Differential Calculus of Vector Functions

February 29, 2000

1. Below $m$ and $n$ are two positive integers. They play the role of the dimensions. In Stewart they usually do not exceed 3. The case $m = 1$ corresponds to a parametric curve in the $n$-dimensional space, while the case $n = 1$ corresponds to a (numeric) function of $m$ variables.

These notes should be studied in conjunction with lectures. Note that vectors in $\mathbb{R}^m$ below are represented as columns of $m$ numbers

$$\mathbf{v} = \left( \begin{array}{c} v_1 \\ \vdots \\ v_m \end{array} \right)$$

where Stewart uses notation $\mathbf{v} = (v_1, \ldots, v_m)$.

2. Functions $L : \mathbb{R}^m \to \mathbb{R}^n$ having the following two properties

   (a) Additivity: $L(\mathbf{v} + \mathbf{w}) = L(\mathbf{v}) + L(\mathbf{w})$ for any vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^m$, and

   (b) Homogeneity: $L(c\mathbf{v}) = cL(\mathbf{v})$ for any vector $v \in \mathbb{R}^m$ and real number $c$.

are called linear transformations. They form one of the simplest classes of vector functions $f : \mathbb{R}^m \to \mathbb{R}^n$. If

$$A = \left( \begin{array}{ccc} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{array} \right)$$
is an \( n \) by \( m \) matrix then the function

\[
L(v) := Av \equiv \begin{pmatrix}
a_{11} & \cdots & a_{1m} \\
\vdots & & \vdots \\
a_{n1} & \cdots & a_{nm}
\end{pmatrix}
\begin{pmatrix}
v_1 \\
\vdots \\
v_m
\end{pmatrix}
\]

is a linear transformation. In fact, every linear transformation is of this form for a unique \( n \times m \) matrix \( A \).

**Exercise 1.** Verify that

\[
a_{ij} = (i^{\text{th}} \text{ row of } A) \cdot e_j
\]

where

\[
e_j = \begin{pmatrix}
0 \\
\vdots \\
1 \\
0 \\
\vdots
\end{pmatrix}
\]

(“1” at the \( j^{\text{th}} \) place).

(Note that \( i = e_1, j = e_2 \) and \( k = e_3 \) in \( \mathbb{R}^3 \).)

3. Consider a function \( f : D \rightarrow \mathbb{R}^m \) which is defined on some subset \( D \) of \( \mathbb{R}^m \). Let \( x^0 \) be an interior point of \( D \). We shall say that \( f \) is **differentiable** at \( x^0 \) if there exists a linear transformation \( L : \mathbb{R}^m \rightarrow \mathbb{R}^n \) such that

\[
f(x) - f(x^0) = L(x - x^0) + u(x)
\]

where \( u(x) \) is negligible compared to \( L(x - x^0) \) when \( x \rightarrow x^0 \). “Negligible” means that \( u(x) \) tends to 0 faster than \( L(x - x^0) \) does. If such a linear transformation \( L \) exists then \( L \) is unique. It will be denoted \( f'(x^0) \) and called the **derivative** of \( f \) at \( x^0 \) and thus (4) can be rewritten as

\[
f(x) = f(x^0) + (f'(x^0)(x - x^0)) + u(x)
\]

where \( u(x) \) is negligible when \( x \) approaches \( x^0 \).
The $n \times m$ matrix corresponding to the linear transformation $f'(x^0) : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is formed by partial derivatives of components of $f$:

$$
\left( \begin{array}{ccc}
\frac{\partial f_1}{\partial x_1}(x^0) & \cdots & \frac{\partial f_1}{\partial x_m}(x^0) \\
\vdots & \ddots & \vdots \\
\frac{\partial f_n}{\partial x_1}(x^0) & \cdots & \frac{\partial f_n}{\partial x_m}(x^0)
\end{array} \right).
$$

(6)

Here $f(x) = (f_1(x), \ldots, f_n(x))$. Each component $f_i$ of $f$ is a (numeric) function $D \rightarrow \mathbb{R}$. Matrix (6) is called Jacobi’s matrix of $f$ at $x^0$ and will be denoted $J_f(x^0)$.

**Exercise 2.** Rewrite (5) in terms of Jacobi’s matrix (6).

*Hint:* Use formula (1).

**Exercise 3.** Calculate Jacobi’s matrix of $f(x, y, z) = (x^2 y - z, e^z \sin x)$.

4. **The case of a parametric curve $\gamma(t)$ in $\mathbb{R}^n$**. Here $\gamma'(t^0)$ is a linear transformation $\mathbb{R} \rightarrow \mathbb{R}^n$.

**Exercise 4.** Show that any linear transformation $L : \mathbb{R} \rightarrow \mathbb{R}^n$ is of the form

$$
L(t) = ta
$$

(7)

for some fixed vector $a \in \mathbb{R}^n$. (This $a$ can be thought of as a vector-valued “slope” of the graph of $L$ which is a line in $\mathbb{R}^n$ passing through $0$.)

According to Exercise 4, the linear transformation $\gamma'(t^0) : \mathbb{R} \rightarrow \mathbb{R}^n$ is given by formula (7) for a suitable vector $a$. This vector $a$ happens to be the same as the **velocity vector** of the parametric curve:

$$
\left( \begin{array}{c}
\frac{d\gamma_1}{dt}(t^0) \\
\vdots \\
\frac{d\gamma_n}{dt}(t^0)
\end{array} \right)
$$

(8)
which also coincides with Jacobi’s matrix of $\gamma$.

5. The case of a numeric function of $m$ variables $f : D \to \mathbb{R}$.

**Exercise 5.** Show that any linear transformation $L : \mathbb{R}^m \to \mathbb{R}$ (a.k.a. a linear functional on $\mathbb{R}^m$) is of the form:

$$L(\mathbf{v}) = \mathbf{a} \cdot \mathbf{v} \quad (9)$$

for some fixed vector $\mathbf{a} \in \mathbb{R}^m$.

**Question:** What do you think the graph of a linear functional $L : \mathbb{R}^2 \to \mathbb{R}$ is? Can you write the equation?

The linear functional $f'(\mathbf{x}^0) : \mathbb{R}^m \to \mathbb{R}$ is usually denoted $d_{\mathbf{x}}f$ or $df(\mathbf{x}^0)$ and called the **differential** of $f$ at $\mathbf{x}^0$. Jacobi’s matrix of $f$ is:

$$
\left( \frac{\partial f}{\partial x_1}(\mathbf{x}^0) \quad \ldots \quad \frac{\partial f}{\partial x_m}(\mathbf{x}^0) \right) \quad (10)
$$

and

$$d_{\mathbf{x}}f(\mathbf{v}) = \nabla f(\mathbf{x}^0) \cdot \mathbf{v} \quad (11)$$

where $\nabla f(\mathbf{x}^0)$ is the column vector:

$$
\nabla f(\mathbf{x}^0) = \left( \begin{array}{c}
\frac{\partial f}{\partial x_1}(\mathbf{x}^0) \\
\vdots \\
\frac{\partial f}{\partial x_m}(\mathbf{x}^0)
\end{array} \right) \quad (12)
$$

Vector (12) is called the **gradient** of $f$ at $\mathbf{x}^0$. Note that it is the transpose of Jacobi’s matrix (10).

In the case of a function $f : D \to \mathbb{R}$, formula (5) becomes

$$f(\mathbf{x}) = f(\mathbf{x}^0) + d_{\mathbf{x}} f(\mathbf{x} - \mathbf{x}^0) + u(\mathbf{x})$$

$$= f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0) + u(\mathbf{x}) \quad (13)$$
where $u(x)$ is negligible when $x$ approaches $x^0$.

6. Chain Rule. Suppose that two functions are given

$$f : D \to \mathbb{R}^n \quad \text{where} \quad D \subseteq \mathbb{R}^m$$

and

$$g : E \to \mathbb{R}^m \quad \text{where} \quad E \subseteq \mathbb{R}^\ell$$

such that the composition $f \circ g$ is well defined. This means that $g(x) \in D$ for every $x \in E$.

Suppose that $g$ is differentiable at $x^0$ and that $f$ is differentiable at $y^0 = g(x^0)$. In other words:

$$g(x) - g(x^0) = g'(x^0)(x - x^0) + v(x) \quad (14)$$

and

$$f(y) - f(y^0) = f'(y^0)(y - y^0) + u(y) \quad (15)$$

where $v(x)$ and $u(y)$ are negligible when $x \to x^0$ and $y \to y^0$, respectively.

Plug (14) into (15):

$$f(g(x)) - f(g(x^0)) = f'(g(x^0))(g(x) - g(x^0)) + u(g(x))$$

$$= f'(g(x^0))(g'(x^0)(x - x^0) + v(x)) + u(g(x)) \quad (16)$$

$$= (f'(g(x^0)) \circ g'(x^0))(x - x^0) + [f'(g(x^0))(v(x)) + u(g(x))]$$

The composition of two linear transformations is linear. Therefore $f'(g(x^0)) \circ g'$ is a linear transformation from $\mathbb{R}^\ell$ to $\mathbb{R}^n$. On the other hand, the expression inside the square brackets is negligible. We conclude that $f \circ g$ is differentiable at $x^0$ and its derivative is given by the following formula:

$$(f \circ g)'(x^0) = f'(g(x^0)) \circ g'(x^0) \quad (17)$$

This is the general form of the Chain Rule. Here is an equivalent statement of the Chain Rule in terms of Jacobi’s matrices:
\[
J_{f \circ g}(x^0) = J_f(g(x^0)) J_g(x^0) \quad .
\]  \hspace{1cm} (18)

7. Let us take a look, for example, at the case \( \ell = m = 2 \) and \( n = 1 \). Let
\[
g(s, t) = \begin{pmatrix} g_1(s, t) \\ g_2(s, t) \end{pmatrix}
\]
and \( f \) be a function of two variables \( x \) and \( y \). Jacobi’s matrix of \( g \) at \((s^0, t^0)\) is
\[
J_g(s^0, t^0) = \begin{pmatrix}
\frac{\partial g_1}{\partial s}(s^0, t^0) & \frac{\partial g_1}{\partial t}(s^0, t^0) \\
\frac{\partial g_2}{\partial s}(s^0, t^0) & \frac{\partial g_2}{\partial t}(s^0, t^0)
\end{pmatrix}
\]  \hspace{1cm} (19)
and Jacobi’s matrix of \( f \) at point \((x^0, y^0) = (g_1(s^0, t^0), g_2(s^0, t^0))\) is
\[
J_f(x^0, y^0) = \begin{pmatrix}
\frac{\partial f}{\partial x}(x^0, y^0) & \frac{\partial f}{\partial y}(x^0, y^0)
\end{pmatrix}
\]  \hspace{1cm} (20)

Jacobi’s matrix of \( f \circ g \) at \((s^0, t^0)\) is, according to Chain Rule (18), equal to the product of (20) and (19):
\[
J_{f \circ g}(s^0, t^0) = J_f(x^0, y^0) J_g(s^0, t^0)
\]
\[
= \left( \frac{\partial f}{\partial x} \frac{\partial g_1}{\partial s} + \frac{\partial f}{\partial y} \frac{\partial g_2}{\partial s} + \frac{\partial f}{\partial x} \frac{\partial g_1}{\partial t} + \frac{\partial f}{\partial y} \frac{\partial g_2}{\partial t} \right)
\]
\[
= \left( f_x (g_1)_s + f_y (g_2)_s, f_x (g_1)_t + f_y (g_2)_t \right)
\]  \hspace{1cm} (21)
where \( f_x = \frac{\partial f}{\partial x} \) and \( f_y = \frac{\partial f}{\partial y} \) are taken at point \((x^0, y^0) = (g_1(s^0, t^0), g_2(s^0, t^0))\)
whereas \((g_k)_s = \frac{\partial g_k}{\partial s}\) and \((g_k)_t = \frac{\partial g_k}{\partial t}\) are taken at point \((s^0, t^0)\).

We can rewrite formula (21) in terms of the gradient vectors (12)
\[
\nabla(f \circ g)(s^0, t^0) = J_g^T(s^0, t^0) \nabla f(x^0, y^0)
\]  \hspace{1cm} (22)
or, in an abbreviated form:

\[
\nabla(f \circ g) = J_g^T \nabla f.
\]

(23)

Here \( J^T \) denotes the transpose of the matrix \( J \):

\[
\text{if } J = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ then } J^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.
\]

(24)

One of the basic properties of the transposition of matrices is that \((AB)^T = B^T A^T\).

(Please, verify that!)

Exercise 6. Derive the following special case of Chain Rule (18):

\[
\nabla f(\gamma(t^0)) \cdot \gamma'(t^0) = \frac{d(f \circ \gamma)}{dt}(t^0)
\]

(25)

for \( f : D \to \mathbb{R} \) and a parametric curve \( \gamma : (a, b) \to D \). (Here \( D \) is a subset of \( \mathbb{R}^m \).)

8. Tangent plane to surface \( z = f(x, y) \). The plane in \( \mathbb{R}^3 \) which is tangent at point \((x^0, y^0, f(x^0, y^0))\) to surface \( z = f(x, y) \) is the graph of the linear function which sends a vector \( \mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \) to the number

\[
f(\mathbf{v}^0) + (d_{x^0} f)(\mathbf{v} - \mathbf{v}^0)
\]

\[
= f(\mathbf{v}^0) + \nabla f(x^0) \cdot (\mathbf{v} - \mathbf{v}^0)
\]

\[
= f(x^0, y^0) + \frac{\partial f}{\partial x}(x^0, y^0)(x - x^0) + \frac{\partial f}{\partial y}(x^0, y^0)(y - y^0)
\]

(26)

where \( \mathbf{v}^0 = \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} \). Formula (26) gives us the equation of the tangent plane at \((x^0, y^0, f(x^0, y^0))\):

\[
f_x(x^0, y^0)(x - x^0) + f_y(x^0, y^0)(y - y^0) - (z - z^0) = 0
\]

(27)
or, in a vector form,

$$\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}^0) = 0$$  \hspace{1cm} (28)

where \( \mathbf{r} = (x, y, z) \), \( \mathbf{r}^0 = (x^0, y^0, f(x^0, y^0)) \) and \( \mathbf{n} = (f_x(x^0, y^0), f_y(x^0, y^0), -1) \), to use Stewart’s notation.

9. Directional derivative \( D_{\mathbf{u}}f \). Let \( f \) be a function from a subset \( D \) of \( \mathbb{R}^m \) to \( \mathbb{R}^n \). The directional derivative of \( f \) at point \( \mathbf{x}^0 \) in the direction of vector

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$$  \hspace{1cm} (29)

is defined as

$$D_{\mathbf{u}}f(\mathbf{x}^0) = \left. \frac{df(\mathbf{x}^0 + t\mathbf{u})}{dt} \right|_{t=0}.$$  \hspace{1cm} (30)

Note that \( f(\mathbf{x}^0 + t\mathbf{u}) = (f \circ \gamma)(t) \) where \( \gamma(t) = \mathbf{x}^0 + t\mathbf{u} \) is a linear function from \( \mathbb{R} \) to \( \mathbb{R}^m \). Thus, the directional derivative \( D_{\mathbf{u}}f(\mathbf{x}^0) \) is the velocity vector of parametric curve \( f \circ \gamma \).

Function \( \gamma \) is a parametrization of the straight line passing through point \( \mathbf{x}^0 \) with the constant velocity \( \mathbf{u} \). Its derivative at \( t = 0 \) (and everywhere else) is \( \mathbf{u} \), see (8) above. Thus Chain Rule (17) gives us

$$D_{\mathbf{u}}f(\mathbf{x}^0) = f'_\mathbf{x}(\mathbf{u}) = J_f(\mathbf{x}^0) \mathbf{u}$$  \hspace{1cm} (31)

Identity (31) has very beautiful applications.

If \( \mathbf{u} \) is tangent to the level set of \( f \) at \( \mathbf{x}^0 \):

$$\{ \mathbf{x} \in D \mid f(\mathbf{x}) = f(\mathbf{x}^0) \}$$  \hspace{1cm} (32)

then \( D_{\mathbf{u}}f(\mathbf{x}^0) = 0 \) and hence

the derivative of \( f \) at \( \mathbf{x}^0 \) vanishes on vectors tangent to level set (32)  \hspace{1cm} (33)

If \( n = 1 \), formula (31) reads as follows:
\[ D_u f(x^0) = \nabla f(x^0) \cdot u = \sum_{j=1}^{m} \frac{\partial f(x^0)}{\partial x_j} u_j = 0. \]  

(34)

In other words,

the gradient vector of \( f \) at \( x^0 \) is orthogonal to level set (32) 

(35)

Vice-versa, among vectors of the same length

the directional derivative of \( x^0 \) attains the largest value on vectors orthogonal to level set (32) 

(36)

10. **Critical points of a function** \( f : D \to \mathbb{R} \) of two variables. Suppose that \( f : D \to \mathbb{R} \) is differentiable at a point \( r^0 = \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} \). We shall say that \( r^0 \) is a critical point of \( f \) if \( d_r f = 0 \). We mean by this that the linear functional \( d_r f \) is zero, i.e.

\[ d_r f (v) = \nabla f(r^0) \cdot v = f_x(r^0) v_1 + f_y(r^0) v_2 = 0 \] 

(37)

for every vector \( v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \) in \( \mathbb{R}^2 \). Equality (37) is satisfied for every \( v \) exactly when

\[ f_x(r^0) = f_y(r^0) = 0 \] 

(38)

or, equivalently, when the gradient of \( f \) vanishes at \( r^0 \).

One expects that the type of a given critical point should be related to the second derivative of \( f \) at \( r^0 \). What does this second derivative look like in our case?

Suppose \( f \) is differentiable at every point \( r \) of \( D \). The first derivative of \( f \) at \( r \), which is called the differential of \( f \) at \( r \), becomes a function

\[ df : D \to \{ \text{linear functionals on } \mathbb{R}^2 \}. \] 

(39)
Any such function is called a **differential form** on $D$.

**Example 1.** The derivative of $f(x, y) = x$ is denoted $dx$. Note that $f_x(r) = 1$ and $f_y(r) = 0$ for all $r \in \mathbb{R}^2$, hence

$$d_x x (v) = v_1 \quad (v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix})$$

and you observe that the value does not depend on $r$. Thus $dx$ is an example of a **constant** differential form.

**Exercise 7.** Define differential form $dy$. Find $d_x y (v)$. Does it depend on $r \in \mathbb{R}^2$?

**Exercise 8.** Express $df$ in the following form

$$df = \frac{\partial f}{\partial x} \, dx + \frac{\partial f}{\partial y} \, dy \quad .$$

*Hint.* Use identity (31), in the case $n = 1$, together with identity (40) and the last exercise.

The space of linear functionals on $\mathbb{R}^2$ can be itself identified with $\mathbb{R}^2$; see Exercise 5 above. Under this identification $f_x'$ corresponds, of course, to gradient vector $\nabla f(r)$. This is, after all, the main reason why we bothered to introduce $\nabla f(r)$ in the first place!

Having made this identification, we are dealing now with the gradient vector function

$$\nabla f : D \to \mathbb{R}^2$$

instead of differential (39). Its derivative $(\nabla f)'(r^0)$ at $r^0$ is thus a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$. Let us calculate its matrix:

$$\left( \begin{array}{cc} \frac{\partial f}{\partial x} (f_x) (r^0) & \frac{\partial f}{\partial y} (f_x) (r^0) \\ \frac{\partial f}{\partial x} (f_y) (r^0) & \frac{\partial f}{\partial y} (f_y) (r^0) \end{array} \right) \left( \begin{array}{c} \frac{\partial^2 f}{\partial x^2} (r^0) \\ \frac{\partial^2 f}{\partial x \partial y} (r^0) \\ \frac{\partial^2 f}{\partial y^2} (r^0) \end{array} \right) = \left( \begin{array}{cc} f_{xx} (r^0) & f_{yx} (r^0) \\ f_{yx} (r^0) & f_{yy} (r^0) \end{array} \right)$$

(43)

If $f_{xy}$ and $f_{yx}$ are **continuous** at $r^0$ then they are equal (**Clairaut’s Theorem**). Therefore
the matrix of the derivative of the gradient function (43) is symmetric. (A matrix \( A = (a_{ij}) \) is symmetric if \( a_{ij} = a_{ji} \) for all \( i \) and \( j \); a symmetric matrix must be a square matrix.)

We shall call

\[
\begin{pmatrix}
  f_{xx}(r^0) & f_{yx}(r^0) \\
  f_{xy}(r^0) & f_{yy}(r^0)
\end{pmatrix}
\]

(44)

the **Hesse matrix** of a function \( f : D \to \mathbb{R} \) at a point \( r^0 \). The determinant of (44)

\[
H_f(r^0) = \begin{vmatrix}
  f_{xx}(r^0) & f_{yx}(r^0) \\
  f_{xy}(r^0) & f_{yy}(r^0)
\end{vmatrix}
\]

(45)

is called the **Hessian** of \( f \) at \( r^0 \).

This concept was introduced for the first time by Ludwig Otto Hesse (1811-1874) in two articles published in 1844 and 1851, respectively. (I do not encourage you to follow Stewart’s notation for Hessian, since it is in conflict with his own notation for the domain of \( f \) and for the directional derivative (30).)

Hessian provides very important information about critical points. If \( r^0 \) is a critical point of \( f \), i.e. \( d_v f = 0 \), then there are the following possibilities.

(i) If \( H_f(r^0) < 0 \), then \( r^0 \) is a **saddle point**;

(ii) If \( H_f(r^0) > 0 \), then there are two further possibilities:

a) \( r^0 \) is a **local minimum** if \( f_{xx}(r^0) > 0 \),

b) \( r^0 \) is a **local maximum** if \( f_{xx}(r^0) < 0 \).

Note that the positivity of \( H_f(r^0) = f_{xx} f_{yy} - (f_{xy})^2 \) requires that \( f_{xx} \) and \( f_{yy} \) have the same sign! Hence one can replace \( f_{xx} \) by \( f_{yy} \) in conditions ii.a) and ii.b) above.

The above three cases exhaust all the possibilities that can occur when the Hessian \( H_f(r^0) \) does not vanish. If \( H_f(r^0) = 0 \) then \( r^0 \) is called a **degenerate** critical point and the situation becomes a lot more complicated in general.
Differential Calculus of Vector Functions

One thing worth remembering: **The Hessian classification of critical points is applicable only at points where \( \nabla f \) is differentiable and \( f_{xy} = f_{yx} \).**

**Example 2.** Let \( f(x, y) = x^2 + 3xy + 2y^2 \). The differential of \( f \) equals (see Exercise 8 above)
\[
df = \frac{\partial f}{\partial x} \, dx + \frac{\partial f}{\partial y} \, dy = (2x + 3y) \, dx + (3x + 4y) \, dy
\]
or, equivalently, the gradient of \( f \) equals
\[
\nabla f = \begin{pmatrix} 2x + 3y \\ 3x + 4y \end{pmatrix} = (2x + 3y) \, \mathbf{i} + (3x + 4y) \, \mathbf{j}.
\]
A point \( (x, y) \) is a critical point of \( f \) if and only if
\[
\begin{cases}
2x + 3y = 0 \\
3x + 4y = 0
\end{cases}
\]  \( (46) \)
The only solution to \( (46) \) is \( x = y = 0 \), i.e., the origin is the only critical point of \( f \).

Hesse’s matrix \( (44) \) of \( f \) does not depend on \( \mathbf{p}^0 \) and equals
\[
\begin{pmatrix}
2 & 3 \\
3 & 4
\end{pmatrix},
\]
Therefore, the Hessian of \( f \) at the origin equals \( 2 \cdot 4 - 3^2 = -1 < 0 \) and it follows that \( f \) has a saddle point at \( \mathbf{0} \). Note, however, that both \( f(x, 0) = x^2 \) and \( f(0, y) = 2y^2 \) at \( \mathbf{0} \) have a minimum!

**11. Another look at the definition of a critical point.** In the opening paragraph of Section 10 we declare a point \( \mathbf{x}^0 \) to be **critical** for a function \( f : D \to \mathbb{R} \) if the differential of \( f \) at \( \mathbf{x}^0 \) vanishes identically:
\[
d_{x^0} f = 0, \tag{47}
\]
i.e. \( \nabla f(\mathbf{x}^0) = 0 \).

**Exercise 9.** Show that every linear transformation \( L : \mathbb{R}^m \to \mathbb{R} \) is either zero or onto.
Differential $d_x f$ is a linear transformation $\mathbb{R}^m \to \mathbb{R}$. So, if $x^0$ is \textit{not} a critical point of $f$ (we call such points \textit{regular}), then $d_x f$ maps $\mathbb{R}^m$ \textit{onto} $\mathbb{R}$. And vice-versa:

\begin{equation} \text{a point } x^0 \text{ is critical for a function } f : D \to \mathbb{R} \text{ if and only if } d_x f : \mathbb{R}^m \to \mathbb{R} \text{ is not onto.} \end{equation} 

Armed with this important observation, we now proceed to discuss critical points of vector valued functions.

\textbf{12. Critical points of vector functions } $f : D \to \mathbb{R}^n$. When is the image of a linear transformation $L : \mathbb{R}^m \to \mathbb{R}^n$ as big as possible? When $L$ is \textit{onto}, of course. Yes, but this is possible only when $m \geq n$. For $m \leq n$, $L$ will have the biggest possible image when $L$ is one-to-one.

This observation, combined with our deepened understanding of what a critical point is (see display (48) above), leads us to the following definition.

\begin{equation} \text{A point } x^0 \text{ is a \textbf{regular} point of a vector function } f : D \to \mathbb{R}^n \text{ if:} \\
\text{Case } m \geq n. \quad f'(x^0) : \mathbb{R}^m \to \mathbb{R}^n \text{ is \textit{onto}.} \\
\text{Case } m \leq n. \quad f'(x^0) : \mathbb{R}^m \to \mathbb{R}^n \text{ is\textit{ one-to-one}.} \end{equation}

Note that these two cases overlap when $m = n$. There is no conflict, however, since a linear transformation $L : \mathbb{R}^m \to \mathbb{R}^m$ is \textit{onto} precisely when it is \textit{one-to-one}.

\[\text{We say that } x^0 \text{ is a \textbf{critical} point if } x^0 \text{ not regular.}\]

General remark: the terminology \textit{regular point, critical point} applies only to points where a function is differentiable (contrary to what Stewart says in §12.7, p. 703).

Let me remind you what have we established in Section 9: the derivative $f'(x)$ of a function $f : D \to \mathbb{R}^n$ vanishes on vectors tangent to the level set of $f$ at point $x$. This holds for any point $x$. However, for points where $f$ is \textit{regular} the reverse is also true.
If \( \mathbf{x} \) is a regular point of a function \( f : D \to \mathbb{R}^n \) then the level set of \( f \) passing through \( \mathbf{x} \) is smooth in the vicinity of \( \mathbf{x} \). Moreover, \( f'(\mathbf{x})(\mathbf{v}) = 0 \) if and only if vector \( \mathbf{v} \) is tangent to the level set of \( f \).

The above statement is among the most important in Multivariable Calculus. Think of it as being the principal reason why you are learning about regular points. Another reason is the role regularity plays in the Lagrange Multipliers method (Section 17 below).

13. Some comments and additions to Theorem (50). Tangent vectors to the level set at a regular point form an \((m - n)\)-dimensional space in \( \mathbb{R}^m \) if \( m \geq n \). This contrasts with the case \( m \leq n \), when the level sets of regular points consist of isolated points. In particular, no non-zero vectors are tangent to such level sets, and therefore Theorem (50) does not say much in this case. One can show, indeed, that when restricted to a sufficiently small neighbourhood of a regular point \( \mathbf{x} \), the function \( f \) becomes one-to-one — exactly like its derivative \( f'(\mathbf{x}) \). All of this forms a basis of a more advanced Multivariable Calculus. You should make your goal to learn this later — after you become familiar with elements of Linear Algebra — it is a fascinating subject and its applications are unlimited!

14. Regularity in some special cases. You already know the meaning of regularity when \( n = 1 \):

point \( \mathbf{x} \) is a regular point of \( f : D \to \mathbb{R} \) if and only if \( \nabla f(\mathbf{x}) \neq 0 \).  \hfill (51)

What about the case \( n = 2 \)? In this case \( f = \langle f_1, f_2 \rangle \) and, assuming that \( m \), i.e. the number of variables, is greater than 1, the answer is as follows.

Point \( \mathbf{x} \) is a regular point of a function \( f : D \to \mathbb{R}^2 \) if and only if the gradient vectors of its component functions \( \nabla f_1(\mathbf{x}) \) and \( \nabla f_2(\mathbf{x}) \) span a plane in \( \mathbb{R}^m \).  \hfill (52)

If they do not — the point is critical. This happens either because gradient vectors \( \nabla f_1(\mathbf{x}) \) and \( \nabla f_2(\mathbf{x}) \) are collinear or, in the most degenerate case, because they both vanish.

In the familiar case of a parametric curve \( \gamma : (a, b) \to \mathbb{R}^n \), the regular points are numbers \( t \in (a, b) \) where the velocity vector \( \gamma'(t) \), introduced in (8), does not vanish. Accord-
ingly, the **critical** points are precisely those numbers \( t \) for which the velocity vector \( \gamma'(t) \) vanishes. Recall that only at such points the curve parametrized by the function \( \gamma \) can have nonsmooth irregularities like “nodes” or “corners”.

15. **Local extrema of a function** \( f : D \to \mathbb{R} \) **along a parametric curve.** Consider a curve parametrized by a function \( \gamma : (a, b) \to D \). We shall say that a function \( f : D \to \mathbb{R} \) has at a point \( x^0 = \gamma(t^0) \) a local maximum (minimum) along a parametric curve \( \gamma \) if the composite function

\[
f \circ \gamma : (a, b) \to \mathbb{R}
\]

has a local maximum (respectively, minimum) at \( t^0 \). In this case, Fermat’s Theorem (Stewart §4.1, p. 257) tells us that the derivative of \( f \circ \gamma \) at \( t^0 \) vanishes and we deduce from Chain Rule (17) — see also Exercise 6 and formula (25) — that

\[
d_{\gamma(t^0)}f \text{ annihilates the velocity vector } \gamma'(t^0), \text{ i.e. } \nabla f(x^0) \cdot \gamma'(t^0) = 0.
\]

(54)

In other words, gradient \( \nabla f(x^0) \) and the velocity vector \( \gamma'(t) \) are **orthogonal** to each other.

16. **Local extrema of a function** \( f : D \to \mathbb{R} \) **on a subset** \( Z \) **of** \( D \). Very often one has to find the maximum or the minimum value that a function \( f \) can take on a given subset \( Z \) of its domain \( D \). From (54) we know that if \( \gamma : (a, b) \to Z \) is **any** differentiable curve passing through a point \( x^0 = \gamma(t^0) \) — where a function \( f \) has its local maximum or minimum on \( Z \) — then differential \( d_{x^0}f \) annihilates the velocity vector \( \gamma'(t^0) \).

Now, any vector tangent to \( Z \) at point \( x^0 \) occurs as the velocity vector of some parametric curve passing through it. Hence we arrive at the following generalization of **Fermat’s Theorem.**

\[
\text{If a function } f \text{ has a local extremum on } Z \text{ at a point } x^0 \text{ then } d_{x^0}f \text{ vanishes on all vectors tangent to } Z \text{ at the point } x^0. \tag{55}
\]

Note that Theorem (55) covers also the case when \( Z \) is the **whole** set \( D \). If \( x^0 \) is an **interior** point of \( D \) then **any** vector \( v \in \mathbb{R}^m \) is tangent to \( D \) at \( x^0 \). Thus, Theorem (55)
has the following corollary.

If \( f \) has a local extremum at an interior point \( x^0 \) then \( d_{x^0} f \) is zero, i.e. \( x^0 \) is a critical point of the function \( f \).

\[(56)\]

17. **Lagrange multipliers.** Now, a practical application of great importance. Suppose that you must find extrema of a function \( f : D \to \mathbb{R} \) where the argument \( x \) is subject to a number of side conditions:

\[ g_1(x) = k_1, \ldots, g_r(x) = k_r \]

called constraints (functions \( g_1, \ldots, g_r \) and numbers \( k_1, \ldots, k_r \) are given in advance). The first thing you should do is to rewrite \( r \) constraints (57) as a single vector constraint:

\[ g(x) = \mathbf{K} \]

(58)

where \( g(x) = (g_1(x), \ldots, g_r(x)) \) and \( \mathbf{K} = (k_1, \ldots, k_r) \). Denote by \( Z \) the corresponding level set of the vector function \( g \):

\[ Z = \{ x \in D \mid g(x) = \mathbf{K} \} . \]

(59)

Theorem (55) tells us that \( d_{x^0} f \) vanishes on vectors tangent to \( Z \) at point \( x^0 \) if the function \( f \) has a local extremum on \( Z \) at \( x^0 \). If \( x^0 \) is a regular point of the vector-constraint function \( g \) then its derivative \( g'(x^0) \) vanishes precisely on vectors tangent to \( Z \).

Now, the derivative \( g'(x^0) \) is a linear transformation from \( \mathbb{R}^m \) to \( \mathbb{R}^r \) and differential \( d_{x^0} f \) is a linear transformation from \( \mathbb{R}^m \) to \( \mathbb{R} \). Since \( g'(x^0) \) vanishes only on those vectors on which \( d_{x^0} f \) vanishes, one can “divide” transformation \( d_{x^0} f \) by transformation \( g'(x^0) \).

The exact meaning of this phrase is: there exists a (unique) transformation \( \Lambda \) from \( \mathbb{R}^r \) to \( \mathbb{R} \) such that \( d_{x^0} f \) is the composition of \( \Lambda \) and \( g'(x^0) \):

\[ d_{x^0} f = \Lambda \circ g'(x^0) . \]

(60)

The form of any linear operator from \( \mathbb{R}^r \) to \( \mathbb{R} \) is given by equality (9), as you already know (see Exercise 5). In our case, this means that

\[ \Lambda(v) = \lambda \cdot v \quad (v \in \mathbb{R}^r) \]

(61)
for a suitable vector \( \lambda = \langle \lambda_1, ..., \lambda_r \rangle \).

\[ \nabla f(x^0) = \lambda_1 \nabla g_1(x^0) + \cdots + \lambda_r \nabla g_r(x^0) \]  \hspace{1cm} (62)

Exercise 10. Verify that equality (61) can be rewritten as follows:

Coefficients \( \lambda_1, ..., \lambda_r \) are called **Lagrange multipliers**. To sum up, we have established the following remarkable theorem which is the essence of the Lagrange multipliers method.

At any point \( x^0 \) where the function \( f \) has a local extremum with \( r \) constraints (57), equality (62) holds for suitable numbers \( \lambda_1, ..., \lambda_r \) provided \( x^0 \) is a regular point of the vector-constraint function \( g(x) = \langle g_1(x), \ldots, g_r(x) \rangle \).

Theorem (63) holds for any values of \( m \) and \( r \). In practice, its usefulness for finding constrained extrema of \( f \) is limited only to situations when the number of constraints is less than \( m \). Here is the reason: if \( r \geq m \) then the level sets of all regular points of \( g \) reduce to isolated points. In this case, one simply checks the values of the function \( f \) at those isolated points that satisfy constraints (57).

Finally, you should be always prepared that there may be no points satisfying given constraints, in which case level set (59) is empty. When this happens then there is no point, of course, in trying to find corresponding constrained extrema of the function \( f \).