}) \end{pmatrix}$$

$$\frac{\mathrm{d}z^{L-1}}{\mathrm{d}w^{L-1}} = a^{L-2}$$

5.2 Softmax and cross entropy

5.2.1 Element form

On the left $\boldsymbol{a}^L$ is left in vector form and the dot product is taken. Instead on the right the summation of the elements is calculated but they are equivalent.

$$\frac{\mathrm{d}c}{\mathrm{d}w_{\mathrm{jk}}^L} = \left(\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L}^T \cdot \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}z_j^L} \right) \frac{\mathrm{d}z_j{}^L}{\mathrm{d}w_{\mathrm{jk}}^L} \qquad \frac{\mathrm{d}c}{\mathrm{d}w_{\mathrm{jk}}^L} = \left(\sum_{i=0}^{n} \frac{\mathrm{d}c}{\mathrm{d}a_i^L} \frac{\mathrm{d}a_i^L}{\mathrm{d}z_j^L} \right) \frac{\mathrm{d}z_j{}^L}{\mathrm{d}w_{\mathrm{jk}}^L}$$

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} = -\frac{1}{\ln(b)} \begin{pmatrix} \frac{t_0}{a_0^L} \\ \frac{t_1}{a_1^L} \\ \vdots \\ \frac{t_n}{a_n^L} \end{pmatrix} \qquad \frac{\mathrm{d}c}{\mathrm{d}a_i^L} = -\frac{1}{\ln(b)} \frac{t_i}{a_i^L}$$

$$\frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}z_j^L} = \begin{pmatrix} a_0^L\,(\delta_{0j} - a_j^L) \\ a_1^L\,(\delta_{1j} - a_j^L) \\ \vdots \\ a_n^L\,(\delta_{nj} - a_j^L) \end{pmatrix} \qquad \frac{\mathrm{d}a_i^L}{\mathrm{d}z_j^L} = a_i^L\,(\delta_{\mathrm{ij}} - a_j^L)$$

$$\frac{\mathrm{d}c}{\mathrm{d}z_j^L} = \frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L}^T \cdot \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}z_j^L} \qquad \frac{\mathrm{d}c}{\mathrm{d}z_j^L} = \sum_{i=0}^{n} \frac{\mathrm{d}c}{\mathrm{d}a_i^L} \frac{\mathrm{d}a_i^L}{\mathrm{d}z_j^L}$$

The multiplication of both forms is too long to write side by side

$$\frac{\mathrm{d}c}{\mathrm{d}z_j^L} = \sum_{i=0}^{n} \frac{\mathrm{d}c}{\mathrm{d}a_i^L} \frac{\mathrm{d}a_i^L}{\mathrm{d}z_j^L} = \sum_{i=0}^{n} \left(-\frac{1}{\ln(b)} \frac{t_i}{a_i^L} a_i^L\,(\delta_{\mathrm{ij}} - a_j^L) \right) = -\frac{1}{\ln(b)} \sum_{i=0}^{n} \left(\frac{t_i}{a_i^L} a_i^L\,(\delta_{\mathrm{ij}} - a_j^L) \right) =$$

$$\frac{\mathrm{d}c}{\mathrm{d}z_j^L} = \frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L}^T \cdot \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}z_j^L} = -\frac{1}{\ln(b)} \begin{pmatrix} \frac{t_0}{a_0^L} \\ \frac{t_1}{a_1^L} \\ \vdots \\ \frac{t_n}{a_n^L} \end{pmatrix}^T \cdot \begin{pmatrix} a_0^L\,(\delta_{0j} - a_j^L) \\ a_1^L\,(\delta_{1j} - a_j^L) \\ \vdots \\ a_n^L\,(\delta_{nj} - a_j^L) \end{pmatrix} = -\frac{1}{\ln(b)} \sum_{i=0}^{n} \left(\frac{t_i}{a_i^L} a_i^L\,(\delta_{\mathrm{ij}} - a_j^L) \right) =$$

The two forms now converge to the same equation. The Kronecker delta can be removed from the equation by taking the single case $i = j$ out of the equation and summing across all other cases $(i \neq j)$.

$$-\frac{1}{\ln(b)} \left(t_j\,(1 - a_j^L) - \sum_{i \neq j}^{n} t_i\,a_j^L \right) = -\frac{1}{\ln(b)} \left(t_j\,(1 - a_j^L) - a_j^L \sum_{i \neq j}^{n} t_i \right)$$

As all the outputs of the softmax function sum to 1 ($\sum_i^n t_i = 1$). We get the following that allows us to remove the summation: $\sum_{i \neq j}^n t_i = \sum_i^n t_i - t_j = 1 - t_j$

$$= -\frac{1}{\ln(b)} (t_j\,(1 - a_j^L) - a_j^L (1 - t_j)) =$$

$$= -\frac{1}{\ln(b)} (t_j - a_j^L\,t_j - a_j^L + a_j^L\,t_j) = -\frac{1}{\ln(b)} (t_j - a_j^L)$$

$$= \frac{\mathrm{d}c}{\mathrm{d}z_j^L} = \frac{1}{\ln(b)} (a_j^L - t_j)$$

Now to use $\frac{\mathrm{d}c}{\mathrm{d}z_j^L}$ in the final calculation

$$\frac{\mathrm{d}z_j{}^L}{\mathrm{d}w_{\mathrm{jk}}^L} = a_k^{L-1}$$

$$\frac{\mathrm{d}c}{\mathrm{d}w_{\mathrm{jk}}^L} = \left(\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L}^T \cdot \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}z_j^L}\right) \frac{\mathrm{d}z_j{}^L}{\mathrm{d}w_{\mathrm{jk}}^L} = \frac{\mathrm{d}c}{\mathrm{d}z_j^L} \frac{\mathrm{d}z_j{}^L}{\mathrm{d}w_{\mathrm{jk}}^L} = \frac{1}{\ln(b)} (a_j^L - t_j)\, a_k^{L-1}$$

5.2.2 Matrix form

$$\frac{\mathrm{d}c}{\mathrm{d}W^L} = \left(\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L}^T \cdot \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L}\right) \cdot \frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}W^L}^T$$

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} = -\frac{1}{\ln(b)} \begin{pmatrix} \frac{t_0}{a_0^L} \\ \frac{t_1}{a_1^L} \\ \vdots \\ \frac{t_n}{a_n^L} \end{pmatrix}$$

$$\frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L} = \begin{pmatrix} a_0^L\,(1-a_0^L) & -a_0^L\, a_1^L & \cdots & -a_0^L\, a_n^L \\ -a_1^L\, a_0^L & a_1^L\,(1-a_1^L) & \cdots & -a_1^L\, a_n^L \\ \vdots & \vdots & \ddots & \vdots \\ -a_n^L\, a_0^L & -a_n^L\, a_1^L & \cdots & a_n^L\,(1-a_n^L) \end{pmatrix}$$

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{z}^L} = \frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L}^T \cdot \frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L} = -\frac{1}{\ln(b)} \begin{pmatrix} t_0\,(1-a_0^L) - t_1\, a_0^L + \cdots + -t_n\, a_0^L \\ -t_0\, a_1^L + t_1\,(1-a_1^L) + \cdots + -t_n\, a_1^L \\ \vdots \\ -t_0\, a_n^L - t_1\, a_n^L + \ldots + t_n\,(1-a_n^L) \end{pmatrix}$$

Please see element form above for the simplification of the summation

$$= -\frac{1}{\ln(b)} \begin{pmatrix} t_0 - a_0^L \\ t_1 - a_1^L \\ \vdots \\ t_n - a_n^L \end{pmatrix} = -\frac{1}{\ln(b)} (\boldsymbol{t} - \boldsymbol{a}^L) = \frac{1}{\ln(b)} (\boldsymbol{a}^L - \boldsymbol{t})$$

$$\frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}W^L} = \begin{pmatrix} a_0^{L-1} \\ a_1^{L-1} \\ \vdots \\ a_n^{L-1} \end{pmatrix}$$

$$\frac{\mathrm{d}c}{\mathrm{d}W^L} = \frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{z}^L} \cdot \frac{\mathrm{d}\boldsymbol{z}^L}{\mathrm{d}W^L}^T = \frac{1}{\ln(b)} \begin{pmatrix} (a_0^L - t_0)\, a_0^{L-1} & (a_0^L - t_0)\, a_1^{L-1} & \cdots & (a_0^L - t_0)\, a_n^{L-1} \\ (a_1^L - t_1)\, a_0^{L-1} & (a_1^L - t_1)\, a_1^{L-1} & \cdots & (a_1^L - t_1)\, a_n^{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ (a_n^L - t_n)\, a_0^{L-1} & (a_n^L - t_n)\, a_1^{L-1} & \cdots & (a_n^L - t_n)\, a_n^{L-1} \end{pmatrix}$$

$$= \frac{1}{\ln(b)} ((\boldsymbol{a}^{\boldsymbol{L}} - \boldsymbol{t}) \cdot \boldsymbol{a}^{L-1^T})$$

A Cross entropy derivation

To get the derivative of cross entropy, the following rules are required:

Derivative of a scalar function with respect to a vector is a vector

$$\frac{\mathrm{d}y}{\mathrm{d}\boldsymbol{x}} = \begin{pmatrix} \frac{\mathrm{d}y}{\mathrm{d}x_0} \\ \frac{\mathrm{d}y}{\mathrm{d}x_1} \\ \vdots \\ \frac{\mathrm{d}y}{\mathrm{d}x_n} \end{pmatrix}$$

Derivative of a logarithm

$$f(x) = \log_b(x) \qquad f'(x) = \frac{1}{x\ln(b)}$$

Chain rule

$$c(z) = f(z) + g(z) \qquad c'(z) = f'(z) + g'(z)$$

The cross entropy function takes a vector as an input and produces a scalar as an output so it is a scalar function.

$$f\colon \mathbb{R}^N \to \mathbb{R}$$

The cross entropy function is:

$$c = -\sum_{i=0}^{n} t_i(\log_b(a_i)) = -(t_0\log_b(a_0^L) + t_1\log_b(a_1^L)) + \cdots + t_n\log_b(a_n^L))$$

The derivative of the a scalar function with with respect to a vector will be a vector of all the partial derivatives.

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} = \begin{pmatrix} \frac{\mathrm{d}c}{\mathrm{d}a_0^L} \\ \frac{\mathrm{d}c}{\mathrm{d}a_1^L} \\ \vdots \\ \frac{\mathrm{d}c}{\mathrm{d}a_n^L} \end{pmatrix}$$

So to find the derivative for an element in $\frac{\mathrm{d}c}{\mathrm{d}a^L}$ where j is the element

$$\frac{\mathrm{d}c}{\mathrm{d}a_j^L} = -\frac{\mathrm{d}}{\mathrm{d}a_j^L}\sum_{i=0}^{n} t_i\log_b(a_i^L)) = -\frac{\mathrm{d}}{\mathrm{d}a_j^L}(t_0\log_b(a_0^L)) + t_1\log_b(a_1^L)) + \cdots + t_n\log_b(a_n^L)))$$

Every term where $i \neq j$ will be 0 so we can ignore all terms except where $i = j$:

$$\frac{\mathrm{d}c}{\mathrm{d}a_j^L} = -\frac{\mathrm{d}}{\mathrm{d}a_j^L}\, t_j(\log_b(a_j)) = -t_j\, \frac{\mathrm{d}}{\mathrm{d}a_j^L}\, (\log_b(a_j^L)) = -\frac{t_j}{a_j^L \ln(b)} = -\frac{1}{\ln(b)}\, \frac{t_j}{a_j^L}$$

So the derivative is:

$$\frac{\mathrm{d}c}{\mathrm{d}\boldsymbol{a}^L} = \begin{pmatrix} \frac{\mathrm{d}c}{\mathrm{d}a_0^L} \\ \frac{\mathrm{d}c}{\mathrm{d}a_1^L} \\ \vdots \\ \frac{\mathrm{d}c}{\mathrm{d}a_n^L} \end{pmatrix} = -\frac{1}{\ln(b)} \begin{pmatrix} \frac{t_0}{a_0^L} \\ \frac{t_1}{a_1^L} \\ \vdots \\ \frac{t_n}{a_n^L} \end{pmatrix} = -\frac{1}{\ln(b)}\, \frac{\boldsymbol{t}}{\boldsymbol{a}^L}$$

B Softmax derivation

To get the derivative of softmax, the following rules are required:

Derivative of a vector function with respect to a vector is a jacobian matrix

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}\boldsymbol{x}} = \begin{pmatrix} \frac{\mathrm{d}y_0}{\mathrm{d}x_0} & \frac{\mathrm{d}y_1}{\mathrm{d}x_0} & \cdots & \frac{\mathrm{d}y_n}{\mathrm{d}x_0} \\ \frac{\mathrm{d}y_0}{\mathrm{d}x_1} & \frac{\mathrm{d}y_1}{\mathrm{d}x_1} & \cdots & \frac{\mathrm{d}y_n}{\mathrm{d}x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}y_0}{\mathrm{d}x_n} & \frac{\mathrm{d}y_1}{\mathrm{d}x_n} & \cdots & \frac{\mathrm{d}y_n}{\mathrm{d}x_n} \end{pmatrix}$$

total derivative is the sum of all partial derivatives

$$f'(x,y) = f'(x) + f'(y)$$

law of independence

$$\frac{\mathrm{d}e^z}{\mathrm{d}e^x} = 0$$

derivative of exponentials

$$f(x) = e^x \qquad f'(x) = e^x$$

chain rule

$$c(z) = f(z) + g(z) \qquad c'(z) = f'(z) + g'(z)$$

quotient rule

$$q(z) = \frac{f(z)}{g(z)} \qquad q'(z) = \frac{g(z)\, f'(z) - f(z)\, g'(z)}{g(z)^2}$$

The softmax function takes a vector as an input and produces a vector as an output so it is a vector function.

$$f: \mathbb{R}^N \rightarrow \mathbb{R}^N$$

The softmax function is:

$$a^L = \frac{e^{\boldsymbol{z}^L}}{\sum_{i=0}^{n} e^{z_i^L}} = \begin{pmatrix} \frac{e^{z_0^L}}{\sum_{i=0}^{n} e^{z_i^L}} \\ \frac{e^{z_1^L}}{\sum_{i=0}^{n} e^{z_i^L}} \\ \vdots \\ \frac{e^{z_n^L}}{\sum_{i=0}^{n} e^{z_n^L}} \end{pmatrix}$$

The derivative of the a vector function with with respect to a vector will be the jacobian matrix of all the partial derivatives.

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}\boldsymbol{z}} = \begin{pmatrix} \frac{\mathrm{d}a_0}{\mathrm{d}z_0} + \frac{\mathrm{d}a_0}{\mathrm{d}z_1} + \ldots + \frac{\mathrm{d}a_0}{\mathrm{d}z_n} \\ \frac{\mathrm{d}a_1}{\mathrm{d}z_0} + \frac{\mathrm{d}a_1}{\mathrm{d}z_1} + \ldots + \frac{\mathrm{d}a_1}{\mathrm{d}z_n} \\ \vdots \\ \frac{\mathrm{d}a_n}{\mathrm{d}z_0} + \frac{\mathrm{d}a_n}{\mathrm{d}z_1} + \ldots + \frac{\mathrm{d}a_n}{\mathrm{d}z_n} \end{pmatrix}$$

Let's find the partial derivative $\frac{\mathrm{d}a_i}{\mathrm{d}z_j}$. To do this we will break up the numerator and denominator into separate functions to make applying the quotient rule easier. Note that $\frac{\mathrm{d}f(z_i)}{\mathrm{d}z}$ has two separate cases for when $j = k$ and $j \neq k$.

$$a_j = S(z_j) = \frac{e^{z_j}}{\sum_{i=0}^{n} e^{z_i}} = \frac{f(z_j)}{g(z_j)}$$

$$\begin{aligned} & & f(z_j) = e^{z_j} \\ & \text{where } j = k & \frac{\mathrm{d}f(z_j)}{\mathrm{d}z_k} = \frac{\mathrm{d}f(z_j)}{\mathrm{d}z_j} = e^{z_j} \\ & \text{where } j \neq k & \frac{\mathrm{d}f(z_j)}{\mathrm{d}z_k} = 0 \\ g(z_j) = \sum_{i=0}^{n} e^{z_i} = e^{z_0} + e^{z_1} + \cdots + e^{z_n} & & \frac{\mathrm{d}g(z_j)}{\mathrm{d}z_k} = e^{z_k} \end{aligned}$$

when $j = k$

$$\begin{aligned} \frac{\mathrm{d}a_j}{\mathrm{d}z_k} = \frac{\mathrm{d}a_j}{\mathrm{d}z_j} = \frac{\mathrm{d}S(z_j)}{\mathrm{d}z_j} = \frac{g(z_j)\, e^{z_j} - e^{z_j}\, e^{z_j}}{g(z_j)^2} = \frac{e^{z_j}\,(g(z_j) - e^{z_j})}{g(z_j)^2} = \frac{e^{z_j}}{g(z_j)} \frac{g(z_j) - e^{z_j}}{g(z_j)} \\ = \frac{e^{z_j}}{g(z_j)} \left(\frac{g(z_j)}{g(z_j)} - \frac{e^{z_j}}{g(z_j)} \right) = S(z_j)\,(1 - S(z_j)) = a_j\,(1 - a_j) \end{aligned}$$

when $j \neq k$

$$\frac{\mathrm{d}a_j}{\mathrm{d}z_k} = \frac{\mathrm{d}S(z_j)}{\mathrm{d}z_k} = \frac{g(z_j)\, 0 - e^{z_j}\, e^{z_k}}{g(z_j)^2} = \frac{-e^{z_j}\, e^{z_k}}{g(z_j)^2} = \frac{-e^{z_j}}{g(z_j)} \frac{e^{z_k}}{g(z_j)} = -S(z_j)\, S(z_k) = -a_j\, a_k$$

We can create the jacobian matrix which contains all the partial derivatives for all elements. Note I_n is an identity matrix of size n*n.

$$\frac{\mathrm{d}\boldsymbol{a}^L}{\mathrm{d}\boldsymbol{z}^L} = \begin{pmatrix} a_0^L\,(1-a_0^L) & -a_0^L\,a_1^L & \cdots & -a_0^L\,a_n^L \\ -a_1^L\,a_0^L & a_1^L\,(1-a_1^L) & \cdots & -a_1^L\,a_n^L \\ \vdots & \vdots & \ddots & \vdots \\ -a_n^L\,a_0^L & -a_n^L\,a_1^L & \cdots & a_n^L\,(1-a_n^L) \end{pmatrix}$$

$$= \begin{pmatrix} a_0^L & a_1^L & \cdots & a_n^L \end{pmatrix} \begin{pmatrix} (1-a_0^L) & -a_1^L & \cdots & -a_n^L \\ -a_0^L & (1-a_1^L) & \cdots & -a_n^L \\ \vdots & \vdots & \ddots & \vdots \\ -a_0^L & -a_1^L & \cdots & (1-a_n^L) \end{pmatrix}$$

$$= \begin{pmatrix} a_0^L & a_1^L & \cdots & a_n^L \end{pmatrix} \left(I_n - \begin{pmatrix} a_0^L & a_1^L & \cdots & a_n^L \\ a_0^L & a_1^L & \cdots & a_n^L \\ \vdots & \vdots & \ddots & \vdots \\ a_0^L & a_1^L & \cdots & a_n^L \end{pmatrix} \right)$$

This can also be written with the Kronecker delta where $\delta_{\mathrm{jk}} \left\{ \begin{smallmatrix} 1 \text{ if } j = k \\ 0 \text{ if } j \neq k \end{smallmatrix} \right.$

$$\frac{\mathrm{d}a_j^L}{\mathrm{d}z_k^L} = a_j^L\,(\delta_{jk} - a_k^L)$$