Neural Networks and Backpropagation Lecture

Grégoire Payen de La Garanderie

February 16, 2020

1 Introduction

This lecture aims to introduce the mathematical concepts at the core of the backpropagation mechanism that is commonly used to optimise neural networks.

Let us consider a neural network whose last layer is a loss function. The output loss can be expressed as a function of the neural network input $\boldsymbol{x}$ and weights $\boldsymbol{W}$ as follows:

$$\begin{aligned} l\colon \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R} \\ \boldsymbol{x}, \boldsymbol{W} &\mapsto l(\boldsymbol{x}, \boldsymbol{W}) \end{aligned}$$

Our goal is to minimise $l(x, W)$ with respect to (*w.r.t.*) W. In mathematical terms, this is written as:

$$\underset{\boldsymbol{W}}{\text{minimise}} \quad l(\boldsymbol{x}, \boldsymbol{W})$$

Compared to most minimisation problems which features many constraints, this problem looks deceptively simple however the difficulty of this optimisation resides is the non-linearity of the function l and the large number of dimensions of $\boldsymbol{W} \in \mathbb{R}^m$ (*i.e.* m is a large number, typically a several millions).

The simplest optimisation strategy for a neural network is called Gradient Descent and is defined as the following update rule:

$$\boldsymbol{W}_{t+1} \leftarrow \boldsymbol{W}_t - lr \cdot \frac{\partial l}{\partial \boldsymbol{W}} l(\boldsymbol{x}, \boldsymbol{W}) \tag{1}$$

where $\frac{\partial l}{\partial \boldsymbol{W}}$ in the derivative of l *w.r.t.* $\boldsymbol{W}$ (also called gradient of l *w.r.t.* $\boldsymbol{W}$) and lr is a learning rate which control the rate of convergence of the algorithm. Figure 1 shows an example of gradient descent on a one-dimensional cost function.

The key question that we will address in this lecture is how to compute the derivative $\frac{\partial l}{\partial \boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{W})$. It is this computation which is commonly called backpropagation in the deep learning jargon. However this is not a new technique, symbolic automatic differentiation (AD) has been studied since the 1960s, long before they were applied to neural networks in the 1980s. The backpropagation

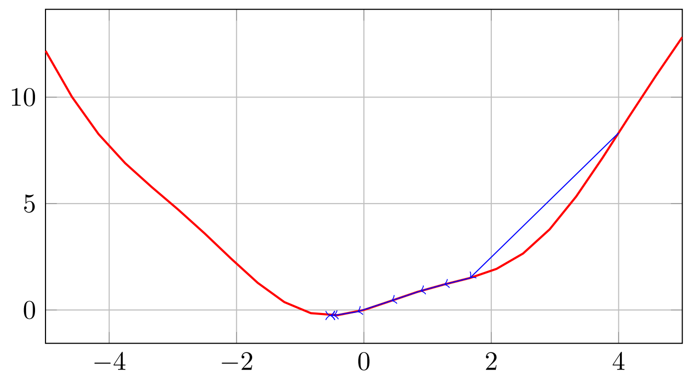


Figure 1: Gradient descent on a one-dimensional function. Red: the cost function to minimise. Blue: Successive gradient updates with $lr = 0.5$.

algorithm is a special case of AD that is particularly well suited to large modern deep neural network.

In the next section, we define function derivatives, both as the refresher of their definition and to define the notation that we use throughout the lecture.

2 Function Derivative

This section starts by introducing the definition of derivative for a simple real-valued function then extend this definition to vector-valued function that are typically used in neural networks. We only consider vector-valued function without loss of generality. Indeed the principles described here for vectors can be trivially generalised to tensor functions because any shape tensor can be flatten into a rank-1 tensor (*i.e.* a vector). We also presents the derivative of commonly used functions in deep learning.

Definition 2.1. Derivative

Let f be a real-valued function:

$$\begin{aligned} f\colon \mathbb{R} &\to \mathbb{R} \\ x &\mapsto f(x) \end{aligned}$$

The derivative of f called f' is a function is defined as:

$$\forall x \in \mathbb{R} \qquad f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x) = \lim_{\delta x \to 0} \frac{f(x+\delta x) - f(x)}{\delta x} \tag{2}$$

Note that this is a very common but slightly confusing and abusive notation. It is important not to forget that the x in $\mathrm{d}x$ or δx is not the same as the x in $f(x)$. An unambiguous equivalent notation would be:

$$\forall a \in \mathbb{R} \qquad f'(a) = \frac{\mathrm{d}f}{\mathrm{d}x}(a) = \lim_{\delta x \to 0} \frac{f(a+\delta x) - f(a)}{\delta x} \tag{3}$$

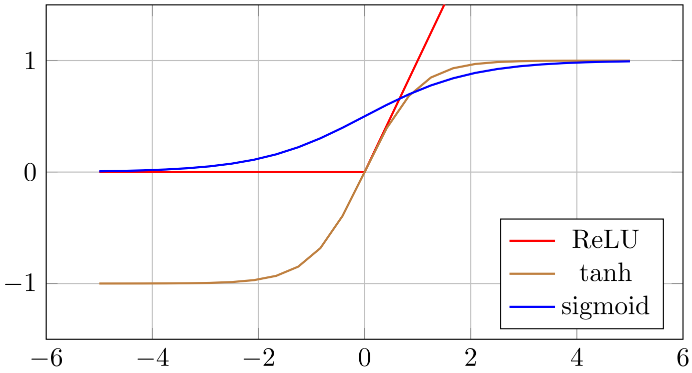


Figure 2: Common activation functions

Example 2.1. Useful derivatives

Here are some useful derivatives for common activation functions shown in Figure 2:

1. $\text{ReLU}'(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$
2. $\tanh'(x) = 1 - \tanh^2(x)$
3. $\text{sigmoid}'(x) = \text{sigmoid}(x) \cdot (1 - \text{sigmoid}(x))$

At first, we extend this definition of a function derivative to functions of two variables in the definition below to introduce the concept of partial derivative. Subsequently we generalise to function of n variables that are typically used in a neural network.

Definition 2.2. Partial derivative (real-valued function of two real-valued variables)

Let f be a function of two real-valued variables:

$$\begin{aligned} f \colon \mathbb{R}^2 &\to \mathbb{R} \\ x, y &\mapsto f(x, y) \end{aligned}$$

The function f admits two partial derivatives called $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ called respectively "partial derivative of f with respect to x" and "partial derivative of f with respect to y" defined as:

$$\forall (x, y) \in \mathbb{R}^2 \qquad \frac{\partial f}{\partial x}(x, y) = \lim_{\delta x \to 0} \frac{f(x + \delta x, y) - f(x, y)}{\delta x} \tag{4a}$$

$$\forall (x, y) \in \mathbb{R}^2 \qquad \frac{\partial f}{\partial y}(x, y) = \lim_{\delta y \to 0} \frac{f(x, y + \delta y) - f(x, y)}{\delta y} \tag{4b}$$

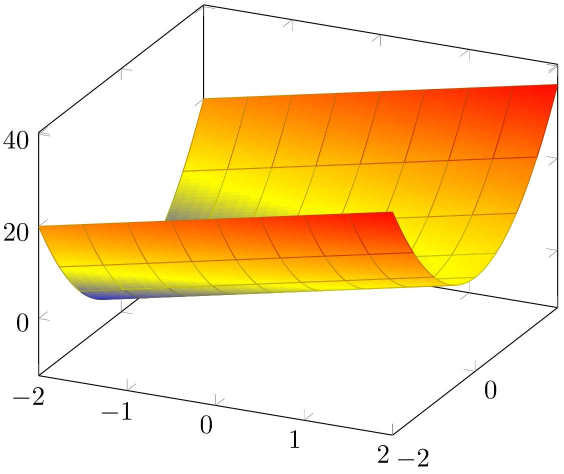


Figure 3: The function $f(x, y) = 4x + 7y^2$

Again, the notation can be confusing and an unambiguous definition would be:

$$\forall (a,b) \in \mathbb{R}^2 \qquad \frac{\partial f}{\partial x}(a,b) = \lim_{\delta x \to 0} \frac{f(a + \delta x, b) - f(a,b)}{\delta x} \tag{5a}$$

$$\forall (a,b) \in \mathbb{R}^2 \qquad \frac{\partial f}{\partial y}(a,b) = \lim_{\delta y \to 0} \frac{f(a, b + \delta y) - f(a,b)}{\delta y} \tag{5b}$$

Example 2.2. Let f be $f(x, y) = 4x + 7y^2$ (as shown in Figure 3). Then, we can prove that:

$$\forall (x,y) \in \mathbb{R}^2 \qquad \frac{\partial f}{\partial x}(x,y) = 4 \quad \text{and} \quad \frac{\partial f}{\partial y}(x,y) = 14y \tag{6}$$

In this example, note that the derivative of f *w.r.t.* x is independent from the value of y and vice-versa. This situation is frequently observed in neural network layers featuring a bias term as we will see later.

We can evaluate the partial derivatives at a specific point:

$$\frac{\partial f}{\partial x}(2,3) = 4 \quad \text{and} \quad \frac{\partial f}{\partial y}(2,3) = 14 \cdot 3 = 42 \tag{7}$$

Example 2.3. Let f be $f(x, y) = x \cdot y$ (as shown in Figure 4. Then, we can prove that:

$$\forall (x,y) \in \mathbb{R}^2 \qquad \frac{\partial f}{\partial x}(x,y) = y \quad \text{and} \quad \frac{\partial f}{\partial y}(x,y) = x \tag{8}$$

This example is an illustration of a simple bilinear function (that is a function that is both linear *w.r.t.* to x and linear *w.r.t.* to y). Many functions used in

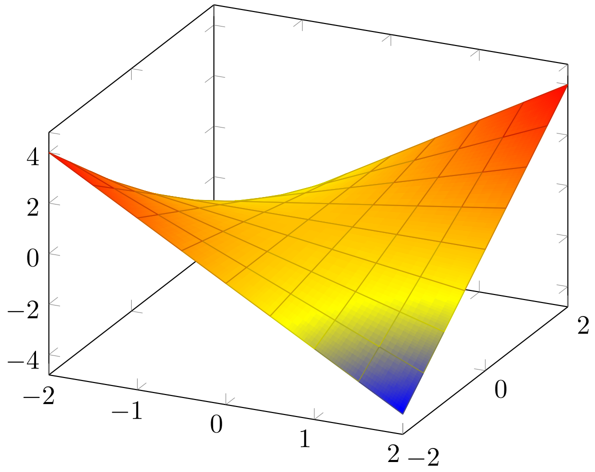


Figure 4: The function $f(x, y) = x \cdot y$

deep learning (*e.g.* fully-connected layers, convolutional layers) are actually multilinear map *w.r.t.* their input, weights and bias terms.

We can evaluate the partial derivatives at a specific point:

$$\frac{\partial f}{\partial x}(2,3) = 3 \quad \text{and} \quad \frac{\partial f}{\partial y}(2,3) = 2 \tag{9}$$

Definition 2.3. Partial derivative (real-valued function of n real-valued variables)

Let f be a function of n real-valued variables:

$$\begin{aligned} f\colon \mathbb{R}^n &\to \mathbb{R} \\ x_1, \ldots, x_n &\mapsto f(x_1, \ldots, x_n) \end{aligned}$$

The function f admit n partial derivatives called $\frac{\partial f}{\partial x_i}$ for $i \in [1, n]$:

$$\begin{aligned} \forall (x_1, \ldots, x_n) \in \mathbb{R}^n, i \in [1, n] \qquad & \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) = \\ \lim_{\delta x_i \to 0} & \frac{f(x_1, \ldots, x_{i-1}, x_i + \delta x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n)}{\delta x_i} \end{aligned} \tag{10}$$

Apart from the verbose notation, this is a straightforward generalisation from a function of 2 variables to a function of n variables.

Example 2.4. Recalling the previous example $f(x, y) = x \cdot y$. We can instead compute a more general form $f(x_1, \ldots, x_n) = \prod_i x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n$. Then:

$$\forall (x_1, \dots, x_n) \in \mathbb{R}^n, i \in [1, n] \qquad \frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = \\ x_1 \times \dots \times x_{i-1} \times x_{i+1} \times \dots \times x_n = \prod_{j \neq i} x_j \tag{11}$$

We can evaluate the partial derivatives at a specific point ($n = 3$):

$$\frac{\partial f}{\partial x_1}(3,7,4) = 7 \cdot 4 \quad \text{and} \quad \frac{\partial f}{\partial x_2}(3,7,4) = 3 \cdot 4 \quad \text{and} \quad \frac{\partial f}{\partial x_2}(3,7,4) = 3 \cdot 7 \tag{12}$$

So far, the previous definitions have considered functions of multiple real-valued input variables however they only had a single real-valued output. We can now have a look at functions of multiple inputs and multiple outputs.

Definition 2.4. Jacobian Matrix
Let f be a function defined as:

$$\begin{aligned} f \colon \mathbb{R}^n & \to \mathbb{R}^m \\ x_1, \dots, x_n & \mapsto (y_1, \dots, y_m) = f(x_1, \dots, x_n) \end{aligned}$$

This function takes multiple values $x_1, \dots, x_n$ as an input and returns multiple values $y_1, \dots, y_n$. For instance, you can think of the multiple x_is as the pixel values of an image and the y_is as output classification probabilities or an intermediate layer feature map.

To simplify the notation, we introduce f_i as the i-th component of the output of f, *i.e.* $y_i = f_i(x_1, \dots, x_n)$. There are m such functions, one for each of our m outputs. Each of the component function f_i admits n partial derivatives called $\frac{\partial f_i}{\partial x_j}$. $\frac{\partial f_i}{\partial x_j}$ can be called the partial derivative of the i-th output of f with respect to the j-th input of f. They are $n \times m$ such partial derivatives for all combinations of inputs and outputs. We can pack all those derivative in a matrix often referred as the Jacobian of f defined as:

$$J_f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f_i}{\partial x_j}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{13}$$

The representation of the derivative of f as a matrix is convenient because it behaves in the same way as the derivative of a real-valued function as shown below.

This function f of multiple inputs and outputs can be represented using a vectorial notation:

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m \qquad \boldsymbol{y} = f(\boldsymbol{x}) \tag{14}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors.

In the next two examples, we show examples of functions commonly used in neural networks and their derivative.

Example 2.5. Linear Map

A linear map is the mathematical formulation of a fully-connected linear layer in a neural network (where $\boldsymbol{x}$ is the layer input, $\boldsymbol{W}$ the weight matrix and $\boldsymbol{b}$ the bias term). Let f be a linear map defined as:

$$\begin{aligned} f\colon \mathbb{R}^n &\to \mathbb{R}^m \\ \boldsymbol{x} &\mapsto \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \end{aligned}$$

where $\boldsymbol{W}$ is a $m \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^m$.

The derivative of f (*i.e.* the Jacobian matrix) is:

$$\forall \boldsymbol{x} \in \mathbb{R}^n \qquad \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{W} \tag{15}$$

That is, the Jacobian of the function f is the matrix $\boldsymbol{W}$. Note that for a linear map, the derivative is constant and does not depend on the value of $\boldsymbol{x}$. This is obviously not always true for all functions.

This example encompasses not only fully-connected layers but convolutional layers as well because convolutions are special form of linear map. The matrix $\boldsymbol{W}$ of a convolution is a Toeplitz matrix defined by the convolution kernel.

Example 2.6. Softmax

The Softmax function is often used as the last layer of a classification neural network. It normalises an un-normalised vector into a probability distribution (ensuring that the sum of all probabilities is 1). The softmax function is defined as:

$$\begin{aligned} s\colon \mathbb{R}^n &\to \mathbb{R}^n \\ \boldsymbol{x} &\mapsto \boldsymbol{y} = \left(\frac{e^{x_i}}{\sum_j e^{x_j}} \right)_{i \in [1,n]} \end{aligned}$$

We can show that its derivative is:

$$\forall i, j, \boldsymbol{x} \qquad \frac{\partial s_i}{\partial \boldsymbol{x}_j}(\boldsymbol{x}) = s_i(\boldsymbol{x}) \cdot (\delta_{ij} - s_j(\boldsymbol{x})) \tag{16}$$

The Jacobian of the softmax function can be defined as:

$$\frac{\mathrm{d}s}{\mathrm{d}\boldsymbol{x}}(\boldsymbol{x}) = \begin{bmatrix} s_1(\boldsymbol{x})(1 - s_1(\boldsymbol{x})) & -s_1(\boldsymbol{x})s_2(\boldsymbol{x}) & \dots & -s_1(\boldsymbol{x})s_n(\boldsymbol{x}) \\ -s_2(\boldsymbol{x})s_1(\boldsymbol{x})) & s_2(\boldsymbol{x})(1 - s_2(\boldsymbol{x})) & & -s_2(\boldsymbol{x})s_n(\boldsymbol{x}) \\ \vdots & & \ddots & \vdots \\ -s_n(\boldsymbol{x})s_1(\boldsymbol{x}) & -s_n(\boldsymbol{x})s_2(\boldsymbol{x}) & \dots & s_n(\boldsymbol{x})(1 - s_n(\boldsymbol{x})) \end{bmatrix} \tag{17}$$

This Jacobian has a common term $-s_i(\boldsymbol{x})s_j(\boldsymbol{x})$ corresponding to the denominator in the softmax function as well as an extra term along the diagonal $s_i(\boldsymbol{x})$ corresponding to the numerator. The softmax function is an example of a non-linear function. As such, the value of the derivative is non-constant and depends on the value of $\boldsymbol{x}$.

3 Function Composition and Derivative

Theorem 3.1. Chain Rule

Let f and g be two functions, either real-valued or vector-valued and let h be the composition of the two: $h = f \circ g$. That is:

$$\forall x \qquad h(x) = f(g(x)) \tag{18}$$

For example, suppose we have a function that normalises the intensity of an image $f(x)$, but before that we Gaussian blur the image $g(x)$, we have the composite $f(g(x))$.

Then the derivative of h is such that:

$$\forall x \qquad h'(x) = f'(g(x)) \times g'(x) \tag{19}$$

This is known as the chain rule. It works with both real-valued and vector-valude functions. However unlike for real-valued functions, the chain-rule is non-commutative for vector-valued function. That is:

$$\begin{array}{lll} \forall \boldsymbol{x} & h'(\boldsymbol{x}) = f'(g(\boldsymbol{x})) \cdot g'(\boldsymbol{x}) & \textbf{Correct} \\ \forall \boldsymbol{x} & h'(\boldsymbol{x}) = g'(\boldsymbol{x}) \cdot f'(g(\boldsymbol{x})) & \textbf{Very wrong} \end{array}$$

If the dimensions of the input and output vectors of g are mismatching, the Jacobian of g will be a rectangular matrix and $g'(\boldsymbol{x}) \cdot f'(g(\boldsymbol{x}))$ will not even be a valid expression. Therefore be extremely careful that you are applying the chain rule in the right order.

The chain rule is often written as:

$$\frac{\mathrm{d}h}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g} \times \frac{\mathrm{d}g}{\mathrm{d}x} \tag{21}$$

This is a convenient notation but slightly misleading notation because it hides that the derivatives of f and g are not applied to the same input. What is really implied by this equation is:

$$\frac{\mathrm{d}h}{\mathrm{d}x}(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(g(x)) \times \frac{\mathrm{d}g}{\mathrm{d}x}(x) \tag{22}$$

Consider the example of a function f applied on itself, a common occurence in machine learning:

$$h(x) = f(f(x)) \tag{23}$$

would lead to this non-sensical notation:

$$\frac{\mathrm{d}h}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}f} \times \frac{\mathrm{d}f}{\mathrm{d}f} \tag{24}$$

This formulation is especially useful is physics where each symbol represents unique physical variables of a system however in machine learning where symbols represents functions, the notation can be quite misleading.

Example 3.1. A simple neural network

Let l be a neural network constituted of three layers f, g, h:

$$l(\boldsymbol{x}) = (f \circ g \circ h)(\boldsymbol{x}) = f(g(h(\boldsymbol{x}))) \tag{25}$$

This network can be represented visually by:

$$\begin{tikzcd}
\boldsymbol{x} \arrow[d] \\
h(\boldsymbol{x}) \arrow[d] \\
(g \circ h)(\boldsymbol{x}) \arrow[d] \\
(f \circ g \circ h)(\boldsymbol{x})
\end{tikzcd}$$

The gradient of l with respect to the input $\boldsymbol{x}$ is:

$$l'(\boldsymbol{x}) = f'((g \circ h)(\boldsymbol{x})) \cdot g'(h(x)) \cdot h'(x) \tag{26}$$

We can see in the above equation that the derivative of each layer depends on the input of its layer. For instance, the derivative of the layer f depends on its input $(g \circ h)(\boldsymbol{x})$.

For the sake of simplicity, we will omit the input of each layer in the notation of the gradients and write:

$$l' = f' \cdot g' \cdot h' \tag{27}$$

We can see that the gradient of this sequential neural network in the product of the gradient of each individual layers (according to the chain rule).

This neural network is not particularly realistic because it does not have any weights. Recall the minimisation goal defined in Section 1, we are interested in minimising the loss *w.r.t.* to the weights, not the input itself. In practice, each layer takes an additional weight parameter. To take this into account, we need one more definition.

Definition 3.1. Function of multiple vectors

Let f be a function of multiple vectors defined as:

$$\begin{aligned} f\colon \mathbb{R}^s \times \mathbb{R}^t &\to \mathbb{R}^u \\ x, y &\mapsto f(x, y) \end{aligned}$$

Analogously to the definition of a partial derivative of a function of multiple real values, this function can be differentiated with respect to each of its input. We will call f'_i the derivative with respect to the i-th input.

Example 3.2. Multilinear Map

In the Example 2.5, we have looked at the derivative of a linear map with fixed weights and bias. However this is not the case of a neural network where the weight matrix and bias vector are themselves inputs of the map that can be adjusted by the optimiser.

Let f be a linear map defined as:

$$\begin{aligned} f\colon \mathbb{R}^n \times \mathbb{R}^{m\times n} \times \mathbb{R}^m &\to \mathbb{R}^m \\ \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b} \qquad\qquad &\mapsto \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \end{aligned}$$

where $\boldsymbol{W}$ is a $m \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^m$.

For a given $\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}$, the derivatives of f are:

$$\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = \boldsymbol{W} \tag{28a}$$

$$\frac{\partial f}{\partial \boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = \boldsymbol{x} \tag{28b}$$

$$\frac{\partial f}{\partial \boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = (1, \ldots, 1) \tag{28c}$$

Example 3.3. A weighted neural network

Let l be a neural network constituted of three layers $f(\boldsymbol{a}, \boldsymbol{b})$, $g(\boldsymbol{a}, \boldsymbol{b})$, $h(\boldsymbol{a}, \boldsymbol{b})$ and weights $\boldsymbol{w_f}$, $\boldsymbol{w_g}$, $\boldsymbol{w_h}$:

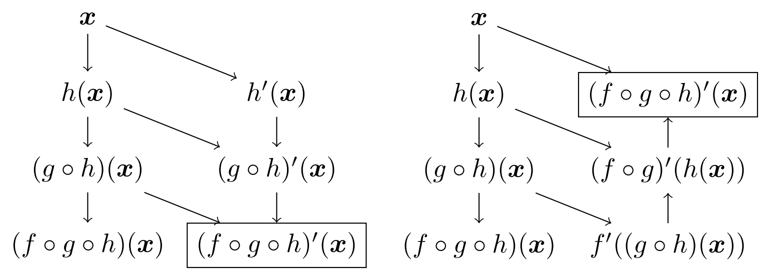
$$l(\boldsymbol{x}, \boldsymbol{w_f}, \boldsymbol{w_g}, \boldsymbol{w_h}) = f(g(h(\boldsymbol{x}, \boldsymbol{w_h}), \boldsymbol{w_g}), \boldsymbol{w_f}) \tag{29}$$

$$\begin{tikzcd}
\boldsymbol{x} \arrow[dd] & \\
 & \boldsymbol{w_h} \arrow[dl] \\
h(\boldsymbol{x}, \boldsymbol{w_h}) \arrow[dd] & \\
 & \boldsymbol{w_g} \arrow[dl] \\
g(h(\boldsymbol{x}, \boldsymbol{w_f}), \boldsymbol{w_g}) \arrow[dd] & \\
 & \boldsymbol{w_f} \arrow[dl] \\
f(g(h(\boldsymbol{x}, \boldsymbol{w_h}), \boldsymbol{w_g}), \boldsymbol{w_f}) &
\end{tikzcd}$$

The gradient with respect to the input $\boldsymbol{x}$ is:

$$\frac{\partial l}{\partial \boldsymbol{x}} = \frac{\partial f}{\partial \boldsymbol{a}} \cdot \frac{\partial g}{\partial \boldsymbol{a}} \cdot \frac{\partial h}{\partial \boldsymbol{a}} \tag{30}$$

However the gradient w.r.t to the input is rarely used in machine learning. Instead the optimiser relies on the gradient of the weights to update each weight at every iteration. Those gradients can also be computed using the chain rule:


$\boldsymbol{x}$
$h(\boldsymbol{x})$
$h'(\boldsymbol{x})$
$(g \circ h)(\boldsymbol{x})$
$(g \circ h)'(\boldsymbol{x})$
$(f \circ g \circ h)(\boldsymbol{x})$
$(f \circ g \circ h)'(\boldsymbol{x})$
$\boldsymbol{x}$
$h(\boldsymbol{x})$
$(f \circ g \circ h)'(\boldsymbol{x})$
$(g \circ h)(\boldsymbol{x})$
$(f \circ g)'(h(\boldsymbol{x}))$
$(f \circ g \circ h)(\boldsymbol{x})$
$f'((g \circ h)(\boldsymbol{x}))$

(a) Forward propagation using Equation 33a

(b) Backpropagation using Equation 33b

Figure 5: Computational graphs for forward and backward propagation. The final result in each graph has been framed.

1. w.r.t to $\boldsymbol{w_f}$: $\frac{\partial l}{\partial \boldsymbol{w_f}} = \frac{\partial f}{\partial \boldsymbol{b}}$
2. w.r.t to $\boldsymbol{w_g}$: $\frac{\partial l}{\partial \boldsymbol{w_g}} = \frac{\partial f}{\partial \boldsymbol{a}} \cdot \frac{\partial g}{\partial \boldsymbol{b}}$
3. w.r.t to $\boldsymbol{w_h}$: $\frac{\partial l}{\partial \boldsymbol{w_h}} = \frac{\partial f}{\partial \boldsymbol{a}} \cdot \frac{\partial g}{\partial \boldsymbol{a}} \cdot \frac{\partial h}{\partial \boldsymbol{b}}$

4 Backpropagation

4.1 Differentiation and Computational Graph

Example 4.1. A weighted neural network

Recall the previous example of a three layers neural network. Let l be a neural network constituted of three layers f, g, h:

$$l(\boldsymbol{x}) = (f \circ g \circ h)(\boldsymbol{x}) = f(g(h(\boldsymbol{x}))) \tag{31}$$

We have shown that the gradient can be computed as:

$$l' = f' \cdot g' \cdot h' \tag{32}$$

Computationally, this gradient can be computed in two different ways:

$$l' = (f' \cdot g') \cdot h' \tag{33a}$$

$$l' = f' \cdot (g' \cdot h') \tag{33b}$$

The first one performs the multiplication between f' and g' first (as shown in Figure 5(a)) while the second performs the multiplication between g' and h' first (as shown in Figure 5(b)). Both of those ways are mathematically equivalent and correct however one of them is not practical from a computational point of view with neural network.

To understand the problem, we need to look at the dimension of each of the layers. During training, the final layer of a neural network (f here) is the loss function which output a single real-value therefore its output has a dimension of 1. In contrast the input and the intermediate layers are much larger. For instance, let us assume that this network is processing an image of size 1000×1000, this image has $1,0000,000$ pixels therefore the input layer has a dimension of $1,000,000$. Let us assume that the intermediate layers have the same size.

This means that we have:

1. $f \colon \mathbb{R}^{1,000,000} \to \mathbb{R}$
2. $g \colon \mathbb{R}^{1,000,000} \to \mathbb{R}^{1,000,000}$
3. $h \colon \mathbb{R}^{1,000,000} \to \mathbb{R}^{1,000,000}$

On one hand, the Jacobian of g is a $1,000,000 \times 1,000,000$ matrix. This matrix would require 1TB of memory. Therefore it is not actually possible to compute and store the Jacobian of g. The Jacobian of h and $g \circ h$ have exactly the same size and are too large to be stored. This means that it is unrealistic to compute $g' \cdot h'$ first with modern technologies.

On the other hand, the Jacobian of f is a $1 \times 1,000,000$ matrix. To store this matrix, we would need about 1Mb of memory. The Jacobian of $f \circ g$ is also a $1 \times 1,000,000$ matrix. This is the intuition between the backpropagation algorithm. It is unrealistic to compute the gradients starting from the input because the intermediate Jacobian matrices would be too large however if we start from the loss value, all the intermediate matrices are kept small.

4.2 Implementation of the Chain Rule

In the previous example, we have shown that it is better to work our way backward, computing f' then $f' \cdot g'$ and finally $f' \cdot g' \cdot h'$. If the intermediate layers are of size $1,000,0000$, each of those derivative is a matrix of size $1 \times 1,000,000$. The simplest way to compute $f' \cdot g'$ is to compute f' and g' and multiply them together however this does not work because the Jacobian g' is also a $1,000,000 \times 1,000,000$ matrix. In practice, the Jacobian of most operations (*e.g.* a convolution) is very sparse and a lot of redundant operations and memory can be saved by computing the derivative g' and the multiplication together.

This leads us to the usual way of defining forward and backward functions in most deep learning frameworks.

Algorithm 1 Sample forward function

```
function Forward(input)
    return $f$(input)
end function
```

Algorithm 2 Sample backward function

```
function Backward(input,grad_output)
    return grad_output · $f'$(input)
end function
```

Example 4.2 (Convolution implementation)**.** Let us define a convolution between a 1D input $\boldsymbol{x} \in \mathbb{R}^n$ and a small kernel $\boldsymbol{h} \in \mathbb{R}^3$:

$$\begin{aligned} f \colon \mathbb{R}^n &\to \mathbb{R}^n \\ \boldsymbol{x} &\mapsto \boldsymbol{y} = \left(\sum_{k \in [-1,1]} x_{i+k} h_{k+2} \right)_{i \in [1,\dots,n]} \end{aligned}$$

In the sum, for the sake of simplicity, we assume that $x_{i+j} = 0$ if $i + j$ is out of range (*i.e.* $i + j \notin [1, n]$).

The partial derivatives of f are:

$$\forall i, j \in [1, n] \qquad \frac{\partial f_i}{\partial x_j} = \begin{cases} h_{j-i+2}, & \text{if } j - i \in [-1, 1] \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

The Jacobian matrix is a band matrix:

$$\forall \boldsymbol{x} \in \mathbb{R}^n \qquad f'(\boldsymbol{x}) = \begin{bmatrix} h_2 & h_3 & 0 & & \dots & & 0 \\ h_1 & h_2 & h_3 & & & & \\ 0 & h_1 & & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & & h_3 & 0 \\ & & & & h_1 & h_2 & h_3 \\ 0 & & \dots & & 0 & h_1 & h_2 \end{bmatrix} \tag{35}$$

This Jacobian is a sparse matrix called a Toeplitz matrix which has mostly zeros everywhere apart from a narrow diagonal band of width 3.

Here is a naive implementation of the backward function that computes the Jacobian first then multiply it with the output gradient:

Algorithm 3 Naive convolution backward function

```
function ConvBackward(input,grad_output)
    #Compute the Jacobian matrix
    $J \leftarrow$ zero matrix of size $n \times n$
    for $i \in [1, n]$ do
        for $j \in [1, n]$ do
            if $j - i \in [-1, 1]$ then
                $J_{i,j} \leftarrow h_{j-i+2}$
            end if
        end for
    end for

    #Multiply the Jacobian with grad_output
    $v \leftarrow$ zero vector of size $n$
    for $i \in [1, n]$ do
        for $j \in [1, n]$ do
            $v_j \leftarrow v_j + \text{grad\_output}_i \cdot J_{i,j}$
        end for
    end for
    return $v$
end function
```

The algorithm has two outer loops with a computational complexity of $\mathcal{O}(n^2)$ and the size of the intermediate Jacobian matrix J is n^2. As we have seen, this is not practical for large arrays and images. In contrast, we can exploit the sparsity of the Jacobian to reduce the computational and spatial complexity to $\mathcal{O}(n)$ in the following algorithm.

Algorithm 4 Optimised convolution backward function

```
function ConvBackward(input,grad_output)
    $v \leftarrow$ zero vector of size $n$
    for $k \in [-1, 1]$ do
        for $j \in [1, n]$ do
            if $0 < k - i \leq n$ then
                $v_j \leftarrow v_j + \text{grad\_output}_{k-i} \cdot h_{k+2}$
            end if
        end for
    end for
    return $v$
end function
```

5 Conclusion

We have seen that there are multiple possible algorithms to compute the gradient of a neural network however a naive approach would have terrible computational and spatial complexity. The backpropagation algorithm provides an efficient way of computing the gradients. Besides for most operations, it is a bad idea to directly compute the Jacobian matrix because it can be very large. It is often much more effective to simplify and compute the differentiation and multiplication together.

While it is often unnecessary to understand how backpropagation works to understand the overall architecture of a neural network; it is a requirement in order to implement new layers, choose adequate architecture to keep the memory requirements down as well as avoid numerous pitfalls related to numerical accuracy and convergence.