Chapter 9

Differential Calculus in Euclidean Space

We have already seen that differential calculus provides many useful tools in the analysis of local properties of real, real-valued functions. The basic idea is that of locally approximating any smooth function by an easier one (an affine function or, more in general by a polynomial) and deduce properties it has from properties of this approximation. The concept of derivative and differentiability play a central role.

9.1 The differential

Motivation. Here we would like to consider general functions $f : D \subset \mathbb{R}^n \to \mathbb{R}^m$ of n variables taking vector values in $\mathbb{R}^m$ for any two $m, n \in \mathbb{N}$. First we need an appropriate concept of derivative. In the real real-valued case we have that the derivative could be thought of as the slope of the tangent line to the graph of the function at the point of interest whenever it is at all possible to obtain a sensible affine approximation. This can be done in this context, too. The general form of an affine function is in this case

$$x \mapsto Ax + b\,,\ D \to \mathbb{R}^m$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. If it is possible to "well" approximate a given function by such an affine map about a given point, we shall say that the function is differentiable there. The approximation has to be good enough in the sense that

$$|f(x+h) - b - Ah|_2 = o(|h|_2) \text{ as } h \to 0$$

Observe that we have to measure size with the respective norms defined on $\mathbb{R}^n$ and $\mathbb{R}^m$. It is easy to see that $b = f(x)$ if f is continuous, which we shall assume. So the problem boils down to whether we are able to find A such that the approximation is better than first order as stated above. This will be the *total* point of view in that we try to understand the function f all at once.
We shall also take a more *partial* view and, upon choosing a curve

$$\gamma : (-1, 1) \to D\,,$$

we can study the behavior of f along that curve, i.e., the behavior of $f \circ \gamma$. The latter has the advantage of being a real function which we know how to deal with better. If $0 \in D$, we could for instance look at

$$t \mapsto f(th, 0, \cdots, 0)\,,\ (-1, 1) \to \mathbb{R}^m$$

by choosing the appropriate path. In this case we would be fixing all variables but one.

Definition 9.1.1. (Differentiability)
Let $f : U \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m$ be given. It is called *differentiable* at $x \in U$ iff

$$\exists\, A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m\times n} \text{ s.t. } f(x + h) = f(x) + Ah + o(|h|_2) \text{ as } h \to 0$$

which simply means

$$\forall\, \varepsilon > 0\, \exists\, \delta > 0 \text{ s.t. } |f(x + h) - f(x) - Ah| \leq \varepsilon |h|_2\, \forall\, h \text{ with } |h|_2 \leq \delta\,.$$

If that is the case, then the linear map A is denoted by $Df(x)$, the *differential*, or simply the derivative, of f at x.

Remarks 9.1.2. (a) f is differentiable at $x \iff f_1, \ldots, f_m$ are differentiable at x. What is the relation between $Df(x)$ and $Df_j(x)\,,\ j = 1, \ldots, m$?
(b) If f is differentiable at x, then it is Lipschitz continuous at x.

Definition 9.1.3. Let $f : D \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m$ be given and assume it is differentiable at each point $x \in D$ in its domain. The function f is called differentiable on D and

$$Df : D \to \mathbb{R}^{m\times n}\,,\ x \mapsto Df(x)$$

is called *derivative* of f. We say f is *continuously differentiable* if Df is continuous. The collection of all continuously differentiable functions defined on D with values in $\mathbb{R}^m$ is denoted by

$$\mathrm{C}^1(D, \mathbb{R}^m) = \left\{ f \in \mathrm{C}(D, \mathbb{R}^m) \,\middle|\, Df \in \mathrm{C}(D, \mathbb{R}^{m\times n}) \right\}$$

in accordance with our previously introduced notation.

Remarks 9.1.4. (a) The operator

$$D : \mathrm{C}^1(D, \mathbb{R}^m) \to \mathrm{C}(D, \mathbb{R}^{m\times n})\,,\ f \mapsto Df$$

is linear.
(b) If $g \in \mathrm{C}^1(D, \mathbb{R})$ and $f \in \mathrm{C}^1(D, \mathbb{R}^m)$, then $fg \in \mathrm{C}^1(D, \mathbb{R}^m)$.

9.2 Partial Derivatives

How do we compute the entries of $Df(x)$? Well, the function

$$f : D \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m$$

has m-components $f_1, \ldots, f_m$. Then, letting $h = e_j$ for $j = 1, \ldots, n$, we have by definition (if the f is differentiable at $x \in D$) that

$$f(x + te_j) = f(x) + tDf(x)e_j + o(|te_j|)\ (t \to 0)$$

and therefore

$$f_k(x + te_j) = f_k(x) + t\big(Df(x)e_j\big)_k + o(|t|)\ (t \to 0)\,,\ k = 1, \ldots, m\,,$$

Definition 9.2.1. (Partial and Directional Derivative)
Since $\big(Df(x)e_j\big)_k = Df(x)_{jk}$ we therefore have that

$$Df(x)_{jk} = \lim_{t\to 0} \frac{f_k(x + te_j) - f_k(x)}{t} = \frac{\partial f_k}{\partial_{x_j}}(x) = \partial_j f_k(x)$$

which is called *j-th partial derivative of f_k at x* and where $j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$. More in general, we can define the derivative of f at x in any direction $u \in \mathbb{R}^n$ by

$$\lim_{t\to 0} \frac{f(x + tu) - f(x)}{t} = \partial_u f(x)$$

called *directional derivative of f at x in direction u.*

Directional derivatives have the advantage that they can be computed just like for real functions. They, however, contain only partial information about the function f, specifically only about its local behavior along a certain line in direction u emanating from the point x.

Theorem 9.2.2. *Let $f : D \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m$ be given such that $Df(x)$ exists for $x \in D$ and let $u \in \mathbb{R}^n$. Then $\partial_u f(x)$ exists and*

$$\partial_u f(x) = Df(x)u\,.$$

The proof is left as an exercise.

Remark 9.2.3. The existence of all partial derivatives $\partial_j f(x)$ at a point $x \in D$, even of all directional derivatives, does not make f differentiable at x. Take the function defined by

$$f\big(r\cos(\theta), r\sin(\theta)\big) := \begin{cases} rg(\theta)\,, & (r,\theta) \in (0,\infty)\times[0,2\pi)\,, \\ 0\,, & r = 0 \end{cases}$$

where g is any function satisfying $g(-\theta) = -g(\theta)$. Why do we need g to be odd? Then it is easily checked that $\partial_u f(0,0) = g(\theta)$ if $u = \big(\cos(\theta), \sin(\theta)\big)$, but f is not even differentiable in the origin if g is not differentiable with respect to θ.

Theorem 9.2.4. *Let $f : D \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m$ and $x \in D$. Assume that*

$$\partial_j f \in \mathrm{C}(U_x, \mathbb{R}^m) \text{ for } j = 1, \ldots, n$$

and some neighborhood $U_x \in \mathcal{U}(x)$. Then f is differentiable at x. Morevover

$$f \in \mathrm{C}^1(D, \mathbb{R}^m) \iff \partial_j f \in \mathrm{C}(D, \mathbb{R}^m)\ \forall\, j \in \{1, \ldots, n\}\,.$$

In particular it follows that

$$\mathrm{C}^1(D, \mathbb{R}^m) = \big\{ f \in \mathrm{C}(D, \mathbb{R}^m) \,\big|\, \partial_j f \in \mathrm{C}(D, \mathbb{R}^m)\ \forall\, j = 1, \ldots, n \big\}\,.$$

Proof. We give the proof only for the case $m = 1$, $n = 2$ since, in the general case, it is perfectly analogous. We need to prove that

$$f(y_1, y_2) - f(x_1, x_2) = \partial_1 f(x_1, x_2)(y_1 - x_1) + \partial_2 f(x_1, x_2)(y_2 - x_2) + o(|(x_1, x_2) - (y_1, y_2)|_2)\,.$$

By using the mean value theorem 5.2.4 on the two differences in the right hand side of

$$f(y_1, y_2) - f(x_1, x_2) = f(y_1, y_2) - f(x_1, y_2) + f(x_1, y_2) - f(x_1, y_1)$$

we find $\xi \in I(x_1, y_1)$ and $\xi_2 \in I(x_2, y_2)$ such that

$$f(y_1, y_2) - f(x_1, x_2) = \partial_1 f(\xi_1, y_2)(y_1 - x_1) + \partial_2 f(x_1, \xi_2)(y_2 - x_2)\,.$$

which is not quite what we need. But, if the error

$$\begin{aligned} \mathrm{E} = \partial_1 f(x_1, x_2)(y_1 - x_1) + \partial_2(x_1, x_2)(y_2 - x_2) + \\ - \partial_1 f(\xi_1, y_2)(y_1 - x_1) - \partial_2 f(x_1, \xi_2)(y_2 - x_2) \end{aligned}$$

incurred, can be proven to be a $o(|(x_1, x_2) - (y_1, y_2)|_2)$, the claim follows. Since

$$\frac{|\mathrm{E}|}{|x - y|} \leq |\partial_1 f(x_1, x_2) - \partial_1 f(\xi_1, y_2)| + |\partial_2 f(x_1, \xi_2) - \partial_2(y_1, y_2)|$$

and since both terms on the right hand side can be made arbitrarily small by making $\sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}$ small (by continuity of the partial derivatives), this is indeed the case and the proof is complete. Why do we say that the proof in general is similar? Can you perform the proof in the general case? $\surd$

9.3 The Chain Rule

Theorem 9.3.1. *(Chain Rule)*
Let $f : D_f \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m$ and $g : D_g \overset{o}{\subset} \mathbb{R}^m \to \mathbb{R}^p$ be such that $f(D_f) \subset D_g$. If f is differentiable at $x \in D$ and g is differentiable at $f(x)$, then $g \circ f$ is differentiable at x and

$$D\big(f \circ g\big)(x) = Dg\big(f(x)\big)Df(x)$$

Moreover, if f and g are continuously differentiable on their respective domain, so is $g \circ f$ on D_f.

Proof. We know that

$$f(y) = f(x) + Df(x)(y - x) + R_1(x, y)$$

and

$$\begin{aligned} g\big(f(y)\big) &= g\big(f(x)\big) + Dg\big(f(x)\big)\big(f(y) - f(x)\big) + R_2\big(f(x), f(y)\big) \\ &= g\big(f(x)\big) + Dg\big(f(x)\big)\big[Df(x)(y - x) + R_1(x, y)\big] + R_2\big(f(x), f(y)\big) \end{aligned}$$

Now observe that

$$|Dg\big(f(x)\big)R_1(x,y)| \leq c|R_1(x,y)| = o(|x-y|_2) \text{ as } y \to x$$

and that

$$|R_2\big(f(x),f(y)\big)| \leq \varepsilon|f(x)-f(y)|$$

for some $\delta > 0$ and whenever $|f(x)-f(y)| \leq \delta$ since $R_2(w,z) = o(|w-z|_2)$. Finally f is (locally) Lipschitz continuous at x, since f is differentiable there which gives

$$|f(x)-f(y)| \leq c|x-y|_2$$

for $|x-y|_2 \leq \delta$ and some (other) $\delta > 0$ which, then, gives

$$|R_2\big(f(x),f(y)\big)| \leq c\varepsilon|x-y|_2 \text{ for } |x-y|_2 \leq \delta$$

and the claim follows. $\surd$

Remark 9.3.2. It follows that

$$\partial_j(g\circ f)(x) = D(g\circ f)(x)e_j = Dg\big(f(x)\big)Df(x)e_j = \sum_{k=1}^{m} \partial_k g\big(f(x)\big)\partial_j f_k(x)\,.$$

Theorem 9.3.3. *Let $f : D_f \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}$ and $x \in D$; if f assumes a maximum or minimum at x and f is differentiable there, then $Df(x) = 0$.*

Proof. Since $x \in D \overset{o}{\subset} \mathbb{R}^n$, there is $t_0 > 0$ such that

$$x + te_j \in D \,\forall\, j = 1,\ldots,n \,\forall\, |t| \leq t_0\,.$$

Since f assumes a minimum or a maximum at x, so do the maps

$$g_j : (-t_0,t_0) \to \mathbb{R}\,,\; t \mapsto f(x+te_j)\;\forall j = 1,\ldots,n$$

at $t = 0$. It then follows from theorem 5.2.2(iii) that

$$0 = \frac{d}{dt}g_j(0) = \partial_j f(x)\;\forall j = 1,\ldots,n$$

which clearly gives $Df(x) = 0$. $\surd$

9.4 Higher Derivatives

Just as in the case of real functions we now move on to higher order derivatives which, when they exist, give us some information about the convexity properties of functions. The latter make it possible to tell maxima, minima and saddle points apart. Taylor's expansion formula will also be generalized.

9.4.1 Mixed Partial Derivatives

First we would like to take a closer look at second derivatives. In particular we need to make sure that we understand what kind of objects they are. On the one hand we could simply say that the are the first derivative of the first derivative. This, albeit correct, might, however, conceal their mapping properties. Let us therefore start with a smooth function

$$f : D_f \overset{o}{\subset} \mathbb{R}^n \to \mathbb{R}^m \,.$$

Its derivative is the map

$$Df : D \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) \hat{=} \mathbb{R}^{m\times n} \,,\; x \mapsto Df(x) \,.$$

Taking a further derivative is going to give us a map

$$D^2 f = D\big(Df\big) : D \to \mathcal{L}\big(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)\big) \hat{=} \mathbb{R}^{m\times n\times n}$$

Since a vector valued map can always be considered component by component and for the sake of simplicity we only consider the case $m = 1$. Then the derivative at a point

$$\mathcal{L}(\mathbb{R}^n, \mathbb{R}) \ni Df(x) = [\partial_1 f(x)\, \partial_2 f(x)\, \ldots\, \partial_n f(x)]$$

is a row vector. It is sometimes useful to think of it as a column vector, in which cases we call it the *gradient of* f and denote it by $\nabla f(x)$. If we now take a further derivative of the gradient we obtain the so-called *Hessian* of the function f

$$D(\nabla f) = \begin{bmatrix} D(\partial_1 f) \\ D(\partial_2 f) \\ \vdots \\ D(\partial_n f) \end{bmatrix} = \begin{bmatrix} \partial_1(\partial_1 f)\partial_2(\partial_1 f)\cdots\partial_n(\partial_1 f) \\ \vdots \\ \partial_1(\partial_n f)\partial_2(\partial_n f)\ldots\partial_n(\partial_n f) \end{bmatrix} = \big[\partial_j\partial_k f\big]_{1\le j,k\le n}$$

We also define the function space

$$\begin{aligned} \mathrm{C}^2(D,\mathbb{R}) &:= \big\{ f \in \mathrm{C}^1(D,\mathbb{R}) \,\big|\, \partial_j f \in \mathrm{C}^1(D,\mathbb{R}) \big\} \\ &= \big\{ f \in \mathrm{C}(D,\mathbb{R}) \,\big|\, \partial_j\partial_k f \in \mathrm{C}(D,\mathbb{R}) \big\} \end{aligned}$$

of twice continuously differentiable functions.

Theorem 9.4.1. *(Mixed Derivatives)*
If $f \in \mathrm{C}^2(D,\mathbb{R})$ *then*

$$\partial_j\partial_k f = \partial_k\partial_j f \;\forall\, j,k \in \{1,\ldots,n\} \,.$$

Proof. For $h \in \mathbb{R}^n$ we introduce the notation $\triangle_h f(x) = f(x+h) - f(x)$ whenever it makes sense. It follows that

$$\partial_j f(x) = \lim_{s \to 0} \frac{1}{s} \triangle_{se_j} f(x)$$

and it is easy to check that

$$\triangle_{se_j} \triangle_{te_k} = \triangle_{te_k} \triangle_{se_j} \; \forall \, j,k = 1, \ldots, n \, .$$

Observe that theorem 5.2.4 implies the existence of $\bar{s} \in I(0,s)$ such that

$$\frac{1}{s} \triangle_{se_j} f(x) = (\partial_j f)(x + \bar{s} e_j)$$

and, consequently

$$\frac{1}{st} \triangle_{se_j} \big(\triangle_{te_k} f(x) \big) = \frac{1}{s} \triangle_{se_j} \big((\partial_k f)(x + \bar{t} e_k) \big) = (\partial_j \partial_k f)(x + \bar{s} e_j + \bar{t} e_k)$$

or

$$\frac{1}{st} \triangle_{te_k} \big(\triangle_{se_j} f(x) \big) = \frac{1}{t} \triangle_{te_k} \big((\partial_j f)(x + \tilde{s} e_j) \big) = (\partial_k \partial_j f)(x + \tilde{s} e_j + \tilde{t} e_k)$$

Since the two representations have to coincide and taking the limits $t, s \to 0$ in whichever order (since the second partial are continuous) we obtain

$$(\partial_j \partial_k f)(x) = (\partial_k \partial_j f)(x)$$

which gives the claim since $x \in D$ was arbitrary. $\surd$

9.4.2 Local Extrema

Definition 9.4.2. (Critical Point)
Let $f \in \mathrm{C}^1(D, \mathbb{R})$ for some $D \overset{o}{\subset} \mathbb{R}^n$. A point $x \in D$ is called *critical point* if $\nabla f(x) = 0$.

Motivation. Since vanishing of the gradient is a necessary condition for a local extremum, we would like to find criteria that would allow us to decide it is a point of minimum, maximum or else. Assume that $f \in \mathrm{C}^2(D, \mathbb{R})$ and denote (abusing the notation) its Hessian by $D^2 f$. By the previous theorem it is symmetric. Let now $x \in D$ be a point of minimum for f. Then so is $t = 0$ for

$$g_u : (-t_0, t_0) \to \mathbb{R} \, , \; t \mapsto f(x + tu) \, .$$

Observe that $g'_u(0) = \big(\nabla f(x)|u\big) = 0$ and that

$$g''_u(t) = \frac{d}{dt}\sum_{j=1}^{n}\partial_j f(x+tu)u_j = \sum_{k=1}^{n}\sum_{j=1}^{n}\partial_k\partial_j f(x+tu)u_k u_j\,.$$

Thus

$$0 \le g''_u(0) = \sum_{k=1}^{n}\sum_{j=1}^{n}\partial_k\partial_j f(x)u_k u_j = u^T D^2 f(x) u$$

and this is valid for any nonzero direction $u \in \mathbb{R}^n$. We also know that if $g'_u(0) = 0$ and $g''_u(0) > 0$, then g_u has a point of minimum at $t = 0$. It is therefore legitimate to hope that if

$$0 < u^T D^2 f(x) u \,\forall\, 0 \neq u \in \mathbb{R}^n$$

we would have that x is a minimum of f.

Definition 9.4.3. Every symmetric matrix $A = A^T \in \mathbb{R}^{n\times n}$ defines a *quadratic form*

$$x \mapsto x^T A x\,,\ \mathbb{R}^n \to \mathbb{R}\,.$$

It is called *positive definite* iff $0 < x^T A x \,\forall\, x \neq 0$ and *positive semi-definite ($A > 0$)* or *nonnegative definite ($A \ge 0$)* if the non strict inequality holds. Can you use your linear algebra knowledge to reformulate the defining condition?

Lemma 9.4.4. *Let $A = A^T, B = B^T \in \mathbb{R}^{n\times n}$. Then*
(i) $A > 0 \iff \exists\, \varepsilon > 0$ s.t. $x^T A x \ge \varepsilon|x|^2 \,\forall\, x \in \mathbb{R}^n$.
(ii) If $A > 0$ and $B - A$ is small, then $B > 0$. This say that the set of positive definite symmetric matrices is open in the set of symmetric matrices.

Proof. (i) By observing that

$$(tx)^T A(tx) = t^2 x^T A x \,\forall\, t > 0$$

we conclude that

$$0 < x^T A x \,\forall\, x \in \mathbb{R}^n \setminus \{0\} \iff 0 < x^T A x \,\forall\, |x| = 1\,.$$

Since the mapping $x \mapsto x^T A x\,,\ \mathcal{S}^{n-1} \to (0,\infty)$ is continuous and $\mathcal{S}^{n-1}$ is compact we see that

$$\varepsilon =: \min_{x\in\mathcal{S}^{n-1}} x^T A x > 0$$

and finally

$$\varepsilon|x|^2 \leq x^T Ax \text{ since } \varepsilon \leq \frac{x^T}{|x|} A \frac{x}{|x|} \,\forall\, x \in \mathbb{R}^n \setminus \{0\}\,.$$

When $x = 0$ the inequality is trivially satisfied.
(ii) If $B - A$ is sufficiently small, that is if $x^T(B-A)x \leq \frac{\varepsilon}{2}|x|^2$, for instance, then

$$x^T Bx = x^T Ax + x^T(B-A)x \geq \varepsilon|x|^2 - \frac{\varepsilon}{2}|x|^2 \geq \frac{\varepsilon}{2}|x|^2$$

and the claim follows. $\surd$

Theorem 9.4.5. *Let $f \in \mathrm{C}^2(D,\mathbb{R})$ for some $D \overset{o}{\subset} \mathbb{R}^n$ and let $x \in D$ be a critical point. Then*
(i) If the point x is a local point of minimum [maximum], then we have that $D^2 f(x) \geq 0$ [≤ 0].
(ii) If $D^2 f(x) > 0$ [< 0], then x is a local strict minimum [maximum].

Proof. (i) The first claim is a direct consequence of the calculations we performed in the motivation.
(ii) Reasoning purely line by line (through x) we would get

$$\frac{d^2}{dt^2} f(x+tu)\big|_{t=0} > 0$$

and, consequently that

$$f(x+tu) > f(x) \,\forall\, |t| \leq t_0(u)$$

for some $t_0(u) > 0$ which depends on u. This is not enough to claim, as we need to, that

$$f(y) > f(x) \,\forall\, y \in \mathrm{B}(x,\varepsilon) \text{ for some } \varepsilon > 0\,.$$

Let therefore $u \in \mathbb{R}^n$ with $|u| = 1$ and define $g_u(t) := f(x+tu)$ for $|t| \leq \varepsilon$ where $\varepsilon > 0$ is chosen so small that $\mathrm{B}(x,\varepsilon) \in D$. Then

$$g_u'(0) = 0\,,\; g_u''(t) = u^T D^2 f(x+tu)u > 0 \,\forall\, |t| \leq \varepsilon$$

for a possibly smaller $\varepsilon > 0$ since $D^2 f$ is continuous at x and lemma 9.4.4. The point x is therefore a strict minimum in $[|t| \leq \varepsilon]$ which concludes the proof (why?). $\surd$

Remark 9.4.6. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if and only if its eigenvalues are positive, that is, if $\lambda_j > 0$ for all $j \in \{1, \dots, n\}$. Recall that a symmetric matrix is always diagonalizable.

Lemma 9.4.7. *Let $0 \neq x \in \mathbb{R}^n$ be a critical point of the map*

$$x \mapsto \frac{x^T A x}{x^T x}\,,\ \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$$

then it is an eigenvector of A.

9.4.3 Taylor Expansion

Let $f \in \mathrm{C}^m(D, \mathbb{R})$ for some $D \overset{o}{\subset} \mathbb{R}^n$. Can we do better than linear approximation? In other words, can we generalize Taylor expansions to multivariable case?

Lemma 9.4.8. *Pick $u \in \mathbb{R}^n$ and $t_0 >$ such that $\mathrm{B}(x, t_0) \overset{o}{\subset} D$ and define $g(t) := f(x + tu)$ for $|t| \leq t_0$. Then, for $k \leq m$,*

$$\frac{1}{k!}\frac{d^k}{dt^k} g(t) = \frac{1}{k!} g^{(k)}(t) = \sum_{|\alpha|=k} \frac{u^\alpha}{\alpha!} \partial_\alpha f(x + tu)$$

where $\alpha! = \alpha_1! \cdots \alpha_n!$ and $0! = 1$.

Proof. The proof is done by induction. We shall, however, look at the cases $k = 1, 2$ to get a better feeling. As for the first

$$g'(t) = \sum_{j=1}^{n} u_j \partial_j f(x + tu)$$

and therefore, for the second

$$g''(t) = \sum_{l=1}^{n}\sum_{j=1}^{n} u_l u_j \partial_l \partial_j f(x + tu) = 2 \sum_{|\alpha|=2} \frac{u^\alpha}{\alpha!} \partial_\alpha f(x + tu)\,.$$

Now, by induction hypothesis, we have that

$$\frac{1}{(k-1)!} g^{(k-1)}(t) = \sum_{|\alpha|=k-1} \frac{u^\alpha}{\alpha!} \partial_\alpha f(x + tu)$$

and, consequently,

$$\frac{1}{k(k-1)!}g^{(k)}(t) = \sum_{|\alpha|=k-1}\sum_{j=1}^{n}\frac{u_j u^\alpha}{k\alpha!}\partial_j\partial_\alpha f(x+tu)$$
$$= \sum_{|\beta|=k}\frac{1}{k}\sum_{j=1}^{n}\frac{\beta_j}{\beta!}u^\beta\partial_\beta f(x+tu)$$

and the claim follows since $|\beta| = k$ implies that $\frac{1}{k}\sum_{j=1}^{n}\beta_j = 1$. $\surd$

By using Taylor expansion of the real function g

$$g(t) = \sum_{k=0}^{m}\sum_{|\alpha|=k}\frac{u^\alpha}{\alpha!}\partial_\alpha f(x)t^k + o(t^m)$$

we will obtain the more general Taylor expansion f

$$f(y) = \sum_{|\alpha|\le m}\frac{(y-x)^\alpha}{\alpha!}\partial_\alpha f(x) + o(|y-x|)$$

by setting

$$u = \frac{y-x}{|y-x|} \text{ and } t = |y-x|\,.$$

This is precisely the idea behind the proof of the next theorem.

Theorem 9.4.9. *(Taylor Expansion)*
Let $f \in \mathrm{C}^m(D,\mathbb{R})$ for some $D \overset{o}{\subset} \mathbb{R}^n$ and $x \in D$. Define the Taylor polynomial $T_m f(x,y)$ of order m of f at x by

$$T_m f(x,y) := \sum_{|\alpha|\le m}\frac{1}{\alpha!}\partial_\alpha f(x)(y-x)^\alpha\,.$$

Then

$$f(y) = T_m f(x,y) + o(|y-x|^m) \text{ as } y \to x\,.$$

Proof. We need to show that

$$\forall\,\varepsilon > 0\;\exists\,\delta > 0 \text{ s.t. } |f(y) - T_m f(x,y)| \le \varepsilon|y-x|^m \text{ if } |y-x| \le \delta\,.$$

If $h(y) := f(y) - T_m f(x,y)$, then $\partial_\beta h(x) = 0$ for each $|\beta| \le m$. This follows from

$$\partial_\beta(x-\cdot)^\alpha = \begin{cases} 0\,, & \beta \neq \alpha \\ \alpha!\,, & \beta = \alpha \end{cases}$$

It follows that, for $\varepsilon > 0$, we can find $\delta > 0$ such that

$$|\partial_\beta h(y)| \leq \varepsilon \text{ if } |y - x| \leq \delta \text{ and } |\beta| \leq m\,.$$

Next, set $g(t) := h\big(x + t(y - x)\big)$ and observe that

$$g^{(k)}(t) = \sum_{|\alpha|=m} \frac{t^k (y-x)^k}{\alpha!} \partial_\alpha h\big(x + t(y - x)\big)\,.$$

It follows that

$$g^{(k)}(0) = 0 \,\forall\, k \leq m\,.$$

Using the simple fact that $|x + t(y - x) - x| = t|y - x| \leq \delta$ for $t \in [0, 1]$ we conclude that

$$|g^{(m)}(t)| \leq \varepsilon \sum_{|\alpha|=m} \frac{t^m}{\alpha!} |y - x|^\alpha \leq c\varepsilon |y - x|^m \text{ if } |y - x| \leq \delta$$

where $c = \sum_{|\alpha|=m} \frac{1}{\alpha!}$ and depends only on m and n. Finally

$$|h(y)| = |g(1)| \leq |g^m(t_1)| \leq c\varepsilon |y - x|^m\,.$$

where the existence of $t_1 \in [0, 1]$ follows from the proof of theorem 5.4.6. $\surd$

Remark 9.4.10. (Lagrange Remainder Formula)
Let $f \in \mathrm{C}^{m+1}(D, \mathbb{R})$ and $x, y \in D$, then z can be found on the segment $\{x + t(y - x) \,|\, t \in [0, 1]\}$ such that

$$f(y) - T_m f(x, y) = \sum_{|\alpha|=m+1} \frac{(y-x)^\alpha}{\alpha!} \partial_\alpha f(z)\,.$$

This, in particular shows, that

$$f(y) - T_m f(x, y) = O(|y - x|^{m+1}) \text{ as } y \to x$$

in this case. The extra regularity assumption is crucial.

Proof. The proof follows the steps of the that of theorem 9.4.9 and exploits the single variable remainder formula. $\surd$