The multidimensional derivative

[1] In the book of Stewart, there is a lot of emphasis on the *partial* derivatives of a function. This is understandable and has its perks. For instance, they are simple to calculate and appeal a lot to what we already know from one-dimensional calculus. It is however cumbersome to solely rely on them to perform calculations in dimensions higher than three (two, or even one, really). Also, the definition of differentiability becomes less transparent. In this short note, we will instead define the multidimensional derivative directly, without using partial derivatives. We will also discuss the chain rule from this perspective, and present the *Schrankensatz*, a theorem which can be used to derive secure measurement error bounds.

1 Definitions

Remember that the derivative $f'(x)$ of a (univariate) function gives the *rate of change* of the function at x, in the meaning that

$$f(x+v) \approx f(x) + f'(x)v$$

for small values of v. That is, the derivative answers the questions "What happens with $f(x)$ when we move away from x slightly?"

We want the multidimensional derivative to do the same thing. Hereby, the displacement $\mathbf{v}$ also is a vector, and the derivative must thus be an object which maps displacements in $\mathbf{x}$ to displacements in $f(\mathbf{x})$. In order to generalize results from 1D to more dimensions easily, we further need that map to be *linear* in the displacement vector. The only object which has this property is a *matrix*. Thus, we end up with the following definition:

Definition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function and $\mathbf{p} \in \mathbb{R}^n$. We say that f is* differentiable at $\mathbf{p}$ *if there exists a matrix $L \in \mathbb{R}^{m,n}$ with*

$$f(\mathbf{p}+\mathbf{v}) = f(\mathbf{p}) + L\mathbf{v} + \epsilon(\mathbf{v})\,|\mathbf{v}|\,, \tag{1}$$

where $\epsilon : \mathbb{R}^n \to \mathbb{R}$ is a function with $\lim_{\mathbf{v}\to\mathbf{0}} \epsilon(v) = \mathbf{0}$. The matrix, if it exists, is called the derivative *of f at p, and we write*

$$f'(\mathbf{p}) = L.$$

Remark 1. 1. *It is a good exercise to convince oneself that this is a way to define the derivative also in the one-variable case ($m = n = 1$).*

2. *Notice that although this is not visible in the notation, the 'error function' ϵ may be different (it almost always is!) for different points $\mathbf{p}$.*

[1] This note has been proofread, but it is very possible that it still contains typos. Please contact `flinth(at)chalmers.se` if such are found!

Example 2. Let us use the definition to calculate the derivative of the function

$$q : \mathbb{R}^n \to \mathbb{R},\ \mathbf{p} \to |\mathbf{p}|^2$$

We need to evaluate q in a point of the form $\mathbf{p}+\mathbf{v}$, and see that it can be written as a sum of a constant term $q(\mathbf{p})$, a linear term $L\mathbf{v}$, and a small error term. We use the formula $|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$

$$q(\mathbf{p}+\mathbf{v}) = |\mathbf{p}+\mathbf{v}|^2 = (\mathbf{p}+\mathbf{v}) \cdot (\mathbf{p}+\mathbf{v}) = \mathbf{p} \cdot \mathbf{p} + \mathbf{v} \cdot \mathbf{p} + \mathbf{v} \cdot \mathbf{p} + \mathbf{v} \cdot \mathbf{v} = q(\mathbf{p}) + 2\mathbf{p}^T\mathbf{v} + |\mathbf{v}| \cdot |\mathbf{v}|\,.$$

We used the fact that $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} = \mathbf{b}^T\mathbf{a}$. We see that this is exactly the formula (1) with $L = 2\mathbf{p}^T$ and $\epsilon(\mathbf{v}) = |\mathbf{v}|$. The latter clearly tends to 0 as $\mathbf{v}$ does. We have thus proven that

$$q'(\mathbf{p}) = 2\mathbf{p}^T.$$

□

The following theorem is useful for determining differentiability.

Theorem 3. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function. If all partial derivatives $\frac{\partial f}{\partial x_i}$, as functions $\mathbb{R}^n \to \mathbb{R}^m$, are defined in a neighbourhood of* $\mathbf{p}$ *and continuous at* $\mathbf{p}$, *f is differentiable in* $\mathbf{p}$.

Example 4. Consider the function

$$f(x, y, z) = x^2 y \cos(z).$$

We can easily compute the partial derivatives:

$$\frac{\partial f}{\partial x} 2xy\cos(z), \quad \frac{\partial f}{\partial y} x^2\cos(z), \quad \frac{\partial f}{\partial z} - x^2 y \sin(z).$$

These three functions are continuous as functions of (x, y, z). Therefore, f is differentiable everywhere.

□

The proof of this theorem is quite technical. We refer to (Appendix F of) Stewart for a proof.

How can we determine $f'(\mathbf{p})$? The key is to realize that the derivative $f'(\mathbf{p})$ maps vectors onto its directional derivatives. This is the meaning of the following theorem.

Theorem 5. *For any vector* $\mathbf{v}$,

$$D_{\mathbf{v}} f(\mathbf{p}) = f'(\mathbf{p})\mathbf{v}.$$

Proof. Let us examine the fraction we use to define the directional derivative.

$$\frac{f(\mathbf{p}+t\mathbf{v}) - f(\mathbf{p})}{t}$$

Here, we can clearly apply the definition of differentiability to rewrite the denominator

$$\begin{aligned}\frac{f(\mathbf{p}+t\mathbf{v}) - f(\mathbf{p})}{t} &= \frac{f(\mathbf{p}) + f'(\mathbf{p})(t\mathbf{v}) + \epsilon(t\mathbf{v}) \cdot |t\mathbf{v}| - f(\mathbf{p})}{t} = \frac{tf'(\mathbf{p})\mathbf{v}}{t} + \frac{\epsilon(t\mathbf{v}) \cdot |t\mathbf{v}|}{t} \\ &= f'(\mathbf{p})\mathbf{v} + \epsilon(t\mathbf{v})\,|\mathbf{v}| \cdot \frac{|t|}{t}.\end{aligned}$$

Now we let $t \to 0$. Then, $\epsilon(t\mathbf{v}) \to 0$, by the definition of differentiablity, and $\frac{|t|}{t}$ remains bounded (either equal to 1 or -1). Thus, the product tends to zero by the squeeze theorem, and

$$\frac{f(\mathbf{p} + t\mathbf{v}) - f(\mathbf{p})}{t} \to f'(\mathbf{p})\mathbf{v}.$$

□

If we apply this formula to the special cases of $\mathbf{v} = \mathbf{e}_i$, we obtain

$$f'(\mathbf{p})\mathbf{e}_i = D_{\mathbf{e}_i} f(\mathbf{p}) = \frac{\partial f(\mathbf{p})}{\partial x_i}.$$

Since for any matrix, $M\mathbf{e}_i$ is the i:th column of M, we thus obtain

$$f'(\mathbf{p}) = \begin{bmatrix} \frac{\partial f(\mathbf{p})}{\partial x_1} & \dots & \frac{\partial f(\mathbf{p})}{\partial x_n} \end{bmatrix}.$$

This matrix is sometimes referred to as the *Jacobian matrix.* We will however mostly refrain from this, as the term 'Jacobian' is used to refer to something slightly different in the book of Stewart.

Caution: Note that it may be possible to write down the Jacobian matrix although the function is not differentiable! If the partial derivatives all are continuous, however, we get differentiability.

Example 6. Let's calculate the derivative of the map

$$f : \mathbb{R}^2 \to \mathbb{R}^2, (x, y) \mapsto \begin{bmatrix} x^2/2 - y^2/2 \\ xy \end{bmatrix}.$$

We start by calculating the partial derivatives of f

$$\frac{\partial f}{\partial x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \frac{\partial f}{\partial y} = \begin{bmatrix} -y \\ x \end{bmatrix}.$$

These are both continuous as functions of x and y – hence, f really is differentiable. We may now write down the multidimensional derivative by putting the two partial derivatives next to each other, like this:

$$f'(x, y) = \begin{bmatrix} x & y \\ -y & x \end{bmatrix}$$

□

Remark 2. *While solving exercises in Stewart's book dealing with calculating partial derivatives, it is a good exercise try writing the derivative in matrix form as above.*

Exercise 1. Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \mathbf{p} \to |\mathbf{p}|.$$

(*i*) Calculate the partial derivatives of f in any point $\mathbf{p} \neq \mathbf{0}$. Show that f is differentiable in such points, and determine the derivative.

(*ii*) Calculate the directional derivatives $D_{\mathbf{v}} f(\mathbf{0})$ for $\mathbf{v} \in \mathbb{R}^2$. Is f differentiable in $\mathbf{0}$?

1.1 Differentiability implies continuity

Let's generalize a classic result from one-dimensional calculus to several variables

Theorem 1.1. *Let f be differentiable at $\mathbf{p}$. Then, f is continuous at $\mathbf{p}$.*

Proof. We will prove that

$$\lim_{\mathbf{v}\to\mathbf{0}} f(\mathbf{p}+\mathbf{v}) = f(\mathbf{p}),$$

which is the definition of continuity.

By the differentiability assumption, we can write $f(\mathbf{p}+\mathbf{v}) = f(\mathbf{p}) + f'(\mathbf{p})\mathbf{v} + \epsilon(\mathbf{v})\left|\mathbf{v}\right|$ for a function ϵ tending to zero as $\mathbf{v}$ does. This means that

$$\lim_{\mathbf{v}\to\mathbf{0}} f(\mathbf{p}+\mathbf{v}) = \lim_{\mathbf{v}\to\mathbf{0}} f(\mathbf{p}) + \lim_{\mathbf{v}\to\mathbf{0}} f'(\mathbf{p})\mathbf{v} + \lim_{\mathbf{v}\to\mathbf{0}} \epsilon(\mathbf{v})\left|\mathbf{v}\right|.$$

It is trivial that $\lim_{\mathbf{v}\to\mathbf{0}} f(\mathbf{p}) = f(\mathbf{p})$. It was proven in the lecture that linear maps $\mathbf{x} \mapsto A\mathbf{x}$ are continuous. This is in particular true for the linear map $\mathbf{v} \mapsto f'(\mathbf{p})\mathbf{v}$. This in particular implies that

$$\lim_{\mathbf{v}\to\mathbf{0}} f'(\mathbf{p})\mathbf{v} = f'(\mathbf{p})\mathbf{0} = \mathbf{0}.$$

Finally, the fact that $\epsilon(\mathbf{v})$ tends to zero as $\mathbf{p}$ tends to zero surely implies that

$$\lim_{\mathbf{v}\to\mathbf{0}} \epsilon(\mathbf{v})\left|\mathbf{v}\right| = \mathbf{0}\cdot 0 = \mathbf{0}.$$

All in all,

$$\lim_{\mathbf{v}\to\mathbf{0}} f(\mathbf{p}) + \lim_{\mathbf{v}\to\mathbf{0}} f'(\mathbf{p})\mathbf{v} + \lim_{\mathbf{v}\to\mathbf{0}} \epsilon(\mathbf{v})\left|\mathbf{v}\right| = f(\mathbf{p}) + \mathbf{0} + \mathbf{0} = f(\mathbf{p}),$$

which is the claim. □

2 The multidimensional chain rule

As a first example on how regarding the multidimensional derivative as a matrix can help us, let us formulate the multidimensional chain rule.

Theorem 7. *Let $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^n$ be differentiable in $\mathbf{p}$ and $g(\mathbf{p})$, respectively. Then, $f \circ g$ is differentiable in $\mathbf{p}$, with*

$$(f \circ g)'(\mathbf{p}) = f'(g(\mathbf{p}))g'(\mathbf{p}) \tag{2}$$

Notice that the above is a statement about equality of matrices. In particular, the product on the right hand side is the matrix product.

Remembering the geometrical meaning of matrix-vector multiplication, the result (2) is very natural. It says that in order to calculate the displacement of $f \circ g$ when moving in $\mathbf{v}$-direction, we need to first calculate the displacement the g-function that is caused. Then, this displacement is mapped by $f'(g(\mathbf{p}))$ to the final displacement. This is essentially also the idea of the proof – but the details are

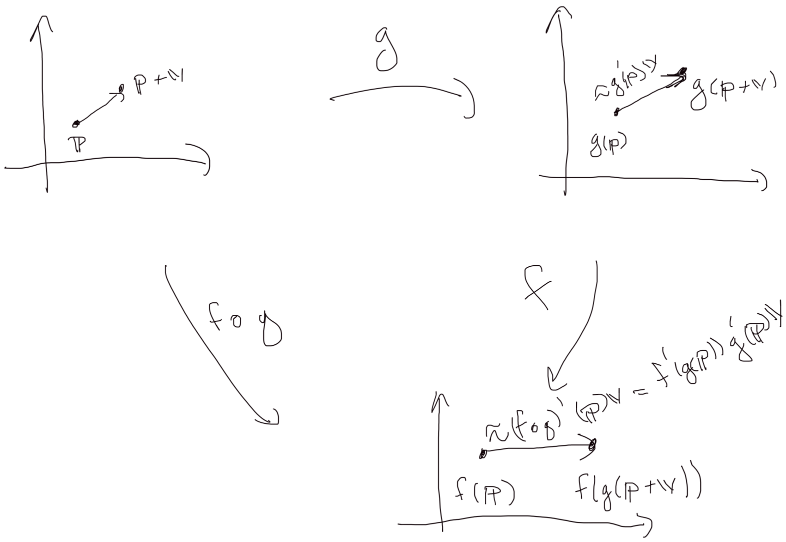

$\mathbf{p}+\mathbf{v}$
$\mathbf{p}$
g
$\approx g'(\mathbf{p})\mathbf{v}$
$g(\mathbf{p}+\mathbf{v})$
$g(\mathbf{p})$
$f \circ g$
f
$\approx (f \circ g)'(\mathbf{p})\mathbf{v} = f'(g(\mathbf{p}))g'(\mathbf{p})\mathbf{v}$
$f(\mathbf{p})$
$f(g(\mathbf{p}+\mathbf{v}))$

Figure 1: A graphical depiction of the chain rule. A displacement $\mathbf{v}$ (approximately) in $\mathbf{p}$-value causes a displacement $g'(\mathbf{p})\mathbf{v}$ in f-value, which in turn causes a displacement $f'(g(\mathbf{p}))g'(\mathbf{p})\mathbf{v}$ in f-value. That displacement is by definition approximately $(f \circ g)'(\mathbf{p})\mathbf{v}$.

quite technical. We postpone it to the end of this section.

The formula (2) compactly contains formulas for all the partial derivative of $f \circ g$, for any special cases discussed in Stewart, and even more. Let's 'unpack it'. We have

$$\frac{\partial f(g(\mathbf{p}))}{\partial x_i} = (f \circ g)'(\mathbf{p})\mathbf{e}_i = f'(g(\mathbf{p}))g'(\mathbf{p})e_i = f'(g(\mathbf{p}))\frac{\partial g}{\partial x_i}(\mathbf{p}).$$

Now we apply the formula for matrix-vector multiplication. Remember that it states that $f'(g(\mathbf{p}))\frac{\partial g}{\partial x_i}(\mathbf{p})$ is a linear combination of the columns $(f'(g(\mathbf{p})))_j$ of $f'(g(\mathbf{p}))$:

$$f'(g(\mathbf{p}))\frac{\partial g}{\partial x_i}(\mathbf{p}) = \sum_{j=1}^{p}(f'(g(\mathbf{p})))_j \frac{\partial g_j(\mathbf{p})}{\partial x_i}(\mathbf{p})$$

If we call the argument of f for $\mathbf{g}$, we have $(f'(g(\mathbf{p})))_j = \frac{\partial f(\mathbf{g}(\mathbf{p}))}{\partial g_j}$. Dropping the arguments, we arrive at the formala

$$\frac{\partial (f \circ g)}{\partial x_i} = \sum_{j=1}^{p} \frac{\partial f}{\partial g_j}\frac{\partial g_j}{\partial x_i},$$

which is a form you will probably encounter more often. Notice the appearant 'cancelling' ∂g_j-terms.

Example 8. Let f denote the map

$$f(x,y) =< \frac{x^2-y^2}{2}, xy >= \begin{bmatrix} \frac{x^2-y^2}{2} \\ xy \end{bmatrix}$$

Let us calculate the derivative of $|f(x,y)|^2$ with the help of the chain rule. We write it as a chain

$$|f(x,y)|^2 = (q \circ f)(x,y),$$

where q is the function $q(\mathbf{p}) = |\mathbf{p}|^2$ we discussed in the previous section. We know that $q'(\mathbf{p}) = 2\mathbf{p}^T$ and

$$f'(x,y) = \begin{bmatrix} x & -y \\ y & x \end{bmatrix}.$$

This has the consequence

$$\begin{aligned}(q \circ f)'(x,y) = q'(f(x,y))f'(x,y) &= 2\begin{bmatrix} \frac{x^2-y^2}{2} & xy \end{bmatrix}\begin{bmatrix} x & -y \\ y & x \end{bmatrix} \\ &= 2\begin{bmatrix} \frac{x^2-y^2}{2} \cdot x + xy \cdot y & \frac{x^2-y^2}{2} \cdot (-y) + xy \cdot x \end{bmatrix} = 2\begin{bmatrix} \frac{x^3+xy^2}{2} & \frac{y^3+x^2y}{2} \end{bmatrix}.\end{aligned}$$

It is surely also possible to calculate an explicit expression for $|q(x,y)|$ and calculate the partial derivatives directly, but this is quite tedious!

□

Exercise 2. *(i)* Calculate the derivative of $f \circ g$ using the chain rule

(a) $f(x,y) = xe^y$, $g(s,t) =< st, s+t >$. (b) $f(x,y,z) =< xy, yz >$, $g(x,y) =< x, y, x^2+y^2 >$.
(c) $f(t) = t^2$, $g : \mathbb{R}^3 \to \mathbb{R}$ arbitrary. (d) $f(x,y) =< x+y^2, y+x^2 >$, $g = f \circ f$.
Only evaluate the derivative of $f \circ g$ in $< 0, 0 >$!

(ii) Use the chain rule to prove the product rule $(f \cdot g) = f' \cdot g + f \cdot g'$, where f, and g are functions $\mathbb{R} \to \mathbb{R}$. *Tip:* Use the functions

$$m : \mathbb{R}^2 \to \mathbb{R}, (x,y) \mapsto xy \text{ and } \Phi : \mathbb{R} \to \mathbb{R}^2, t \mapsto < f(t), g(t) > .$$

Can you generalize the result to functions $f, g : \mathbb{R}^n \to \mathbb{R}^m$? (Here, '$\cdot$' is the dot product).

2.1 Excursion: Neural networks

This section is *entirely optional*, and above the intended learning outcomes of the course. It however concerns an increasingly important application, whence it should be of some interest to many of you. It might be a very good idea to skip it on first reading, and save it as an intellectual challenge while revising for the exam.

Let us apply the chain rule to *learning of neural networks.* Somewhat simplified[2], a one-layer neural network is a map of the following form:

$$\Phi_{A,\mathbf{b}} : \mathbb{R}^n \to \mathbb{R}^m, \mathbf{x} \mapsto \sigma(A\mathbf{x} + \mathbf{b}),$$

[2] For those who are familiar with the neural network lingo: we skip the final linear layer.

where $A \in \mathbb{R}^{m,n}$ is a matrix, $\mathbf{v} \in \mathbb{R}^m$ is a vector, and $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is a non-linearity applied pointwise. That is, σ is of the following form:

$$\sigma(x_1, \dots, x_n) =< f(x_1), \dots, f(x_n) > .$$

We want to assume that f is differentiable.

Neural networks has become extremely popular in machine learning, or AI if you will, in the last few years. They are used for many things – the most simple example is classification tasks. The idea hear is to choose A and $\mathbf{b}$ in such a way that

$$\Phi_{A,\mathbf{b}}(\mathbf{x}_i) \approx \mathbf{y}_i$$

for some data vectors $\mathbf{x}_i$ and label vectors $\mathbf{y}_i$ that is available to us. The hope is that the function then will be able to correctly predict the output $\mathbf{y}$ also when we feed it vectors $\mathbf{x}$ that it hasn't seen before. As a concrete and not so unrealistic example, we can imagine a hospital recording test results from blood tests from a lot of patients, and what deceases they were finally diagnosed with. In this case

- The $\mathbf{x}$-vectors are the results from the blood tests – note that we can test for a plethora of different substances.
- The $\mathbf{y}$-vectors are corresponding to the diagnoses. An element y_i being equal to 1 means that patient was diagnosed with illness i, and a zero that he/she was not.

Another example is calculating which move ($\mathbf{y}$) to perform given a game state ($\mathbf{x}$) in a game of go. The latter is actually one of the early examples of the success of neural networks – look up *AlphaGo*.

There is a lot to be said about the properties and intuition behind neural networks. We do not want to do this here – we refer to more specialized texts and courses for that. A short intro that I can recommend is the following `youtube series` by the (excellent) channel ThreeBlueOneBrown.

What we instead want to do here is to look at one mathematically very heavy step of machine learning using neural networks: the choice of the correct $\mathbf{A}$ and $\mathbf{b}$. This is namely done through *optimization*, i.e. assigning $\mathbf{A}$ and $\mathbf{b}$ (the *weights*) of the network so that

$$R(A, \mathbf{b}) = \sum_{i=1}^{N} |\Phi_{A,\mathbf{b}}(\mathbf{x}_i) - \mathbf{y}_i|^2$$

is as small as possible. Note that this quantity only is small if $\Phi_{A,\mathbf{b}}(\mathbf{x}_i) \approx \mathbf{y}_i$ for all i. Thus, minimizing it seems a like a good strategy.

In order to perform this optimization, we need to find the derivative of the above map with respect to the pair $(\mathbf{A}, \mathbf{b})$ (more information about this in the coming weeks). Let us calculate that derivative with the help of the chain rule. We make two simplifying assumptions:

- We only have one data-point, e.g. $N = 1$. This is actually not as unrealistic as is might seem – it is in fact standard to only consider one data-point *at a time* when adjusting the weights (the technical term is *stochastic gradient descent*). This is the only way to make the optimization feasible to perform on a computer when the set of data is huge. For each data point, the weights are slightly adjusted. Doing this for many data points, we end up with a good choice of weights.

- The vector(s) $\mathbf{y}_i$ are (a) one-dimensional scalar(s) y_i. This in particular implies that the matrix A is a vector $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b}$ is a scalar b. This simplification is in essence only made to simplify the notation – remember that differentiation of a vector-valued function can always be reduced to calculating the derivatives of the components.

Now notice that

$$R = c \circ \sigma \circ L,$$

where

$$\begin{aligned}
&L : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, (\mathbf{a}, b) \mapsto \mathbf{a} \cdot \mathbf{x} + b \\
&\sigma : \mathbb{R} \to \mathbb{R}, \sigma(x) = f(x) \\
&c : \mathbb{R} \to \mathbb{R}, t \mapsto (t - y)^2.
\end{aligned}$$

That means that

$$R'(\mathbf{a}, b) = c(\sigma(L(\mathbf{a}, b))\sigma'(L(\mathbf{a}, b))L'(\mathbf{a}, b).$$

In the exercises of this section, you will show that

$$L'(\mathbf{a}, b) = [\mathbf{x}^T 1], \quad \sigma'(x) = f'(x), \qquad c'(t) = 2(t - y).$$

Thus, defining $t = \mathbf{a} \cdot \mathbf{x} + b$

$$R'(\mathbf{a}, b) = 2(\sigma(\mathbf{t}) - y)f'(t)^T[\mathbf{x}^T, 1].$$

If we transpose this, we obtain

$$\nabla R(\mathbf{a}, b) = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} f'(t) \cdot 2(\sigma(t) - y) \tag{3}$$

The formula (3) is commonly referred to as the *back-propagation formula.* Notice that the current error $\sigma(t) - y$ is being multiplied with the derivative of the f-functio , to then be mapped ('fed back') to 'change directions' $f'(t) \cdot 2(\sigma(t) - y)\mathbf{x}$ and $f'(t) \cdot 2(\sigma(t) - y)$ in the weights $\mathbf{a}$ and b, respectively.

Remark 3. *The chain rule is also the key to understanding how to optimize weights for bigger networks. In fact, neural networks really start to be able to do amazing things when objects like the ones described above (*layers*) are put together in a chain. Such a* deep *network can be elegantly defined as* $\Phi_{A_0,\ldots,A_{k-1},\mathbf{b}_0,\ldots,\mathbf{b}_{k-1}} = \sigma(\mathbf{t}_k)$*, where*

$$\begin{aligned}
\mathbf{t}_\ell &= A_{\ell-1}\sigma(\mathbf{t}_{\ell-1}) + \mathbf{b}_{\ell-1}, \ell \geq 1 \\
\mathbf{t}_0 &= A_0\mathbf{x} + \mathbf{b}_0.
\end{aligned}$$

To write down the formulas for the gradient for the above with respect to any A_ℓ *or* $\mathbf{b}_\ell$ *requires a lot more notation and algebra than what we did above, but it should be clear that we can in principle do it using the chain rule. After all, we have calculated the derivative of one 'layer function'* $\sigma(\mathbf{t}_k)$*, and a neural network is just a chain of such layers.*

Not surprisingly, the result is a new kind of back propagation. Again, the current error $\mathbf{y}-\sigma(\mathbf{t}_k)$ *is fed through a derivative map* $\sigma'(\mathbf{t}_k)^T$*, and from there mapped onto change directions for* A_{k-1}*,* $\mathbf{b}_{k-1}$ *and (this is different from before) also one in* $\sigma(\mathbf{t}_{k-1})$*. The latter is then fed through* $\sigma'(\mathbf{t}_{k-1})^T$*, and so on. Note that we are deliberately staying inprecise here. If we would not, we would probably just be overwhelmed by the notation. We refer to textbooks and/or courses on deep learning for a more comprehensive treatise of back-propagation. The above mentioned youtube series is also a nice alternative.*

Exercise 3. Check that the formula for R, and for the derivatives for the functions L and c are correct.

2.2 Proof of the chain rule

For completeness, here is the proof of the multidimensional chain rule.

Proof of the chain rule. Let's first introduce a notation to increase readability: letting $\mathbf{h} = g(\mathbf{p}+\mathbf{v}) - g(\mathbf{p})$, we trivially have $g(\mathbf{p}+\mathbf{v}) = g(\mathbf{p}) + \mathbf{h}$. We can now use the fact that f is differentiable in $g(\mathbf{p})$ to write

$$f(g(\mathbf{p}+\mathbf{v})) = f(g(\mathbf{p})+\mathbf{h}) = f(g(\mathbf{p})) + f'(g(\mathbf{p}))\mathbf{h} + |\mathbf{h}|\,\epsilon_f(\mathbf{h})$$

Now we need to 'unpack' $\mathbf{h}$. The definition of differentiability proves that

$$\mathbf{h} = g(\mathbf{p}+\mathbf{v}) - g(\mathbf{p}) = g(\mathbf{p}) + g'(\mathbf{p})\mathbf{v} + \epsilon_g(\mathbf{v})\,|\mathbf{v}| - g(\mathbf{p}) = g'(\mathbf{p})\mathbf{v} + \epsilon_g(\mathbf{v})\,|\mathbf{v}|$$

Thus

$$\begin{aligned} f(g(\mathbf{p})) + f'(g(\mathbf{p}))\mathbf{h} + |\mathbf{h}|\,\epsilon_f(\mathbf{h}) &= f(g(\mathbf{p})) + f'(g(\mathbf{p}))\left(g'(\mathbf{p})\mathbf{v} + \epsilon_g(\mathbf{v})\,|\mathbf{v}|\right) + |\mathbf{h}|\,\epsilon_f(\mathbf{h}) \\ &= f(g(\mathbf{p})) + f'(g(\mathbf{p}))g'(\mathbf{p})\mathbf{v} + |\mathbf{v}|\, f'(g(\mathbf{p}))\epsilon_g(\mathbf{v}) + |\mathbf{h}|\,\epsilon_f(\mathbf{h}) \end{aligned}$$

We notice the constant term $f(g(\mathbf{p}))$ and the linear term $f'(g(\mathbf{p}))g'(\mathbf{p})\mathbf{v}$. If we can prove that the error term can be written in the form $\epsilon(\mathbf{v})\,|v|$, we are done. We have

$$|\mathbf{v}|\, f'(g(\mathbf{p}))\epsilon_g(\mathbf{v}) + |\mathbf{h}|\,\epsilon_f(\mathbf{h}) = |\mathbf{v}| \cdot \left(f'(g(\mathbf{p}))\epsilon_g(\mathbf{v}) + \frac{|\mathbf{h}|}{|\mathbf{v}|}\epsilon_f(\mathbf{h})\right)$$

We hence define

$$\epsilon(\mathbf{v}) = \left(f'(g(\mathbf{p}))\epsilon_g(\mathbf{v}) + \frac{|\mathbf{h}|}{|\mathbf{v}|}\epsilon_f(\mathbf{h})\right)$$

If we can prove this function goes to zero as $\mathbf{v} \to \mathbf{0}$, we are done. First note that $\epsilon_g(\mathbf{v}) \to 0$ as $\mathbf{v} \to \mathbf{0}$. This has the consequence that

$$f'(g(\mathbf{p}))\epsilon_g(\mathbf{v}) \to f'(g(\mathbf{p}))\mathbf{0}) = \mathbf{0}.$$

To deal with the rest, first notice that due to the continuity of g in $\mathbf{p}$ (it is differentiable, therefore continuous),

$$\lim_{\mathbf{v}\to\mathbf{0}} \mathbf{h} = \lim_{\mathbf{v}\to\mathbf{0}} g(\mathbf{p}+\mathbf{v}) - g(\mathbf{p}) = g(\mathbf{p}) - g(\mathbf{p}) = \mathbf{0}.$$

This implies that $\lim_{\mathbf{v}\to\mathbf{0}} \epsilon_f(\mathbf{h}) = \lim_{\mathbf{h}\to\mathbf{0}} \epsilon_f(\mathbf{h}) = 0$.

We still need to control the quotient in front of $\epsilon_f(\mathbf{h}$. Remember that $g'(\mathbf{p})$ is a matrix. This implies that there exists a constant $C \geq 0$ so that $|g'(\mathbf{p})\mathbf{v}| \leq C\,|\mathbf{v}|$ (this was mentioned in lecture 2). This implies that

$$\frac{|\mathbf{h}|}{|\mathbf{v}|} = \frac{|g'(\mathbf{p})\mathbf{v} + \epsilon_g(\mathbf{v})\,|\mathbf{v}||}{|\mathbf{v}|} \leq \frac{C\,|\mathbf{v}| + \epsilon_g(\mathbf{v})\,|\mathbf{v}|}{|\mathbf{v}|} = C + \epsilon_g(\mathbf{v}).$$

That is, the quotient $\frac{|\mathbf{h}|}{|\mathbf{v}|}$ is bounded for $\mathbf{v} \to \mathbf{0}$. We may therefore apply the squeeze theorem to prove

$$\lim_{\mathbf{v}\to\mathbf{0}} \frac{|\mathbf{h}|}{|\mathbf{v}|}\epsilon_f(\mathbf{h}) = 0.$$

□

3 The 'Schrankensatz'

As we have seen, the derivative gives us information about what happens with $f(\mathbf{p})$ when we change the value of $\mathbf{p}$ slightly. If have some sort of bound on the 'size' of the derivative over an entire domain D, we should therefore be able to give a bound on how much the f-values vary throughout D. The following theorem gives us such a bound

Theorem 9. *Let $f : D \to \mathbb{R}^m$ be a differentiable function, where $D \subseteq \mathbb{R}^n$. Let $M_1, \ldots, M_n$ be defined through*

$$M_i = \max_{\mathbf{p}\in D} \left|\frac{\partial f}{\partial x_i}\right|(\mathbf{p}).$$

Now let $\mathbf{a}$ and $\mathbf{b}$ be two points in D such that the entire line segment between them is contained in D. Then,

$$|f(\mathbf{a}) - f(\mathbf{b})| \leq \sum_{i=1}^{n} M_i\,|a_i - b_i|$$

We will refer to this theorem through its german name, the 'Schrankensatz', which literally translates to the 'bound theorem'. The only reason for this is that there is no universally agreed term for this in the English literature!

Proof of the theorem in the case $m = 1$. We will only be able to prove the theorem in the case $m = 1$. Since this arguably is the most important case, it will more than suffice.

Define the function

$$g : [0,1] \to \mathbb{R}, t \mapsto f(\mathbf{a} + t(\mathbf{b} - \mathbf{a})).$$

We then aim to give a bound on $|g(1) - g(0)|$. Note that g is differentiable everywhere and in particular continuous in the endpoints. We can thus apply the mean-value theorem to it:

$$g(1) - g(0) = g'(\xi)(1 - 0)$$

for some $\xi \in [0,1]$. By the chain rule, we have $g'(t) = f'(\mathbf{a} + t(\mathbf{b} - \mathbf{a})(\mathbf{b} - \mathbf{a})$. Now we can estimate

$$|g(1) - g(0)| = \left|g'(\xi)\right| = \left|f'(\mathbf{a} + t(\mathbf{b} - \mathbf{a}))(\mathbf{b} - \mathbf{a})\right| = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(b_i - a_i)\right| (\mathbf{a} + t(\mathbf{b} - \mathbf{a})$$

$$\leq \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}(\mathbf{a} + t(\mathbf{b} - \mathbf{a})\right| \cdot |b_i - a_i| \leq \sum_{i=1}^{n} M_i \, |b_i - a_i|$$

We applied the triangle inequality, and the definition of the M_i's. □

Remark 4. *The proof will not work for $m > 1$, since there is no mean value theorem for functions f mapping into $\mathbb{R}^m$ for $m > 1$. As a counter-example, consider*

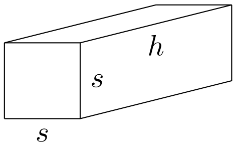
$$f(t) =< \cos(t), \sin(t) > .$$

Then, $f(2\pi) - f(0) = 0$, but there is no ξ with $f'(\xi) = 0$.

This theorem is of high interest to an engineer, as it gives a way to quantify the effect of measurement errors on certain quantities on the precision of estimation of other quantities. Consider the following concrete example.

Example 10. Let us assume that we are given the task of determining the volume of box with two equal sidelenghts. We measure, using a ruler with cm-marks, the equal side length s and the height h of it to be equal to $s_0 = 1.1$dm and $h_0 = 3.2$dm and thus determine a volume measurement of

$$V_0 = s_0^2 h_0 = 1.1^2 \cdot 3.2 = 3.872\text{dm}^3.$$



Can we however be sure that that is the actual value of the box? After all, we used a ruler to determine s_0 and h_0. Any actual value between 10.5 and 11.5cm would yield a measurement of 11 cm for the equal side length, and similarly for h. Thus, we are only sure that the *actual* side length and height $(\tilde{s}, \tilde{h})$ fulfill $|\tilde{s} - s_0| \leq 0.05$dm and $\left|\tilde{h} - h_0\right| \leq 0.05$dm. That is, we are only sure that

$$(\tilde{s}, \tilde{h}) \in S = [1.05, 1.15] \times [3.15, 3.25].$$

We can however give a bound on the volume using the Schrankensatz. Consider the function

$$V(s,h) = s^2 h.$$

We then need to give a bound on $\left|V(\tilde{s}, \tilde{h}) - V(s_0, h_0)\right|$. Let us begin by calculating the M_i-numbers:

$$M_1 = \max_{(s,h) \in S} \left|\frac{\partial V}{\partial s}\right| = \max_{(s,h) \in S} 2sh = 2 \cdot 3.25 \cdot 1.15,$$

$$M_2 = \max_{(s,h) \in S} \left|\frac{\partial V}{\partial h}\right| = \max_{(s,h) \in S} s^2 = 1.15^2.$$

Notice that we may safely remove the absolute values, since all variables are positive. Also notice that it in this case was easy to see which values for (s, h) to choose in order to maximize the partial derivatives. In situations where this not is the case, it is enough to give upper bounds for the quantities – the theorem will surely remain true using these values for M_i, since the bound becomes 'worse'.

The Schrankensatz then says

$$\left|V(\tilde{s}, \tilde{h}) - V(s_0, h_0)\right| \leq M_1 \left|\tilde{s} - s_0\right| + M_2 \left|\tilde{h} - h_0\right| \leq 2 \cdot 3.25 \cdot 1.15 \cdot 0.05 + 1.15^2 \cdot .05 \leq 0.44$$

where the final $\leq$ is again due to rounding. We can thus safely say that the actual volume of the box is

$$V(\tilde{s}, \tilde{h}) = 3.872 \pm 0.44 \text{ m}^3.$$

□

Remark 5. *In this concrete task, we can of course give a lower bound on the volume by simply using the smallest values possible for s and h, and similarly an upper bound. In a more general scenario,however, it might however not be so easy to determine the smallest possible value of the function over the range of the possible inputs. Also, the method will surely not be applicable when we try to estimate the measure error of a vector-valued quantity. Hence, the Schrankensatz surely has its raison d'être.*

Exercise 4. A company is producing peanuts. They sell them in bags with 500 grams in each bag. Their machine is not flawless, so it that the weight of the bag is subject to a measurement error of 10 grams. The company's peanuts weigh 0.45 grams each, with outliers of up to 0.05 grams in each direction possible.

Use the Schrankensatz to give an estimate of the number of peanuts in a bag, with error bounds.