# Differential Calculus of Vector Functions

October 9, 2003

*These notes should be studied in conjunction with lectures.*[1]

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

$$\mathbf{f}(\mathbf{x}) \ \textit{tends to} \ \mathbf{f}(\mathbf{a}) \qquad \text{whenever} \qquad \mathbf{x} \ \textit{tends to} \ \mathbf{a}\,. \tag{1}$$

If function $\mathbf{f}$ is continuous at *every* point of its domain, then we simply say that $\mathbf{f}$ 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 $\mathbf{a}$ be an *interior* point of $D$.[2] We shall say that $\mathbf{f}$ is **differentiable** at $\mathbf{a}$ if there exists a linear transformation $L\colon \mathbb{R}^m \to \mathbb{R}^n$ such that

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a}) \;=\; L(\mathbf{x}-\mathbf{a}) + \mathbf{u}(\mathbf{x}) \tag{2}$$

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

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{\|\mathbf{u}(\mathbf{x})\|}{\operatorname{dist}(\mathbf{x},\mathbf{a})} \;=\; \lim_{\mathbf{x}\to\mathbf{a}} \frac{\|\mathbf{u}(\mathbf{x})\|}{\|\mathbf{x}-\mathbf{a}\|} \;=\; 0\,. \tag{3}$$

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

$$\mathbf{f}(\mathbf{x}) \;=\; \mathbf{f}(\mathbf{a}) + (\mathbf{f}'(\mathbf{a}))(\mathbf{x}-\mathbf{a}) + \mathbf{u}(\mathbf{x}) \tag{4}$$

[1] Abbreviations Prelim and Problembook stand for *Preliminaries* and *Problembook*, respectively.
[2] A point $\mathbf{a}$ is an *interior* point of a set $D$ if $D$ containes some ball with center at $\mathbf{a}$.

where $\mathbf{u}(\mathbf{x})$ is negligible when $\mathbf{x}$ approaches $\mathbf{a}$.

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

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

$$\begin{aligned}\|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{a})\| &= \|(\mathbf{f}'(\mathbf{a}))(\mathbf{x}-\mathbf{a})+\mathbf{u}(\mathbf{x})\| \\ &\leqslant \|(\mathbf{f}'(\mathbf{a}))(\mathbf{x}-\mathbf{a})\|+\|\mathbf{u}(\mathbf{x})\| \quad \text{(Triangle Inequality, cf. Sect. 6 of Prelim)} \\ &\leqslant \|\mathbf{f}'(\mathbf{a})\|\,\|\mathbf{x}-\mathbf{a}\|+\|\mathbf{u}(\mathbf{x})\| \qquad \text{(inequality (34) in Prelim).} \end{aligned} \tag{5}$$

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

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

$$(\mathbf{f}+\mathbf{g})'(\mathbf{a})=\mathbf{f}'(\mathbf{a})+\mathbf{g}'(\mathbf{a})\,; \tag{6}$$

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

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

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

$$\frac{\partial f}{\partial x_j}(\mathbf{a}) \tag{7}$$

is defined as the ordinary derivative

$$\left.\frac{d\phi_j}{dt}\right|_{t=a_j} \tag{8}$$

of the function of single real variable

$$\phi_j(t) := f\left(\begin{pmatrix} a_1 \\ \vdots \\ t \\ \vdots \\ a_m \end{pmatrix}\right) \quad \text{j -th coordinate} \tag{9}$$

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

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

$$\boldsymbol{\gamma}_j(t) := \begin{pmatrix} a_1 \\ \vdots \\ t \\ \vdots \\ a_m \end{pmatrix} = \mathbf{a} + (t - a_j)\mathbf{e_j} . \tag{10}$$

**6 Theorem** *The $n \times m$ matrix corresponding to the linear transformation*

$$\mathbf{f}'(\mathbf{a}) : \mathbb{R}^m \to \mathbb{R}^n$$

*is formed by partial derivatives of components of* $\mathbf{f}$*:*

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_1}{\partial x_m}(\mathbf{a}) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_n}{\partial x_m}(\mathbf{a}) \end{pmatrix} . \tag{11}$$

Here $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \dots \\ f_n(\mathbf{x}) \end{pmatrix}$; each component $f_i$ of $\mathbf{f}$ is a scalar-valued function $D \to \mathbb{R}$. Matrix (11) is called the **Jacobi**[3] **matrix** of $\mathbf{f}$ at $\mathbf{a}$ and will be denoted $J_{\mathbf{f}}(\mathbf{a})$. [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 $\mathbf{a} \in D$ is an interior point of $D$.

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

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

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

**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 $\mathbf{f}\colon D \to \mathbb{R}^n$ is differentiable at a point $\mathbf{a}$ if it is of class $C^1$ *on some neighborhood* of $\mathbf{a}$, i.e., on some *open* ball

$$\mathbf{B}_r(\mathbf{a}) := \left\{\mathbf{x} \in \mathbb{R}^m \mid \operatorname{dist}(\mathbf{x}, \mathbf{a}) < r\right\}. \tag{12}$$

**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 $t \mapsto \mathbf{a}t$ for a suitable column-vector $\mathbf{a}$. In the case of linear transformation $\gamma'(a) : \mathbb{R} \to \mathbb{R}^n$ that vector happens to be the **velocity**

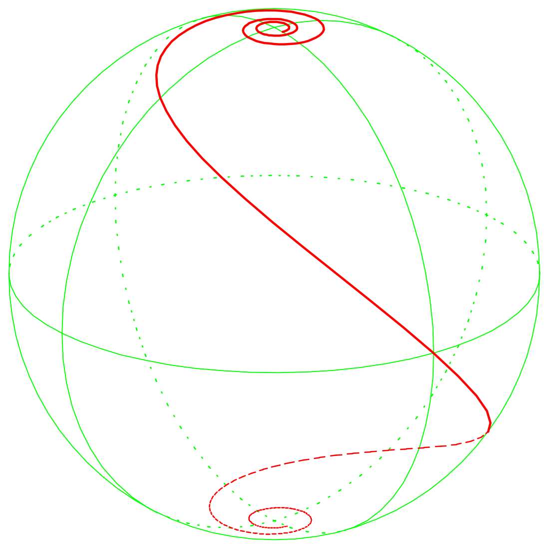


Figure 1: A parametric curve contained in the unit sphere in $\mathbb{R}^3$:

$$\boldsymbol{\gamma}\colon \mathbb{R} \to \mathbb{R}^3, \qquad \boldsymbol{\gamma}(\theta) = \frac{1}{\sqrt{1+\theta^2}} \begin{pmatrix} \cos\theta \\ \sin\theta \\ \theta \end{pmatrix}$$

**vector** of the parametric curve:

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

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

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

## 11 The case of a scalar-valued function of $m$ variables $f\colon D \to \mathbb{R}$

A scalar-valued function of $m$ scalar variables

$$(x_1, \ldots, x_m) \mapsto f(x_1, \ldots, x_m) \tag{14}$$

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

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

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

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

$$\begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{a}) & \ldots & \frac{\partial f}{\partial x_m}(\mathbf{a}) \end{pmatrix} \tag{15}$$

and

$$df_{\mathbf{a}}(\mathbf{v}) = \nabla f(\mathbf{a}) \cdot \mathbf{v} \tag{16}$$

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

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

☞ Vector (17) is called the **gradient** of $f$ at $\mathbf{a}$. Note that it is the **transpose** of Jacobi's matrix (15).

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

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{a}) + df_{\mathbf{a}}(\mathbf{x} - \mathbf{a}) + \mathbf{u}(\mathbf{x}) \\ &= f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}) + \mathbf{u}(\mathbf{x}) \end{aligned} \tag{18}$$

where $\mathbf{u}(\mathbf{x})$ is negligible when $\mathbf{x}$ approaches $\mathbf{a}$.

**12 Chain Rule** Suppose that two functions are given

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

and

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

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

Suppose that $\mathbf{g}$ is differentiable at $\mathbf{a}$ and that $\mathbf{f}$ is differentiable at $\mathbf{b} = \mathbf{g}(\mathbf{a})$. In other words:

$$\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{a}) = \mathbf{g}'(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{u}(\mathbf{x}) \tag{19}$$

and

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{b}) = \mathbf{f}'(\mathbf{b})(\mathbf{y} - \mathbf{b}) + \mathbf{v}(\mathbf{y}) \tag{20}$$

where $\mathbf{u}(\mathbf{x})$ and $\mathbf{v}(\mathbf{y})$ are negligible when $\mathbf{x} \to \mathbf{a}$ and $\mathbf{y} \to \mathbf{b}$, respectively.

Plug $\mathbf{y} = \mathbf{g}(\mathbf{x})$ and $\mathbf{b} = \mathbf{g}(\mathbf{a})$ into (20) and use identity (19):

$$\begin{aligned} \mathbf{f}(\mathbf{g}(\mathbf{x})) - \mathbf{f}(\mathbf{g}(\mathbf{a})) &= \mathbf{f}'(\mathbf{g}(\mathbf{a}))\,(\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{a})) + \mathbf{v}(\mathbf{g}(\mathbf{x})) \\ &= \mathbf{f}'(\mathbf{g}(\mathbf{a}))\,\big(\mathbf{g}'(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{u}(\mathbf{x})\big) + \mathbf{v}(\mathbf{g}(\mathbf{x})) \\ &= \big(\mathbf{f}'(\mathbf{g}(\mathbf{a})) \circ \mathbf{g}'(\mathbf{a})\big)\,(\mathbf{x} - \mathbf{a}) + [\mathbf{f}'(\mathbf{g}(\mathbf{a}))(\mathbf{u}(\mathbf{x})) + \mathbf{v}(\mathbf{g}(\mathbf{x}))] \end{aligned} \tag{21}$$

The composition of two linear transformations is linear. Therefore $\mathbf{f}'(\mathbf{g}(\mathbf{a})) \circ \mathbf{g}'(\mathbf{a})$ 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 $\mathbf{f} \circ \mathbf{g}$ is differentiable at $\mathbf{a}$ and its derivative is given by the following formula:

$$\boxed{(\mathbf{f} \circ \mathbf{g})'(\mathbf{a}) = \mathbf{f}'(\mathbf{g}(\mathbf{a})) \circ \mathbf{g}'(\mathbf{a})} \ . \tag{22}$$

☞ This is the general form of the **Chain Rule.** Here is an equivalent statement of the Chain Rule in terms of Jacobi's matrices:

$$\boxed{J_{\mathbf{f} \circ \mathbf{g}}(\mathbf{a}) = J_{\mathbf{f}}(\mathbf{g}(\mathbf{a}))\, J_{\mathbf{g}}(\mathbf{a})} \ . \tag{23}$$

**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:

$$\epsilon_j\colon \mathbb{R} \to \mathbb{R}^m. \qquad \epsilon_j(t) := t\mathbf{e_j}, \tag{24}$$

and

$$\pi_i\colon \mathbb{R}^n \to \mathbb{R}, \qquad \pi_i\left(\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}\right) := v_i\,. \tag{25}$$

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

$$t \mapsto a_{ij}t \qquad (t \in \mathbb{R})\,.$$

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

$$(\pi_i \circ \mathbf{f} \circ \boldsymbol{\gamma}_j)'(a_j) = \pi_i'(\mathbf{f}(\mathbf{a})) \circ \mathbf{f}'(\mathbf{a}) \circ \boldsymbol{\gamma}_j'(a_j)\,. \tag{26}$$

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

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

$$(\pi_i \circ \mathbf{f} \circ \boldsymbol{\gamma}_j)'(a_j) = \pi_i \circ \mathbf{f}'(\mathbf{a}) \circ \epsilon_j\,. \tag{27}$$

The left-hand side of (27) multiplies $t \in \mathbb{R}$ by $\frac{\partial f_i}{\partial x_j}(\mathbf{a})$, while the right-hand side multiplies $t$ by entry $a_{ij}$ of the matrix corresponding to derivative $\mathbf{f}'(\mathbf{a})$. Therefore these two numbers must be equal, as asserted in Theorem 9.

14 Let us take a closer look, for example, at the case $\ell = m = 2$ and $n = 1$. Let

$$\mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} : E \longrightarrow \mathbb{R}^2$$

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 $\mathbf{g}$ at point $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ is

$$J_{\mathbf{g}}(\mathbf{a}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{a}) & \frac{\partial g_1}{\partial x_2}(\mathbf{a}) \\ \frac{\partial g_2}{\partial x_1}(\mathbf{a}) & \frac{\partial g_2}{\partial x_2}(\mathbf{a}) \end{pmatrix} \tag{28}$$

and Jacobi's matrix of $f$ at point $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} := \begin{pmatrix} g_1(\mathbf{a}) \\ g_2(\mathbf{a}) \end{pmatrix}$ is

$$J_f(\mathbf{b}) = \begin{pmatrix} \frac{\partial f}{\partial y_1}(\mathbf{b}) & \frac{\partial f}{\partial y_2}(\mathbf{b}) \end{pmatrix} . \tag{29}$$

Jacobi's matrix of $f \circ \mathbf{g}$ at point $\mathbf{a}$ is, according to Chain Rule (23), equal to the product of (29) and (28):

$$\begin{aligned} J_{f\circ\mathbf{g}}(\mathbf{a}) &= J_f(\mathbf{b})\, J_{\mathbf{g}}(\mathbf{a}) \\ &= \begin{pmatrix} \frac{\partial f}{\partial y_1}\frac{\partial g_1}{\partial x_1} + \frac{\partial f}{\partial y_2}\frac{\partial g_2}{\partial x_1} & \frac{\partial f}{\partial y_1}\frac{\partial g_1}{\partial x_2} + \frac{\partial f}{\partial y_2}\frac{\partial g_2}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} f_{y_1}\,(g_1)_{x_1} + f_{y_2}\,(g_2)_{x_1} & f_{y_1}\,(g_1)_{x_2} + f_{y_2}\,(g_2)_{x_2} \end{pmatrix} \end{aligned} \tag{30}$$

where $f_{y_1} := \partial f/\partial y_1$ and $f_{y_2} := \partial f/\partial y_2$ are taken at point $\mathbf{b} = \begin{pmatrix} g_1(\mathbf{a}) \\ g_2(\mathbf{a}) \end{pmatrix}$ whereas $(g_i)_{x_1} := \partial g_i/\partial x_1$ and $(g_i)_{x_2} := \partial g_i/\partial x_2$ are taken at point $\mathbf{a}$.

We can rewrite formula (30) in terms of the corresponding gradient vectors, see (17),

$$\nabla(f \circ \mathbf{g})\,(\mathbf{a}) \;=\; J^{\mathsf{T}}_{\mathbf{g}}(\mathbf{a})\, \nabla f(\mathbf{b}) \tag{31}$$

or, in an abbreviated form:

$$\boxed{\nabla(f \circ \mathbf{g}) \;=\; J^{\mathsf{T}}_{\mathbf{g}}\, \nabla f} \;. \tag{32}$$

Here $J^T$ denotes the *transpose* of the matrix $J$:

$$\text{if } J = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{then} \qquad J^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}. \tag{33}$$

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):*

$$\boxed{\nabla f(\boldsymbol{\gamma}(a)) \cdot \frac{d\boldsymbol{\gamma}}{dt}(a) \;=\; \frac{d(f \circ \boldsymbol{\gamma})}{dt}(a)} \tag{34}$$

*for* $f\colon D \to \mathbb{R}$ *and a parametric curve* $\boldsymbol{\gamma}\colon I \to D$. *(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* $\mathbf{v}$ is **tangent** to a set $Z$ at point $\mathbf{a}$ if there exists a curve $\boldsymbol{\gamma}\colon I \to \mathbb{R}^n$ contained in $Z$, and an interior point $a$ of $I$, such that

$$\boldsymbol{\gamma}(a) = \mathbf{a} \qquad \text{and} \qquad \frac{d\boldsymbol{\gamma}}{dt}(a) = \mathbf{v}. \tag{35}$$

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

$$\frac{d\tilde{\boldsymbol{\gamma}}}{dt}(a) = \mathbf{v}\frac{d(a + c(t-a))}{dt}(a) = c\mathbf{v}. \tag{36}$$

Properly speaking, being tangent is rather a property of vectors anchored at point $\mathbf{a}$: an *anchored vector* $\overrightarrow{\mathbf{ab}}$ is said to be **tangent** to $Z$ if the corresponding column-vector $\mathbf{b} - \mathbf{a}$ is tangent at $\mathbf{a}$ to $Z$.[5]

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

[5] Reread Section 9 of Prelim!

**16** *If* $\mathbf{a}$ *is an interior point of set* $Z$ *then any column-vector is tangent to* $Z$ *at* $\mathbf{a}$.

Indeed, since $\mathbf{a}$ 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

$$\boldsymbol{\gamma}\colon (-\epsilon, \epsilon) \to \mathbb{R}^n, \qquad \text{where} \qquad \boldsymbol{\gamma}(t) = \mathbf{a} + t\mathbf{v}, \tag{37}$$

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

**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 $\mathbf{a}$, the tangent space, $T_{\mathbf{a}}Z$, is the plane

$$\left\{\overrightarrow{\mathbf{ab}} \mid \mathbf{b} \text{ is any point of } \mathbb{R}^2\right\}.$$

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

$$\gamma_1(t) \geqslant b_1 = \gamma_1(a),$$

$t = a$ 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

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

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

$$\boldsymbol{\gamma}(t) := \begin{pmatrix} b_1 \\ ct \end{pmatrix},$$

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

$$\mathbf{v} = \begin{pmatrix} 0 \\ c \end{pmatrix}$$

is indeed the velocity vector for some path passing through $\mathbf{b}$. To sum up: $T_{\mathbf{b}}Z$ is the line tangent to the boundary of $Z$ at $\mathbf{b}$.

[6] Pierre de Fermat (1601–1665)

[7] See, e.g., Stewart §4.1, p. 226.

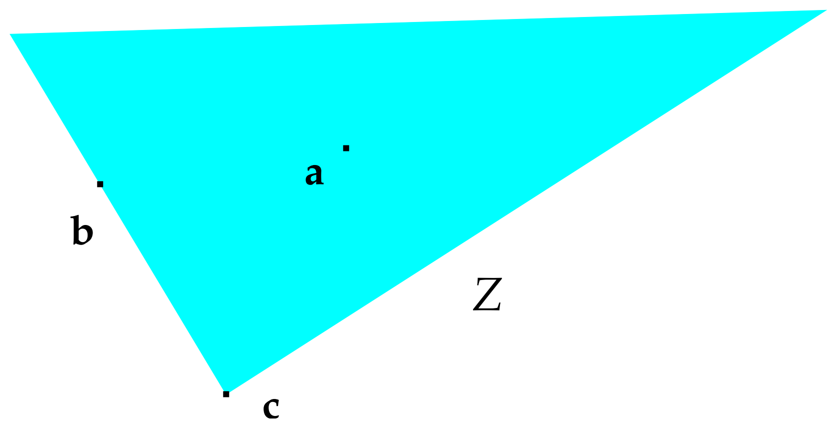



Figure 2: A subset Z of the plane and three points with different types of tangent spaces.

Finally, let us consider a corner point $\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$. Let $\boldsymbol{\gamma}\colon I \to \mathbb{R}^2$ be a path contained in Z such that $\boldsymbol{\gamma}(a) = \mathbf{c}$ for some $a \in I$. Since $\boldsymbol{\gamma}(t) \in Z$ for any $t \in I$, the following two inequalities:

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

show that the both component functions of $\boldsymbol{\gamma}$ have absolute minima at $t = a$. By the above mentioned Fermat's Theorem, $\gamma_1'(a) = \gamma_2'(a) = 0$, i.e.,

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

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

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

$$(\mathbf{f} \circ \boldsymbol{\gamma})_a'(1) = \mathbf{f}'_{\gamma(a)}(\boldsymbol{\gamma}'_a(1)) = \mathbf{f}'_{\mathbf{a}}(\mathbf{v})\,. \tag{38}$$

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

☞ The **directional derivative** of f at point $\mathbf{a}$ in the direction of a column-vector $\mathbf{v}$ is defined as

$$D_{\mathbf{v}}\mathbf{f}(\mathbf{a}) = \frac{d\mathbf{f}(\mathbf{a} + t\mathbf{v})}{dt}\Big|_{t=0}\,. \tag{39}$$

Note that $\mathbf{f}(\mathbf{a}+t\mathbf{v}) = (\mathbf{f}\circ\boldsymbol{\gamma})(t)$ where $\boldsymbol{\gamma}$ is the path introduced in (37). By using identity (38), we therefore get the identity:

$$\boxed{D_{\mathbf{v}}\mathbf{f}(\mathbf{a}) \;=\; \mathbf{f}'_{\mathbf{a}}(\mathbf{v}) \;=\; J_{\mathbf{f}}(\mathbf{a})\,\mathbf{v}} \tag{40}$$

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

If $\mathbf{v}$ is tangent to the level set of $\mathbf{f}$ at $\mathbf{a}$:

$$Z_{\mathbf{a}} := \{\mathbf{x}\in D \mid \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a})\} \tag{41}$$

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

the derivative of $\mathbf{f}$ at $\mathbf{a}$ vanishes on vectors tangent to level set (41) . (42)

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

$$\boxed{D_{\mathbf{v}}f(\mathbf{a}) = \nabla f(\mathbf{a})\cdot\mathbf{v} = \sum_{j=1}^{m}\frac{\partial f}{\partial x_j}(\mathbf{a})v_j = 0\,.} \tag{43}$$

In other words,

the gradient vector of $f$ at $\mathbf{a}$ is *orthogonal* to level set (41) (44)

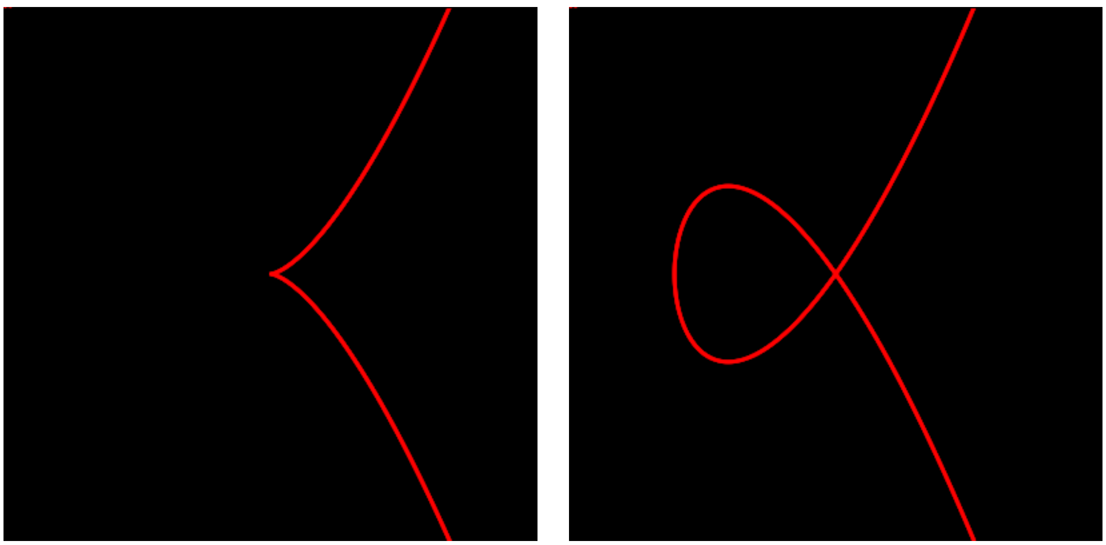
Vice-versa, among vectors of the same length

the directional derivative of $f$ 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 $\mathbf{e_j}$:

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

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



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

Vice-versa, points $\mathbf{a} \in D$ where $\nabla f(\mathbf{a}) \neq 0$ are called the **regular** points of $f\colon D \to \mathbb{R}$. Their importance is expressed by the following fundamental fact. ☜

> If $\mathbf{a}$ is a regular point of a function $f\colon D \to \mathbb{R}$ then the level set of $f$ passing through $\mathbf{a}$ is **smooth** in the vicinity of $\mathbf{a}$. Moreover, vector $\overrightarrow{\mathbf{ab}}$ is tangent to the level set of $f$ **if and only if** $\nabla f(\mathbf{a}) \cdot (\mathbf{b} - \mathbf{a}) = 0$. (46)

Note that the differential, $df_{\mathbf{a}}$, 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 $\mathbf{a}$ is critical, the former—if $\mathbf{a}$ is regular.

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

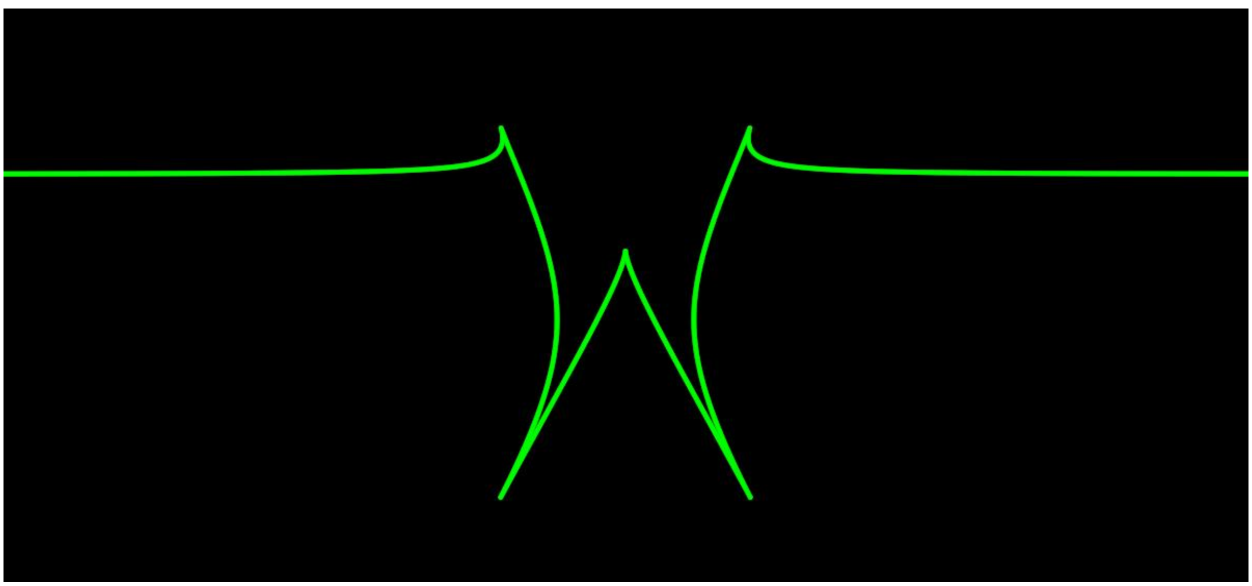


Figure 4: The polynomial function

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

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.

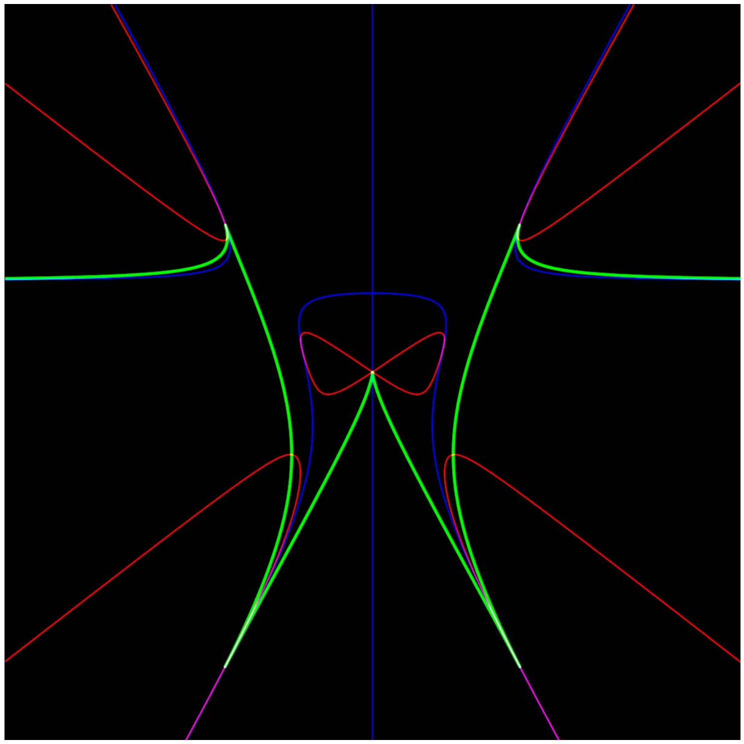


Figure 5: The **blue** curve indicates points where $\frac{\partial f}{\partial x}$ vanishes and the **red** curve indicates points where $\frac{\partial f}{\partial y}$ vanishes (where the two curves approach each other the color becomes *violet*; where the blue curve approaches the level set (green 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

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

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 $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$. Since

$$df_{\mathbf{a}}(\mathbf{v}) = \nabla f(\mathbf{a}) \cdot \mathbf{v} = f_x(\mathbf{a})\, v_1 + f_y(\mathbf{a})\, v_2 \qquad \left(\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right), \tag{47}$$

differential $df_{\mathbf{a}}$ is identically zero (we express this by writing $df_{\mathbf{a}} = 0$) if and only if the partial derivatives of $f$ vanish:

$$\boxed{f_x(\mathbf{a}) = f_y(\mathbf{a}) = 0} \tag{48}$$

or, equivalently, when the gradient of $f$ vanishes at $\mathbf{a}$.

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

$$df\colon D \longrightarrow \{\text{linear functionals on } \mathbb{R}^2\}\,. \tag{49}$$

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

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

$$dx_{\mathbf{a}}(\mathbf{v}) = v_1 \qquad \left(\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right) \tag{50}$$

and you observe that, for every $\mathbf{a} \in \mathbb{R}^2$, one has $dx_{\mathbf{a}} = \pi_1$ 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* $dy_{\mathbf{x}}(\mathbf{v})$. *Does it depend on* $\mathbf{x} \in \mathbb{R}^2$*?*

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

$$\boxed{df \;=\; \frac{\partial f}{\partial x}\, dx + \frac{\partial f}{\partial y}\, dy}\;. \tag{51}$$

*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, $f'_{\mathbf{a}}$ corresponds, of course, to gradient vector $\nabla f(\mathbf{a})$. This is, after all, the main reason why we bothered to introduce $\nabla f(\mathbf{a})$ in the first place!

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

$$\nabla f\colon D \to \mathbb{R}^2 \tag{52}$$

instead of differential (49). Its derivative $(\nabla f)'(\mathbf{a})$ at $\mathbf{a}$ is thus a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$. Let us calculate its matrix:

$$\begin{pmatrix} \frac{\partial(f_x)}{\partial x}(\mathbf{a}) & \frac{\partial(f_x)}{\partial y}(\mathbf{a}) \\ \frac{\partial(f_y)}{\partial x}(\mathbf{a}) & \frac{\partial(f_y)}{\partial y}(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(\mathbf{a}) & \frac{\partial^2 f}{\partial y \partial x}(\mathbf{a}) \\ \frac{\partial^2 f}{\partial x \partial y}(\mathbf{a}) & \frac{\partial^2 f}{\partial y^2}(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} f_{xx}(\mathbf{a}) & f_{yx}(\mathbf{a}) \\ f_{xy}(\mathbf{a}) & f_{yy}(\mathbf{a}) \end{pmatrix} \tag{53}$$

**21 Clairaut's Theorem** If $f_{xy}$ and $f_{yx}$ are **continuous** at $\mathbf{a}$ 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

$$\begin{pmatrix} f_{xx}(\mathbf{a}) & f_{yx}(\mathbf{a}) \\ f_{xy}(\mathbf{a}) & f_{yy}(\mathbf{a}) \end{pmatrix} \tag{54}$$

☞ the **Hesse**[10] **matrix** of a function $f\colon D \to \mathbb{R}$ at a point $\mathbf{a}$. The determinant of (54)

$$H_f(\mathbf{a}) = \begin{vmatrix} f_{xx}(\mathbf{a}) & f_{yx}(\mathbf{a}) \\ f_{xy}(\mathbf{a}) & f_{yy}(\mathbf{a}) \end{vmatrix} \tag{55}$$

☞ is called the **Hessian** of $f$ at $\mathbf{a}$.

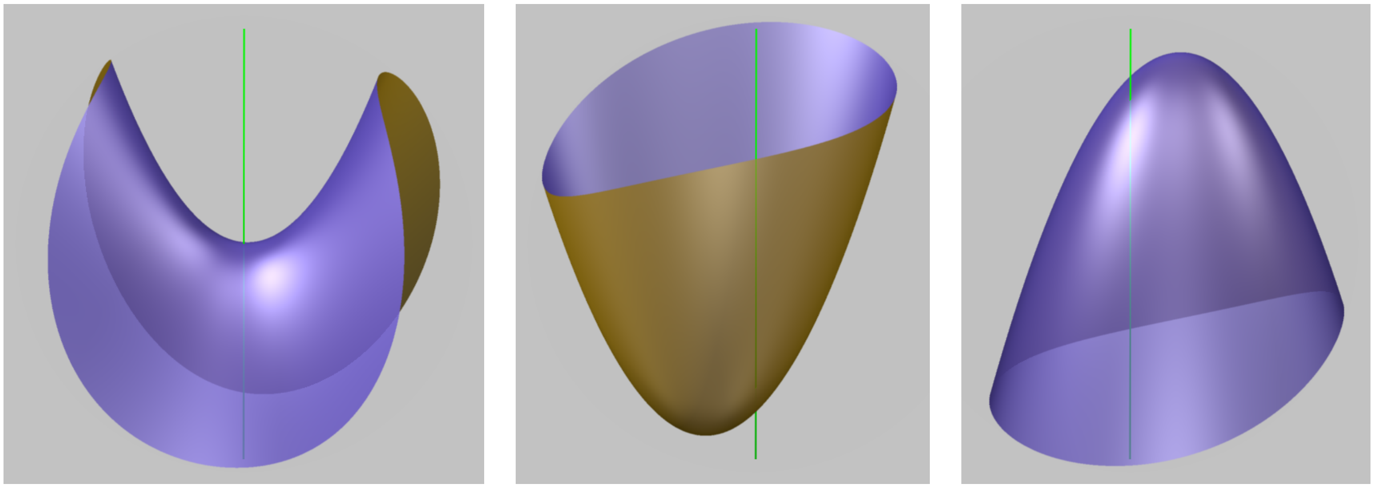
This concept was introduced for the first time by Ludwig Otto Hesse (1811-1874) in two articles published in 1844 and 1851, respectively.

Hessian provides very important information about critical points. If $\mathbf{a}$ is a critical point of $f$, i.e. $df_{\mathbf{a}} = 0$, 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.

[10] Ludwig Otto Hesse (1811–1874)



(a) $H_f(\mathbf{a}) < 0$ (b) $H_f(\mathbf{a}) > 0$ and $f_{xx} > 0$ (c) $H_f(\mathbf{a}) > 0$ and $f_{xx} < 0$

Figure 6: The graph of a function $f\colon D \to \mathbb{R}$, 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 $H_f(\mathbf{a}) < 0$, then $\mathbf{a}$ is a **saddle point**;[11]

(ii) If $H_f(\mathbf{a}) > 0$, then there are two further possibilities:

a) $\mathbf{a}$ is a **local minimum**[12] if $f_{xx}(\mathbf{a}) > 0$,

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

Note that the positivity of $H_f(\mathbf{a}) = f_{xx}f_{yy} - (f_{xy})^2$ 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 $H_f(\mathbf{a})$ does not vanish. If $H_f(\mathbf{a}) = 0$ then $\mathbf{a}$ 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)

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial g}{\partial y}\,dy = (2x+3y)dx + (3x+4y)dy$$

or, equivalently, the gradient of $f$ equals

$$\nabla f = \begin{pmatrix} 2x+3y \\ 3x+4y \end{pmatrix}.$$

A point $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ is a critical point of $f$ if and only if

$$\begin{cases} 2a_1 + 3a_2 = 0 \\ 3a_1 + 4a_2 = 0 \end{cases} . \tag{56}$$

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 $\mathbf{a}$ 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 o. 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 $\mathbf{0}$ is a saddle point while $\begin{pmatrix} -2 \\ 0 \end{pmatrix}$ is a local maximum.

**23 Another look at the definition of a critical point** In Section 19 we declare a point $\mathbf{a} \in D$ to be **critical** for a function $\mathbf{f} : D \to \mathbb{R}$ if the differential of $\mathbf{f}$ at $\mathbf{a}$ identically vanishes:

$$df_{\mathbf{a}} = 0 , \tag{57}$$

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

> a point $\mathbf{a}$ is critical for a function $\mathbf{f}\colon D \to \mathbb{R}$ if and only if $df_{\mathbf{a}}\colon \mathbb{R}^m \to \mathbb{R}$ is **not** onto. (58)

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

**24 Critical points of functions $f\colon D \to \mathbb{R}^n$** When is the image of a linear transformation $L\colon \mathbb{R}^m \to \mathbb{R}^n$ as big as possible? When $L$ is **onto**, of course. Yes, but this is possible only when $m \geqslant n$ . For $m \leqslant n$ , $L$ 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.

> A point $\mathbf{a}$ is a **regular** point of a vector function $\mathbf{f}\colon D \to \mathbb{R}^n$ if:
>
> **Case** $m \geqslant n$ . $\mathbf{f}'(\mathbf{a}) : \mathbb{R}^m \to \mathbb{R}^n$ is **onto**.
>
> **Case** $m \leqslant n$ . $\mathbf{f}'(\mathbf{a}) : \mathbb{R}^m \to \mathbb{R}^n$ is **one-to-one.** (59)

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

☞ We say that $\mathbf{a}$ is a **critical** point if $\mathbf{a}$ is not regular.[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, $\mathbf{f}'(\mathbf{a})$, of a function $\mathbf{f}\colon D \to \mathbb{R}^n$ vanishes on vectors tangent to the level set of $\mathbf{f}$ at point $\mathbf{a}$. This holds for any point $\mathbf{a}$. However, for points where $\mathbf{f}$ is *regular* the reverse is also true.

> If $\mathbf{a}$ is a regular point of a function $\mathbf{f}\colon D \to \mathbb{R}^n$ then the level set of $\mathbf{f}$ passing through $\mathbf{a}$ is **smooth** in the vicinity of $\mathbf{a}$. Moreover, $\mathbf{f}'(\mathbf{a})(\mathbf{v}) = 0$ if **and only if** vector $\mathbf{v}$ is tangent to the level set of $\mathbf{f}$. (60)

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* $m \geqslant n$. This contrasts with the case $m \leqslant n$, 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 $\mathbf{a}$, function $\mathbf{f}$ becomes *one-to-one* — exactly like its derivative $\mathbf{f}'(\mathbf{a})$ — and the image, $\mathbf{f}(N)$, is **smooth**. (61)

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 $\mathbf{n = 1}$:

$$\text{a point } \mathbf{a} \text{ is a regular point of } f : D \to \mathbb{R} \text{ if and only if } \nabla f(\mathbf{a}) \neq 0. \tag{62}$$

What about the case $\mathbf{n = 2}$? In this case $\mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$ and, assuming that $m$, i.e. the number of variables, is *greater* than 1, the answer is as follows.

$$\text{A point } \mathbf{a} \text{ is a regular point of a function } \mathbf{f}\colon D \to \mathbb{R}^2 \text{ if and only if the gradient vectors of its component functions } \nabla f_1(\mathbf{a}) \text{ and } \nabla f_2(\mathbf{a}) \textbf{ span a plane} \text{ in } \mathbb{R}^m. \tag{63}$$

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

**Case $m = 1$.** In the familiar case of a parametric curve $\gamma\colon I \to \mathbb{R}^n$, 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*[14] singularities like "cusps" or "corners".

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

$$J_{\mathbf{f}}(\mathbf{a}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \frac{\partial f_1}{\partial x_2}(\mathbf{a}) \\ \vdots & \vdots \\ \frac{\partial f_2}{\partial x_1}(\mathbf{a}) & \frac{\partial f_2}{\partial x_2}(\mathbf{a}) \end{pmatrix}. \tag{64}$$

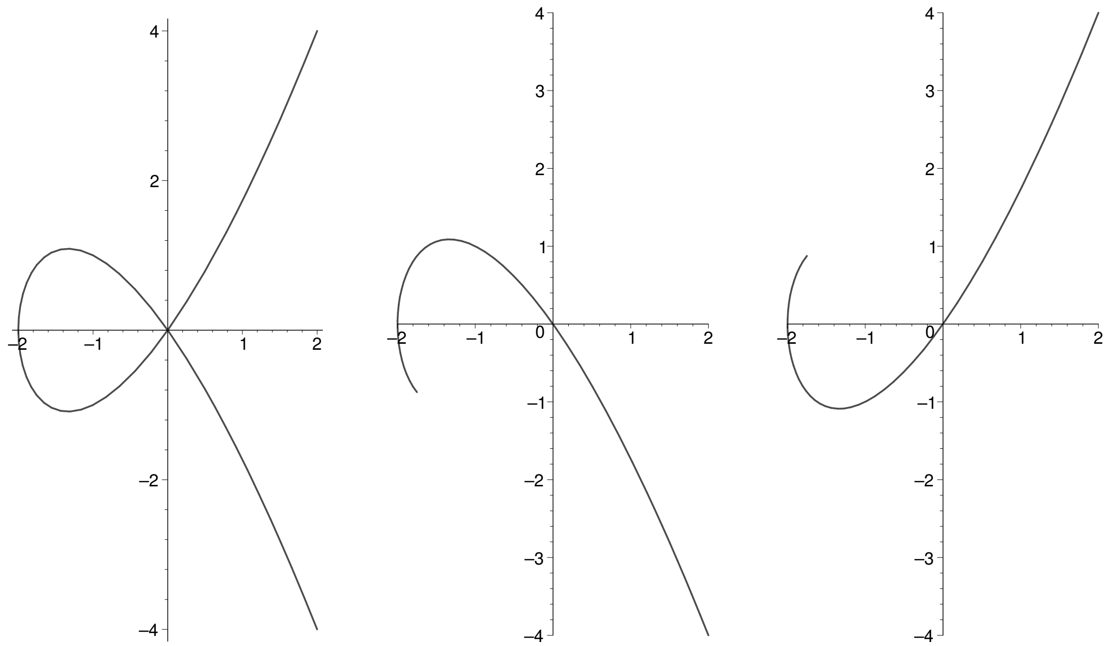
Assuming $n \geqslant 2$, we have the following characterization of regular points:

$$\text{A point } \mathbf{a} \text{ is a regular point of a function } \mathbf{f}\colon D \to \mathbb{R}^n, \text{ defined on a subset } D \subseteq \mathbb{R}^2, \text{ if and only if the two columns of Jacobi matrix (64) } \textbf{span a plane} \text{ in } \mathbb{R}^n. \tag{65}$$

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 $\mathbf{a} \in D$ is a regular point of function $\mathbf{f}$ when the Jacobi matrix, $J_{\mathbf{f}}(\mathbf{a})$, has**

[14] 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 $\boldsymbol{\gamma}$ (b) $\boldsymbol{\gamma}((-,1/2))$ (c) $\boldsymbol{\gamma}((-1/2,))$

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

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

*The image of* $\boldsymbol{\gamma}$*, i.e., set* $\boldsymbol{\gamma}(\mathbb{R})$*, has a singularity at the origin and* $\boldsymbol{\gamma}$ *is* ***not one-to-one,*** *since* $\boldsymbol{\gamma}(-\sqrt{2}) = \mathbf{0} = \boldsymbol{\gamma}(\sqrt{2})$*, see Subfigure (a). Function* $\boldsymbol{\gamma}$ *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.*

the largest possible rank.[15] When $m$ or $n$ equals $1$ the largest possible value of rank of $J_{\mathbf{f}}(\mathbf{a})$ is $1$. When the smaller of the two numbers $m$ and $n$ equals $2$, the largest possible value of rank of $J_{\mathbf{f}}(\mathbf{a})$ is $2$.

[15]**Rank** 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 $f\colon D \to \mathbb{R}$ along a path** Consider a path $\boldsymbol{\gamma}\colon I \to D$. We shall say that a function $f\colon D \to \mathbb{R}$ has, at a point $\mathbf{a} = \boldsymbol{\gamma}(a)$, *a local maximum (minimum) along path* $\boldsymbol{\gamma}$ if the composite function

$$f \circ \boldsymbol{\gamma}\colon I \to \mathbb{R} \tag{66}$$

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 $f \circ \boldsymbol{\gamma}$ at $a$ vanishes and we deduce from Chain Rule (22) — see also Exercise 5 and formula (34) — that

$$\boxed{df_{\boldsymbol{\gamma}(a)} \text{ annihilates the velocity vector } \frac{d\boldsymbol{\gamma}}{dt}(a), \text{ i.e. } \nabla f(\mathbf{a}) \cdot \frac{d\boldsymbol{\gamma}}{dt}(a) = 0} \quad . \tag{67}$$

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

**28 Local extrema of a function $f\colon D \to \mathbb{R}$ 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 $\boldsymbol{\gamma}\colon I \to Z$ is *any* differentiable path passing through a point $\mathbf{a} = \boldsymbol{\gamma}(a)$ — where function $f$ has its local maximum or minimum on $Z$ — then differential $df_{\mathbf{a}}$ annihilates velocity vector $\frac{d\boldsymbol{\gamma}}{dt}(a)$.

Now, any vector tangent to $Z$ at point $\mathbf{a}$ 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 $\mathbf{a}$ then $df_{\mathbf{a}}$ **vanishes on all vectors tangent** to $Z$ at point $\mathbf{a}$. (68)

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

> If $f$ has a local extremum at an *interior* point $\mathbf{a}$ then $df_{\mathbf{a}}$ **is zero,** i.e. $\mathbf{a}$ is a **critical point** of the function $f$. (69)

**29 Example** Let $f\colon E \to \mathbb{R}$ be a function on the ellipse

$$E := \left\{ \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; \left(\frac{x-c_1}{a}\right)^2 + \left(\frac{y-c_2}{b}\right)^2 \leqslant 1 \right\}. \tag{70}$$

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

$$\begin{cases} \nabla f(\mathbf{x}) \cdot \begin{pmatrix} a^2(y-c_2) \\ -b^2(x-c_1) \end{pmatrix} = 0 \\ \left(\frac{x-c_1}{a}\right)^2 + \left(\frac{y-c_2}{b}\right)^2 = 1 \end{cases}. \tag{71}$$

The *second* equation expresses the fact that point $\mathbf{x}$ belongs to the boundary, $\partial E$, of ellipse $E$. The *first* equation expresses the fact that $df_{\mathbf{x}}$ 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 $f\colon D \to \mathbb{R}$ where argument $\mathbf{x}$ is subject to a number of side conditions:

$$g_1(\mathbf{x}) = k_1 \;,\; \ldots \;,\; g_r(\mathbf{x}) = k_r \tag{72}$$

☞ 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:

$$\mathbf{g}(\mathbf{x}) = \mathbf{K} \tag{73}$$

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

$$Z = \{\mathbf{x} \in D \mid \mathbf{g}(\mathbf{x}) = \mathbf{K}\}. \tag{74}$$

Theorem (68) tells us that $df_{\mathbf{a}}$ vanishes on vectors tangent to $Z$ at a point $\mathbf{a}$ if function $f$ has a local extremum on $Z$ at $\mathbf{a}$. If $\mathbf{a}$ is a **regular** point of vector-constraint function $\mathbf{g}$ then its derivative $\mathbf{g}'(\mathbf{a})$ vanishes precisely on vectors tangent to $Z$.

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

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

$$df_{\mathbf{a}} = \Lambda \circ \mathbf{g}'(\mathbf{a}) \,. \tag{75}$$

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

$$\Lambda(\mathbf{v}) = \boldsymbol{\lambda} \cdot \mathbf{v} \qquad\qquad (\mathbf{v} \in \mathbb{R}^r) \tag{76}$$

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

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

$$\boxed{\nabla f(\mathbf{a}) = \lambda_1 \nabla g_1(\mathbf{a}) + \cdots + \lambda_r \nabla g_r(\mathbf{a})} \,. \tag{77}$$

Equality (77) expresses the fact that gradient vector $\nabla f(\mathbf{a})$ is a *linear combination* of gradient vectors $\nabla g_1(\mathbf{a}), \ldots, \nabla g_r(\mathbf{a})$ 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 $\mathbf{a}$ where function $f$ has a local extremum *with r constraints* (72), gradient vector $\nabla f(\mathbf{a})$ can be expressed as a linear combination (77) of gradient vectors $\nabla g_1(\mathbf{a}), \dots, \nabla g_r(\mathbf{a})$ for *suitable* numbers $\lambda_1, \dots, \lambda_r$ **provided** $\mathbf{a}$ is a **regular** point of the vector-constraint function:

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_r(\mathbf{x}) \end{pmatrix}. \tag{78}$$

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 $r \geqslant m$ then the level sets of all regular points of $\mathbf{g}$ 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 $f$.

◇
Typeset in 12pt Lido STF CE, a True Type font designed by František Štorm of Czech Republic and implemented to use with pdfeLaTeX by tha author
◇