A Conceptual Introduction to Several Variable Calculus

S. Kumaresan
School of Math. and Stat.
University of Hyderabad
Hyderabad 500046
kumaresa@gmail.com

We start with some preliminaries on linear maps from $\mathbb{R}^n$ to $\mathbb{R}$.

Let $T\colon V \to W$ be a linear map from a finite dimensional real vector space V to another W. Then T is determined if we know Tv_i, $1 \le i \le n$, where $\{v_i : 1 \le i \le n\}$ is a basis of V. For, if $v \in V$, we can write $v = \sum_i a_i v_i$ for some scalars $a_i \in \mathbb{R}$ so that using linearity of T, we deduce

$$Tv = T(\sum_{i=1}^{n} a_i v_i) = \sum_i T(a_i v_i) = \sum_i a_i T v_i.$$

The right hand side of this equation involves Tv_i and the coefficients a_i of v as a linear combination of v_i's.

We now take $V = \mathbb{R}$ and $W = \mathbb{R}$. Let $A\colon \mathbb{R} \to \mathbb{R}$ be linear. Then as a basis of $\mathbb{R}$, we take $v_1 = 1$. Any real number $x \in \mathbb{R}$ can be written as $x = x \cdot 1$. We therefore have,

$$Tx = T(x \cdot 1) = xT(1) = \alpha x, \quad \text{for all } x \in \mathbb{R}, \quad \text{where } \alpha = T(1). \tag{1}$$

Thus any linear map $T\colon \mathbb{R} \to \mathbb{R}$ is of the form $T(x) = \alpha x$ where $\alpha := T(1)$. Conversely, if we define $T\colon \mathbb{R} \to \mathbb{R}$ by setting $Tx := \alpha x$ for some real scalar α, then it is easy to verify that T is a linear map.

We extend this result for linear maps $A\colon \mathbb{R}^n \to \mathbb{R}$. As a basis of $\mathbb{R}^n$, we choose the so-called standard basis $\{e_i : 1 \le i \le n\}$ where $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ with 1 at the i-th place. Then any $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is expressed as $x = \sum_{i=1}^n x_i e_i$. We apply A to both sides of this equation to get

$$Ax = \sum_{i=1}^{n} x_i A e_i = \sum_{i=1}^{n} \alpha_i x_i, \quad \text{where } \alpha_i = Ae_i,\ 1 \le i \le n. \tag{2}$$

Thus, if $A\colon \mathbb{R}^n \to \mathbb{R}$ is any linear map, there exists a vector $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ such that $Ax = \alpha \cdot x$ where $\alpha_i = Ae_i$ for all i. Conversely, if $\alpha \in \mathbb{R}^n$ is given and if we set $Ax = \alpha \cdot x$, then $A\colon \mathbb{R}^n \to \mathbb{R}$ is linear with $Ae_i = \alpha_i$. Here and in the sequel, $v \cdot w$ and $\langle v, w \rangle$ stand for the dot product of vectors in $\mathbb{R}^n$.

The most important observation in (1) and (2) is that in the expression for T or A, the scalars α or α_i are completely determined as $T(1)$ or as Te_i.

Given two vector spaces V and W, a map of the form $x \mapsto Ax + w_0$ where $A: V \to W$ is linear and $w_0 \in W$ is a fixed vector, is called an *affine function.*

The basic idea of differential calculus is this: Given a function $f: U \subset \mathbb{R}^n \to \mathbb{R}$ and a point $a \in U$, we wish to find a linear map $A: \mathbb{R}^n \to \mathbb{R}$ such that $f(x)$ is "approximately equal to" $f(a) + A(x-a)$ for points x sufficiently near to a, that is, $f(a+h) \approx f(a) + Ah$ for all h near to $0 \in \mathbb{R}^n$. ($\approx$ is read as approximately equal to.) In other words, if we think of h as an increment to the independent variable, then the increment $f(a+h) - f(a)$ in the dependent variable is approximately near to Ah. If we can find such an A, we shall say that f is differentiable at a and the derivative at a is A.

What do we mean by Ah is near to $f(a+h) - f(a)$ for h near zero? An obvious and naive answer would be $f(a+h) - f(a)$ approaches zero (in $\mathbb{R}$) as h approaches 0 in $\mathbb{R}^n$. This is not the correct formulation. For, if we assume that f is continuous at a, we can take A to be the zero map $Ah = 0$ for all $h \in \mathbb{R}^n$. With this choice we have

$$f(a+h) - f(a) - Ah = f(a+h) - f(a) \to 0, \text{ as } h \to 0,$$

by continuity of f at a. What is wrong with this? For example, if we take $f: \mathbb{R} \to \mathbb{R}$ given by $f(x) = |x|$, then we all know that f is continuous at $x = 0$ but not differentiable at $x = 0$. This is clearly not acceptable to us.

A correct definition runs as follows. The 'error term' $f(a+h) - f(a) - Ah$ should approach 0 much 'faster' than h as $h \to 0$. How do we make sense out of this?

Let us look at examples. If we take $f(x) = x$ and $g(x) = x^2$, it is intuitively clear to us that $g(x) \to 0$ much faster than $f(x)$ as $x \to 0$. More generally, if $0 < \alpha < \beta$, then $g(x) := x^\beta$ approaches zero much faster than $f(x) := x^\alpha$ as $x \to 0$. How do we make this rigorous? Observe that $\lim_{x \to 0} \frac{g(x)}{f(x)} = 0$.

With this observation, we go back to $f: U \subset \mathbb{R}^n \to \mathbb{R}$. We say that f is differentiable at a if we can find a linear map $A: \mathbb{R}^n \to \mathbb{R}$ such that $[f(a+h) - f(a) - Ah]/h$ goes to zero as $h \to 0$. This does not make sense, as the denominator h is a vector. An obvious way out is to require that

$$\frac{f(a+h) - f(a) - Ah}{\|h\|} \to 0 \text{ as } h \to 0.$$

Before the precise definition, a small point. Since we want to find an approximation of f near a, we want the domain U of f to contain a neighbourhood of a, that is, we want an $r > 0$ so that the open ball $B(a, r) \subset U$. If we want to talk of differentiability at each point $a \in U$, it is natural to demand that U be open in $\mathbb{R}^n$. In the following, we shall assume that U is open.

We are now ready for the precise definition.

Definition 1. Let $U \subset \mathbb{R}^n$ be open and $a \in U$. A function $f: U \to \mathbb{R}$ is said to be *differentiable* at a if there exists a linear map $A: \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{|f(a+h) - f(a) - Ah|}{\|h\|} \to 0, \text{ as } h \to 0. \tag{3}$$

We can cast this in ε-δ notation.

The function f is differentiable at a iff there exists a linear map $A\colon \mathbb{R}^n \to \mathbb{R}$ such that for any given $\varepsilon > 0$, there exists a $\delta > 0$ with the following property:

$$|f(a+h) - f(a) - Ah| < \varepsilon \|h\| \text{ for all } h \text{ with } \|h\| < \delta. \tag{4}$$

Two remarks are in order. The first is that even though the domain U of f may be a proper subset of $\mathbb{R}^n$, the domain of A is $\mathbb{R}^n$, (as it should be, since the domain of a linear map should be vector spaces!)

The second point is that the linear map appearing in the definition is unique. (This is a place where we need U to be open.) Let us explain the second point in detail. Our claim is this: If $B\colon \mathbb{R}^n \to \mathbb{R}$ is a linear map satisfying (3) with A replaced by B (possibly with different δ for a given ε), then $B = A$, that is, $Ah = Bh$ for all $h \in \mathbb{R}^n$. Using the fact that both A and B satisfy (3), we shall prove that $Ah = Bh$ for all h with $\|h\| = 1$.

Given $\varepsilon > 0$ there exist δ_1 and δ_2 such that the following hold:

$$\begin{aligned} |f(a+h) - f(a) - Ah| &< \frac{\varepsilon}{2}\|h\| \text{ for } \|h\| < \delta_1, \\ |f(a+h) - f(a) - Bh| &< \frac{\varepsilon}{2}\|h\| \text{ for } \|h\| < \delta_2. \end{aligned}$$

Let $\delta = \min\{\delta_1, \delta_2\}$. Let $v \in \mathbb{R}^n$ be given with $\|v\| = 1$. Choose any $t \in \mathbb{R}$ such that $|t| < \delta$. Set $h := tv$. We have

$$\begin{aligned} \|Ah - Bh\| = |t|\,\|Av - Bv\| &= \|(f(a+h) - f(a) - Ah) - (f(a+h) - f(a) - Bh)\| \\ &\le \|f(a+h) - f(a) - Ah\| + \|f(a+h) - f(a) - Bh\| \\ &< \varepsilon \|h\| \\ &= \varepsilon\,|t|\,. \end{aligned}$$

It follows that $\|Av - Bv\| < \varepsilon$ for $\|v\| = 1$ for any $\varepsilon > 0$. Hence we conclude that $Av = Bv$ whenever $\|v\| = 1$. If $w \in \mathbb{R}^n$ is any nonzero vector, if we set $v := w/\|w\|$, then $\|v\| = 1$ so that

$$Av = Bv \implies A(\frac{w}{\|w\|}) = B(\frac{w}{\|w\|}) \implies \frac{1}{\|w\|}Aw = \frac{1}{\|w\|}Bw \implies Aw = Bw.$$

This completes the proof of our claim. The unique A is called the *total derivative* of f at a and is denoted by $Df(a)$.

We now face the following question. If $f\colon (a,b) \subset \mathbb{R} \to \mathbb{R}$ is differentiable at $c \in (a,b)$ in the usual calculus sense, is it differentiable according to our new definition and if so, is there any relation between $f'(c)$ of calculus and $Df(c)$? The answers to these questions is contained in the following

Theorem 2. *Let $f\colon (a,b) \to \mathbb{R}$ be given and $c \in (a,b)$. Then f is differentiable at c in the usual calculus sense iff f is differentiable according to Def. 1 and we have*

$$Df(c)h = f'(c)h, \quad \textit{that is, } f'(c) = Df(c)(1). \tag{5}$$

Proof. Do you understand (5)? The left hand side is the value of the linear map $Df(c)$ at $h = 1$ and the equation says that the real number $f'(c)$ is $Df(c)(1)$.

Let us assume that f is differentiable at c in the usual sense. We shall show that if we define $Df(c)(h) := f'(c)h$, then we have

$$\frac{|f(c+h) - f(c) - Df(c)(h)|}{\|h\|} \to 0 \text{ as } h \to 0.$$

Since $f'(c)$ exists, given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left| \frac{f(c+h) - f(c)}{h} - f'(c) \right| < \varepsilon \text{ for } 0 < |h| < \delta.$$

This is same as saying that

$$\left| \frac{f(c+h) - f(c) - f'(c)h}{h} \right| < \varepsilon \quad \text{for } 0 < |h| < \delta,$$

or,

$$\left| f(c+h) - f(c) - f'(c)h \right| < \varepsilon \, |h| \quad \text{for } |h| < \delta.$$

This shows that f is differentiable at c according to Def. 1 with derivative given by $Df(c)(h) = f'(c)h$.

The converse is proved in an analogous way. We set $\alpha := Df(c)(1)$ and show that

$$\frac{f(c+h) - f(c)}{h} \to \alpha \text{ as } h \to 0.$$

This will prove that f is differentiable at c and the derivative as α. Note that $Df(c)(h) = \alpha h$. Since f is differentiable at c according to Def. 1, given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f(c+h) - f(c) - Df(c)(h)| < \varepsilon \, |h| \quad \text{for } |h| < \delta.$$

Recalling that $Df(c)(h) = \alpha h$ and dividing by $|h|$, we get

$$\left| \frac{f(c+h) - f(c)}{h} - \alpha \right| < \varepsilon \quad \text{for } 0 < |h| < \delta.$$

This proves that $\lim_{h \to 0} \frac{f(c+h) - f(c)}{h}$ exists and the limit is α. $\square$

Remark 3. Note that (5) brings out the intimate relation between $f'(c)$ and $Df(c)$.

Definition 4. Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable at $a \in U$ with $Df(a) = A$. From (2), we know that there exists a unique vector $\alpha \in \mathbb{R}^n$ such that $Ah = \alpha \cdot h \equiv \sum_{i=1}^n \alpha_i h_i$ if $h = (h_1, \dots, h_n)$. This unique vector α is called the *gradient* of f at a. It is denoted by $\operatorname{grad} f(a)$ or $\nabla f(a)$. In view of (2), we have

$$\operatorname{grad} f(a) = (Df(a)(e_1), \dots, Df(a)(e_n)). \tag{6}$$

Later we shall identify $Df(a)(e_j)$. (See (7) and (8).)

Let us now look at some examples.

Example 5. Let $f\colon U \subset \mathbb{R}^n \to \mathbb{R}$ be a constant. Let $a \in U$ be arbitrary. If we look at $f(a+h) - f(a) = 0$, an obvious choice for $Df(a)$ is the zero linear map. We see that

$$\frac{|f(a+h) - f(a) - 0h|}{\|h\|} = \frac{0}{\|h\|} = 0.$$

Hence f is differentiable at a with $Df(a) = 0$, the zero linear map.

Example 6. Let f be the restriction to an open $U \subset \mathbb{R}^n$ of a linear amp $A\colon \mathbb{R}^n \to \mathbb{R}$. It should be intuitively clear that $Df(a)$ exists and equals A for *any* $a \in U$. Let us verify this.

$$|f(a+h) - f(a) - Ah| = |Aa + Ah - Aa - Ah| = |0|.$$

Example 7. Let $f\colon \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$. Note that from Theorem 2, we know that f is differentiable according to our new definition and that $Df(a)(h) = 2ah$. Let us verify this from first principles.

$$\begin{aligned} |f(a+h) - f(a)| &= (a+h)^2 - a^2 \\ &= a^2 + 2ah + h^2 - a^2 \\ &= 2ah + h^2. \end{aligned}$$

The term on the right hand side which is 'linear' in h is $2ah$. It suggests that we may take as $Df(a)$ the linear map $h \mapsto 2ah$. We do so and find that

$$\frac{|f(+h) - f(a) - 2ah|}{|h|} = \frac{|h^2|}{|h|} \to 0 \text{ as } h \to 0.$$

Example 8. Consider $f\colon \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = x \cdot x \equiv \sum_{i=1}^n x_i^2$. Again, we shall compute the derivative from the first principles.

$$\begin{aligned} f(a+h) - f(a) &= (a+h)\cdot(a+h) \\ &= a\cdot a + 2a\cdot h + h\cdot h - a\cdot a \\ &= 2a\cdot h + h\cdot h. \end{aligned}$$

The term 'linear' in h on the right hand side of the above equation is $2a \cdot h$. This suggests that we may take $Df(a)(h) := 2a \cdot h$. Using this, we see that

$$\frac{|f(a+h) - f(a) - Df(a)(h)|}{\|h\|} = \frac{|h\cdot h|}{\|h\|} = \frac{\|h\|^2}{\|h\|},$$

which goes to 0 as $h \to 0$. Note that $\operatorname{grad} f(a) = 2a$.

Example 9. Let $f, g\colon \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) := x + y$ and $g(x,y) := xy$. We show that f and g are differentiable.

We have $f(a+h, b+k) - f(a,b) = h + k$. So, if we define $A(h,k) := h + k$, then $A\colon \mathbb{R}^2 \to \mathbb{R}$ is linear and $f(a+h, b+k) - f(a,b) - A(h,k) = 0$.

As for g, we find that $g(a+h, b+k) - g(a,b) = ak + bh + hk$. This suggests as a possible candidate for the derivative B, we can try $B(h,k) := ak + bh$. We find that the 'error' is then hk. We need to show that

$$\frac{|g(a+h, b+k) - f(a,b) - B(h,k)|}{\sqrt{h^2+k^2}} = \frac{hk}{\sqrt{h^2+k^2}} \to 0 \text{ as } (h,k) \to 0.$$

This follows easily if we observe that $|h| \le \sqrt{h^2+k^2}$ and $|k| \le \sqrt{h^2+k^2}$.

In particular, we find that $\operatorname{grad} f(a,b) = (1,1)$ and $\operatorname{grad} g(a,b) = (b,a)$.

Example 10. One last example in the same vein. Let A be an $n \times n$-symmetric matrix with real entries. Define $f(x) := Ax \cdot x$. Here Ax denotes the column vector got from the matrix multiplication of A and the column vector x. Let us show that f is differentiable at each $a \in \mathbb{R}^n$ and compute its derivative.

$$\begin{aligned} f(a+h) - f(a) &= A(a+h)\cdot(a+h) - Aa\cdot a \\ &= (Aa + Ah)\cdot(a+h) - a\cdot a \\ &= Aa\cdot a + 2Aa\cdot h - Aa\cdot h. \end{aligned}$$

An obvious choice for $Df(a)$ is by setting $Df(a)h := 2Aa \cdot h$, for $h \in \mathbb{R}^n$. Let us check whether this choice works.

$$\frac{|f(a+h) - f(a) - Df(a)(h)|}{\|h\|} = \frac{|2Aa\cdot h|}{\|h\|} \le \frac{\|Ah\|\,\|h\|}{\|h\|} = \|Ah\|\,,$$

which goes to zero as $h \to 0$ by continuity of A at $h = 0$. Incidentally, we have $\operatorname{grad} f(a) = 2Aa$.

Ex. 11. Let $M(2,\mathbb{R})$ denote the set of all 2×2 matrices with real entires. We may identify $M(2,\mathbb{R})$ with $\mathbb{R}^4$ as a vector space with the norm $\|A\| := \|(a_{11}, a_{12}, a_{21}, a_{22})\|$. Consider $f\colon M(2,\mathbb{R}) \to \mathbb{R}$ be given by $f(X) = \det(X)$. Show that f is differentiable at I and that $Df(I)(H) = \operatorname{Tr}(H)$.

Ex. 12. Keep the notation of the last exercise. Show that the map $f\colon M(2,\mathbb{R}) \to M(2,\mathbb{R})$ given by $X \mapsto X^2$ is differentiable at A and $Df(A)(H) = AH + HA$.

Let us now go back to the general theory.

> The single most important trick in several variable differential calculus is the reduction of the problems to one-variable case.

We explain this principle by exhibiting its employment in various results in differential calculus.

Let $U \subset \mathbb{R}^n$ be open and $a \in U$. Then there exists an $r > 0$ such that $B(a,r) \subset U$. Let $v \in \mathbb{R}^n$ be any vector. If we take, $\varepsilon := r/\|v\|$, ($\varepsilon = \infty$ if $v = 0$). Then $a + tv \in B(a,r) \subset U$ for all t with $|t| < \varepsilon$. We can thus define a one-variable function $g\colon (-\varepsilon,\varepsilon) \to \mathbb{R}$ by setting $g(t) := f(a+tv)$. (Note that g depends on v also.) Observe that $g(0) = f(a)$. We claim that g is differentiable at a and that $g'(0) = Df(a)(v)$. We prove it thus. First of all, observe that g is a constant if $v = 0$ and in this case, the claim is obvious. So we assume that $v \ne 0$. We need to prove that given $\varepsilon > 0$, there exists $\delta > 0$ such that $\left|\frac{g(h)-g(0)}{h} - Df(a)(v)\right| < \varepsilon$ for all h with $0 < |h| < \delta$. For this $\varepsilon > 0$, by differentiability of f at a, we can find a $\delta > 0$ such that $\frac{|f(a+hv)-f(a)-Df(a)(hv)|}{\|hv\|} < \varepsilon/\|v\|$, provided $0 < \|hv\| < \delta$.

$$\begin{aligned}\left|\frac{g(h)-g(0)}{h} - Df(a)(v)\right| &= \left|\frac{f(a+hv)-f(a)-hDf(a)(v)}{h}\right| \\ &= \left|\frac{f(a+hv)-f(a)-Df(a)(hv)}{h}\right| \quad \text{(using linearity of } Df(a)) \\ &= \|v\| \frac{f(a+hv)-f(a)-Df(a)(hv)}{\|hv\|} \\ &< \varepsilon,\end{aligned}$$

by differentiability of f at a. This proves the claim.

Before summarizing what we have done, we need a definition.

Definition 13. Let $f\colon U \subset \mathbb{R}^n \to \mathbb{R}$ be any function. Fix a vector $v \in \mathbb{R}^n$. We say that f has *directional derivative* at a in the direction of v if the limit $\lim_{t\to 0} \frac{f(a+tv)-f(a)}{t}$ exists. Note that since U is open, as observed earlier, $a + tv \in U$ for all t in a sufficiently small interval around 0. The limit, if exists, is denoted by $D_v f(a)$.

What we have proved just before this definition is the following theorem.

Theorem 14. *Let $f\colon U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable at $a \in U$. Then $D_v f(a)$ exists for all $v \in \mathbb{R}^n$ and we have*

$$D_v f(a) = Df(a)(v). \tag{7}$$

□

Remark 15. There exist functions $f\colon U \to \mathbb{R}$ such that $D_v f(a)$ exist for all $v \in \mathbb{R}^n$ but f is not differentiable.

Let us now specialize the vector v, in the definition of the directional derivatives, by taking $v = e_i$. In this case,

$$f(a + te_i) = f(a_1, \ldots, a_{i-1}, a_i + t, a_{i+1}, \ldots, a_n),$$

so that

$$\lim_{t\to 0} \frac{f(a+te_i)-f(a)}{t} = \lim_{t\to 0} \frac{f(a_1,\ldots,a_{i-1},a_i+t,a_{i+1},\ldots,a_n) - f(a_1,\ldots,a_n)}{t}.$$

The limit, namely, the direction derivative $D_{e_i} f(a)$, if exists, is called the i-th *partial derivative* of f at a and is usually denoted $\frac{\partial f}{\partial x_i}(a)$ or at times by $D_i f(a)$.

We now go back to the question raised earlier. Can we find a concrete expression for the vector $\operatorname{grad} f(a)$? Recall from (2), if $T\colon \mathbb{R}^n \to \mathbb{R}$ is linear, then $Tv = (Te_1, \ldots, Te_n) \cdot (v_1, \ldots, v_n)$. Hence, we have

$$\begin{aligned}Df(a)(h) &= (Df(a)(e_1), \ldots, Df(a)e_n) \cdot (h_1, \ldots, h_n) \\ &= (D_{e_1} f(a), \ldots, D_{e_n} f(a)) \cdot (h_1, \ldots, h_n) \\ &= (\frac{\partial f}{\partial x_1}(a), \ldots, \frac{\partial f}{\partial x_n}(a)) \cdot (h_1, \ldots, h_n).\end{aligned}$$

Thus we have proved

$$\operatorname{grad} f(a) = (\frac{\partial f}{\partial x_1}(a), \ldots, \frac{\partial f}{\partial x_n}(a)). \tag{8}$$

Let us understand what we have achieved. Computing $\frac{\partial f}{\partial x_i}(a)$ is a one variable job: it is $g'(0)$ where $g(t) := f(a + te_i)$. Thus, if we somehow know that f is differentiable, say, by means of some theoretical considerations, then we can compute $Df(a)$ simply by finding the directional derivatives $D_v f(a)$ or the partial derivatives $\frac{\partial f}{\partial x_i}(a)$ for $1 \leq i \leq n$. See Example 16 below.

Example 16. Let A be an $n \times n$-real matrix. Let $R \colon \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ be defined by setting

$$R(x) := \frac{\langle Ax, x\rangle}{\langle x, x\rangle}, \quad x \neq 0.$$

Then it follows that both numerator and the denominator are differentiable functions and hence their quotient is differentiable on $\mathbb{R}^n \setminus \{0\}$. The derivative can be computed using the algebra of differentiable functions. However, we shall show how to compute the derivative in a simpler way. Since it is already known that f is differentiable, it is enough to compute $Df(x)(v)$ for $x \neq 0$ for an arbitrary vector $v \in \mathbb{R}^n$. By (7) we know $Df(x)(v)$ is the directional derivative $D_v f(x)$. If we set $g(t) := R(x + tv)$, then $D_v f(x) = g'(0)$. Computing this is easy, since $g(t) = \frac{\varphi(t)}{\psi(t)}$ where

$$\varphi(t) := \langle A(x + tv), x + tv\rangle = \langle Ax, x\rangle + t\,\langle Ax, v\rangle + t\,\langle Av, x\rangle + t^2\,\langle Av, v\rangle$$

and

$$\psi(t) := \langle x, x\rangle + 2t\,\langle x, v\rangle + t^2\,\langle v, v\rangle\,.$$

Hence from the quotient rule of one variable calculus,

$$\begin{aligned} g'(0) &= \frac{\varphi'(0)\psi(0) - \varphi(0)\psi'(0)}{\psi(0)^2} \\ &= \frac{\langle Ax, v\rangle + \langle Av, x\rangle}{\langle x, x\rangle} - 2\frac{\langle Ax, x\rangle\,\langle x, v\rangle}{(\langle x, x\rangle)^2}. \end{aligned}$$

We hope that the simplicity impresses about the utility of our principle!

Ex. 17. In Example 8, the function is $f(x) := x_1^2 + \cdots + x_n^2$. From our work over there we know $\operatorname{grad} f(a) = 2a$. Assuming the algebra of differentiable functions, f is differentiable. Compute the partial derivatives and hence $\operatorname{grad} f(a)$. Compare your work with Example 8

Carry out a similar investigation with $f(x) := \sum_{i,j} a_{ij} x_i x_j$ of Example 10.

Let us look at another instance of the principle of reduction to the one-variable case. Let $f \colon U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable at all point of U. Assume that $p \in U$ is a point of local maximum, that is, $f(x) \leq f(p)$ for all $x \in B(p, r)$ for some $r > 0$. We claim that $Df(p) = 0$. In view of (7) and (8), it suffices to show that $\frac{\partial f}{\partial x_i}(p) = 0$ for $1 \leq i \leq n$. Consider the one-variable function $g(t) := f(p + te_i)$ defined on $(-\varepsilon, \varepsilon)$ for sufficiently small $\varepsilon > 0$. Since $g(0) = f(p)$ and $p + te_i \in B(p, r)$, we see that $t = 0$ is a maximum of g on $(-\varepsilon, \varepsilon)$. From one-variable result, it follows that $g'(0) = 0$. What is $g'(0)$? As done earlier,

$$\lim_{t \to 0} \frac{g(t) - g(0)}{t} = \lim_{t \to 0} \frac{f(p + te_i) - f(p)}{t} = D_{e_i} f(p) = \frac{\partial f}{\partial x_i}(p).$$

Hence the claim is proved.

A third instance of the principle in action is seen in the proof of the mean value theorem for differentiable functions $f\colon U \to \mathbb{R}$.

Theorem 18 (Mean Value Theorem)**.** *Let $f\colon U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable on U. Let $x, y \in U$. Assume that the line segment $[x, y] := \{(1-t)x + ty : 0 \le t \le 1\} \subset U$. Fix any $v \in \mathbb{R}^n$. Then there exists $t_0 \in [0,1]$ such that if we set $z := (1-t_0)x + t_0 y$, then*

$$f(y) - f(x) = Df(z)(y-x) \equiv \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(z)(y_i - x_i). \tag{9}$$

Proof. Consider $g(t) := f((1-t)x + ty)$ on $[0,1]$. Then g is continuous on $[0,1]$ and differentiable on $(0,1)$. (To show that g is differentiable, adapt the computation of $g'(t_0)$ below.) Hence by mean value theorem of one-variable calculus, there exists $t_0 \in (0,1)$ such that $g(1) - g(0) = g'(t_0)(1-0) = g'(t_0)$. What is $g'(t_0)$?

$$\begin{aligned}
\lim_{h\to 0} \frac{g(t_0+h) - g(t_0)}{h} &= \lim_{h\to 0} \frac{f((1-t_0-h)x + (t_0+h)y) - f((1-t_0)x + t_0 y)}{h} \\
&= \lim_{h\to 0} \frac{f([(1-t_0)x + t_0 y] + h(y-x)) - f((1-t_0)x + t_0 y)}{h} \\
&= D_{y-x} f(z) \equiv Df(z)(y-x).
\end{aligned}$$

(Note that the above computation proves that g is differentiable at t_0 and computes the derivative.) ☐

As an immediate consequence, we have

Corollary 19. *Assume that U is star-shaped at $a \in U$, that is, the line segment $[a, x] \subset U$ for all $x \in U$. Assume that $f\colon U \to \mathbb{R}$ be differentiable on U and that $Df(x) = 0$ for $x \in U$. Then f is a constant.*

Proof. Let $x \in U$ be arbitrary. By the mean value theorem, the exists $z \in [a, x]$ such that

$$f(x) - f(a) = Df(z)(x-a) = 0.$$

Thus, $f(x) = f(a)$ for all $x \in U$, that is f is a constant. ☐

A fourth example of an application of the principle is lies in the derivation of Taylor's formula for functions $f\colon U \subset \mathbb{R}^n \to \mathbb{R}$. Since our aim is to bring out the ideas clearly, we shall restrict ourselves to C^2-functions. If you wish, you may assume that $n = 2$. A function f is said to be C^2 if all partial derivatives of the form

$$\frac{\partial^{\alpha_1 + \cdots + \alpha_n} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}$$

exist where $\alpha_j \in \mathbb{Z}_+$ and $\alpha_1 + \cdots + \alpha_n \le 2$ and are continuous. When $n = 2$, adopting the usual notation this means that all partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial y^2}$, $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ exist and are continuous. (Note that $\frac{\partial f}{\partial x}$ etc. are functions from U to $\mathbb{R}$ so we can speak of their partial derivatives w.r.t. x or y.)

To make things simple, let us also assume that $0 \in U$, U is star-shaped at 0 and that we wish to find a 'Taylor expansion' of f at 0. Consider the one-variable function $g(t) := f(tx)$. We first observe that this is defined for t in $(-\varepsilon, 1+\varepsilon)$ for some sufficiently small $\varepsilon > 0$. For, since U is open and $0 \in U$, there exists a $\varepsilon_1 > 0$ such that $B(0, \varepsilon_1) \subset U$. Hence $0 + tx \in B(0, \varepsilon_1)$ provided that $\|tx\| = |t| \|x\| < \varepsilon_1$, that is, when $|t| < \varepsilon_1 / \|x\|$. A similar consideration will show that $x + tx = (1+t)x \in B(x, \varepsilon_2) \subset U$ if $|t| < \varepsilon_2 / \|x\|$. If we choose $\varepsilon := \min\{\varepsilon_1 / \|x\|, \varepsilon_2 / \|x\|\}$, then $tx \in U$ for $t \in (-\varepsilon, 1+\varepsilon)$. Let us show that g is differentiable on this interval and compute its derivative.

$$\begin{aligned} g'(t) = \lim_{h \to 0} \frac{g(t+h) - g(t)}{h} &= \lim_{h \to 0} \frac{f((t+h)x) - f(tx)}{h} \\ &= \lim_{h \to 0} \frac{f(tx + hx) - f(tx)}{h} \\ &= D_x f(tx) \equiv Df(tx)(x) \quad \text{by (7)} \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(tx) x_i \quad \text{by (8)}. \end{aligned}$$

In particular,

$$g'(0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(0) x_i. \tag{10}$$

Let $g_i(t) := \frac{\partial f}{\partial x_i}(tx)$. If we proceed as above, we find that

$$g_i'(t) = \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}(tx) x_j = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(tx) x_j, \tag{11}$$

so that we have

$$\begin{aligned} g''(t) &= \sum_{i=1}^{n} g_i'(t) \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(tx) x_j \right) x_i \\ &= \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(tx) x_j x_i. \end{aligned} \tag{12}$$

Note that the above computation show that g is twice continuously differentiable and so we can apply Taylor's theorem of one-variable calculus to g. We get

$$g(t) = g(0) + g'(0)t + g''(0)t^2 + R,$$

where the remainder R is such that $\lim_{t \to 0} R/t^2 = 0$. Taking $t = 1$ in the last displayed equation, we deduce the Taylor's formula for f.

$$f(x) = f(0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(0) x_i + \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(0) x_i x_j + R. \tag{13}$$

From (13), it is easy to deduce the sufficient condition (in terms of second order partial derivatives) for the local maximum/minimum of f. If we assume, for instance $x = 0$ is a point of local maximum, then $t = 0$ is a point of local maximum for g. Hence $g''(0) \leq 0$. Coming back to f, this implies that

$$\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(0) x_i x_j \leq 0$$

for all choices of x in a neighbourhood of 0. This is same as saying that the matrix

$$D^2 f(0) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(0) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(0) \\ & \vdots & & \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(0) & \frac{\partial^2 f}{\partial x_n \partial x_2}(0) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(0) \end{pmatrix}$$

is negative semi-definite. (A symmetric matrix A of size n is *negative semi-definite* if $Ax \cdot x \leq 0$ for all $x \in \mathbb{R}^n$. It is said to be negative definite if $Ax \cdot x < 0$ for all $x \neq 0$. In the above, we use the fact that for C^2-functions, $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$ and hence the matrix $D^2 f(x)$ is symmetric.)

A sufficient condition for $t = 0$ to be a point of local maximum is that $g''(0) < 0$. This translates into the following sufficient condition for $x = 0$ to a local maximum for the function f. The above matrix $D^2 f(0)$ of second order partial derivatives at 0 must be negative-definite. The cases of local minima are left to the reader. The last two paragraphs need a careful analytical treatment which is not very difficult. Since our aim is to show the principle in action, we shall be content with this.

Before we wind up, we shall state two versions of the chain rule in our set-up and give applications. For one version, we need to introduce the concept of differentiability of a vector valued function of a scalar variable.

If $f\colon (a,b) \to U \subset \mathbb{R}^n$ is a function, we think of it as a parametrized curve in U. A physical interpretation of f is that it is the path traversed by a particle from time a to time b. We write $f(t) = (f_1(t), \ldots, f_n(t))$. We say that f is differentiable at $t_0 \in (a,b)$ if each of the one-variable function $f_j\colon (a,b) \to \mathbb{R}$ is differentiable at t_0 for $1 \leq i \leq n$ and the derivative of f is defined by the equation $f'(t) = (f_1'(t), \ldots, f_n'(t))$. This derivative has a physical interpretation too. If we think of f as the path traversed by a particle from time $t = a$ to time $t = b$, then $f'(t)$ is the (instantaneous) velocity (vector) of the particle at t. This is justified in view of the observation that

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

While this easy to prove, we do not give a proof.

Theorem 20 (Chain Rule)**.** (a) *Let $f\colon (a,b) \to U \subset \mathbb{R}^n$ be differentiable at $t \in (a,b)$ and $g\colon U \to \mathbb{R}$ be differentiable at $f(t)$. Then $g \circ f$ is differentiable at t and we have*

$$(g \circ f)'(t) = Dg(f(t))(f'(t)) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(f(t)) f_i'(t). \tag{14}$$

(b) *Let $g\colon U \to \mathbb{R}$ be differentiable at $p \in U$. Let $J \subset \mathbb{R}$ be an open interval such that $q := g(p) \in J$. Let $\varphi\colon J \to \mathbb{R}$ be differentiable at q. Then $\varphi \circ g$ is differentiable at p and we have*

$$\operatorname{grad}(\varphi \circ g)(p) = \varphi'(q) \operatorname{grad} \varphi(p). \tag{15}$$

Proof. We do not prove this. The aim here is to make sure that you understand (14) and (15).

In case (a), the function $g \circ f$ is a function from $\mathbb{R}$ to $\mathbb{R}$ and hence it is enough to know its derivative in the usual calculus sense. (14) says that this derivative is the real number got by letting the linear map $Dg(f(t)) \colon \mathbb{R}^n \to \mathbb{R}$ act on the vector $f'(t)$.

In case (b), the function $\varphi \circ g$ is from $\mathbb{R}^n$ to $\mathbb{R}$. Its derivative is known if we know the gradient. (15) says that this gradient is $\operatorname{grad} g(p)$ multiplied by the scalar $\varphi'(q)$. □

Let us give a typical application of (15).

Theorem 21. *Let $f \colon U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable at a with $f(x) \neq 0$ for $x \in B(a,r)$ for some $r > 0$. Then $g \colon B(a,r) \to \mathbb{R}$ given by $g(x) = 1/f(x)$ is differentiable at a and we have*

$$\operatorname{grad} g(a) = -\frac{1}{f(a)^2} \operatorname{grad} f(a).$$

Proof. Let $\mathbb{R}^*$ stand for the set of nonzero real numbers. Let $\varphi \colon \mathbb{R}^* \to \mathbb{R}$ be given by $\varphi(t) = 1/t$. Then φ is differentiable on $\mathbb{R}^*$ and we observe that $g = \varphi \circ f$. Hence g is differentiable at a and we deduce from (15),

$$\operatorname{grad} g(a) = \varphi'(f(a)) \operatorname{grad} f(a) = -\frac{1}{f(a)^2} \operatorname{grad} f(a).$$

□

A typical application of (14) is the following.

Theorem 22. *Let $f \colon U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable. Let $c \colon (a,b) \to \mathbb{R}^n$ be differentiable. Assume that for all $t \in (a,b)$, $f(c(t)) = \alpha$, a constant. Then* $\operatorname{grad} f(c(t))$ *is perpendicular to the tangent vector $c'(t) = (c_1'(t), \ldots, c_n'(t))$ for all t.*

Proof. Since $f \circ c$ is a constant, we see that $(f \circ c)'(t) = 0$ for all t. By (14), we have

$$(f \circ c)'(t) = \left\langle \operatorname{grad} f(c(t)), c'(t) \right\rangle .$$

The result follows. □

Remark 23. Let us bring out the geometric significance of the last result. Let us assume that $n = 3$. Then the set of points $\{x \in U : f(x) = \alpha\}$, if non-empty, can be considered as a 'surface'. Then the hypothesis on c says that the curve c lies entirely in S. Hence $c'(t)$ can also be thought of a vector tangent to the surface S. Thus, we see that $\operatorname{grad} f(a)$ is 'normal' to the surface at $a \in S$. Take some linear and quadratic functions on $\mathbb{R}^3$ and try to understand this remark.

Geometric Meaning of Derivatives and Partial Derivatives

To facilitate geometric imagination, we shall consider functions $f\colon \mathbb{R}^2 \to \mathbb{R}$. Look at the graph of f, the set of points $((x, y), f(x, y)) \in \mathbb{R}^3$. This can be viewed as a surface in $\mathbb{R}^3$. A geometric view point of the derivative of f at a is obtained by looking at the surface explicitly defined by f and the tangent plane to the surface at the point $(a, b, f(a, b))$. The tangent plane under consideration is given as the graph of the affine function

$$A : (x, y) \mapsto f(a, b) + Df(a, b)(x - a, y - b) \equiv f(a, b) + \operatorname{grad} f(a, b) \cdot (x - a, y - b).$$

Thus the difference $f(x, y) - A(x, y)$ is a good approximation to the increment $f(x, y) - f(a, b)$. If we use (8), then, in terms of the coordinates, the equation to the tangent plane at (a, b) is

$$z = f(a, b) + (x - a)\frac{\partial f}{\partial x}(a, b) + (y - b)\frac{\partial f}{\partial y}(a, b). \tag{16}$$

Now, let us give the geometric interpretation of partial derivatives.

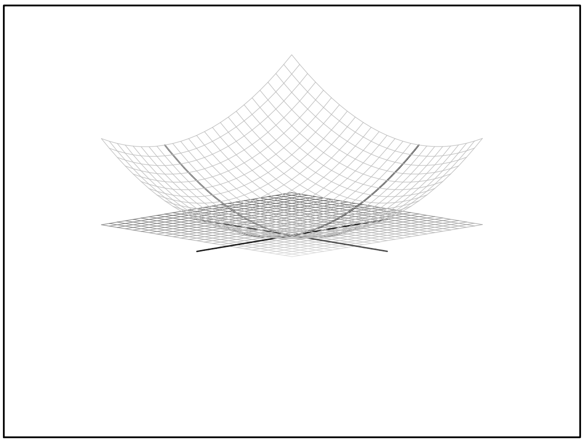

Figure 1: Graph of $f(x, y) = x^2 + y^2$ and its Tangent plane at $(1, 1, 2)$

The partial derivative $\frac{\partial f}{\partial x}(a, b)$ is the derivative at a of the one variable function $x \mapsto f(x, b)$. The intersection of the surface with the vertical plane determined by the equation $y = b$ is a curve determined by the equations

$$z = f(x, y) \text{ and } y = b.$$

The partial derivative $\frac{\partial f}{\partial x}(a, b)$ is the slope of the tangent line at a to this curve considered as the graph of $x \mapsto f(x, b)$. Similar interpretation holds for the other partial derivative. For this reason, it is natural to define a tangent plane to the surface at (a, b) as the plane containing the tangent lines to the curves

$$z = f(x, y) \;\&\; y = b \quad \text{and} \quad z = f(x, y) \;\&\; x = a.$$

It is easily verified that the set of points satisfying the equation

$$z = f(a, b) + (x - a)\frac{\partial f}{\partial x}(a, b) + (y - b)\frac{\partial f}{\partial y}(a, b)$$

is a plane containing the two tangent lines. (Observe that, by taking $y = b$ or $x = a$ we get the tangent lines!) The figure illustrates these points.