Derivation of the BackPropagation Algorithm

Or Fishman

April 4, 2025

1 Introduction

Note: We'll be dealing with the case of a fully connected neural network (DNN), and use vector-matrix notation very often.

1.1 Notation

To establish clear communication, I'll mention the mathematical notation used.
$\vec{x}$ - The input vector.
ℓ - An arbitrary layer in the neural network.
$\mathcal{L}$ - The last layer of the neural network(output layer).
$\mathbf{W}_{L,L-1}$ - The weights of the neural network between the last two layers. b_L - The bias scalar value of the last layer.
$\vec{z}_2 = \mathbf{W}_{1,2} \cdot \vec{x} + b_1$ - a linear transformation
f_ℓ - The activation function of an arbitrary layer l.
$\vec{a}_\ell = f_\ell(\vec{z}_\ell)$ - activation $\vec{y}$ - The output of the DNN.
$\vec{t}$ - The expected value of the DNN.
E - Error loss of the neural network model.

2 Dive Into The Chain Rule

We'll start with finding the derivative to a composition of this very easy function:

$$g(x) = (5x + 2)^3$$

$$g'(x) = 3(5x + 2)^2 \cdot 5 = 15(5x + 2)^2$$

So far so good next, I will introduce Leibniz's derivative notation: $\frac{\mathrm{d}g}{\mathrm{d}x}$ which refers to how g changes in regards to x, or the corresponding partial derivative notation: $\frac{\partial g}{\partial x}$ when talking about the function that receives as input multiple variables.

$$\frac{\mathrm{d}g}{\mathrm{d}x} = 15(5x + 2)^2$$

Now, we can use this to find a general formula for the following derivative: $(f(g(x))'$ being the following:

$$(f(g(x)))' = f'(g(x)) \cdot g'(x)$$

As always what does this derivative express? how f changes with respect to x therefore, we can express it this way:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g} \cdot \frac{\mathrm{d}g}{\mathrm{d}x}$$

Example:
Let $f(g) = g^3, \quad g(x) = 5x + 2$ what is $\frac{\mathrm{d}f}{\mathrm{d}x}$?

$$\frac{\mathrm{d}f}{\mathrm{d}g} = 3g^2$$

$$\frac{\mathrm{d}g}{\mathrm{d}x} = 5$$

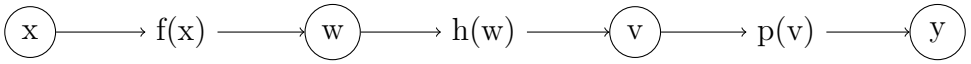
$$\frac{\mathrm{d}f}{\mathrm{d}x} = 15g^2 = 15(5x + 2)^2$$

which is what we got earlier.

This general formula has a special name called the Chain Rule which we'll use.

3 Computational Graphs with the Chain Rule

Given this artibrary computational graph that corresponds to applying some functions, how do we express $\frac{\mathrm{d}y}{\mathrm{d}x}$?



Well, we can use the chain rule!

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}w}{\mathrm{d}x} \cdot \frac{\mathrm{d}v}{\mathrm{d}w} \cdot \frac{\mathrm{d}y}{\mathrm{d}v}$$

Notice that here we do not mention the functions themselves at all, but their values, that is merely for clarify but we could actually remove w, and v, and just stick with the function compositions.

4 Derivation

Let start with the general equations associated with forward propagation:

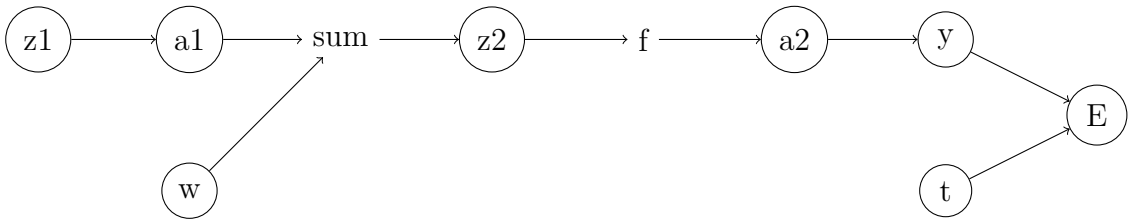
$$\vec{z}_\ell = \mathbf{W}_{\ell,\ell-1} \cdot \vec{a}_{\ell-1} + b_\ell \tag{1}$$

$$\vec{a}_\ell = f_\ell(\vec{z}_\ell) \tag{2}$$

$$E_{MSE} = \frac{1}{2}(\vec{t} - \vec{y})^2 \tag{3}$$

$$\frac{\partial E}{\partial a} = (\vec{t} - \vec{y}) \cdot f'(\vec{z}_L) \tag{4}$$

Now, we'll display the computational graph:



$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial W}$$

However, often times, we subsitue part of it with delta for more clarity:

$$\vec{\delta}_\ell = \frac{\partial E}{\partial a} \cdot \frac{\partial a}{\partial z}$$

$$\frac{\mathrm{d}E}{\mathrm{d}W} = \vec{\delta}_\ell \cdot \frac{\mathrm{d}z}{\mathrm{d}w}$$

such that

$$\vec{\delta} =$$

Is The formula using the chain rule, next we'll evaluate these derivatives:
*remember $\vec{y} = \vec{a_L}$

$$\frac{\mathrm{d}E}{\mathrm{d}a} = (\vec{t} - \vec{y})$$

For Sigmoid:

$$\frac{\mathrm{d}a}{\mathrm{d}z} = \sigma' = \sigma(1 - \sigma)$$

$$\frac{\mathrm{d}z}{\mathrm{d}w} = \vec{a}_{\ell-1}$$

$$\vec{\delta}_\ell = (\vec{t} - \vec{y}) \cdot \sigma'$$

$$\vec{\delta}_\ell = \mathbf{W}_{\ell+1,\ell} \cdot \vec{\delta}_{\ell+1} \cdot \sigma'$$

4.1 Putting everything together

$$\frac{\mathrm{d}E}{\mathrm{d}W} = \vec{\delta}_\ell \cdot \vec{a}_{\ell-1} = (\vec{t} - \vec{y}) \cdot \sigma' \cdot \vec{a}_{\ell-1}$$

Therefore,

$$\Delta \mathbf{W}_{\ell,\ell-1} = -\alpha((\vec{t} - \vec{y}) \cdot \sigma' \cdot \vec{a}_{\ell-1})$$

5 Auto Differentiation

6 Introduction

Although we have fully derived the backpropagation algorithm, it poses some problems. To begin with, when calculating it, we must be able to calculate the derivative of the activation function with respect to the input of the neuron. Another problem is that we would need to manually adjust our derivation for almost every change we make to the network's architecture, from new activation functions, to possibly new layers such as a convolution layer, pooling layer, etc.
This is clearly impractical, and not how it works in real world deep learning libraries such as PyTorch, or Tensorflow as well as Keras. In these frameworks, there are modules for the concept of automatically differentiating any functions specified with code.

6.1 Dual Numbers

One of the methods for doing so involves the use of specially crafted type of numbers, called dual numbers.

A Dual number is a number similar to a complex number, that is composed of a real part, and an infinitesimal part that will be our derivative.

$$d = a + b\epsilon$$

such that the ϵ part has these two properties:

$$\epsilon^2 = 0 \tag{5}$$

$$\epsilon \neq 0; \tag{6}$$

From here on, we can further derive some other useful properties such as:

$$(a + b\epsilon) + (c + d\epsilon) = (a + c) + (b + d)\epsilon \tag{7}$$

$$(a + b\epsilon) \cdot (c + d\epsilon) = ac + (ad + bc)\epsilon \tag{8}$$

$$\tag{9}$$

These numbers very special as they allow to calculate the derivative term, for two dual numbers:

$$d_1 = a_1 + b_1\epsilon \tag{10}$$
$$d_2 = a_2 + b_2\epsilon \tag{11}$$

$$\frac{d}{dt}(d_1 + d_2) = b_1 + b_2 \tag{12}$$
$$\frac{d}{dt}d_1 \cdot d_2 = b_2a_1 + b_1a_2 \tag{13}$$
$$\frac{d}{dt}(d_1/d_2) = \frac{b_1a_2\epsilon - b_2a_1\epsilon}{a_2^2} \tag{14}$$

6.1.1 How Dual Numbers can be used

Let f be a function defined like this:

$$f(x) = 2x^2 + 3x + 1$$

$$f(x+\epsilon) = 2(x+\epsilon)^2 + 3(x+\epsilon) + x + 1$$
$$f(x+\epsilon) = 2x^2 + 4x\epsilon + 3x + 3\epsilon + 1$$

$$f(x+\epsilon) = (2x^2 + 3x + 1) + (4x + 3)\epsilon$$

Therefore, for x=4

$$f(4+\epsilon) = 45 + 19\epsilon$$

such that

$$f'(4) = 19$$