Differential Calculus of Vector Functions

October 9, 2003

These notes should be studied in conjunction with lectures.

1 Continuity of a function at a point
Consider a function \( f : D \rightarrow \mathbb{R}^n \) which is defined on some subset \( D \) of \( \mathbb{R}^m \). Let \( a \) be a point of \( D \). We shall say that \( f \) is continuous at \( a \) if

\[
f(x) \text{ tends to } f(a) \quad \text{whenever} \quad x \text{ tends to } a.
\]

If function \( f \) is continuous at every point of its domain, then we simply say that \( f \) is continuous.

Exercise 1 Any linear transformation is continuous. Show this using inequality (34) in Prelim.

2 Differentiability of a function at a point
Now, let \( a \) be an interior point of \( D \). We shall say that \( f \) is differentiable at \( a \) if there exists a linear transformation \( L : \mathbb{R}^m \rightarrow \mathbb{R}^n \) such that

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

where \( u(x) \) is negligible, compared to \( \text{dist}(x, a) \), when \( x \to a \). “Negligible” means that \( \|u(x)\| \) approaches 0 faster than \( \text{dist}(x, a) \) does, i.e., that

\[
\lim_{x \to a} \frac{\|u(x)\|}{\text{dist}(x, a)} = \lim_{x \to a} \frac{\|u(x)\|}{\|x - a\|} = 0.
\]

If such a linear transformation \( L \) exists then \( L \) is unique. It will be denoted \( f'(a) \) and called the derivative of \( f \) at \( a \) and thus (2) can be rewritten as

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

Abbreviations Prelim and Problembook stand for Preliminaries and Problembook, respectively.

A point \( a \) is an interior point of a set \( D \) if \( D \) contains some ball with center at \( a \).
where \( u(x) \) is negligible when \( x \) approaches \( a \).

In the interest of keeping notation as transparent as possible, we shall be denoting \( f'(a) \) also \( f'_a \). For example, in this alternate notation \((f'(a))(v)\) becomes \( f'_a(v) \) (which uses one instaed of three pairs of parentheses).

3  If \( f \) is differentiable at a point \( a \), then it is also continuous at \( a \). This follows from the following estimate for the distance between \( f(x) \) and \( f(a) \):

\[
\|f(x) - f(a)\| = \|(f'(a))(x - a) + u(x)\| \\
\leq \|(f'(a))(x - a)\| + \|u(x)\| \quad \text{(Triangle Inequality, cf. Sect. 6 of Prelim)} \\
\leq \|f'(a)\| \|x - a\| + \|u(x)\| \quad \text{(inequality (34) in Prelim).} \tag{5}
\]

4  Basic properties of the derivative  The following properties follow directly from the definition given in Section 2:

a) if \( f: D \to \mathbb{R} \) and \( g: D \to \mathbb{R} \) are differentiable at a point \( a \) then so is their sum \( f + g \) and

\[
(f + g)'(a) = f'(a) + g'(a) \; ; \tag{6}
\]

b) for any scalar \( c \in \mathbb{R} \), one has \((cf)(a) = cf'(a) \); 

c) if \( f \) is a linear transformation then \( f'(a) = f \) for all \( a \).

5  Partial derivatives  Consider a scalar-valued function \( f: D \to \mathbb{R} \), where \( D \subseteq \mathbb{R}^m \), and a point \( a \in D \). Let \( j \) be any integer between 1 and \( m \). The partial derivative

\[
\frac{\partial f}{\partial x_j}(a) \tag{7}
\]

is defined as the ordinary derivative

\[
\frac{d\phi_j}{dt} \bigg|_{t=a_j} \tag{8}
\]
of the function of single real variable

$$\phi_j(t) := f\left( \begin{pmatrix} a_1 \\ \vdots \\ t \\ \vdots \\ a_m \end{pmatrix} \right)$$

obtained by freezing all but the $j$-th coordinate of a variable point $x \in D$.

Note that function $\phi_j$ is the composite $f \circ \gamma_j$ of $f$ and the parametric curve $\gamma_j: \mathbb{R} \to \mathbb{R}^m$,

$$\gamma_j(t) := \begin{pmatrix} a_1 \\ \vdots \\ t \\ \vdots \\ a_m \end{pmatrix} = a + (t - a_j)e_j.$$

6 Theorem The $n \times m$ matrix corresponding to the linear transformation $f'(a): \mathbb{R}^m \to \mathbb{R}^n$ is formed by partial derivatives of components of $f$:

$$\frac{\partial f_1}{\partial x_1}(a) \quad \ldots \quad \frac{\partial f_1}{\partial x_m}(a) \\
\vdots \\
\frac{\partial f_n}{\partial x_1}(a) \quad \ldots \quad \frac{\partial f_n}{\partial x_m}(a)$$

Here $f(x) = \begin{pmatrix} f_1(x) \\ \ldots \\ f_n(x) \end{pmatrix}$; each component $f_i$ of $f$ is a scalar-valued function $D \to \mathbb{R}$.

Matrix (11) is called the Jacobi\(^3\) matrix of $f$ at $a$ and will be denoted $J_f(a)$. \(^4\)

\(^3\)Carl Gustav Jacob Jacobi (1804–1851). He was equally good at each of the three greatest subjects of all: Greek, Latin and Mathematics. In May 1832 he was promoted to full professor after being subjected to a four hour disputation in Latin.

\(^4\)We shall prove Theorem 6 in Section 13 below.
Exercise 2  Rewrite (4) in terms of Jacobi’s matrix (11).

Hint: Use formula (26) of Prelim.

7 Functions of class $C^1$  A subset $D \subseteq \mathbb{R}^m$ is said to be open if every point $a \in D$ is an interior point of $D$.

We say, in this case, that a function $f : D \to \mathbb{R}^n$ is of class $C^1$ if partial derivatives

$$\frac{\partial f_i(a)}{\partial x_j}(1 \leq i \leq n, 1 \leq m)$$

exist at all points $a \in D$ and are continuous as functions of $a$.

8 Theorem  A function of class $C^1$ on $D$ is differentiable at every point of $D$.

As a corollary, we obtain the following useful criterion.

9 Criterion of differentiability  A function $f : D \to \mathbb{R}^n$ is differentiable at a point $a$ if it is of class $C^1$ on some neighborhood of $a$, i.e., on some open ball

$$B_r(a) := \{x \in \mathbb{R}^m \mid \text{dist}(x,a) < r\}. \quad (12)$$

10 The case of a parametric curve $\gamma(t)$ in $\mathbb{R}^n$  Any continuous function $\gamma : I \to \mathbb{R}^n$, where $I$ is a subset of real line $\mathbb{R}$, will be called a parametric curve in $\mathbb{R}^n$. By abuse of language, we shall say that a curve $\gamma$ is contained in a subset $Z \subseteq \mathbb{R}^n$ if $\gamma(t) \in Z$ for all $t \in I$.

A particularly important case occurs when $I$ is an interval of the real line. A curve parametrized by an interval will be called a path.

For a parametric curve $\gamma$, derivative $\gamma'(a)$ is a linear transformation $\mathbb{R} \to \mathbb{R}^n$.

Any linear transformation $\mathbb{R} \to \mathbb{R}^n$ is of the form $t \mapsto$ at for a suitable column-vector $a$. In the case of linear transformation $\gamma'(a) : \mathbb{R} \to \mathbb{R}^n$ that vector happens to be the velocity
Figure 1: A parametric curve contained in the unit sphere in $\mathbb{R}^3$:

$$\gamma: \mathbb{R} \rightarrow \mathbb{R}^3, \quad \gamma(\theta) = \frac{1}{\sqrt{1 + \theta^2}} \begin{pmatrix} \cos \theta \\ \sin \theta \\ \theta \end{pmatrix}$$

vector of the parametric curve:

$$\frac{d\gamma}{dt}(a) := \begin{pmatrix} \frac{d\gamma_1(a)}{dt} \\ \vdots \\ \frac{d\gamma_n(a)}{dt} \end{pmatrix}.$$ (13)

This is Jacobi’s matrix of $\gamma$. It has one column because $m = 1$. Note that the velocity vector is just the value of linear transformation $\gamma'(a) = \gamma'_a$ at 1:

$$\frac{d\gamma}{dt}(a) = \gamma'_a(1).$$

11 The case of a scalar-valued function of $m$ variables $f: D \rightarrow \mathbb{R}$ A scalar-valued function of $m$ scalar variables

$$(x_1, \ldots, x_m) \mapsto f(x_1, \ldots, x_m)$$ (14)
is best viewed as a function $f: D \to \mathbb{R}$ defined on some suitable subset $D \subseteq \mathbb{R}^m$. In this case, we use the notation $f(x)$, instead of $f(x_1, \ldots, x_m)$, where

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

is the corresponding point of $\mathbb{R}^m$.

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

$$\left( \frac{\partial f}{\partial x_1}(a) \; \ldots \; \frac{\partial f}{\partial x_m}(a) \right)$$

and

$$df_a(v) = \nabla f(a) \cdot v$$

where $\nabla f(a)$ is the column-vector:

$$\nabla f(a) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(a) \\ \vdots \\ \frac{\partial f}{\partial x_m}(a) \end{pmatrix}.$$  \hfill (17)

Vector (17) is called the gradient of $f$ at $a$. Note that it is the transpose of Jacobi's matrix (15).

In the case of a function $f: D \to \mathbb{R}$, formula (4) becomes

$$f(x) = f(a) + df_a(x - a) + u(x)$$

$$= f(a) + \nabla f(a) \cdot (x - a) + u(x)$$  \hfill (18)

where $u(x)$ is negligible when $x$ approaches $a$. 
12 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 \( a \) and that \( f \) is differentiable at \( b = g(a) \).
In other words:
\[ g(x) - g(a) = g'(a)(x - a) + u(x) \] (19)
and
\[ f(y) - f(b) = f'(b)(y - b) + v(y) \] (20)
where \( u(x) \) and \( v(y) \) are negligible when \( x \to a \) and \( y \to b \), respectively.
Plug \( y = g(x) \) and \( b = g(a) \) into (20) and use identity (19):
\[
f(g(x)) - f(g(a)) = f'(g(a))(g(x) - g(a)) + v(g(x)) \\
= f'(g(a))(g'(a)(x - a) + u(x)) + v(g(x)) \] (21)
\[ = (f'(g(a)) \circ g'(a))(x - a) + [f'(g(a))(u(x)) + v(g(x))] \\
\]
The composition of two linear transformations is linear. Therefore \( f'(g(a)) \circ g'(a) \) 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 \( a \) and its derivative is given by the following formula:
\[
(f \circ g)'(a) = f'(g(a)) \circ g'(a) \] (22)

This is the general form of the Chain Rule. Here is an equivalent statement of the Chain Rule in terms of Jacobi’s matrices:
\[
J_{f \circ g}(a) = J_f(g(a))J_g(a) \] (23)
Exercise 3  Explain why the expression inside the square brackets in the last row of (21) is negligible.

Hint: use identity (19) in conjunction with inequality (34) from Prelim.

13 As an application of the Chain Rule we shall now prove Theorem 9. Consider the following two simple yet very useful linear transformations:

\[ \epsilon_j : \mathbb{R} \to \mathbb{R}^m, \quad \epsilon_j(t) := t e_j, \quad (24) \]

and

\[ \pi_i : \mathbb{R}^n \to \mathbb{R}, \quad \pi_i \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} := v_i. \quad (25) \]

Exercise 4  Let \( L : \mathbb{R}^m \to \mathbb{R}^n \) be a linear transformation with matrix \( A \), cf. (26) in Prelim. Verify that the composite transformation \( \pi_i \circ L \circ \epsilon_j : \mathbb{R} \to \mathbb{R} \) has the form

\[ t \mapsto a_{ij} t \quad (t \in \mathbb{R}). \]

The \( i \)-th component \( f_i : \mathbb{R}^m \to \mathbb{R} \) of \( f \) is the composite \( \pi_i \circ f \), and we know that partial derivative \( \frac{\partial f_i}{\partial x_j}(a) \) is the ordinary derivative of the composite function \( f_i \circ \gamma_j \), cf. Section 5. Chain Rule gives us

\[ (\pi_i \circ f \circ \gamma_j)'(a_j) = \pi_i'(f(a)) \circ f'(a) \circ \gamma_j'(a_j). \quad (26) \]

Now, \( \pi_i \) is linear, hence \( (\pi_i)'(f(a)) = \pi_i \). On the other hand, \( \gamma_j = a - a_j e_j + \epsilon_j \), as follows from (10). Using the basic properties of the derivative we thus get \( \gamma'_j(a) = \epsilon_j \).

By plugging this into (26), we obtain the following equality of linear transformations \( \mathbb{R} \to \mathbb{R} \):

\[ (\pi_i \circ f \circ \gamma_j)'(a_j) = \pi_i \circ f'(a) \circ \epsilon_j. \quad (27) \]

The left-hand side of (27) multiplies \( t \in \mathbb{R} \) by \( \frac{\partial f_i}{\partial x_j}(a) \), while the right-hand side multiplies \( t \) by entry \( a_{ij} \) of the matrix corresponding to derivative \( f'(a) \). Therefore these two numbers must be equal, as asserted in Theorem 9.
Let us take a closer look, for example, at the case $\ell = m = 2$ and $n = 1$. Let
\[ g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} : E \to \mathbb{R}^2 \]
be a function from a subset $E \subseteq \mathbb{R}^2$ to $\mathbb{R}^2$, and $f$ be a scalar-valued function $D \to \mathbb{R}$.

Jacobi’s matrix of $g$ at point $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ is
\[ J_g(a) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(a) & \frac{\partial g_1}{\partial x_2}(a) \\ \frac{\partial g_2}{\partial x_1}(a) & \frac{\partial g_2}{\partial x_2}(a) \end{pmatrix} \quad (28) \]
and Jacobi’s matrix of $f$ at point $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} g_1(a) \\ g_2(a) \end{pmatrix}$ is
\[ J_f(b) = \begin{pmatrix} \frac{\partial f}{\partial y_1}(b) & \frac{\partial f}{\partial y_2}(b) \end{pmatrix} . \quad (29) \]

Jacobi’s matrix of $f \circ g$ at point $a$ is, according to Chain Rule (23), equal to the product of (29) and (28):
\[ J_{f \circ g}(a) = J_f(b) J_g(a) = \begin{pmatrix} \frac{\partial f}{\partial y_1} \frac{\partial g_1}{\partial x_1} + \frac{\partial f}{\partial y_2} \frac{\partial g_2}{\partial x_1} & \frac{\partial f}{\partial y_1} \frac{\partial g_1}{\partial x_2} + \frac{\partial f}{\partial y_2} \frac{\partial g_2}{\partial x_2} \\ f_{y_1}(g_1)_{x_1} + f_{y_2}(g_2)_{x_1} & f_{y_1}(g_1)_{x_2} + f_{y_2}(g_2)_{x_2} \end{pmatrix} \quad (30) \]
where $f_{y_1} := \frac{\partial f}{\partial y_1}$ and $f_{y_2} := \frac{\partial f}{\partial y_2}$ are taken at point $b = \begin{pmatrix} g_1(a) \\ g_2(a) \end{pmatrix}$ whereas $(g_i)_{x_1} := \frac{\partial g_i}{\partial x_1}$ and $(g_i)_{x_2} := \frac{\partial g_i}{\partial x_2}$ are taken at point $a$.

We can rewrite formula (30) in terms of the corresponding gradient vectors, see (17),
\[ \nabla (f \circ g)(a) = J^T_g(a) \nabla f(b) \quad (31) \]
or, in an abbreviated form:
\[ \nabla (f \circ g) = J^T_g \nabla f . \quad (32) \]
Here $J^T$ denotes the transpose of the matrix $J$:

$$
J = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{then} \quad J^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.
$$

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

(Please, verify that!)

**Exercise 5** Derive the following special case of Chain Rule (23):

$$
\nabla f(\gamma(a)) \cdot \frac{d\gamma}{dt}(a) = \frac{d(f \circ \gamma)}{dt}(a)
$$

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

15 **Tangent vectors to a subset $Z \subseteq \mathbb{R}^n$**

We shall say that a column-vector $v$ is tangent to a set $Z$ at point $a$ if there exists a curve $\gamma: I \to \mathbb{R}^n$ contained in $Z$, and an interior point $a$ of $I$, such that

$$
\gamma(a) = a \quad \text{and} \quad \frac{d\gamma}{dt}(a) = v.
$$

Note that, for any number $c \in \mathbb{R}$, “reparametrized” curve $\tilde{\gamma}(t) := \gamma(a + c(t - a))$ passes through point $a$ at $t = a$ with the velocity $c$ times “faster” than $\gamma$ does (this follows from Chain Rule):

$$
\frac{d\tilde{\gamma}}{dt}(a) = v \frac{d(a + c(t - a))}{dt}(a) = cv.
$$

Properly speaking, being tangent is rather a property of vectors anchored at point $a$: an anchored vector $\overrightarrow{ab}$ is said to be tangent to $Z$ if the corresponding column-vector $b - a$ is tangent at $a$ to $Z$.

The set of all vectors tangent to $Z$ at point $a$ is usually denoted $T_aZ$ and called the tangent space to $Z$ at point $a$. It follows from equality (36) that $T_aZ$ contains for every vector $\overrightarrow{ab}$, all its multiples $c \overrightarrow{ab}$, $c \in \mathbb{R}$.

---

5Reread Section 9 of Prelim!
16 If \( a \) is an interior point of set \( Z \) then any column-vector is tangent to \( Z \) at \( a \).

Indeed, since \( a \) is an interior point of \( Z \), it is contained in \( Z \) together with a ball of radius \( \epsilon \) if one chooses number \( \epsilon \) to be sufficiently small. Thus, the path

\[
\gamma: (-\epsilon, \epsilon) \rightarrow \mathbb{R}^n, \quad \text{where} \quad \gamma(t) = a + tv,
\]

(37)
passes through \( a \) at \( t = 0 \) and is contained in \( Z \). Its velocity is constant, i.e. does not depend on \( a \in (-\epsilon, \epsilon) \), and equals \( v \). Note that function (37) is a parametrization of a straight line segment, passing through point \( a \) with constant velocity \( v \).

17 Three examples Let \( Z \) be a rectangle in the plane like the one in Figure 2. We already know (see the previous section) that at any interior point \( a \), the tangent space, \( T_aZ \), is the plane

\[
\{ \overrightarrow{ab} \mid b \text{ is any point of } \mathbb{R}^2 \}.
\]

Let us determine the tangent space, \( T_bZ \), for a point \( b \) which lies on the edge of \( Z \). Suppose that \( \gamma: I \rightarrow \mathbb{R}^2 \) is a path that is contained in \( Z \) and such that \( \gamma(a) = b \) for some \( a \in I \). Since for all \( t \in I \) one has the obvious inequality

\[
\gamma_1(t) \geq b_1 = \gamma_1(a),
\]

\( t = a \) is the absolute minimum of function \( \gamma_1 \). In such a situation, Fermat’s Theorem\(^6\) from Freshman Calculus\(^7\) tells us that \( \gamma_1'(a) = 0 \). In other words, the velocity vector

\[
\frac{d\gamma}{dt}(a) = \begin{pmatrix} 0 \\ \gamma_2'(a) \end{pmatrix}
\]

is vertical (i.e. tangent to the edge of \( Z \)). By considering the vertical path along the edge:

\[
\gamma(t) := \begin{pmatrix} b_1 \\ ct \end{pmatrix},
\]

for a given number \( c \in \mathbb{R} \), we see that any column-vector tangent to the edge at \( b \):

\[
v = \begin{pmatrix} 0 \\ c \end{pmatrix}
\]

is indeed the velocity vector for some path passing through \( b \). To sum up: \( T_bZ \) is the line tangent to the boundary of \( Z \) at \( b \).

\(^6\)Pierre de Fermat (1601–1665)

\(^7\)See, e.g., Stewart §4.1, p. 226.
Finally, let us consider a corner point $c = \left( \begin{array}{c} c_1 \\ c_2 \end{array} \right)$. Let $\gamma: I \to \mathbb{R}^2$ be a path contained in $Z$ such that $\gamma(a) = c$ for some $a \in I$. Since $\gamma(t) \in Z$ for any $t \in I$, the following two inequalities:

$$\gamma_1(t) \geq c_1 = \gamma_1(a) \quad \text{and} \quad \gamma_2(t) \geq c_2 = \gamma_2(a),$$

show that the both component functions of $\gamma$ have absolute minima at $t = a$. By the above mentioned Fermat’s Theorem, $\gamma'_1(a) = \gamma'_2(a) = 0$, i.e.,

$$\frac{d\gamma}{dt}(a) = 0.$$

Thus, the tangent space to $Z$ at corner point $c$ consists of zero vector $\overrightarrow{cc}$ alone.

18 Directional derivative $D_vf$ Let $f$ be a function from a subset $D$ of $\mathbb{R}^m$ to $\mathbb{R}^n$ and $\gamma$ be a parametric curve satisfying (35) and contained in $D$. Then the composite $f \circ \gamma$ is a parametric curve in $\mathbb{R}^n$ and Chain Rule tells us that its velocity vector equals

$$f'_{\gamma(a)}(1) = f'_{\gamma'(a)}(\gamma'_a(1)) = f'_a(v).$$

(38)

We immediately notice that the right-hand side of (38) depends only on vector $v$ and not on any particular choice of parametric curve $\gamma$ satisfying (35).

The directional derivative of $f$ at point $a$ in the direction of a column-vector $v$ is defined as

$$D_vf(a) = \frac{df(a + tv)}{dt} \bigg|_{t=0}.$$
Note that \( f(a + tv) = (f \circ \gamma)(t) \) where \( \gamma \) is the path introduced in (37). By using identity (38), we therefore get the identity:

\[
D_v f(a) = f'_a(v) = J_f(a)v
\]  

(40)

Identity (40) combined with (38) has the following very beautiful application.

If \( v \) is tangent to the level set of \( f \) at \( a \):

\[
Z_a := \{ x \in D \mid f(x) = f(a) \}
\]  

(41)

then, by definition given in Section 15, there exists a parametrized curve contained in \( Z_a \) and satisfying (35). Function \( f \) is, of course, constant on any level set and, since \( \gamma \) is contained in \( Z_a \), composite function \( f \circ \gamma \) is constant. Thus, its derivative \( f \circ \gamma \)' is the zero linear transformation and the left-hand-side of identity (38) therefore vanishes. But the right-hand side of (38) equals \( D_v f \), in view of boxed identity (40). Hence,

\[
\text{the derivative of } f \text{ at } a \text{ vanishes on vectors tangent to level set (41)}
\]  

(42)

If \( n = 1 \), formula (40) reads as follows:

\[
D_v f(a) = \nabla f(a) \cdot v = \sum_{j=1}^{m} \frac{\partial f}{\partial x_j}(a)v_j = 0
\]  

(43)

In other words,

\[
\text{the gradient vector of } f \text{ at } a \text{ is orthogonal to level set (41)}
\]  

(44)

Vice-versa, among vectors of the same length

\[
\text{the directional derivative of } f \text{ attains the largest value on vectors orthogonal to level set (41)}
\]  

(45)
Note that the $j$-th partial derivative is simply the directional derivative of $f$ in the direction of the $j$-th basis vector $e_j$:

$$\frac{\partial f}{\partial x_j}(a) = D_{e_j} f(a).$$

19 Regular versus critical points of a scalar valued function $f : D \to \mathbb{R}$ If gradient vector $\nabla f(a)$ vanishes then $\nabla f(a) \cdot v = 0$ for any column-vector $v \in \mathbb{R}^n$, while it is generally not true that any column-vector is tangent at point $a$ to level set (41) (look at the singular points of two level sets shown in figure 3).

![Figure 3](image)

(a) $f_0 \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) = x^3 - y^2$
(b) $f_1 \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) = x^2(x + 2) - y^2$

Figure 3: Both functions have a critical point at the origin; This produces visible singularities of the corresponding level sets. The tangent spaces at the corresponding critical points are determined in Problembook (Solved Exercises ?? and ??).

Points $a \in D$ where $\nabla f(a) = 0$ are called the critical points of function $f : D \to \mathbb{R}$. At such points formula (43) is of no use.

Vice-versa, points $a \in D$ where $\nabla f(a) \neq 0$ are called the regular points of $f : D \to \mathbb{R}$. Their importance is expressed by the following fundamental fact.
If \( a \) is a regular point of a function \( f: D \to \mathbb{R} \) then the level set of \( f \) passing through \( a \) is smooth in the vicinity of \( a \). Moreover, vector \( \overrightarrow{ab} \) is tangent to the level set of \( f \) if and only if \( \nabla f(a) \cdot (b - a) = 0 \). \hfill (46)

Note that the differential, \( df_a \), which is a linear functional \( \mathbb{R}^m \to \mathbb{R} \), is always either onto or identically zero (see Exercise 6). The latter happens when point \( a \) is critical, the former—if \( a \) is regular.

**Exercise 6** Show that every linear functional \( L: \mathbb{R}^m \to \mathbb{R} \) is either zero or onto.

---

**Figure 4:** The polynomial function

\[
    f \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) = 516x^4y - 340x^2y^3 + 57y^5 - 640x^4 - 168x^2y^2 + 132y^4 - 384x^2y + 292y^3 + 1024x^2
\]

of degree 5 has exactly seven critical points—five belonging to the level set passing through the origin, located at the center, which has indeed five singular points (cusps), cf. Figure 5 below.
Figure 5: The **blue** curve indicates points where $\frac{\partial f}{\partial x}$ vanishes and the **red** curve indicates points where $\frac{\partial f}{\partial y}$ vanishes (where the two curves approach each other the color becomes **violet**; where the blue curve approaches the level set (green curve) the color becomes **cyan**). Their intersection consists of critical points of function $f$ from Figure 4. You can see that there are exactly seven such points, and five of them coincide with the cusps of the level set of $f$. All seven critical points are **degenerate**, cf. Section 22, p. 19.

20 Special case: critical points of a scalar-valued function of two variables  Recall from Section 11 that a function

$$f: D \to \mathbb{R}^2, \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right),$$

defined on a subset $D \subseteq \mathbb{R}^2$ is the same as a function of two scalar variables $x$ and $y$. Let $f$ be differentiable at a point $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$. Since

$$df_a(v) = \nabla f(a) \cdot v = f_x(a)v_1 + f_y(a)v_2 \quad \left(v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right),$$

(47)
differential $df_a$ is identically zero (we express this by writing $df_a = 0$) if and only if the partial derivatives of $f$ vanish:

$$f_x(a) = f_y(a) = 0$$  \hspace{1cm} (48)

or, equivalently, when the gradient of $f$ vanishes at $a$.

For scalar-valued functions of one variable, the type of a critical point (a local maximum, a local minimum, an inflection point) is related to the behavior of the second derivative of $f$ at that point. We expect the same for functions of two variables. What does this second derivative look like in our case?

Suppose $f$ is differentiable at every point $x$ of $D$. The first derivative of $f$ at $x$, which is called the differential of $f$ at $x$, becomes a function

$$df : D \to \{\text{linear functionals on } \mathbb{R}^2\}. \hspace{1cm} (49)$$

Any such function is called a differential form on $D$.

**Example 1.** The differential of function $f \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = x$ is denoted $dx$. Note that $f_x(x) = 1$ and $f_y(x) = 0$ for all $x \in \mathbb{R}^2$, hence

$$dx_a(v) = v_1 \quad \left( v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right) \hspace{1cm} (50)$$

and you observe that, for every $a \in \mathbb{R}^2$, one has $dx_a = \pi_1$ where $\pi_1$ is linear functional $\mathbb{R}^2 \to \mathbb{R}$ defined in (25). Thus, $dx$ is an example of a constant differential form.

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

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

$$df = \frac{\partial f}{\partial x} \, dx + \frac{\partial f}{\partial y} \, dy \hspace{1cm} (51)$$

*Hint. Use identities (16-17), in the case $m = 2$, together with identity (50) and the last exercise.*
The space of linear functionals on \( \mathbb{R}^2 \) can be itself identified with \( \mathbb{R}^2 \); see Section 13 of Prelim. Under this identification, \( f'_a \) corresponds, of course, to gradient vector \( \nabla f(a) \). This is, after all, the main reason why we bothered to introduce \( \nabla f(a) \) 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 (49). Its derivative \( (\nabla f)'(a) \) at \( a \) is thus a linear transformation from \( \mathbb{R}^2 \) to \( \mathbb{R}^2 \). Let us calculate its matrix:

\[
\begin{pmatrix}
\frac{\partial (f_x)}{\partial x}(a) & \frac{\partial (f_x)}{\partial y}(a) \\
\frac{\partial (f_y)}{\partial x}(a) & \frac{\partial (f_y)}{\partial y}(a)
\end{pmatrix} =
\begin{pmatrix}
\frac{\partial^2 f}{\partial x^2}(a) & \frac{\partial^2 f}{\partial y \partial x}(a) \\
\frac{\partial^2 f}{\partial x \partial y}(a) & \frac{\partial^2 f}{\partial y^2}(a)
\end{pmatrix} =
\begin{pmatrix}
f_{xx}(a) & f_{yx}(a) \\
f_{xy}(a) & f_{yy}(a)
\end{pmatrix}
\]

21 Clairaut’s Theorem If \( f_{xy} \) and \( f_{yx} \) are continuous at \( a \) then they are equal.\(^8\)

22 The Hesse Matrix By Clairaut’s Theorem, under mild conditions on a function \( f \), the matrix of the derivative of the gradient function (53) is symmetric.\(^9\)

We shall call

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

the Hesse\(^\text{10}\) matrix of a function \( f : D \to \mathbb{R} \) at a point \( a \). The determinant of (54)

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

is called the Hessian of \( f \) at \( a \).

This concept was introduced for the first time by Ludwig Otto Hesse (1811-1874) in two articles published in 1844 and 1851, respectively.

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

\(^8\)Alexis Claude Clairaut (1713–1765)
\(^9\)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.
\(^\text{10}\)Ludwig Otto Hesse (1811–1874)
Figure 6: The graph of a function $f: D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^2$, in the neighborhood of a nondegenerate critical point; there are three possibilities: a saddle point, a local minimum and a local maximum.

(i) If $H_f(a) < 0$, then $a$ is a saddle point;\(^{11}\)

(ii) If $H_f(a) > 0$, then there are two further possibilities:
   a) $a$ is a local minimum\(^{12}\) if $f_{xx}(a) > 0$,
   b) $a$ is a local maximum if $f_{xx}(a) < 0$.

Note that the positivity of $H_f(a) = f_{xx}f_{yy} - (f_{xy})^2$ 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 $H_f(a)$ does not vanish. If $H_f(a) = 0$ then $a$ is called a degenerate critical point and the situation becomes a lot more complicated in general.

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

\(^{11}\)See also Figure ?? in Problembook.

\(^{12}\)See also Figure ?? in Problembook.
Example 2. Let \( f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = x^2 + 3xy + 2y^2 \). The differential of \( f \) equals (see Exercise 8 above)
\[
\text{df} = \frac{\partial f}{\partial x} \text{dx} + \frac{\partial f}{\partial y} \text{dy} = (2x + 3y)\text{dx} + (3x + 4y)\text{dy}
\]
or, equivalently, the gradient of \( f \) equals
\[
\nabla f = \begin{pmatrix} 2x + 3y \\ 3x + 4y \end{pmatrix}.
\]
A point \( a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \) is a critical point of \( f \) if and only if
\[
\begin{cases}
2a_1 + 3a_2 = 0 \\
3a_1 + 4a_2 = 0
\end{cases}
\]
(56)
The only solution to (56) is \( a_1 = a_2 = 0 \), i.e., the origin is the only critical point of \( f \).

Hesse’s matrix (54) for \( f \) does not depend on \( a \) 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 \( 0 \). Note, however, that the restriction of \( f \) to the \( x \)-axis, \( f\left(\begin{pmatrix} x \\ 0 \end{pmatrix}\right) = x^2 \), and the restriction to the \( y \)-axis, \( f\left(\begin{pmatrix} 0 \\ y \end{pmatrix}\right) = 2y^2 \), both have a minimum at the origin!

Example 3. Function \( f_0 \) from Figure 3(a) has only one critical point \( \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \) where the Hesse matrix equals \( H_{f_0}(\mathbf{0}) = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} \). In particular, critical point \( \mathbf{0} \) is degenerate.

Example 4. Function \( f_1 \) from Figure 3(b) has two critical points: \( \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \) and \( \begin{pmatrix} -2 \\ 0 \end{pmatrix} \).
The Hesse matrices are
\[
H_{f_1}(\mathbf{0}) = \begin{pmatrix} 4 & 0 \\ 0 & -2 \end{pmatrix} \qquad H_{f_1}\left(\begin{pmatrix} -2 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} -10 & 0 \\ 0 & -2 \end{pmatrix}
\]
which means that 0 is a saddle point while \((-2, 0)\) is a local maximum.

### 23 Another look at the definition of a critical point

In Section 19 we declare a point \(a \in D\) to be **critical** for a function \(f: D \to \mathbb{R}\) if the differential of \(f\) at \(a\) identically vanishes:

\[
df_a = 0 ,
\]

i.e., if \(\nabla f(a) = 0\). Differential \(df_a\) is a linear functional \(\mathbb{R}^m \to \mathbb{R}\). So, if \(a\) is *not* a critical point of \(f\) (recall that such points are called **regular**), then \(df_a\) maps \(\mathbb{R}^m\) **onto** \(\mathbb{R}\) (see Exercise 6). And vice-versa:

| a point \(a\) is critical for a function \(f: D \to \mathbb{R}\) if and only if \(df_a: \mathbb{R}^m \to \mathbb{R}\) is not onto. |

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

### 24 Critical points of 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 **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 (58) above), leads us to the following definition.

<table>
<thead>
<tr>
<th>A point (a) is a <strong>regular</strong> point of a vector function (f: D \to \mathbb{R}^n) if:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case (m \geq n). (f'(a): \mathbb{R}^m \to \mathbb{R}^n) is <strong>onto</strong>.</td>
</tr>
<tr>
<td>Case (m \leq n). (f'(a): \mathbb{R}^m \to \mathbb{R}^n) is <strong>one-to-one</strong>.</td>
</tr>
</tbody>
</table>

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 **onto** precisely when it is **one-to-one**.

We say that \(a\) is a **critical** point if \(a\) is not regular.\(^{13}\)\(^{13}\)

\(^{13}\)Terminology: **regular point** and **critical point** applies only to points where the function is differentiable (contrary to what Stewart says in §15.7, p. 990).
Let me remind you what have we established in Section 18: the derivative, \( f'(a) \), of a function \( f: D \to \mathbb{R}^n \) vanishes on vectors tangent to the level set of \( f \) at point \( a \). This holds for any point \( a \). However, for points where \( f \) is regular the reverse is also true.

If \( a \) is a regular point of a function \( f: D \to \mathbb{R}^n \) then the level set of \( f \) passing through \( a \) is smooth in the vicinity of \( a \). Moreover, \( f'(a)(v) = 0 \) if and only if vector \( v \) 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 30 below).

25 Some comments and additions to Theorem (60) 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 (60) does not say much in this case. One can show, however, that

when restricted to a sufficiently small neighbourhood, \( N \), of a regular point \( a \), function \( f \) becomes one-to-one — exactly like its derivative \( f'(a) \) — and the image, \( f(N) \), is smooth.

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!

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

\[ \text{a point } a \text{ is a regular point of } f: D \to \mathbb{R} \text{ if and only if } \nabla f(a) \neq 0. \]
What about the case \( n = 2 \)? In this case \( f = \left( \begin{array}{c} f_1 \\ f_2 \end{array} \right) \) and, assuming that \( m \), i.e. the number of variables, is greater than \( r \), the answer is as follows.

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

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

**Case \( m = 1 \).** In the familiar case of a parametric curve \( \gamma: I \to \mathbb{R}^n \), the regular points are numbers \( a \in I \) where the velocity vector, \( \frac{d\gamma}{dt}(a) \), introduced in (13), does not vanish. Accordingly, the critical points are precisely those numbers \( a \in I \) for which the velocity vector, \( \frac{d\gamma}{dt}(a) \), does vanish. Recall that only at such points the curve parametrized by function \( \gamma \) can have local singularities like “cusps” or “corners”.

**Case \( m = 2 \).** For a function \( f: D \to \mathbb{R}^n \), defined on a subset \( D \subseteq \mathbb{R}^2 \), the Jacobi matrix has two columns:

\[
J_f(a) = \begin{pmatrix}
\frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) \\
\vdots & \vdots \\
\frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a)
\end{pmatrix}.
\]

Assuming \( n \geq 2 \), we have the following characterization of regular points:

A point \( a \) is a regular point of a function \( f: D \to \mathbb{R}^n \), defined on a subset \( D \subseteq \mathbb{R}^2 \), if and only if the two columns of Jacobi matrix (64) span a plane in \( \mathbb{R}^n \).

If they do not — the point is critical. This happens either because the two columns of matrix (64) are collinear or, in the most degenerate case, because they both vanish.

**Comment.** You must have noticed parallels between cases \( m = 1 \) and \( n = 1 \), as well as between cases \( m = 2 \) and \( n = 2 \). This is not accidental, one can rephrase the definition of a regular point by saying that a point \( a \in D \) is a regular point of function \( f \) when the Jacobi matrix, \( J_f(a) \), has

\[\text{collinear}\]

This does not preclude that the global image of \( \gamma \) may have singularities like “nodes” even though \( \gamma \) has no critical points; cf. Figure 7(a).
(a) The image of $\gamma$

(b) $\gamma((−1/2))$

(c) $\gamma((−1/2,))$

**Figure 7:** Every point $a \in \mathbb{R} = (−, )$ is regular for the function $\gamma: \mathbb{R} \to \mathbb{R}^2$ given by

$$\gamma(t) = \begin{pmatrix} t^2 - 2 \\ t(t^2 - 2) \end{pmatrix}.$$  

The image of $\gamma$, i.e., set $\gamma(\mathbb{R})$, has a singularity at the origin and $\gamma$ is not one-to-one, since $\gamma(−\sqrt{2}) = 0 = \gamma(\sqrt{2})$, see Subfigure (a). Function $\gamma$ is one-to-one when restricted to the neighborhood $(-1/2)$ of point $−\sqrt{2}$, see Subfigure (b), or to the neighborhood $(-1/2,)$ of point $\sqrt{2}$, see Subfigure (c). In either case, the image of the restricted function is a smooth arc.

When $m$ or $n$ equals 1 the largest possible value of rank of $J_f(a)$ is 1. When the smaller of the two numbers $m$ and $n$ equals 2, the largest possible value of rank of $J_f(a)$ is 2.

---

15Rank of an $n \times m$ matrix $A$ is the dimension of the space spanned by the rows of $A$ (equivalently, by the columns of $A$). As such, the largest value the rank can take is $\min(m, n)$, the smaller of the two numbers $m$ and $n$.

In the case of square matrices, an $n \times n$ matrix $A$ has rank $n$ if and only if $\det A \neq 0$. Rank of a matrix is one of the fundamental concepts of Linear Algebra.
27 Local extrema of a function \( f: D \to \mathbb{R} \) along a path Consider a path \( \gamma: I \to D \). We shall say that a function \( f: D \to \mathbb{R} \) has, at a point \( a = \gamma(a) \), a local maximum (minimum) along path \( \gamma \) if the composite function \( f \circ \gamma: I \to \mathbb{R} \) has a local maximum (respectively, minimum) at \( a \). In this case, Fermat’s Theorem mentioned a few times before tells us that the derivative of \( f \circ \gamma \) at \( a \) vanishes and we deduce from Chain Rule (22) — see also Exercise 5 and formula (34) — that

\[
\left. \frac{df}{d\gamma(a)} \right|_{a} \text{ annihilates the velocity vector } \frac{d\gamma}{dt}(a), \text{ i.e. } \nabla f(a) \cdot \frac{d\gamma}{dt}(a) = 0.
\]

In other words, gradient \( \nabla f(a) \) and the velocity vector \( \frac{d\gamma}{dt}(a) \) are orthogonal to each other.

28 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 (67) we know that if \( \gamma: I \to Z \) is any differentiable path passing through a point \( a = \gamma(a) \) — where function \( f \) has its local maximum or minimum on \( Z \) — then differential \( df_a \) annihilates velocity vector \( \frac{d\gamma}{dt}(a) \).

Now, any vector tangent to \( Z \) at point \( a \) occurs as the velocity vector of some path passing through it. Hence we arrive at the following generalization of Fermat’s Theorem.

If a function \( f \) has a local extremum on \( Z \) at a point \( a \) then \( df_a \) vanishes on all vectors tangent to \( Z \) at point \( a \).

Note that Theorem (68) covers also the case when \( Z \) is the whole set \( D \). If \( a \) is an interior point of \( D \) then any vector \( v \in \mathbb{R}^m \) is tangent to \( D \) at \( a \). Thus, Theorem (68) has the following corollary.

If \( f \) has a local extremum at an interior point \( a \) then \( df_a \) is zero, i.e. \( a \) is a critical point of the function \( f \).
29 Example Let \( f : E \to \mathbb{R} \) be a function on the ellipse

\[
E := \left\{ x = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \left| \left( \frac{x - c_1}{a} \right)^2 + \left( \frac{y - c_2}{b} \right)^2 \leq 1 \right. \right\}.
\]

with center at \( c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \). Local extrema of \( f \) on \( E \) are either critical points of \( f \) belonging to \( E \) or points \( x = \begin{pmatrix} x \\ y \end{pmatrix} \) satisfying the following two equations:

\[
\begin{cases} 
\nabla f(x) \cdot \begin{pmatrix} a^2(y - c_2) \\ -b^2(x - c_1) \end{pmatrix} = 0 \\
\left( \frac{x - c_1}{a} \right)^2 + \left( \frac{y - c_2}{b} \right)^2 = 1 
\end{cases}
\]

The second equation expresses the fact that point \( x \) belongs to the boundary, \( \partial E \), of ellipse \( E \). The first equation expresses the fact that \( df_x \) vanishes on any column-vector tangent to \( \partial E \) at point \( \begin{pmatrix} x \\ y \end{pmatrix} \). This is so, because any such column-vector is a multiple of column-vector \( \begin{pmatrix} a^2(y - c_2) \\ -b^2(x - c_1) \end{pmatrix} \) (cf. Solved Exercise ?? in Problembook).

30 Lagrange multipliers Now, a practical application of great importance. Suppose that you must find extrema of a function \( f : D \to \mathbb{R} \) where 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 \( k_1, \ldots, k_r \) being given in advance). The first thing you should do is to rewrite \( r \) constraints (72) as a single vector constraint:

\[
g(x) = K
\]

where \( g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_r(x) \end{pmatrix} \) and \( K = \begin{pmatrix} k_1 \\ \vdots \\ k_r \end{pmatrix} \). Denote by \( Z \) the corresponding level set of vector-constraint function \( g \):

\[
Z = \{ x \in D \mid g(x) = K \}.
\]
Theorem (68) tells us that \( df_a \) vanishes on vectors tangent to \( Z \) at a point \( a \) if function \( f \) has a local extremum on \( Z \) at \( a \). If \( a \) is a regular point of vector-constraint function \( g \) then its derivative \( g'(a) \) vanishes precisely on vectors tangent to \( Z \).

Now, derivative \( g'(a) \) is a linear transformation from \( \mathbb{R}^m \) to \( \mathbb{R}^r \) and differential \( df_a \) is a linear functional on \( \mathbb{R}^m \). Since \( g'(a) \) vanishes only on those vectors on which \( df_a \) vanishes, one can “divide” linear functional \( df_a \) by linear transformation \( g'(a) \). The exact meaning of this phrase is:

there exists a (not necessarily unique)\(^{16} \) linear functional \( \Lambda \) on \( \mathbb{R}^r \) such that \( df_a \) is the composition of \( \Lambda \) and \( g'(a) \):

\[
\text{df}_a = \Lambda \circ g'(a). \tag{75}
\]

Any linear functional on \( \mathbb{R}^r \) is conveniently described by formula (35) in Section 13 of Prelim, as you already know. In our case, this means that

\[
\Lambda(v) = \lambda \cdot v \quad (v \in \mathbb{R}^r) \tag{76}
\]

for a suitable vector \( \lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_r \end{pmatrix} \).

**Exercise 9** Verify that equality (75) can be rewritten as follows:

\[
\nabla f(a) = \lambda_1 \nabla g_1(a) + \cdots + \lambda_r \nabla g_r(a) \tag{77}
\]

Equality (77) expresses the fact that gradient vector \( \nabla f(a) \) is a linear combination of gradient vectors \( \nabla g_1(a), \ldots, \nabla g_r(a) \) with coefficients \( \lambda_1, \ldots, \lambda_r \). Coefficients \( \lambda_1, \ldots, \lambda_r \) are called Lagrange multipliers.\(^{17} \) To sum up, we have established the following remarkable theorem which is the essence of the Lagrange multipliers method.

\(^{16}\) \( \Lambda \) is unique if the number of constraints, \( r \), does not exceed dimension \( m \). Incidentally, this is the only interesting case.

\(^{17}\) Giuseppe Lodovico Lagrangia (1736–1813), his name is better known in its French form.
At any point \( a \) where function \( f \) has a local extremum with \( r \) constraints (72), gradient vector \( \nabla f(a) \) can be expressed as a linear combination (77) of gradient vectors \( \nabla g_1(a), \ldots, \nabla g_r(a) \) for suitable numbers \( \lambda_1, \ldots, \lambda_r \) provided \( a \) is a regular point of the vector-constraint function:

\[
\begin{pmatrix}
    g_1(x) \\
    \vdots \\
    g_r(x)
\end{pmatrix}
\]

(78)

Theorem (78) 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 (72).

Finally, you should be always prepared that there may be no points satisfying given constraints, in which case level set (74) is empty. When this happens then there is no point, of course, in trying to find corresponding constrained extrema of function \( f \).