Review of differential calculus theory

Winter 2017

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

$$f'(x)$$
$$\frac{df}{dx}$$
$$\frac{\partial f}{\partial x}$$
$$\nabla_x f$$

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** $(a|b) = \sum_{i=1}^{n} a_i b_i$
- **dot-product** $a^T \cdot b = (a|b) = \sum_{i=1}^{n} a_i b_i$
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 $(\cdot | \cdot)$. 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:

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

and $df_x : \mathbb{R}^n \mapsto \mathbb{R}$ is a linear form, which means that for all $x, y \in \mathbb{R}^n$, we have $df_x(x + y) = df_x(x) + df_x(y)$.

Example
Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $f\left(\begin{array}{l} x_1 \\ x_2 \end{array}\right) = 3x_1 + x_2^2$. 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

$$f\left(\begin{array}{l} a + h_1 \\ b + h_2 \end{array}\right) = 3(a + h_1) + (b + h_2)^2$$
$$= 3a + 3h_1 + b^2 + 2bh_2 + h_2^2$$
$$= 3a + b^2 + 3h_1 + 2bh_2 + h_2^2$$
$$= f(a, b) + 3h_1 + 2bh_2 + o(h)$$

Then, $df_x\left(\begin{array}{l} h_1 \\ h_2 \end{array}\right) = 3h_1 + 2bh_2$

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 $\forall h \in \mathbb{R}^n$

$$a(h) = (u_a|h)$$

In particular, for the differential $df_x$, we can find a vector $u \in \mathbb{R}^n$ such that

$$df_x(h) = (u|h)$$

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

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

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

\[
f(x + h) = f(x) + d_x f(h) + o_{h \to 0}(h) \tag{2}
\]

\[
= f(x) + \langle \nabla_x f | h \rangle + o_{h \to 0}(h) \tag{3}
\]

**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

\[
\frac{d}{dx}\begin{pmatrix} a \\ b \end{pmatrix} f\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = 3h_1 + 2bh_2
\]

We can rewrite this as

\[
\frac{d}{dx}\begin{pmatrix} a \\ b \end{pmatrix} f\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \langle \begin{pmatrix} 3 \\ 2b \end{pmatrix} | \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \rangle
\]

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 \( (e_1, \ldots, e_n) \) of \( \mathbb{R}^n \). Let’s define the partial derivative

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

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

\[
\frac{\partial f}{\partial x_1}(x) \in \mathbb{R}^n \text{ instead of } \frac{\partial f}{\partial x_1}(x)
\]
\[ \frac{\partial f}{\partial x_1} \left( \begin{array}{c} a \\ b \end{array} \right) = \lim_{h \to 0} \frac{f \left( \begin{array}{c} a + h \\ b \end{array} \right) - f \left( \begin{array}{c} a \\ b \end{array} \right)}{h} \]
\[ = \lim_{h \to 0} \frac{3(a + h) + b^2 - (3a + b^2)}{h} \]
\[ = \lim_{h \to 0} \frac{3h}{h} \]
\[ = 3 \]

In a similar way, we find that
\[ \frac{\partial f}{\partial x_2} \left( \begin{array}{c} a \\ b \end{array} \right) = 2b \]

### 2.4 Link with the partial derivatives

**Formal definition**

It can be shown that
\[ \nabla_x f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) e_i \]
\[ = \left( \begin{array}{c} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{array} \right) \]

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

**Example**

We showed that
\[ \left\{ \begin{array}{l} \frac{\partial f}{\partial x_1} \left( \begin{array}{c} a \\ b \end{array} \right) = 3 \\ \frac{\partial f}{\partial x_2} \left( \begin{array}{c} a \\ b \end{array} \right) = 2b \end{array} \right. \]

and that
\[ \nabla \left( \begin{array}{c} a \\ b \end{array} \right) f = \left( \begin{array}{c} 3 \\ 2b \end{array} \right) \]

That’s why we usually write
\[ \nabla_x f = \frac{\partial f}{\partial x}(x) \]
(same shape as \( x \))

\( e_i \) is an orthonormal basis. For instance, in the canonical basis
\[ e_i = (0, \ldots, 1, \ldots, 0) \]
with 1 at index \( i \)
\[
\n\nabla \begin{pmatrix} a \\ b \end{pmatrix} f = \begin{pmatrix} \frac{\partial f}{\partial x_1}(a) \\ \frac{\partial f}{\partial x_2}(a) \end{pmatrix}
\]

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

\[
f(x + h) = f(x) + \text{d}_xf(h) + o_{h \to 0}(h)
\]

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

\[
\begin{align*}
\text{differential} & \quad f(x) + \langle \nabla_x f| h \rangle + o_{h \to 0}(h) \\
\text{gradient} & \quad f(x) + \langle \frac{\partial f}{\partial x}(x) | h \rangle + o_{h \to 0}(h) \\
\text{partial derivatives} & \quad f(x) + \left( \begin{array}{c}
\frac{\partial f}{\partial x_1}(x) \\
\vdots \\
\frac{\partial f}{\partial x_n}(x)
\end{array} \right) | h \rangle + o_{h \to 0}(h)
\end{align*}
\]

Remark

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

\[
\text{d}_xf(\cdot) = \frac{\partial f}{\partial x}(x) \text{d}x(\cdot)
\]

because

\[
\text{d}_xf(h) = \frac{\partial f}{\partial x}(x) \text{d}x(h) = \frac{\partial f}{\partial x}(x) h
\]

In higher dimension, we write

\[
\text{d}_xf = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x) \text{d}x_i
\]

4 Jacobian: Generalization to \( f : \mathbb{R}^n \mapsto \mathbb{R}^m \)

For a function
Then, we have that

\[
\frac{\partial f_i}{\partial x}(x) = \begin{pmatrix}
\frac{\partial f_i}{\partial x_1}(x) \\
\frac{\partial f_i}{\partial x_2}(x) \\
\vdots \\
\frac{\partial f_i}{\partial x_n}(x)
\end{pmatrix}
\]

Putting all this in the same vector yields

\[
f_i(x + h) = f_i(x) + \frac{\partial f_i}{\partial x}(x) \cdot h + o(h)
\]

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

In the case where \( m = 1 \), the Jacobian is a row vector

\[
\frac{\partial f}{\partial x}(x)
\]

Remember that our gradient was defined as a column vector with the same elements. We thus have that

\[
f(x) = \nabla_x f^T
\]
Example 2 : \( g : \mathbb{R}^3 \mapsto \mathbb{R}^2 \)

Let’s define
\[
g(\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) = \begin{pmatrix} y_1 + 2y_2 + 3y_3 \\ y_1y_2y_3 \end{pmatrix}
\]

Then, the Jacobian of \( g \) is
\[
J_g(y) = \begin{pmatrix}
\frac{\partial(y_1 + 2y_2 + 3y_3)}{\partial y_1} (y) \\
\frac{\partial(y_1 + 2y_2 + 3y_3)}{\partial y_2} (y) \\
\frac{\partial(y_1 + 2y_2 + 3y_3)}{\partial y_3} (y)
\end{pmatrix}^T
\]
\[
= \begin{pmatrix}
1 & 2 & 3 \\
y_2y_3 & y_1y_3 & y_1y_2
\end{pmatrix}
\]

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

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.
6 Generalization to }\mathbb{R}^{n\times p}\mapsto \mathbb{R}^{m}\text{ }
l
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 \( \delta = f(x) \cdot h \in \mathbb{R}^{m} \) means that the \( i \)-th component of \( \delta \) is

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

7 Chain-rule

Formal definition

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

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

\[ d_x(f \circ g) = d_{g(x)} f \circ d_x g \]

Where \( \circ \) is the composition operator. Here, }d_{g(x)} f \text{ and }d_x g\text{ 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,

\[ J_h(x) = J_f(g(x)) \cdot J_g(x) \]

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

\[ \langle \nabla_x f \rangle h = \langle \nabla_X f \rangle H \]

\[ \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i}(x) h_i = \sum_{i,j} \frac{\partial f_i}{\partial X_{ij}}(x) H_{ij} \]
\[ J_h(x)_{ij} = \sum_{k=1}^{n} J_f(g(x))_{ik} \cdot J_g(x)_{kj} \]

**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 \( g : (\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) = (y_1 + 2y_2 + 3y_3) / \begin{pmatrix} y_1y_2y_3 \\ y_1y_2^2 \\ y_1^2y_3 \end{pmatrix} \).

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

\[ h(\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) = f(\begin{pmatrix} y_1 + 2y_2 + 3y_3 \\ y_1y_2y_3 \end{pmatrix}) = 3(y_1 + 2y_2 + 3y_3) + (y_1y_2y_3)^2 \]

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

\[
\begin{align*}
\frac{\partial h}{\partial y_1}(y) &= 3 + 2y_1y_2^2y_3^2 \\
\frac{\partial h}{\partial y_2}(y) &= 6 + 2y_2^3y_3^2 \\
\frac{\partial h}{\partial y_3}(y) &= 9 + 2y_3^3y_1^2y_2^2
\end{align*}
\]

And then our gradient is

\[
\nabla_y h = \begin{pmatrix}
3 + 2y_1y_2^2y_3^2 \\
6 + 2y_2^3y_3^2 \\
9 + 2y_3^3y_1^2y_2^2
\end{pmatrix}
\]

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

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

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

\[
J_g(y) = \begin{pmatrix}
1 & 2 & 3 \\
y_2y_3 & y_1y_3 & y_1y_2
\end{pmatrix}
\]
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,

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

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 \( \frac{\partial}{\partial y} \) is often ambiguous and can refer to either the gradient or the Jacobian.