Review of differential calculus theory

Keywords: Differential, Gradients, partial derivatives, Jacobian, chain-rule

This note is optional and is aimed at students who wish to have a deeper understanding of differential calculus. It defines and explains the links between derivatives, gradients, jacobians, etc. First, we go through definitions and examples for $f : \mathbb{R}^n \mapsto \mathbb{R}$. Then we introduce the Jacobian and generalize to higher dimension. Finally, we introduce the chain-rule.

1 *Introduction*

We use derivatives all the time, but we forget what they mean. In general, we have in mind that for a function $f : \mathbb{R} \mapsto \mathbb{R}$, we have something like

$$f(x+h) - f(x) \approx f'(x)h$$

Some people use different notation, especially when dealing with higher dimensions, and there usually is a lot of confusion between the following notations

$$f'(x)$$
$$\frac{\mathrm{d}f}{\mathrm{d}x}$$
$$\frac{\partial f}{\partial x}$$
$$\nabla_x f$$

However, these notations refer to different mathematical objects, and the confusion can lead to mistakes. This paper recalls some notions about these objects.

Scalar-product and dot-product
Given two vectors a and b,

- **scalar-product** $\langle a|b\rangle = \sum_{i=1}^{n} a_i b_i$
- **dot-product** $a^T \cdot b = \langle a|b\rangle = \sum_{i=1}^{n} a_i b_i$

2 *Theory for $f : \mathbb{R}^n \mapsto \mathbb{R}$*

2.1 *Differential*

Formal definition

Let's consider a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined on $\mathbb{R}^n$ with the scalar product $\langle \cdot | \cdot \rangle$. We suppose that this function is **differentiable**, which means that for $x \in \mathbb{R}^n$ (fixed) and a small variation h (can change) we can write:

$$f(x+h) = f(x) + \mathrm{d}_x f(h) + o_{h \to 0}(h) \tag{1}$$

Notation
$\mathrm{d}_x f$ is a **linear form** $\mathbb{R}^n \mapsto \mathbb{R}$
This is the best **linear approximation** of the function f

$\mathrm{d}_x f$ is called the **differential** of f in x

$o_{h \to 0}(h)$ (Landau notation) is equivalent to the existence of a function $\epsilon(h)$ such that $\lim_{h \to 0} \epsilon(h) = 0$

and $\mathrm{d}_x f : \mathbb{R}^n \mapsto \mathbb{R}$ is a linear form, which means that $\forall x, y \in \mathbb{R}^n$, we have $\mathrm{d}_x f(x+y) = \mathrm{d}_x f(x) + \mathrm{d}_x f(y)$.

Example

Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $f(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) = 3x_1 + x_2^2$. Let's pick $\begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2$ and $h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \in \mathbb{R}^2$. We have

$$\begin{aligned} f(\begin{pmatrix} a+h_1 \\ b+h_2 \end{pmatrix}) &= 3(a+h_1) + (b+h_2)^2 \\ &= 3a + 3h_1 + b^2 + 2bh_2 + h_2^2 \\ &= 3a + b^2 + 3h_1 + 2bh_2 + h_2^2 \\ &= f(a,b) + 3h_1 + 2bh_2 + o(h) \end{aligned}$$

$h^2 = h \cdot h = o_{h \to 0}(h)$

Then, $\mathrm{d}_{\begin{pmatrix} a \\ b \end{pmatrix}} f(\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}) = 3h_1 + 2bh_2$

2.2 *Link with the gradients*

Formal definition

It can be shown that for all linear forms $a : \mathbb{R}^n \mapsto \mathbb{R}$, there exists a vector $u_a \in \mathbb{R}^n$ such that $\forall h \in \mathbb{R}^n$

$$a(h) = \langle u_a | h \rangle$$

Notation for $x \in \mathbb{R}^n$, the gradient is usually written $\nabla_x f \in \mathbb{R}^n$

The dual of a vector space E^* is isomorphic to E
See Riesz representation theorem

In particular, for the **differential** $\mathrm{d}_x f$, we can find a vector $u \in \mathbb{R}^n$ such that

$$\mathrm{d}_x(h) = \langle u | h \rangle$$

.

The gradient has the **same shape** as x

We can thus define the **gradient** of f in x

$$\nabla_x f := u$$

Then, as a conclusion, we can rewrite equation 2.1

Gradients and **differential** of a function are conceptually very different. The **gradient** is a vector, while the **differential** is a function

$$f(x+h) = f(x) + \mathrm{d}_x f(h) + o_{h\to 0}(h) \tag{2}$$
$$= f(x) + \langle \nabla_x f | h \rangle + o_{h\to 0}(h) \tag{3}$$

Example

Same example as before, $f : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $f(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) = 3x_1 + x_2^2$. We showed that

$$\mathrm{d}_{\begin{pmatrix} a \\ b \end{pmatrix}} f(\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}) = 3h_1 + 2bh_2$$

We can rewrite this as

$$\mathrm{d}_{\begin{pmatrix} a \\ b \end{pmatrix}} f(\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}) = \langle \begin{pmatrix} 3 \\ 2b \end{pmatrix} | \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \rangle$$

and thus our gradient is

$$\nabla_{\begin{pmatrix} a \\ b \end{pmatrix}} f = \begin{pmatrix} 3 \\ 2b \end{pmatrix}$$

2.3 *Partial derivatives*

Formal definition

Now, let's consider an orthonormal basis $(e_1, \dots, e_n)$ of $\mathbb{R}^n$. Let's define the partial derivative

$$\frac{\partial f}{\partial x_i}(x) := \lim_{h\to 0} \frac{f(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_n)}{h}$$

Notation

Partial derivatives are usually written $\frac{\partial f}{\partial x}$ but you may also see $\partial_x f$ or f'_x

- $\frac{\partial f}{\partial x_i}$ is a **function** $\mathbb{R}^n \mapsto \mathbb{R}$
- $\frac{\partial f}{\partial x} = (\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n})^T$ is a **function** $\mathbb{R}^n \mapsto \mathbb{R}^n$.
- $\frac{\partial f}{\partial x_i}(x) \in \mathbb{R}$
- $\frac{\partial f}{\partial x}(x) = (\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x))^T \in \mathbb{R}^n$

Note that the partial derivative $\frac{\partial f}{\partial x_i}(x) \in \mathbb{R}$ and that it is defined *with respect to the i-th component and evaluated in x.*

Example

Same example as before, $f : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $f(x_1, x_2) = 3x_1 + x_2^2$. Let's write

Depending on the context, most people omit to write the (x) evaluation and just write $\frac{\partial f}{\partial x} \in \mathbb{R}^n$ instead of $\frac{\partial f}{\partial x}(x)$

$$\begin{aligned}\frac{\partial f}{\partial x_1}(\begin{pmatrix}a\\b\end{pmatrix}) &= \lim_{h\to 0}\frac{f(\begin{pmatrix}a+h\\b\end{pmatrix}) - f(\begin{pmatrix}a\\b\end{pmatrix})}{h}\\ &= \lim_{h\to 0}\frac{3(a+h)+b^2-(3a+b^2)}{h}\\ &= \lim_{h\to 0}\frac{3h}{h}\\ &= 3\end{aligned}$$

In a similar way, we find that

$$\frac{\partial f}{\partial x_2}(\begin{pmatrix}a\\b\end{pmatrix}) = 2b$$

2.4 Link with the partial derivatives

Formal definition

It can be shown that

$$\begin{aligned}\nabla_x f &= \sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(x)e_i\\ &= \begin{pmatrix}\frac{\partial f}{\partial x_1}(x)\\ \vdots \\ \frac{\partial f}{\partial x_n}(x)\end{pmatrix}\end{aligned}$$

That's why we usually write

$$\nabla_x f = \frac{\partial f}{\partial x}(x)$$

(**same shape as** x)

e_i is a orthonormal basis. For instance, in the canonical basis

$$e_i = (0,\dots,1,\dots 0)$$

with 1 at index i

where $\frac{\partial f}{\partial x_i}(x)$ denotes the partial derivative of f with respect to the ith component, evaluated in x.

Example

We showed that

$$\begin{cases}\frac{\partial f}{\partial x_1}(\begin{pmatrix}a\\b\end{pmatrix}) &= 3\\ \frac{\partial f}{\partial x_2}(\begin{pmatrix}a\\b\end{pmatrix}) &= 2b\end{cases}$$

and that

$$\nabla_{\begin{pmatrix}a\\b\end{pmatrix}} f = \begin{pmatrix}3\\2b\end{pmatrix}$$

and then we verify that

$$\nabla \begin{pmatrix} a \\ b \end{pmatrix} f = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\begin{pmatrix} a \\ b \end{pmatrix}) \\ \\ \frac{\partial f}{\partial x_2}(\begin{pmatrix} a \\ b \end{pmatrix}) \end{pmatrix}$$

3 *Summary*

Formal definition

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, we have defined the following objects which can be summarized in the following equation

Recall that $a^T \cdot b = \langle a|b\rangle = \sum_{i=1}^n a_i b_i$

$$\begin{aligned} f(x+h) &= f(x) + \mathrm{d}_x f(h) + o_{h\to 0}(h) && \textbf{differential} \\ &= f(x) + \langle \nabla_x f | h \rangle + o_{h\to 0}(h) && \textbf{gradient} \\ &= f(x) + \langle \frac{\partial f}{\partial x}(x) | h \rangle + o_{h\to 0} \\ &= f(x) + \langle \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} | h \rangle + o_{h\to 0} && \textbf{partial derivatives} \end{aligned}$$

Remark

Let's consider $x : \mathbb{R} \mapsto \mathbb{R}$ such that $x(u) = u$ for all u. Then we can easily check that $\mathrm{d}_u x(h) = h$. As this differential does not depend on u, we may simply write $\mathrm{d}x$. That's why the following expression has some meaning,

The $\mathrm{d}x$ that we use refers to the differential of $u \mapsto u$, the identity mapping!

$$\mathrm{d}_x f(\cdot) = \frac{\partial f}{\partial x}(x)\mathrm{d}x(\cdot)$$

because

$$\begin{aligned} \mathrm{d}_x f(h) &= \frac{\partial f}{\partial x}(x)\mathrm{d}x(h) \\ &= \frac{\partial f}{\partial x}(x)h \end{aligned}$$

In higher dimension, we write

$$\mathrm{d}_x f = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x)\mathrm{d}x_i$$

4 ***Jacobian**: Generalization to $f : \mathbb{R}^n \mapsto \mathbb{R}^m$*

For a function

$$f : \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{pmatrix}$$

We can apply the previous section to each $f_i(x)$:

$$\begin{aligned} f_i(x+h) &= f_i(x) + \mathrm{d}_x f_i(h) + o_{h\to 0}(h) \\ &= f_i(x) + \langle \nabla_x f_i | h \rangle + o_{h\to 0}(h) \\ &= f_i(x) + \langle \frac{\partial f_i}{\partial x}(x) | h \rangle + o_{h\to 0} \\ &= f_i(x) + \langle (\frac{\partial f_i}{\partial x_1}(x), \dots, \frac{\partial f_i}{\partial x_n}(x))^T | h \rangle + o_{h\to 0} \end{aligned}$$

Putting all this in the same vector yields

$$f \begin{pmatrix} x_1 + h_1 \\ \vdots \\ x_n + h_n \end{pmatrix} = f \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1}{\partial x}(x)^T \cdot h \\ \vdots \\ \frac{\partial f_m}{\partial x}(x)^T \cdot h \end{pmatrix} + o(h)$$

Now, let's define the **Jacobian** matrix as

The **Jacobian** matrix has dimensions $m \times n$ and is a generalization of the gradient

$$J(x) := \begin{pmatrix} \frac{\partial f_1}{\partial x}(x)^T \\ \vdots \\ \frac{\partial f_m}{\partial x}(x)^T \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) \dots \frac{\partial f_1}{\partial x_n}(x) \\ \ddots \\ \frac{\partial f_m}{\partial x_1}(x) \dots \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}$$

Then, we have that

$$\begin{aligned} f \begin{pmatrix} x_1 + h_1 \\ \vdots \\ x_n + h_n \end{pmatrix} &= f \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) \dots \frac{\partial f_1}{\partial x_n}(x) \\ \ddots \\ \frac{\partial f_m}{\partial x_1}(x) \dots \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix} \cdot h + o(h) \\ &= f(x) + J(x) \cdot h + o(h) \end{aligned}$$

Example 1 : $m = 1$

In the case where $m = 1$, the **Jacobian** is a **row vector**
$\frac{\partial f_1}{\partial x_1}(x) \dots \frac{\partial f_1}{\partial x_n}(x)$
Remember that our **gradient** was defined as a column vector with the same elements. We thus have that
$J(x) = \nabla_x f^T$

Let's take our first function $f : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $f(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) = 3x_1 + x_2^2$. Then, the Jacobian of f is

$$\begin{aligned} \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) & \frac{\partial f}{\partial x_2}(x) \end{pmatrix} &= \begin{pmatrix} 3 & 2x_2 \end{pmatrix} \\ &= \begin{pmatrix} 3 \\ 2x_2 \end{pmatrix}^T \\ &= \nabla_f(x)^T \end{aligned}$$

Example 2 : $g : \mathbb{R}^3 \mapsto \mathbb{R}^2$ Let's define

$$g(\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) = \begin{pmatrix} y_1 + 2y_2 + 3y_3 \\ y_1 y_2 y_3 \end{pmatrix}$$

Then, the Jacobian of g is

$$\begin{aligned} J_g(y) &= \begin{pmatrix} \frac{\partial(y_1+2y_2+3y_3)}{\partial y}(y)^T \\ \frac{\partial(y_1 y_2 y_3)}{\partial y}(y)^T \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial(y_1+2y_2+3y_3)}{\partial y_1}(y) & \frac{\partial(y_1+2y_2+3y_3)}{\partial y_2}(y) & \frac{\partial(y_1+2y_2+3y_3)}{\partial y_3}(y) \\ \frac{\partial(y_1 y_2 y_3)}{\partial y_1}(y) & \frac{\partial(y_1 y_2 y_3)}{\partial y_2}(y) & \frac{\partial(y_1 y_2 y_3)}{\partial y_3}(y) \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 & 3 \\ y_2 y_3 & y_1 y_3 & y_1 y_2 \end{pmatrix} \end{aligned}$$

5 *Generalization to* $f : \mathbb{R}^{n \times p} \mapsto \mathbb{R}$

If a function takes as input a matrix $A \in \mathbb{R}^{n \times p}$, we can transform this matrix into a vector $a \in \mathbb{R}^{np}$, such that

$$A[i, j] = a[i + nj]$$

Then, we end up with a function $\tilde{f} : \mathbb{R}^{np} \mapsto \mathbb{R}$. We can apply the results from 3 and we obtain for $x, h \in \mathbb{R}^{np}$ corresponding to $X, h \in \mathbb{R}^{n \times p}$,

$$\tilde{f}(x + h) = f(x) + \langle \nabla_x f | h \rangle + o(h)$$

where $\nabla_x f = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_{np}(x)} \end{pmatrix}$.

Now, we would like to give some meaning to the following equation

$$f(X + H) = f(X) + \langle \nabla_X f | H \rangle + o(H)$$

The gradient of f wrt to a matrix X is a matrix of same shape as X and defined by
$\nabla_X f_{ij} = \frac{\partial f}{\partial X_{ij}}(X)$

Now, you can check that if you define

$$\nabla_X f_{ij} = \frac{\partial f}{\partial X_{ij}}(X)$$

that these two terms are equivalent

$$\langle \nabla_x f | h \rangle = \langle \nabla_X f | H \rangle$$
$$\sum_{i=1}^{np} \frac{\partial f}{\partial x_i}(x) h_i = \sum_{i,j} \frac{\partial f}{\partial X_{ij}}(X) H_{ij}$$

6 *Generalization to $f : \mathbb{R}^{n \times p} \mapsto \mathbb{R}^m$*

Let's generalize the generalization of the previous section

Applying the same idea as before, we can write

$$f(x+h) = f(x) + J(x) \cdot h + o(h)$$

where J has dimension $m \times n \times p$ and is defined as

$$J_{ijk}(x) = \frac{\partial f_i}{\partial X_{jk}}(x)$$

Writing the 2d-dot product $\delta = J(x) \cdot h \in \mathbb{R}^m$ means that the i-th component of δ is

You can apply the same idea to any dimensions!

$$\delta_i = \sum_{j=1}^{n} \sum_{k=1}^{p} \frac{\partial f_i}{\partial X_{jk}}(x) h_{jk}$$

7 *Chain-rule*

Formal definition

Now let's consider $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ and $g : \mathbb{R}^p \mapsto \mathbb{R}^n$. We want to compute the **differential** of the composition $h = f \circ g$ such that $h : x \mapsto u = g(x) \mapsto f(g(x)) = f(u)$, or

$$\mathrm{d}_x (f \circ g)$$

.

It can be shown that the differential is the composition of the differentials

$$\mathrm{d}_x (f \circ g) = \mathrm{d}_{g(x)} f \circ \mathrm{d}_x g$$

Where $\circ$ is the composition operator. Here, $\mathrm{d}_{g(x)} f$ and $\mathrm{d}_x g$ are linear transformations (see section 4). Then, the resulting differential is also a linear transformation and the **jacobian** is just the dot product between the jacobians. In other words,

The **chain-rule** is just writing the resulting **jacobian** as a dot product of **jacobians**. Order of the dot product is very important!

$$J_h(x) = J_f(g(x)) \cdot J_g(x)$$

where $\cdot$ is the dot-product. This dot-product between two matrices can also be written component-wise:

$$J_h(x)_{ij} = \sum_{k=1}^{n} J_f(g(x))_{ik} \cdot J_g(x)_{kj}$$

Example

Let's keep our example function $f : (\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) \mapsto 3x_1 + x_2^2$ and our function $g : (\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) = \begin{pmatrix} y_1 + 2y_2 + 3y_3 \\ y_1 y_2 y_3 \end{pmatrix}$.

The composition of f and g is $h = f \circ g : \mathbb{R}^3 \mapsto \mathbb{R}$

$$\begin{aligned} h(\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) &= f(\begin{pmatrix} y_1 + 2y_2 + 3y_3 \\ y_1 y_2 y_3 \end{pmatrix}) \\ &= 3(y_1 + 2y_2 + 3y_3) + (y_1 y_2 y_3)^2 \end{aligned}$$

We can compute the three components of the gradient of h with the partial derivatives

$$\begin{aligned} \frac{\partial h}{\partial y_1}(y) &= 3 + 2y_1 y_2^2 y_3^2 \\ \frac{\partial h}{\partial y_2}(y) &= 6 + 2y_2 y_1^2 y_3^2 \\ \frac{\partial h}{\partial y_3}(y) &= 9 + 2y_3 y_1^2 y_2^2 \end{aligned}$$

And then our gradient is

$$\nabla_y h = \begin{pmatrix} 3 + 2y_1 y_2^2 y_3^2 \\ 6 + 2y_2 y_1^2 y_3^2 \\ 9 + 2y_3 y_1^2 y_2^2 \end{pmatrix}$$

In this process, we did not use our previous calculation, and that's a shame. Let's use the chain-rule to make use of it. With examples 2.2 and 4, we had

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, the Jacobian is the transpose of the gradient

$$\nabla_x f^T = J_f(x)$$

$$\begin{aligned} J_f(x) &= \nabla_x f^T \\ &= \begin{pmatrix} 3 & 2x_2 \end{pmatrix} \end{aligned}$$

We also need the jacobian of g, which we computed in 4

$$J_g(y) = \begin{pmatrix} 1 & 2 & 3 \\ y_2 y_3 & y_1 y_3 & y_1 y_2 \end{pmatrix}$$

Applying the chain rule, we obtain that the **jacobian** of h is the product $J_f \cdot J_g$ (**in this order**). Recall that for a function $\mathbb{R}^n \mapsto \mathbb{R}$, the jacobian is formally the transpose of the gradient. Then,

$$\begin{aligned} J_h(y) &= J_f(g(y)) \cdot J_g(y) \\ &= \nabla^T_{g(y)} f \cdot J_g(y) \\ &= \begin{pmatrix} 3 & 2y_1y_2y_3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \\ y_2y_3 & y_1y_3 & y_1y_2 \end{pmatrix} \\ &= \begin{pmatrix} 3 + 2y_1y_2^2y_3^2 & 6 + 2y_2y_1^2y_3^2 & 9 + 2y_3y_1^2y_2^2 \end{pmatrix} \end{aligned}$$

and taking the transpose we find the same gradient that we computed before!

Important remark

- The gradient is only defined for function with values in $\mathbb{R}$.
- Note that the chain rule gives us a way to compute the **Jacobian** and not the **gradient**. However, we showed that in the case of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, the **jacobian** and the **gradient** are directly identifiable, because $\nabla_x J^T = J(x)$. Thus, if we want to compute the gradient of a function by using the chain-rule, the best way to do it is to compute the Jacobian.
- As the gradient must have the same shape as the variable against which we derive, and
  - we know that the Jacobian is the transpose of the gradient
  - and the Jacobian is the dot product of Jacobians

  an efficient way of computing the gradient is to find the ordering of jacobian (or the transpose of the jacobian) that yield correct shapes!
- the notation $\frac{\partial \cdot}{\partial \cdot}$ is often ambiguous and can refer to either the gradient or the Jacobian.