The Mathematics of Dense Neural Networks

Trevor Dick

January 2021

1 Notation

Let L be the number of layers, not including the input layer, which will be designated layer 0. Let $g(z)$ be an activation function for some node in the network. The input layer will accept a vector $\vec{x} \in \mathbb{R}^n$ and the target will be a vector $\vec{y} \in \mathbb{R}^m$. Keep in mind that a dense net is designed to accept one sample at a time. The process of training a dense net can be thought of in time steps, and these time steps are coming from the set of samples in the training set. For example in batch training, each batch uses a subset of the training samples, and if the batch size is not one, then the number of batches is necessarily less than the number of total samples. In this example we have one set of time steps corresponding to each batch and another set of time steps for the samples used within a given batch.

2 Constructing the Network

For each layer $\ell \in \{0, 1, \ldots, L\}$, we have the number of nodes per layer $n_\ell \in \{n_0, n_1, \ldots, n_L\}$. We also have activation functions $g^\ell \in \{g^1, \ldots, g^L\}$ for each layer excluding the input layer. A single node in one layer is connected to all the nodes in the previous layer by a weight matrix $W^\ell_{[n_\ell \times n_{\ell-1}]}$ and each node may add a bias $b^\ell_{\nu_\ell}$ before the activation g^ℓ is applied. In total the network may be expressed by two indices, the layer ℓ and node in that layer ν_ℓ. For each pair of indices (ℓ, ν_ℓ) we can define the following functions that describe the values any node in the network may have:

$$z^\ell_{\nu_\ell} = \left(\sum_{\nu_{\ell-1}=1}^{n_{\ell-1}} w^\ell_{\nu_\ell \nu_{\ell-1}} a^{\ell-1}_{\nu_{\ell-1}} \right) + b^\ell_{\nu_\ell}$$

$$a^\ell_{\nu_\ell} = g^\ell(z^\ell_{\nu_\ell})$$

In vector notation the equations become the following:

$$\vec{z^\ell} = W^\ell \cdot \overrightarrow{a^{\ell-1}} + \vec{b^\ell}$$

$$\vec{a^\ell} = g^\ell(\vec{z^\ell})$$

3 Derivatives w.r.t W and $\vec{b}$

Consider the final activation layer, $\overrightarrow{a^L} = \overrightarrow{y_{pred}}$, since it is used to compare against the target $\vec{y}$. We will consider the index representation for the following calculations.

Differentiate any node in the final layer L, with respect to the p, q-th weights in any layer ℓ. Note: δ_{ij} is the Kronecker delta function.

$$\begin{aligned}
\frac{\partial a^L_{\nu_L}}{\partial w^\ell_{pq}} &= \frac{\partial g^L}{\partial z^L_{\nu_L}} \frac{\partial z^L_{\nu_L}}{\partial w^\ell_{pq}} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \frac{\partial}{\partial w^\ell_{pq}} \left[\left(\sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} a^{L-1}_{\nu_{L-1}} \right) + b^L_{\nu_L} \right] \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial a^{L-1}_{\nu_{L-1}}}{\partial w^\ell_{pq}} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \frac{\partial z^{L-1}_{\nu_{L-1}}}{\partial w^\ell_{pq}} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \frac{\partial}{\partial w^\ell_{pq}} \left[\left(\sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1} \nu_{L-2}} a^{L-2}_{\nu_{L-2}} \right) + b^{L-1}_{\nu_{L-1}} \right] \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1} \nu_{L-2}} \frac{\partial a^{L-2}_{\nu_{L-2}}}{\partial w^\ell_{pq}} \\
&= \vdots \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1} \nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1} \nu_\ell} \frac{\partial a^\ell_{\nu_\ell}}{\partial w^\ell_{pq}} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1} \nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1} \nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \frac{\partial z^\ell_{\nu_\ell}}{\partial w^\ell_{pq}} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1} \nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1} \nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \sum_{\nu_{\ell-1}=1}^{n_{\ell-1}} \frac{\partial w^\ell_{\nu_\ell \nu_{\ell-1}}}{\partial w^\ell_{pq}} a^{\ell-1}_{\nu_{\ell-1}} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1} \nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1} \nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \sum_{\nu_{\ell-1}=1}^{n_{\ell-1}} \delta_{\nu_\ell p} \delta_{\nu_{\ell-1} q} a^{\ell-1}_{\nu_{\ell-1}}
\end{aligned}$$

The result is a $[n_L \times n_\ell \times n_{\ell-1}]$ array, arising from the free indices (ν_L, p, q).

This shape matches our expectations that there should be enough information to encode changes of each node in the final layer w.r.t. changes of any element in the W^ℓ matrix. Somewhat more precisely, during training the loss function is aggregating over the index ν_L leaving a matrix with shape $[n_\ell \times n_{\ell-1}]$, the same as W^ℓ. This new matrix must be of the same shape as W^ℓ so that it can update W^ℓ during learning, typically through back propagation.

We can reuse much of the above work in deriving the equations for derivatives w.r.t $\vec{b^\ell}$. Since w^ℓ_{pq} is just another variable, we can easily replace it with b^ℓ_p, for every place where any function does not explicitly depend on it as we apply the chain rule.

$$
\begin{aligned}
\frac{\partial a^L_{\nu_L}}{\partial b^\ell_p} &= \frac{\partial g^L}{\partial z^L_{\nu_L}} \frac{\partial z^L_{\nu_L}}{\partial b^\ell_p} \\
&= \vdots \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1}\nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1}\nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \frac{\partial z^\ell_{\nu_\ell}}{\partial b^\ell_p} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1}\nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \\
&\quad \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1}\nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \frac{\partial}{\partial b^\ell_p} \left[\left(\sum_{\nu_{\ell-1}=1}^{n_{\ell-1}} w^\ell_{\nu_\ell \nu_{\ell-1}} a^{\ell-1}_{\nu_{\ell-1}} \right) + b^\ell_{\nu_\ell} \right] \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1}\nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1}\nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \frac{\partial b^\ell_{\nu_\ell}}{\partial b^\ell_p} \\
&= \frac{\partial g^L}{\partial z^L_{\nu_L}} \sum_{\nu_{L-1}=1}^{n_{L-1}} w^L_{\nu_L \nu_{L-1}} \frac{\partial g^{L-1}}{\partial z^{L-1}_{\nu_{L-1}}} \sum_{\nu_{L-2}=1}^{n_{L-2}} w^{L-1}_{\nu_{L-1}\nu_{L-2}} \frac{\partial g^{L-2}}{\partial z^{L-2}_{\nu_{L-2}}} \cdots \frac{\partial g^{\ell+1}}{\partial z^{\ell+1}_{\nu_{\ell+1}}} \sum_{\nu_\ell=1}^{n_\ell} w^{\ell+1}_{\nu_{\ell+1}\nu_\ell} \frac{\partial g^\ell}{\partial z^\ell_{\nu_\ell}} \delta_{\nu_\ell p}
\end{aligned}
$$

The resultant object is a $[n_L \times n_\ell]$ array, or matrix in this case. This is also according to the free indices (ν_L, p).

A few things to note. The Kronecker delta function when summed over does behave as expected by selecting out only the one index such that $\delta = 1$. I have chosen not to apply this operation in the formulas, since there is a situation where one might attempt to convert all the above notation over to index (Einstein) notation and arrive at incorrect array shapes. Specifically, the common index notation convention that treats repeated indices as summed over is not compatible with these equations. There are repeated indices that correctly appear more than twice, barring the use of Einstein notation.

3.1 Vector Representation

Let us introduce some minor notation changes to aid in simplifying expressions. Let $\gamma^{\ell}_{\nu_\ell} = \frac{\partial g^{\ell}}{\partial z^{\ell}_{\nu_\ell}}$, and we will begin by rearranging the two derivative equations by bringing all terms within the summations.

$$\frac{\partial a^L_{\nu_L}}{\partial w^{\ell}_{pq}} = \sum_{\nu_{L-1}=1}^{n_{L-1}} \cdots \sum_{\nu_\ell=1}^{n_\ell} \sum_{\nu_{\ell-1}=1}^{n_{\ell-1}} \gamma^L_{\nu_L} w^L_{\nu_L \nu_{L-1}} \gamma^{L-1}_{\nu_{L-1}} w^{L-1}_{\nu_{L-1}\nu_{L-2}} \cdots \gamma^{\ell+1}_{\nu_{\ell+1}} w^{\ell+1}_{\nu_{\ell+1}\nu_\ell} \gamma^{\ell}_{\nu_\ell} \delta_{\nu_\ell p} \delta_{\nu_{\ell-1} q} a^{\ell-1}_{\nu_{\ell-1}}$$

$$\frac{\partial a^L_{\nu_L}}{\partial b^{\ell}_{p}} = \sum_{\nu_{L-1}=1}^{n_{L-1}} \cdots \sum_{\nu_\ell=1}^{n_\ell} \gamma^L_{\nu_L} w^L_{\nu_L \nu_{L-1}} \gamma^{L-1}_{\nu_{L-1}} w^{L-1}_{\nu_{L-1}\nu_{L-2}} \cdots \gamma^{\ell+1}_{\nu_{\ell+1}} w^{\ell+1}_{\nu_{\ell+1}\nu_\ell} \gamma^{\ell}_{\nu_\ell} \delta_{\nu_\ell p}$$

Notice the pattern of terms $\gamma^k_{\nu_k} w^k_{\nu_k \nu_{k-1}}$. Since we are explicitly using summations, this is just another term with indices (ν_k, ν_{k-1}), so let us define:

$$d^k_{\nu_k \nu_{k-1}} = \gamma^k_{\nu_k} w^k_{\nu_k \nu_{k-1}}$$

From this index notation it is clear that the $\overrightarrow{\gamma^k}$ vector is multiplying each column of W^k. Since the product is taken on each element, irrespective of the column index, we can define a new matrix:

$$G^k = \begin{bmatrix} | & & | \\ \overrightarrow{\gamma^k} & \cdots & \overrightarrow{\gamma^k} \\ | & & | \end{bmatrix}_{[n_k \times n_{k-1}]} = \begin{bmatrix} \gamma^k_1 & \cdots & \gamma^k_1 \\ \vdots & \ddots & \vdots \\ \gamma^k_{\nu_k} & \cdots & \gamma^k_{\nu_k} \end{bmatrix}$$

Which is the matrix whose column vectors are made of n_{k-1} copies of $\overrightarrow{\gamma^k}$.

If we look at the Hadamard product, $G^k \odot W^k$, we see that it elements are precisely $\gamma^k_{\nu_k} w^k_{\nu_k \nu_{k-1}}$. Therefore, if we define D^k as the matrix with elements $d^k_{\nu_k \nu_{k-1}}$, then:

$$D^k = G^k \odot W^k$$

For completeness and later necessity, lets also define another matrix that is compatible with multiplication along column indices, rather than just row indices as we just saw:

$$\tilde{G}^k = \begin{bmatrix} - & \overrightarrow{\gamma^k} & - \\ & \vdots & \\ - & \overrightarrow{\gamma^k} & - \end{bmatrix}_{[n_{k+1} \times n_k]}$$

With the tilde, we will say this is the matrix whose row vectors are made of n_{k+1} copies of $\overrightarrow{\gamma^k}$.

Going back to our two derivative equations, we replace the terms $\gamma^k_{\nu_k} w^k_{\nu_k \nu_{k-1}}$ with $d^k_{\nu_k \nu_{k-1}}$ and apply the appropriate sums over the Kronecker delta functions.

$$\frac{\partial a^L_{\nu_L}}{\partial w^\ell_{pq}} = \sum_{\nu_{L-1}=1}^{n_{L-1}} \cdots \sum_{\nu_{\ell+1}=1}^{n_{\ell+1}} d^L_{\nu_L \nu_{L-1}} d^{L-1}_{\nu_{L-1} \nu_{L-2}} \cdots d^{\ell+2}_{\nu_{\ell+2} \nu_{\ell+1}} d^{\ell+1}_{\nu_{\ell+1} p} \gamma^\ell_p a^{\ell-1}_q$$

$$\frac{\partial a^L_{\nu_L}}{\partial b^\ell_p} = \sum_{\nu_{L-1}=1}^{n_{L-1}} \cdots \sum_{\nu_{\ell+1}=1}^{n_{\ell+1}} d^L_{\nu_L \nu_{L-1}} d^{L-1}_{\nu_{L-1} \nu_{L-2}} \cdots d^{\ell+2}_{\nu_{\ell+2} \nu_{\ell+1}} d^{\ell+1}_{\nu_{\ell+1} p} \gamma^\ell_p$$

Notice that for each pair $d^{k+1}_{\nu_{k+1} \nu_k}$ and $d^k_{\nu_k \nu_{k-1}}$, we sum over ν_k explicitly like so $\sum_{\nu_k=1}^{n_k} d^{k+1}_{\nu_{k+1} \nu_k} d^k_{\nu_k \nu_{k-1}}$. This is exactly the dot product of the two matrices, D^{k+1} and D^k, and we have one dot product for each summation remaining in each expression.

This now leaves the matter of handling $\gamma^\ell_p a^{\ell-1}_q$ and γ^ℓ_p, which remain unaccounted for, thus far. Let us examine the term $d^{\ell+1}_{\nu_{\ell+1} p} \gamma^\ell_p$ and see that the vector $\overrightarrow{\gamma^\ell}$ is multiplying every row in $D^{\ell+1}$. This can be expressed as another Hadamard product of two matrices. We already have a well defined matrix operation representing $d^{\ell+1}_{\nu_{\ell+1} p} \gamma^\ell_p$, which is $D^{\ell+1} \odot \tilde{G}^\ell$. Lastly the term $a^{\ell-1}_q$ has an index independent of all others. In fact it multiplies every element distinctly for every q. The tensor product $\otimes$ is precisely the operation used here.

Note, alternatively, we could have defined a left and right sided Hadamard product, but it seems better to preserve the commutative nature of the Hadamard product and instead use the $\tilde{G}$ transpose notation to the matrix G.

Finally we can reduce our two equations to only matrix dot product, Hadamard product, and tensor product operations.

$$\frac{\partial \overrightarrow{a^L}}{\partial W^\ell} = \left(D^L \cdot D^{L-1} \cdots D^{\ell+2} \cdot (D^{\ell+1} \odot \tilde{G}^\ell) \right) \otimes \overrightarrow{a^{\ell-1}}$$

$$\frac{\partial \overrightarrow{a^L}}{\partial \overrightarrow{b^\ell}} = D^L \cdot D^{L-1} \cdots D^{\ell+2} \cdot (D^{\ell+1} \odot \tilde{G}^\ell)$$

$$D^k = G^k \odot W^k$$

3.2 Examples

1. $L = 5$, $\ell = 3$

$$\begin{aligned}\frac{\partial a^5_{\nu_5}}{\partial w^3_{pq}} &= \frac{\partial g^5}{\partial z^5_{\nu_5}} \sum_{\nu_4=1}^{n_4} w^5_{\nu_5\nu_4} \frac{\partial g^4}{\partial z^4_{\nu_4}} \sum_{\nu_3=1}^{n_3} w^4_{\nu_4\nu_3} \frac{\partial g^3}{\partial z^3_{\nu_3}} \sum_{\nu_2=1}^{n_2} \delta_{\nu_3 p}\delta_{\nu_2 q} a^2_{\nu_2} \\ &= \frac{\partial g^5}{\partial z^5_{\nu_5}} \sum_{\nu_4=1}^{n_4} w^5_{\nu_5\nu_4} \frac{\partial g^4}{\partial z^4_{\nu_4}} w^4_{\nu_4 p} \frac{\partial g^3}{\partial z^3_p} a^2_q\end{aligned}$$

   Array with shape $[n_5 \times n_3 \times n_2]$

2. $L = 2$, $\ell = 2$

$$\begin{aligned}\frac{\partial a^2}{\partial w^2_{pq}} &= \frac{\partial g^2}{\partial z^2_{\nu_2}} \sum_{\nu_1=1}^{n_1} \delta_{\nu_2 p}\delta_{\nu_1 q} a^1_{\nu_1} \\ &= \frac{\partial g^2}{\partial z^2_{\nu_2}} \delta_{\nu_2 p} a^1_q\end{aligned}$$

   Array with shape $[n_2 \times n_2 \times n_1]$

$$\frac{\partial a^2}{\partial b^2_p} = \frac{\partial g^2}{\partial z^2_{\nu_2}} \delta_{\nu_2 p}$$

   Array with shape $[n_2 \times n_2]$

3. $L = 2$, $\ell = 1$

$$\begin{aligned}\frac{\partial a^2_{\nu_2}}{\partial w^1_{pq}} &= \frac{\partial g^2}{\partial z^2_{\nu_2}} \sum_{\nu_1=1}^{n_1} w^2_{\nu_2\nu_1} \frac{\partial g^1}{\partial z^1_{\nu_1}} \sum_{\nu_0=1}^{n_0} \delta_{\nu_1 p}\delta_{\nu_0 q} a^0_{\nu_0} \\ &= \frac{\partial g^2}{\partial z^2_{\nu_2}} w^2_{\nu_2 p} \frac{\partial g^1}{\partial z^1_p} a^0_q\end{aligned}$$

   Array with shape $[n_2 \times n_1 \times n_0]$

$$\begin{aligned}\frac{\partial a^2_{\nu_2}}{\partial b^1_p} &= \frac{\partial g^2}{\partial z^2_{\nu_2}} \sum_{\nu_1=1}^{n_1} w^2_{\nu_2\nu_1} \frac{\partial g^1}{\partial z^1_{\nu_1}} \delta_{\nu_1 p} \\ &= \frac{\partial g^2}{\partial z^2_{\nu_2}} w^2_{\nu_2 p} \frac{\partial g^1}{\partial z^1_p}\end{aligned}$$

   Array with shape $[n_2 \times n_1]$

4 Loss and Cost

4.1 Loss

If we first define the metric function on $\mathbb{R}^n$,

$$d(\vec{u}, \vec{v}) : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$$

and a projection function

$$\pi : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

then loss is the image of the projection of the metric:

$$L : \mathbb{R}^n \longrightarrow \mathbb{R}$$

where $d(\vec{u}, \vec{v}) = (L \circ \pi)(\vec{u}, \vec{v}) = L(\pi(\vec{u}, \vec{v})) = L(\vec{u})$.

If we have a prediction vector, $\overrightarrow{y_{pred}}$ and a truth vector $\vec{y}$, then the loss is $L(\overrightarrow{y_{pred}}) = d(\overrightarrow{y_{pred}}, \vec{y})$

4.2 Cost

Cost is essentially loss, but now taking into account a set of samples $\{\overrightarrow{a^{L1}}, \ldots, \overrightarrow{a^{Lm}}\}$ of the output vectors, $\overrightarrow{a^L} \in \mathbb{R}^n$. The cost function is

$$\mathcal{C} : \Omega \longrightarrow \mathbb{R}$$

where the domain Ω is a finite, $|\Omega| = m$, collection of vectors in $\mathbb{R}^n$.

We define a a few component functions that help to fully describe the various ways in which the cost can be used to aggregate the m vectors while still incorporating the loss.

First define,

$$\phi : \Omega \longrightarrow \mathbb{R}^m$$

such that

$$\phi(\overrightarrow{a^{L1}}, \ldots, \overrightarrow{a^{Lm}}) = (L(\overrightarrow{a^{L1}}), \ldots, L(\overrightarrow{a^{Lm}}))$$

Notice that a tuple of m vectors in $\mathbb{R}^n$, $(\overrightarrow{a^{L1}}, \ldots, \overrightarrow{a^{Lm}})$ can be made isomorphic to a vector $A \in \mathbb{R}^{mn}$, where $A = (a_1^{L1}, \ldots, a_n^{L1}, \ldots, a_1^{Lm}, \ldots, a_n^{Lm})$. Thus, we can also define a map

$$\tilde{\phi} : \mathbb{R}^{mn} \longrightarrow \mathbb{R}^m$$

such that if $A \in \mathbb{R}^{mn}$ and a projection map,

$$\pi_k : \mathbb{R}^{mn} \longrightarrow \mathbb{R}^n$$

where $\pi_k(A) = (a_1^{Lk}, \ldots, a_n^{Lk}) = \overrightarrow{a^{Lk}}$, then $L(\overrightarrow{a^{Lk}}) = (L \circ \pi_k)(A)$. So let

$$L_k = L \circ \pi_k$$

Thus,

$$\begin{aligned}
\tilde{\phi}(A) &= (L_1(A), \ldots, L_m(A)) \\
&= ((L \circ \pi_1)(A), \ldots, (L \circ \pi_m)(A)) \\
&= (L(\overrightarrow{a^{L1}}), \ldots, L(\overrightarrow{a^{Lm}})) \\
&= \phi(\overrightarrow{a^{L1}}, \ldots, \overrightarrow{a^{Lm}})
\end{aligned}$$

Next define the differentiable function (note the similarities to a functional), which gives rise to cost,

$$\varphi : \mathbb{R}^m \longrightarrow \mathbb{R}$$

such that

$$\mathcal{C}(\overrightarrow{a^{L1}}, \ldots, \overrightarrow{a^{Lm}}) = (\varphi \circ \phi)\,(\overrightarrow{a^{L1}}, \ldots, \overrightarrow{a^{Lm}})$$

and

$$\mathcal{C}(A) = (\varphi \circ \tilde{\phi})(A)$$

5 Total Derivative of Cost

We can simply apply the chain rule on cost as follows:

$$\begin{aligned}
D_A\mathcal{C} &= D_A(\varphi \circ \tilde{\phi}) \\
&= D_{\tilde{\phi}(A)}\varphi \cdot D_A\tilde{\phi} \\
&= \nabla\varphi \cdot D_A\tilde{\phi}
\end{aligned}$$

where $\nabla\varphi$ is a $[1 \times m]$ gradient vector.

Let us examine $D_A\tilde{\phi}$ further. Since we know $\tilde{\phi} : \mathbb{R}^{mn} \longrightarrow \mathbb{R}^m$, then the total derivative is a linear map represented by a $[m \times mn]$ matrix. Now recall $\tilde{\phi}(A) = (L_1(A), \ldots, L_m(A))$, therefore:

$$D_A\tilde{\phi} = \begin{bmatrix}
\frac{\partial L_1}{\partial a_1^{L1}} & \cdots & \frac{\partial L_1}{\partial a_n^{L1}} & \cdots & \frac{\partial L_1}{\partial a_1^{Lm}} & \cdots & \frac{\partial L_1}{\partial a_n^{Lm}} \\
\vdots & & & \ddots & & & \vdots \\
\frac{\partial L_m}{\partial a_1^{L1}} & \cdots & \frac{\partial L_m}{\partial a_n^{L1}} & \cdots & \frac{\partial L_m}{\partial a_1^{Lm}} & \cdots & \frac{\partial L_m}{\partial a_n^{Lm}}
\end{bmatrix}$$

$$= \begin{bmatrix}
- & D_AL_1 & - \\
 & \vdots & \\
- & D_AL_m & -
\end{bmatrix}$$

Next we will compute the total derivatives for L_k,

$$\begin{aligned} D_A L_k &= D_A(L \circ \pi_k) \\ &= D_{\pi_k(A)} L \cdot D_A \pi_k \\ &= \nabla L(\overrightarrow{a^{Lk}}) \cdot D_A \pi_k \end{aligned}$$

and since $\pi_k(A) = (a_1^{Lk}, \ldots, a_n^{Lk})$,

$$D_A \pi_k = \left[\begin{array}{c|c|c|c|c} 0 & \cdots & \mathrm{I}_{n\times n} & \cdots & 0 \end{array} \right]_{n \times mn}$$

where the identity matrix is k-th block matrix (m total block matrices of size $n \times n$).

Therefore, we can also express,

$$D_A L_k = \nabla L_k = \left(0, \ldots, \frac{\partial L}{\partial a_1^{Lk}}, \ldots, \frac{\partial L}{\partial a_n^{Lk}}, \ldots, 0\right)$$

where only the k-th, sub-vector of $\mathbb{R}^n$ is non-zero.

Putting these pieces together:

$$\begin{aligned} D_A \mathcal{C} &= \nabla \varphi \cdot D_A \tilde{\phi} \\ &= \nabla \varphi_{[1\times m]} \cdot \begin{bmatrix} - & \nabla L_1 & - \\ & \vdots & \\ - & \nabla L_m & - \end{bmatrix}_{[m \times mn]} \end{aligned}$$

Recall that each element of A is a function of the weights and biases of the network. Using the chain rule once again we compute the total derivatives of cost w.r.t. W^ℓ and $\vec{b^\ell}$:

$$\begin{aligned} D_{W^\ell} \mathcal{C} &= D_{W^\ell}(\varphi \circ \tilde{\phi} \circ A) \\ &= D_{\phi(A)} \varphi \cdot D_{A(W^\ell)} \tilde{\phi} \cdot D_{W^\ell} A \\ &= \nabla \varphi \cdot D_A \tilde{\phi} \cdot D_{W^\ell} A \end{aligned}$$

where $D_{W^\ell} A$ has shape $mn \times n_\ell \times n_{\ell-1}$.

$$\begin{aligned} D_{\vec{b}} \mathcal{C} &= D_{W^\ell}(\varphi \circ \tilde{\phi} \circ A) \\ &= D_{\phi(A)} \varphi \cdot D_{A(W^\ell)} \tilde{\phi} \cdot D_{\vec{b}} A \\ &= \nabla \varphi \cdot D_A \tilde{\phi} \cdot D_{\vec{b}} A \end{aligned}$$

where $D_{\vec{b}} A$ has shape $mn \times n_\ell$.

The formula for $D_{W^\ell} A$ is exactly the same as that for $\frac{\partial \overrightarrow{a^L}}{\partial W^\ell}$, only with a different sized vector. Ignoring a change of notation, where $D_{W^\ell} \overrightarrow{a^L} = \frac{\partial \overrightarrow{a^L}}{\partial W^\ell}$, the vector $\overrightarrow{a^L} \in \mathbb{R}^n$ is simply being replaced with a vector $A \in \mathbb{R}^{mn}$. Thus, we see

in the expression for the shape of the total derivative, $n \times n_\ell \times n_{\ell-1}$, we are only making a substitution of n for mn. The same idea holds for $D_{\vec{b}}\vec{\mathcal{C}}$, since it is still just a substitution of vectors from a different dimension vector space. The derivatives of the final activation layer w.r.t. the weights or biases are true for any finite dimensional Euclidean vector space.

6 Optimizing Parameters

6.1 Gradient Descent