1 Partial Derivatives and the Chain Rule

Vectors in $\mathbb{R}^N$ will be expressed interchangeabley as row as well as column vectors. It will be clear from the context whic way a vector is represented. For a function $f : \mathbb{R}^N \to \mathbb{R}^1$, the ith partial derivative of f at a point $a \in \mathbb{R}^N$ is defined by

$$f_{,i}(a) = \lim_{h \to 0} \frac{f(a + e_i h) - f(a)}{h},$$

where e_i is the ith unit vector in $\mathbb{R}^N$

$$e_i = \begin{array}{ccccccccc} \big(& 0, & 0, & \cdots, & 0, & 1, & 0, & \cdots, & 0 \big) \\ & & & & & \uparrow i\text{th} & & & \end{array}$$

<u>Notation</u>: Suppose $a, b \in \mathbb{R}^N$. By $[a, a+b]$ we mean the line segment joining a to $a+b$. The parametric representation of this segment is

$$[a, a+b] = \left\{ (a + tb) \in \mathbb{R}^N : 0 \le t \le 1 \right\}.$$

1.1 The mean value theorem for partial derivatives

Theorem 1 *Suppose f is continuous on $[a, a + e_i h]$ and that $f_{,i}$ exists on $(a, a + e_i h)$ then there is a $\tau \in (0,1)$ such that*

$$f(a + e_i h) - f(a) = h f_{,i}(a + e_i \tau h).$$

Proof. Define the function $\varphi : [0,1] \to \mathbb{R}^1$ by

$$\varphi(t) = f(a + e_i t h).$$

Then φ is continuous on $[0,1]$ because f is continuous and $g(t) = a + e_i t h$ is also continuous. Furthermore, φ is differentiable on $(0,1)$. To show this, let $t \in (0,1)$.

$$\begin{aligned} \varphi'(t) &= \lim_{s \to 0} \frac{\varphi(t+s) - \varphi(t)}{s} \\ &= \lim_{s \to 0} \frac{f(a + e_i (t+s) h) - f(a + e_i t h)}{s}. \end{aligned}$$

Let $b = a + e_i t h$. Then

$$\varphi'(t) = \lim_{s \to 0} \frac{f(b + e_i s h) - f(b)}{s}.$$

Let $sh = r$, then

$$\begin{aligned} \varphi'(t) &= h \lim_{s \to 0} \frac{f(b + e_i r) - f(b)}{r} \\ &= h f_{,i}(b) = h f_{,i}(a + e_i t h). \end{aligned}$$

From the Mean Value Theorem applied to the function φ we get the existence of a $\tau \in (0,1)$ such that

$$\varphi(1) - \varphi(0) = \varphi'(\tau).$$

Substituting the function f we get

$$f(a+e_ih)-f(a)=hf_{,i}(a+e_i\tau h).$$

■

<u>Notation:</u> Suppose $f:\mathbb{R}^M\to\mathbb{R}^N$. Then

$$f(x)=\begin{pmatrix} f^1(x)\\ f^2(x)\\ \vdots\\ f^N(x)\end{pmatrix}.$$

By $f'(x)$ we will denot the $N\times M$ matrix of parital derivatives

$$f'(x)=\begin{pmatrix} f^1_{,1}(x) & f^1_{,2}(x) & \cdots & f^1_{,M}(x)\\ f^2_{,1}(x) & f^2_{,2}(x) & \cdots & f^2_{,M}(x)\\ \vdots & \vdots & \cdots & \vdots\\ f^N_{,1}(x) & f^N_{,2}(x) & \cdots & f^N_{,M}(x)\end{pmatrix}.$$

So, if $f:\mathbb{R}^N\to\mathbb{R}^1$, f' is $1\times N$ matrix, i.e., a row vector.

1.2 The fundamental lemma of differentiation

We generalise the fundamental lemma of differentiation studied for functions of one variable to functions of several variables.

Theorem 2 *Suppose $f:\mathbb{R}^N\to\mathbb{R}^1$ and that f' exists and is continuous ar $a\in\mathbb{R}^N$(i.e., all components of f' are continuous functions). Then there exist N functions $\varepsilon_i:\mathbb{R}^N\to\mathbb{R}^1$, $i=1,2,\ldots,N$ all continuous at 0 with $\varepsilon_i(0)=0$ such that for all $h\in\mathbb{R}^N$,*

$$f(a+h)-f(a)=(f'(a)+\varepsilon(h))\,h, \tag{1}$$

where

$$\varepsilon(h)=(\varepsilon_1(h),\varepsilon_2(h),\cdots,\varepsilon_N(h)).$$

Proof. For $h=(h_1,h_2,\cdots,h_N)\in\mathbb{R}^N$,

$$\begin{aligned} f(a+h)-f(a) &= f\left(a+\sum_{i=1}^N e_ih_i\right)-f(a)\\ &= \sum_{k=1}^N\left(f\left(a+\sum_{i=k}^N e_ih_i\right)-f\left(a+\sum_{i=k+1}^N e_ih_i\right)\right).\end{aligned}$$

Let $b_k=a+\sum_{i=k+1}^N e_ih_i$, we can rewrite the above equation as

$$f(a+h)-f(a)=\sum_{k=1}^N f(b_k+e_kh_k)-f(b_k).$$

Using the Mean Value Theorem for partial derivatives, we get

$$f(a+h)-f(a)=\sum_{k=1}^{N} h_k f_{,k}\left(b_k+e_k\tau_k h_k\right),$$

where $\tau_k \in (0,1), k=1,2,\ldots,N$. Therefore,

$$f(a+h)-f(a)=\sum_{k=1}^{N} h_k\left(f_{,k}(a)+\varepsilon_k(h)\right),$$

where

$$\varepsilon_k(h)=f_{,k}\left(b_k+e_k\tau_k h_k\right)-f_{,k}(a).$$

Clearly $b_k=a+\sum_{i=k+1}^{N} e_i h_i \to a$ as $h\to 0$. From this and the continuity of $f_{,k}$ we get the continuity of $\varepsilon_k(h)$ and that $\varepsilon_k(0)=0$. We can rewrite the above equation in matrix form as

$$f(a+h)-f(a)=\left(f'(a)+\varepsilon(h)\right)h.$$

■

Observe that the functions $\varepsilon_k, k=1,2,\ldots,N$ defined in the proof of the above theorem depend also on the point a under consideration.

Corollary 3 *Suppose $f:\mathbb{R}^N\to\mathbb{R}^M$ and that (the $M\times N$ matrix) f' exists and is continuous at $a\in\mathbb{R}^N$, then there exist $M\times N$ functions $\varepsilon_{ij}:\mathbb{R}^N\to\mathbb{R}^M$ all continuous at 0 with $\varepsilon_{ij}(0)=0$ such that for all $h\in\mathbb{R}^N$,*

$$f(a+h)-f(a)=\left(f'(a)+\varepsilon(h)\right)h,$$

where $\varepsilon(h)$ is the $M\times N$ matrix $\varepsilon(h)=\left(\varepsilon_{ij}(h)\right)$.

Proof. Apply the Fundamental Lemma of Differentiation to each component of the function $f(x)=\begin{pmatrix} f^1(x)\\ f^2(x)\\ \vdots\\ f^M(x)\end{pmatrix}$ and write the result in matrix form. The details are left as an excercise.

■

Corollary 4 *Suppose $f:\mathbb{R}^N\to\mathbb{R}^1$ and that f' exists and is continuous at $a\in\mathbb{R}^N$ then f is continuous at a.*

Proof. Taking the limit on both sides of (1) as $h\to 0$ and using the continuity of $\varepsilon(h)$ at 0 and $\varepsilon(0)=0$ we get

$$\lim_{h\to 0} f(a+h)-f(a)=0.$$

Thus

$$\lim_{h\to 0} f(a+h)=f(a),$$

which means that f is continuous at a. ■

1.3 The chain rule

The chain rule for functions of several variables has the following form.

Theorem 5 *Suppose $g : \mathbb{R}^M \to \mathbb{R}^N$ and $f : \mathbb{R}^N \to \mathbb{R}^1$. Suppose also that (the $N \times M$ matrix) g' exists and is continuous at $a \in \mathbb{R}^M$ and that (the $1 \times N$ matrix) f' exists and is continuous at $b = g(a) \in \mathbb{R}^N$. Then (the $1 \times M$ matrix) $(f \circ g)'(a)$exists and*

$$\begin{aligned}(f \circ g)'(a) &= f' \circ g(a)\, g'(a) \\ &= f'(g(a))\, g'(a) \\ &= f'(b)\, g'(a).\end{aligned}$$

Proof. For $h \in \mathbb{R}^M$ we have

$$f \circ g(a+h) - f \circ g(a) = f(g(a+h)) - f(g(a)).$$

Let $k = g(a+h) - g(a) = g(a+h) - b$, then

$$f \circ g(a+h) - f \circ g(a) = f(b+k) - f(b).$$

Applying the Fundamental Lemma of Differentiation twice, we get

$$\begin{aligned}f \circ g(a+h) - f \circ g(a) &= (f'(b) + \varepsilon(k))\, k \\ &= (f'(g(a)) + \varepsilon(k))\,(g(a+h) - g(a)) \\ &= (f'(g(a)) + \varepsilon(k))\,(g'(a) + \varepsilon_1(h))\, h.\end{aligned}$$

In particular, replacing h by $e_i h$ gives

$$\begin{aligned}f \circ g(a + e_i h) - f \circ g(a) &= (f'(g(a)) + \varepsilon(k))\,(g'(a) + \varepsilon_1(e_i h))\, e_i h \\ &= h\,(f'(g(a)) + \varepsilon(k))\,(g'(a) + \varepsilon_1(e_i h))\, e_i\end{aligned}$$

(because h is now a scalar). Therefore,

$$\begin{aligned}(f \circ g)_{,i}(a) &= \lim_{h \to 0} \frac{f \circ g(a + e_i h) - f \circ g(a)}{h} \\ &= \lim_{h \to 0} (f'(g(a)) + \varepsilon(k))\,(g'(a) + \varepsilon_1(e_i h))\, e_i \\ &= f'(g(a))\, g'(a)\, e_i.\end{aligned}$$

Hence,

$$\begin{aligned}(f \circ g)'(a) &= \Big((f \circ g)_{,1}(a), (f \circ g)_{,2}(a), \cdots, (f \circ g)_{,M}(a)\Big) \\ &= (f'(g(a))\, g'(a)\, e_1, f'(g(a))\, g'(a)\, e_2, \cdots, f'(g(a))\, g'(a)\, e_M) \\ &= f'(g(a))\, g'(a)\, (e_1, e_2, \cdots, e_M) \\ &= f'(g(a))\, g'(a)\, I_{M \times M} \\ &= f'(g(a))\, g'(a).\end{aligned}$$

■

1.4 Examples

We give here some example on how to compute partial derivatives efficiently.

1. Let A be an $M \times N$ matrix and define $g : \mathbb{R}^N \to \mathbb{R}^M$ by $g(x) = Ax$. To compute $g'(x)$ we proceed as follows:

$$\begin{aligned} g(x+h) - g(x) &= A(x+h) - Ax \\ &= Ah. \end{aligned}$$

   In particular, if we replace h by $e_i h$ we get

$$g(x + e_i h) - g(x) = hAe_i.$$

   Thus

$$g_{,i}(x) = \lim_{h \to 0} \frac{g(x + e_i h) - g(x)}{h} = \lim_{h \to 0} Ae_i = Ae_i.$$

   Therefore,

$$\begin{aligned} g'(x) &= (Ae_1, Ae_2, \cdots, Ae_N) \\ &= A(e_1, e_2, \cdots, e_N) \\ &= AI_{N \times N} = A. \end{aligned}$$

   A special case of this example is when A is an $M \times 1$ matrix (a column vector). In this case $g : \mathbb{R}^1 \to \mathbb{R}^M$ is given by $g(x) = Ax = xA$ because now x is a scalar. Here also $g'(x) = A$.

2. Define the function $g : \mathbb{R}^N \to \mathbb{R}^1$ by $g(x) = x \cdot x = x^T x = \|x\|^2$. Following the same steps as above, we get

$$\begin{aligned} g(x+h) - g(x) &= (x+h)^T (x+h) - x^T x \\ &= 2x^T h + h^T h \\ g(x + e_i h) - g(x) &= 2hx^T e_i + h^2 e_i^T e_i \\ &= 2hx^T e_i + h^2 \\ g_{,i}(x) &= 2x^T e_i \\ g'(x) &= 2x^T (e_1, e_2, \cdots, e_N) \\ &= 2x^T. \end{aligned}$$

   As generalization of this example we define $g : \mathbb{R}^N \to \mathbb{R}^1$ by $g(x) = Ax \cdot Bx = (Ax)^T Bx$

where A and B are $M \times N$ matrices.

$$\begin{aligned}
g(x+h)-g(x) &= (A(x+h))^T B(x+h)-(Ax)^T Bx \\
&= (Ah)^T B(x+h)+(Ax)^T Bh \\
&= (B(x+h))^T Ah+(Ax)^T Bh \\
&= \left((B(x+h))^T A+(Ax)^T B\right) h \\
g(x+e_i h)-g(x) &= h\left((B(x+e_i h))^T A+(Ax)^T B\right) e_i \\
g_{,i}(x) &= \left((Bx)^T A+(Ax)^T B\right) e_i \\
g'(x) &= \left((Ax)^T B+(Bx)^T A\right)(e_1, e_2, \cdots, e_N) \\
&= (Ax)^T B+(Bx)^T A. \\
&= x^T\left(A^T B+B^T A\right).
\end{aligned}$$

The first part of this example can be obtained from this general case by taking $M=N$ and $A=B=I$.

3. A "product rule" for functions of several variables may be obtained as follows. Let $f, g: \mathbb{R}^M \to \mathbb{R}^N$ and define $h: \mathbb{R}^N \to \mathbb{R}^1$ by

$$\begin{aligned}
h(x) &= f(x) \cdot g(x)=f(x)^T g(x). \\
h(x+r)-h(x) &= f(x+r)^T g(x+r)-f(x)^T g(x) \\
&= f(x+r)^T g(x+r)-f(x+r)^T g(x)+f(x+r)^T g(x)-f(x)^T g(x). \\
&= f(x+r)^T(g(x+r)-g(x))+\left(f(x+r)^T-f(x)^T\right) g(x). \\
h(x+e_i r)-h(x) &= f(x+e_i r)^T(g(x+e_i r)-g(x))+\left(f(x+e_i r)^T-f(x)^T\right) g(x) \\
h_{,i}(x) &= f(x)^T g_{,i}(x)+f_{,i}(x) g(x). \\
h'(x) &= f(x)^T g'(x)+f'(x) g(x).
\end{aligned}$$

4. Let $f: \mathbb{R}^M \to \mathbb{R}^N$ and define $g: \mathbb{R}^N \to \mathbb{R}^1$ by $g(x)=\|f(x)\|^2$. We apply the chain rule to find g' and get

$$g'(x)=2 f(x)^T f'(x).$$

As an exercise, one can verfiy the above formula directly.

5. Let $f: \mathbb{R}^M \to \mathbb{R}^N$ be (positive) homogeneous of degree n, i.e., for any (positive) scalar k, $f(kx)=k^n f(x)$. Also let $g: \mathbb{R}^1 \to \mathbb{R}^M$ be defined by $g(x)=Ax$, where A is an $M \times 1$. matrix. Then $f \circ g: \mathbb{R}^1 \to \mathbb{R}^N$ is given by $f \circ g(x)=f(Ax)=x^n f(A)$. From the chain rule,

$$\begin{aligned}
(f \circ g)'(x) &= f'(g(x)) g'(x) \\
&= f'(Ax) A \text{ (see example 1)}.
\end{aligned}$$

On the other hand, using the explicit definition of $f \circ g(x) = x^n f(A)$, we get that $(f \circ g)'(x) = nx^{n-1} f(A)$. Thus we have the equation

$$nx^{n-1} f(A) = f'(Ax) A.$$

In particular, for $x = 1$, we have

$$nf(A) = f'(A) A,$$

which is Euler's formula for homogeneous functions.