

# Mathematics for Deep Learning

## Partial Derivatives for Backpropagation

---

### Inputs and Output

Let  $x$  denote an input vector. For the moment we'll leave its dimension and how we index its components unspecified.

Similarly, let  $y$  denote the output vector corresponding to the input vector,  $x$ . Its dimension and how we index its components will for a moment also remain unspecified, but note that the dimensions of  $x$  and  $y$  are generally completely different.

---

### Neural Nets and Layers

We will conceive of a neural net as consisting of  $L$  layers. Each layer takes an input vector and produces an output vector. Then

$$y = x^{(L)} \leftarrow x^{(L-1)} \leftarrow x^{(L-2)} \leftarrow \dots \leftarrow x^{(2)} \leftarrow x^{(1)} \leftarrow x^{(0)} = x$$

The parenthesized superscripts denote layer numbers, not exponents. A net with one layer ( $L = 1$ ) has two vectors, consisting of one input vector, one output vector, and no vectors in between. An  $L = 4$  neural net has five vectors  $x^{(0)}$  to  $x^{(4)}$ .

We still haven't said what the dimensions of the vectors are or how we index their components, but again note that the dimensions of the various  $x^{(l)}$ ,  $l = 0, \dots, L$  are in general unrelated.

---

### Functions to Represent Layer Operations

The  $l$ th function representing the layer  $l$  operation transforms  $x^{(l)}$  to  $x^{(l+1)}$ . For clarity — not compactness — we'll write this operation as  $x^{(l+1)} = f^{(l+1 \leftarrow l)}(x^{(l)})$ .

In a neural net with  $L$  layers, there are  $L$  functions  $f$ . The one that operates on the input vector is  $f^{(1 \leftarrow 0)}(x^{(0)})$ , and the one producing the output vector is  $f^{(L \leftarrow L-1)}(x^{(L-1)})$ .

## Parameters

The  $l$ th function also depends on parameters which we will call  $\alpha^{(l)}$ . So for completeness, we need to capture this too:

$$x^{(l+1)} = f^{(l+1 \leftarrow l)}(x^{(l)}; \alpha^{(l)})$$

When we chain the transformation of the  $L$  layers in the neural net together, the complete transformation is

$$x^{(L)} = f^{(L \leftarrow L-1)}(f^{(L-1 \leftarrow L-2)}(\dots f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x^{(0)}; \alpha^{(0)}); \alpha^{(1)}); \dots; \alpha^{(L-2)}); \alpha^{(L-1)})$$

The ellipses lack clarity, so a better way to see the pattern is to write out a non-trivial case, such as  $L = 4$ :

$$L = 4: x^{(3)} = f^{(4 \leftarrow 3)}(f^{(3 \leftarrow 2)}(f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x^{(0)}; \alpha^{(0)}); \alpha^{(1)}); \alpha^{(2)}); \alpha^{(3)})$$

This can be made even clearer by using  $y = x^{(L)}$  and  $x = x^{(0)}$ :

$$y = f^{(4 \leftarrow 3)}(f^{(3 \leftarrow 2)}(f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x; \alpha^{(0)}); \alpha^{(1)}); \alpha^{(2)}); \alpha^{(3)})$$

Actually, let's write out the  $L = 3$ ,  $L = 2$ , and  $L = 1$  cases too, since that is barely more than a bit of copy-and-paste:

$$L = 3: y = f^{(3 \leftarrow 2)}(f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x; \alpha^{(0)}); \alpha^{(1)}); \alpha^{(2)})$$

$$L = 2: y = f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x; \alpha^{(0)}); \alpha^{(1)})$$

$$L = 1: y = f^{(1 \leftarrow 0)}(x; \alpha^{(0)})$$

This very special form of composite function is central to the simplicity of the mathematics that follows.

## Training

The neural net makes predictions that depend on  $x = x^{(0)}$  by computing all the intermediaries  $x^{(l)}$ . However, during training, we are much more keenly interested on the net's dependence on the parameters,  $\alpha^{(l)}$ , which are tuned on the training data to make the predictions better and better, than we are on the values of the  $x^{(l)}$ . A measure of goodness (or badness, which is most commonly called "loss") is computed, and then derivatives of the loss are taken w.r.t. (with respect to) the parameters. Taking the derivatives of the loss function immediately requires us (after one application of the chain rule) to know the derivatives w.r.t. any of the parameters  $\alpha^{(l)}$  of:

$$x^{(L)} = f^{(L \leftarrow L-1)}(f^{(L-1 \leftarrow L-2)}(\dots f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x^{(0)}; \alpha^{(0)}); \alpha^{(1)}) \dots; \alpha^{(L-2)}); \alpha^{(L-1)})$$

Remember that each of the vectors  $x^{(l)}$  has a still-unspecified dimension and each of the vectors  $\alpha^{(l)}$  also has a still-unspecified dimension, and that none of these dimensions in general have anything to do with any of the others.

## A Special Case

### $L = 4$ , and Entirely One-Dimensional

To get the hang of this, let's take a very special case. We'll do  $L = 4$ , and we'll assume that each and every one of  $y = x^{(4)}$ ,  $x^{(3)}$ ,  $x^{(2)}$ ,  $x^{(1)}$ , and  $x = x^{(0)}$ , are one-dimensional, and that  $\alpha^{(3)}$ ,  $\alpha^{(2)}$ ,  $\alpha^{(1)}$ , and  $\alpha^{(0)}$  also happen to be one-dimensional. We want to know the derivatives w.r.t. any of  $\alpha^{(3)}$ ,  $\alpha^{(2)}$ ,  $\alpha^{(1)}$ , and  $\alpha^{(0)}$  of

$$y = f^{(4 \leftarrow 3)}(f^{(3 \leftarrow 2)}(f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x^{(0)}; \alpha^{(0)}); \alpha^{(1)}); \alpha^{(2)}); \alpha^{(3)})$$

Well, let's do one:

$$\frac{\partial y}{\partial \alpha^{(3)}} = \frac{\partial f^{(4 \leftarrow 3)}}{\partial \alpha^{(3)}}$$

That wasn't so bad. Let's do another, but this one will require one application of the chain rule:

$$\frac{\partial y}{\partial \alpha^{(2)}} = \frac{\partial f^{(4 \leftarrow 3)}}{\partial x^{(3)}} \frac{\partial f^{(3 \leftarrow 2)}}{\partial \alpha^{(2)}}$$

Hopefully, you already see the pattern, but if not, write a few more out, and you will get:

$$\frac{\partial y}{\partial \alpha^{(1)}} = \frac{\partial f^{(4 \leftarrow 3)}}{\partial x^{(3)}} \frac{\partial f^{(3 \leftarrow 2)}}{\partial x^{(2)}} \frac{\partial f^{(2 \leftarrow 1)}}{\partial \alpha^{(1)}}$$

$$\frac{\partial y}{\partial \alpha^{(0)}} = \frac{\partial f^{(4 \leftarrow 3)}}{\partial x^{(3)}} \frac{\partial f^{(3 \leftarrow 2)}}{\partial x^{(2)}} \frac{\partial f^{(2 \leftarrow 1)}}{\partial x^{(1)}} \frac{\partial f^{(1 \leftarrow 0)}}{\partial \alpha^{(0)}}$$

## Re-Introducing Dimensionality

Sticking with  $L = 4$ , let's allow the vectors  $y = x^{(4)}$ ,  $x^{(3)}$ ,  $x^{(2)}$ ,  $x^{(1)}$ , and  $x = x^{(0)}$ , to regain dimensionality, and similarly, the vectors  $\alpha^{(3)}$ ,  $\alpha^{(2)}$ ,  $\alpha^{(1)}$ , and  $\alpha^{(0)}$  will also regain dimensionality.

We'll call the dimensions of the  $x^{(l)}$  in the  $L = 4$  case,  $d_4$ ,  $d_3$ ,  $d_2$ , and  $d_1$  (in the partial derivatives, we won't need to refer to  $d_0$ ) and the dimensions of the  $\alpha^{(l)}$  in the  $L = 4$  case,  $n_3$ ,  $n_2$ ,  $n_1$ , and  $n_0$ .

The  $i_l$ th component of  $x^{(l)}$  will be  $x_{i_l}^{(l)}$  and the  $\mu_l$ th component of  $\alpha^{(l)}$  will be  $\alpha_{\mu_l}^{(l)}$ . We want to know:

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_l}^{(l)}} \equiv \frac{\partial x_{i_4}^{(4)}}{\partial \alpha_{\mu_l}^{(l)}}$$

Let's do  $l = 3$  first because it is the easiest:

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_3}^{(3)}} \equiv \frac{\partial x_{i_4}^{(4)}}{\partial \alpha_{\mu_3}^{(3)}} = \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial \alpha_{\mu_3}^{(3)}}$$

We will finally need some multi-variable calculus to actually write out the rest. Here is  $l = 2$ :

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_2}^{(2)}} \equiv \frac{\partial x_{i_4}^{(4)}}{\partial \alpha_{\mu_2}^{(2)}} = \sum_{i_3} \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial x_{i_3}^{(2)}} \frac{\partial f_{i_3}^{(3 \leftarrow 2)}}{\partial \alpha_{\mu_2}^{(2)}}$$

The summation is over  $i_3$  which goes from 1 to  $d_3$  (or if you prefer, 0 to  $d_3 - 1$ ).

The indices  $i_4$  and  $\mu_2$  in this expression are known as free indices, and the index  $i_3$ , which appears in exactly two places, is summed over its  $d_3$  possible values. From here on, whenever an index appears in two places, we will assume that it is summed over its possible values, and not bother with writing the summation symbol. Thus our result is simply

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_2}^{(2)}} = \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial x_{i_3}^{(3)}} \frac{\partial f_{i_3}^{(3 \leftarrow 2)}}{\partial \alpha_{\mu_2}^{(2)}}$$

Hopefully, you already see the pattern, but if not, write a few more out, and you will get:

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_1}^{(1)}} = \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial x_{i_3}^{(3)}} \frac{\partial f_{i_3}^{(3 \leftarrow 2)}}{\partial x_{i_2}^{(2)}} \frac{\partial f_{i_2}^{(2 \leftarrow 1)}}{\partial \alpha_{\mu_1}^{(1)}}$$

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_0}^{(0)}} = \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial x_{i_3}^{(3)}} \frac{\partial f_{i_3}^{(3 \leftarrow 2)}}{\partial x_{i_2}^{(2)}} \frac{\partial f_{i_2}^{(2 \leftarrow 1)}}{\partial x_{i_1}^{(1)}} \frac{\partial f_{i_1}^{(1 \leftarrow 0)}}{\partial \alpha_{\mu_0}^{(0)}}$$

Remember that there are 2 and 3 suppressed summations in these expressions, respectively.

Notice that for  $l = 3$  we had no  $x$  derivatives and one  $\alpha$  derivative. For  $l = 2$ , we had one  $x$  derivative and one  $\alpha$  derivative. For  $l = 1$ , we had two  $x$  derivatives and still one  $\alpha$  derivative. Finally for  $l = 0$ , we had three  $x$  derivatives and still one  $\alpha$  derivative.

The pattern is that you have  $L - l - 1$  suppressed summations, the same number of  $x$  derivatives, and finally one  $\alpha$  derivative.

## Tensors (Actually just Matrices)

Tensors can have 0, 1, 2, or more indices and the number of indices is called the rank. When they have 0 indices, we call them scalars. When they have 1 index, we call them vectors. When they have 2 indices, we call them matrices. When they have  $r$  indices with  $r > 2$ , we say tensor of rank  $r$ , but these are conventions. All these are tensors just with different ranks, and the ones we have immediate use for are matrices.

With matrices at our disposal, the expressions that are already getting messy in even the  $L = 4$  case are just matrix multiplications.

We will define

$$M_{i_l, i_{l-1}}^{(l \leftarrow l-1)} \equiv \frac{\partial f_{i_l}^{(l \leftarrow l-1)}}{\partial x_{i_{l-1}}^{(l-1)}}$$

This will take hide the plethora of partial derivatives with respect to the  $x$ 's. For example,

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_1}^{(1)}} = \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial x_{i_3}^{(3)}} \frac{\partial f_{i_3}^{(3 \leftarrow 2)}}{\partial x_{i_2}^{(2)}} \frac{\partial f_{i_2}^{(2 \leftarrow 1)}}{\partial \alpha_{\mu_1}^{(1)}}$$

becomes

$$\frac{\partial y_{i_4}}{\partial \alpha_{\mu_1}^{(1)}} = M_{i_4, i_3}^{(4 \leftarrow 3)} M_{i_3, i_2}^{(3 \leftarrow 2)} \frac{\partial f_{i_2}^{(2 \leftarrow 1)}}{\partial \alpha_{\mu_1}^{(1)}}$$

Let's keep going and hide the partial derivatives with respect to the  $\alpha$ 's too by defining

$$O_{i_4, \mu_1}^{(4 \leftarrow 1)} \equiv \frac{\partial f_{i_4}^{(4 \leftarrow 3)}}{\partial \alpha_{\mu_1}^{(1)}}$$

and

$$O_{i_2, \mu_1}^{(2 \leftarrow 1)} \equiv \frac{\partial f_{i_2}^{(2 \leftarrow 1)}}{\partial \alpha_{\mu_1}^{(1)}}$$

Putting those two definitions in, we have

$$O_{i_4, \mu_1}^{(4 \leftarrow 1)} = M_{i_4, i_3}^{(4 \leftarrow 3)} M_{i_3, i_2}^{(3 \leftarrow 2)} O_{i_2, \mu_1}^{(2 \leftarrow 1)}$$

Because **people that do matrix multiplications all the time** know that (1) each matrix carries two indices, (2) that the left-most index in an expression is a free index (in this case,  $i_4$ ), (3) that the right-most index in an expression is also free (in this case,  $\mu_1$ ), and (4) that all the in-between indices are repeated twice, are dummy indices (in this case,  $i_3$  and  $i_2$ ) and are summed over all their allowed values. So they **don't bother writing any indices!** With that understanding, we have simply

$$O^{(4 \leftarrow 1)} = M^{(4 \leftarrow 3)} M^{(3 \leftarrow 2)} O^{(2 \leftarrow 1)}$$

## Our Final Result

### The General Case

The partial derivatives we need to know when doing gradient descent are given by matrix multiplications. If there are  $L$  layers and we are looking for the derivative of one of the components of  $y$ , specifically,  $y_{i_L}$ , w.r.t. one of the components of  $\alpha^{(l)}$ , specifically,  $\alpha_{\mu_l}^{(l)}$ , then we need to compute the following matrix product

$$O^{(L \leftarrow l)} = M^{(L \leftarrow L-1)} M^{(L-1 \leftarrow L-2)} \dots M^{(l+2 \leftarrow l+1)} O^{(l+1 \leftarrow l)}$$

where

$$M_{i_l, i_{l-1}}^{(l \leftarrow l-1)} \equiv \frac{\partial f_{i_l}^{(l \leftarrow l-1)}}{\partial x_{i_{l-1}}^{(l-1)}}$$

$$O_{i_L, \mu_l}^{(L \leftarrow l)} \equiv \frac{\partial y_{i_L}}{\partial \alpha_{\mu_l}^{(l)}}$$

and

$$O_{i_{l+1}, \mu_l}^{(l+1 \leftarrow l)} \equiv \frac{\partial f_{i_{l+1}}^{(l+1 \leftarrow l)}}{\partial \alpha_{\mu_l}^{(l)}}$$

There are  $L - l - 1$  of the  $M$  matrices in this product. Having taken all the products, which including the right-most  $O$ , is a total of  $L - l$  products, we take the  $i_L, \mu_l$  entry.

This is insanely tidy for what could have turned into a much worse mess. The keys to the elegant and compact result were

(1) the rather special assumed functional form of our neural net,

$$y \equiv x^{(L)} = f^{(L \leftarrow L-1)}(f^{(L-1 \leftarrow L-2)}(\dots f^{(2 \leftarrow 1)}(f^{(1 \leftarrow 0)}(x^{(0)}; \alpha^{(0)}); \alpha^{(1)}) \dots; \alpha^{(L-2)}); \alpha^{(L-1)}),$$

(2) our hiding of decently complicated partial derivatives in matrices, and

(3) the conventions for matrix multiplication, especially summation of repeated indices.