Ju. I. Neimark

Mathematical Models in Natural Science and Engineering

Springer

Foundations of Engineering Mechanics

Ju. I. Neimark

Springer

Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Juri I. Neimark

Mathematical Models in Natural Science and Engineering

Translated by V. I. Gloumov and M. M. Kogan

With 223 Figures

Springer

Series Editors:

V. I. Babitsky
Department of Mechanical Engineering
Loughborough University
Loughborough LE11 3TU, Leicestershire
Great Britain

J. Wittenburg
Institut für Technische Mechanik
Universität Karlsruhe (TH)
Kaiserstraße 12
76128 Karlsruhe / Germany

Authors:

Ju. I. Neimark
Nizhny Novgorod State University
after N. I. Lobatchevsky
Gagarin Ave., 23
603950 Nizhny Novgorod
Russian Federation

Translators:

V. I. Gloumov
Research Institute for Applied Mathematics
and Cybernetics of Nizhny Novgorod
State University
Ulyanov Str., 10
603005 Nizhny Novgorod
Russian Federation

M. M. Kogan
Architecture and Civil Engineering University
Il'yinskaya Str., 65
603950 Nizhny Novgorod
Russian Federation

Dedicated to
my mother,
Elena Nickolaevna,
my father,
Isaac Goustavovitch,
and my wife,
Valentina Vasilyevna

without whom I would not live in this world,

and also to

Alexandr Alexandrovitch Andronov

who has opened for me a way to science.

PREFACE

This book has come into being as a result of the author's lectures on mathematical modelling rendered to the students, BS and MS degree holders specializing in applied mathematics and computer science and to post-graduate students in exact sciences of the Nizhny Novgorod State University after N.I. Lobatchevsky. These lectures are adapted and presented as a single whole about mathematical models and modelling.

This new course of lectures appeared because the contemporary Russian educational system in applied mathematics rested upon a combination of fundamental and applied mathematics training; this way of training oriented students upon solving only the exactly stated mathematical problems, and thus there was created a certain estrangement to the most essential stages and sides of real solutions for applied problems, such as thinking over and deeply piercing the essence of a specific problem and its mathematical statement. This statement embraces simplifications, adopted idealizations and creating a mathematical model, its correction and matching the results obtained against a real system. There also existed another main objective, namely to orient university graduates in their future research not only upon purely mathematical issues but also upon comprehending and widely applying mathematics as a universal language of contemporary exact science, and mathematical modelling as a powerful means for studying nature, engineering and human society.

The author of this book is a pupil of A.A. Andronov's world-known scientific school on nonlinear oscillations theory. This very fact has determined the choice of the models under consideration. They are evolutionary models, i.e. the models for time-varying processes: dynamics, control, training, computational processes, recognition and optimization. These models describe systems of various nature, viz. mechanical, physical, chemical, biological, technical, ecological, social, etc. A unifying basis for this entire collection is a universal all-embracing mathematical model, i.e. a dynamical system.

The book may be interesting and useful for a vast spectrum of readers. To the students of various fields of natural science, engineering and other educational establishments, to researchers encountering difficulties in modelling or facing the situations where modelling may be of use, and also to those being curious and willing to find out how mathematics studies the surrounding world making it possible for us to understand and explain mysterious phenomena, foresee future changes, comprehend properties and foretell the behaviour of not only existing systems and processes but also those that can be imagined or invented.

To its utmost degree, modelling is an art and so beauty plays in it not a last role, the beauty of the model, the beauty of studying it, and, at last, the beauty of explaining through it the mysterious and nontrivial phenomena and properties of the system under study. Perhaps, creating and studying some models this book suggests will bring you some aesthetic pleasure.

CONTENTS

Introduction

In this book you will become familiar with various mathematical models for mechanical, electrical, physical, astronomical, chemical, biological, ecological, cybernetical and other systems and processes. In addition, this book will help you to form your own scientific understanding and attitude towards "applied mathematics", and other mathematical subjects you are studying.

You would like to study and comprehend the surrounding world and to this you are driven by your heredity. Why are you then trained in abstract and hard understandable mathematics isolated from the living world, i.e. in numbers, vectors, matrices, functions, operations upon them, differentiating, integrating, etc ?

I guess you know how to answer this question in general – mathematics is a basis for the scientific study of the world. However, why is this abstract mathematics lying so far from our specific, touchable and beautiful world so essential for world study? Why is it not, say, astrology that is capable of foretelling the future on the basis of positions of stars?

What is the method for applying mathematics to studying the world and foretelling the future?

Perhaps, for this last question you also know the answer – we are comprehending the world through creating and studying its mathematical models.

So together we shall construct mathematical models and study them, and in this way we shall pierce the secrets of the world we are living in.

What is a model? The word "model" is borrowed from Italian and means "a copy", "a template", "a prototype" upon studying which we have the original already studied. Therefore, a mathematical model makes up a prototype of some fragment of the world – that of some system, device, machine, process, apparatus, and through studying this prototype we are understanding this fragment of the world.

What is then mathematics and what is its role in creating this auxiliary model? Why can this model be studied by us? Why then upon studying it are we capable of studying the world? Finally, what is, after all, mathematics? I think you are not aware of all of the answers to these questions.

To the answers to these questions we shall be approaching from far away.

Once long ago, in the last century at one scientific sitting there was noisily discussed the issue as to the role of languages and mathematics in training students. The debates lasted long. One side stood up for the role of languages, and another for the importance of mathematics. Among disputants was the great scientist J. W. Gibbs. This is the very scientist who stated that a whole is simpler than its

part; that to study some accumulation of milliards of molecules is simpler than a single molecule. He did not only state but proved it by his wonderful scientific discoveries. He was respected by everybody and was a great taciturn. However here, to everybody's astonishment, he asked to give him the floor and stated: "Mathematics is a language".

What did he imply? What is common between Russian, English, Greek, French and other languages, and mathematics? This common lies in the following: any natural language describes human actions, feelings, wishes, reminiscences, suggestions, opinions, orders, etc., and mathematics is a language for describing nature; it is a language of scientific knowledge of the world. If you wish to speak to a Frenchman, you need to have the French language studied, to speak to an Englishman the English language, and to nature, the mathematical language. Nature discloses its mysteries to us only in terms of the mathematical language, and if you wish to comprehend these mysteries you need to have the mathematical language studied, to have mathematics studied. Speaking to nature and engineering, being also a portion of nature but created by man, is performed in the mathematical language.

Human languages are specific; each word in them is assigned some specific meaning. In contrast to this, the mathematical language is abstract. However, they both are languages and this difference is not so essential. Imagine that you have seen an inscription on the fence:

<p align="center">"Nhiseful mitteler rims qoptly"</p>

At first, you took it as something gibberish, but upon thinking it over you realise that there exists a certain "mitteler" which is "nhiseful", and that this "mitteler" "rims" and performs this action "qoptly". Right now, match this phrase against the following mathematical phrases

$$(a+b)^2 = a^2 + 2ab + b^2,$$

$$0 < \frac{d}{dt}x(t) < \frac{d}{dt}y(t),$$

$$\int u\,dv = uv - \int v\,du.$$

In the second mathematical phrase it is said that some magnitude $x(t)$ increases with t slower than the magnitude $y(t)$ does. What x and y represent is not known, nor what a and b are equal to in the first phrase.

Linguistic descriptions of our life – stories, life-stories, novels – are its language models. Descriptions of natural phenomena in the mathematical language are their mathematical models, i.e. mathematics is a language of exact sciences, and mathematical models are a description of systems and processes of nature or engineering in terms of the mathematical language.

For writing a good school composition you need to know well what you are writing about. The same owns for mathematical models – to have a robust mathematical model you need to have a clear idea about what you want to describe; to distinguish essential from secondary. In addition, you need to understand the es-

sence and regularities of the running processes, phenomena and their interrela-
tions. Finally, you need to be skilful in describing all these things in the mathe-
matical language.

Writing a composition is not simple, but reading and understanding it is usually
simple. Though at times this simplicity is deceptive. Compiling a model is also
rather difficult, and understanding it is sometimes not so simple, and at times it
requires even a lot of effort. To understand the model and what it describes, it is
necessary to have it studied, and this study may turn out to be very complicated
and difficult.

Now I think you understand in general what my book is about.

However, there remains the following question not yet answered: why in
studying a mathematical model of a real system or process do we also get the pos-
sibility to study them themselves? As said above they are so different and seem-
ingly have nothing in common. This question is not so simple. One of the well-
known mathematicians called this wonderful possibility "incomprehensible".
Though, this seems not absolutely so. Something here can be comprehended.

Everything here rests upon isomorphism. We shall touch this point later on, but
for now still note only that differences between a real system and a model are not
so terrible. How a TV set runs can be understood through its radio circuit looking
absolutely different from the TV set itself. Streets in the city can be made out
through its maps.

Now I am coming to the story about one of basic models in natural science and
engineering, a dynamical system.

1 Dynamical systems

Mathematics as a language. Mathematical models. Mathematical modelling as a method of studying the surrounding world. The Laplace determinism. A dynamical system as a basic mathematical model in the natural sciences. A phase portrait. Examples of dynamical systems and their phase portraits. A phase portrait as a means of geometrical representation of our knowledge about a dynamical system and as a means of its study.

Archimedes, having perceived the law of the lever and being excited by the mystery of nature unveiled to him, exclaimed: "Give me a fulcrum and I shall turn over the world".

Many centuries ago, in the 18th century, the great mathematician and astronomer Pierre S. Laplace, being impressed by a revealed scientific picture of the universe and by causal relationships between all phenomena, expressed his enlightment with the words: "I shall predict in all details the entire future, everything that will come for all centuries ahead, from the smallest to the biggest things, if you tell me or completely describe what state the contemporary world is in".

It was an apotheosis of determinism, a full causative relationship of the future with the present.

Clear it is that nobody could accuse Laplace of lying or being wrong, since who is able to describe the universe in full? It would be nice to describe a portion even partially. And what then? Isn't a human being capable of predicting ? Well, sometimes prediction is possible. Now let us try to clear it up for ourselves when it is possible and when not, and what it depends upon. Meanwhile, let us understand why it was stated so by Laplace.

It was stated in this way thanks to great astronomical achievements. In astronomy, a full understanding of the causes guiding the movement of planets was gained and it became possible to calculate planets' orbits for many years ahead. Alongside, a mechanistic picture of the universe was established, in which all things occurring in the world were explained through a mechanical motion of the tiniest material particles, molecules. Such are the caprices of the history that the greatest enlightments came into science "from the heavens", thanks to the astronomical achievements.

The movement of planets, the travelling stars, attracted the attention of mankind for a long time. At present, we have nothing to do with them, but in that distant past it was far from so. Well known that time was the Ptolemaic system of the universe. It described sufficiently well visible the movements of planets (Mercury,

Venus, Mars, Moon, Jupiter, Saturn) and of the Sun. A more simple description was given by the Copernicus system, though it was Newton who explained them via his laws for the mechanical motion and universal gravity. He revealed that this mechanical motion of a material point is described by a second-order differential equation and is determined uniquely through its initial position and velocity. From this it followed that the movement of all planets and of the Sun can be exactly calculated if their masses, initial positions and velocities are known. The calculation itself was carried out by Newton only for two mutually attracting free solids, but in principle this is not essential. Very soon people learned how to approximately calculate, but with a magnificent accuracy, the movement of all planets. Nowadays this is done easily with use of computers. Thus, there exists some description for planets through which the future can be predicted, i.e. how this description will vary in the future. Such a description was called a state of the mechanical system. To some extent this notion of a state can be generalized and applied not to mechanical systems only. It is in this way that exact natural sciences proceeded to study evolutionary processes and predict the future.

The idea of determinism that so much excited Laplace, the idea of a temporal causality and a unique conditionality of events in our world, can be thought as a token of relationships existing between the descriptions of the past and the present.

Here it is senseless to talk about the relation of the entire past with the entire present, since a possibility to actually investigate these relations arises only when a certain part of something is chosen both from the description and the medium. Choosing a subject for investigation in such a way was described by Newton as concentrating attention upon separate attractive beautiful pebbles on a boundless pebbly sea-shore. The "pebble" we will choose will be called a system, and what in it catches our attention will be called a description of the system. This description may be denoted by x. It is time-dependent, and it is not obligatory that the principle of determinism should be true for it, i.e. the fact that the past description $x(t_1)$ defines the future description $x(t_2)$, $(t_2 > t_1)$. This is so only when we are very lucky both in choosing a pebble and in choosing what is attracting us in it.

Now I am coming to some examples on how to choose the "pebbles" and describe them. Prior to this, I would like to tell you some more about the Laplace determinism, to protect him, so to say, against natural claims that commonly observable randomness was not supposedly observed and ignored by him. No, he observed it. Moreover, it was he who introduced into science the well-known random magnitude distributed by the widely employed normal law, the Laplace law. Though, as well as Albert Einstein, he did not believe that God plays dice. He did not know how to explain the observable randomness, but this did not hamper him accepting the deterministic picture of the world. That was right, since random processes admit a deterministic description, but we shall touch on this point later. Now we return to our previous talk.

Here is a small piece of chalk in my hands. I may be interested in knowing how it writes on the blackboard and why it does this. It means that I shall be interested

in its property and its description from the point of view of its capability for writing on a blackboard. However I can use this chalk for simply throwing it or releasing it and see how it is falling down. In this last case I can be lucky and reveal a law for heavy solids falling onto earth, thus revealing a triumph of determinism.

I can also pay attention to a glass of hot tea and ponder over why it has such a taste and smell and what it is. I can be curious about when the tea at last gets cold and one can drink it. In the last case, I am interested only in the tea temperature described by the Centigrade scale. If I am persistent enough, I shall find, perhaps, a law of its decrease, thus confirming a general idea of determinism.

Similar examples can be given endlessly – a burning candle that attracted Michael Faraday's attention; the motion of Earth around the Sun studied by Johannes Kepler and Isaac Newton; an electrical circuit of a capacitance and a self-inductance, whose mystery of oscillation was unveiled by Benjamin Thompson; an atom whose first model (description of the structure) was suggested by Nils Bohr; Earth's atmosphere that affects weather which we are still not able to forecast; a living organism that until now is yet full of secrets for us; and the absolutely mysterious human brain.

Upon these sketchy examples let us now consider some in more details. Let us follow two cyclists on a cycle track. They started at the same time, and he who finishes first will be winner. How much time they ride does not matter, and this yields the situation that neither rushes forward but is manoeuvring to try to deceive his rival and arrive first at the finish. Simply to dash to the finish headlong is not good, since his rival can stay at his tai spending substantially less efforts, and will easily outrun him before the finish.

To describe this situation, let us take the angles φ_1 and φ_2 to represent the first and second cyclists (figure 1.1).
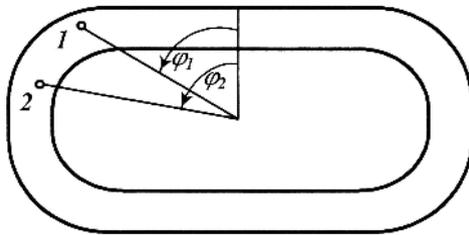
Fig.1.1. A view from above of the racing track and a racing event of the cyclists 1 and 2.

The plots of the functions $\varphi_1(t)$ and $\varphi_2(t)$ provide a good representation of how the race was running (figure 1.2). For this case, two curves are drawn. The course of the competition can be expressed through a single curve in the plane φ_1, φ_2. (Here a start corresponds to the angle $\varphi = 0$ and a finish to $\varphi = 2\pi$).
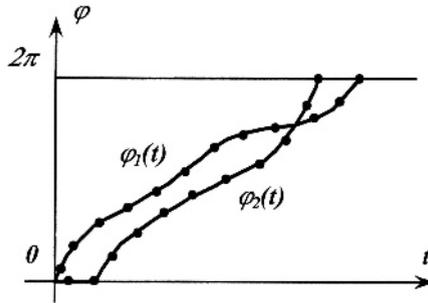
Fig. 1.2. The time plots of $\varphi_1 (t)$ and $\varphi_2 (t)$ for one of the races.

Upon some time, the point M of the coordinates φ_1 and φ_2 will describe some curve (figure1.3). Its form also demonstrates how the competition was running.
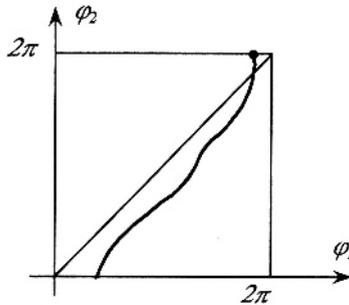
Fig. 1.3. Depicting a race of the cyclists on the (φ_1, φ_2) – plane.

According to figure 1.3, the first to leave the start line was the first cyclist, and the second cyclist was behind his rival all the way until the finish, and just before the finish he overran the rival and came first.

In figure 1.4 another variant of the racing competition is presented, where cyclists overran each other many times and the first was the winner.

It is not difficult to see that the above description through the angles φ_1 and φ_2 will not allow, until one of them is equal to 2π, to uniquely predict an outcome of the cyclist racing, though a skilful specialist could express more or less likely estimates.
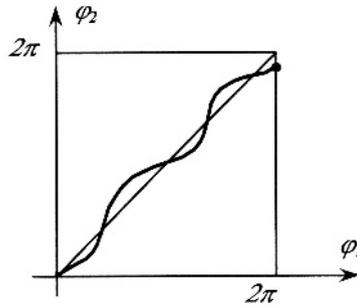
Fig. 1.4. Another race of the cyclists on the (φ_1, φ_2) – plane.

Let us consider one more example pertaining to describing a disease through a plot of a patient's temperature (figure 1.5). An experienced doctor will find in this plot a lot of information concerning the flow of the disease, though the initial fragment of the plot is insufficient for a trustworthy prediction of the entire curve or a final stage of the disease only.

The next example is a free vertical falling of a solid body at height h_0 above the earth and having vertical velocity v_0. Under the gravitational law, we get

$$h = h_0 + v_0 t + \frac{gt^2}{2} \, ,$$

where t is the time of the fall, and h is the position counted down along the vertical line. The admissible types of the plots for $h(t)$ are presented in figure 1.6. These plots show how the solid is falling when it has been thrown up, simply dropped or thrown down from height h_0. Therefore, the magnitude h describes the fall of the solid well. Is it possible through a single h to predict a further fall of the solid body? Evidently not, since the magnitude $\dot{h}(t)$ must be known as well.

Having this in mind, let us take two plots for $h(t)$ and $\dot{h}(t)$ (figure 1.7) as a description. With values of $h(t)$ and $\dot{h}(t) = v$ known at any instant t, it becomes possible to predict their values at any subsequent instant $\bar{t} > t$ with use
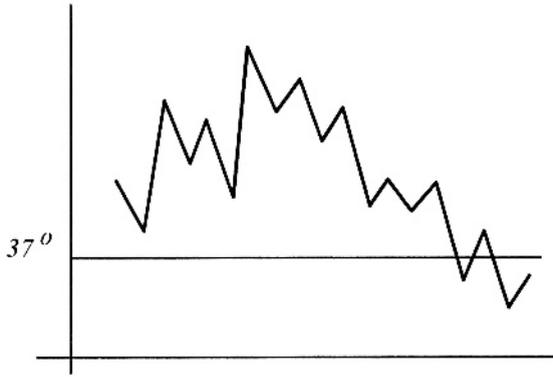
Fig. 1.5. The plot of morning and evening temperatures of the patient.
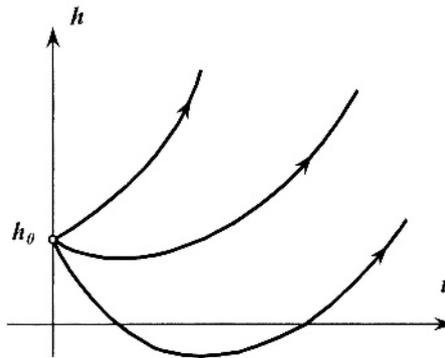
Fig. 1.6. The plots of the body falling from the same height but at different initial velocities.

of the below formulae

$$h(\bar{t}) = h(t) + v(t)(\bar{t} - t) + \frac{1}{2} g(\bar{t} - t)^2$$

$$(1.1)$$

$$\dot{h}(t) = \dot{h}(t) + \dot{g}(\bar{t} - t) \quad ,$$

being familiar to you from school physics.
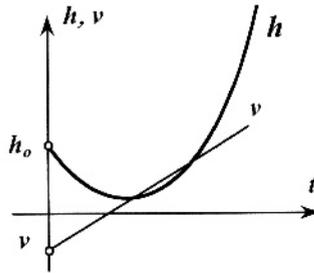
Fig. 1.7. The time plots of the height h and the velocity v for the body thrown vertically up.

Instead of two plots $h(t)$ and $\dot{h}(t)$, we can take only a single curve being travelled with the time t in the plane h, v by the point M with coordinates h and v

$$h = h(t) \ , \ v = \dot{h}(t) \, .$$

Equations for this curve may be derived in the following way.
From the obvious expressions

$$v = \dot{h} \ , \ \dot{v} = g \ , \tag{1.2}$$

implying that a velocity is a time derivative of the displacement and that the acceleration due to gravity is equal to g, it will follow that

$$\frac{dh}{dv} = \frac{v}{g}$$

or

$$h = \frac{v^2}{2g} + C \ ,$$

where C is an arbitrary constant.

Hence, the trajectories circumscribed by the point $M(h, v)$ on the plane h, v for various C are of the form shown in figure 1.8.
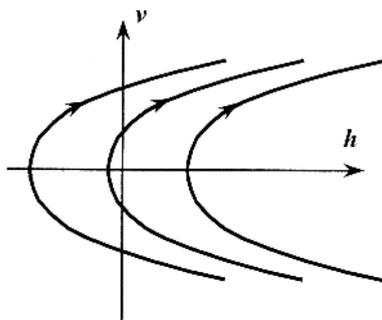
Fig. 1.8. The phase portrait of the body fall.

When time increases, the point M will run along these curves in the directions indicated by arrows (it follows from the fact that for $v > 0$ the magnitude h increases and for $v < 0$, decreases). The description of a falling solid via two magnitudes h and v, adopted by us, possesses a wonderful property of self-sufficiency and enables to produce a unique prediction. Indeed, setting any h and v will define uniquely a single parabola with the point $M(h, v)$ on it. In this way, a further evolution of magnitudes h and v is defined uniquely.

The plane of the variables h, v has a special name, a phase plane. Accordingly, the variables h and v are called phase variables, and the trajectories being drawn by the points $M(h, v)$ on this phase plane will be called phase trajectories.

The motion of the phase point along trajectories is remarkable by the fact that knowledge of its position at any instant t makes it possible to find its position to which it will have moved at any subsequent instant $\bar{t} > t$, and namely, according to (1.1) or (1.2),

$$v(\bar{t}) = v(t) + g \cdot (\bar{t} - t) \ ,$$

$$h(\bar{t}) = h(t) + v(t)(\bar{t} - t) + \frac{1}{2} g(\bar{t} - t)^2 \ .$$

This property of self-sufficiency of the variables h and v, implying that in order to determine their values upon any time interval Δt it is enough to know their values at an initial moment, makes them especially important. Due to this, such a description was called a state, and the variables describing this state were called phase variables. This very property of the state was used as a basis for defining a mathematical model being called a dynamical system. This model describes the determinative evolutionary processes and is one of the most essential, if not the most essential, in contemporary natural sciences and engineering.

Let us look once more at the description of the cycling competition on the plane φ_1 and φ_2 and that of a freely falling body on the plane h, v. In the first case, descriptions of different racing heats are represented by all possible curves within the square $0 \leq \varphi_1 \leq 2\pi$, $0 \leq \varphi_2 \leq 2\pi$. Naturally we assume only that φ_1 and φ_2 are nondecreasing functions of time. Through each point of this square may pass across any number of curves. It is due to this very fact that setting the values φ_1 and φ_2 at some instant of time will not determine a further running of the curve passing through this point. On the contrary, on the plane h, v the curves depicting different cases of falling will not intersect each other. It is due to their inability to intersect, and due to each point being intersected by its unique curve, that a unique prediction of further changes of magnitudes h and v becomes possible.

Thus, description by the variables φ_1, φ_2 does not provide a unique prediction and this very fact is reflected in that that each point (φ_1, φ_2) may be passed by some curves. On the other hand, describing a free falling object through the variables h and v will produce a determinism, because each point (h, v) is intersected with a single curve only. The given mathematical description for a solid freely falling in a gravity field constitutes an example of a mathematical model called a dynamical system, whereas a mathematical description of the competing cyclists through the angles φ_1 and φ_2 does not.

Now let us give a general abstract definition of a mathematical model, called a dynamical system, with use of geometrical interpretation suggested by the great mathematician Jules H. Poincaré.

A dynamical system is defined by the space X and the single-valued operator $T(\Delta t)$, depending upon the parameter $\Delta t \geq 0$ in such a way that to each point $x \in X$ the operator T puts into correspondence the point $\overline{x}$, i.e. $\overline{x} = T(\Delta t)x$. Here the operator $T(\Delta t)$ is supposed to satisfy, for any admissible $\Delta t_1 \geq 0$ and $\Delta t_2 \geq 0$, the expression

$$T(\Delta t_2)\, T(\Delta t_1) = T(\Delta t_1 + \Delta t_2).$$

The notional matter of the space X and the operator $T(\Delta t)$ is as follows. X is a space of all possible states of the system under study. Usually x is a multidimentional vector with the components $x_1, x_2,..., x_n$. The description x is called a state or a phase point, and a space of states X is called a phase space. For the given state x at an initial moment, the operator $T(\Delta t)$ determines the state $\overline{x}$ at the time Δt. Here it is clear that a transference from the state x to $\overline{x}$, being executed first during the time $\Delta t_1 \geq 0$ and then during the time $\Delta t_2 \geq 0$, should be the same as that performed during the time $\Delta t_1 + \Delta t_2$. This is the meaning of the above requirements assigned to the operator $T(\Delta t)$.

The description x is remarkable in the sense that by knowing it at the present time, one can determine such a description upon any time $\Delta t \geq 0$. It is because of this that it is called a state, and the point x a phase point.

The above discussed example concerning a freely falling body is completely appropriate to the definition given. The state x here is a two-dimensional vector with the components h and v. A space of states or a phase space is a two-dimensional plane of variables h and v. The height h and the velocity v vary during the time Δt and, according to the above, their new values will be equal to $\overline{h}$ and $\overline{v}$, where

$$\overline{h} = h + v\Delta t + \frac{1}{2} g(\Delta t)^2 \ , \qquad \overline{v} = v + g\Delta t \ . \tag{1.3}$$

These last formulae determine the operator $T(\Delta t)$. It may be immediately checked that this operator $T(\Delta t)$ meets the requirement imposed by the definition of a dynamical system.

From the relations determining a change of variables h and v for Δt, it follows that

$$\frac{\overline{v} - v}{\Delta t} = g \ , \qquad \frac{\overline{h} - h}{\Delta t} = v + \frac{1}{2} g\Delta t \ .$$

As $\Delta t \to 0$, we find that the functions $h(t)$ and $v(t)$ satisfy the differential equations

$$\frac{dh}{dt} = v \ , \qquad \frac{dv}{dt} = g \ .$$

This conclusion also takes place in a common case if the limit

$$\lim_{\Delta t \to 0} \frac{T(\Delta t) - T(0)}{\Delta t} = L$$

is supposed to exist. Then the state x as a function of the time t satisfies the differential equation

$$\frac{dx}{dt} = Lx .$$

Indeed, let us write the expression

$$\overline{x} = T(\Delta t)x$$

in the form

$$x(t + \Delta t) = T(\Delta t)x(t)$$

and further in the form

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} = \frac{T(\Delta t)x(t) - x(t)}{\Delta t} \quad ,$$

or, taking into account

$$x(t) = T(0)x(t)$$

and passing to the limit as $\Delta t \to 0$, we conclude that

$$\frac{dx(t)}{dt} = \lim_{\Delta t \to 0} \frac{T(\Delta t) - T(0)}{\Delta t} x(t) = Lx(t) \quad ,$$

as had to be proved.

Thus, a change of the state of a dynamical system will satisfy some system of first-order differential equations that, for the components of the vector x, can be written as

$$\frac{dx_1}{dt} = l_1(x_1, x_2, ..., x_n) \ ,$$

$$.\qquad .\qquad .\qquad .$$

$$\frac{dx_n}{dt} = l_n(x_1, x_2, ..., x_n) \ .$$

On the contrary, integrating these equations enables us to find the state-transferring operator $T(\Delta t)$. Here, these equations are certainly assumed to be uniquely solvable, and then their solution $x_i(t)$ ($i = 1, 2, ..., n$) at any instant $t \geq t_0$ is defined by the initial conditions, i.e. by the values $x_i(t_0)$ ($i = 1, 2, ..., n$). In this way, the assignment of the differential equations, that are satisfied by the state as a function of the time t, determines a dynamical system operator. The opposite takes place only in the case of differentiability of the dynamical system operator, and this may be often not the case. Therefore, the above definition for a dynamical system is somewhat more general than in the case when changes of a dynamical system state are supposed to be subject to some differential equations. Note also that the operator $T(\Delta t)$ can be defined not for all $\Delta t \geq 0$, but only for some set of values $\Delta t \geq 0$.

A central and visual geometrical image of a dynamical system is its phase portrait depicting all possible motions, i.e. all possible time evolutions of its states (descriptions).

A knowledge about a phase portrait gives a full representation of the dynamicals (possible changes) for the dynamical system; it is a portrait of its dynamics.

We already encountered a phase portrait in the two-dimensional plane h, v when describing the fall of a body. Each separate fall was depicted on this plane by the parabola

$$h = \frac{v^2}{2g} + C \ .$$

The set of all possible falls is described by parabolas for various values of C. This has led to figure 1.8. Separate curves of this phase portrait are called phase trajectories, and an entire set of possible phase trajectories constitutes a phase portrait.

On the basis of this obvious example it is not difficult to give general definitions for a phase trajectory and a phase portrait and reveal the most essential property of phase trajectories, viz. the fact that they cannot bifurcate, i.e. from a phase point there can necessarily come out no more than a single phase trajectory.

Indeed, let x be an arbitrary point of the phase space X. During the time Δt the point will displace to the point $\overline{x}$. With Δt changing from 0 to ∞, the point $\overline{x}$ will circumscribe some "curve" coming out from the point x. This is the very phase trajectory coming out from the phase point x. This phase trajectory is single, because of the uniqueness of the operator $T(\Delta t)$. A combination of all phase trajectories makes up a phase portrait.

Let us illustrate the notion of a phase portrait on the examples describing a sledge riding. Here, where we shall observe the two types of profiles, a pit with an ever ascending edges and a pit on a horizontal surface. The profiles are given in figures 1.9a and 1.9b. The sledges are taken symmetrical and sliding without friction, both forward and backward.
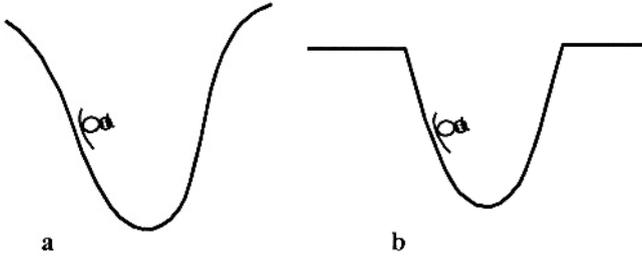
Fig. 1.9. The two types of skiing profiles, **a** and **b**.

The description has to be chosen in such a way as to have it as a state. With the help of the experience already gained in the example about a free fall, it is now natural for us to choose the position s and the velocity v of the sledge as a description. The variable s is chosen along the horizontal line, and the deepest place of the pit is taken as the origin of the count. The velocity of the sledge along the terrain profile will be chosen as the velocity v. No formulae and equations will be written. Without their help, let us think about what form the phase portrait will have, i.e. what phase trajectories in the plane s, v will be. In the case of the pit with ever ascending edges, any motion of the sledge will be represented either by the motionless position of the sledge in the bottom of the pit or by a periodic driving from one slope to another. The associated phase portrait is depicted in figure 1.10 a. This portrait includes oval, closed, nested phase trajectories.
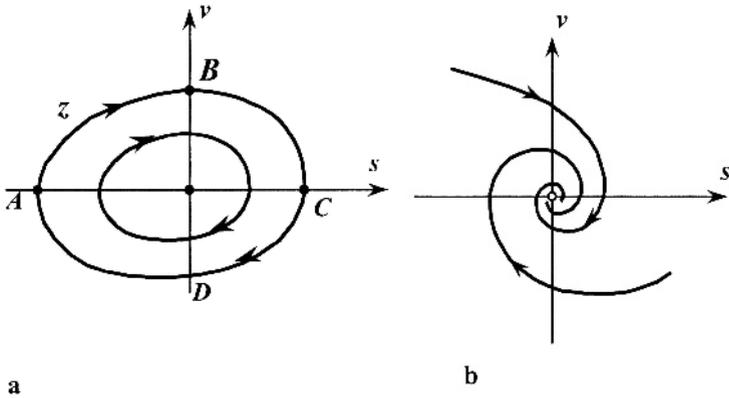
Fig. 1.10. The phase portraits of skier's movements corresponding to the profile in figure 1.9a , **a** (without friction) and **b** (with friction accounted).

Inside them, there lies the phase trajectory represented by the single point $s = 0$, $v = 0$ corresponding to the equal position of the sledge on the bottom of the pit. The points A, B, C and D on the phase trajectory Z respectively indicate the subsequent positions held by the sledges, i.e. in the extreme left top position, on the bottom of the pit, in the extreme right top position, again on the bottom of the pit, and, at last, again in the extreme left top position. Such motion is periodically repeated with no limit.

The second case of the phase portrait depicts a more complicated situation, since along the horizontal parts the sledge moves with a constant velocity and each of the points in horizontal parts for $v = 0$ is an equilibrium state. This fact results in a phase portrait given in figure 1.11a.

Let us ponder over how the phase portraits in figures 1.10a and 1.11a will change, if ever existing friction is also accounted. In this case, the oscillations inside the pit will dissipate always and will convert to the equilibrium on the bottom of the pit. The motion along the horizontal part will dissipate as well. If all this taken into consideration, then the phase portraits in figures 1.10a and 1.11a will assume the forms shown in figures 1.10b and 1.11b.

Now I think you are already ready to find a rather complicated phase portrait for a plane pendulum. Here again, we will not resort to equations and formulae, and, instead, will try to outguess its form. Later on, in the proper place, we shall give you the formulae and specify this portrait quantitatively; right now we will restrict ourselves to its qualitative shape only. It is, perhaps, most difficult to understand what a phase space of the pendulum is. If the angle of the pendulum deviation from the horizontal line is φ, then a state will be a collection of this angle φ and the angular velocity $\omega = \dot{\varphi}$ of the pendulum rotation. A phase space will
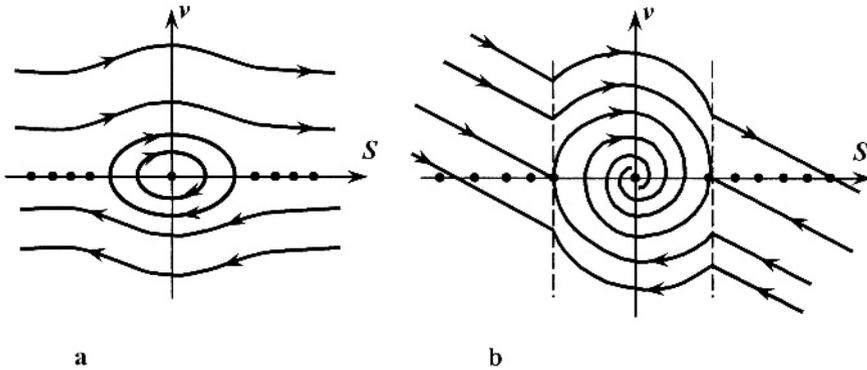
Fig. 1.11. The phase portraits of skier's movements corresponding to the profile in figure 1.9b , **a** (without friction) and **b** (with friction taken into account).

be the space whose points are all possible points $M(\varphi, \omega)$. The angle φ varies from $-\pi$ to π, the values $\dot\varphi = -\pi$ and $\varphi = \pi$ describing the same position of the pendulum. Therefore, the infinite band $-\pi \le \varphi \le \pi$, $-\infty < \omega < \infty$ will serve as a phase space. Besides, the points $\varphi = -\pi$, ω and $\varphi = \pi$, ω will indicate the same pendulum state. Thus, in the phase space they should be represented by a single phase point only, and not by two. Here, it is possible to adopt the simple convention – the points $(-\pi, \omega)$ and (π, ω) constitute the same point. Also, their real merging is possible, by sticking down the band along
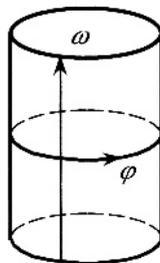
Fig. 1.12. A phase cylinder

the lines $\varphi = -\pi$ and $\varphi = \pi$ into a cylinder. It should be certainly done in such a way as to have the points of equal ω stuck together. This will yield a two-dimensional cylinder as a pendulum phase space. The cylinder is shown in figure

1.12. Upon cutting it along the line $\varphi = \pm\pi$ we come to the band (Fig. 1.13) with identified sides.
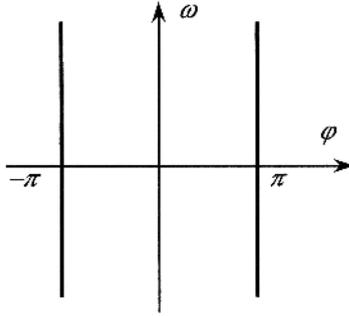
Fig. 1.13. The development of the cylinder on a plane.

To draw phase trajectories just upon the cylinder is not comfortable. Therefore, we will do it on the band remembering always that this is a cut out cylinder expanded on the plane.

The pendulum has two equilibrium states, a bottom stable equilibrium corresponding to the point $\varphi = 0$, $\omega = 0$ and a top unstable one corresponding to the point $\varphi = \pm\pi$, $\omega = 0$. Each of these points is a whole phase trajectory, its beginning, its end and it itself are entirely held in a single point.

If a motionlessly hanging pendulum, represented by the equilibrium $\varphi = 0$, $\omega = 0$, is pushed slightly, then it will start swinging with a small amplitude about this equilibrium state. Such a motion will be represented through the small oval embracing the point of the bottom equilibrium. With the push increased, the pendulum will swing with a larger and larger span (an amplitude), and at last its oscillations will turn to a rotating motion, either clockwise or counterclockwise. The said is depicted in figure 14a. In this phase portrait: k_0 is a bottom stable equilibrium; k_∞ is a top unstable equilibrium; k_1, k_2, k_3 are periodic oscillating motions about the bottom equilibrium; B_1, B_2, B_3 are quicker and quicker rotating counterclockwise motions; $\overline{B}_1, \overline{B}_2, \overline{B}_3$ are quickening clockwise rotations; (the pendulum deviation angle φ is counted counterclockwise from the low position of the pendulum); P_1 and $\overline{P}_2$ are the motions coming close to the point k_∞ at $t \to \infty$ and $t \to -\infty$. These two last phase trajectories, P_1 and $\overline{P}_2$, separate the oscillating and rotating motions.

In this initial stage, you will be supposed to possess some knowledge about a phase portrait, if you feel now capable of drawing, without peeping into figure 1.14b, how the portrait in figure 1.14a will be changed, if dissipations appear, say, caused by air friction. Then, you will have to explain also what motions are

a   **b**
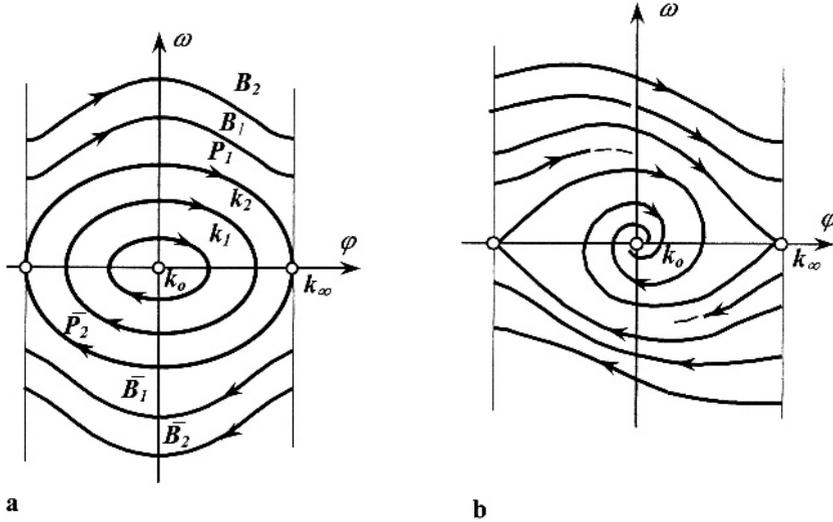
Fig. 1.14. Phase portraits of the pendulum : **a** with no friction ; **b** with friction.

represented by its separate phase trajectories.

All the mathematical models to further appear in our text and to be studied are nothing but the specifications and the particular cases of the same more general mathematical model of the dynamical system described above. Also, so much they will be unlike each other, so much they will differ both by the nature of processes running in them and by their physical nature! Any process, phenomenon or system of any nature being described by differential equations is a dynamical system. Therefore, the great laws of nature – the laws of mechanical motion of solid bodies, fluids, elastic media, the theory of electromagnetic field, the laws of electrodynamics and quantum physics – are described by mathematical models representing themselves the dynamical systems. Though, our further narration will be started not from them. Instead, first we shall consider and study very simple models, and only upon this you will be suggested more complicated models, which retain, nevertheless, their specificity and simplicity. In spite of their simplicity, sooner, thanks to their simplicity, they are most admissible for training and contribute to the most extent to your gaining the general and intuitive comprehension. The specific examples to be suggested to you are supposed to ensure your many-

sided understanding. Sometimes, very simple mathematical models will be suggested for simple physical systems, and, sometimes, simple models for very complicated systems. Though, even in very complicated cases these models will remain simple demonstrating their significance and cognitive power.

As it was noted above, each system or process described by differential equations always represents itself a dynamical system. This follows from the theorem of existence and uniqueness of the solution for the differential equations with the given initial conditions. Its state is a collection of the initial conditions which uniquely determine the solution. Its operator is determined by its solution within the time interval Δt.

Alongside with it, I would not wish you to think that the above general definition of a dynamical system is nothing but simply another treatment of differential equations. In order to eliminate such a thinking, I will give you an example where this definition is easily seen to be far from the differential equations. These equations here have nothing to do with it. This example is the game "Life" by J. Conway.

In this game played on a chess board, a state is defined through the position of counters. A space of states will be a set of all possible positions of counters. A state may be set up by the 8 x 8 matrix whose elements are units and zeroes, depending upon whether there is or there is no counter on the associated square. A phase space consists of all possible matrices of the above type. At each time $\Delta t = 1$, the positions of counters are changed by the operator being defined by the below three rules – survival, death and birth. These rules are as follows:

1) a counter retains, if nearby there are two or three other counters;

2) a counter is removed, if nearby there are more than three or less than two counters;

3) a free square is occupied with a new counter, if there are three counters nearby.

There are many possible changes in the positions of counters depending upon their initial position. For example, the below three counters of the configuration
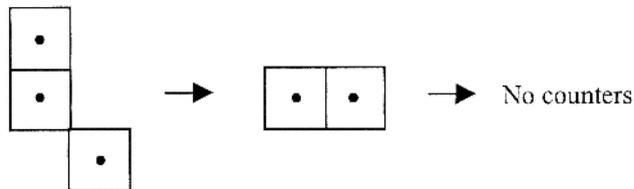
turn, as shown, first, into two counters and then vanish. The square-shaped configuration of four counters
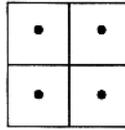
will lead to no changes.

Three counters aligned
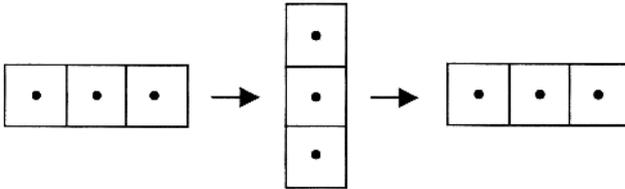
will oscillate, as shown, with a two-time period.

Five counters, forming a "glider"-like configuration, are repeated each four times, shifting one square to the right and down.
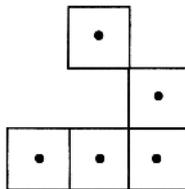
There exist such positions of counters which through oscillating each period will generate a "glider".

Also, there are the positions of counters which will "devour" gliders, and so on.

This variety of possibilities, imitating in a distant and simplified way the real life generated via combining chemical molecules, will be represented, as said already before, by the dynamical system whose state x is a matrix of zeroes and units. The matrix will hold so many rows and columns, so many they are on the playing board used for the game. The entry a_{ij} of this matrix is equal to 1 or 0, depending upon whether there is or there is no counter on the intersection of the i-th row and j-th column. A set of all possible such matrices with 1- and 0- en-

tries will constitute its state space. The operator $T(\Delta t)$ of this dynamical system will be determined only for discrete values of Δt equal to 0, 1, 2,

If n is an integer, then

$$T(n) = T^n(1) \ ;$$

$T(0)$ will be the operator which does not change the counter positions.

Before finishing our narration on a mathematical model of the dynamical system and its phase portrait, let us find a phase portrait for the physical pendulum mounted on a rotating base (figure 1.15).
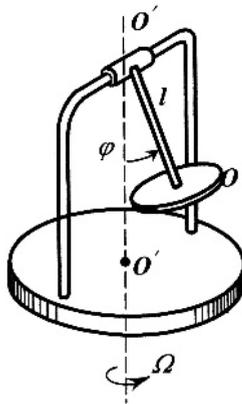
Fig. 1.15. The phase pendulum on a rotating base.

For the case of the pendulum base being fixed, we have already found a phase portrait of the pendulum without writing out the equations of its motions, since the pendulum is well familiar for us and the differential equations of its motions are not thus needed. As for the case with the pendulum on a rotating base, to treat it in the same way is impossible, for the dynamics of this pendulum is a mystery for us and one may unvail this mystery only through studying its mathematical model. This is the very thing we are doing right now. For this, some information from mechanics will be needed for us. Take it on trust.

The state of the pendulum with a base rotating around a vertical axis at the angular velocity Ω, as well as that of the fixed-base pendulum, will be expressed by the angle of its deviation from the vertical line φ and by its angular velocity $\omega = \dot{\varphi}$. We need to derive the differential equations of its motions. Here, theoretical mechanics can be of help for us. Its prescription looks like this: find the kinetic energy T and the potential energy V of the pendulum and compile the

Lagrange function L expressing it through the state variables φ and ω; upon it, the motion equations will be written in such a beautiful form as

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \omega} \right) - \frac{\partial L}{\partial \varphi} = 0 \ .$$

Upon the computations, being still not clear for you at present, we find the function L of the form

$$L = \frac{A}{2}(\omega^2 + \Omega^2 \sin^2 \varphi) + \frac{C}{2}\Omega^2 \cos^2 \varphi + Mgl \cos \varphi \ ,$$

where A and C are inertia moments of the pendulum, and L and M are its length and mass.

The elementary differentiations will yield the wanted second-order differential equation

$$A\ddot{\varphi} + \frac{C-A}{2}\Omega^2 \sin 2\varphi + Mgl \sin \varphi = 0 \ . \tag{1.4}$$

Take it on trust how this equation is derived. One needs only to understand this equation, i.e. to see that for $\Omega = 0$ it is transferred to the equations for the usual physical pendulum on a fixed base. Also, one should have an idea concerning the physical sense of the inertia moments A and C. The pendulum looks like a solid rotating with respect to its suspension line (the line connecting a suspension point with a mass centre). Its centre of gravity is at the distance l from the suspension axis; the pendulum itself is of the mass M. C is an inertia moment with respect to its suspension line, and A is a moment of its inertia with respect to its axis being perpendicular to the suspension line and coming through the suspension point. For the solid being stretched along the suspension axis, we have $C < A$. A reverse situation takes place for the disk-like pendulum whose radius is sufficiently longer than the pendulum length.

To solve the equation (1.4) with use of elementary functions is impossible. Though, it is possible to integrate it once. For this, multiply it by $\dot{\varphi}$ and perform the below calculations:

$$A\ddot{\varphi}\dot{\varphi} + \frac{C-A}{2}\Omega^2 \sin 2\varphi\dot{\varphi} + Mgl \sin \varphi\dot{\varphi} =$$

$$= \frac{d}{dt}\left(\frac{A}{2}\dot{\varphi}^2 - \frac{C-A}{4}\Omega^2 \cos 2\varphi - Mgl \cos \varphi \right) = 0 \ ,$$

from which we find that

$$\frac{A}{2}\dot{\varphi}^2 - \frac{C-A}{4}\Omega^2 \cos 2\varphi - Mgl \cos \varphi = h \quad ,$$

where h is an integration constant.

From the integral derived we find that

$$\sqrt{\frac{A}{2}}\dot{\varphi} = \pm\sqrt{h + \frac{C-A}{4}\Omega^2 \cos 2\varphi + Mgl \cos \varphi} \qquad (1.5)$$

This is the very equation of phase trajectories in the cylindrical phase space of the state variable φ and $\dot{\varphi} = \omega$. These phase trajectories have to be constructed, for each h there should be found its own phase trajectory. Though, for you it will be not so easy. Therefore, with the help of the equation (1.5) let us, first, create the already known phase portrait of the pendulum on a fixed base ($\Omega = 0$). In this more simple, case we obtain

$$\sqrt{\frac{A}{2}}\dot{\varphi} = \pm\sqrt{h + Mgl \cos \varphi} \qquad (1.6)$$

Assuming the subradical expression to be the function φ., let us construct a plot for the constant h and the plot for the function $Mgl \cos \varphi$ (figure 1.16).

The first plot will be the straight line being parallel to the axis φ, at the distance h from it, and the second will be the single period of the sinusoid of the amplitude Mgl. In figure 1.16, arrows indicate the values of the subradical function for various φ. With the length of this arrow indicated as ρ (with its direction considered), the phase trajectory equation (1.16) will be written in a very simple form

$$\dot{\varphi} = \pm\sqrt{\frac{2\rho}{A}} \qquad (1.7)$$

For the value of h given in figure 1.16 , the possible values of φ will lie between φ_1 and $-\varphi_1$ (outside of them we have $\rho < 0$; and $\dot{\varphi}$ is imaginary). With φ changing from $-\varphi_1$ to φ_1 , we easily find, from figure 1.16, the form of the corresponding phase trajectory. It will be an oval drawn below the plots

used by us in the plane φ, $\dot{\varphi}$ (to be more exact, within the band $-\pi \leq \varphi \leq \pi$, $-\infty < \varphi < \infty$). Through changing h from $-Mgl$ to ∞, let us find all phase trajectories. This phase portrait is already familiar for us and once more drawn in figure 1.16.
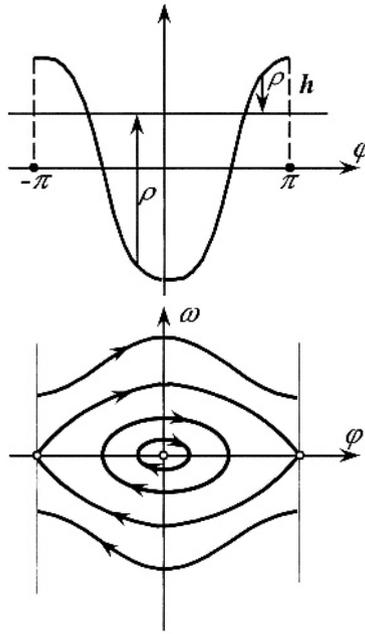
Fig. 1.16. Constructing the phase portrait of the pendulum on a rotating base at $\Omega = 0$.

To pendulum oscillations there will correspond $-Mgl < h < Mgl$, to the low stable equilibrium there will correspond $-h = -Mgl$; to the upper unstable equilibrium there will correspond $-h = Mgl$; and to the rotations, $-h > Mgl$. Now, let us similarly construct a more complicated phase portrait appropriate to the equation (1.5) for $\Omega \neq 0$.

For

$$(C - A)\Omega^2 > 4Mgl$$

figure 1.16 will be changed and take the form of figure 1.17.
Here, the plot

$$-Mgl\cos\varphi - \frac{C-A}{4}\Omega^2 \cos 2\varphi$$

is drawn and the line corresponding to the value h is depicted. As earlier, the values of $\rho(\varphi)$ are shown with arrows and the phase trajectory is described by the equation (1.7). By changing h, we, with some patience, shall find the phase
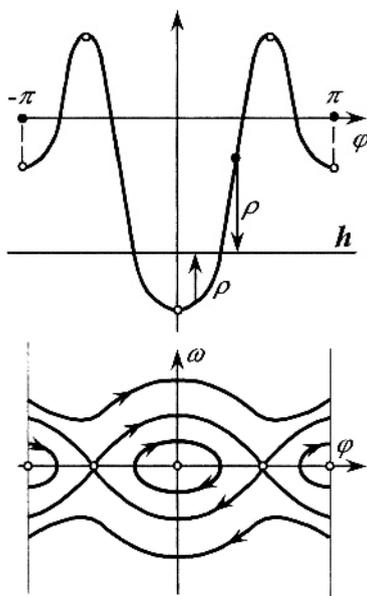
Fig. 1.17. Constructing the phase portrait of the pendulum on a rotating base at $\Omega \neq 0$.

portrait suggested in figure 1.17. This phase portrait exhibits exceptionally wonderful properties of the pendulum, viz. its ability to stably hang both up and down. This wonderful ability has appeared due to the above assumption that

$$(C - A)\Omega^2 > 4Mgl   ,$$

i.e. when $C > A$ and the rotating velocity Ω of the pendulum base is sufficiently large. It may be similarly revealed that for $C < A$ an increase of the rotating velocity Ω will retain the instability of the upper position, and will make the lower equilibrium position unstable as well.

2 Fluid outflow from a vessel

The Torrichelli law and a simplest model of the fluid outflow. Compression of the outflowing water jet. Insufficiency of a simplest model and improvements with the account of the fluid outflow speedup. The phase portrait for a fast speedup and a slow outflow.

Let us consider a very simple phenomenon of a fluid outflow from a cylindrical vessel with a small hole in its bottom (Fig. 2.1).
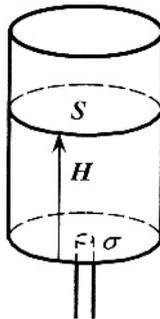
Fig. 2.1. The cylindrical vessel with a hole in its bottom for a fluid outflow.

Let S be a sectional area of the vessel; σ a square of the hole; H is a height of the fluid level. How will the fluid level height H be changed, if the fluid is flowing out and the initial value of the level is equal to $H = H_0$? In order to answer this question, it is sufficient to know the velocity v of the outflow through the hole. Indeed, the fluid effluence during the time interval dt will be equal to $\sigma v dt$, and, hence, the velocity of the fluid level sinking in the vessel will be equal to

$$\dot{H} = -\frac{\sigma}{S}v .$$

With v known, the above expression will be a differential equation from which H can be found as a function of t, i.e. $H(t)$. The first who outguessed three centuries ago the value of v was Torricelli. He said: "Water will flow out with the same velocity as if it would have dropped from the height H". You see that from above water is vanishing, and from below it is flowing out, as if it had been dropped from the height H. A solid body dropped from the height H will gain the velocity

$$v = \sqrt{2gH} .$$

This is the Torricelli formula famous in his time. It is impossible not to admit a wit of Torricelli's considerations. However, you see that from above one water is being dropped and another water is flowing out through the hole. Here, not everything is clear; though, the formula is true as it was proved experimentally. If we have trust to this formula, let us then come to the differential equation

$$\dot{H} = -\frac{\sigma}{S}\sqrt{2gH} , \qquad (2.1)$$

from which it follows that

$$\frac{dH}{\sqrt{H}} = -\frac{\sigma}{S}\sqrt{2g}\, dt ,$$

and, hence,

$$-2\sqrt{H} = -\frac{\sigma}{S}\sqrt{2g}t + C ,$$

where C is an unknown constant.
   From the initial conditions

$$H_{t=0} = H_0$$

we find that

$$C = -2\sqrt{H_0}$$

and, finally,

$$H = (\sqrt{H_0} - \frac{\sigma}{S}\sqrt{\frac{g}{2}}t)^2 . \tag{2.2}$$

Now, a decrease of H may be shown diagrammatically (Fig. 2.2) and we find the time of a complete outflow

$$t_{fl} = \frac{S}{\sigma}\sqrt{\frac{2H_0}{g}} . \tag{2.3}$$
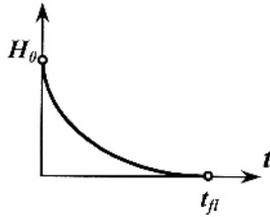
Fig. 2.2. The time graph for the changes of the   water level during the outflow .

From the plot it is seen that, at first , water is flowing out faster, then its velocity is slowing down to zero and this occurs during the time t_{fl} being proportional to a square root of the initial water height H .

   How much do these conclusions coincide with the experiment? The Torricelli formula is confirmed in practice. Indeed, in all cases water flows out with the same velocity as if it was dropped freely from the top water level. Though, the outflowing time t_{fl} , calculated by the formula (2.3), turns out to be approximately two times less. Accordingly, two times less than the calculated velocity will be the velocity of the sinking level H in the vessel. Why does such a large discrepancy appear, whereas the velocity v of water outflow through the hole was determined correctly? This discrepancy may appear only due to the incorrectness of the formula

$$\dot{H} = -\frac{\sigma}{S}v,$$

which means that how much water has flowed out and so much of it has become lower in the vessel. Why may this formula be incorrect? Let us formulate its exact sense – the ratio of velocities $v/\dot{H}$ is equal to the ratio of the cylinder section to the section of the outflowing jet. We assumed, without thinking over, the jet section to coincide with the size of the hole; but , to be more exact, this is abso- lutely not so. In reality, the water jet section will be   smaller than that of the hole, since this   the water approaching the hole from different directions will compress the jet, and, therefore, the jet section will be less than that of the hole. This compression of the water jet may be eliminated, if the hole is supplied with a gradually converging   funnel.   A complicated hydrodynamic calculation con- firmed experimentally shows   that the compression coefficient of the water jet flowing out through a hole in a horizontal bottom is approximately equal to 2 .

The fluid outflow mathematical model   constructed by us is a dynamical sys- tem. Its phase space will be the half-line $H \geq 0$. Its single phase trajectory repre- sents this half-line being travelled from $H = \infty$ to $H = 0$ (Fig. 2.3).
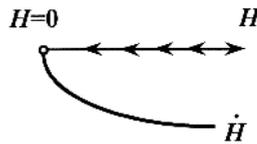
Fig. 2.3. Constructing an one-dimensional phase portrait for the fluid outflow.

A vessel with a water outflow was used in ancients times for time countdown. This device is a so-called "water clock", clepsydrae. If you choose S , σ and H_0 so that t_{fl} is equal to 24 hours, then, upon filling up the vessel with water, one can determine time during the subsequent days through the level H . In order to mark the levels corresponding to different hours of the day, the section of the axis t (Fig. 2.2) must be divided into 24 equal parts and, according to the plot, the appropriate H 's may be then found. It is not difficult to see that the resulting time scale marked   through water levels will be non-uniform, i.e. at first, one-hour gap of time r will correspond to a large sinking of water, and fur- ther, will correspond to a smaller and smaller sinking. Let us think how the water clock must be modified in order to make   its scale more handy, i.e. uniform. Look at the formula (2.1) again : the scale will be uniform, if irrespective of H , equal dH 's will be assigned   equal dt 's, i.e. it is necessary to do so that

$$\frac{\sigma}{S}\sqrt{2gH} = const.$$

At first sight this seems impossible, but, upon some thinking , one can realize
that it is sufficient to have the cylindrical vessel (with S constant and inde-
pendent of H) replaced by the vessel with

$$S = a\sqrt{H}.$$

 What is the form of such a vessel? Let it have the form of a revolving body;
then its horizontal section at the height H will be a circle of the radius r and of
the square $S = \pi r^2$. Consequently, we get

$$\pi r^2 = a\sqrt{H}$$

or

$$H = (\frac{\pi}{a})^2 r^4,$$

which corresponds to the form of the vessel shown in figure 2.4, the form similar
to a mess-tin.
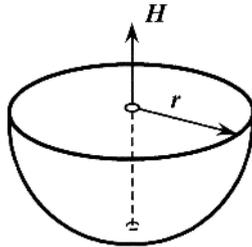
Fig. 2.4. The shape of the vessel from which water is flowing out with the level
decreasing at a constant velocity.

    Thus, the problem has been solved, the outflow law   has been found and can
be utilized for constructing a water clock with a convenient uniform scale. Al-
though, at present nobody needs it.
    The problem of water outflow is solved on the basis of the Torricelli law,
whose conclusion is somewhat vague. Let us try to make it clear. For this, let us

resort to something that was not known by Torricelli, to the energy conservation law. When water is floating out and its level is sinking, the potential energy of its thin layer of the mass dm is converted into the kinetic energy of the effluent water in such a way that

$$dmgH = dm\frac{v^2}{2}$$

or

$$v = \sqrt{2gH} \ .$$

This is just exactly the Torricelli formula. So, everything   is proved.

However, let us test in practice the above result (2.2), i.e. let us see what it will yield for $\sigma = S$, when water is not flowing out but is merely falling down from a bottomless cylinder. As it is falling freely, then it is obvious that

$$H = H_0 - \frac{gt^2}{2} \ .$$

This   is not consistent with the Torricelli law accepted by us. For comparison, both plots   for $H(t)$ are given in figure 2.5 . The first of them corresponds to the Torricelli law for $\sigma = S$, the second to a free fall of water. They are not coincident, though in both cases water is flowing out from the vessel during the similar time t_{fl} .
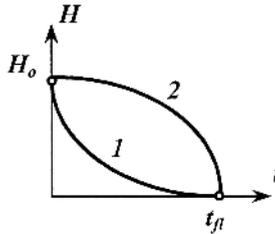
Fig. 2.5. The time graphs 1 and 2 show the times of water outflow from a bottomless cylinder   (equation (2.1)) and the free-fall law, respectively.

What is then the matter here? Where is an error? Is the Torricelli law not true? To doubt the law for a falling solid is difficult. Let us see where lies the differ-

ence between the plots 1 and 2 in figure 2.5. According to the plot 1, water, is , at first, flowing out rapidly; then its outflowing velocity is slowing down. According to the plot 2, it behaves conversely, i.e. at first it is running slowly and then faster and faster. With a hole as large as the vessel itself, the second case seems more truthful. Where is then an error? Perhaps, the energy conservation law is invalid? It is hardly believed. May be the law has been applied incorrectly? Yes, it seems applied incorrectly. But in what place? You see that here everything is so simple and clear. Let us return to the main assumption: energy is conserved (friction of water can be neglected); therefore, the potential energy of the top layer is converted to the kinetic energy of the effluent fluid. But this is really true, though, only approximately, for $\sigma << S$ only. Indeed, the top layer of the fluid is also sinking and gets some kinetic energy. Thus, it would be better to write as below

$$dmgH + dm\frac{\dot{H}^2}{2} = dm\frac{\dot{v}^2}{2} \quad .$$

Accordingly, in view of $\sigma v = -S\dot{H}$, we have

$$\dot{H} = -\frac{\sqrt{2gH}}{\sqrt{(\frac{S}{\sigma})^2 - 1}} = -\frac{\sigma}{S}\sqrt{\frac{2gH}{1-(\frac{\sigma}{S})^2}} \quad . \qquad (2.4)$$

For $\sigma << S$, this new formula turns into the Torricelli one. Hence,   the needed   improvement has been found. Though, we should not be in a hurry. Again, let us assume   that $\sigma = S$; then, the result will be worse, $\dot{H} = \infty$. There is an error here again. Let us   return to our initial consideration and assume that we have   at once $\sigma = S$. When water is falling down its potential energy, being actually equal to $dmgH$, converts into the kinetic energy but not only of the effluent fluid (escaping from the vessel), but also of the sinking fluid left inside the vessel, more precisely, into the increment of its kinetic energy, because the motion is accelerated. Thus, the error lies in the fact that the increment of the kinetic energy of the sinking fluid in the vessel should be additionally accounted as well. For $\sigma << S$, this increment will be actually small; though, for σ being congruent to S , this increment will be absolutely not so. Let us again return to the initial consideration and insert necessary improvements.

Thus, let V be the potential energy of the fluid in the vessel, T be its kinetic energy, T_1 the kinetic energy of the fluid escaped from the vessel during the time dt . Then, under the energy conservation law, we have

$$\frac{d}{dt}(T+V)+T_1 = 0.$$

The expression for T_1 is already known for us. It is as follows

$$T_1 = dm\frac{v^2}{2} = -\rho S\dot{H}dt\frac{v^2}{2},$$

where ρ is a cubic density of the fluid. The potential energy may be calculated as follows

$$V = \int gHdm = \int_0^H \rho SHgdH = \frac{\rho gSH^2}{2}.$$

Then, the kinetic energy is equal to

$$T = \int \frac{u^2}{2} dm,$$

where integrating is done over all volume of the fluid in the vessel and u is a current velocity of the mass element dm of this volume. The velocities u in the different points of the vessel are different and calculating them means to find how the fluid is flowing out from the vessel. This is very difficult to do. What to do then? How it should be done is clear for the two extreme cases, i.e. for $\sigma << S$, when an magnitude of the order u^2 may be neglected, and for $\sigma = S$, when $u = \dot{H}$. The last approximation is also applied to the case when $H >> r$, where r is a radius of the cylindrical vessel filled with fluid. Let us consider this case; then

$$T = \rho SH\frac{\dot{H}^2}{2}.$$

After that, the energy conservation law, according to the above, will be of the form

$$\frac{d}{dt}(\rho SH\frac{\dot{H}^2}{2}+\frac{\rho gSH^2}{2})-\rho S\dot{H}(\frac{S}{\sigma})^2\frac{\dot{H}^2}{2} = 0$$

or

$$\rho S \frac{\dot{H}^3}{2} + \rho S H \dot{H} \ddot{H} - \rho S (\frac{S}{\sigma})^2 \frac{\dot{H}^3}{2} + \rho g S H \dot{H} = 0,$$

from which

$$\ddot{H} = -\frac{1}{H} \{gH - [(\frac{S}{\sigma})^2 - 1] \frac{\dot{H}^2}{2}\} \qquad (2.5)$$

As was expected, the differential equation derived now is of the second order and for $\sigma = S$ it is checked successfully, since from (2.5) it follows from this examination that

$$\ddot{H} = -g .$$

This is just what was required.

Now, it remains to be understood how for $\sigma << S$ the earlier derived equation, based on the Torricelli law or its improvement, follows the equation (2.5). To reveal this is not easy. To make our considerations easier, let us introduce $u = -\dot{H}$ and write the equation under study in the form

$$\dot{u} = \frac{1}{H} \{gH - [(\frac{S}{\sigma})^2 - 1] \frac{u^2}{2}\} \equiv f(u); \qquad (2.6)$$

and see how u will vary. When $f(u) < 0$, the velocity of u will decrease and, for $f(u) > 0$, conversely, it will increase, i.e. there will exist some value

$$u^* = \sqrt{2gH[(\frac{S}{\sigma})^2 - 1]^{-1}} ,$$

such that for $u < u^*$ u will increase and for $u > u^*$ it will decrease. As a result, u will turn out to be close to u^*. For visibility purposes, let us depict the semiaxis $u \geq 0$ and draw the plot of $f(u)$ as a function in u. In accordance with the plot, u varies as shown in figure 2.6. For $S / \sigma >> 1$, u will approach u^* fast. Note that when u is approaching u^*, the point u^* displaces slightly, but its displacement occurs much slower than u approaches u^*, since

$$\overset{*}{u} = \frac{d}{dt}\sqrt{\frac{2gH}{(\frac{S}{\sigma})^2 - 1}} = -\frac{1}{2}\frac{\sigma u}{S}\sqrt{\frac{2g}{H[1 - (\frac{\sigma}{S})^2]}} . \qquad (2.7)$$

As a result, we have approximately

$$u = \frac{\sigma}{S}\sqrt{2gH} . \qquad (2.8)$$

This will be the more precisely, the smaller σ / S is , that completely coincides with the above formulae (2.1) and (2.4).
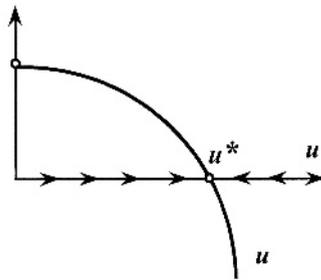
Fig. 2.6. The phase portrait for the differential equation (2.6) at $H = const$.

What are the conclusions from the fact that the equation

$$\dot{H} = -\frac{\sigma}{S}\sqrt{2gH}$$

was replaced by the equation

$$\ddot{H} = -\frac{1}{H}\{gH - [(\frac{S}{\sigma})^2 - 1]\frac{\dot{H}^2}{2}\} .$$

First of all, at the initial moment $t = 0$ we have now $\dot{H} = 0$. Then, $\dot{H}$ is fast approaching the value determined by the first equation and further $\dot{H}$ is varying in accordance with this equation. The said is depicted in figure 2.7, where the firm line depicts $\dot{H}$ fluctuating in accordance with the first equation and the dotted line, with the second equation. The part of the plot corresponding to the interval from $t = 0$ to t_{ac} describes the acceleration of the outflowing jet, and

the remaining segment, from t_{ac} to t_{fl}, to the lengthy outflow with a progressive slowdown.
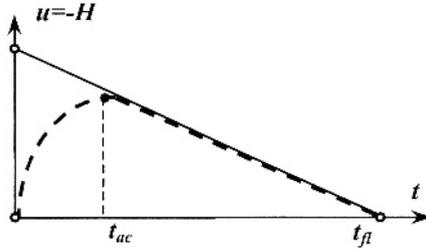
Fig. 2.7. The time graph of the water level fall upon opening the hole: without firm line) and with the of speedup phase taken into account.

In order to correctly depict the velocity fluctuations , one should determine the acceleration time t_{ac}, whereas the outflowing time is already known to be as

$$t_{fl} = \frac{S}{\sigma} \sqrt{\frac{2H_0}{g}} .$$

After the acceleration, the outflowing velocity reaches the value $\sqrt{2gH}$; besides , as follows from (2.6), the acceleration velocity $\dot{v}$, at least in the beginning, will be equal to

$$\dot{v} = \frac{S}{\sigma} g .$$

Therefore, the acceleration time, by the order of values, will be approximately equal to

$$t_{ac} = \frac{\sigma}{S} \sqrt{\frac{2H_0}{g}} .$$

Hence, we approximately have

$$t_{ac} / t_{fl} = (\frac{\sigma}{S})^2 \ .$$

In particular, if the ratio of the cylindrical vessel radius to the radius of the bottom hole is equal to 100, then the acceleration time will be approximately 10^8 times less than the outflowing time, i.e. if the outflowing time is about an hour, then the acceleration time will be 0.0001 second. Even with the hole being relatively large, when the ratio of radii is about 10, the ratio of times will be of the order 10^4. Thus, for $\sigma / S \ll 1$ the acceleration will occur actually instantly and the further outflow will meet the Torricelli law. Here, the brief story about the water outflow from the vessel with a hole in its bottom could be finished. But I would like also to show you how the notion of a phase portrait can be utilized by you for studying the motions of the improved model (2.5).

Let us write its differential equations in the variables of the fluid level H and of the outflowing velocity v. A collection of these variables will constitute a state, and, according to the above, its changes will satisfy the below differential equations

$$\dot{H} = -\frac{\sigma}{S}v, \qquad \dot{v} = \frac{S}{\sigma}\{g - \frac{1}{H}[1-(\frac{\sigma}{S})^2]\frac{v^2}{2}\} \ . \qquad (2.9)$$

As $\sigma / S \to 0$, these differential equations will assume the form

$$\dot{H} = 0, \qquad \dot{v} = \begin{cases} \infty & for \ \ v^2 < 2gH \\ 0 & for \ \ v^2 = 2gH \ , \\ -\infty & for \ \ v^2 > 2gH \end{cases} \qquad (2.10)$$

according to which the phase portrait consists of the vertical phase trajectories along which the phase points are moving, at the infinitely large velocity, to the curve of the equilibrium states (Fig. 2.8). For $\sigma / S \ll 1$, i.e. when it is very small, the instantaneous motions will turn to the fast ones slowing down as they are approaching the curve $v^2 = 2gH$; simultaneously, each of these phase points, including the former equilibrium states, will be moving to the left at a very small velocity. As a result of this, the phase portrait will be changed and will take the form shown in figure 2.9.
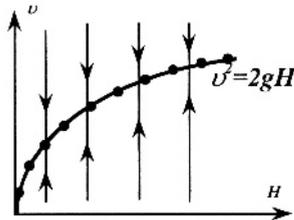
Fig. 2.8. The phase portrait for the limiting differential equation (2.10). The curve of black points is a manifold of equilibrium states.
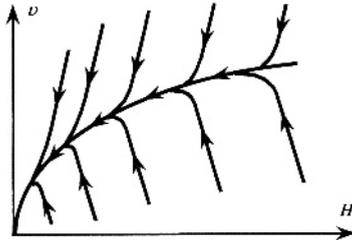
Fig. 2.9. The phase portrait for the differential equation (2.9) at $\sigma/S \ll 1$.

On this portrait (Fig. 2.9), any motion of a phase point, upon a fast change, turns to a slow motion nearby and along the phase curve $v^2 = 2gH$, where the Torricelli law is valid. In particular, if the initial point is $H = H_0$, $v = 0$, it, at first, will arrive at the curve $v^2 = 2gH$ very fast and then will move along this curve to the left remaining nearby. Diagrammatically, the fluctuations of the velocity v (more precisely, $H = \dfrac{\sigma}{S}v$) corresponding to this motion of the phase point have been already presented in figure 2.7. In the same place, we have very roughly estimated the time of the phase point arrival from the initial state $H = H_0$, $v = 0$ at the small neighbourhood of the curve $v^2 = 2gH$. Now, we may improve this estimate. Indeed, let we want to know the time during which the above initial phase point arrives at the ε-neighbourhood of the curve $v^2 = 2gH$ represented by the inequalities

$$-2gH\varepsilon < v^2 - 2gH < 2gH\varepsilon,$$

where $\varepsilon > 0$ is small. From the second differential equation (2.9) it follows that

$$\mu dv = [g - \frac{v^2}{2(1-\mu^2)H}]dt,$$

where $\mu = \sigma / S.$ is introduced. Through neglecting both the small changes of H, being equal to H_0 at the initial moment, and μ^2, let us write this relation in the form

$$\frac{\mu dv}{g(1-\frac{v^2}{2gH_0})} = dt.$$

By integrating its left-hand side from $v = 0$ to the value $v = \sqrt{2gH(1-\varepsilon)}$, corresponding to the arrival at the boundary of the ε-neighbourhood of the curve $v^2 = 2gH$, and its right-hand side from 0 to τ, we shall find that the desired time τ of the phase point travel will be approximately equal to

$$\tau = \frac{\mu}{g}\int_0^{\sqrt{2gH(1-\varepsilon)}} \frac{dv}{1-\frac{v^2}{2gH}} = \frac{1}{2}\frac{\sigma}{S}\sqrt{\frac{2H}{g}}\ln\frac{1+\sqrt{1-\varepsilon}}{1-\sqrt{1-\varepsilon}}.$$

This magnitude is of the same order as the earlier roughly estimated value τ being equal to $\sigma / S\sqrt{\frac{2H}{g}}$.

   Thus, when the hole is small ($\sigma / S << 1$), the fluid outflowing process is clearly divided into the two phases: a fast acceleration phase, i.e. a speed-up occurs until the velocity is close to $\sqrt{2gH_0}$, and a lengthy outflow phase with the velocity decreasing for a finite time up to zero. Moreover, from the estimates derived it follows that the acceleration time is unrestrictedly decreasing as the section σ is decreasing. This gives rise to doubts, since it corresponds to an unbounded growth of acceleration. Apparently, this paradox arises as a consequence of the approximations assumed for calculating the energy of the fluid in the vessel, when the fluid velocity in all the points is assumed to be equal to $\dot{H}$. It is clear

that this is not the case near the hole; the greater S/σ is, the more it is not so. At the same time, a capillar surface tension of the fluid makes it impossible for the fluid to flow out, if the radius of the hole is less than $2\mu/\rho g H_0$, where μ is a surface tension coefficient. If the fluid is water and $H_0 = 50$ cm, then this hole will be much less than 1 mm. Owing to this effect, the outflow can halt, when the level H is decreasing. In general, a small-size hole may not only halt but also stop the outflow. However, let us leave these phenomena aside and try to improve the calculation of the vessel fluid kinetic energy on the basis of the hydrodynamical model describing an outflow of ideal fluid. This calculating improvement becomes most essential near the hole, where the velocity is much more than $\dot{H}$.

Let $\dot{H}$ be sufficiently small so that for the fluid inside the vessel and not far from the hole the field of velocities will vary slowly and it can be considered quasi-stationary. In this case, to have a possibility of comparing the expression derived for the kinetic energy T let us write it in the form (Fig. 2.10)

$$T = \int_{\sigma} d\sigma \int_{0}^{H} \frac{\rho u^2}{2} \frac{d\overline{\sigma}}{d\sigma} dx = \frac{1}{2}\rho v^2 \sigma K = \frac{1}{2}\rho K\sigma(\frac{S}{\sigma})^2 \dot{H}^2 = $$

$$= (\frac{1}{2}\rho S H \frac{\dot{H}^2}{2})\frac{KS}{\sigma H}$$

The factor inside the brackets is the previous expression for the kinetic energy. The new expression differs from the previous one by the factor $KS(\sigma H)^{-1}$, where

$$K = \frac{1}{\sigma}\int_{\sigma} d\sigma \int_{0}^{H} \frac{d\sigma}{d\overline{\sigma}} dx$$

is a mean of the integral over the fluid tube. The magnitude $\dfrac{d\sigma}{d\overline{\sigma}}$ is decreasing with x from 1 to $\dfrac{d\sigma}{dS}$, where dS is a value of $d\overline{\sigma}$ on the fluid surface. The length of the fluid tube is nearly equal to H. Therefore, K is the magnitude being less than H, and the more the magnitude $\dfrac{d\sigma}{d\overline{\sigma}}$ decreases along the fluid tube with x increasing, the more K varies. Thus, the factor of distinction be-

tween the initial and improved magnitudes for the kinetic energy is less than S / σ and its value grows as σ decreases. It is obvious that our error in the initial calculation can be great, but it is decreasing as H is growing and as $\dfrac{d\sigma}{d\bar{\sigma}}$ is decreasing rapidly when x is growing.
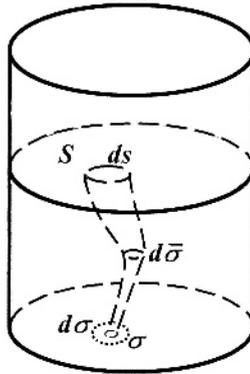
Fig.2.10. The fluid tube.

Now, let us find how the differential equation (2.5) will change. We have that

$$\frac{d}{dt}(\rho S^2 K \sigma^{-1} \frac{\dot{H}^2}{2} + \frac{\rho g S H^2}{2}) - \rho S \dot{H} (\frac{S}{\sigma})^2 \frac{\dot{H}^2}{2} = 0 \quad .$$

Therefore,

$$\ddot{H} = -\frac{\sigma}{SK}[g - (\frac{S}{\sigma})\frac{\dot{H}^2}{2}] \qquad (2.10)$$

and, as earlier, we come to the following estimation of the acceleration time

$$t_{sp} \approx K\sqrt{\frac{2H_0}{g}} \quad , \qquad (2.11)$$

according to which t_{sp} does not converge to zero as σ is decreasing. Note that

$\sqrt{\dfrac{2H_0}{g}}$ is the time of a free fall from the height H_0 and $K < 1$. The time of

the free fall from the height 50 cm is approximately equal to 0.3 sec, i.e. the acceleration retains fast.

3 Equilibrium and auto-oscillations of fluid level in the vessel with simultaneous inflow and outflow

Dynamics of fluid level when the outflow through a bottom hole or a siphon and when a constant inflow are present.

Let we now have a cylindrical vessel of the cross section S and with the bottom hole of the efficient section σ, and simultaneously with water outflow there occurs the water inflow of the intensity Q. For this case, the equation of water balance in the vessel will be of the form

$$S\dot{H} = -\sigma \upsilon + Q \ , \tag{3.1}$$

where υ is the velocity of the water outflow through the hole of the cross section σ; and H is , as before, the height of the fluid level in the vessel. Here the fluid is considered incompressible. If the fluid viscosity is also neglected, then for $\sigma / S << 1$ we have $\upsilon = \sqrt{2gH}$ and the differential equation (3.1) may be written as

$$\dot{H} = -\frac{\sigma}{S}\sqrt{2gH} + \frac{Q}{S} \tag{3.2}$$

This equation is easily integrated, but let us consider its phase portrait through depicting the plot of the velocity $\dot{H}$ as the function of H (figure 3.1). To the left of the point

$$H^* = \frac{Q^2}{2\sigma^2 g} \tag{3.3}$$

where $\dot{H} = 0$, the phase points move to the right along the phase half-line $H \geq 0$ and H increases; and vice versa, to the right from the point H^* the function H decreases. It means that the point $H = H^*$ is a stable equilibrium state. This equilibrium state will be approached by all the points on the phase half-line $H \geq 0$.

From formula 3.3 it follows that the height of the equilibrium level is directly proportional to the squared intensity Q of the fluid inflow and inversely proportional to the squared cross section of the effluent jet.

This is all concerning the entire unpretentious dynamics of this system – the system always approaches its stable dynamic equilibrium $H = H^*$ for the fluid inflow and outflow.
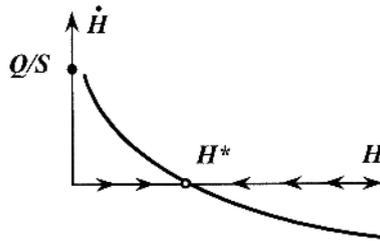
Fig. 3.1. The phase portrait of the system (3.2) for water outflow with a constant water inflow available.

Now let us assume the fluid outflow to be performed not through the vessel bottom hole but through a so-called siphon, i.e. the σ-section tube bent in the way shown in figure 3.2. The tube does not reach the bottom at the distance H_1, and above at the height H_2 it is bent, leaves the vessel and goes down till the distance H_3 from the bottom of the siphon (figure 3.2). The siphon is a wonderful device to empty an incompletely filled barrel over its brims, i.e. when $H < H_2$. Though, this may be only done when the siphon itself is filled with water. When it is empty, at the water level $H < H_2$ no water will flow through it. Therefore, this system – the barrel with a siphon; and the water inflow – may be described in the following way: the water level H and the variable ξ assuming the value equal to 1, if the siphon is filled with water; and the value equal to 0, if the siphon is empty. This description of the system is sufficient for prediction. Thus, H and ξ are chosen by us as a description pretending to be a state of the system. Here, we should apparently distinguish the following cases: $H < H_1$, $H_1 < H < H_2$ and $H > H_2$. For $H < H_1$ we get $\xi = 0$ and $S\dot{H} = Q$.

For $H > H_2$ the siphon is being filled (if it was empty) and hence we obtain $\xi = 1$ and $S\dot{H} = Q - \sigma\sqrt{2g(H + H_3)}$.

For the intermediate case $H_1 < H < H_2$ the values $\xi = 0$ and $\xi = 1$ are possible. For $\xi = 0$ we get $S\dot{H} = Q$. And for $\xi = 1$ we obtain

$$S\dot{H} = Q - \sigma\sqrt{2g(H + H_3)} \ .$$

For the case $H_1 < H < H_2$ we get $\xi = 1$, if immediately prior to it we had $\xi = 1$ or $H > H_2$; and $\xi = 0$ is obtained, if prior to it we had $\xi = 0$ or $H < H_1$.
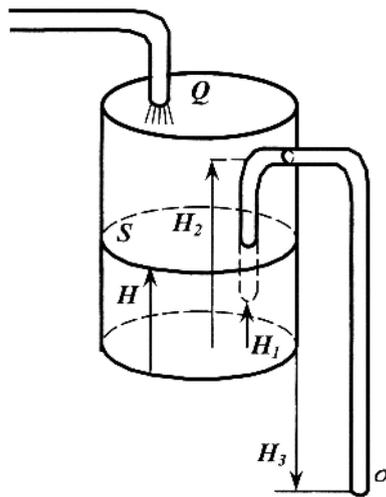
Fig. 3.2. The cylinder with a solid bottom, with fluid inflow and outflow through the siphon.

The above dependence of ξ upon the value and the change of H can be diagrammatically depicted in figure 3.3. In it, there are given the conditions to change by ξ its values from 0 to 1 and conversely, from 1 to 0. Namely, ξ changes from 0 to 1, if H, when increasing, turns into H_2; and vice versa, ξ is transformed from 1 to 0, if H, when decreasing, passes through the value H_1. The description just made is a mathematical model for the siphon. Here, ξ is not a function of H, and it is not a two-value function even. It is nothing but a functional from the previous values of $H(\tau)$ for $\tau \leq t$ diagrammatically shown in figure 3.3.
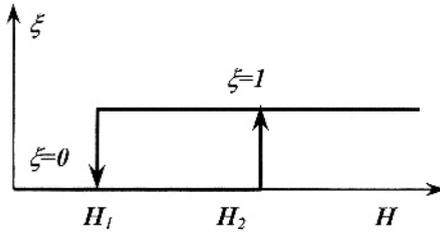
Fig. 3.3. The graph of dependence of ξ upon the fluid level H.

In accordance with the above, the equations for changes of H are written in the following way:

$$
\dot{H} = \begin{cases}
\dfrac{Q}{S} & for \quad H < H_1 \\[2mm]
\dfrac{Q}{S} & for\ H_1 \le H \le H_2 \quad and\ \xi = 0 \\[2mm]
\dfrac{Q}{S} - \dfrac{\sigma}{S}\sqrt{2g(H + H_3)} & for\ H_1 \le H \le H_2 \quad and\ \xi = 1 \\[2mm]
\dfrac{Q}{S} - \dfrac{\sigma}{S}\sqrt{2g(H + H_3)} & for \quad H > H_2
\end{cases}
$$

$$(3.4)$$

The phase space of the system under consideration consists of the two parts: the segment $0 \le H \le H_2$ and the half-line $H_1 \le H < \infty$, with transitions from one to another diagrammatically shown in figure 3.3.

To construct a phase portrait in this phase space, one needs to show by an arrow the motion direction of each of its points according to the differential equations (3.4). For this, let us combine figure Fig. 3.3 with the plots of the functions Q/S and $Q/S - (\sigma/S)\sqrt{2g(H + H_3)}$; the first of them pertains to the segment $(0, H_1)$, the second to (H_1, ∞). Here, their three different interpositions (figure 3.4) are possible. They are distinguished by the place where the equilibrium state O occurs.

The **a** case accounts for the appearance of water level periodic oscillations from H_1 to H_2 and vice versa. The **b** and **c** cases describe the stable equilibrium levels $H*$ between H_1 and H_2 and higher than H_2.

Let all the parameters of the system under study be fixed and H_3 may be variable. What case it will take place will depend upon the value of the root $H*$ in the equation

$$Q - \sigma\sqrt{2g(H + H_3)} = 0$$

equal to

$$H* = \frac{Q^2}{2g\sigma^2} - H_3 \quad .$$

According to figure 3.4, $H* > H_1$ will bring a stable equilibrium of the
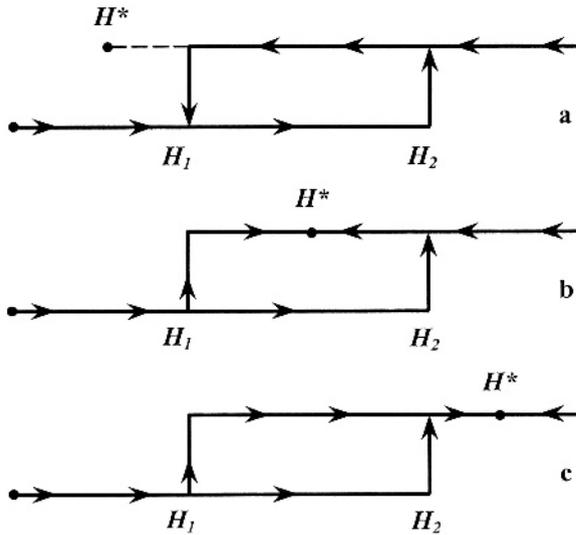
Fig. 3.4. Possible types of the phase portrait for the system in figure 3.2 described by the differential equation (3.4).

water level in the vessel, while $H* < H_1$ will bring stable periodic oscillations of the water level called auto-oscillations.

The dynamics of the water level will vary in a jerky way at the value of H_3 determined from the condition

$$\frac{Q^2}{2g\sigma^2} - H_3 = H_1,$$

or at

$$H_3 = H_3{}^* = \frac{Q^2}{2g\sigma^2} - H_1.$$

That is, $H_3 < H_3{}^*$ will bring a stable equilibrium, and it will retain with increase of H_3; and $H_3 > H_3{}^*$ will bring auto-oscillations being observed all the time.

4 Transitive processes, equilibrium states and auto-oscillations

Possible types of motions – transient processes, stable equilibrium states and auto-oscillations.

In the mathematical models just discussed above we have become acquainted with three types of motions of dynamic systems – a transitive process, an equilibrium state and an auto-oscillation.

A transitive process took place when water was flowing out of the vessel, i.e. the system makes a transition from its original state corresponding to the initial water level to another state corresponding to the vessel being empty. These transitive processes also took place in the cases when there have been obtained an equilibrium level or a periodically oscillating level in the vessel. The latter case occurred when water was flowing out through a siphon.

The equilibrium was of dynamic nature, as a result of a stable equalization of the water inflow and outflow. As for the stability, its appearance was stipulated by the water level increase causing the excess of the outflow over the inflow, and by the level decrease, conversely, resulting in the excess of the inflow over the outflow.

Here, most sudden and interesting is the case when stable oscillations are arising. The matter is that they are arising without any visible impulsive reason, absolutely themselves, and that is why they are called auto-oscillations (or self-oscillations). Not so long ago the appearance of such oscillations was thought to be attributed to any periodic action. With great difficulty the researchers then assumed the periodic oscillations to be able to appear themselves, being not stipulated by this periodic action.

In science and engineering, that the existence of auto-oscillations had been acknowledged brought about revolutionary changes in consciousness. These changes then sequentially happened in electrical engineering, mechanics, chemistry, biology and economy. The existence of auto-oscillations seemingly contradicted the nature of things, since any kinds of mechanical oscillations – say, the oscillations of a pendulum – die down; and also dissipative are the oscillations of an electric circuit. Where from can the oscillations arise in the chemical reaction? Any chemical reaction strives to its termination, to a certain dynamic equilibrium. Though, everything turned out to be more complicated. Auto-oscillations were revealed in various physical, chemical and biological systems. They could be harmful and useful; and soon , upon a lot of stubborn and severe debates they began to be observed everywhere. At present these things are very trivial already. Though, in 1930s – 1950s these things were treated not so. For a

long time, various auto-oscillating phenomena were habitually explained through the assumptions concerning the existence of some driving force and resonance. These assumptions were exploited for explaining such phenomena as the car front wheel shimmy, the aircraft wing flutter resulting in a wing damage or a loss of the aircraft control; the oscillations of a cutting instrument in the metal-processing machine-tool, etc. The above erroneous assumptions exposed the feebleness of the theory in its fight against very harmful and dangerous phenomena.

Earlier than in mechanics, the auto-oscillations were generally acknowledged in radio engineering, since these were namely the oscillations that were laid in the foundation of a coming radio-signal transmission, radio engineering and, later on, television.

Especially stubborn and persistent was the confrontation against auto-oscillations in chemistry and biology. Here, conservatism manifested itself more fiercely and longer. Even at the time when Belousov had discovered a reaction of an obviously oscillating nature – this nature was observable through periodic changes of solution colours – the chemists were not giving up nevertheless and stated the impossibility of the reaction on the basis that it is never possible.

It is a habitual way of thinking that great discoveries and turns in science demand some new complicated methods. In practice, it happens most often that simple but substantially new methods are sooner most wanted.

Such a substantially new approach is a phase portrait of a dynamic system. It helps to understand a natural way of appearance of auto-oscillations and this approach makes us accept them despite our erroneous intuition. It was namely this approach that has easily persuaded us in the siphon's ability to give rise to auto-oscillations. Though, more essential than separate examples seem those general considerations that touch the point of how the auto-oscillations can be depicted on a phase portrait and how this image can appear on the portrait. This simple geometric image, a closed phase trajectory, had been discovered by J. H. Poincaré; and this scientific mathematical discovery was connected with practice and the theory of oscillations by A.A. Andronov. His well-known publication carried the same name – "The limit cycles of J. H. Poincaré and auto-oscillations".

How can the phenomenon of appearance of auto-oscillations in dynamic systems be explained to-day? How can one explain the natural and usual nature of this phenomenon?

Let a dynamic system be described by differential equations

$$\dot{x} = X(x) \ ,$$

where x is a vector; and its phase trajectories are envelopes of the system field vectors, i.e. the velocity vector $\dot{x}$ is tangent in each of their points. If the field of velocities, i.e. the vector-function $X(x)$ is such that the phase trajectory is closed, then auto-oscillations will arise. For this purpose, the vector-function $X(x)$ should be chosen properly, and this thing is done easily. The vector-function should be chosen so as to have the bunch of phase trajectories entered its own tail and compressed itself, as shown in figure 4.1.

The compression of phase trajectories ensures not only the existence of a closed phase trajectory but its stability as well. In life this happens very often: this very situation creates the possibility that a fiddle string is singing, an organ is ringing, a flute is lifting, a nightingale is singing, we are speaking and singing, our heart is running, we are walking and running, your watch is running, radio signals are being transmitted, grasshoppers are chirring, a car brake is squeaking
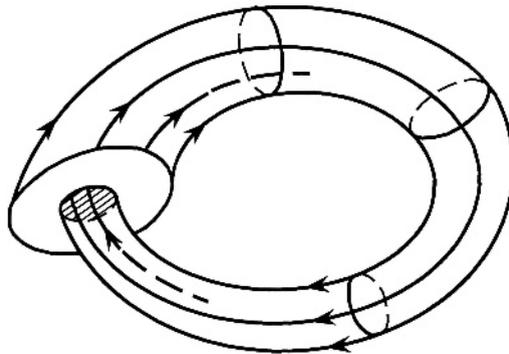
Fig. 4.1. A behaviour of phase trajectories giving rise to an auto-oscillation in the three-dimensional case.

disgustingly, and more disgustingly squeaking is the finger-nail when being drawn across the glass.

For a two-dimensional case, i.e. for the case of a two-dimensional phase space, it becomes possible to depict more vividly a general behaviour of phase trajectories yielding the auto-oscillations (i.e. yielding a stable closed phase trajectory, to which the neighbouring trajectories are arriving). Let one phase trajectory γ_1 be winding up, and another phase trajectory γ_2, lying inside, be winding off, as shown in figure 4.2.

It is seen here that between them there lies the closed phase trajectory γ onto which they are being wound (shown by a broken line). Later on you will be suggested some specific examples describing a real object. They will describe such a behaviour of phase trajectories. Right now, only an example of a differential equation is given. Let φ, r be polar coordinates satisfying differential equations

$$\dot{\varphi} = 1 , \quad \dot{r} = r - r^2 .$$

From this equation it is seen that the phase trajectory, starting at $\varphi = 0$ and $r = 1/2$, is winding off, and that, starting at the point $\varphi = 0$, $r = 2$, is winding up. Both trajectories are separated with the circle $r = 1$ corresponding to a stable periodic rotation, i.e. an auto-oscillation.
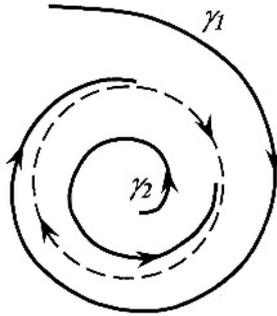
Fig. 4.2. The behaviour of phase trajectories giving rise to an auto-oscillation in the two-dimensional case.

5 Dynamics of the water surface level in a reservoired hydropower station

A phase portrait and equilibrium regimes for a hydropower station. Critical values. Bifurcation diagram.

In our previous chapters we were constructing and studying mathematical models for a water-filled barrel with water flowing out through a bottom hole or a siphon;water inflow is also possible. A barrel is the object deserving our attention. Except his astrological investigations, Johannes Kepler studied the problem of measuring a volume of the water-filled barrel without pouring out its water. Among his contemporaries, the solution of this problem was one of his most popular investigations. Though, it was long ago and since that time the "barrel became less urgent". So here, we also jump from a barrel simulation to the mathematical models of more significant objects, and namely, to simulating a reservoired hydropower station and also a priceless fluid pump held in our breast, our heart. These objects are of complicated nature, but, nevertheless, their models will remain simple, primitively simple. Despite their simplicity, these models are able to reveal us something important, deeply veiled and interesting.

Figure 5.1 schematically shows the water reservoir of the water level H and the dam to sustain this level; the dam is equipped with a hydraulic turbine and an electrical generator. From the reservoir, water runs through a tubular corridor to the turbine and turns it round. The generator connected to the turbine produces electrical current of the needed power W. The reservoir has some water inflow whose total inflow intensity is equal to Q. The magnitude Q is time-dependent but assumed constant. Water in the reservoir is spent for rotating the turbine. Also, there should be taken into consideration both water evaporation and filtration, which will increase as soon as the reservoir water level increases. Our target is to study the water level variation H. The water discharge required for electrical current production will depend upon the intensity of using the water by the turbine and also upon the water pressure determined by the water level H. With all these things taken into account, it becomes possible to write the following equation

$$S\frac{dH}{dt} = Q - I - F - \frac{W}{k\rho g(H+h)} \quad , \qquad (5.1)$$

where $S(H)$ is a reservoir surface square; I and F are evaporation and filtration intensities, respectively. Here, the last term stands for the water discharge

intensity needed to produce the electrical power W, provided that the power station efficiency coefficient is equal to k.
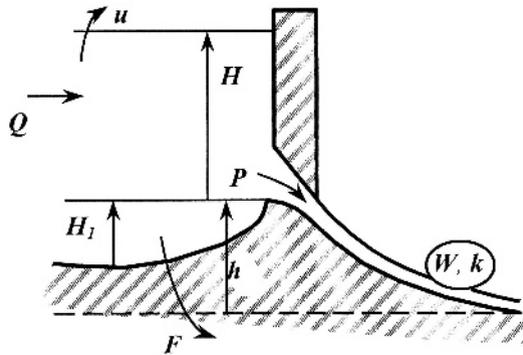
Fig. 5.1. The scheme of the reservoired power station with a dam.

The last term (the water discharge intensity) in formula (5.1) is calculated in the following way. Let the desired total volumetric water discharge be equal to P. If the water overfall before the hydroturbine is equal to $H + h$, then the water outflow from the reservoir to the hydroturbine for the time dt will release the energy equal to

$$\rho g P(H + h)dt \ .$$

One portion $k\rho g P(H + h)$ of this energy will be transformed to electrical energy of the needed amount W. (Here k stands for the efficiency coefficient of the hydroturbine, with the electrical generator included). Thus, we obtain

$$kpgP(H + h) = W \ ,$$

from which the needed intensity of the hydroturbine volumetric discharge

$$P = \frac{W}{k\rho g(H + h)}$$

may be found.

Now, we will study the differential equation (5.1) derived. For this purpose, we need to depict the graph of the right-hand side of the equation as the function of

H. The right-hand side of this equation involves the constant component Q, the magnitudes I and F increasing with the growth of the level H, and the last term which, on the contrary, increases with the decrease of H. For $H < 0$ the last term will vanish, since the water delivery to the turbine will cease. Accordingly, the more complete form of the equation (5.1) will be as follows

$$S\dot{H} = \begin{cases} Q - I - F - \dfrac{W}{k\rho g(H + h)} & for \quad H > 0 \\ Q - H - F & for \quad H < 0 \end{cases}. \tag{5.2}$$

The plot for the right-hand side of the equation (5.1) will be of the form given in figure 5.2.
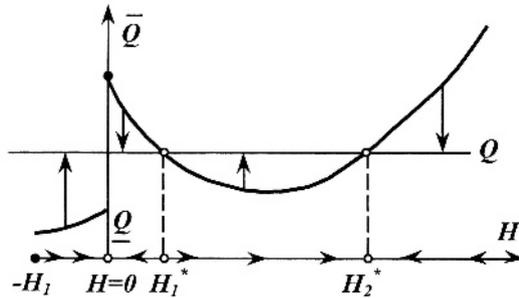
Fig. 5.2. The phase portrait of the system "reservoir – hydroelectric station with a dam" described by the differential equation (5.1).

In this figure, the right-hand side magnitude indicated through arrows is the difference between Q and the rest terms. In accordance with this plot, there are three equilibrium states on the phase half-line $H \geq -H_1$. The equilibrium state $O_2 (H = H_2)$ is a stable one, in which a fully efficient running of the hydroelectric station is provided, with the required power W being produced. The equilibrium state O, with $H = 0$, is also stable but fails to ensure the full power W and ensures only a portion of it equal to

$$k\rho g(Q - I - F)\Big|_{H=0} < W,$$

because for $H = 0$ we, as seen in figure 5.2, obtain

$$Q - I - F - \frac{W}{kg\rho h} < 0 \ .$$

At last, the unstable equilibrium state $O_1(H = H_1*)$. This state is interesting by the water level increasing, for $H > H_1*$, up to the magnitude $H = H_2*$; and here the needed energy output is provided. On the contrary, at $H < H_1*$ the level H falls down to zero and this results in the inefficient operation of the hydroelectric station. Note here that the hydroelectric station fails in its running not due to the lack of the water needed but rather because of the incorrect choice of the operating regime done. You know that with the same inflow Q, when the system lies in the equilibrium O_2, the station is just able to provide the electric power W.

From the above said it follows that no drop of water level lower than $H = H_1*$ is admitted. It is the level H_1* that becomes crucial; if the case is below this level, then some emergent measures will be needed. Namely, electrical energy consumption and station output have to be temporarily reduced; only upon the level H having   exceeded H_1* we should again return to the normal exploitation of the hydroelectric station.

Now let us see what emergencies may arise due to the temporary decreases of the inflow intensity Q or due to the cases when the temporary increase of the power W is needed. As seen from the plots in figure 5.2, the drop of the inflow Q below some value $Q_{m\,arg}$ can result in the decrease of H below the critical level H_1*. With Q decreasing, the equilibrium states O_2 and O_1 will approach each other; for $Q = Q_{cr}$ they will merge; for $Q < Q_{cr}$ will vanish. Upon this, the level H will approach $H = 0$ and may become less than H_1*. An identical result can be also achieved through the temporary increase of W. In these both cases, to return to the previous normal regime of the hydroelectric station it is necessary to temporarily drop the power output and wait until the level H has exceeded its critical value H_1*.

Now let us assume that the hydroelectric station cannot vary its output, whereas the inflow Q is varied by the humanly-uncontrollable nature. Let us draw the diagram showing how the equilibrium operating conditions of the station depend upon the inflow Q. At one value of Q (depicted in figure 5.2) we obtain the below equilibrium states $O_1(H_1*)$, $O_2(H_2*)$ and $O(H_3* = 0)$ (figure 5.3).

Here, the circles stand for stable equilibrium states and crosses for unstable. Now we shall start to vary Q. With Q decreasing, the equilibrium states O_1
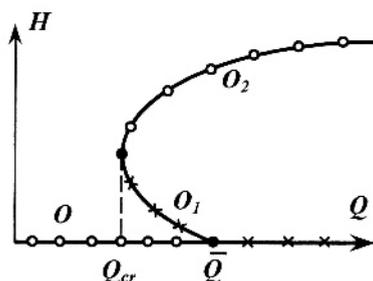
Fig. 5.3. The bifurcation diagram of the balanced levels in the reservoir with a hydroelectric station.

and O_2 will approach each other (H_2* is decreasing and H_1* increasing). For the equilibrium state O, the magnitude H will still retain equal to zero. At some critical value $Q = Q_{m\,\mathrm{arg}}$, the equilibrium states O_1 and O_2 will merge, and with Q decreasing further these states will vanish. Upon their having vanished, there will retain only the single stable equilibrium O. In this equilibrium state $H = 0$, the hydroelectric station will not be able to produce the full electric power W, since the water expenditure P, determined from the relation

$$Q - I - F - P = 0 ,$$

will be less than the expenditure $\dfrac{W}{kg\rho h}$ required to produce the power W. At the reverse increase of Q, the hydroelectric station will assume the non-complete operating regime $H = 0$ until we achieve $Q = \overline{Q}$, where

$$\overline{Q} = I + F + \frac{W}{kg\rho h} \quad .$$

Upon Q growing above $\overline{Q}$, the operating regime O will cease to be balanced. Therefore, the hydroelectric station will arrive at the favourable operating regime Q_2, at which it will be able to supply the needed power W.

In figure 5.4, repeating figure 5.3, it is shown how the operating regime of the station will vary under, first, the slow decrease of the inflow Q (described above) and then its increase.
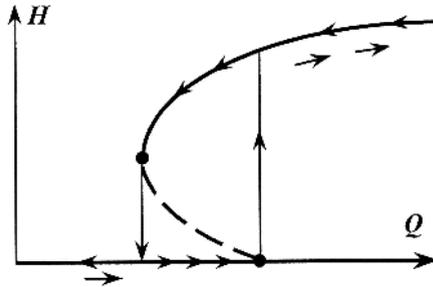
Fig. 5.4. The hysteresis phenomenon for the dependence of the reservoir level H upon the inflow Q (illustrated graphically).

These changes are specific in the following: the same inflow Q for $Q_{cr} < Q < \overline{Q}$ yields the two substantially different operating regimes for the hydroelectric station, corresponding to the two stable equilibriums, O_2 and O. From the viewpoint of the hydroelectric station functioning, the first regime will be good and the second poor. Which of these operating regimes will occur depends upon the history of the variations of Q. It is easily seen here that the nature of this dependence looks similar to that we had when were dealing with the siphon.

The diagrams in figures 5.3 and 5.4 help us to easily trace the dependence of possible equilibrium operating regimes upon the inflow Q and reveal how these regimes replace each other at the slow changes of Q. This diagram is usually called a bifurcation diagram for the equilibrium states with respect to the parameter Q, i.e. the diagram for changes of equilibrium states, when Q varies.

From the above said, it follows that it is hardly possible to realize the control strategy, under which the constant output W of the station could be obligatorily demanded despite the fluctuations in the level H and the inflow Q. Such strategy will be poor as well, because it may entail the situation, in which the station will fail to ensure the output W. Though, another control strategy may exist, under which this output could be provided absolutely.

It is clear that the most favourable operating regime for the station is its running at the maximally possible H_{max} (being determined by the dam height), since this

case will provide the minimal water discharge per a unit of the power produced. This very operating regime is the most efficient not only from the viewpoint of water discharge but also it produces a maximally possible power for the given inflow Q. This operating regime is ideal., so to say. Though, with the given H_{max}, this regime is realizable in the case of the absolutely certain inflow Q equal to

$$I + F + \frac{W}{kg\rho(h + H_{max})} \ .$$

If Q decreases, then this optimal operating regime will be disturbed and will transfer to a less favourable one (with the output W retained), providing the level H smaller than H_{max}. Here, it is clearly seen that this situation brings a loss of possible total energy. In order to avoid this, the electrical power W can be reduced so as to retain the water level equal to H_{max}. You see here that we encounter a problem of the optimal control for the station. To study the similar problems theoretically is an objective for the special science that has appeared not long ago and is still under development. This science is the theory of optimal control. It is an independent part of the more general science called the operations research.

6 Energetic model of the heart

Types of crisis states and crucial values. Narrowing the vital capabilities.

We will construct and study a very simple energetic model of the heart. This very complicated organ will be studied via a very simple model. It turns out to be possible. A very simple model turns out to be useful and is able to tell us something very important and interesting.

From the functional point of view, the heart is a four-chamber pump supplying blood for the entire organism. One half of the heart pumps blood along the so-called small circle, through the lungs enriching them with oxygen. The blue venous blood turns into red, its red corpuscles, upon absorbing oxygen, change its colour. Another half of the heart is responsible for supplying all human organs and tissues with arterial blood filled with oxygen. This is the so-called big circle of the blood circulation. The heart cannot cease its work, it must run continuously, day and night during the entire life.

Though, the heart is far not a simple pump. It is the pump controlled by the commands from the vegetative and the central nervous systems. Its functioning is coordinated with a physical and psychological loads of its owner. The heart should sustain the balance between its big and small circles of blood circulation. Its rhythm and operational intensity are regulated by commands from the nervous system. This nervous regulation is a multistaged one and of great complexity. The heart performs its mechanical work at the expense of the chemical energy accumulated by it. This energy is continuously replenished by the blood incoming as a result of heart's functioning. Thus, the heart, so to say, feeds itself and works thanks to this very feeding.

The heart is a complicated controllable pump responsible for pushing the blood through the small and big circles. Here, the blood is accepted by the heart auricles and pumped further by the heart ventricles. For this purpose there should exist a system of valves (a mitral valve, an aorta valve and others) and also some complicated networks of arterial and venous vessels, right down to the thinnest capillary vessels transporting the blood to the tissue cells. Converting the chemical energy of the adenosindiphoshatic acid to the mechanical work is a very complicated process. Also very complicated is the control system of the heart being responsible for sufficiently supplying the blood to the body without any excession or increase in the time of physical, mental and stress loads. The control system of the heart performs a coordination of the blood pumping across the big and the small circles causing, thus, no blood overfill in organs and tissues of the body. Elastic contrac-

tions and relaxations of vessels are of great help to this process. The control system provides also a synchronous contraction of some heart muscle tissues and a possibility of heart's autonomous activity. All these things are too complicated. Mathematical models of the heart are greatly needed for the contemporary medicine in order to understand how to help the heart, especially in the so-called emergency situations – when performing reanimation, surgery on the heart and when treating this or that disease.

Contemporary models are able to simulate only some separate aspects of this very complicated activity of the heart. Right now I am willing to describe you a very simple model which simulates only the fact that the heart is controlled by the nervous system and by some chemical substances transported into the heart, and also the fact that the heart lives only because it feeds itself through its functioning. Accordingly, the heart will be described by two variables only – by the control command u being executed by the heart unconditionally and by its current energy stock Q being spent for heart's running and replenished by the blood circulating through the heart. The energetic stock Q is even consumed by the idle heart, i.e. when it is not running. This stock is spent for sustaining the life of heart's tissues. In accordance with this, we may write the below differential equation

$$\frac{dQ}{dt} = -a - f(u,Q) + g(u,Q) , \qquad (6.1)$$

where a is an intensity of consuming the energetic stock Q in order to sustain the life state of the nonfunctioning heart; $f(u,Q)$ is an intensity of spending the energetic stock by the heart for its blood pumping; and $g(u,Q)$ is an intensity of replenishing the stock Q by the blood incoming the heart . The magnitudes u and Q may vary within some limits $0 \le u \le u_{\max}$, $0 \le Q \le Q_{\max}$ (figure 6.1).
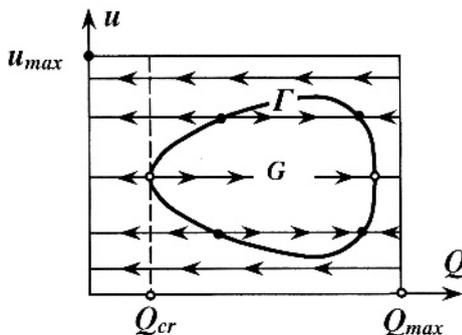
Fig. 6.1. The geometrical interpretation of the heart energetic model described by the differential equation (6.1) on the plane of the heart energetic stock Q and the control u .

In order to make the written model specific, one should be aware of the functions $f(u,Q)$ and $g(u,Q)$, at least on a qualitative basis. It is clear that both at $u = 0$ and $Q = 0$ these functions are nullified, i.e.

$$f(0,Q) = f(u,0) = g(0,Q) = g(u,0) = 0 \ .$$

Further, $f(u,Q)$ is a nondecreasing function of both arguments. Since Q cannot exceed $Q_{\max}$, then for $Q = Q_{\max}$ we get $\dfrac{dQ}{dt} \le 0$ and therefore obtain

$$- a - f(u,Q_{\max}) + g(u,Q_{\max}) \le 0$$

It is natural to assume that for $u = u_{\max}$, i.e. at maximum loads, the heart will be fast exhausted and here we get $\dfrac{dQ}{dt} < 0$. Therefore, on the sides $Q = 0$, $u = 0$ and $u = u_{\max}$ of the rectangle (Fig. 6.1) the right-hand side of the differential equation (6.1) will be negative.

On the fourth side $Q = Q_{\max}$ it will be not positive. And meanwhile, we live a long life and ,therefore, there should exist some domain inside the rectangle where the right-hand side of the differential equation (6.1) is positive. It will then lead us to the picture (figure 6.1) where arrows show the direction of replacement of the point $M(u,Q)$, with u being fixed. An exact boundary of the domain, within which the internal energetic stock Q of the heart will grow, is certainly not known , but for our purpose this knowledge is not needed. Our further conclusions will rest only upon our general ideas concerning the shape of the domain G inside which the stock Q will grow until its boundaries. Beyond this domain G , the point $M(u,Q)$, when travelling under the fixed control command u, will arrive either at the segment with $Q = 0$ or at the portion Γ of the domain boundary .

Observing the possible motions of the point $M(u,Q)$ as u varies, it may be noticed that until the point $M(u,Q)$ lies to the left of the line $Q = Q_{cr}$ it may arrive at any place and may be brought to any place where $Q > Q_{cr}$. Conversely, for $Q < Q_{cr}$ its fate will be predetermined; independently of the variations of u the point $M(u,Q)$ will arrive at the boundary $Q = 0$. This will bring death to the heart. Death will be inevitable, if no associated reanimating measures are adopted. Thus, it is clear here that only until the point $M(u,Q)$ lies to the right from the line $Q = Q_{cr}$ it may stay beyond the domain G within that its part

where it arrives at the segment of $Q = 0$. Besides, these urgent measures to eliminate the fatal displacement of the point $M(u, Q)$ must be different for the case of this point lying above the domain G and for the case when it lies below this domain. In the first case, further high loads upon the heart should be timely cancelled, i.e. u should be reduced. In contrast to this, the second case implies a necessity to stimulate heart's activity, i.e. u should be increased. It is not difficult to see here that the first case implies prolonged loads which should be stopped. The second case represents an insufficient activity of the heart, insufficiency of its stimulation; this case needs an increase of u, with possible exciting medicines administered as well.

Thus, our model distinguishes only the two absolutely different crucial states of the heart: the first state is called a durable load upon the heart and the second implies a durable extreme weakening of the heart's activity. Note here that a durable intensive activity of the heart may be caused not only by actual physical overloads but also by some stress and overexcitation of the nervous system.

Now, let us see how the domain G will undergo variations in the situations when the heart-feeding vessels are constricted and the efficiency coefficient of the heart decreases. This is caused, in particular, by a lack of training, intoxication or by a general exhaustion of the organism.

The first case will decrease the value of the feeding function $g(u, Q)$; the second will increase the value of the function $f(u, Q)$.

Both cases will produce the same effect – the lessening of the domain G which will reduce the vital capabilities of the organism. In particular, here Q_{cr} will increase and there will appear a decrease in the loading capabilities and simultaneously in the heart's relaxational capabilities. However paradoxical it may seem but in this situation the heart will have to work more intensively and at the narrowed restrictions of overloads. As a result of this, the heart will be deprived of any relaxation and will possess a small power stock and a very limited reserve. This situation may be partially bettered via artificially widening the coronary vessels by medicines and, possibly, by moderate training, in order to increase the efficiency coefficient of the heart's muscle and that of the entire heart.

Such are the conclusions possible to be made on the basis of this primitive energetic model of the heart. This model covers only the fact that the heart's functional intensity is governed by the control u and that the heart supports its life through its unceasing work. It may be noted here that any living creature placed amidst merciless nature is granted the same specificity – it is forced to constantly escape from its enemies, search for food and get it. The intensity of these searches and escapes is dictated by it itself and circumstances. The difference between this poor creature and our human heart lies only in the rigidity of the external demands which humans have successfully managed to soften, more or less.

7   Soiling a water reservoir with a bay and the Caspian Sea puzzles

Polluting a water reservoir with sewage of dissolvable pollutants. The Caspian Sea puzzles. Soiling a water reservoir through a bay. Regulating a water regime and salinity.

The next mathematical model to be discussed is soiling or solvable pollution of a water reservoir by incoming waters. Besides its ecological aspect, this model relates to the mystery of the substantially different salinities of the Black and Caspian Seas, having a common origin. The soiling model for the Caspian Sea is constructed with use of our previous consideration of the equilibrium water level when an inflow and outflow take place. Though, the Caspian Sea equilibrium level may carry substantial specificity to be discussed below. When discussing the soiling problem of some certain closed restricted water reservoir, it is clear that its incoming waters with dissolvable substances make this reservoir more and more soiled and this results in saturation and sedimentation .

Let Q be an intensity of the incoming stream and v a concentration of the polluting substances. Further, let υ be a volume of the reservoir and I an intensity of evaporation from its surface. The evaporating intensity depends upon many aspects, among them are air humidity, air temperature and wind. Though, its average value is determined mainly by the size of the evaporating surface being dependent upon the volume υ. Analogously, the incoming stream intensity Q will vary and depend upon many reasons, though its average value Q will retain sufficiently constant.

Let the equation for the balance of the incoming and evaporating water be of the form

$$\frac{d\upsilon}{dt} = Q - I(\upsilon)$$

and the equation regulating the accumulation of common pollutants in the reservoir have the form

$$\frac{dM}{dt} = vQ \quad .$$

Since $I(\upsilon)$ is an increasing function of υ, the first equation will yield the stable equilibrium level υ * determined by the equation

$$Q - I(\upsilon) = 0 \ .$$

This is explained by the plot in figure 7.1.
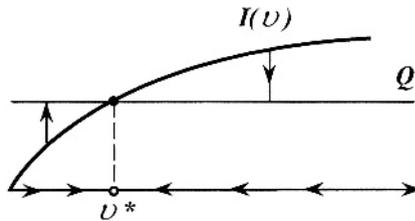
Fig. 7.1. The phase portrait of the water level dynamics in the reservoir with the evaporation dependent upon a water volume.

From the second equation it follows that the size of pollution grows steadily. First, this growth causes a growth of polluting concentration equal to M/υ, and upon it, when saturation has occurred, the concentration ceases to grow. Then, the process of bottom sedimentation is initiated.

It is seen here that this model is not capable of revealing the reasons for existing differences in the salinity of the Black and Caspian Seas. Some other circumstances have to be taken into consideration. What are they exactly? The Black Sea is not completely closed. Through the narrow Bosporus strait it is connected with the Mediterranean Sea. At present, through the Bosporus strait, water is chiefly flowing out . It is so at present, but earlier this sea was closed. As for the Caspian Sea, it is always isolated from the ocean, though, in spite of it its salinity is greatly less than that in the Black Sea.

Where then does the basic difference lie? This difference at first is not striking our eyes. The deference is explained by the presence of the Cara-Bogaz-Gol bay of the Caspian Sea. At the first sight, this presence seams of no great weight, for the Caspian Sea remains closed so far. But actually it is not absolutely so, because the waters of the bay and the Caspian Sea are not being intermixed. The Caspian Sea waters flow into the bay always. The strait connecting the Caspian Sea and its bay is narrow and long; a water drop in it makes up about 4.5 metres. This specificity will be accounted by us in our more complete mathematical model which we will construct now.

Let υ and υ_1 be total water volumes in the reservoir and the bay; Q and v will be given the previous values; I and I_1 will be the evaporating intensities of

the reservoir and the bay; q will be the intensity of the water outflow from the reservoir into the bay (figure 7.2).

In this notation we immediately obtain

$$\frac{d\upsilon}{dt} = Q - I - q$$

$$\frac{d\upsilon_1}{dt} = q - I_1$$

(7.1)

Besides, we have

$$\frac{dM}{dt} = vQ - \mu q \ ,$$

(7.2)

$$\frac{dM_1}{dt} = \mu q \ ,$$

where M and M_1 are general masses of the dissolvable pollutants in the reservoir and the bay; μ is a pollution concentration in the reservoir. Assuming the reservoir pollution having not yet arrived at its saturation, we shall obtain that $\mu = M / \upsilon.$
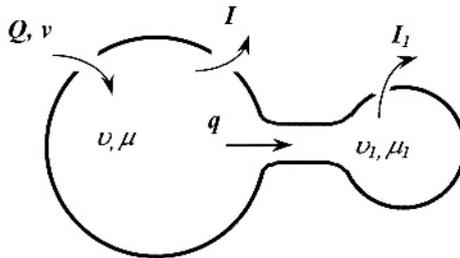
Fig. 7.2. The Caspian Sea with its Kara-Bogaz-Gol bay (represented schematically).

In contrast to this, the saturation in the bay has been taking place for a long time already; gigantic sodium, magnesium and other sediments have been already formed on its bottom.

The system of differential equations (7.1) and (7.2) obtained describes a water balance and a balance of the dissolvable pollutants. The first two equations for the water balance are independent of the remaining two; thus, they may be considered separately. However, this consideration is hampered by I and I_1 being dependent upon υ and υ_1 , respectively, and by q being a function of υ and υ_1 . Besides, to write these specific dependences is extremely difficult. Alongside with it , these equations are evident to have the stable equilibrium state $\upsilon *$, $\upsilon_1 *$ being found from the equations

$$I(\upsilon) + q(\upsilon,\upsilon_1) = Q$$
$$I_1(\upsilon_1) - q(\upsilon,\upsilon_1) = 0$$

.

(7.3)

We will deal with this later; and now, assuming a presence of such a stable equilibrium, we come to the analysis of the differential equations (7.2) for the mathematical model of the water reservoir with the bay. The last equation is already well known and brings us nothing new: a mass of the pollutants in the bay will infinitely increase with time. Thus, the concentration of pollutants will grow and then this growth will cease as soon as saturation and sedimentation are achieved.

Conversely, the first differential equation informs us about a drastic change of the situation. It informs us about the fact that an infinite growth of pollutants has been transferred to the equilibrium concentration $\mu *$ found from the conditions under which the right-hand side of the differential equation is nullified.

$$\mu q(\upsilon,\upsilon_1) - vQ = 0 \ .$$

With $\upsilon \rightarrow \upsilon *$, $\upsilon_1 \rightarrow \upsilon_1 *$ taken into account, we shall find the below equilibrium concentration

$$\mu* = \frac{vQ}{q(\upsilon*,\upsilon_1*)} \ .$$

It is worthy of mentioning here that the water equilibrium arises more rapidly than the equilibrium concentration does. Therefore, the general process of the arrival at the equilibrium may be divided into two stages: a relatively rapid arrival at the equilibrium volumes $\upsilon *$ and $\upsilon_1 *$ and the comparatively slow arrival at the equilibrium concentration $\mu *$. Upon arriving at the equilibrium volumes $\upsilon *$

and $\upsilon_1{}^*$, the equilibrium concentration, due to (7.2), will be described by the differential equation

$$\upsilon^* \frac{d\mu}{dt} = vQ - q(\upsilon^*, \upsilon_1{}^*)\mu .$$

This equation is solved easily and its solution with the zero initial condition will be of the form

$$\mu = \frac{vQ}{q(\upsilon^*, \upsilon_1{}^*)} \left[1 - \exp(-\frac{q(\upsilon^*, \upsilon_1{}^*)}{\upsilon^*} t) \right] =$$

$$(7.4)$$

$$= \mu^* \left[1 - \exp(-\frac{q(\upsilon^*, \upsilon_1{}^*)}{\upsilon^*} t) \right] .$$

We notice here that, as it follows from the formula (7.4), the concentration μ will constitute the $1 - e^{-1}$ part of the equilibrium concentration μ^* at the time υ^* / q^*. This time is equal to the time during which the equilibrium volume υ^* will outflow from the Caspian Sea into the Cara-Bogaz-Gol bay. With this time doubled, the concentration μ will be already equal to $(1 - e^{-2})\mu^*$; hence, as an estimate for a transition time of the equilibrium concentration μ^* one may take the time $\tau = \dfrac{\upsilon^*}{q^*}$.

Now we are return to the equilibrium concentration μ^*. It is equal to

$$\mu^* = \frac{vQ}{q(\upsilon^*, \upsilon_1{}^*)} .$$

and seemingly demands the knowledge of v, Q and q. However, making use of the equations (7.3) determining the volumes υ^* and $\upsilon_1{}^*$, we shall easily find that

$$\mu^* = \frac{vQ}{q(\upsilon^*,\upsilon_1{}^*)} = v\frac{q^*+I^*}{q^*} = v\left(1+\frac{I^*}{q^*}\right) = v\left(1+\frac{I^*}{I_1{}^*}\right), \quad (7.5)$$

where stars indicate the associated magnitudes being taken at the equilibrium values $\upsilon = \upsilon^*$ and $\upsilon_1 = \upsilon_1{}^*$.

The ratio $I^*/I_1{}^*$ in (7.5) is roughly approximately equal to the ratio of the reservoir and bay surfaces, i.e. S/S_1. For the Caspian Sea and the Cara-Bogaz-Gol bay, this ratio is approximately equal to 37. Therefore, the equilibrium concentration in the bay exceeds only as much as about 40 times the average salt concentration in the sweet water streams of the rivers and rains running into the Caspian Sea.

The Cara-Bogaz-Gol bay, thus, plays for the Caspian Sea a role of some sedimentation tank sucking off the dissolvable pollutants from the sea. The same role for soiling reservoirs may be played by artificial bays. For this, there should be provided an unceasing level drop of the water from the reservoir to the bay. Otherwise, a reverse stream from the bay to the reservoir will spoil everything.

Therefore, a smaller salinity in the Caspian Sea as compared against that in the Black Sea may be explained via the existence of the Cara-Bogaz-Gol bay. However, our assumption concerning the equilibrium of the volumes υ^* and $\upsilon_1{}^*$ does not hold entirely. The Caspian Sea level varies in a very surprising way – it is fluctuating about one magnitude or the other. The difference between these fluctuating levels significantly exceeds the oscillations about each of them. These changes are hardly explained by either weather changes or tectonic changes in the bed of the Caspian Sea. This is also one of the puzzles of the Caspian Sea.

The relatively slight fluctuations of the Caspian Sea level do not eliminate the above explanation for the small salinity of the sea and the formula (7.5), but these fluctuations are significant for the exploitation of the Caspian Sea and interesting in themselves. Prior to starting to outguess this wonderful phenomenon, let us investigate the seemingly evident supposition concerning an existence of the stability of the equilibrium volumes υ^* and $\upsilon_1{}^*$ in the Caspian Sea and the bay. Let us also find out whether this equilibrium is unique or not and whether there exist any auto-oscillations as well.

We return now to the differential equations (7.1) of the water balance in the Caspian Sea and the bay. A phase space of this system is the first octant of the plane υ , υ_1, for which $\upsilon \geq 0$ and $\upsilon_1 \geq 0$. First of all, let us find that all the phase trajectories arrive at its finite part of the form

$$\upsilon + \upsilon_1 = C > 0 \ .$$

It follows from the fact that for sufficiently large υ and υ_1 (large C) we obtain

$$\frac{d}{dt}(\upsilon + \upsilon_1) = Q - I(\upsilon) - I_1(\upsilon_1) < 0 \ .$$

Figure 7.3 shows the cross sections of the Caspian Sea, its bay and the strait connecting them, demonstrates the highest level of the strait.
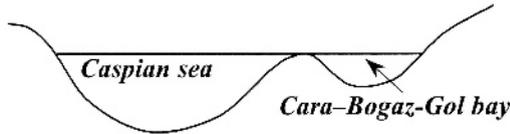
Fig. 7.3. The bottom profile of the sea, its bay and the narrow strait connecting them.

From this figure it follows that the phase space, the first octant $\upsilon > 0$, $\upsilon_1 > 0$ of the plane υ, υ_1 is decomposed into the parts where $q = 0$, $q > 0$ and $q < 0$. It is clear that $q = 0$ when the Caspian Sea level and the bay level are lower than those in figure 7.3. Let it take place for $\upsilon < \underline{\upsilon}$ and $\upsilon_1 < \underline{\upsilon}_1$. The case $q = 0$ takes place also when the level of the Caspian Sea and that of its bay are equal to each other but $\upsilon > \underline{\upsilon}$ and $\upsilon_1 > \underline{\upsilon}_1$. The set of the phase plane points, satisfying $q = 0$, consists of the rectangle $\upsilon \leq \underline{\upsilon}$ and $\upsilon_1 \leq \underline{\upsilon}_1$ and some curve γ presented in figure 7.4 .
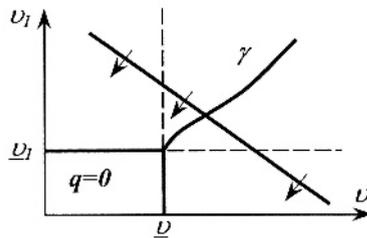
Fig. 7.4. The graphically illustrated search for the domain at which arrive all the trajectories of the phase space $\upsilon \geq 0$, $\upsilon_1 \geq 0$.

This figure also holds the straight line $\upsilon + \upsilon_1 = C$ intersected by phase trajectories from outside to inside as shown by arrows. This line together with the axes $\upsilon = 0$ and $\upsilon_1 = 0$ confines the domain, into which all phase trajectories income, i.e. upon some sufficient time any phase point (υ, υ_1) will fall into this domain.

Now, upon our preliminary considerations, let us come to studying the steady-state motions of the system (7.1). It is clear that they depend upon the inflow Q, the functions $I(\upsilon)$ and $I(\upsilon_1)$, and the function $q(\upsilon, \upsilon_1)$.

For $Q < I(\underline{\upsilon})$ all the phase trajectories will enter the quadrangle, where $q = 0$, and will then come close to the equilibrium state $\upsilon = \upsilon^* < \underline{\upsilon}$ and $\upsilon_1 = 0$, where υ^* is a root of the equation

$$Q - I(\upsilon) = 0 \ ;$$

the root is unique, provided that $I'(\upsilon) > 0$. As far as only an inflow from the Caspian Sea was included into the second equation (7.1), the bay will dry up. Notice that this very event has happened with the Caspian Sea – to the bay strait having been temporarily blocked with a dam.

For $Q > I(\underline{\upsilon})$, all the phase trajectories enter the triangle confined by the straight lines $\upsilon = \underline{\upsilon}_1$, $\upsilon_1 = 0$ and $\upsilon + \upsilon_1 = C$; thus, this triangle holds the equilibrium states. In one part of this triangle we have $q > 0$ and in another $q < 0$. They are separated with the curve γ, on which $q = 0$. Above and on this curve we have $q \leq 0$ and $\dot{\upsilon}_1 < 0$. Therefore, all the phase trajectories, starting in the curvilinear triangle confined by the lines $\upsilon = \underline{\upsilon}$, $\upsilon + \upsilon_1 = C$ and by the curve γ, depart from this triangle; so, this triangle can hold no equilibrium state.

Now we will show this equilibrium to be unique and stable. Further, it will be shown that there exist no closed phase trajectories and, thus, this unique equilibrium state is stable globally. Certainly, this rests upon certain properties of the functions $I(\upsilon)$, $I(\upsilon_1)$ and $q(\upsilon, \upsilon_1)$.

We will accept, except $Q > I(\underline{\upsilon})$, that

$$\frac{dI(\upsilon)}{d\upsilon} > 0 \ , \qquad \frac{dI_1(\upsilon_1)}{d\upsilon_1} > 0 \ , \qquad \frac{\partial q}{\partial \upsilon} > 0 \ , \qquad \frac{\partial q}{\partial \upsilon_1} \leq 0 \qquad (7.6)$$

Let us reduce (7.3) to the form

$$I(\upsilon) + I_1(\upsilon_1) = Q$$
$$I(\upsilon) - q(\upsilon, \upsilon_1) = 0$$

(7.7)

Differentiating the first equation (7.7) in υ we shall get

$$\frac{dI}{d\upsilon} + \frac{dI_1}{d\upsilon_1}\frac{d\upsilon_1}{d\upsilon} = 0 ;$$

therefore, due to (7.6), we obtain $\dfrac{d\upsilon_1}{d\upsilon} < 0$.

Similarly, from the second equation (7.7) it follows that

$$\frac{dI}{d\upsilon} + \frac{\partial q}{\partial \upsilon} + \frac{\partial q}{\partial \upsilon_1}\frac{d\upsilon_1}{d\upsilon} = 0$$

and then $\dfrac{d\upsilon_1}{d\upsilon} > 0$; besides, it may turn to infinity.

The first relation (7.7) defines υ_1 as the decreasing function υ; and the second relation (7.7) as the increasing function being able to have vertical segments or be a vertical line in general. The intersection of these curves is unique and defines the unique equilibrium state $\upsilon, \upsilon_1 *$.

Thus, under natural assumptions, the equilibrium state exists; let it take place for $\upsilon = \upsilon^*$ and $\upsilon_1 = \upsilon_1 *$.

Now we will show the equilibrium state to be stable. For this purpose, let ξ and η be such small deviations from the equilibrium that $\upsilon = \upsilon^* + \xi$, $\upsilon_1 = \upsilon_1 * + \eta$.

The deviations ξ and η meet the differential equations

$$\frac{d(\upsilon^* + \xi)}{dt} = Q - I(\upsilon^* + \xi) - q(\upsilon^* + \xi, \ \upsilon_1 * + \eta)$$

$$\frac{d(\upsilon_1 * + \eta)}{dt} = q(\upsilon^* + \xi, \ \upsilon_1 * + \eta) - I_1(\upsilon_1 * + \eta)$$

and, upon linearization, we obtain

$$\dot{\xi} = \left(-\frac{\partial q}{\partial \upsilon} - \frac{dI}{d\upsilon} \right) \xi - \frac{\partial q}{\partial \upsilon_1} \eta$$

$$\dot{\eta} = \frac{\partial q}{\partial \upsilon} \xi + \left(\frac{\partial q}{\partial \upsilon_1} - \frac{dI_1}{d\upsilon_1} \right) \eta \ ,$$

where the derivatives are calculated for $\upsilon = \upsilon*$, $\upsilon_1 = \upsilon_1*$.

Whether the equilibrium will be stable or unstable is determined by the roots of the characteristic equation, which in our case is of the second degree and has the form

$$\lambda^2 + A\lambda + B = 0 \ .$$

The stability will occur, if the roots of this equation lie to the left of the imaginary axis. For this, it is necessary and sufficient to have

$$A > 0 \ , \qquad B > 0 \ .$$

Through calculating the coefficients A and B, due to the above inequalities, we immediately find that

$$A = \frac{\partial q}{\partial \upsilon} + \frac{dI}{d\upsilon} - \frac{\partial q}{\partial \upsilon_1} + \frac{dI_1}{d\upsilon_1} > 0$$

$$B = -\frac{dI}{d\upsilon}\frac{\partial q}{\partial \upsilon_1} + \frac{\partial q}{\partial \upsilon}\frac{dI_1}{d\upsilon_1} + \frac{dI}{d\upsilon}\frac{dI_1}{d\upsilon_1} > 0 \ .$$

Thus, the equilibrium state is existent, stable and unique.

It is intuitively clear that this study was unnecessary; though, on the other hand, this study makes it impossible to conclude that except the stable equilibrium there exist no other steady-state motions that can be only stable periodic motions. Nevertheless, no such motions are existent, as it follows from the well known Bendixon criterion, under which valid is the below inequality

$$\frac{\partial P}{\partial \upsilon} + \frac{\partial Q}{\partial \upsilon_1} > 0 \ ,$$

or its reverse (where P and Q are the right-hand sides of the differential equations (7.1)), is sufficient for the absence of the limiting cycles within the domain, where this inequality holds. In the case under consideration, on the basis of our assumptions we get :

$$\frac{\partial P}{\partial \upsilon} + \frac{\partial Q}{\partial \upsilon_1} = -\frac{dI}{d\upsilon} - \frac{\partial q}{\partial \upsilon} + \frac{\partial q}{\partial \upsilon_1} - \frac{dI_1}{d\upsilon_1} < 0.$$

Hence, the equilibrium state $(\upsilon^*, \upsilon_1{}^*)$ is globally stable and all the motions will converge to it, i.e. for any initial conditions we shall obtain

$$\upsilon(t) \to \upsilon^* \ , \ \upsilon_1(t) \to \upsilon_1{}^*$$

as $t \to \infty$.

Thus, under our natural assumptions it seems not only intuitively but it has been proved that the equilibrium is unique and globally stable. Why then do these observable transitions of the Caspian Sea equilibrium level from one level to another arise? This is an actual puzzle. The assumptions done look so natural. But for us here there is no way out and we should, nevertheless, find what of the assumptions is wrong.

We recall you here that the above suggested assumptions were $Q > I(\underline{\upsilon})$ and the inequality (7.6). The puzzle, certainly, would have been fully eliminated, if the water inflow Q, averaged throughout some years, had changed drastically or the Caspian Sea bed had changed, and if these changes had been similar, more or less. But this was not observed. What are then any other faulty points? Let us now look at the inequalities (7.6) adopted by us: here the weakest place is seemingly the assumption $\dfrac{dI}{d\upsilon} > 0$. Indeed, we can think out the situation when this assumption may be false. Let the Caspian Sea bottom be of the shape shown in figure 7.5.
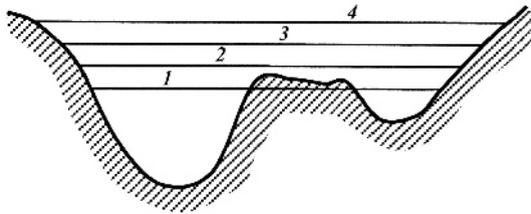
Fig. 7.5. The hypothetical profile of the Caspian Sea bottom and different possible levels 1, 2, 3 and 4 of the sea.

The hump in the centre of the figure implies possible shoal banks. Now let us see how the evaporation intensity changes in hot weather, during the transition from the lowest level (in figure 7.5 denoted through 1) to the top level 4.

During the transition from 1 to 2, evaporation will grow drastically. In the course of the transition from 2 to 4, evaporation, despite a general increase of the sea surface, may decrease, because shallow waters have become deeper and substantially cooler, and, hence, less vaporizable. In the course of the further transition from the level 3 to 4, evaporation will increase again. Thus, the existence of shoal banks may lead to the shape of the graph $I(\upsilon)$ presented in figure 7.6.
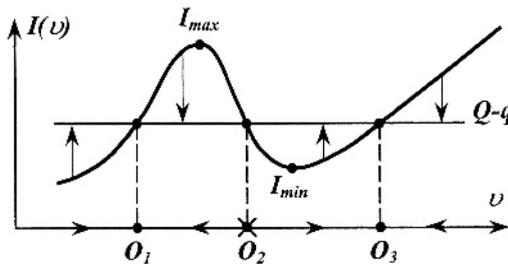
Fig. 7.6. The phase portrait of the variations of the Caspian Sea water volumes for the incoming stream Q, when the intersection takes place in the three points, to which there correspond the two stable equilibrium states, O_1, O_3, and the one unstable equilibrium stat, O_2.

Now, as seen from the same figure, the equilibriums here may be already three in number. Neglecting the magnitude $q \ll Q$ for simplicity, let us study the behaviour of these equilibriums at time changes of Q. This incoming stream sometimes falls beyond the limits I_{max} and I_{min} shown in figure 7.6.

Here, a phase space is the half-line $\upsilon \geq 0$ and a phase portrait is determined by the below equation

$$\frac{d\upsilon}{dt} = Q - I(\upsilon) \ .$$

The intersecting points of the plot $I(\upsilon)$ with the line Q in figure 7.6 determine the equilibrium states, and the arrows describe the magnitude and the

sign of the difference $Q - I(\upsilon)$, i.e. $d\upsilon / dt$. For the case in figure 7.6, there will exist the three equilibrium states, O_1, O_2 and O_3, O_1 and O_3 being stable and O_2 unstable.

How O_1, O_2 and O_3 depend upon Q may be now shown diagrammatically. The coordinates of the equilibrium states O_1, O_2, O_3 will be denoted as u_1 *, u_2 * and u_3 *, respectively. Then, we shall arrive at the dependences of u_1 *, u_2 * and u_3 * upon Q shown in figure 7.7.

The circles on this plot, being called a bifurcation diagram, stand for stable and crosses for unstable equilibrium states. For $I_{min} < Q < I_{max}$, there will exist the two stable equilibrium states, u_1 * and u_3 *. With the inflow Q decreasing below I_{min}, the both equilibrium states will turn to u_1 *; besides, the state u_3 * in a leap-like fashion. With a further increase of Q above I_{max}, the state u_1 * will turn to u_3 *, by a leap as well. With a further increase of Q, the equilibrium state u_1 * will also jump to u_3 *, if Q increases above I_{max}.
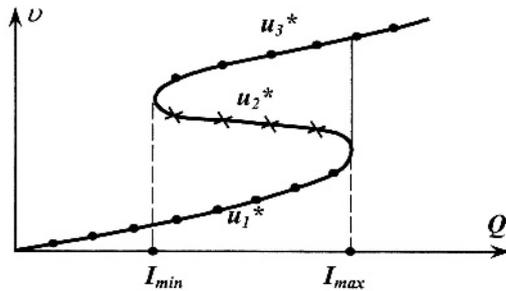
Fig. 7.7. The bifurcation diagram of the Caspian Sea equilibrium levels; stable equilibriums indicated by circles and unstable by crosses.

The above said can explain the existing changes of the Caspian Sea water levels. These changes occur by leaps from the low to the top level when the inflow Q exceeds I_{max}; conversely, the high sea level drops abruptly to the low level, when Q drops lower than I_{min}. The talk here is about leap-like changes of the stable equilibriums caused by the changes in the inflow Q. As for the Caspian

Sea level, its changes are far from of a leaping nature and it varies sufficiently slowly, in accordance with the rate of transition to a new equilibrium state.

8 Exponential processes

A mathematical model. Half-decay and halving periods. Examples of exponential processes: reproduction and destruction; radioactivity; chain reactions; capacitor discharge; missile speed-up; retarding; radiation absorption; cooling; spreading of epidemics and rumour; growth of human population, production and knowledge; arriving at and departing from an equilibrium state, etc. Specifying a model: taking into consideration the saturation and the explosive development. Phenomena of sudden crisis, collapse and exceptionality.

From narrating some separate mathematical models, we now jump to describing a sufficiently wide class of dynamic processes, the so-called exponential processes and some other processes like logistic, explosive, etc.

The exponential processes can be of two types: exponentially increasing and exponentially decreasing. These two types are described by the below very simple differential equation:

$$\frac{dx}{dt} = \lambda x \tag{8.1}$$

An exponentially increasing process occurs for $\lambda > 0$, and for $\lambda < 0$ an exponentially decreasing one. The exponentially increasing or decreasing process may occur with respect to both time t and any other physical variable being different from t.

The solution of the differential equation (8.1), taking the value x_0 for $t = 0$, is of the form

$$x = x_0 e^{\lambda t} \tag{8.2}$$

and possesses the following remarkable characteristic property: upon the time interval $\tau = |\lambda|^{-1} \ln 2$ from any initial instant t one obtains

$$\frac{x(t + \tau)}{x(t)} = 2 \quad , \tag{8.3}$$

provided that $\lambda > 0$, and the process is an increasing one, and

$$\frac{x(t+\tau)}{x(t)} = \frac{1}{2} \ , \tag{8.4}$$

provided that $\lambda < 0$, and the process is a decreasing one. This implies that to an arithmetic progression of increasing values $t + n\tau(n = 0,1,2,...)$ there will correspond a geometric progression with the denominator 2 or 1/2. These properties of the increasing and decreasing exponential processes are presented in figure 8.1.
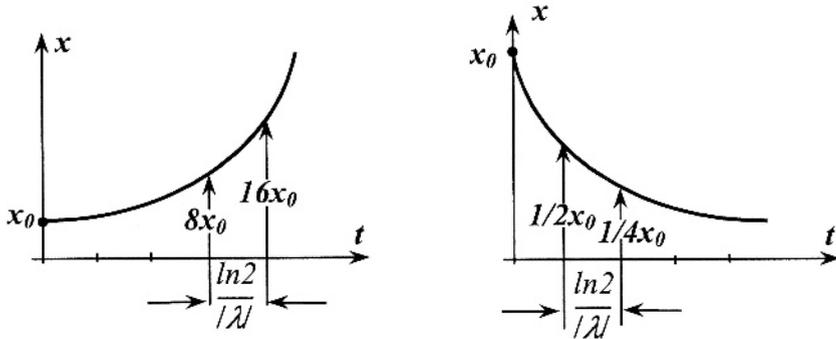
Fig. 8.1. The graphical illustration of the exponential increase and decrease and ,accordingly, of the time doubling and the time of double decrease.

The basic thing that strikes our imagination in the exponential processes is their fantastically fast, not yet humanly comprehended, increase or decrease with a larger and larger velocity. Here, as a very vivid illustration of this may serve a fairy tale about a king and a chess inventor. The king wishing to award the inventor said: "You may ask from me anything you like, even a half-country, even my daughter". But the chess inventor, being very timid, asked the king to give him a tiny trifle, and, namely, to fill the first check of the chess board with a single wheat grain, the second check with 2 grains, the third with 4, etc. The king was astonished: "What trifle things you are asking!", he exclaimed. "But let it be as you wish". Having approached only the middle of the chess board, the king's servants suddenly found that the king's granaries were already emptied. To have all the 64 checks filled with grains, one would have had the amount of grains much more than those existing on the entire planet.

The same fantastic increasing rate accompanies the limitless free reproduction of living cells, microbes and viruses. Cells and microbes are halved every one-two hours and,thus, upon 3-6 days and nights their number will be of the same order as the number of wheat grains on all the checks of the chess board. As for viruses, their halving period is much less than a single hour. Hence, a catastrophe will occur in less than 24 hours, if their reproduction is not blocked. The same manner of

acceleration and resource exhaustion is also specific for such processes as nuclear and chemical reactions, avalanche-like and avalanching processes. Such processes possess the halving time measured by the thousandth portions of a second. Therefore, these processes behave like explosive, i.e. fantastic sizes are achieved in the course of negligible portions of a second.

Similarly are spread the epidemics and rumour. In their initial stage, these processes flow exponentially too. As a vivid example of our poor orientation in the rate of growth of the exponential process there may serve our thoughtless answer to the question about by how many hand shakes a resident of a small Russian town is connected with the US President. Usually, in our hasty answer we suppose this shaking chain to be too long, whereas in reality this chain contains less than 10 shakes. We can estimate this, if every person is assumed to shake hands with not less than 100 persons in the course of his life. Hence, in a 5-person chain there will be (if repeated shakes not considered) 10^{10} persons, i.e. more than the number of all human beings on this planet. This fact demonstrates, in particular, a swift propagation of the epidemic during its short premonitory period and a greater rumour spreading rate.

In the above examples, the rate for the growth of some magnitude will be the more, the more will be this magnitude. Until the material, this magnitude is made of, has been exhausted and the conditions of the growth of this magnitude retain good, this process will retain its exponential nature and will be governed by the below differential equation

$$\frac{dN}{dt} = \lambda N \ ,$$

where N is the number of the reproduced cells or viruses or the number of the people fallen ill during the epidemic, or the number of the neutrons or molecules reproduced in the nuclear or chemical reaction, or the number of the people, whom the rumour or some sensation has approached, or the number of inhabitants in our country or that of the entire planet, etc. To somewhat extent, this law covers such situations as knowledge acquisition, studying foreign languages, since the knowledge already gained makes its further acquisition easier and quicker. This pertains both to knowledge acquisition by a single person and by the entire mankind. This pertains to the development of science, production, growth of pollution and nature destruction.

The fast growth of the exponential process brings a sudden crisis and its disappearance, resulting in fast collapse of the resources. Imagine that once you were settling down in a tent on a bank of a beautiful lake where you were going to swim. Let this lake be 16384 square metres. It is a big good lake of 200 metres in length and 100 metres in width. On that day of your coming here there were some green water-plants occupying 1 square metre only. They were not noticed by you at all. Let their halving period during their reproduction be equal to 24 hours. In the course of the first 13 days they did not disturb you, though during the last 3-4 days they caught your eye already. At last, the 14-th day came and the entire lake

was completely filled with green water-plants that made it impossible for you to swim anywhere. The catastrophe occurred during a single night. Yesterday nothing manifested the coming trouble and one half of the lake was free from water-plants. This half makes up 8192 square metres and this square is sufficiently enough for swimming. The square of good water was collapsed during 24 hours but before this, during a 13-day period, the lake had good swimming square and no troubles were expected.

The exponential growth of knowledge, science, industry, wealth and the like reveals one more specificity. Let us explain it on the individual accumulation of knowledge, the knowledge of foreign languages in particular. In the period of his active life and favourable circumstances, each person accumulates his knowledge exponentially, for the knowledge already acquired by him helps him to gain new knowledge, i.e. roughly approximately we obtain

$$\frac{dZ}{dt} = \lambda Z \quad .$$

By virtue of this, the knowledge obtained is of a threshold nature, i.e. as soon as the knowledge arrives at some threshold $Z*$ its further acquisition will assume an extremely fast nature and a human will be able to gain fantastic results during his short life. Though, to have it happened, he should arrive at some threshold reachable with great difficulties. Not everybody manages it to do. Though, those who have overcome it, then strive to the top. This brings the situation that only the limited number of people, a negligible portion of percentage, is distinguished by the size and quality of their knowledge. Hence, there occurs a colossal gap between them and the remaining people. This fact can be considered as an explanation for the phenomenon of exceptionality.

The next example touches on a suddenly arising crisis. It may be, perhaps, from your own experience. For example, you were infected a week ago. For six days you had a so-called premonitory period of the disease; the disease was not yet revealing itself. On the seventh day you felt some cold, had high temperature and your state was bad and even hard. Certainly, microbes or viruses in your organism are not reproduced without obstacles. Your powerful immune system is fighting against them, but it has not yet been fully mobilized and is, perhaps, insufficient to suppress the infection. To do this, the organism needs some time. That is why the exponential catastrophe is substantially smoothened for the time being; though, exposed. The more weakened and unprepared your immune system is and the more delayed will be its reaction to the disease, the quicker the number of pathogens and intoxication will arrive at their marginal quantities. After the first faint features of the disease your body will assume the state threatening to your life.

In the end of the 18th century T.R. Malthus was the first to tell about the exponential growth of the world population. He had a predecessor living yet in the 12th century. It was famous mathematician L. Fibonacci who put and answered the question: "How many pairs of rabbits are born yearly by one pair of rabbits?"

The exponential reproduction implies an absence of any limitations. Actually, these limitations always exist (for example, a limit of resources) and may be sim-

ply taken into account through introducing a negative quadratic term. Thus, the corresponding differential equation

$$\frac{dx}{dt} = \lambda x - \varepsilon x^2 \tag{8.5}$$

will have the solution expressed by the so-called logistic curve depicted in figure 8.2.
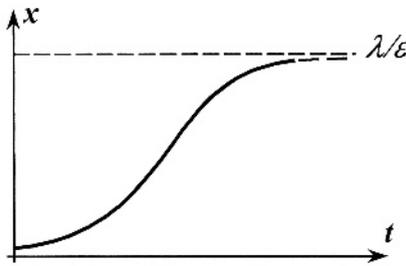
Fig. 8.2. The logistic curve.

Here, an initial increase, near to exponential, is slowed down and constrained by the limit approachable magnitude $x_\infty = \dfrac{\lambda}{\varepsilon}$ (see figure 8.2).

Many processes, being of exponential nature in their initial stage, are then slowed down and come close to those represented by a logistical curve. This curve determines the growth of world population, animals, industry, science, education and many other processes.

There exist some other processes increasing even faster than the exponential processes. For example, reproduction of insects where the size of posterity is determined by a frequency of copulations which is, within a limited square, proportional to the square of the number of its inhabitants, i.e.

$$\frac{dN}{dt} = aN^2 \ . \tag{8.6}$$

The solution of this equation

$$N = \frac{1}{N_0^{-1} - at}$$

at the initial condition

$$N\bigg|_{t=0} = N_0 .$$

unboundedly increases with t approaching the final value $\dfrac{1}{\alpha N_0}$; this will corre-

spond to an explosive growth of the population, say, a well known growth of the locusts (grass hoppers). The frequency of copulations within a limited area is proportional to N^2, because the probability of copulation of a single insect with another will be proportional to N. Therefore, the total probability of copulation of N insects will be proportional to N^2.

The exponential nature of the growth, at least, at the initial stage, is typical for the non-equilibrium departures from unstable equilibrium states. Let the system be described by the below differential equation

$$\dot{x} = f(x) \tag{8.7}$$

and $x = x*$ be its unstable equilibrium state. Figure 8.3 shows a small mountain with the top point standing for an unstable equilibrium. When rolling down the mountain, the point is gaining speed faster and faster.
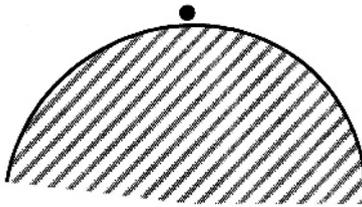
Fig. 8.3. The unstable equilibrium illustrated in the form of a material point on the top of a convex curve.

Let us introduce the new variable ξ under the assumption that

$$x = x* + \xi .$$

Then, having put this variable into the equation (8.7), we shall find that

$$\dot{\xi} = f(x* + \xi) = f(x*) + f'(x*)\xi + ..., \qquad f(x*) = 0 ,$$

where three dots denote the terms of the second and higher orders of smallness with respect to ξ. On the initial time interval of the departure from the unstable equilibrium (when ξ is still small), we may approximately assume that

$$\dot{\xi} = \lambda \xi \ ,$$

where $\lambda = f'(x^*) > 0$, i.e. ξ is increasing exponentially, at least, within the initial time interval.

Until now we described the exponentially increasing processes only. Now let us turn to the exponentially decreasing ones. They are also numerous and various: for example, hot tea cooling, discharging an electrically charged condenser and, in particular, that of a TV electronic tube after a switch off; a decreasing of the emitting intensity when emitting is done through a partially transparent medium; a decreasing of the mass of a speed-gaining missile; a retarding of the descent upon deploying a parachute; a perishment of the population under unfavourable conditions, say, because of the sharp shortage of food; arriving at an equilibrium, and many other things.

The exponential decrease implies a continuously decelerating decrease and it is exercised so fast that the entire decelerating process will be always limited. It is the very finiteness of time that was the very reason of why quick-legged Achill failed to overtake a turtle, because he was running first a meter, then a half meter, then a quarter meter, and so on, and not more than two meters in all. Achill was not overtaking the turtle but, nevertheless, was coming closer to it very quickly, in the sense that the distance to the turtle was shortened two times at first, then again two times, and in total as much as four times, and then again two times, and in total as much as eight times, etc. From the said it follows that the exponential decrease may be interpreted as a continuous arrival at the final value.

We are passing now to specific examples. Figure 8.4 shows a circuit for
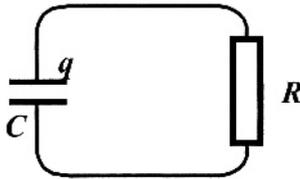
Fig. 8.4. The electric circuit with a resistor and a capacitor.

the condenser of capacitance C discharged by the resistance R. In accordance with the Kirchhoff law, the differential equation for the condenser discharge is written in the form

$$R\dot{q} = -\frac{q}{C}$$

or

$$\frac{dq}{dt} = -\frac{1}{RC}q \ .$$

Therefore, it is described by the exponentially decreasing process of the time being twice less and equal to

$$\tau = RC \ln 2 \ .$$

The time τ is proportional to the capacitance C and the resistance R. With the initial charge $q = q_0$, we have

$$q(t) = q_0 e^{-t/RC} \ .$$

The next example pertains to a parachute deceleration. Let us imagine that upon arriving at the descending velocity v_0 the parachute was deployed and its descent is retarded proportionally to its velocity. In accordance with the Newton law, we obtain

$$m\dot{v} = -mg - hv \ .$$

The solution of this equation is as follows

$$v = \frac{mg}{h} + (v_0 - \frac{mg}{h})e^{\frac{-h}{m}t} \ .$$

Therefore, the initial velocity $v_0 (v_0 > 0)$ is exponentially slowed down to the constant descending velocity equal to $\frac{mg}{h}$. This process is demonstrated in figure 8.5.

Up to now, all the exponential processes discussed were of time nature, i.e. they varied exponentially in time. Now we will look at the processes which are also exponential but not variable in time.

At present, only few are fond of reading Jules Verne's science-fiction novels. Though, there was time when his readers were absorbed in his books. In his fantastic novels he foresaw a lot, though at times was mistaken. One of his vivid errors was the described flight to the Moon on a cannon-ball. The flight was predicted by him, but on a cannon-ball – and it was an error. For Jules Verne it seemed that the bigger the cannon, the faster it will eject a cannon-ball. But it is not absolutely so. There exists a limiting velocity of ejection and it cannot be exceeded.

Let powder (as was at Jules Verne) or some other explosive be used. Imagine that its mass m, when exploded, will discharge the energy E_{chem} per a unit of mass. This energy can transfer to this mass m the velocity v (without a nucleus mass taken into consideration). In the best case we shall obtain

$$\frac{mv^2}{2} = mE_{chem} \quad ,$$

i.e. the velocity of the ejected ball cannot exceed

$$v_{max} = \sqrt{2E_{chem}} \quad .$$

This is an absolute theoretical limit for the velocity obtained when firing a cannon. This limit is greatly less than the first cosmic velocity (8 km/sec).
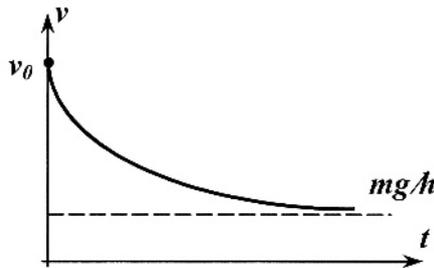
Fig. 8.5. The time plot of the falling body when viscous resistance is available.

K. E. Tsiolkovsky was the first to understand that the same chemical substance should be used in some other way and that this way may provide theoretically any arbitrary large velocity. This way is the application of a missile jet engine. The jet engine gathers traction by ejecting a gas jet in the direction being reverse to that of the missile movement. Let the missile of the mass m have the velocity v and the ejected mass move with the velocity c (figure 8.6). The mass being ejected (dm) gathers the velocity c at the expense of the chemical energy produced by the burnt propellant of the missile. Here, the entire chemical energy is assumed to be converted to the kinetic energy. Generally speaking, this is not true. Though, for our final conclusion, the fact that a portion of the energy is transferred to heat does not weigh much. Thus, we obtain

$$\frac{dmc^2}{2} = dE_{chem} \quad , \tag{8.8}$$

where $d\,E_{chem}$ is the loss of the chemical energy reserved in the missile and is converted into the kinetic energy of the gas outflow. That the missile, while ejecting from itself in the reverse direction the mass dm, gathers a forward acceleration may be revealed and the value of this velocity gain may be calculated. This calculation may be done with use of the law of energy conservation, i.e. through equalizing the energies prior to and after the mass ejection.
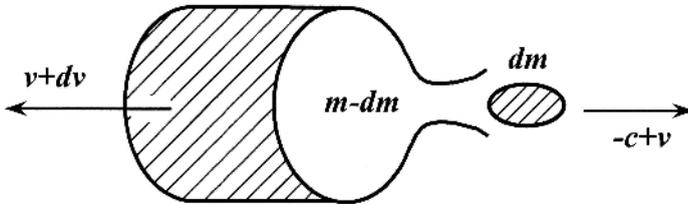
Fig. 8.6. The scheme explaining how to derive a missile speed-up equation when gas is escaping from the missile nozzle.

Prior to the mass ejection, the missile possessed the energy equal to

$$E_{chem} + \frac{mv^2}{2} \ .$$

This energy included the chemical energy E_{chem} of the missile propellant and kinetic energy of the missile movement. Upon the mass ejection, the general energy will include the reduced chemical energy, the changed kinetic energy of the missile and the kinetic energy of the ejected gas mass dm. In accordance with this, this energy will be equal to

$$E_{chem} + dE_{chem} + \frac{(m+dm)(v+dv)^2}{2} - \frac{dm(v-c)^2}{2} \ .$$

Equalizing the pre-ejected and post-ejected general energies, we will find that

$$E_{chem} + dE_{chem} + \frac{(m+dm)(v+dv)^2}{2} - \frac{dm(v-c)^2}{2} =$$

$$= E_{chem} + \frac{mv^2}{2} \ .$$

From this, taking into account (8.8) and naturally omitting the terms with $dmdv$, $(dv)^2$ of the higher order of smallness, we shall come to the below simple relation

$$mvdv + cvdm = 0$$

or

$$\frac{dm}{dv} = -\frac{1}{c}m \quad . \tag{8.9}$$

The mass m of the missile will, thus, decrease in the way similar to the exponential function of the velocity v

$$m = m_0 e^{-v/c} \tag{8.10}$$

As we want to find out how the velocity will increase with a decrease of the mass, since it is ejected backward, it is natural to write the last ratio with use of the Tsiolkovsky formula

$$v = c \ln \frac{m_0}{m} \quad . \tag{8.11}$$

With the velocity v expressed as the ratio of the initial mass m_0 to the mass m, we obtain that the missile, via ejecting a greater and a greater portion of its mass, will get a possibility to gain higher and higher velocities. Besides, with each lessening of the mass two times the missile's velocity will be increased by the magnitude $c \ln 2$. In principle, this regularity enables to gain any arbitrary large velocities, but, since the remaining mass cannot be very small, the initial mass m_0 accordingly should be increased significantly. Hence, for travelling to the Moon, no big cannon is needed (as to Jules Verne), but a big missile. Not simply big, but an enormous one. From the ratio (8.10), it follows that the initial mass m_0 is the exponentially increasing function of the velocity v

$$m_0 = m e^{v/c} \quad .$$

Here, the initial mass (with the final mass given) will be doubled at each increase of the velocity by the magnitude $c \ln 2$. Hence, in order to decrease the initial weight of the missile, its final weight has to be decreased maximally. The final weight is substantially decreased, if a multistage design of the missile is used. During the flight, the utilized stages will be dropped down to have only the minimally needed stage remained.

Another example. The dynamics of the amount of the population N, i.e. the value of the derivative $\dot{N}$, is determined by the correlation of the birth rate B and the death rate D. Thus, we have

$$\dot{N} = B - D \ ,$$

where B and D are functions of the quantity N and the external medium. With the external conditions remaining unchanged and with natural simplest dependence upon N ($B = aN$, $D = bN$), we shall come to the processes of the exponential reproduction or destruction, due to the sign of the magnitude $\lambda = a - b$.

Analogous to the population destruction is the radioactive decay resulting in an exponential decrease of the indecomposable substance. The rate of the radioactive decay is determined by the time of the half-decay. The less this time, the more quickly vanishes the radioactive radiation being dangerous for humans.

How a partially transparent medium absorbs a radiation may be represented as a destructive process only. Here, it is convenient to simulate the radiation as a stream of the large number N of photons; the medium may be assumed to possess some independent absorbing probabilities for each photon of the stream per a unit of the covered distance. Using this interpretation for very large N, we obtain

$$\frac{dN}{dt} = -\lambda N \quad .$$

This will bring us to the exponentially decreasing process, where a two-fold decrease of the radiation intensity will occur for the time $\lambda^{-1} \ln 2$.

It is through the exponential decrease of the radiation intensity that the pitch-darkness in the oceanic depths is explained, in spite of the upper layers of the ocean being illuminated sufficiently. Indeed, let the illumination drop as much as two times per each 20 meters. On a fair and sunny day such a drop is not noticed absolutely. Though at the depth of one kilometre, there will be darkness, since the radiation intensity will be $2^{50} \approx 10^{12}$ times less.

Such a model of the stream of particles can lead us to the exponentially increasing avalanching process, if a single high-velocity particle runs into a gas molecule and this collision gives birth to more than one particles of the same velocities. Similar avalanching processes are observed, when cosmic particles are coming into our atmosphere at the velocities being close to the light velocity. These particles carry enormous energy.

Cooling a not so much heated body, say, a glass of tea or coffee, also makes up the exponentially decreasing process. Indeed, if the temperature of the environmental medium is taken as an origin, then the temperature T of the heated body will be described by the differential equation

$$\frac{dT}{dt} = -\lambda T \quad ,$$

because the heated body will return its heat into the medium approximately proportionally to the difference of the temperatures in the body and the medium.

At last, let us note that all the processes of arriving at the equilibrium state (and it is only in this manner that one must interpret a discharging of the charged condenser, a retarding of the moving body, death of the population, and cooling the heated body) will possess the exponentially decreasing nature. This happens so, because the movement equation, being linearized within the neighbourhood of the equilibrium state, will, in the general case, coincide with the differential equation for the exponentially decreasing process.

Therefore, for the one-dimensional process

$$\frac{dx}{dt} = f(x) \quad ,$$

with the equilibrium state $x*$, the deviation ξ from $x*$ will be described by the equation

$$\frac{d\xi}{dt} \approx f'(x*)\xi \quad ,$$

where, due to the stability, $f'(x*) < 0$.

For the multidimensional case, the picture will be more complicated. The decreasing process will be of the form

$$\Sigma C_S e^{-\lambda_S t} + \Sigma D_k e^{-\mu_k t} \cos(\omega_k t + \varphi_k) \quad ,$$

where λ_S and μ_k are positive, i.e. this process is a superposition of the exponentially decreasing and the exponentially decreasing oscillating components. It is clear that a final arrival at the equilibrium will be described by the term of the smallest value λ_S or μ_k. In this sense, it may be also interpreted as the exponentially decreasing or the oscillating exponentially decreasing process.

9 Dynamics in coexistence of populations

The mathematical models for the coexistences of the types "predator – prey", competition (opposition) and symbiosis. The phase and bifurcation portraits.

The above cases touched on a limitless reproduction of the population. The reproduction was performed exponentially. Though, in reality, the population has various relationships with other populations and inside itself. These relationships may be of different types: "predator – prey" (antagonism), mutual assistance (symbiosis) and internal competition. In the simplest case, there exist only the interaction of two populations. This type of interaction is the most decisive and essential, whereas all the remaining types of interaction are secondary.

The first ecological model for interacting populations is the predator-prey model constructed by Volterra and Lotck. This antagonistic model assumes that the population of preys can exist independently and the population of predators cannot; the latter exists only through eating the preys. Thus, with the predator not available, the prey will be exponentially reproduced under the equation

$$\dot{x} = ax,$$

whereas with the prey not available, the predator will die out. Thus, we have

$$\dot{y} = -cy \quad .$$

The predator eats the more preys, the bigger it is and the more numerous the predators are. Hence, with the predator available, we obtain

$$\dot{x} = ax - bxy \quad . \tag{9.1}$$

The quantity of the preys eaten will favour a reproduction of the predator, and, therefore, we obtain

$$\dot{y} = -cy + dxy \quad . \tag{9.2}$$

The differential equations (9.1) and (9.2) are the very famous and widely known differential equations by Volterra and Lotck. These are the very equations that have given birth to the theory of interacting populations.

As the state of the dynamic system described by the equations (9.1) and (9.2) there will serve the quantities x (preys) and y (predators). As the phase space there will be assumed the octant $x \geq 0$, $y \geq 0$. For $y = 0$, the quantity x will exponentially increase, whereas for $x = 0$ the quantity y will exponentially decrease. The point $x = y = 0$ will be the unstable equilibrium state of the saddle type. This follows from the fact that in the neighborhood of the equilibrium point $x = y = 0$ the terms, containing the product of small magnitudes x and y, may be neglected. Hence, we approximately obtain

$$\dot{x} = ax$$

$$\dot{y} = -cy \ .$$

Except the equilibrium $x = y = 0$, there is also possible the equilibrium $x = x*$, $y = y*$, where $x*$ and $y*$ is a non-zero solution for the equations

$$ax - bxy = 0 \ , \qquad -cy + dxy = 0 \ ,$$

i.e.

$$x* = c/d \ , \qquad y* = a/b \ .$$

The type of equilibrium $(x*, y*)$ may be found, as usually, through constructing its characteristic equation

$$\begin{vmatrix} a - by* - \lambda & -bx* \\ dy* & -c + dx* - \lambda \end{vmatrix} = 0$$

or

$$\lambda^2 + \frac{ac}{bd} = 0 \ .$$

The roots of this characteristic equation are purely imaginary, $\pm i\sqrt{\dfrac{ac}{bd}}$.
In the linear approximation, they will correspond to the center-type equilibrium. This equilibrium, if the terms dropped in the linearization are now considered, can become both a stable and unstable focus. In the given case, a complete analysis reveals that the nonlinear terms will save the type of the equilibrium, i.e. the equilibrium will retain to be a center-type equilibrium. A more specific analysis becomes possible, because the Volterra-Lotck differential equations can be integrated. The Volterra-Lotck equation may be written as

$$\frac{dx}{ax - bxy} = -\frac{dy}{-cy + dxy} \ .$$

From here, separating the variables x and y, we obtain

$$\frac{(-c + dx)dx}{x} = \frac{(a - by)dy}{y} \ ;$$

upon integration we have

$$-a \ln y - c \ln x + dx + by = h \ ,$$

where h is an arbitrary constant. The last equation is the equation of the phase trajectories. Each trajectory will be assigned its own value of the integrating constant h. To qualitatively construct the phase trajectories on the plane x, y will be possible, if each phase trajectory is assumed to be constructed by projecting onto the plane x, y the line

$$z = dx + by - c \ln x - a \ln y \equiv f(x, y)$$

of intersecting the surface by the plane $z = h$ (Fig. 9.1).

The shape of this surface shown in figure 9.1 may be represented through the following conclusion. On the axes x and y, we get $f(x, y) = \infty$ and $f(x, y) \to \infty$, with at least one of the variables, x or y, being increased. To the minimum point of this surface $z = f(x, y)$ there will correspond the equilibrium state (x^*, y^*). All the rest phase trajectories will be the oval curves contained in each other and encircling the equilibrium state (x^*, y^*). Hence, the phase portrait obtained will be of the shape depicted in figure 9.2.
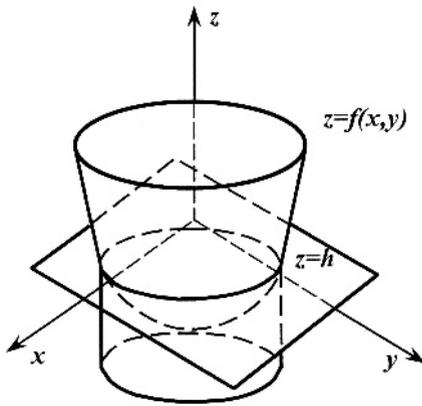
Fig. 9.1. The geometrically interpreted construction of the phase trajectories in the Volterra-Lotck model.

In spite of its simplicity, the Volterra-Lotck model reflects in a qualitatively correct way the oscillating nature of the quantities x and y in the predator-prey coexistence. This oscillating nature was exposed, for example, during the long-time observations of hares and bobcats in Canada and in many other cases.
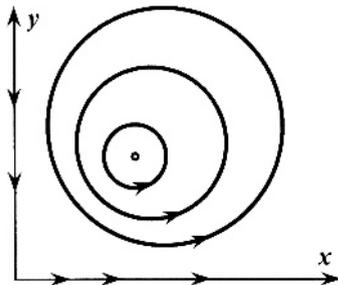
Fig. 9.2. The phase portrait of the Volterra-Lotck model.

According to the phase portrait, the fluctuations of population quantities will be periodic, but they may occur with various swings – from the zero swings to the infinite ones. Though, as was noticed by J. Volterra, the time-average values of

the quantities x and y usually retain fixed, equal to the coordinates x^* and y^* of the equilibrium state.

This immediately follows from averaging over the period of the ratios

$$\frac{\dot{x}}{x} = a - by , \qquad\qquad \frac{\dot{y}}{y} = -c + dx ,$$

since average values of their left-hand parts are equal to zero.

According to this model, neither the variations of the parameters a, b, c and d nor the disturbances, causing a variation in initial conditions, can damage the unlucky fate of the preys – to become a food for the predators – until y turns to zero. The most favourable thing for the preys is seemingly to remain in the neighborhood of the equilibrium state, because at large swings the prey population (and, by the way, the predator population as well) will arrive at the periods of very small quantities.

When applied to an existence of humans (a prey) with a microbe or a virus (a predator) (this application admissible with large qualifications only), this model makes it possible to adopt some curious conclusions on how the predator-killing medicines must be administered and how must not. First of all, a predator, if possible, has to be killed fully, i.e. if left even in small quantities it will be shortly reproduced again, and possibly, in greater quantities. And so we shall fail to achieve out target. If a full destruction of a predator is impossible, it is expedient then to reduce its quantity to some extent in the top position of the phase point (above the equilibrium) only, i.e. near its maximum quantity. A partial destruction in the time of a small-size population will only lead us to a further great worsening of the decease.

These are the conclusions possible to be adopted from the very simple Volterra-Lotck "predator-prey" model. The model was further specified and made more complicated. From the mathematical point of view the Volterra-Lotck model is not satisfactory because a phase portrait may undergo substantial qualitative changes in response to indefinitely small changes in the right-hand sides of its differential equations. It means that even insignificant factors can result in significant changes of the portrait. Thus there exists no basis to consider this model correct since ignoring secondary factors is inevitable.

Such an undesirable property of the Volterra-Lotck model is accounted for the fact that its equilibrium is a centre-type equilibrium to which there correspond purely imaginary roots of the characteristic equation; at the very insignificant additions $\delta(x,y)$ and $\varepsilon(x,y)$ to the right-hand sides of the equations (9.1), (9.2) the equilibrium may be replaced by a stable or unstable focus, and around it there may arise closed stable periodic motions, which will essentially change the phase portrait and accordingly the behaviour of the coexisting predator-prey population following from it.

With this drawback eliminated and if a predator's satiation and prey's struggle for living sources taken into account, the model will then assume the form

$$\dot{x} = ax - \frac{bxy}{1 + Ax} - Bx^2$$

(9.4)

$$\dot{y} = -cy + \frac{dxy}{1 + Ax} \quad .$$

For A=0 and $B = 0$ the model is reduced to the Volterra-Lotck model. The parameter A takes into account a predator's satiation, i.e. that even at very large quantity x of preys a predator cannot eat its preys more than some certain number of them; whereas for $A = 0$ it is able to eat up infinitely many for $x \to \infty$. The parameter $B > 0$ takes into account a limited nature of living resources for prey reproduction.

In contrast to the Volterra-Lotck model, to investigate the differential equations (9.4) is already not an easy job. This investigation is better to be started from eliminating some parameters. Now, they are six. It turns out so that they may be reduced to two. Replacing the variables $t = \frac{\tau}{\alpha}$, $x = \frac{c}{d}u$, $y = \frac{a}{b}v$ will lead the differential equations (9.4) to the form

$$\dot{u} = u - \frac{uv}{1 + \alpha u} - \varepsilon u^2$$

(9.5)

$$\dot{v} = -\gamma v(1 - \frac{u}{1 + \alpha u}) \quad .$$

A study has revealed a qualitative shape of the model phase portrait to be dependent only upon the positive parameters α and ε; the parameter γ produces no effects upon the portrait ($\alpha = A\frac{c}{d}$, $\varepsilon = \frac{cB}{ad}$).

The search for a two-dimensional phase portrait (the variables $u \geq 0$, $v \geq 0$) of these equations and for a bifurcation portrait includes some stages. First, subject to search and investigation are equilibrium states. This stage prompts a possibility of auto-oscillations. Further, there should be found the parameter domains
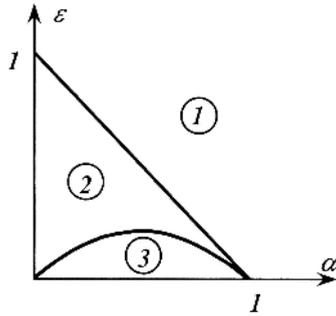
Fig. 9.3. The parametric portrait for the improved "predator-prey" model.

corresponding to various qualitative structures of the phase portrait. As a result of this stage, we arrive at the bifurcation portrait given in figure 9.3.

It consists of three domains, 1, 2 and 3. For each of them there is its shape of the phase portrait shown in figure 9.4; here, the number stands for the number of the domain of the bifurcation portrait.

At the parameters taken from the domain 1, the predator will perish; the prey will retain only in the quantity $u^* = \dfrac{1}{\varepsilon}$ corresponding to a stable equilibrium state of the phase portrait 1 (Fig. 9.4).

With the parameters lying within the domain 2, the prey's isolated existence will become unstable and there will appear some equilibrium interaction between the predator and the prey.

Upon coming into the domain 3 the predator-prey balanced coexistence is broken and it obtains an auto-oscillating nature.

Now, upon our brief description, we are coming to describing our investigation. The equilibrium states are found from the equations

$$ u - \frac{uv}{1+\alpha u} - \varepsilon u^2 = 0 \ , \qquad \gamma v (1 - \frac{u}{1+\alpha u}) = 0 \eqno(9.6) $$

Here, there are 3 solutions:

1) $u = v = 0$; 2) $u = 1/\varepsilon$, $v = 0$; and 3) $u = \dfrac{1}{1-\alpha}$, $v = \dfrac{1-\alpha-\varepsilon}{(1-\alpha)^2}$.

The first two equilibrium states occur for any positive α and ε , and the last one only in the domain 2 and 3, where $1-\alpha-\varepsilon > 0$. On the boundary of the domains 1 and 2 the third and the second equilibrium states are merged.

In order to study the stability of the revealed equilibrium states let us write down the characteristic equation

$$X(\lambda) = \begin{vmatrix} 1 - 2\xi u - \dfrac{v}{1+\alpha u} + \dfrac{\alpha uv}{(1+\alpha u)^2} - \lambda & -\dfrac{u}{1+\alpha u} \\[2em] -\gamma v\left[\dfrac{\alpha u}{(1+\alpha u)^2} - \dfrac{1}{1+\alpha u}\right] & -\gamma(1 - \dfrac{u}{1+\alpha u}) - \lambda \end{vmatrix} = 0 \ ,$$

in which u and v will be replaced with coordinates of the equilibriums being
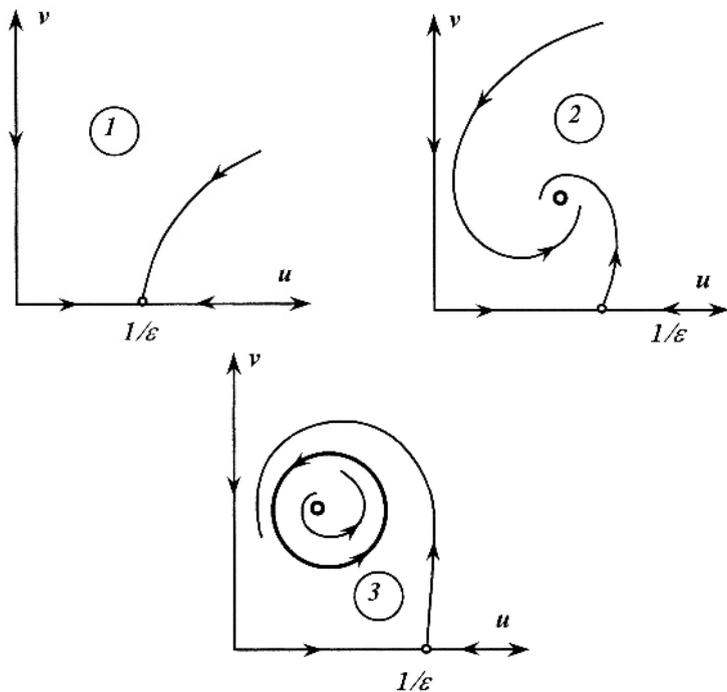
Fig. 9.4. Phase portraits of the improved "predator -prey" model for the parameters from different domains 1, 2 and 3 of the parametric portrait (Fig. 9.3).

checked for the type and stability. For the equilibrium states 1, 2 and 3 we sequentially find the following roots of this characteristic equation:

1. $\lambda_1 = 1, \lambda_2 = -\gamma$.

2. $\lambda_1 = -1$, $\lambda_2 = \dfrac{-\gamma(\alpha + \varepsilon - 1)}{\alpha + \varepsilon}$.

3. $\lambda_{1,2} = -\sigma \pm i\omega$.

where σ turns to zero on the boundary curve $\varepsilon = \dfrac{\alpha(1-\alpha)}{1+\alpha}$ of the domains 2 and 3 (in the domain 2 it is negative, and in the domain 3 positive). By virtue of the bifurcation theory, during the transition from the domain 2 into the domain 3 there should be born from the third equilibrium state a stable periodic motion or the unstable periodic motion should vanish by getting stuck at it. To distinguish which of the cases occurs is possible through calculating a sign of the so-called Lyapunov magnitude. This process gives birth to a stable periodic motion (i.e. an auto-oscillation) and accordingly there occurs the phase portrait 3 in figure 9.4 .

Justice demands to say that a strict realization of this approach needs also a determination of whether there exists a possibility for complicated limit cycles to appear. One can easily make sure of it by looking at them on a computer display; all the above given pictures of the phase portraits were taken from the display.

We have thus analysed the mathematical models of interacting predator-prey populations. Similarly, there can be constructed the models for two competing populations and two cooperating populations. The two competing populations, with internal competition within each of them also accounted, may be described by the following form of differential equations

$$\dot{x}_1 = a_1 x_1 - a_{11} x_1^{\,2} - a_{12} x_1 x_2$$

$$\dot{x}_2 = a_2 x_2 - a_{22} x_2^{\,2} - a_{21} x_2 x_1$$

or , upon replacing the variables $t = \dfrac{\tau}{a_1}$, $x_1 = \dfrac{a_1}{a_{11}} u_1$, $x_2 = \dfrac{a_2}{a_{22}} u \ u_2$, by the below equivalent form

$$\dot{u}_1 = u_1(1 - u_1 - \varepsilon_1 u_2)$$

(9.7)

$$\dot{u}_2 = \gamma u_2(1 - u_2 - \varepsilon_2 u_1) .$$
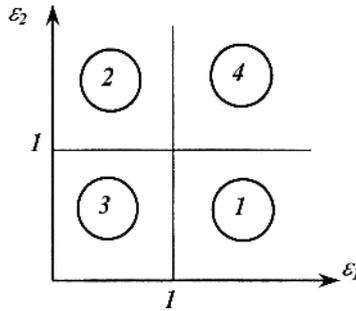
Fig. 9.5. The parametric portrait for the model of the competing species.

Here, as well as in the previous model reduced to the same form, the depend ence upon γ is not substantial, not affecting the quality of a phase portrait.
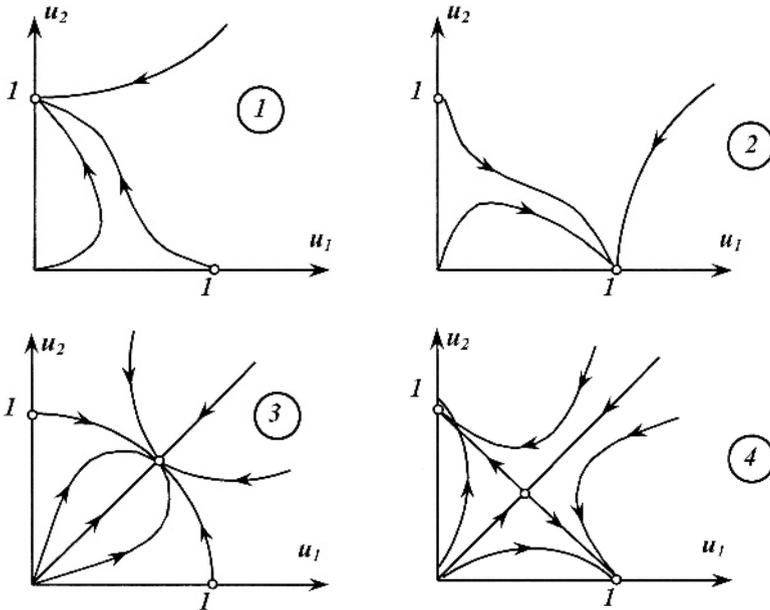
Fig. 9.6. Phase portraits for the model of the competing species corresponding to the domains in the parametric portrait (Fig. 9.5).

If the substantial parameters ε_1 and ε_2 are used, then the bifurcation diagram will contain 4 domains (1, 2, 3 and 4 in Fig. 9.5), to each of which there will correspond its own shape of the phase portrait. These phase portraits are shown in figure 9.6 under the same numbers which carry the associated domains of the bifurcation portrait. The general idea here is that among the competing populations there will survive only such a population that produces greater pressing effects upon other populations, which is counted in the parameters ε_1 and ε_2 (Fig. 9.6).

This situation occurs in the domains 1 and 2, where one of parameters ε_1 and ε_2 is more than unity and the other less. In the domain 3 there occurs an equilibrium coexistence. In the domain 4 there survives one of the populations, depending upon the initial conditions.

In order to think these conclusions over, it will be not bad to have some knowledge concerning the relationships between the initial parameters of the model a_1, a_2, a_{11}, a_{12}, a_{21}, a_{22} and the reduced ε_1 and ε_2

$$\varepsilon_1 = \frac{a_2 a_{12}}{a_1 a_{22}}, \qquad \varepsilon_2 = \frac{a_1 a_{21}}{a_2 a_{11}}.$$

The following two models describe the cooperation between populations known in biology as symbiosis. The first model covers the case when mutual assistance constitutes a necessary condition of existence since when living independently each population will perish. This model also describes an internal competition within each population. With its initial parameters, this model has the form

$$\dot{x}_1 = -a_1 x_1 + \frac{b x_1 x_2}{1 + A_1 x_2} - c_1 x_1^{\,2}$$

(9.8)

$$\dot{x}_2 = -a_2 x_2 + \frac{b x_1 x_2}{1 + A_2 x_1} - c_2 x_2^{\,2}.$$

As earlier all the parameters here are not negative. Here, only two shapes of phase portraits are possible: the one when for any initial conditions the quantities of both populations x_1 and x will tend to zero, and, thus, the populations will die out; and the other shape of a phase portrait is represented by figure 9.7. In the shaded part of the phase space, all the phase trajectories will tend to the coordinate origin and this will mean death for both populations. At large quantities x_1 and x_2 of

the populations, relating to the non-shaded part, there will be established a balanced mutually beneficial coexistence. This equilibrium state is denoted in the
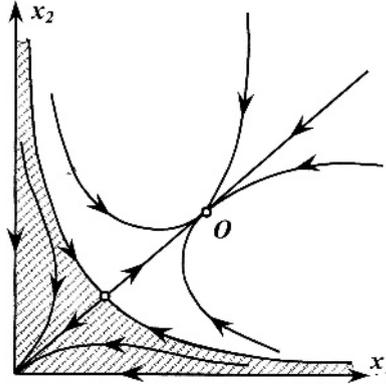
Fig. 9.7. The phase portrait for the model (9.8) of the two-species symbiosis.

phase portrait by the letter O. A substantial variety of dynamics of symbiosis of the two populations is observed in the model, where each population can exist independently, though its quantity must overcome a certain threshold – since a small-size population will die out. Such specificity in the biological behaviour of the isolated population may be represented by the model of the form

$$\dot{x} = -ax(x - L_1)(x - L_2) \quad (0 < L_1 < L_2)$$

(Fig. 8). Figure 8 shows the plot of the right-hand side of this equation and the
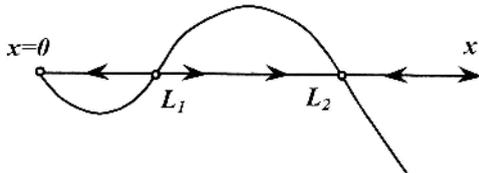
Fig. 9.8. The one-dimensional phase portrait for the one-species existence (the equation (9.9) for $b_1 = 0$).

associated phase portrait.

In accordance with this, the mathematical model for the two-population symbiosis will be written as follows

$$\dot{x}_1 = a_1 x_1 (x_1 - L_1)(L_2 - x_1) + b_1 x_1 x_2$$

(9.9)

$$\dot{x}_2 = a_2 x_2 (x_2 - K_1)(K_2 - x_2) + b_2 x_1 x_2 \ .$$

Here, one of possible phase portraits (it, as well as others, may be displayed on a computer) is depicted in figure 9.9.
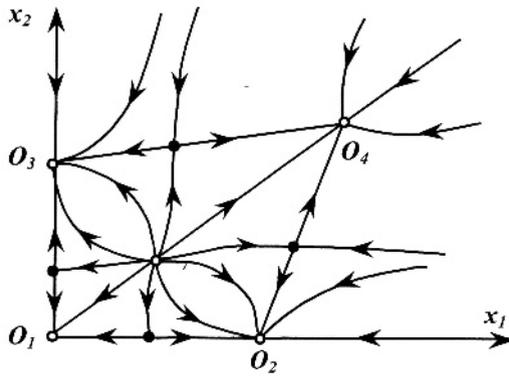
Fig. 9.9. The two-species symbiosis phase portraits for the model (9.9).

In the symbiosis described, in accordance with the behaviour of the phase trajectories, both populations can die out, or any of the populations can survive and the other can perish; or, at last, both can enjoy coexistence at the quantities larger than each separately.

Another possible case is represented in figure 9.10. In it, as well as in the previous cooperative model, possible only is the coexistence of both populations or their mutual destruction.

We have thus considered several specific mathematical models simulating a coexistence of two populations. The ideas laid in their foundations are natural and simple but the conclusions made are sudden and instructive.

Could we expect that the predator – prey coexistence will lead to the birth of auto-oscillations? Could we imagine that a competition between the preys will favour the preys to save themselves from the predators? Could we think that in the case of the coexisting competing populations the same internal competition (that makes life more difficult) will favour a survival?

Less sudden, though also interesting is our conclusion on the fact that in the case of the symbiosis the populations will survive only at sufficiently large quantities.
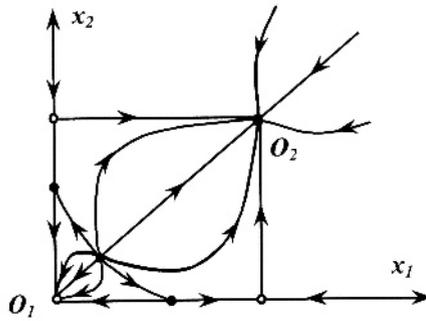
Fig. 9.10. The phase portraits of the two-species symbiosis for the model (9.9).

10 Flow biological reactor

The simplest model of a biological reactor (a chemostat); a phase portrait; optimization.

A flow biological reactor, a chemostat, represents itself a vessel filled with some nutrient medium where microbes are reproduced. Its specificity lies in the fact that its medium where this unceasing reproduction occurs is continuously replenished with a fresh nutrient solution and simultaneously at the same rate some fluid with microorganisms and substrata is discharged from the chemostat. A chemostat medium is intensively stirred up so as to make similar the concentration of the nutrient substrata and microbes in the vessel. Replenishing and picking up is performed at similar volume velocities. The chemostat has constant temperature sustained.

Our objective is constructing a mathematical model for the dynamics of this chemostat, then studying this model and optimizing the chemostat operation with respect to its controllable parameters.

Reproducing microbes in some nutrient medium of the chemostat looks like a reproduction of a predator with due account of its satiation. In accordance with it, the reproduction rate will be assumed to be proportional to the mass of microbes x with the proportion coefficient of the below form

$$\mu = \frac{\mu_m S}{K + S} \quad , \tag{10.1}$$

where S is a concentration of substrata in the chemostat, and μ_m and K are parameters. Let the nutrient substrata supplied into the chemostat be of the concentration S_0 and the incoming volume rate be equal to D. It is clear that the subsidence of the nutrient substance is proportional to the velocity of microbe reproduction. The nutrient substance is delivered and the chemostat is discharged at the constant rate D. According to the above said and making use of (10.1), we find that

$$\dot{x} = \frac{\mu_m S}{K + S} x - \frac{D}{V} x$$

$$VS\dot{} = D(S_0 - S) - \frac{l\mu_m S}{K + S} x \ ,$$

where V is a fluid volume in the chemostat. Now, let us introduce new parameters – D instead of D/V and l instead of l/V, write the above equations as

$$\dot{x} = \frac{\mu_m S}{K + S} x - Dx$$

(10.2)

$$\dot{S} = D(S_0 - S) - \frac{l\mu_m S}{K + S} x$$

and start studying them.

The phase space of the chemostat, more exactly, that of its mathematical model (10.2) is the first octant $x \geq 0$, $S \geq 0$ of the plane x, s. Besides the trivial equilibrium $x = 0$, $S = S_0$, there exists one more equilibrium determined from the equations

$$\frac{\mu_m S}{K + S} - D = 0 \ , \qquad D(S_0 - S) - \frac{l\mu_m S}{K + S} x = 0 \quad .$$

It is directly found that its coordinates $x*$ and S* are equal to

$$S* = \frac{KD}{\mu_m - D} \ , \quad x* = \left[(\mu_m - D)S_0 - KD\right] l^{-1} (\mu_m - D)^{-1} \qquad (10.3)$$

To these values S* and $x*$ there will correspond an equilibrium, if they are not negative, i.e. if the below inequality holds

$$D \leq \frac{\mu_m S_0}{K + S_0} < \mu_m$$

(10.4)

or, in other words, if the point (S_0, D) is located in the domain depicted in figure 10.1.

The stability of equilibriums is determined by the roots of the characteristic equation

$$\lambda(\lambda) = \begin{vmatrix} \mu - D - \lambda & \dfrac{\mu_m S}{(K+S)^2}\,x \\[4mm] -l\mu & -D - \dfrac{\mu_m Kl}{(K+S)^2}\,x - \lambda \end{vmatrix} = 0 \;, \qquad (10.5)$$

where the notation of (10.1) is used, S and x are the values of the coordinates of the equilibrium under study. For the equilibrium $x = 0$, $S = S_0$
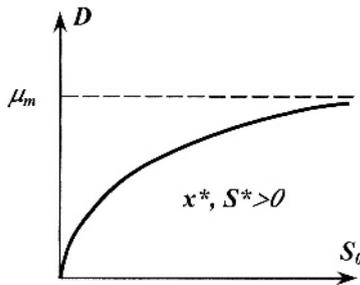
Fig. 10.1. The domain for the existence of the chemostat equilibrium (x^*, S^*) in the parameters S_0 and D.

the roots are real and of different signs; for the equilibrium $x = x^*$, $S = S^*$, determined by (10.3), the coefficients of the quadratic equation with respect to λ (10.5) are positive; hence, it is stable. The associated phase portrait is given in figure 10.2.

It is very simple: any phase point x, S ($x > 0$) asymptotically arrives at the equilibrium of the coordinates S^*, x^*. According to (10.2), the chemostat parameters are μ_m, K, D, S_0, l; besides, the values of the parameters D and S_0 may be changed, they are regulared. It is better to vary them so as not to disrupt the stable functioning of the chemostat, i.e. to obey the conditions (10.4) and naturally they should be changed so as to maximize Dx^*. From (10.3) it follows that

$$Dx^* = D\big[(\mu_m - D)S_0 - KD\big]l^{-1}(\mu_m - D)^{-1} =$$

$$(10.6)$$

$$= \frac{DS_0}{l} - \frac{KD^2}{l(\mu_m - D)} \ .$$
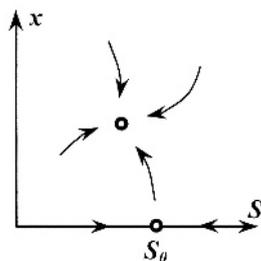
Fig. 10.2. The phase portrait of the chemostat for the parameter values of S_0 and D from the equilibrium existence domain depicted in figure 10.1.
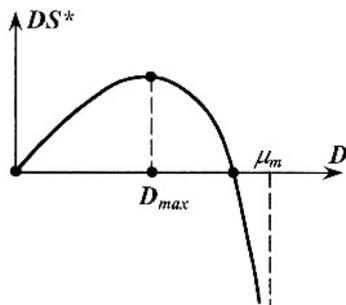
Fig. 10.3. The dependence of chemostat productivity upon the discharging rate D.

For S_0 chosen, a maximum over D is attained at some $D(S_0) < \mu_m$, $D(S_0)$ being an increasing function of S_0. That is, S_0 is better to be chosen as large as possible; at S_0 chosen there exists the optimal discharging rate D.

This very fact is explained in figure 10.3, where there is shown the plot for microbe reproduction rate $DS*$ as a function of the discharging rate D. The point where this plot intersects the abscissa axis has the coordinate $\mu_m S_0 (K + S_0)^{-1}$; therefore, after this point the plot becomes physically senseless, since the inequality (10.4) for the chemostat functioning is not obeyed.

11 Mathematical model for the immune response of a living organism to an infectious invasion

The simplified phenomenological model in the form of the fourth-order differential equation system for the immune response of the organism to an infection. Basic variants of the disease progress, its dependence upon parameters; the extent of the infection and the treating counteractions.

The animal and human immune systems are very complicated. Complicated and diverse is their structure, also complicated and multi-variant are their responses to the attack by infection. Their counteracting facilities are complicated and multisided as well. A contemporary description of the immune system occupies a pair of volumes, and a lot still remains untouched and unknown. Alongside with this, the immune system, whatever complicated it is, is responsible for counteracting and destructing a foreign infection intruded into the organism and harmful to it. In general notions, these counteractions against the infection and its harmful effect become possible to be described, with omitting details of how this actually occurs.

Thus, there exists some infection tending to be reproduced and it is attacking an organism suppressing its living viability. There exists an organism trying to destruct this infection, exerting initially its counteractions and arranging some supplementary counteraction as soon as this infection is spotted. A high-speed response of the organism and its efficiency are dependent upon its inherited features, the "life experience" acquired and also upon the state of the organism and its reserves for the moment.

So, despite all the complexity of the immune response to an infectious attack there arise the following three fundamental factors: the infection; its reproduction and its infecting of the organism; and the counteraction of the organism and its potential counteracting capabilities dependent upon its state. Quantitatively, these factors may be represented by three magnitudes – x (a quantity of the infection), y (an extent of the organism counteraction) and z (a potential of the organism). Besides its quantity x, any infection is characterized by the rate of its reproduction within the organism environment and by its suppressive action upon the organism. As for the counteraction, it is characterized by its extent y and its counteracting efficiency, fastness and the replenishing rate, which depend upon the organism and its potential z.

All the above said makes it possible to describe the dynamics of the immune response – the variation of the magnitudes x, y and z – by the following four differential equations

$$\dot{x} = \lambda x - \frac{axy}{1+ax} - \bar{\varepsilon}x^2$$

$$\dot{y} = \begin{cases} -\dfrac{\bar{b}xy}{1+ax} & for \quad x < \bar{x}_0 \\[3mm] -\dfrac{\bar{b}xy}{1+xy} + w = K & for \quad x \geq \bar{x}_0 \\[3mm] 0 & for \quad y = 0 \quad and \quad K < 0 \end{cases} \qquad (11.1)$$

$$\tau\dot{w} + w = \bar{\bar{B}}z(2z - z_0)(x + \bar{\beta}x^2)$$

$$\dot{z} = \begin{cases} \dfrac{c(z - z_0)}{1+\gamma x} - dy - e = F & for \quad z > 0 \quad or \quad z = 0 \; and \; F > 0 \\[3mm] 0 & for \quad z = 0 \quad and \quad F \leq 0 \end{cases}$$

In the equations (11.1), $\bar{x}_0$ is the threshold of the organism sensitivity against the infection; τ is the time delay of the supplementary immune response. The extent of this response will depend upon the infection quantity x and the organism potential z. For $\bar{\beta} = 0$, this dependence upon x will be linear; for $\bar{\beta} > 0$, it will be a forced one; for $\bar{\beta} < 0$, it will be, on the contrary, a reduced one. The last equation in the system (11.1) describes the variation of the potential z; here, we have $0 \leq z \leq z_0$. The first term in the last equation (11.1) will be the rate of replenishing the potential z till its extreme value z_0; the second and third terms will be the potential expenditures for the immune response y and for the needs of the organism itself. Converting z to zero is here understood as the complete organism exhaustion bringing about its destruction.

The four equations (11.1) involve 13 parameters. The parameter λ may have any sign: $\lambda < 0$ will indicate that the infection introduced into the body is perishing; $\lambda > 0$ will imply that the infection is being, nevertheless, reproduced, though this reproduction is hampered by the organism counteraction y. Though, this counteraction may be not available. Further, the case $\lambda > 0$ is possible. All the other parameters, except β, are positive. The process of the growth and the

dynamics of the counteraction y against the infection x are described by the last three equations.

The model (11.1) is rather complicated for studying; though, it may be sufficiently simplified by making use of the variables x and y only. It is the more so as the role of the rest equations is rather evident and may be mentally accounted and understood. This simplified model will be of the form

$$\dot{x} = \lambda x - \frac{axy}{1+ax} - \bar{\varepsilon}x^2$$

$$\dot{y} = \begin{cases} -\dfrac{\bar{b}xy}{1+ax} & for \quad x < \bar{x}_0 \\[2mm] -\dfrac{\bar{b}xy}{1+ax} + \overline{B}(x+\overline{\beta}x^2) = K & for \quad x \geq \bar{x}_0 \\[2mm] 0 & \\[2mm] & for \quad y = 0 \quad and \quad K < 0 \end{cases} \qquad (11.2)$$

The equations (11.2) differ from (11.1) in that that the magnitude B now becomes constant, while in the equations (11.1) it varies in accordance with the last two equations. A formal jump from the system (11.1) to (11.2) is possible by setting $\tau = 0$ and assuming z to be constant. The magnitude $\tau > 0$ describes the rate of the increase of w from zero to the extreme value $\overline{\overline{B}}z(2z-z_0)(x+\overline{\beta}x^2)$; $\overline{B}$ in the equations (11.2) is equal to $\overline{\overline{B}}z(2z-z_0)$, with z assumed constant. Thus, the role of the last two equations in (11.1) may be accounted as the variability of B in the equations (11.2).

In the simplified equations (11.2), the number of the parameters may be reduced by scaling the time t and the variables x and y. Upon this, these equations will assume the form

$$\dot{x} = x - \frac{xy}{1+x} - \varepsilon x^2$$

$$\dot{y} = \begin{cases} -\dfrac{bxy}{1+x} & for \quad x < x_0 \\[2mm] -\dfrac{bxy}{1+x} + Bx(1+\beta x) = K & for \quad x \geq x_0 \\[2mm] 0 & \\[2mm] & for \quad y = 0 \quad and \quad K < 0 \end{cases} \qquad . (11.3)$$

Let us understand here what will be obtained from the simplified model (11.3), if at first we assume $\varepsilon = \beta = 0$. In this case, there will remain only the three parameters, x_0, b and B. Here, the two significantly different cases $b > B$ and $b < B$ are possible. The second case ($b < B$), in its turn, will be splitted into other two cases, $b < 1$ and $b > 1$. As a result of this, we shall obtain the four different phase portraits depicted in figure 11.1.
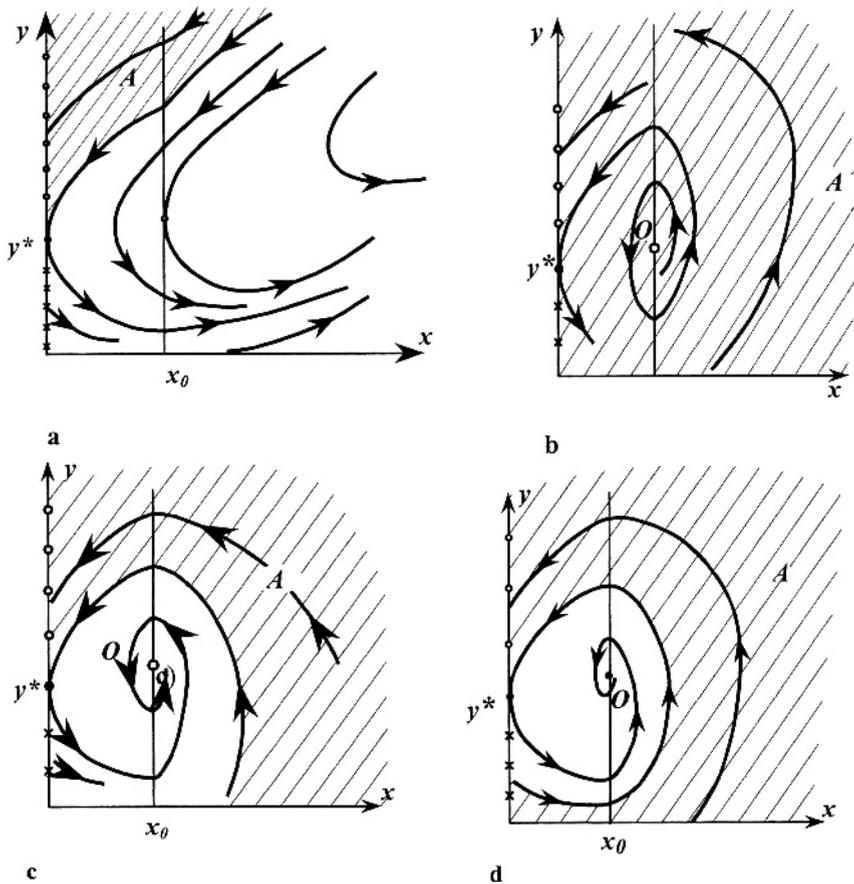
Fig. 11.1. The phase portraits of the simplified model (11.3) for $\beta = \varepsilon = 0$ and for **a** $b > B$, **b** $b < B$ and $b > 1$, **c, d** $b < B$ and $b < 1$. A is the domain of the initial conditions, in which a recovery takes place. In the cases **c** and **d**, beyond the A-domain the infection will retain at the sensitivity threshold x_0 and a bacilli-carrying will take place.

As seen from figure 11.1, in the case of $b > B$ the organism will escape from the destruction only at the initial conditions in the domain A, i.e. at its sufficiently high initial counteraction y. With the initial insufficient counteraction y, the organism will perish (i.e. x will increase infinitely and this, due to the equation (11.1), will yield the nullification of the potential z). This even will occur when the initial quantity of the infection is very insignificant. On the contrary, for $b < B$, the organism will never perish. Though, the counteraction may either fully destroy the infection (the domain A and the arrival of the phase point at one of the equilibriums $x = 0$, $y > y*$) or may retain the infection at the sensitivity threshold (the entire positive octant, except the domain A), because of the phase point having arrived at the stable equilibrium O. With the initial conditions lying in the domain A, a complete recovery of the organism will be accompanied with a single or, perhaps, some aggravations (Figs. 11.1b and c). How vividly these aggravations are expressed will depend upon the infection quantity x and upon the initial value of the immune counteraction y. A bacilli-carrying in the organism and an equilibrium coexistence with the infection occurs via an oscillatory arrival at the stable equilibrium O (Fig. 11.1c).

Figure 11.1 makes it possible to observe the consequences of the decrease of the infection quantity x (for example, with the help of antibiotics or other medicines), if it is supposed that this will exert no sufficient effect upon the extent of the immune response, i.e. upon the magnitudes b and β in the equations (11.3) and the value of y. In the case of $b > \beta$ and $y > y*$, a sufficient decrease of x will result in a faster recovery or will even save the organism.

For $y < y*$, the above recovery will be also possible but only upon complete destruction of the infection ($x = 0$). Here, this recovery is of an unstable nature – the smallest infection may cause a repeated disease. It should be also noted that in the case of $y > y*$ the infection should be completely destructed, otherwise, the disease, upon some temporal amelioration, can revive and become even more threatening than earlier (Fig. 11.1b). For the case $b < B$, the effect of such medical therapy will be small and always not obligatorily positive, since the therapy may lead to a cohabitation with the infection. Though, it will make a flow of the disease easier. Alongside with it, it should be remembered that a considerable and prolonged exacerbation (through which, due to figures 11.1b and c, a recovery may come) may not actually occur, due to the neglected decrease of the potential z that will cause a decrease of the magnitude B in the equations (11.3). We should also mind one more thing: in accordance with the complete model (11.1), we initially (i.e. at the instant of the infection) have $B = 0$; then the magnitude B reaches the value we consider. Besides, B will increase during the time interval of the τ-order, only after x having exceeded x_0.

The above given figures of the phase portraits, as a rule, demonstrate a need in the externally introduced increase of y, i.e. a need in the immune system stimulation. In general, this usually quickens the recovery and allows to eliminate a transition to the chronic infectiousness. It will be also seen from below when the role of $\beta > 0$ is exposed. Nevertheless, the most important conclusion (except the most need of providing $b < B$) is the expediency of sustaining the initial value of y greater than the value of $y*$ (see Figs. 11.1).

Indeed, if at the initial moment, prior to the infection, the immune system is in the equilibrium $x = 0$, $y \geq 0$, then for $y < y*$, due to the instability of the equilibrium, even insignificant increase of x will result in a disease, whereas for $y > y*$ the infection will be destroyed and a new similar equilibrium state will appear. The counteraction y may decrease because of the absence of the asymptotically stable equilibriums $x = 0$, $y > y*$. Conversely, an existing cohabitation with the infection will make the equilibrium state O asymptotically stable, and in this sense this will protect the organism against the second infection. Certainly, it happens so only for the parameters b and B being invariant; if these parameters vary, then this infection , conversely, may initiate the second disease.
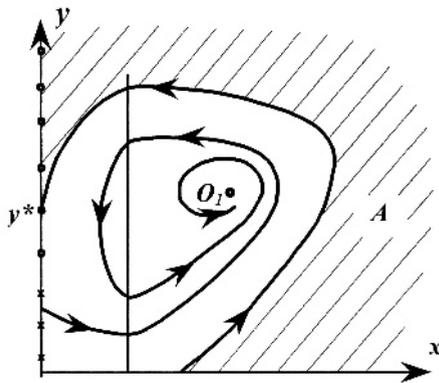
Fig. 11.2. The shape of the phase portrait for the simplified model (11.3) for $x_0 = 1.0$, $b = 1.0$, $B = 0.9$, $\beta = 0.05$ and $\varepsilon = 0.01$. The initial conditions lying in the domain A bring a recovery; if beyond this domain, then a chronic disease (the stable equilibrium state O_1).

Now, let us see what new may be brought by the parameters β and ε, earlier not taken into account ($\varepsilon = \beta = 0$). Here, we restrict ourselves to illustrating only

some phase portraits taken from the computer display. The phase portraits in figures 11.2, 11.3 and 11.4 demonstrate an importance of the forced immune response ($\beta > 0$) for the very unfavourable case, $b > B$.
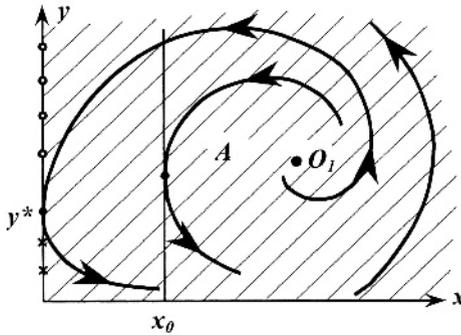
Fig. 11.3. The shape of the phase portrait for the simplified model (11.3) for $x_0 = 1.0$, $b = 0.8$, $B = 0.5$, $\beta = 0.2$ and $\varepsilon = 0$. The equilibrium state O_1 is unstable and for any initial conditions this brings a recovery.
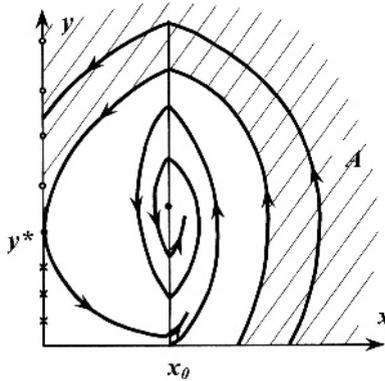
Fig. 11.4. The shape of the phase portrait for the simplified model (11.3) for $x_0 = 1.0$, $b = 0.9$, $B = 0.9$, $\beta = 0.5$ and $\varepsilon = 0.1$. The initial conditions when lying in the domain A bring a recovery; if beyond this domain, a bacilli-carrying.

Here, everything will end up with a chronic disease, a bacilli-carrying or a recovery. Of course, a long chronic disease can substantially decrease the potential

z and a resultant decrease of the magnitude B. All this may lead to fatal consequences. In figure 11.5 there is given another variant of the
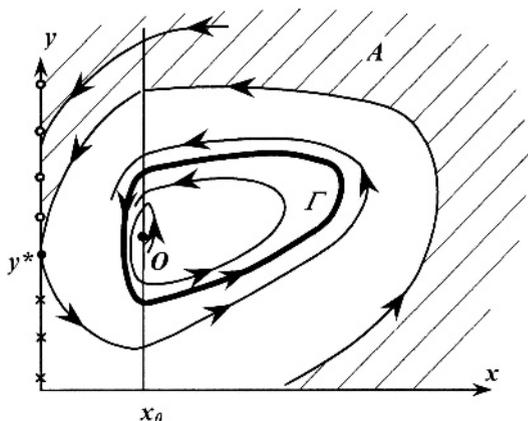
Fig. 11.5. The shape of the phase portrait for the simplified model (11.3) for $x_0 = 1.0$, $b = 0.96$, $B = 0.9$, $\beta = 0.05$ and $\varepsilon = 0.01$. The initial conditions when lying in the domain A bring a recovery; and when beyond this domain, a periodic progress of the disease (the auto-oscillation Γ).
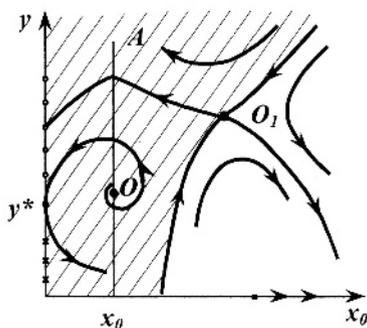
Fig. 11.6. The shape of the phase portrait for the simplified model (11.3) for $x_0 = 1.0$, $b = 0.5$, $B = 2.0$, $\beta = -0.1$ and $\varepsilon = 0$. Being within the A - domain implies a recovery; beyond it death.

immune response speed-up ($\beta > 0$), for which the disease can flow with the periodically repeated relapses (the auto-oscillation Γ). What a weakening of the immune response ($\beta < 0$) can bring is shown in figures 11.6 and 11.7; here, despite $B > b$ a lethal outcome is possible.
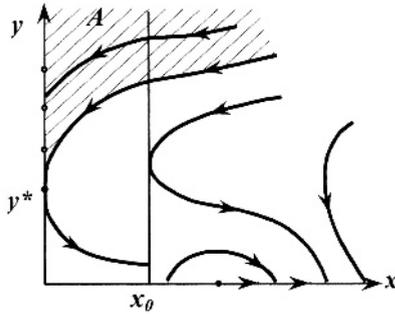
Fig. 11.7. The shape of the phase portrait for the simplified model (11.3) for $x_0 = 1.0$, $b = 0.8$, $B = 1.0$, $\beta = -0.2$ and $\varepsilon = 0$). Being within the A - domain implies a recovery;beyond it death.

The immune response may be weakened not only by $\beta < 0$ but also by a drop of the potential z , due to the last equation of the system (11.1), which was left out of account in the simplified system (11.3).

12 Mathematical model for the community "Producers – Products – Managers"

Possible kinds of phase portraits. Evolution of the community with the growth of technologies and its dependence upon other parameters.

The human society is extremely complicated, each person in it is an entire world, his behaviour is many-sided and unpredictable. The relations between people are of the very complicated nature. Besides, in the human society there exist numerous systems, such as a governmental system (a state system), a system of management, a distributing system, finance, industry, consumptive establishments, transport, communications, educational and medical establishments, etc. To describe all this mathematically seems impossible. This is just really so. However, one can try to describe, at least, some aspects of this unspeakably complicated human community, i.e. some of its peculiarities. For example, for his living man needs to get food; and to obtain it efficiently he needs management. Also, man shall not live by bread alone. He needs a shelter, a family, production tools, entertainments and pleasures. When getting his necessities of life, he starts to clash against other members of his community (society) demanding their own necessities. Nevertheless, the members of the community have to unite their efforts in order to more efficiently achieve what they want to achieve. In such a community, let us distinguish its three basic components – producers, products and managers. The producers are the immediate manufacturers of the products; the managers produce no products, though promote their production; the products imply the things needed for humans' living, i.e. what man eats and utilizes.

Let us specify these chosen aspects and denote them as the magnitudes x, y and z, i.e. as a quantity of the producers, the managers and an entire quantity of the products the community has available for use. The producers and the managers compete between themselves, and also within their own groups. Nevertheless, the producers produce their products together; and this production can be promoted by the managers. The producers are immediate feeders for the managers and themselves. Thus, there exist competitive interrelations, natural reproduction and a joint production. The efficiency of this production is determined by the level of the technologies available in the community.

Figure 12.1 shows the links and the cooperation existent between the producers x, the managers y and the products z. Here, each arrow indicates the action of the variable, from which the arrow is outcoming, upon the variable at which the arrow is arriving. For example, the arrow running from the variable x to the variable z implies that the fluctuation rate of the amount of the products z will de-

pend upon x, i.e. upon the amount of the producers. Such arrows will be only two: one will imply that the variable x produces the products z; the another similar arrow will indicate that the producers utilize these products. The arrow running from z to x will imply that the fluctuation rate of x is dependent upon the products z available. All the arrows from x, y, z to z will indicate that the fluctuation rate of the accumulated
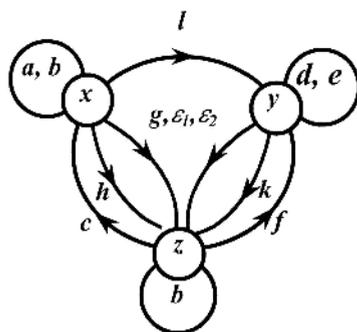
Fig. 12.1. The scheme of influences in the "producers-products-managers" model.

products z will depend upon x, y and z. The scheme discloses no specific dependences and mutual links; it exhibits only their presence and our wish to account them in our model. It is clear that to account them in some precise fashion is very difficult. With the variables x, y and z aggregated it is impossible. Here, we restrain ourselves to their rough qualitative consideration described by the following system of three differential equations

$$\dot{x} = (a - bx - ly + cz)x$$

$$\dot{y} = (-d - ey + fz)y \tag{12.1}$$

$$\dot{z} = \begin{cases} F = g\dfrac{1 + \varepsilon_1 y}{1 + \varepsilon_2 y}\dfrac{x}{1 + \delta z} - hx - ky & \begin{array}{l} for \quad z > 0 \\ or \quad z = 0 \quad and \quad F > 0 \end{array} \\ 0 & for \quad z = 0 \quad and \quad F \le 0 \end{cases}$$

The above system is the mathematical model suggested by us for simulating the community "producers – products – managers" (the PPM-community). In this model, very many things are left beyond our consideration. Now, let us see what

dynamics of the model this fact will lead us to. Let us describe again and in more details our general ideas making a basis of this model. The fluctuation rate of x is assumed to be generally proportional to x, with the proportionality coefficient $a - bx - ly + cz$ taken as decreasing with the growth of x and y and as increasing with the growth of z. Such assumptions are supposed to reflect the internal competition among the producers x, the pressure exerted upon them from the side of the managers y and the managers' promotion for the production of the accumulated products z. Without accounting these influences of the variables x, y and z upon the fluctuations of x, the first equation of the system (12.1) would have assumed the form

$$\dot{x} = ax$$

that would express an unlimited reproduction of the producers.

The similar considerations were laid into the second equation of the system (12.1); though here, the producers' pressure upon the managers is assumed to be absent or negligible. Possibly, it is usually not so but, if needed, it will not be difficult to take this aspect into account. While the producers for $z = 0$ are able to some extent for an independent existence, the independent existence of the managers is excluded. Therefore, the sign before the coefficient d is given negative; accordingly, the "isolated" managers will be described by the differential equation

$$\dot{y} = -dy$$

that implies their destruction.

The third equation simulates the accumulation of the products in the course of their production and consumption. The products are utilized by the producers and managers. The products are subject to depreciation; they are broken down themselves and produced by the joint efforts of the producers and managers.

The extent of promoting the products production by the producers and managers is different: the producers produce the products immediately, while the managers promote (or hamper) this production. Accordingly, the right-hand side of the differential equation for z involves three terms: the first term describes the product production rate with due account of such aspects as the increasing difficulties of the production, a growth of the volumes of the products, an amortization of the products (their becoming outdated); the second and the third terms represent a consumption by the producers and managers, respectively. The products consumption by the producers and managers is assumed to be proportional to x and y. As for the production, here the problem is more complicated. The production rate is assumed proportional to the number of the producers x with its proportionality coefficient. This coefficient includes, as multipliers, the community's technological level g, the magnitude $\dfrac{(1 + \varepsilon_1 y)}{(1 + \varepsilon_2 y)}$ describing the influence of the managers y upon the production, the multiplier $(1 + \delta z)^{-1}$ that decreases with

the increase of z. This multiplier is incorporated to simulate an amortization (depreciation) of the accumulated products and the difficulties brought about by the increase of the volumes of products. For $\dfrac{\varepsilon_1}{\varepsilon_2} > 1$, the managers are thought to promote the production within the interval from 1 to $\dfrac{\varepsilon_1}{\varepsilon_2}$; conversely, for $\dfrac{\varepsilon_1}{\varepsilon_2} < 1$ they will hamper it within the same limits.

In our model, the concepts of the producers, the products and the managers are aggregated, i.e. these components do not account their possible complicated and multisided differentiation. For the ancient community, under the products there were implied the first consumer products, and, possibly, a shelter. In a more developed society, these things were also augmented with the hunting tools, cattle-breeding and land-cultivating tools covering the cattle itself and the land. In the contemporary society, the products cover an entire aggregation of the material and intellectual wealth of the society. In our case, the variables x, y and z are interpreted by us in an averaged and generalized manner and are associated with the productive and public activities reflected in the productive and competitive inter-relations.

Such generalized, averaged and fuzzy definitions of the variables x, y and z will naturally bring some dissatisfaction. Much is left out not specified, not differentiated but the numerous specifications, if injected by us, may kill the model's simplicity and generality. That is why let us not be so strict. Perhaps, this model still describes something; possibly, the most significant. Let us first see what it will bring us and later on formulate our opinion upon it.

Any model is defined by a large number of parameters; in nature these parameters vary, and not slowly; at times they fluctuate very fast. Their behaviour depends in a complicated fashion upon the current and the past states of the society. This is their life in reality. Though, in the model they are assumed constant. The analysis of the model may reveal the events occurring at this or that fixed values of the parameters. It may also show how the model's behaviour varies in response to the parameter variations.

How the current values of the variables x, y and z affect the model parameters will depend upon whether the contemporaries of this model understand the laws of the dynamic development of the society they live in, what their actual thinking about this society is, what management system is used in the society, what morals, ethic, beliefs and breeding of the members of this society are. It is also clear that the nature of the dependence of the parameters may be strongly affected by the news about the community obtained as a result of the study of the mathematical model of this community. All these things are very difficult to be accounted; therefore, we restrain ourselves to a study of the model at different constant values of its parameters. All the rest will be left to our intellect.

In spite of its simplicity, to study this model is far not so easy, since it is three-dimensional and involves a lot of parameters. In our case, studying its dynamics

on a qualitative basis is reduced to creating the phase and the bifurcation portraits. The phase portrait is three-dimensional and can be drawn. As for the bifurcation portrait, things here are more difficult with it. Only common representations of this portrait can be obtained. Even a large reduction of the parameters will still leave not less than nine. Thus, we think it more comfortable to retain 13 initial parameters.

The possible types of the phase portraits and their dependence upon the number of the parameters have been mainly revealed thanks to the old-fashioned labour-consuming analytical research. Visualizing the phase trajectories on a PC display was also of some help.

We have managed to reveal the following. One of the basic parameters responsible for describing the structure of the PPM-community is its technological level g. In accordance with this level of technologies, all the communities may be divided into the three types – a low-technology community, an middle-technology community and a high-technology community. The technological level will be

low, if $g < h$; middle, if $h < g < h(1 + \delta \dfrac{d}{f})$; and high, if $g > h(1 + \delta \dfrac{d}{f})$.

The phase portraits for the low- and middle-technology communities will also depend upon the magnitudes ε_1 and ε_2 describing the managers' impact upon the productive efficiency. For $\varepsilon_1 < K(\varepsilon_2)$, where $K(\varepsilon_2)$ depends also upon the rest parameters, it will become possible to create the phase portraits presented in figures 12.2 and 12.3. The first of them represents a low-technology community, where only the producers without any accumulated products and without
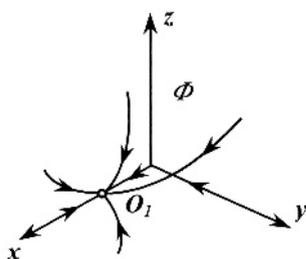
Fig. 12.2. The phase portrait of the PPM-model for $g < h$, when the globally stable equilibrium O_1 holds no products and managers.

managers have remained – all the produced has been eaten. The second portrait shows a middle-technology community where only the producers with accumu-

lated products are living. The states, at which the managers exist, vary in such a way that the managers vanish with time. As for the low-technology community, the accumulated products vanish as well.
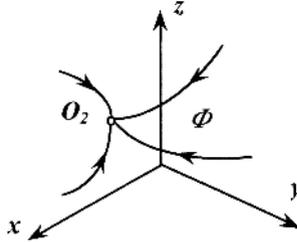
Fig. 12.3. The phase portrait of the PPM-model for $h < g < h(1 + \delta d / f)$, when the globally stable equilibrium O_2 holds no managers.

If the managers are able to substantially increase the productive efficiency so that $\varepsilon_1 > K(\varepsilon_2)$, then the phase portraits will assume the form shown in figures 12.4 and 12.5. In this case, alongside with the preservation of the previous stable communities of the types "producers" or "producers – products", there may also arise a community with the accumulated products and managers, i.e. the PPM - community.

The PPM-communities can be stable and unstable. All this will lead us to four types of the phase portraits depicted in figures 12.4, 12.4a, 12.5 and 12.5a. Here, stability takes place for $\Delta > 0$ and instability for $\Delta < 0$.

An unstable community may be thought to be arising from a corresponding stable community, as a result of the equilibrium O (related to the PPM-community) having lost its stability, because of the stable periodic motion Γ born from this equilibrium O. The periodic motion Γ may lie entirely above the plane $z = 0$ of the complete exhaustion of the accommodated products and also may contain the part lying on the plane $z = 0$.

The model for the high-technology community with the technological level g always assumes two equilibrium states, O_2 and O_4. The earlier stable equilibrium state O_2 corresponding to the "producers – products" community becomes unstable. There appears a new equilibrium state, O_4.
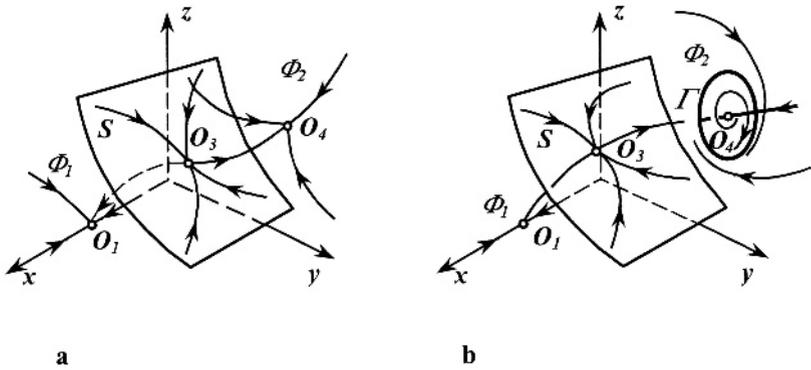
Fig. 12.4. The PPM-model phase portraits: **a** the two stable equilibrium states, O_1 and O_4 whose attracting domains Φ_1 and Φ_2 are separated by the separatrix surface S of the saddle equilibrium O_3; **b** one equilibrium, O_1, is stable and the other, O_4, having become unstable, has given a birth to the stable periodic motion, Γ; their (of O_1 and Γ) attracting domains Φ_1 and Φ_2 are separated by the separatrix surface S of the saddle equilibrium O_3.
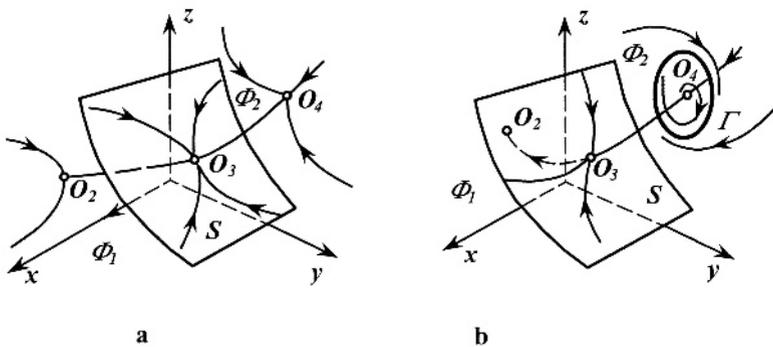
Fig. 12.5. Here, the phase portraits are similar to those shown in figure 12.4 but the stable equilibrium state O_1, holding only producers, has been replaced with the stable equilibrium state O_2, where only managers are absent.

Depending upon the sign of the magnitude Δ, this state will be stable or unstable – the first occurs for $\Delta > 0$ and the second for $\Delta < 0$. Being locally stable, this equilibrium state will be stable also globally and will describe the PPM-community. With the magnitude Δ changing its sign and with the equilibrium state having lost its stability, this equilibrium state will give birth to the stable periodic motion Γ. Its stability will also be global and this motion will correspond to the PPM-community. As earlier, this periodic motion, pertaining to the unstable auto-oscillating PPM-community, will lie above the plane $z = 0$, but also can hold some components lying on the plane $z = 0$.

   The phase portraits, constructed according to the above descriptions, are shown in figures 12.6 and 12.7.
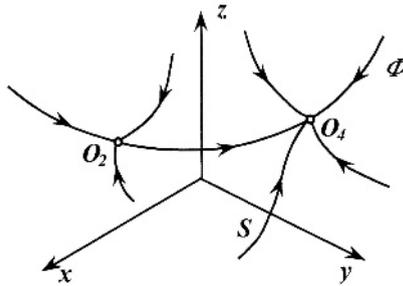
Fig. 12.6. The phase portrait of the globally stable equilibrium PPM-community.
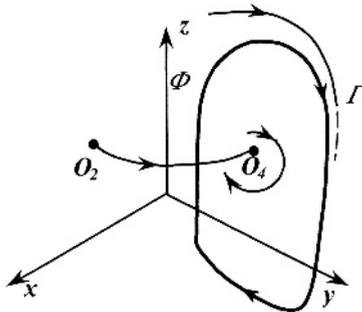
Fig. 12.7. The same as in figure 12.6 but the globally stable equilibrium community is replaced here with the auto-oscillation community.

Besides these qualitative differences, it may be noticed that a stable or an unstable equilibrium community, or rather the equilibrium $x^* \neq 0$, $y \neq 0$, $z^* \neq 0$ relative to it, may behave itself in a different way for an unlimited growth of the parameter g. The magnitudes x^* y^*, z^* either grow infinitely or remain restrained. The first takes place for $ce - lf > 0$, the second for the reverse inequality.

The above described equilibrium states and periodic motions entirely lie within the parallelepiped defined by the inequalities

$$0 \leq x \leq A , \qquad 0 \leq y \leq B , \qquad 0 \leq z \leq C ,$$

where finite A, B and C are determined by the inequalities

$$g \max(1, \varepsilon_1 / \varepsilon_2) - (1 + \delta C)h < 0$$

$$- d - eB + fC < 0$$

$$a - bA + cC < 0 ,$$

because through its sides $x = A$, $y = B$ and $z = C$ the phase trajectories are coming inside it.

Of some interest for us are the quantitative values of the coordinates x^*, y^* and z^* respondent to the stable equilibriums for the stable communities. These values are easily calculated for the "producers" and "producers – products" communities as below

$$x^* = \frac{a}{b} , \; y^* = z^* = 0 \; . \tag{12.2}$$

Accordingly,

$$x = \frac{a}{b} + \frac{c(g - h)}{\delta bh} , \qquad y^* = 0 , \qquad z^* = \frac{g - h}{\delta h} \; . \tag{12.3}$$

For the PPM-community, to find the coordinates x^*, y^*, z^* is more complicated. These coordinates are found in the following way: first we find the maximal positive root y^* of the equation

$$g \frac{1+\varepsilon_1 y}{1+\varepsilon_2 y} \frac{a_1 + a_2 y}{1+\delta b_1 + \delta b_2 y} - ha_1 - (ha_2 + k)y = 0 \quad, \tag{12.4}$$

where

$$a_1 = \frac{a}{b} + \frac{cd}{bf} \ , \quad a_2 = \frac{ce}{bf} - \frac{l}{b} \ , \quad b_1 = \frac{d}{f} \ , \quad b_2 = \frac{e}{f} \ .$$

Then, x^*, y^* are determined through the formulas:

$$x^* = a_1 + a_2 y^* \ , \qquad z^* = b_1 + b_2 y^* \ . \tag{12.5}$$

For

$$g < h(1 + \frac{\delta d}{f}) \tag{12.6}$$

the equation (12.4) will have either two positive roots or no one. The roots will merge at

$$\varepsilon_1 = K(\varepsilon_2) \ ; \tag{12.7}$$

for $\varepsilon_1 < K(\varepsilon_2)$, they will not exist.

For the condition

$$g > h(1 + \frac{\delta d}{f}) \ , \tag{12.8}$$

the equation (12.4) will have one and only one positive root.

Stability of the equilibrium (x^*, y^*, z^*) where all coordinates are different from zero and y^* is a maximal root of the equation (12.4) will occur if the below inequality holds

$$\Delta = (ey^* + bx^*)(bBx^* - Afy^* + eBy^* + bex^* y^*) -$$
$$- (bkfx^* y^* + lkfx^* y^{*2} - cekx^{*2} y^{*2}) > 0 \qquad , \tag{12.9}$$

where

$$A = g \frac{\varepsilon_1 - \varepsilon_2}{(1 + \varepsilon_2 y^*)^2} \frac{x^*}{1 + \delta z^*} \qquad B = \frac{\delta(hx^* + ky^*)}{1 + \delta z^*} .$$

With Δ decreasing at the instant $\Delta = 0$, the equilibrium state (x^*, y^*, z^*) will give birth to the stable limit cycle Γ.

Now, let us sum up the results discussed. They are diagrammatically presented in figure 12.8.

The diagram also holds the numbers of the pictures of the related phase portraits. Let us comment on the diagram.

$\varepsilon_1 < k$ $\varepsilon_1 > k$

$g < h$ $g < h$
Fig. 12.2   *Fig. 12.4a, $\Delta > 0$; Fig. 12.4b, $\Delta < 0$*

$h < g < h(1 + \delta d / f)$ $h < g < h(1 + \delta d / f)$
Fig. 12.3   *Fig. 12.4c, $\Delta > 0$; Fig. 12.4b, $\Delta < 0$*

$g > h(1 + \delta d / f)$

$\Delta > 0$ $\Delta < 0$
Fig. 12.5   *Fig. 12.6*

$c e - l f < 0$ $c e - l f > 0$
of limited   *of unlimited*
development   *development*
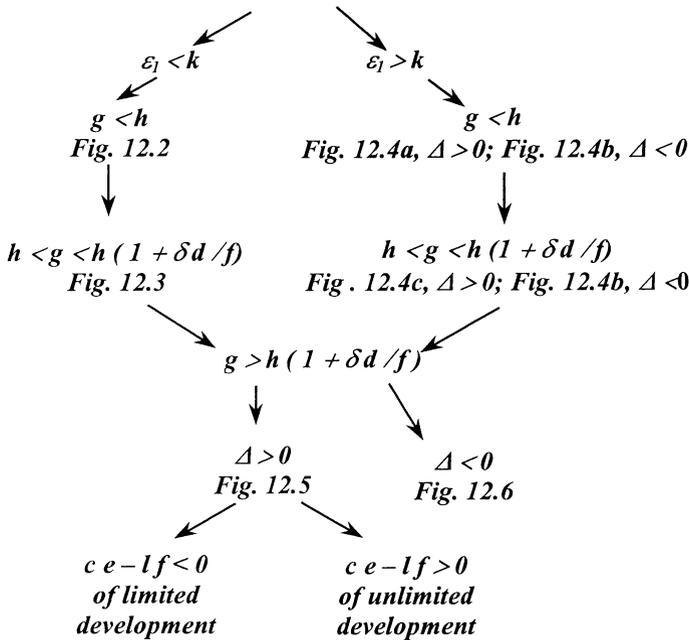
Fig. 12.8. The changes in the community structure in accordance with the parameters.

When we have $g < h$, the phase portrait, depending upon other parameters, may be of three types shown in figures 12.2, 12.4 and 12.4a. The first case stands for the globally stable equilibrium O_1 ($x^* > 0$, $y^* = 0$, $z^* = 0$) corresponding to the "producers" community. In the second case, the entire phase space is divided in two parts:

- the part Φ_1 is an attraction domain of the stable equilibrium O_1 ($x^* > 0$, $y^* = z^* = 0$); and

- the part Φ_2 is an attraction domain of either the stable equilibrium $O_4(x^* > 0,\ y^* > 0,\ z^* > 0)$ (Fig. 12.4) or the stable periodic motion (auto-oscillation) Γ (Fig. 12.4 a).

A birth of the second domain (Φ_2) and its attracting steady-state motion at small g is explained by its actual increase due to the managers who may appear because of their positive effect upon the community productivity. When this positive role of the managers ceases, they also vanish and the accumulated products vanish as well.

At the middle-technology level (when $h < g < h(1 + \dfrac{\delta d}{f})$) and also depend-

ing upon ε_1, there will be possible three different types of the phase portraits depicted in figures 12.3, 12.5 and 12.5a. The case represented in figure 12.3 describes an existence of the "producers – products" community. In this community, the managers are expelled. This happens when $\varepsilon_1 < K$; on the contrary, with $\varepsilon_1 > K$ there will arise the community holding the managers as well, in addition to the producers and the accumulated products. Depending upon the sign of the magnitude Δ, this community may be stable or unstable; its steady-state regime may be either an equilibrium state or an auto-oscillation. In the third case (high-level technologies), the managers will appear independently of their positive or negative impact upon the society. They arise since the society's productivity allows them to appear.

The arising variants of the phase portrait will depend upon the sign of the magnitude Δ and will cover a stable equilibrium community and an unstable auto-oscillatory community. In these both cases, the entire phase space will be the domain of attraction of an equilibrium state or auto-oscillation, i.e. they will be globally stable.

Note here that except the communities with the steady-state motion depending upon values of the parameters, there will be also possible the cases in which the steady-state motion will depend upon initial conditions as well.

Dividing the stable societies into unlimitedly developing and limitedly developing societies is not of a qualitative nature. This division is rather determined by a nature of variations of coordinates in the globally stable equilibrium x^*, y^*, z^* as $g \rightarrow \infty$. Note here that the same effect is exerted by the sign of the magnitude $ce - lf$ upon the nature of the variations undergone by the average values of the periodic functions $x^*(t), y^*(t), z^*(t)$ that describe auto-oscillations. This effect follows from the ratios for average values of these functions obtainable from the differential equations of the model

$$a - bx^*(t) - ly^*(t) + cz^*(t) = 0$$

$$- d - ey*(t) + fz*(t) = 0 \quad ,$$

since from these equations we obtain

$$\tilde{x}* = \frac{a}{b} + \frac{cd}{bf} + \frac{1}{bf}(ce - lf)\tilde{y}* \quad .$$

In conclusion, we want you to look at figure 12.8 again. It holds the conditions assigned for different types of phase portraits and shows how one type transfers to another when the parameters are changed. The parameter g and partly ε_1 determine a general structure of the community. Here, it is worthy of mentioning that, while at the high technological level the managers arise independently of their utility or harmfulness for the community , at the lower technological level their appearance is stipulated by their sufficiently high utility only. The sign of the parameter Δ determines a stability of the community, i.e. whether its limited steady-state motion is a stable equilibrium or a stable auto-oscillation. At last, important is also the parameter $ce - lf$ which determines a quantitative nature of the community development with increase of the technologies.

The diagram and the phase pictures may be imagined as a play of mathematical intelligence and imagination. Also, one may ponder here about how they pertain to many-century history of mankind and our fate.

13 Linear oscillators

A mathematical model for the linear oscillator. Possible kinds of motions. Phase and bifurcation portraits. The motions described by a linear oscillator: equilibrium states, harmonic, dissipative and divergent oscillations.

A linear oscillator is a very simple mathematical model wonderful by both its variety and width of specific interpretations and by the multiplicity of the phenomena it describes. In mechanics, an oscillator is anywhere where interacting mass and elasticity are available, as well, in those sectors of electrodynamics where there are capacitance and self-induction. Schematically they are given in figures 13.1 and 13.2.

   A linear oscillator describes periodic harmonic oscillations, dissipative and divergent oscillations of various frequencies, various equilibriums, namely, stable and unstable, and those like a node, a focus and a saddle. A mathematical model for a linear oscillator is a linear differential second-order equation of the form

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = 0 \ . \tag{13.1}$$

   This equation is of two parameters only, δ and ω^2; its phase space is two-dimensional.

   The simplest physical plants described by the equation (13.1) are a spring-attached mass and an electrical circuit of the self-inductor and a capacitor shown in figures 13.1 and 13.2.

   The mass m can move only along the axis x, and the elasticity force of the spring (when extended at x from its nonextended state $x = 0$) is equal to $-kx$. Here, "minus" stands for the force being opposite to the displacement of the mass m. According to the Newton equation,

$$mx = -kx$$

or

$$\ddot{x} + \frac{k}{m}x = 0$$

which for $\delta = 0$ and $\dfrac{k}{m} = \omega^2$ will coincide with (13.1). In a closed-loop circuit including the capacitance C and the self-inductance L (Fig. 13.2), q will be a capacitor charge; and hence, $I = q$ will be a current force. On the capacitor plates, there will be induced the voltage $\dfrac{q}{C}$ and in the self-inductor
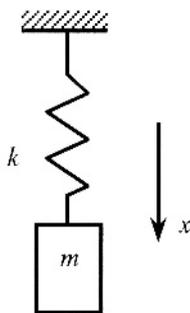
Fig. 13.1. The weight suspended to a spring.
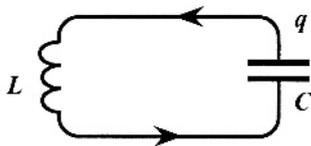
Fig. 13.2. The electrical circuit with a capacitor and a self-inductor.

the electromotive force (emf) equal to $-L\dfrac{dI}{dt}$. Again, here "minus" stands for the self-inductor emf being opposite to the current growth. Since the capacitor closes on the self-inductor, the voltages $\dfrac{q}{C}$ and $-L\dfrac{dI}{dt}$ will be equal to each other, i.e.

$$-L\frac{dI}{dt} = \frac{q}{C} \ ,$$

or, in view of $I = \dot{q}$, we obtain

$$\ddot{q} + \frac{q}{LC} = 0 \ .$$

This equation will coincide with (13.1), when $\delta = 0$ and $\omega^2 = \dfrac{1}{LC}$. In any real electrical circuit, there is an ohmic resistor (only, if we have no extraordinary case of superconductivity). The scheme of the corresponding circuit with such a resistor is represented in figure 13.3. In this case, the capacitor emf $\dfrac{q}{C}$ is opposite to emf $-L\dfrac{dI}{dt}$ of the self-inductor and the emf $-RI$ of the resistor being sequentially connected.
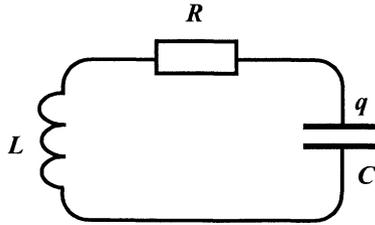
Fig. 13.3. The electrical circuit with a condenser, a self-inductor and a resistor.

Thus, we obtain

$$-L\frac{dI}{dt} - RI = \frac{q}{C}$$

or

$$\ddot{q} + \frac{R}{L}\dot{q} + \frac{1}{LC}q = 0$$

which coincides with (13.1) when $2\delta = \dfrac{R}{L}$ and $\omega^2 = \dfrac{1}{LC}$. Similarly, in the case of a spring-attached mass there may appear some move-hampering resistance proportional with the coefficient h to the velocity $\dot{x}$. Then, we get

$$m\ddot{x} = -kx - hx ,$$

or

$$\ddot{x} + \frac{h}{m}\dot{x} + \frac{k}{m}x = 0 ,$$

which again brings us to the linear oscillator equation (13.1) with $2\delta = \dfrac{h}{m}$ and $\omega^2 = \dfrac{k}{m}$.

The differential equation (13.1) has particular solutions of the form $x = e^{\lambda t}$ where λ is any of the roots of the so-called characteristic equation

$$\lambda^2 + 2\delta\lambda + \omega^2 = 0 .$$

This equation is derived through immediately placing these presupposed solutions into (13.1). In general, such roots are two; the general solution obtained under the superposition principle for linear dynamical systems will be as

$$x = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} . \qquad (13.2)$$

We remind here that the superposition principle when applied to the oscillator (13.1) implies that if $x_1(t)$ and $x_2(t)$ are the arbitrary solutions to the differential equation (13.1), then any of their linear combination $C_1 x_1(t) + C_2 x_2(t)$ will be a solution as well.

The general solution to the equation (13.1) at $\lambda_1 = \lambda_2$, i. e. at $\delta^2 = \omega^2$, can be found with the help of this principle as well. Indeed, let λ_1 be close to λ_2 and two independent solutions be $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$. However, as $\lambda_2 \rightarrow \lambda_1$ they get identical and only a single solution remains. Though, it is possible to consider the solution

$$(\lambda_1 - \lambda_2)^{-1}(e^{\lambda_1 t} - e^{\lambda_2 t})$$

which, as $\lambda_2 \to \lambda_1$, does not disappear and does not coincide with $e^{\lambda_1 t}$ – this solution becomes equal to $t\, e^{\lambda_1 t}$. Therefore, if $\lambda_1 = \lambda_2 = \lambda$, the general solution may be written as

$$C_1 e^{\lambda t} + C_2 t e^{\lambda t} \quad . \tag{13.3}$$

This simple consideration speaks about the fact that when solving this equation we may assume $\lambda_1 \neq \lambda_2$. In the opposite case, as $\lambda_2 \to \lambda_1$, one should then pass to the limit

The roots λ_1 and λ_2 of the characteristic equation can be real or complex. In any case, the general solution (13.3) with appropriate C_1 and C_2 can satisfy any initial conditions of the form

$$x\big|_{t=0} = x_0 \quad , \qquad \dot{x}\big|_{t=0} = x_1 \quad . \tag{13.4}$$

Indeed, these conditions, if specified, will bring us the next two linear equations regarding C_1 and C_2:

$$C_1 + C_2 = x_0 \qquad \lambda_1 C_1 + \lambda_2 C_2 = x_1 \quad .$$

From this, one can easily derive the below:

$$C_1 = \frac{\lambda_2 x_0 - x_1}{\lambda_2 - \lambda_1} \qquad C_2 = \frac{\lambda_1 x_0 - x_1}{\lambda_1 - \lambda_2} \quad .$$

Therefore, the corresponding will be as

$$x = (\lambda_2 - \lambda_1)^{-1}\left\{(\lambda_2 x_0 - x_1)e^{\lambda_1 t} - (\lambda_1 x_0 - x_1)e^{\lambda_2 t}\right\} \,.. \tag{13.5}$$

For complex roots, when

$$\lambda_{1,2} = -\delta \pm i\Omega \quad , \qquad \Omega = \sqrt{\omega^2 - \delta^2} \quad , \tag{13.6}$$

from (13.5) we find that

$$x = -\frac{1}{2}i\Omega\left\{\left[(-\delta - i\Omega)x_0 - x_1\right]e^{-\delta t}(\cos\Omega t + i\sin\Omega t) - \right.$$

$$\left. - \left[(-\delta + i\Omega)x_0 - x_1\right]e^{-\delta t}(\cos\Omega t - i\sin\Omega t)\right\} \tag{13.7}$$

$$= e^{-\delta t}(x_0 \cos\Omega t + \Omega^{-1}(\delta x_0 + x_1)\sin\Omega t).$$

This very result (13.7) and a general form for the real solution of the characteristic equation with complex roots may be obtained in another way. Again, due to the superposition principle,

$$\frac{1}{2}(e^{\lambda_1 t} + e^{\lambda_2 t}) = e^{-\delta t} \cos \Omega t$$

and

$$\frac{1}{2i}(e^{\lambda_1 t} - e^{\lambda_2 t}) = e^{-\delta t} \sin \Omega t$$

will be the solutions as well as e$^{\lambda 1 t}$ and e$^{\lambda 2\, t}$; but they are real already. The superposition of these real solutions gives us a general solution.

What are then the derived general solutions

$$C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} \tag{13.8}$$

with real λ_1 and λ_2 and

$$e^{-\delta t} (C_1 \cos \Omega t + C_2 \sin \Omega t) \tag{13.9}$$

with complex λ_1 and λ_2 .?

We are interested both in the plots of $x(t)$ as functions of time t and the phase portraits corresponding to the cases (13.8) and (13.9). Both this and that are dependent upon the parameters δ and ω^2. First, which of these cases, (13.8) or (13.9), occurs will depend upon these parameters.

For the sake of visibility, let us introduce a plane of the parameters ω^2 and δ (Fig. 13.4). We emphasise that both parameters here ,δ and ω^2, can have any sign. To the marginal case separating (13.8) and (13.9) there will correspond the equality $\delta^2 = \omega^2$. On the plane in figure 13.4, to this equality there will correspond the parabola s. It is inside this parabola that $\delta^2 < \omega^2$ occurs; it is beyond this parabola, $\delta^2 > \omega^2$, there occurs the case (13.8).

Let us , first, consider the oscillators corresponding to the points lying within the parabola. To the points on the axis $\delta = 0$ within the parabola (indicated by 1 in figure 13.4), there will correspond the harmonic oscillators

$$\ddot{x} + \omega^2 x = 0 , \tag{13.10}$$

whose motions are the well-known harmonic oscillations

$$x = C_1 \cos \omega t + C_2 \sin \omega t = A \cos(\omega t + \varphi), \tag{13.11}$$

where A is an oscillation amplitude, ω is an oscillation frequency, and φ is a phase. The amplitude A and the phase φ are related to the constants C_1 and C_2 as

$$A \sin \varphi = C_1 \ , \qquad A \cos \varphi = C_2 \ ,$$

from which it follows that

$$A = \sqrt{C_1^2 + C_2^2} \ , \qquad tg\varphi = \frac{C_1}{C_2} \ .$$
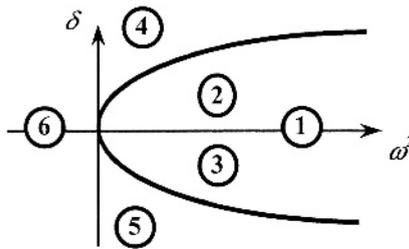
Fig. 13.4. Decomposing the plane of parameters ω^2 , δ of the linear oscillator into the domains of different roots λ_1 and λ_2 of the characteristic equation: **1** both roots are purely imaginary; **2** complex with negative real parts; **3** complex with positive real parts; **4** real negative; **5** real positive; and **6** real of different signs.
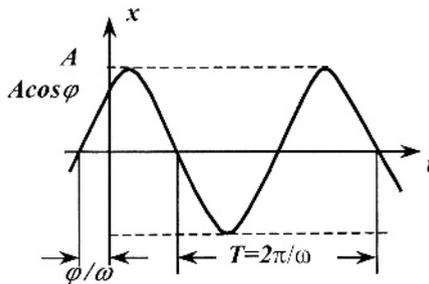
Fig. 13.5. The time graph of harmonic oscillations.

The plot of harmonic oscillations is given in figure 13.5. In it, also shown are the amplitude A, the phase φ and the period $T = \dfrac{2\pi}{\omega}$.

From (13.11), we derive the below parametric equations of phase trajectories

$$x = A\cos(\omega t + \varphi)$$
$$\dot{x} = -\omega A \sin(\omega t + \varphi) \ .$$

The point $(x, \dot{x})$, being a function of t, runs clockwise along the ellipse

$$\frac{x^2}{A^2} + \frac{\dot{x}^2}{\omega^2 A^2} = 1 \ .$$

To different A there will correspond different ellipses, and this will result in the phase portrait given in figure 13.6.
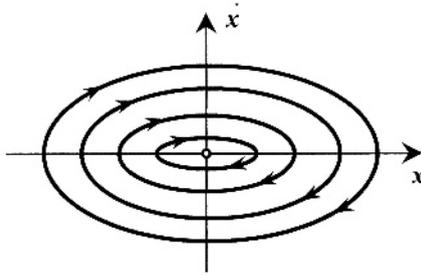
Fig. 13.6. The phase portrait of the harmonic oscillator (domain 1 in Fig. 1.4).

Note that the phase trajectories could be also found through immediately integrating the differential equation (13.10) of the harmonic oscillator. Namely, through multiplying (13.10) by $\dot{x}$ and integrating the result we find that

$$\frac{\dot{x}^2}{2} + \frac{\omega^2 x^2}{2} = h,$$

where the integrating constant h, corresponding to the scaled oscillator energy, will be related to the oscillation amplitude of the oscillator as follows

$$h = \frac{1}{2}\omega^2 A^2 .$$

Let us pass now to the domain above the abscissa in the plane of parameters ω^2, δ, yet remaining within the parabola. Then, due to (13.9), we obtain

$$x = Ae^{-\delta t} \cos(\Omega t + \varphi) \quad . \tag{13.12}$$

This will correspond to the dissipative harmonic oscillations shown in figure 13.7.
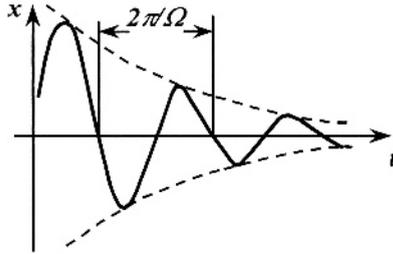
Fig. 13.7. The exponentially dissipative harmonic oscillation (domain 2 in Fig. 13.4).

Their oscillation period $\dfrac{2\pi}{\Omega}$ is a little larger than that of the associated harmonic oscillator being equal to $\dfrac{2\pi}{\omega}$ ($\Omega = \sqrt{\omega^2 - \delta^2}$). For $\delta \langle 0$, i.e. when displacing down from the abscissa axis in the plane of parameters ω^2, δ (Fig. 13.4), we shall get the same solution (13.12). Though, this time the associated oscillations will be the increasing ones, as shown in figure 13.8.
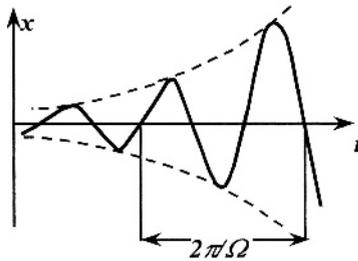
Fig. 13.8. The exponentially divergent harmonic oscillations (domain 3 in Fig. 13.4).

In the cases discussed, the phase portraits consist of the twisting $(\delta > 0)$ or untwisting $(\delta < 0)$ spirals represented in figures 13.9a and 13.9b.

Thus, we have constructed the plots of $x(t)$ and the phase portraits within the domain 1 $(\delta = 0, \omega^2 > 0)$, 2 $(\delta > 0, \delta^2 \langle \omega^2 \rangle)$ and 3 $(\delta < 0, \delta^2 \langle \omega^2 \rangle)$ (Fig. 13.4).

Now let us draw the plots of $x(t)$ and the phase portraits for the oscillators from the domains 4 $(\delta > 0, \delta^2 > \omega^2 > 0)$, 5 $(\delta < 0, \delta^2 > \omega^2 > 0)$, 6 ($\omega^2 < 0$) of the same figure 13.4. In contrast to previous cases corresponding to complex roots of the characteristic equation, now these roots are real and the general solution will be represented by the formula (13.8). This solution is a superposition of the two solutions, $x = e^{\lambda_1 t}$ and $x = e^{\lambda_2 t}$. The plots of these solutions are increasing or falling exponents depending upon the signs of λ_1 and λ_2. This makes it possible for us to imagine the shapes of the plots for their superpositions with any coefficients C_1 and C_2. To construct the phase portraits, let us first draw the phase trajectories for specific solutions $x = e^{\lambda_1 t}$ and $x = e^{\lambda_2 t}$.
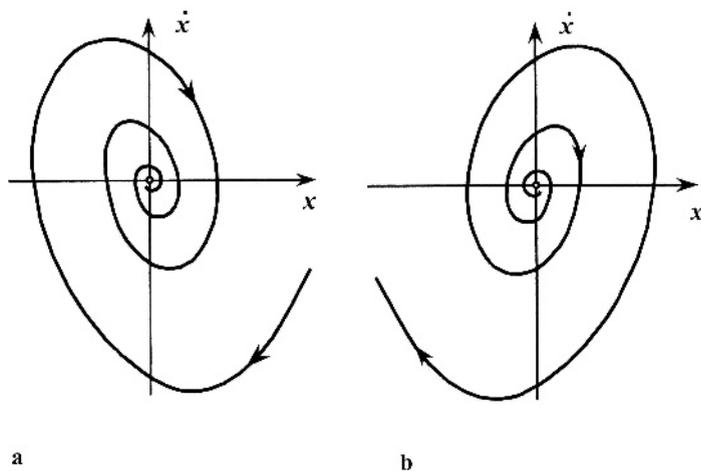
Fig. 13.9. The phase portraits for **a** the domain 2 in Fig. 13.4 and **b** the domain 3 in the same figure.

Let λ be either λ_1 or λ_2, then in both cases the parametric equation of the phase trajectory is written as

$$x = e^{\lambda t} \ , \quad \dot{x} = \lambda e^{\lambda t} \quad (-\infty < \ t \ < \ \infty) \ . \tag{13.13}$$

The phase trajectories are represented as rays

$$\dot{x} = \lambda x \quad (\,x > 0 \text{ or } x < 0)$$

running by the phase point, according to (13.13), from the origin to infinity, when $\lambda > 0$, or vice versa, from infinity to zero, when $\lambda < 0$. It is not difficult to understand that the below vector equality

$$(x, \dot{x}) = C_1(1, \lambda_1)e^{\lambda_1 t} + C_2(1, \lambda_2)e^{\lambda_2 t} \tag{13.14}$$

immediately follows (13.13). The geometrical interpretation appropriated is represented in figure 13.10.
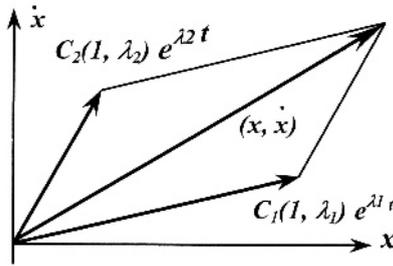
Fig. 13.10. The subsidiary construction made to create the phase portrait for a nodal equilibrium state.
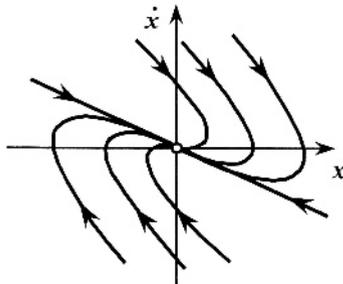
Fig. 13.11. The phase portrait of a stable node (corresponds to the domain 4 in Fig. 13.4).

Through changing signs and values of C_1 and C_2, we obtain a possibility to construct a phase portrait. In the domain 4 (Fig. 13.4), λ_1 and λ_2 are negative and the corresponding phase portrait is given in figure 13.11. In the domain 5, λ_1 and λ_2 are positive; the associated phase portrait will be of the shape shown in figure 13.12.
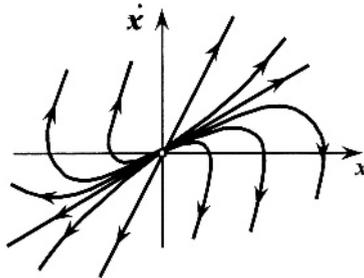
Fig. 13.12. The phase portrait of an unstable node (corresponds to the domain 5 in Fig. 13.4).

In the domain 6, λ_1 and λ_2 are of different signs and the phase portrait is of the form shown in figure 13.13. Thus, we have found all possible forms of the phase portraits for a linear oscillator. In each of them, there will be available a special phase trajectory converging to a unique point, viz. to the equilibrium state $x = \dot{x} = 0$. The remaining trajectories will lie, as it were, around it; therefore, one may speak about various kinds of equilibrium states.
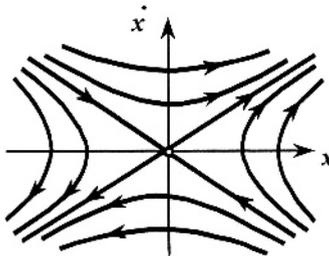
Fig. 13.13. The phase portrait of a saddle (corresponds to the domain 6 in Fig. 13.4).

They correspond to the domains 1,2,3,4,5 and 6 in figure 13.4 and are called, respectively, a centre, a stable and unstable foci, a stable and unstable nodes and a saddle. In this listing, the boundary 1 between the domains 2 and 3 is distinguished and the rest boundaries were left without our attention. This was done due to its essential role. This is the boundary between exponential stability and instability, the dissipative and divergent oscillations. The case of the crucial boundary is a general one for conservative and, in particular, Hamiltonian systems.

   The phase portraits of the harmonic oscillators covered by this boundary case carry some specificity that is extremely essential for statistical mechanics. This specificity is revealed when, instead of a single oscillator, there is taken an ensemble of identical oscillators, each of which being represented by its running phase point. Here, a phase portrait may be interpreted as a stream of particles representing oscillators or as a stream of phase fluid. Phase particles move in time and are deformed in some way. If there is assigned the domain G_0 in phase space , then, upon the time t, this domain will be transformed to the domain G_t, in accordance with the transformation

$$x = e^{-\delta t}\left\{(\cos\Omega t + \frac{\delta}{\Omega}\sin\Omega t)x_0 + \sin\frac{\Omega t}{\Omega}\dot{x}_0\right\},$$

$$\dot{x} = e^{-\delta t}\left\{-(\Omega + \frac{\delta^2}{\Omega})\sin\Omega t x_0 + (-\frac{\delta}{\Omega}\sin\Omega t + \cos\Omega t)\dot{x}_0\right\}.$$

This transformation is received from the formula (13.7) for the solution $x(t)$ with the initial conditions x_0 and x_1. The jacobian of this transformation

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial x_0} & \dfrac{\partial x}{\dot{x}_0} \\ \dfrac{\partial \dot{x}}{\partial x_0} & \dfrac{\partial \dot{x}}{\partial \dot{x}_0} \end{vmatrix}$$

is computed in a simple way and turns out to be equal to $e^{-2\delta t}$. This result implies that the phase fluid, while floating, is compressed when $\delta > 0$, and when $\delta < 0$ it will expand; when $\delta = 0$ it will behave as incompressed. To be more exact, for the time t any element of the phase volume $d\sigma_0$ will be transformed into the element $d\sigma_t$ and

$$d\sigma_t = e^{-2\delta t}d\sigma_0 \ .$$

Therefore, $e^{-2\delta}$ is a compression coefficient ($\delta > 0$) or an extension coefficient ($\delta < 0$) of the floating phase liquid. For $\delta = 0$, this extension coefficient is equal to unity; this corresponds to the incompressed fluid flow.

The above listed kinds of the linear oscillator phase portrait – a centre, a focus, a node and a saddle – are available not only for linear oscillator equilibriums but also for the equilibriums $O(x^*, y^*)$ of any two-dimensional dynamical systems represented by two second-order differential equations

$$\dot{x} = X(x, y) \ , \qquad \dot{y} = Y(x, y) \ . \qquad (13.15)$$

We already encountered this when dealing with the phase portraits of the interacting populations. In that case, these phase portraits occurred not in the entire plane but within some neighbourhood of the equilibriums only. As well as for the oscillator, the type of the phase portrait within the equilibrium neibourhood of x^*, y^* will be determined by the roots λ_1 and λ_2 of the characteristic equation. As well known, this equation for the equation (13.15) takes the form:

$$X(\lambda) = \begin{vmatrix} (\dfrac{\partial X}{\partial x})^* - \lambda & (\dfrac{\partial X}{\partial y})^* \\[2mm] (\dfrac{\partial Y}{\partial x})^* & (\dfrac{\partial Y}{\partial y})^* - \lambda \end{vmatrix} =$$

$$= \lambda^2 - ((\dfrac{\partial X}{\partial x})^* + \left(\dfrac{\partial Y}{\partial y}\right)^*)\lambda + \left(\dfrac{\partial X}{\partial x}\right)^* \left(\dfrac{\partial Y}{\partial y}\right)^* - \left(\dfrac{\partial X}{\partial y}\right)^* \left(\dfrac{\partial Y}{\partial x}\right)^* = 0 \ .$$

Here, the stars stand for the derivatives being computed in the equilibrium point x^*, y^* .

In the neibourhood of the equilibrium state, the linearized equations (13.15) are of the form

$$\dot{\xi} = \left(\dfrac{\partial X}{\partial x}\right)^* \xi + \left(\dfrac{\partial X}{\partial y}\right)^* \eta$$

$$\dot{\eta} = \left(\dfrac{\partial Y}{\partial x}\right)^* \xi + \left(\dfrac{\partial Y}{\partial y}\right)^* \eta \ ,$$

where ξ and η are deviations from the equilibriums x^*, y^*, i.e. $\xi = x - x^*$ and $\eta = y - y^*$.

Excluding ξ or η from these differential equations leads to the linear oscillator equation for ξ

$$\ddot{\xi} - \left[\left(\frac{\partial X}{\partial x} \right)^* + \left(\frac{\partial Y}{\partial y} \right)^* \right] \dot{\xi} + \left[+ \left(\frac{\partial X}{\partial x} \right)^* \left(\frac{\partial Y}{\partial y} \right)^* - \left(\frac{\partial X}{\partial y} \right)^* \left(\frac{\partial Y}{\partial x} \right)^* \right] \xi = 0$$

or to exactly the same equation for η.

Below, we shall give you some specific examples of the linear oscillators for all domains 1-6 of the bifurcation portrait held in figure 13.4. Here, the linear oscillator model (13.1) is, as a rule, this or that approximation of the real system. Very often, this system is admissible only for small or not very large deviations from an equilibrium. Say, in the linear approximation, the lower state of the pendulum equilibrium is described through the linear harmonic oscillator with the centre-type equilibrium. The movement close to the upper equilibrium is described through the oscillator with the unstable saddle-type equilibrium. Indeed, the circle pendulum

$$\ddot{\varphi} + \frac{g}{l} \sin \varphi = 0 \tag{13.16}$$

will have two equilibriums: the lower $\varphi^* = 0$ and the upper $\varphi^* = \pi$. Supposing that $\varphi = \varphi^* + x$, we obtain

$$\ddot{x} + \frac{g}{l} \sin(\varphi^* + x) = 0 \ ,$$

or

$$\ddot{x} + \frac{g}{l} \sin \varphi^* + \frac{g}{l} x \cos \varphi^* + ... = 0 \qquad , \tag{13.17}$$

where the dots stand for the magnitudes of the third and greater order of smallness with respect to the supposedly small x. From (13.17), for $\varphi^* = 0$ it will follow that

$$\ddot{x} + \omega^2 x = 0, \quad \omega^2 = \frac{g}{l} \ ;$$

and for $\varphi^* = \pi$,

$$\ddot{x} - \omega^2 x = 0 \ ,$$

as was stated.

The pendulum determined by the mathematical model (13.16) will be represented by us as a solid suspended in the gravitational field. Though, this pendulum may be also represented as a dipole in the electrical field. If I is an inertia moment with respect to the mass centre of the $2l$ - length dipole, e and $-e$ are the charges on its terminals, E is an electrostatic intensity, then the angle φ of the dipole deviation from the electrical field direction will satisfy the equation

$$I\ddot{\varphi} = -2le\sin\varphi \ .$$

Here again we obtain the classical equation of the pendulum. In the linear approximation at small deviations from the equilibrium values $\varphi = 0$ and $\varphi = \pi$, it will be a linear oscillator.

With a good approximation, the linear oscillator describes small oscillations of a boat or some other body swimming in water. A boat or a ship have board and keel oscillations, each of which possessing its own frequency. Through a linear oscillator it is possible to approximately describe the oscillations of a string or a building. Here, for the building we mean the vertical oscillations of its base and its transversal oscillations. A 100-storey skyscraper oscillates so that transversal displacements of its top make up about 1 meter. Through the oscillator it is possible to describe also the elastic oscillations of gas contained inside the Helmholtz resonator-like vessel, the oscillations of molecules in a crystal, the oscillations of electromagnetic waves within a transmitting cavity. This case also covers many other oscillations, such as those of a car, a train, an air bubble in water, and the oscillations of the various electrical circuits in a radio-circuit board; the oscillations of water in the communicating vessels, the oscillations of fluid in water ions, the oscillations in a vestibular system, the twisting oscillations of the flywheel mounted on a stiff shaft.

In all the examples of the oscillators listed above, the energy dissipation will take place and the oscillations will dissipate. Here, we shall obtain $\delta > 0$ or $\delta = 0$ if this dissipation may be neglected. Below, you will be given some examples with $\delta < 0$, with the so-called negative friction when, in contrast, there occurs some energy pump-up from some source. Besides, we shall also discuss some nonlinear oscillators for which a linear description will be not sufficient; their behaviour is different essentially.

In conclusion, you are invited to look at the below common figure 13.14 as a general result of all the above said. It shows the evolution of $x(t)$ and the types of the phase portraits for the parameters ω^2 and δ taken from the different domains of the bifurcation diagram.
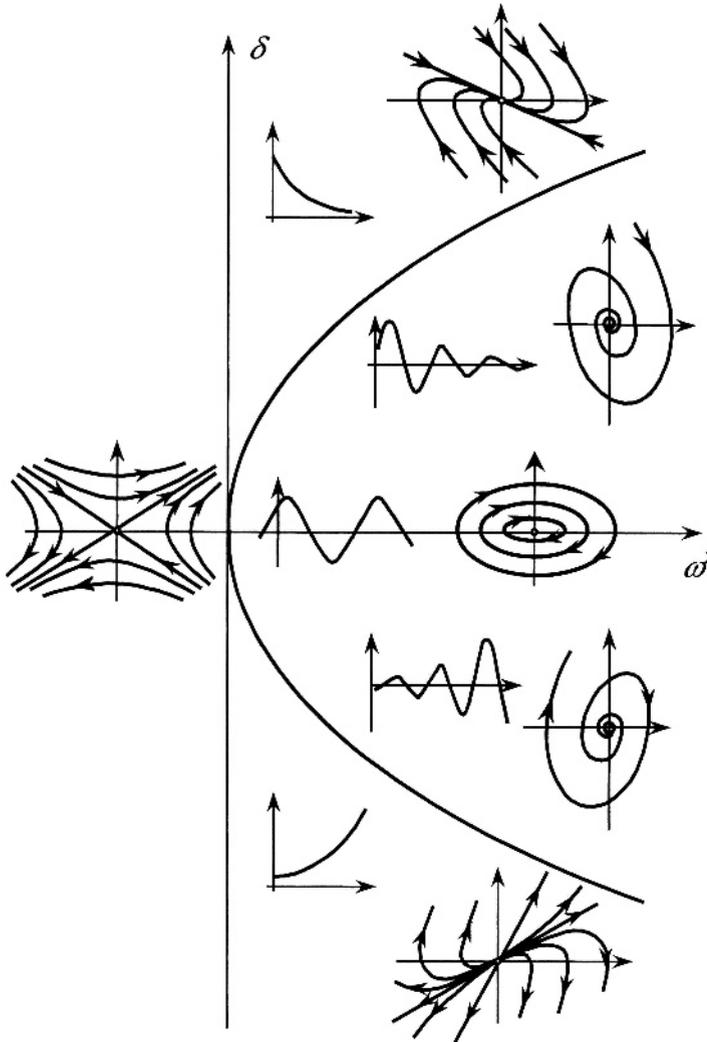
Fig. 13.14. The final parametric portrait of a linear oscillator. Each of its domains contains both time graphs and phase portraits.

14 Electromechanical analogies. Lagrange-Maxwell equations

Mechanical and electrical oscillators. Electromechanical analogies. The Lagrange-Maxwell equations. Invariance of the Lagrange equations. The least-action principle as a variation form of mathematical models.

Let us return to the simplest examples of mechanical and electrical linear oscillators. The first physical model is described by the differential equation

$$m\ddot{x} + h\dot{x} + kx = 0 \qquad (14.1)$$

and the second by the differential equation

$$L\ddot{q} + R\dot{q} + (1/C)q = 0 \qquad (14.2)$$

The models (14.1) and (14.2) turn into each other via replacing the variable x by q ($\dot{x}$ by $\dot{q}$, $\ddot{x}$ by $\ddot{q}$) and the parameters m by L, h by R, and k by $1/C$. Therefore, if $m = L$, $h = R$, $k = 1/C$ and if, at some initial point of time, $x = q$ and $\dot{x} = \dot{q}$, then the latter equalities will take place for all subsequent time. The variable x is a coordinate of mass position and q is a coordinate of capacitor electric charge, and under the above conditions these variables will vary similarly. Analogously, all the time the value of the mass displacement velocity $\dot{x}$ will coincide with one of the electrical current $\dot{q}$ in the circuit. Thus, one may speak about the analogy between the spring - attached mass with viscous friction and an electric circuit of a self-inductor, a capacitor and a resistor. The analogy here is implied in the sense that the absolutely different physical magnitudes x and q vary identically, provided $m = L$, $h = R$, $k = 1/C$. And analogous are mechanical displacement and electrical charge, velocity and current, mass and self-induction, viscous friction coefficient and ohmic resistance, spring stiffness coefficient and the magnitude inverse to the capacitance.

Let us proceed with this analogy. The mass kinetic energy is equal to $m\dot{x}^2/2$. Its analogue will be the magnitude $L\dot{q}^2/2$. It is known to be the energy of a self-inductor magnetic field. The potential energy of the spring is $kx^2/2$, its

analogy is $q^2/2C$, and it is, as we saw before, the energy of capacitor electric field. The product kx is a spring compressing force and its corresponding magnitude is the capacitor electrical voltage q/C.

Thus, we have revealed the analogy between the magnitudes of displacement, velocity, force, kinetic and potential energies and, respectively, charge, current, voltage, self-inductor magnetic field energy and capacitor electric field energy. At the same time, the mass m, the viscous resistance h, and the elasticity k are similar to the self-induction L, the ohmic resistance R, and the value $1/C$ being inverse to capacitance.

This analogy does not only relate to the spring-attached mass and electric circuit. It is much wider. To make sure of that, let us imagine a mechanical and an electrical system of the elements: masses, "elasticities", "viscous resistors" and, respectively, self-inductors, capacitors and electric resistors. Each of these simplest elements is represented by its mathematical model. They are as follows. The mass m "ties together" the acceleration $\ddot{x}$ and the force F by $m\ddot{x} = F$. And similarly, the self-inductor L ties together the current change rate $\ddot{q}$ and the voltage E so that $L\ddot{q} = E$. The same are the analogies between stiffness and capacitance, mechanical viscous and electrical resistances. Let us write down all of them:

$$m\ddot{x} = F \longleftrightarrow L\ddot{q} = E$$

$$h\dot{x} = F \longleftrightarrow R\dot{q} = E \qquad (14.3)$$

$$kx = F \longleftrightarrow q/C = E.$$

Here, in each of the analogies (14.3), the force F and the voltage E have their own sense and value.

After that, it is clear that, if one has two systems, mechanical and electrical, composed of the above elements in the way that there appear the same links between the analogous elements, then the systems as a whole will be analogous too. This makes up the basis of an analogous modelling that creates a possibility for studying a purely mechanical system through its electrical analogue, and vice versa.

Let us omit this subject for a while and restrict ourselves with the above said. Only think it over what relates to what in this remarkable analogy. Let us write it once more

$$
\begin{array}{lll}
x \longleftrightarrow q & mv^2/2 \longleftrightarrow LI^2/2 & m \longleftrightarrow L \\
v = \dot{x} \longleftrightarrow I = \dot{q} & kx^2/2 \longleftrightarrow q^2/2C & m \longleftrightarrow 1/C \quad (14.4) \\
F \longleftrightarrow E & & h \longleftrightarrow R.
\end{array}
$$

And now let us pass to a difficult question of constructing mathematical models for mechanical and electromechanical systems.

For mechanical systems, Lagrange invented a general simple and convenient universal way of creating their dynamical mathematical models in the shape of the famous Lagrange differential equations. In general, this is a great and beautiful theory. You will become familiar with it from appropriate lecturing courses. Right now, without going into details, I will try to describe you an algorithm of how to construct these differential equations for an arbitrary mechanical system in which the positions of all its material points are determined by a finite collection of scalars $q_1, q_2, ..., q_n$ named as generalized coordinates.

Let, at first, all forces acting in the system be potential, i.e., there exists some potential energy $V(q_1, q_2, ..., q_n, t)$ for them, depending on the generalized coordinates and, possibly, on the time t. It implies that the work of these forces along any virtual displacement $\delta q_1, \delta q_2, ..., \delta q_n$ will be equal to reduction of potential energy, i.e.,

$$\delta A = -\delta V = -\frac{\partial V}{\partial q_\alpha} \delta q_\alpha .$$

Further, it is clear then that at any moment of time the kinetic energy T of the system can be expressed via the generalized coordinates $q_1, q_2, ..., q_n$ and the generalized speeds $\dot{q}_1, \dot{q}_2, ..., \dot{q}_n$. Because the kinetic energy, as you know, is a sum of the kinetic energies of all system material points whose positions uniquely depend upon $q_1, q_2, ..., q_n$ and t.

According to Lagrange, it is necessary to write the below function

$$L = T - V , \tag{14.5}$$

and with the help of it, the motion equations desired are then written as

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_\alpha} - \frac{\partial L}{\partial q_\alpha} = 0 , \ (\alpha = 1, 2, ..., n) . \tag{14.6}$$

The analogy explained above makes it possible to transfer this very beautiful general technique of writing motion equations to electrical systems. In his time, this was done by Maxwell. This can be done not only for electrical systems but for electromechanical systems as well. As the above analogy demands, you are only requested to do a single thing, and this thing is understanding that the kinetic energy T now is a sum of the mechanical kinetic energy and the energy of magnetic fields of inductors; and under the potential energy V you have to

understand a sum of the mechanical potential energy and the energies of electric fields of capacitors. Here, the generalized coordinates $q_1, q_2, \ldots, q_n$ are thought to be any scalars that define not only a mechanical position but also all the charges whose derivatives determine all electrical currents. We emphasize that this time the system contains neither mechanical nor electrical resistors.

Now through a very simple example of the mathematical model for motion, let us explain how the notation of the differential equations in the form of the Lagrange equation (14.6) comes into being.

Let us take the Newton equation describing motions of the linear mechanical oscillator

$$m\ddot{x} + kx = 0 \qquad (14.7)$$

and carry out the following transformations

$$m\ddot{x} = \frac{d}{dt}\frac{\partial}{\partial \dot{x}}\frac{m\dot{x}^2}{2}, \quad kx = \frac{\partial}{\partial x}\frac{kx^2}{2},$$

and, hence, we obtain

$$m\ddot{x} + kx = \frac{d}{dt}\frac{\partial}{\partial \dot{x}}\left(\frac{m\dot{x}^2}{2} - \frac{kx^2}{2}\right) - \frac{\partial}{\partial x}\left(\frac{m\dot{x}^2}{2} - \frac{kx^2}{2}\right) = 0, \qquad (14.8)$$

which coincides in its form with the Lagrange equation (14.6). Here, x is a generalized coordinate, $m\dot{x}^2 / 2$ is a kinetic energy, and $kx^2 / 2$ is a potential energy.

Similarly, to the Lagrange form there is reduced the following equation of the electrical oscillator

$$L\ddot{q} + \frac{q}{C} = \frac{d}{dt}\frac{\partial}{\partial \dot{q}}\left(\frac{L\dot{q}^2}{2} - \frac{q^2}{2C}\right) - \frac{\partial}{\partial q}\left(\frac{L\dot{q}^2}{2} - \frac{q^2}{2C}\right) = 0, \qquad (14.9)$$

where the generalized coordinate is the charge q of capacitor, $L\dot{q}/2$ is the magnetic energy of the self-inductor, and $q/2C$ is the electric energy of the capacitor.

These computations look artificial and proving nothing. It is so indeed. In order to give them some sense, it is necessary to penetrate into the secret of the Lagrange notation for the motion equations. Its secret is in the fact that these equations correspond to some variation problem on minimum of a functional. As known, the necessary condition for the minimum is that the first variation of the functional must be equal to zero. These conditions under which the first variation is reduced to zero turn out to be the Lagrange equations. This makes up the secret

of the surprising form and universality of the Lagrange equations. However, let us not to be in a hurry and, at first, get acquainted with the surprising property of invariance in the Lagrange notation for motion equations, the independence of their form upon the generalized coordinates chosen. In particular, it takes place for the equations (14.8) or (14.9). It turns out that, with any change of variables of the kind $x = f(y)$ ($q = f(p)$), the form of these equations still retains, in contrast to the Newton or Kirchhoff equations.

Indeed, replacing the variables in the Newton differential equation (14.7) reduces it to the form

$$m(\frac{\partial^2 f}{\partial y^2}\dot{y}^2 + \frac{\partial f}{\partial y}\ddot{y}) + kF(y) = 0, \tag{14.10}$$

which is different from the initial writing (14.7). Now, let us perform the same change of variables in the Lagrange notation (14.8). We need to make sure that

$$\frac{d}{dt}\frac{\partial}{\partial\dot{x}}(\frac{m\dot{x}^2}{2} - \frac{kx^2}{2}) - \frac{\partial}{\partial x}(\frac{m\dot{x}^2}{2} - \frac{kx^2}{2}) = 0 \tag{14.11}$$

implies

$$\frac{d}{dt}\frac{\partial}{\partial\dot{y}}(\frac{m(\frac{\partial f}{\partial y}\dot{y})^2}{2} - \frac{kf^2(y)}{2}) - \frac{\partial}{\partial y}(\frac{m(\frac{\partial f}{\partial y}\dot{y})^2}{2} - \frac{kf^2(y)}{2}) = 0 \tag{14.12}$$

and vice versa, i.e., a change of variables in the Lagrange equation is reduced to a change of variables in the Lagrange function (the Lagrange function must be a function of the generalized coordinates and speeds we accepted). For this, it is sufficient to show that (14.12) coincides with (14.10), since (14.10) coincides with (14.7) and the latter coincides with (14.11). Indeed,

$$\frac{d}{dt}\frac{\partial}{\partial\dot{y}}(\frac{mf'^2\dot{y}^2}{2} - \frac{f^2}{2}) - \frac{\partial}{\partial y}(\frac{mf'^2\dot{y}^2}{2} - \frac{f^2}{2}) =$$

$$= [m(f\ddot{y} + f''\dot{y}^2) + kf]\frac{\partial f}{\partial y} = 0$$

and, as $\partial f / \partial y \neq 0$, we come to the equation (14.10).

Thus, the Lagrange notation possesses a surprising property of invariance for the writing form with respect to a change of variables. This notation does not depend on the variables chosen; it is invariant with respect to a change of variables. It is interesting where this invariance comes from. To outguess it is not

so easy. However, in research it is often very important to reveal a certain toe-hold and then you have to pierce until you have caught the idea. And what should be pierced is, perhaps, clear – evidently there exists a certain   coordinate-free approach to separating the actual   system motions from all imaginably possible. You know that it is this, after all, that differential equations do.

From a historic point of view, the variation formulation of the mechanical motion laws, independent of coordinates, was in the air for a long time,   and it rested upon theological conceptions about the expedience of Providence. Much has been written about it. Most likely, even nowadays it is difficult to come to this purely logically . Therefore, without playing cunning, we will tell point-blank what the matter is.

Let us take a time integral of the Lagrange function from t_0 to t_1

$$\int\limits_{t_0}^{t_1} L dt$$

and call this magnitude an action. Let us denote it by W and write in detail what the action W depends upon

$$W = \int\limits_{t_0}^{t_1} L(q_1(t),...,q_n(t); \dot{q}_1(t),...,\dot{q}_n; t) dt .$$

It is determined by the time-dependent functions $q_\alpha(t)$ $(\alpha = 1,2,...,n)$ within the time interval from t_0 to t_1, i.e., this is a functional. Let us consider how this functional will vary at little variations of the functions $q_\alpha(t)$. For this purpose, let us slightly change these functions   by substituting $q_\alpha(t)$ with $q_\alpha(t) + \delta q_\alpha(t)$ and find the variation of the action W up to magnitudes of the first-order smallness with respect to the little variations $\delta q_\alpha(t)$ and $\delta \dot{q}_\alpha$.

We shall then have the following

$$\delta W = \delta \int\limits_{t_0}^{t_1} L dt = \int\limits_{t_0}^{t_1} (\sum \frac{\partial L}{\partial q_\alpha} \delta q_\alpha + \frac{\partial L}{\partial \dot{q}_\alpha} \delta \dot{q}_\alpha) dt =$$

$$= \int\limits_{t_0}^{t_1} (\sum_\alpha \frac{\partial L}{\partial q_\alpha} \delta q_\alpha + \frac{\partial L}{\partial \dot{q}_\alpha} \frac{d}{dt} \delta q_\alpha) dt =$$

$$= \sum_{\alpha} \frac{\partial L}{\partial \dot{q}_{\alpha}} \delta q_{\alpha} \Big|_{t_0}^{t_1} + \int_{t_0}^{t_1} \sum_{\alpha} (\frac{\partial L}{\partial q_{\alpha}} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_{\alpha}}) \delta q_{\alpha} dt .$$

The last operation in these calculations is a partial integration. As a result, we reveal to our great surprise that integrands here are the below expressions

$$\frac{\partial L}{\partial q_{\alpha}} - \frac{d \partial L}{dt \partial \dot{q}_{\alpha}}, \tag{14.13}$$

and getting them equal to zero will be just exactly the Lagrange equations. What then should be required to have them turned to zero? What must then be the assumption concerning the action W that would bring such a conclusion? To outguess it is not difficult. Let us require that

$$\delta W = 0,$$

provided that there are no changes of the functions $q_{\alpha}(t)$ on the terminals, and, hence, $\delta q_{\alpha} = 0$ for $t = t_0$ and $t = t_1$. Indeed, from these assumptions it follows that

$$\int_{t_0}^{t_1} [\sum_{\alpha} (\frac{\partial L}{\partial q_{\alpha}} - \frac{d \partial L}{dt \partial \dot{q}_{\alpha}}) \delta q_{\alpha}] dt = 0 \tag{14.14}$$

for the arbitrary ∂q_{α} $(\alpha = 1,...,n)$ reducing to zero for $t = t_0$ and $t = t_1$. It is not difficult to understand that this will result in all the expressions (14.13) being equal to zero. This will be so just due to the arbitrariness of δq_{α} which , by virtue of this, may be, in particular, chosen equal to

$$\delta q_{\alpha} = (\frac{\partial L}{\partial q_{\alpha}} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_{\alpha}}) \varepsilon^2(t),$$

where $\varepsilon^2(t)$ reduces to zero at terminals of the closed interval $[t_0, t_1]$ only. For such variations of the functions $q_{\alpha}(t)$, from (14.14) it follows that

$$\int_{t_0}^{t_1} \left[\sum_{\alpha} \left(\frac{\partial L}{\partial q_{\alpha}} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_{\alpha}} \right)^2 \right] \varepsilon^2(t) dt = 0, \tag{14.15}$$

and, if one of the expressions (14.13) had been nonzero at any value t inside the interval $[t_0, t_1]$, then the integral (14.13) would be also positive and nonzero. Thus, the Lagrange equations are equivalent to the requirement

$$\delta W = 0 \qquad\qquad (14.16)$$

for any variations of the functions $q_\alpha(t)$, for which their values at $t = t_0$ and $t = t_1$ do not vary.

This is a very desired variation statement for the principle to choose the actual motion among all imaginary possible. The actual motion from the initial point $q_\alpha = q_\alpha^0$ to the terminal point $q_\alpha = q_\alpha^1$ during the interval $[t_0, t_1]$ is distinguished among all imaginary by the action variation for it being equal to zero.

It can also be shown that for the actual motion not only the first variation is to be zero, but the action reaches a minimum. However, what we have learned is sufficient for us; namely, we came to the motion law formulation independent of the variables chosen and to the fact how we describe the law. The invariance of the Lagrange notation for motion equations just follows from this independence.

Meanwhile, we have found out that as a mathematical model there may be utilized not only differential equations for motion but also, perhaps, more general variation principles. These principles are more general, since they do not depend upon a coordinate description. They are as if above it. Now we are repeating what this variation mathematical model is.

A state space does not constitute its basis, but the basis now is a space of positions, configurations, a space of variables $q_1, ..., q_n$. For a mechanical system, they will be the positions of all its material points. For electrical systems, considered above, they will be electric charges whose time derivatives are electric currents.

A motion under the variation principle is defined not through its initial state, but through its initial and final positions. In any case, they are supposed to be, as it were, given or fixed. And the variation principle itself is that an actual motion from a given initial point to a final one is distinguished by its action variation being equal to zero.

We may look at it as some mathematical technique for a new notation for the differential equations of motions, or as a new principle. This very principle is called the least action principle a discovery of which belongs to Helmholtz . This is a principle of some expediency, and in its idea it is different from the principle of determinism. Generally speaking, the determinism principle does not follow from the expediency principle, but specifically, when applied to laws of nature, this retains so always. It is this very fact that makes both points of view consistent, complementary to each other.

Now I wish to give you some examples of how to use the Lagrange-Maxwell equations. They will convince you in their efficiency and also even in the cases when you do not know how to approach to constructing a mathematical model.

As a first example, let us take a usual circular pendulum. Its kinetic energy is equal to $J\dot{\varphi}^2 / 2$ and the potential energy is $-mgl\cos\varphi$, where φ is an angle of deviation from a vertical line, J and m are an inertia moment and a mass,

respectively, and l is a distance between the point of suspension and the mass centre. Thus, the Lagrange function is equal to

$$L = \frac{J\dot{\varphi}^2}{2} + mgl\cos\varphi \,.$$

It is then easy to find that

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\varphi}} - \frac{\partial L}{\partial \varphi} = J\ddot{\varphi} + mgl\sin\varphi = 0\,.$$

Now let a pendulum be not circular but parabolic, i.e., the mass m can move along the parabola $z = ax^2$ placed in the vertical plane (Fig. 14.1).
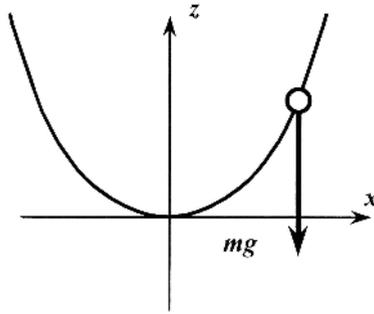
Fig. 14.1. A parabolic mathematical pendulum.

In this case,

$$T = \frac{mv^2}{2} = \frac{m(\dot{x}^2 + \dot{y}^2)}{2} = \frac{m}{2}(1 + 4a^2x^2)\dot{x}^2\,,$$

$$V = mgz = mgax^2$$

and

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = m(1 + 4a^2x^2)\ddot{x} + 2mgax = 0\,.$$

Finally, let a material point of mass m be moving in the vertical plane Oxz (Fig. 14.2) along an arbitrary curve whose equation in a parametric form will look like

$$x = x(s), \quad z = z(s),$$

where the parameter s is a curve arc length being counted from some point O. We find immediately that

$$L = \frac{m\dot{s}^2}{2} + mgz(s),$$

and, therefore,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{s}} - \frac{\partial L}{\partial s} = m\ddot{s} + mgz'(s) = 0.$$
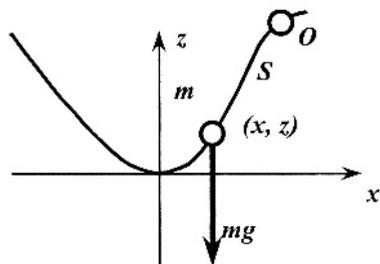
Fig. 14.2. An illustration of a material point motion along an arbitrary smooth plane curve in the gravity field.

Let us take an electric circuit consisting of the self-inductance L and the capacitance C. At this,

$$L = \frac{L\dot{q}^2}{2} - \frac{q^2}{2C}$$

and, hence,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = L\ddot{q} + \frac{q}{C} = 0,$$

that is well known for us.

Now, let us complicate the task by supposing that there is a ferrous stick of mass m suspended by means of a spring inside a self-inductor. Since the stick can swing, the self-inductance L will depend upon the stick displacement x from its equilibrium position $x = 0$ (Fig.14.3).
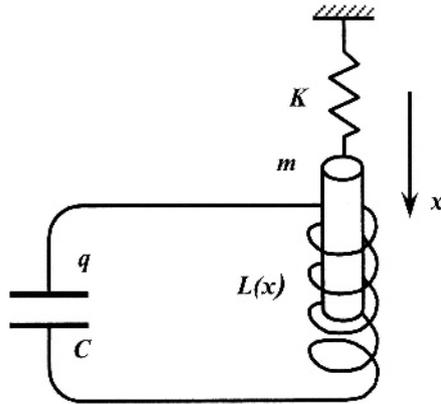
Fig. 14.3. The interacting mechanical and electrical oscillators.

This system is rather complicated. It has two oscillators (a closed loop of the self-inductor and the capacitor, and a spring-attached mass) which interact. A motion of the stick causes changes in the self-inductor magnetic field inducing in it the electromotive force (emf) , and a varying magnetic field of the self-inductor coil acts upon the ferrous core. In order to write the motion equations for these oscillators one needs to find both the emf inducted in the coil by the moving stick and the varying force with which the magnetic field of the self-inductor coil acts upon the ferrous core placed in it. This is not so simple for a mathematician. Whereas to write the Lagrange-Maxwell function is very easy.

Besides the coordinate x, let us introduce also the capacitor charge q. Then we obtain

$$L = \frac{L(x)\dot{q}^2}{2} + \frac{m\dot{x}^2}{2} - \frac{q^2}{2C} - \frac{kx^2}{2}$$

and come to the two equations for the coordinates x and q as functions of time

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = m\ddot{x} - \frac{1}{2}L'(x)\dot{q}^2 + kx = 0,$$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = L(x)\ddot{q} - L'(x)\dot{x}\dot{q} + \frac{q}{C} = 0.$$

As it had to be expected, from these equations there is obtained the energy conservation law of the form

$$\frac{L(x)\dot{q}^2}{2} + \frac{m\dot{x}^2}{2} + \frac{q^2}{2C} + \frac{kx^2}{2} = h = const.$$

This expression is obtained through multiplying the first equation by $\dot{x}$, the second equation by $\dot{q}$, and, then, by adding them and integrating.

Let us rewrite the above Lagrange-Maxwell equations in the form

$$m\ddot{x} + kx = \frac{1}{2}L'\dot{q}^2,$$

$$L\ddot{q} + \frac{q}{C} = -L'\dot{x}\dot{q}.$$

When $L' = 0$, these are decomposed into two independent oscillators, mechanical and electrical. Their interaction is the stronger, the greater L' is. For $L' \neq 0$, the electrical oscillator will act upon the mass of the mechanical oscillator with the force $\frac{1}{2}L'\dot{q}^2$, and the mechanical oscillator will induce in the electrical circuit the emf equal to $-L'\dot{x}\dot{q}$.

Perhaps, you have already noticed that in all above-mentioned examples mechanical and electrical resistors are absent. This is so, because the Lagrange equations presuppose the acting forces to be potential. With the potential function being independent of time, the energy conservation law, as was discussed in the above example, follows from this presupposition.

If forces are not potential and, in particular, there are viscous frictions and ohmic resistors available, then the energy does not retain and the Lagrange equations become not applicable. However, they can be generalized. This is not difficult to do. One will need only to proceed not from the potential energy of the acting forces and the energy of electric field, but from the expression for the elementary work δA. With V being a potential energy, the elementary work δA will be equal to

$$\delta A = -\delta V = -\sum \frac{\partial V}{\partial q_\alpha}\delta q_\alpha. \tag{14.17}$$

In a general case or, more exactly, in a more general case, we have

$$\delta A = \sum Q_\alpha \delta q_\alpha ,\qquad (14.18)$$

i.e. the derivatives $-\dfrac{\partial V}{\partial q_\alpha}$ are replaced by the so-called generalized forces Q_α.
The generalized Lagrange equation turns out to be reduced to this formal replacement. Let us write the Lagrange equations in the form

$$\frac{d}{dt}\frac{\partial T}{\partial \dot q_\alpha} - \frac{\partial T}{\partial q_\alpha} = -\frac{\partial V}{\partial q_\alpha},$$

and now their generalization will be as follows

$$\frac{d}{dt}\frac{\partial T}{\partial \dot q_\alpha} - \frac{\partial T}{\partial q_\alpha} = Q_\alpha .\qquad (14.19)$$

Note also that by separating potential and nonpotential forces , the Lagrange equation can be written as

$$\frac{d}{dt}\frac{\partial T}{\partial \dot q_\alpha} - \frac{\partial T}{\partial q_\alpha} = -\frac{\partial V}{\partial q_\alpha} + Q_\alpha$$

or

$$\frac{d}{dt}\frac{\partial L}{\partial \dot q_\alpha} - \frac{\partial L}{\partial q_\alpha} = Q_\alpha ,\qquad (14.20)$$

where, as before, $L = T - V$, but V relates to the potential forces only and nonpotential forces are represented by the generalized forces Q_α.

Let us illustrate, at first, the equations (14.19) and (14.20) with a very simple example of a spring-attached mass in the presence of viscous friction. In this case the kinetic energy T is equal to $m\dot x^2 / 2$. At the virtual displacement of x there exist two forces performing work, the elasticity force $-kx$ and the friction force $-h\dot x$. Their work for the displacement δx is equal to

$$\delta A = -kx\,\delta x - h\dot x\,\delta x ,$$

so the generalized force corresponding to the coordinate x (simply, the force in our case) is equal to

$$Q = -kx - h\dot x .$$

The Lagrange equation (14.19) takes the form

$$m\ddot{x} = -kx - h\dot{x}.$$

When constructing motion equations in the form (14.20), the potential elasticity force should be referred to the Lagrange function L equal to

$$L = \frac{m\dot{x}^2}{2} - \frac{kx^2}{2},$$

and the viscous friction generalized force Q will be $-h\dot{x}$. By virtue of this, the equation (14.20) will be of the form

$$\frac{d}{dt}\frac{\partial}{\partial \dot{x}}\left(\frac{m\dot{x}^2}{2} - \frac{kx^2}{2}\right) - \frac{\partial}{\partial x}\left(\frac{m\dot{x}^2}{2} - \frac{kx^2}{2}\right) = -h\dot{x}$$

or

$$m\ddot{x} + kx = -h\dot{x}.$$

For an electric circuit with a self-inductor, a capacitor and an ohmic resistor, all will be similar, since the work of dissipative forces of the ohmic resistor in transferring the charge δq is equal to $-R\dot{q}\,\delta q$.

15 Galileo-Huygens clock

How and why did the Galileo-Huygens clock appear? What is fundamentally new in it? What characterizes its accuracy? Analysis of errors and ways to eliminate them. A simplest mathematical model of the Galileo-Huygens clock. A phase portrait. The Poincare mapping and the Keniks-Lamerey point mapping diagram. A clock as an auto-oscillating system and as a feedback system.

Computing time is the most ancient problem of humanity. Millions of years, centuries, years, months, days, hours, minutes, a second and its shortest portions, nanoseconds are those various scales, in which a man has learned to measure time for his different purposes. Initially, time measurements were based on the Earth's revolutions about the Sun and on its own revolutions. The signs of the zodiac were used for indicating year seasons, and the Sun in the daytime and stars at night for indicating time.

A need of locating ships, caused by sea navigation, required exact time measurements. A latitude used to be determined through astronomic observations only. For measuring a longitude, the exact knowledge of time was required as well. To determine a longitude for a position with up to a one-kilometer accuracy is equivalent to a three-angle second error. But people had to sail without being able to correct ship's clock indications for months and even longer.

In the XVII-th century, the UK Admiralty announced a competition for inventing a high-precision clock. Such a clock was designed and manufactured by X. Huygens. It was a first pendulum clock where there were used pendulum oscillations for countdown. Pendulum oscillations were studied by G. Galileo who revealed an invariability of their period (Galileo was not able to observe frequency variations, since pendulum deviations were small). It is on this basis that Huygens created his clock. Therefore, this new clock of his may be called the Galileo-Huygens clock. Before Galileo and Huygens, there existed a clock with a balance-wheel, however the balance did not possess its own oscillating frequency; its oscillating frequency was made dependent upon and determined by the actions of the clock mechanism upon the balance.

A use of the device with its own oscillating frequency turned out to be exceptionally fruitful for clocks and resulted in a series of more and more precise clocks of various constructions. On the basis of this principle, there were produced the following types of clocks: a spring-attached balance clock with cylindrical, anchor, and chronometer movements; a quartz-crystal clock using own elastic

vibrations of a quartz plate; a clock employing an electrical oscillatory circuit with its own frequency; and finally, a phenomenally accurate molecular and atomic clock enabling to reveal and measure the irregularity of the Earth's rotation.

In order to turn an oscillating pendulum into a clock, one needs to count its oscillations and indicate their number, for example, with the help of a hand on a clock-face in a corresponding scale. Besides, its oscillations have to be sustained, for they are damping themselves. Thus, there arises a necessity in converting damping oscillations to undamping auto-oscillations.
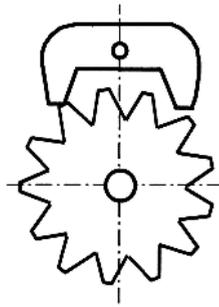
Fig. 15.1. A schematic representation of a travelling gear in a clock.

These both problems are successfully solved with a use of a so-called clock movement, whose simplest variant is diagrammatically shown in figure 15.1. After each pendulum oscillation there and back, a ratchet moves one tooth ahead under the action of a wound-up spring or of a falling weight; simultaneously it imparts to the pendulum a pushing impulse. In this way, a rotating velocity of the ratchet wheel is determined by the pendulum oscillating frequency, and at the moment, when the wheel is performing a turn, its teeth are pushing the pendulum, thus sustaining its oscillations.

A running accuracy of the clock equipped with such a device is determined by a stability of pendulum oscillations. An oscillation period of the pendulum depends upon many reasons, and each of them can entail its changes. The crucial factors determinative to the oscillation period are its length, mass, inertia moment, medium resistance and friction in the suspension point, acceleration of gravity, impacts of the ratchet-wheel , and also the amplitude of pendulum oscillations.

Consider at first a dependence of the pendulum oscillation period upon its amplitude. Let us write the equations of pendulum oscillations in the form

$$J\ddot{\varphi} + mgl \sin \varphi = 0,$$ (15.1)

where J is its moment of inertia, m and l are a mass and a length (the distance from the suspension axis to the centre of gravity). By denoting $\omega^2 = mgl / J$ and multiplying the equation (15.1) by $\dot\varphi$ and integrating, we find that

$$\frac{\dot\varphi^2}{2} - \omega^2 \cos\varphi = h, \qquad (15.2)$$

where h is a constant of integration. Assuming an angular amplitude of the pendulum to be equal to α, we obtain $\dot\varphi = 0$ when $\varphi = \alpha$. Let us express, in accordance with (15.2), the integration constant through α and write (15.2) as

$$\frac{\dot\varphi^2}{2} - \omega^2 \cos\varphi = -\omega^2 \cos\alpha$$

or

$$\dot\varphi = \omega\sqrt{2(\cos\varphi - \cos\alpha)}. \qquad (15.3)$$

From the equation (15.3) it follows that

$$\frac{d\varphi}{\omega\sqrt{2(\cos\varphi - \cos\alpha)}} = dt. \qquad (15.4)$$

Let the oscillation period be T at the amplitude α. Then, with the pendulum deviation angle T varying from zero to α, the clock will run a quarter-period, and according to (15.4), we shall get

$$\frac{T}{4} = \int_0^\alpha \frac{d\varphi}{\omega\sqrt{2(\cos\varphi - \cos\alpha)}}.$$

This integral is not computed in elementary functions. It can be expressed numerically. For small α, being suitable for us, we may take the Taylor-series expansion in the parameter α. With accuracy up to α^2, this representation will be as follows

$$[2(\cos\varphi - \cos\alpha)]^{-1/2} =$$

$$= \{2[(1 - \frac{\varphi^2}{2} + \frac{\varphi^4}{24} - ...) - (1 - \frac{\alpha^2}{2} + \frac{\alpha^4}{24} - ...)]\}^{-1/2} =$$

$$= (\alpha^2 - \varphi^2)^{-1/2}(1 - \frac{\varphi^2 + \alpha^2}{12} + ...)^{-1/2} =$$

$$= (\alpha^2 - \varphi^2)^{-1/2}(1 + \frac{\varphi^2 + \alpha^2}{24} + ...)$$

and, hence,

$$T = \frac{4}{\omega}\int_0^\alpha (1 + \frac{\varphi^2 + \alpha^2}{24} + ...)\frac{d\varphi}{\sqrt{\alpha^2 - \varphi^2}}$$

or, after the substitution of $\varphi = \alpha \sin \theta$,

$$T = \frac{4}{\omega}\int_0^{\pi/2}(1 + \frac{\alpha^2(1 + \sin^2 \theta)}{24} + ...)d\theta =$$

$$= \frac{2\pi}{\omega}(1 + \frac{\alpha^2}{16} + ...) . \tag{15.5}$$

As for the accuracy of this formula, it can be seen from the plot in figure 15.2, where the continuous curve represents the values calculated by the formula placed in brackets in the right-hand side of (15.5), i.e. $\omega T / 2\pi$, and the broken curve represents the plot of the first two terms in these brackets, i.e. $1 + \alpha^2 / 16$. From this figure it is seen that these plots are indistinguishable for $\alpha < \pi / 2$.
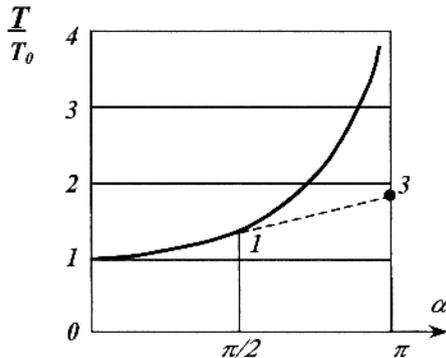
Fig.15.2. A dependence of the pendulum oscillating period T upon its amplitude.

From the formula (15.5) itself, it is seen that for small amplitudes α, the period of the pendulum oscillations is changing a little. For example, for $\alpha \leq 0.1$, it changes not more than 0.7%. Meanwhile, with the amplitude α changing by $\Delta\alpha$, the period T will change as below

$$\frac{dT}{T} \approx \frac{\alpha\Delta\alpha}{8}.$$

For $\alpha = 0.1$ and for the desired accuracy $dT/T = 10^{-5}$, $\Delta\alpha$ should not exceed 4×10^{-4}, i.e. a large invariance of amplitude is needed. In this connection, people used to ponder upon the problem of whether it is possible to design a pendulum of the period being independent upon its amplitude. Such a pendulum was devised. Its idea lay in the fact that the pendulum was made swinging not along a circle but along some curve, and this can be attained in the way as shown in figure 15.3, where Π is a flexible plate, and H are the guides restricting its motions. Here, the line L directed along the pendulum suspension touches all the time the guide H.
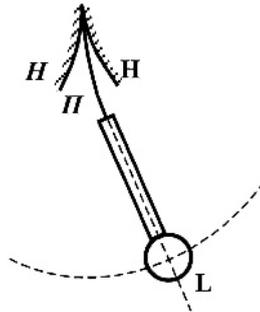
Fig. 15.3. The guide restriction H allowing to change a trajectory of pendulum oscillations.

Let the desired curve, along which the pendulum (a material point of the mass m) is to swing, be described by the equation

$$x = x(s), \; y = y(s), \tag{15.6}$$

where s is an arch length being counted from the bottom point of the curve (Fig.15.4).

Let us write the motion equation of the mass m. The Lagrange function is

$$L = \frac{m}{2}\dot{s}^2 - mgy(s) .$$

Hence, we have

$$m\ddot{s} + mgy'(s) = 0 . \qquad (15.7)$$
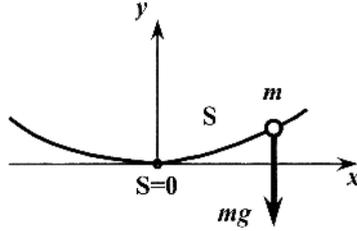
Fig. 15.4. An illustration of a search for a curve along which the pendulum must oscillate so as to have its oscillation period not dependent upon its amplitude.

The solution to this equation will be periodic, with a period being independent upon initial conditions, and, therefore, independent upon its amplitude, if this equation coincides with the linear harmonic oscillator equation, i.e. when

$$gy'(s) = \omega^2 s \qquad (15.8)$$

with any ω. Solving this equation, we find

$$y = \frac{\omega^2}{2g}s^2 , \qquad (15.9)$$

where any constant is absent, since according to figure 15.3, we have $y = 0$ for $s = 0$.

It remains to find the function $x(s)$. This is easy to make starting from the expression

$$dx^2 + dy^2 = ds^2$$

or

$$(\frac{dx}{ds})^2 + (\frac{dy}{ds})^2 = 1 . \qquad (15.10)$$

From (15.9) and (15.10) it follows that

$$\frac{dx}{ds} = \sqrt{1 - (\frac{\omega^2}{2g})^2 s^2} \qquad (15.11)$$

or

$$x = \int \sqrt{1 - (\frac{\omega^2}{2g})^2 s^2}\, ds = \int \sqrt{1 - a^2 s^2}\, ds . \qquad (15.12)$$

By using the substitution $s = a^{-1} \sin \varphi$, we take the integral (15.12), and then, taking into account (15.9), come to the following parametric equations of the curve desired

$$x = \frac{g}{2\omega^2} \varphi + \frac{g}{4\omega^2} \sin 2\varphi, \qquad y = \frac{2g}{\omega^2} \sin^2 \varphi . \qquad (15.13)$$

This is a well-known cycloid, and its form is shown in figure 15.5. The cycloid differs from a circle in that that it is ascending from the point $x = y = 0$ faster and faster than the circle, and hereby rolling the pendulum down it is accelerated as its deviation is increasing. For small φ, this distinction is, naturally, small, as small also is the difference between the circle pendulum oscillation equation and that of a linear harmonic oscillator, i.e. similar to the small distinction between $\sin \varphi$ and φ.
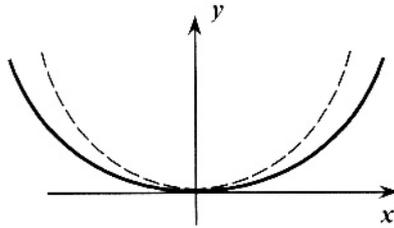
Fig. 15.5. A comparison of motion trajectories for circular and cycloidal (broken curve) pendulums.

Thus, we have considered a dependence of the oscillation period of the pendulum upon its oscillation amplitude and have also estimated the influence of amplitude variations upon running accuracy of the clock. It was revealed that this dependence can be escaped and the possible reason of lessening the clock running accuracy can be eliminated, if a circle pendulum is replaced by a cycloid one. However, a change of the amplitude is far from being a unique reason affecting

the running accuracy of the clock. The pendulum length can vary, and substantial displacements of the clock can result in a change of gravity acceleration. In our considerations, we have omitted the inevitable presence of viscous and dry frictions. The instability of the period can be also caused by a pushing mechanism responsible for sustaining pendulum oscillations.

Thus, instability and a loss of accuracy can be brought about by many reasons. If we are wishing to design a high-precision clock, we need to have all these instability factors investigated, estimate their magnitudes and find the ways to eliminate them.

By virtue of their smallness, all these instability factors can be investigated separately and independently upon each other, taking into account not all of them together, but each individually. This large simplification assumed can be substantiated as follows. Since the oscillation period T is a function of many parameters $u_1, u_2, ..., u_m$ of a clock mechanism , we obtain

$$T = T(u_1, u_2, ..., u_m) .$$

Let $u_1^*, u_2^*, ..., u_m^*$ be the nominal values of the parameters and $\delta u_1, \delta u_2, ..., \delta u_m$ be their possible variations. Then, with accuracy of up to these smallest changes, the below

$$\delta T = (\frac{\partial T}{\partial u_1})^* \delta u_1 + (\frac{\partial T}{\partial u_2})^* \delta u_2 + ... + (\frac{\partial T}{\partial u_m})^* \delta u_m$$

will be true. Here, the derivatives of the function T are calculated for the nominal values of the parameters, i.e. the derivative $\dfrac{\partial T}{\partial u_s}$ is calculated by assuming it independent upon the rest parameters $u_1, u_2, ..., u_m$ except u_s.

In view of this, the dependence upon a length, a viscous friction and a gravity acceleration can be calculated   within the assumed accuracy through making use of the simplified pendulum oscillation model, which does not consider the dependence of pendulum period upon the amplitude, i.e. from the linear oscillator equation

$$\ddot{\varphi} + 2\delta\dot{\varphi} + \omega^2\varphi = 0 , \quad (\omega^2 = \frac{mgl}{J}) \tag{15.14}$$

with the oscillation period being equal to

$$T = \frac{2\pi}{\sqrt{\omega^2 - \delta}} . \tag{15.15}$$

For the sake of simplicity, a basic mass of the pendulum is assumed to be concentrated in a single place. Then, $J = ml^2$ and (15.15) will be of the form

$$T = \frac{2\pi}{\sqrt{\dfrac{g}{l} - \delta^2}} = 2\pi\sqrt{\frac{l}{g}}(1 + \frac{1}{2}\frac{l\delta^2}{g} + \ldots),$$

from which the following estimate can be found

$$\frac{dT}{T} = \frac{1}{2}\frac{dl}{l} - \frac{1}{2}dg + \frac{l\delta d\delta}{g}. \tag{15.16}$$

These estimates demonstrate a strong dependence of the movement accuracy upon changes in length, gravity and in viscous friction as well, if the latter is not very small.

In general, the pendulum length is affected by temperature changes. Therefore, this movement accuracy loss can be eliminated through sustaining the surrounding temperature constant. However, there exists another way of sustaining l constant in spite of temperature changes. This way, as shown in figure 15.6, lies in a special design of the pendulum. With the temperature being increased, the lengths l_1 and l_2 are increasing by δl_1 and δl_2, and meanwhile the pendulum length is increasing by $\delta l_1 - \delta l_2$. The rods of lengths l_1 and l_2 are made of different materials of the thermal expansion coefficients α_1 and α_2, and these coefficients are chosen so as to nullify the extension magnitude equal to

$$\delta l = \delta l_1 - \delta l_2 = (\alpha_1 l_1 - \alpha_2 l_2)\delta T$$

with δT being a temperature change, i.e. to choose these coefficients so that to obtain $\alpha_1 l_1 - \alpha_2 l_2 = 0$.
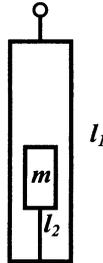
Fig. 15.6. The structural scheme of a suspension for a pendulum of the length independent upon temperature.

For the clock with a fast change of its position, i.e. when, say, being mounted on a fast sailing ship, it seems impossible to kill this dependence upon the gravity acceleration. One can only compensate this dependence through introducing associated computational corrections, or replace the clock pendulum with a spring-attached balance. This very technique came soon into being, the Huygens clock was replaced with a ship chronometer.

Also we should estimate the influence imposed by dry friction. According to the above, it is enough to consider the model

$$\ddot{\varphi} + \omega^2 \varphi = -f \, sign\dot{\varphi} \,, \qquad (15.17)$$

where f is a parameter informing about the availability of Coulomb friction. As we see below, dry friction leads to decreasing oscillations, but their period is not changing. Indeed, at each half-cycle of oscillations, the magnitude $\dot{\varphi}$ does not change its sign and hence, for example, for $\dot{\varphi} > 0$, the equation (15.17) is written as

$$\ddot{\varphi} + \omega^2 \varphi = -f$$

or, upon replacing φ by $\varphi - f / \omega^2$, as follows

$$\ddot{\varphi} + \omega^2 \varphi = 0 \,,$$

where the dependence upon friction f vanished, i.e. it is not there.

Let us now turn to considering a technique of how to sustain pendulum oscillations and how this technique affects the stability of clock running.

Clock drivings are very different in construction (cylindrical, anchor, chronometer and others) and their exact description is rather difficult. Let us assume an idealized scheme of a clock driving by thinking that this drive gives a jog at the angle $\varphi = \alpha$ $(\dot{\varphi} > 0)$ supplying the impulse p, and beyond this pushing impulse a motion of pendulum will follow the equation

$$\ddot{\varphi} + 2\delta\dot{\varphi} + \omega^2 \varphi = 0 \,. \qquad (15.18)$$

The jog occurs at $\varphi = \alpha$ and $\dot{\varphi} > 0$, and it entails an instant increase of pendulum velocity $\dot{\varphi}$ by the quantity p, and, therefore, we get

$$\dot{\varphi}_+ = \dot{\varphi}_- + p \,, \qquad (15.19)$$

where $\dot{\varphi}_-$ is a prejog angle velocity, and $\dot{\varphi}_+$ is a postjog velocity. During the impact, the angle φ has no time to vary. Let us draw a phase trajectory representing a single oscillation from the point $\varphi = \alpha$ to the same point $\varphi = \alpha$ again. Let M_0 be an initial point and M_1 be the reiterated point, where $\varphi = \alpha$,

provided that the motion is described by the differential equation (15.18). At the instant when the pendulum arrives to the point M_1 there occurs a transfer of the impact p, and the point M_1 will displace along the axis $\dot{\varphi}$ at the distance p (Fig.15.7) to the point M_2.
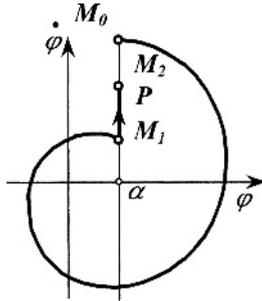
Fig.15. 7. A phase trajectory of the clock and a technique to study it with the help of a point mapping of the point M_0 to the point M_2.

After this, the entire cycle described will be repeated (except the case when the phase trajectory outcoming from the point M_2 will not cross the line $\varphi = \alpha$ any more).

Let us analytically express the dependence of the described motion first from the phase point M_0 to M_1 and then to M_2. Suppose that the points M_0, M_1, M_2 have the coordinates $\varphi_0 = \alpha$ and $\dot{\varphi}_0$, $\varphi_1 = \alpha$ and $\dot{\varphi}_1$, $\varphi_2 = \alpha$ and $\dot{\varphi}_2$. According to the equation (15.7), the points M_0 and M_1 are mutually dependent as

$$\varphi_1 = e^{-\delta\tau}(\varphi_0 \cos\Omega\tau + \frac{\dot{\varphi}_0 + \delta\varphi_0}{\Omega}\sin\Omega\tau)$$

$$\dot{\varphi}_1 = e^{-\delta\tau}[-\delta\varphi_0 \cos\Omega\tau + \frac{\dot{\varphi}_0 + \delta\varphi_0}{\Omega}\sin\Omega\tau) -$$

$$- \Omega\varphi_0 \sin\Omega\tau + (\dot{\varphi}_0 + \delta\varphi_0)\cos\Omega\tau], \qquad (15.20)$$

where τ is a time of motion from the point M_0 to the point M_1, and $\Omega = \sqrt{\omega^2 - \delta^2}$. According to (15.19), for the points M_1 and M_2, we have

$$\varphi_2 = \varphi_1, \ \dot{\varphi}_2 = \dot{\varphi}_1 + p. \qquad (15.21)$$

Besides, we should remember that

$$\varphi_0 = \varphi_1 = \varphi_2 = \alpha. \qquad (15.22)$$

Denoting $\dot{\varphi}_0 = u$ and $\dot{\varphi}_2 = \overline{u}$, from (15.20)-(15.22) it follows that

$$\alpha = e^{-\delta\tau}(\alpha\cos\Omega\tau + \frac{u + \delta\alpha}{\Omega}\sin\Omega\tau)$$

$$(15.23)$$

$$\overline{u} = p + e^{-\delta\tau}(u\cos\Omega\tau + \frac{u + \delta\alpha - \Omega^2\alpha}{\Omega}\sin\Omega\tau).$$

The first equation in (15.23) determines the time τ as a function of u and the second one allows to find $\overline{u}$.

Analysis of these equations is rather complicated, but for $\alpha = 0$ it is substantially simplified, since in this case the formulas (15.23) will be of the form

$$\sin\Omega\tau = 0 \quad (u \neq 0), \quad \overline{u} = e^{-\delta\tau}u\cos\Omega\tau + p,$$

from which it follows that $\tau = 2\pi/\Omega$ (the positive root of second value to the first equation was taken, because for the another root we have $\dot{\varphi}_1 < 0$, which also is seen from figure (15.7)) and that

$$\overline{u} = e^{-2\pi\delta/\Omega}u + p. \qquad (15.24).$$

Prior to analysing this surprisingly simple result, let us show that it is this very result that is worthy of our attention first of all. The matter is that it is at $\alpha = 0$ that the dependence of the oscillation period τ upon the velocity u vanishes. The period is constant and equal to $2\pi/\Omega$. On the contrary, for $\alpha \neq 0$ such a dependence exists and follows from the equation (15.23). Indeed, differentiating the first equation (15.23) in u, we find that

$$0 = -e^{-\delta\tau}\frac{d\tau}{du}(\alpha\cos\Omega\tau + \frac{u + \delta\alpha}{\Omega}\sin\Omega\tau) +$$

$$+ e^{-\delta\tau}[-\alpha\Omega\sin\Omega\tau\frac{d\tau}{du} + \frac{1}{\Omega}\sin\Omega\tau + (u + \delta\alpha)\cos\Omega\tau\frac{d\tau}{du}].$$

Consequently, we have

$$\frac{d\tau}{du} = \frac{\sin\Omega\tau}{\Omega[(\alpha - u - \delta\alpha)\cos\Omega\tau + (\frac{u + \delta\alpha}{\Omega} + \alpha\Omega)\sin\Omega\tau]}$$

and, hence, $d\tau/du$ does not vanish, since τ is different from $2\pi/\Omega$.

This very fact can be revealed immediately from the form of the phase trajectory in figure 15.7. From this figure it is seen that $\tau > 2\pi/\Omega$, since we have there $\alpha > 0$. On the contrary, for $\alpha < 0$ it is seen that $\tau < 2\pi/\Omega$. The said becomes obvious, if one observes that any phase trajectory crosses any ray $\dot{\varphi} = \alpha\varphi$ repeatedly after one and the same time $2\pi/\Omega$.

Let us return to the earlier found relation (15.24) that determines a connection between sequential angle velocities at which the pendulum passes upon jogs the vertical equilibrium position. This relation, i.e. this connection between u and $\overline{u}$, can be supplied with a very simple geometrical interpretation that will be useful in many cases. Let us take the Cartesian coordinate system, by placing u and $\overline{u}$ on its axes, and draw the plot of the mapping (15.24) (Fig.15.8). Now, let us supplement it with a bisector and determine $\overline{u}$ for the given u. Then transfer it to the axis u again, as shown in figure 15.8. Repeating this way of construction, we will consecutively find $\overline{u}, \overline{\overline{u}}$, etc. From this figure it is seen that the sequential points $u, \overline{u}, \overline{\overline{u}}$,... are approaching the point u^*, corresponding to the intersection of the curve (15.24) with the bisection. The value u^* itself passes into itself since $\overline{u}^* = u^*$ and, hence, according to (15.24), we obtain

$$u^* = e^{-2\pi\delta/\Omega}u^* + p$$

and

$$u^* = \frac{p}{1 - e^{-2\pi\delta/\Omega}} . \tag{15.25}$$

Thus, with the pushing impacts supplying the equilibrium-passing pendulum with a constant impulse, its oscillations will asymptotically approach the periodic oscillations of the period $2\pi/\Omega$ and of the velocity amplitude equal to u^*. Thereby, the damping pendulum oscillations are converted into auto-oscillations. Such a conversion is achieved through regulating the pendulum motions by a feedback, i.e., more precisely, through a clock movement that implements the feedback. A general idea of a control and a feedback is of fundamental importance in science and engineering. We shall encounter it many times, and right now, let

us restrict ourselves to the explanation of what control itself is and why it is implemented via a feedback.
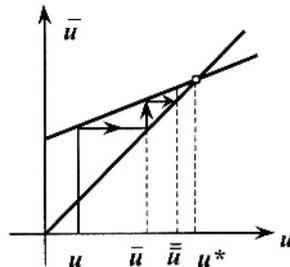
Fig. 15.8. The Keniks-Lamerey diagram for the Galileo-Huygens clock ; u^* is a stable fixed point corresponding to the auto-oscillations in the clock; $u, \overline{u}, \overline{\overline{u}}$,... are successive approximations to the point u^* .

In our case, regulating a pendulum implies the fact that the pendulum is acted upon by impacts coming from some exterior source of energy. These impacts are imposed upon the pendulum when it is passing through its equilibrium position, i.e. an influence upon the pendulum is exerted depending upon the motion of the pendulum itself. The pendulum is acting , as it were, upon itself implementing a feedback and exploiting at this instant an exterior energy of the main spring or a falling bob.

Generally speaking, control exists in nature and is used by a human being in order to implement somewhat purposes. In many cases it is applied to correct the things that are happening with our object under control, i.e. control is executed with the help of feedback and some energy source needed for this feedback. Though there exist the cases when control implements a preassigned plan or obeys to somewhat commands from the outside.

In conclusion, some words touching the relationship between the mapping (15.24) and the diagram in figure 15.8.

The technique through which we have come to the relation (15.24) is of a general character. It was invented by J. H. Poincare; and it is widely used in the oscillations theory and the theory of dynamical systems. In its general formulation, it implies that a description and an investigation of dynamic system motions are reduced to a derivation and a study of some point mapping. From this general point of view, the relation (15.24) will be a point mapping of a line into itself that transforms any its point with a coordinate u into another point of the coordinate $\overline{u}$. This very point mapping is constructed with a use of the technique shown in figure 15.8. The diagram of figure 15.8 was first employed by Lamerey, and a line-to-line point mapping was studied by Keniks. In this

connection, this diagram was called the Keniks-Lamerey diagram. This name was given by A.A. Andronov who improved and widely used it in solving a wide range of urgent problems in the oscillations theory and the control theory.

It is enough, perhaps, for a first talk about the Poincare method and point mappings. Further knowledge about it you will get from the appropriate courses on the control theory and the theory of dynamical systems.

16 Generator of electric oscillations

A generator of electric oscillations as an electric analogue for the Galileo-Huygens clock. The Van der Pol equation and its phase portrait. An approximate investigation of the Van der Pol equation and reducing it to a point mapping. A soft excitation of auto-oscillations.

As an electric analogue for a pendulum there serves an oscillating circuit composed of a capacitor and a self-inductor. Similar to the way when pendulum oscillations in the Galileo-Huygens clock are sustained through pushing impacts of the clockwork, the oscillations in an electric circuit can be sustained by means of special devices, an electronic tube or a transistor. These devices should be responsible for pumping up energy in order to compensate its dissipation for unavoidable ohmic resistance. A pushing impact upon a clock pendulum has to be agreed with its phase, and, in this sense, serves as some control. Analogously, sustaining electric circuit oscillations, i.e. a generator of electric oscillations, through an electronic tube or a transistor also carries a regulating nature. An electric scheme for one of such a controllable energy pump-up is shown in figure 16.1.
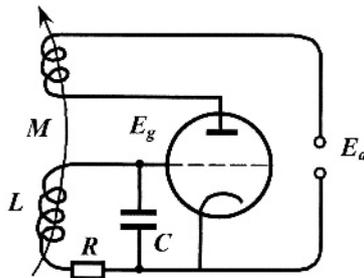
Fig. 16.1. A scheme of a vacuum-tube generator.

Here, an electronic tube, as a component of the scheme, has a cathode, an

anode and a grid. Through a grid voltage, the resistance of the tube between the cathode and the anode can be changed practically without any energy supply. Thus, if some voltage is applied between the cathode and the anode, then the current I_a, running from the cathode to the anode, will be determined by the grid voltage E_g . This dependence is of the form shown in figure 16.2. You probably remember from the secondary school why the plot of the dependence $I_a(E_g)$ has this very shape. This is so because a heated cathode emits electrons, and the voltage E_g on the grid, placed between the cathode and the anode, supports or resists their arrivals at the anode, thus increasing or decreasing the anode current.

Let us write now a mathematical model for the electric scheme presented in figure 16.1. Let I_a be an anode current, E_g be a grid voltage (relative to the cathode), and q be a charge of the capacitor. Also let C, R and L be capacitance, resistance and self-induction, respectively, and M be a coefficient of mutual induction between coils in the anode and grid circuits. For the oscillating circuit consisting of the self-induction L, the capacitance C and the resistance R , we shall have

$$I = \dot{q}, \ E_g = \frac{q}{C},$$

$$(16.1)$$

$$L\frac{dI}{dt} + RI + \frac{q}{C} = M\frac{dI_a}{dt}.$$

In order to go further, the form of the dependence $I_a(E_g)$ should be specified. For this purpose, assume that

$$I_a = \alpha + \beta E_g - \gamma E_g{}^3. \qquad (16.2)$$

In doing so we neglect the capacitance in the anode circuit and assume the anode current to be determined by the formula (16.2) approximating the dependence of figure 16.2. Now we get

$$L\ddot{q} + R\dot{q} + \frac{q}{C} = M(\frac{\beta}{C} - \frac{3\gamma}{C^3}q^2)\dot{q}$$

or

$$\ddot{q} + (\frac{R}{L} - \frac{M\beta}{CL})\dot{q} + \frac{3\lambda}{LC^3}\dot{q}q^2 + \frac{q}{CL} = 0. \qquad (16.3)$$
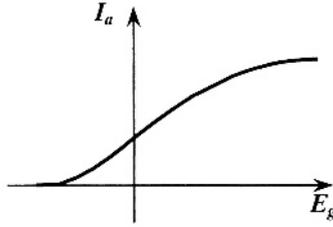
Fig. 16.2. Dependence of the anode current upon the grid voltage in the triode electronic tube.

The equation (16.3) is reduced to the well-known Van der Pol equation

$$\ddot{u} - 2\delta(1 - au^2)\dot{u} + \omega^2 u = 0,\qquad(16.4)$$

where

$$2\delta = \frac{M\beta}{CL} - \frac{R}{L},\quad a = (\frac{M\beta}{CL} - \frac{R}{L})^{-1}\frac{3\gamma}{LC^3},\quad \omega^2 = \frac{1}{LC}.$$

It is assumed here that $\delta > 0$, $a > 0$ and $\omega^2 > 0$. For $a = 0$ and very small u, the Van der Pol equation will turn into a linear oscillator with the negative friction $\delta > 0$. For $u^2 > 1/a$, the coefficient in front of $\dot{u}$ becomes positive, and we can rightfully suppose that in this case the oscillations are damping. Thus, the equilibrium state of the Van der Pol oscillator is unstable and small oscillations are growing, while very large oscillations are damping, and, therefore, between them there should exist a stable periodic motion Γ (Fig. 16.3). One may be persuaded by this truth-like assertion, if he looks at the phase portrait of the equation (16.3) on a computer display . Analytically this fact can be easily proved for a sufficiently small δ, and in doing so we are proving not only an auto-oscillatory nature of this system at $\delta > 0$ (i.e. for a sufficiently large M) but also shall reveal an approximate value of the auto-oscillation amplitude.

Let us write the Van der Pol equation (16.3) in the form

$$\ddot{u} + \omega^2 u = -\varepsilon u^2 \dot{u} + 2\delta\dot{u},\qquad(16.5)$$

where the magnitude

$$\varepsilon = \frac{3\gamma}{LC^3} > 0$$

is assumed to be small, as well as δ. This allows us to write the solution to the equation (16.5) as

$$u = A\cos(\omega t + \varphi) , \qquad (16.6)$$

where, due to smallness of ε and δ, A and φ are slowly changing functions of time. That such a representation is possible does not give rise to any special doubts, but, nevertheless, a request for its specific form will entail some difficulties. These difficulties, first of all, will be caused by the fact that such a representation is defined not uniquely. Though, due to this very lack of uniqueness, we may to the condition (16.6) add another one, for example,

$$\dot{u} = -\omega A\sin(\omega t + \varphi) , \qquad (16.7)$$

whereupon A and φ will be expressed through u and $\dot{u}$. Further, one may be convinced immediately of the derivatives $\dot{A}$ and $\dot{\varphi}$ along the solutions of the equation (16.5) being small in view of smallness of δ and ε.
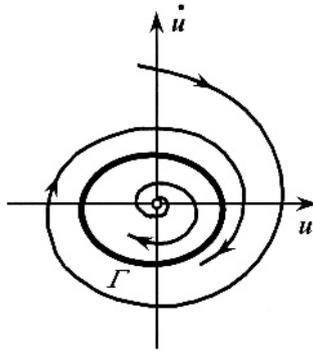
Fig. 16.3. A phase portrait for the Van der Pol equation; Γ is an auto-oscillation, a stable periodic motion.

Consider now how energy of the oscillator relating to the equation (16.5) with $\varepsilon = \delta = 0$ varies with time. Its energy is equal to

$$E = \frac{1}{2}(\dot{u}^2 + \omega^2 u^2), \qquad (16.8)$$

and its derivative, according to (16.5), is of the form

$$\frac{dE}{dt} = \dot{u}\ddot{u} + \omega^2 u\dot{u} = (2\delta - \varepsilon u^2)\dot{u}^{2`},$$

from which, in view of (16.6), (16.7), we get

$$\frac{dE}{dt} = [2\delta - \varepsilon^2 A^2 \cos^2(\omega t + \varphi)]A^2 \omega^2 \sin^2(\omega t + \varphi)). \qquad (16.9)$$

By integrating the equation (16.9) under the same assumptions $\dot{A} = \dot{\varphi} = 0$, we will determine that the change of the energy for the period $2\pi/\omega$ is equal to

$$\Delta E = \frac{4\pi\delta}{\omega} A^2 - \frac{\pi\varepsilon}{\omega} A^4 . \qquad (16.10)$$

On the other hand, according to (16.8) and (16.6), one has $E = \omega^2 A^2 / 2$. Therefore,

$$\Delta E = \frac{\omega^2}{2}\Delta A^2 = 2\pi\delta\omega A^2 - \pi\varepsilon\omega\frac{A^4}{2}$$

or

$$\Delta A^2 = \frac{4\pi\delta}{\omega} A^2 - \frac{\pi\varepsilon}{\omega} A^4 . \qquad (16.11)$$

By denoting through A_n and A_{n+1} the A- amplitude successive values, lying in the interval $2\pi/\omega$ distant from each other , we come, with accordance to (16.11), to the relation

$$A_{n+1}{}^2 = (1 + \frac{4\pi\delta}{\omega})A_n{}^2 - \frac{\pi\varepsilon}{\omega} A_n{}^4 , \qquad (16.12)$$

which can be investigated by means of the already known Keniks-Lamerey diagram. From the form of these diagrams (Fig. 16.4) for $\delta < 0$ and $\delta > 0$ and with $\varepsilon > 0$ being constant, it follows that for $\delta < 0$ (Fig. 16.4a) the amplitude A of the motion is decreasing down to zero, that is, the equilibrium state is globally stable. On the contrary, in the case of our interest ($\delta > 0$) (Fig. 16.4b), the form of the corresponding diagram betrays an instability of the equilibrium and an availability of a globally stable auto-oscillation (a periodic motion). The amplitude A^* of this periodic motion is determined from the condition $A_{n+1} = A_n$ and equal to

$$A^* = 2\sqrt{\frac{\delta}{\varepsilon}} \ .$$ (16.13)
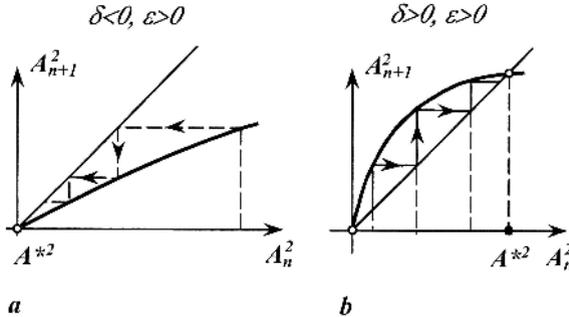
Fig. 16.4. The Keniks-Lamerey diagrams for soft excitation of auto-oscillations as the parameter δ is growing up **a** for $\delta < 0$ and **b** for $\delta > 0$. For $\delta = 0$, from the equilibrium losing its stability, there is born a stable fixed point A^* corresponding to an auto-oscillation.

Under the assumptions done regarding the smallness of δ and ε, a period of the corresponding auto-oscillation is approximately equal to $2\pi / \omega$. For the sake of visualization, let us also construct a bifurcation diagram for the equilibrium and the auto-oscillation born from it when δ changes its sign (see figure 16.5). Here, the points on the abscissa axis correspond to the equilibrium states which are stable for $\delta < 0$ and unstable for $\delta > 0$. A loss of stability occurs at $\delta = 0$ and together with it, when δ is increasing, there appears an auto-oscillation whose amplitude is increasing from zero, i.e. it is being born, so to say, by the equilibrium changing its stability.

All the above conclusions have been derived by us from an approximate form of the point mapping (16.12), assumed to be accurate. Couldn't it lead us to an error? In general this is possible, but only with $\dot{\varphi}$ or $\dot{A}$ being sufficiently large. For $\dot{\varphi}$ or $\dot{A}$ being small, the point mappings in figure 16.4 vary a little and the general conclusions remain, as previously, valid. These conclusions remain valid for sufficiently small $\dot{\varphi}$ or $\dot{A}$, i.e. when δ and ε are sufficiently small. On the

contrary, a change in the form of the anode-grid characteristic $I_a = I_a(E_g)$ can result in considerable consequences. We will consider them, because it will allow us to reveal how auto-oscillations appear, if a equilibrium state is losing its stability, and to mathematically describe, first, the phenomena of a soft and a hard excitations of auto-oscillations and, second, the so-called hysteresis phenomenon.
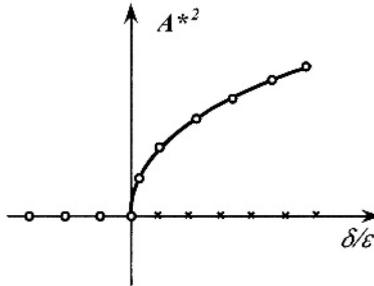
Fig. 16.5. A bifurcation diagram for soft excitation of auto-oscillations. As the parameter δ/ε is growing up and passing through zero, the equilibrium loses its stability and becomes unstable. Simultaneously, there is born an auto-oscillation whose amplitude will grow under a further increase of the magnitude δ/ε.

A soft excitation of auto-oscillations implies a smooth (from zero) appearance of auto-oscillations being born by the equilibrium state losing its stability because of changes in system parameters – this exciting way was described above. This excitation is soft, for it appears smoothly from zero. It is worth to emphasize here that auto-oscillations are born as a result of a change of system parameters.

If system parameters are changed in the inverse direction, the entire picture of auto-oscillation birth is repeated in the inverse order. As an adequate mathematical model for the soft excitation there may be used the sequences of phase portraits or of point mappings corresponding to the bifurcation diagram shown in figure 16.5.

A sequence of point mappings and phase portraits is shown in a series of three figures (Fig. 16.6). The first plot relates to a stable equilibrium; the second corresponds to the situation arising immediately after an instability appears; and the third takes place when further change of the parameter causes a growth of auto-oscillation amplitude. Such a soft excitation of oscillations caused by a change of parameters is observed in real systems, but along with it there exists a quite different way of exciting the auto-oscillations, called a hard excitation. In this case, the picture is changed drastically: auto-oscillations appear by jump and

immediately take a finite and possibly large value. If changed inversely, the picture is not repeated in a reverse fashion. It becomes quite another.
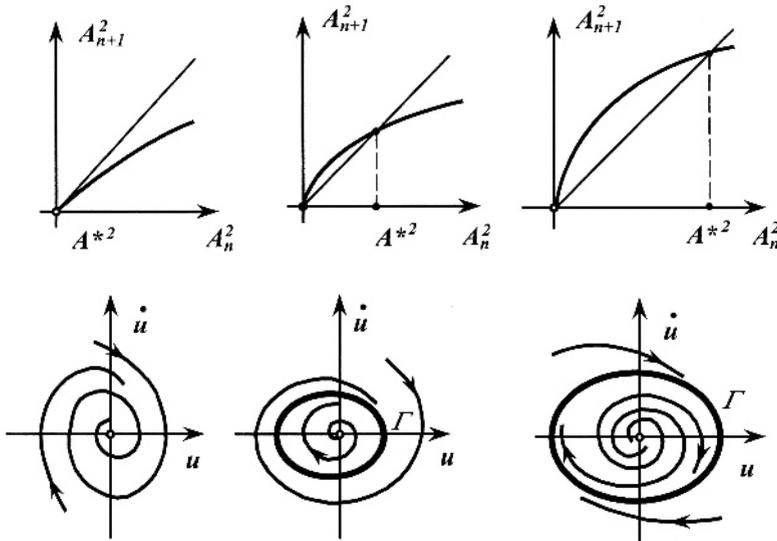
Fig. 16.6. Successive shapes of the Keniks-Lamerey diagrams and the corresponding phase portraits under soft excitation of auto-oscillations with δ / ε growing up.

17 Soft and hard regimes of exciting auto-oscillations

Constructing phase and bifurcation portraits for soft and hard excitations of auto-oscillations on the basis of a point mapping study.

We have already become acquainted with the phenomenon of softly exciting auto-oscillations with the help of the example of a tube generator (this example has played a historical role at the beginning of the previous century) and with the Van der Pol differential equation describing its dynamics. This phenomenon is of a general nature and relates to the bifurcation of bearing a stable periodic motion (i.e. an auto-oscillation) arising from a stability-losing equilibrium state. This bifurcation was named after A.A. Andronov who was the first to have revealed it. You were shown an associated   figure containing   successive stages (picture fragments on a film) of the softly excited auto-oscillations in the two-dimensional phase space. These pictures are suggested to be considered a geometrical mathematical model for this phenomenon, while the Van der Pol equation may be considered to be its simplest specific analytic model.

However, auto-oscillations can be excited by parametric variations not only in a soft but also in a so-called hard way. This process can be observed on the same tube generator, whose anode characteristic $I_a = I_a(E_g)$ is of another form, for example, like

$$I_a = \beta E_g + \gamma E_g{}^3 + \mu E_g{}^5,\tag{17.1}$$

where the absence of even degrees is inessential because they do not reveal themselves at all in our approximate investigations.

We have to repeat the preceding computations for this more general case, when one more term is inserted into the characteristic (17.1). Then we come to a point mapping of the form

$$A_{n+1}{}^2 = (1 + \frac{4\pi\delta}{\omega})A_n{}^2 + \varepsilon_1 A_n{}^4 + \varepsilon_2 A_n{}^6.\tag{17.2}$$

Here, the first two terms coincide with those from the previous case and the third is new. The coefficients ε_1 and ε_2 are proportional to γ and μ, and, by the assumption, they and δ are small. Besides, simulating the anode characteristic via   the polynomial (17.1) is acceptable only within   some finite domain of

changing E_g , and this constraint yields a boundedness of the amplitude A_n in (17.2).

Here, we shall have to investigate the point mapping (17.2), with the above constraints taken into account. If all possible cases must be distinguished by signs of parameters δ, ε_1 and ε_2 , then this   will lead us to eight different variants. For each of them there is given in figure 17.1 a respective plot of point mapping, the Keniks-Lamerey diagram. These diagrams are divided into four couples,   **a, b, c, d**, each of which corresponding to $\delta < 0$ and $\delta > 0$ when ε_1 and ε_2 are identical.

The case **d** conforms to the already known soft excitation of auto-oscillations during a continuous passage from $\delta < 0$ to $\delta > 0$; the case **c** represents a new hard excitation we are interested in. The cases **a** and **b** are rather of a theoretical nature, since they do not correspond to any actual performances of an electronic tube.

Let us focus our attention upon the case **c** representing it explicitly in the form of five sequent plots of the point mapping (17.2) for the increasing values of the magnitude δ (Fig. 17.2).

From these plots it is seen that initially there is the stable equilibrium O , and then there appears the semi-stable periodic motion M_{12} (more precisely, the semi-stable periodic motion corresponding to the fixed point M_{12}) being decomposed   into the stable periodic motion M_2 and the unstable periodic motion M_1 . With δ increasing further, the amplitude of the unstable periodic motion M_1 will be decreasing and that of the stable motion M_2 increasing; and this will produced a confluence of M_1 with the equilibrium state O which will, because of this, turn unstable.

The phase portraits depicting   these transformations are presented in figure 17.3. On the basis of these pictures, let us now construct a bifurcation diagram. It is depicted in figure 17.4. As you see, it differs from the bifurcation diagram constructed for the soft excitation (Fig. 16.5). What follows from this difference? First of all, if earlier, with δ growing across zero, the equilibrium state   was losing its stability and was giving birth to a stable periodic motion, then now an unstable periodic motion is merging with this unstable equilibrium; as a result, the system, being earlier in the stable equilibrium, will leave it and pass at once into the periodic motion M_2 of a finite amplitude. Here lies the idea of a hard mode for exciting oscillations. In this mode, auto-oscillations of a finite amplitude arise immediately, on a stepwise basis. Of course, they are established in a steady way, not instantly, but upon some period of   time, and during this period,   the oscillation amplitude is growing continuously and, as a parameter is varying, a steady-state (or stationary) motion will arrive by jump.
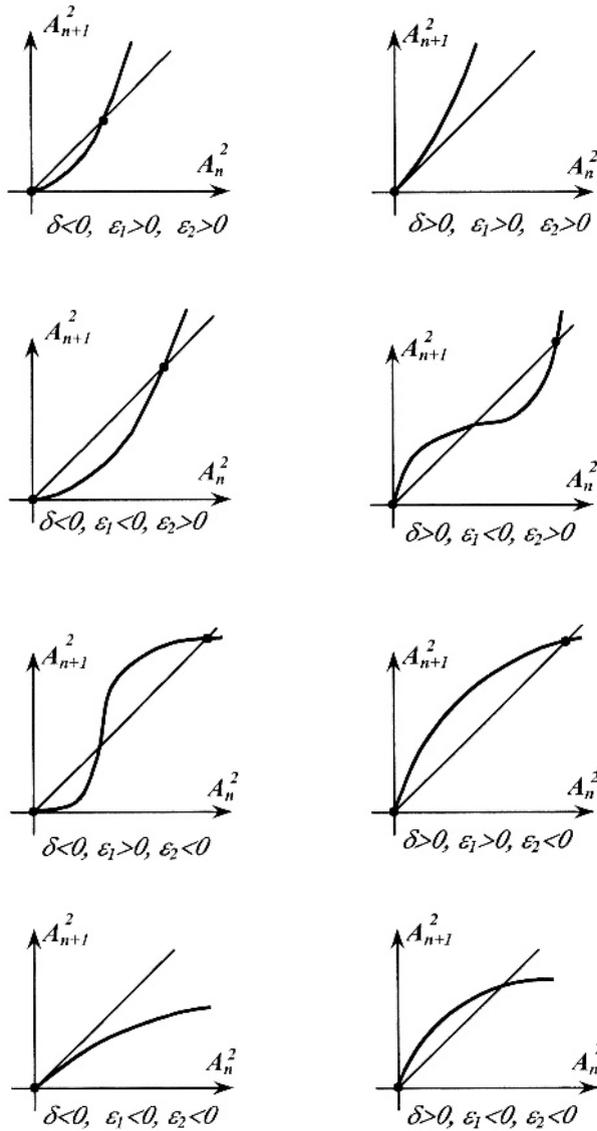
Fig. 17.1. Possible shapes of the Keniks-Lamerey diagrams for the polynomial anode characteristic (17.1) as consistent with the shape of the point mapping (17.2).
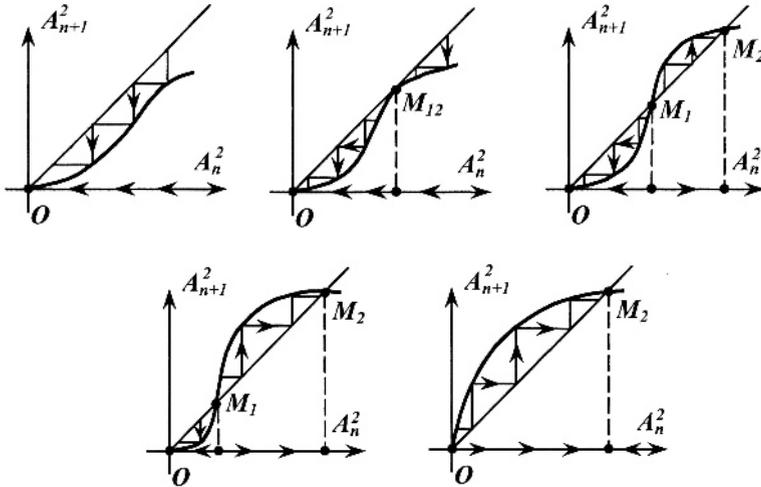
Fig. 17.2. A sequence of the Keniks-Lamerey diagrams for hard excitation of oscillations: at first, the equilibrium O is globally stable; then the semi-stable fixed point M_{12} is born, and then it is decomposed into two points M_1 and M_2; then the point M_1 merges with the equilibrium O and makes it unstable; and, as a result, the fixed point M_2 turns out to be globally stable.

Thus, the first specificity of the hard excitation of auto-oscillations as a physical phenomenon consists in a jump-like change of the nature of the motion, but not smoothly, with a parameter being varied. Mathematically, this specificity is reflected by the fact that an equilibrium state does not give birth to a stable periodic motion; instead, an unstable periodic motion merges with the equilibrium state, and the points, lying close to the now becoming unstable equilibrium state, are now tending to the earlier existing stable periodic motion.

The second essential specificity lies in an irreversible nature of a hard transition from an equilibrium to a steady-state motion, viz. when δ is increasing across $\delta = 0$ there appear hard auto-oscillations, but at the inverse decrease of δ across zero, the auto-oscillations do not vanish and remain varying their amplitude continuously. In order to pass from hard arising auto-oscillations to a stable equilibrium state, the parameter should be decreased until some negative value $\delta = \delta^*$ (Fig. 17.4), whereupon the auto-oscillations will vanish by jumps and a transition to a stable equilibrium will occur. The disappearance of a stable auto-oscillation is accompanied by its merge with the unstable periodic motion and by their joint disappearance (for $\delta = \delta^*$ the stable periodic motion M_2 will

merge with the unstable motion M_1 and at $\delta < \delta^*$ they disappear). Thus, as it is shown in figure 17.4, in the course of the hard excitation, auto-oscillations appear and disappear at different values of the parameter δ. This phenomenon looks like a hysteresis of ferromagnetic bodies.
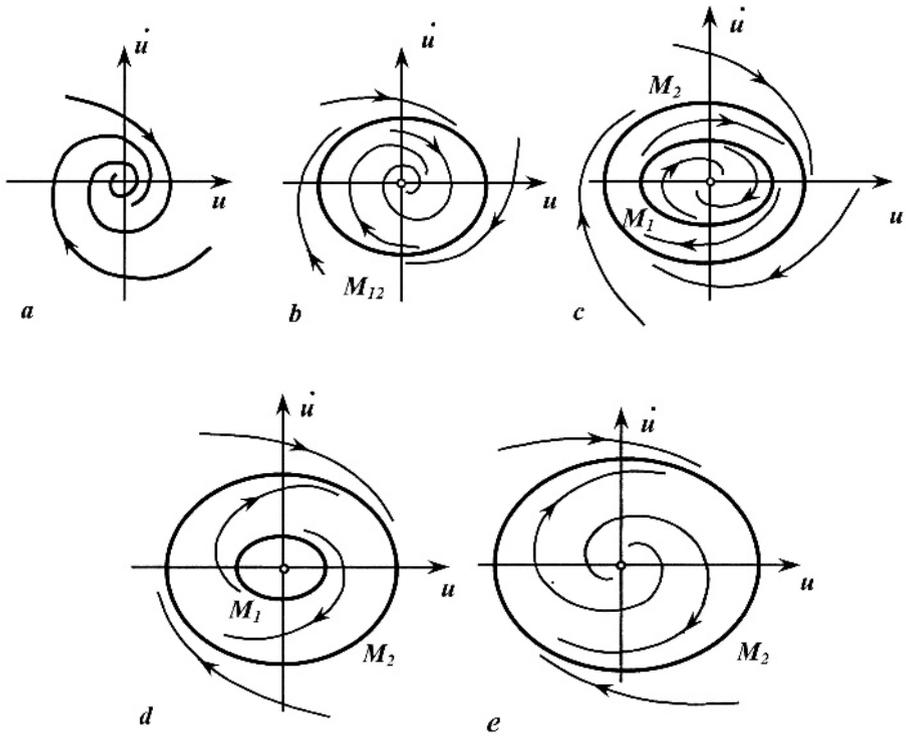
Fig. 17.3. The changes of the phase portraits corresponding to the changes in the Keniks-Lamerey diagrams in figure 17.2.
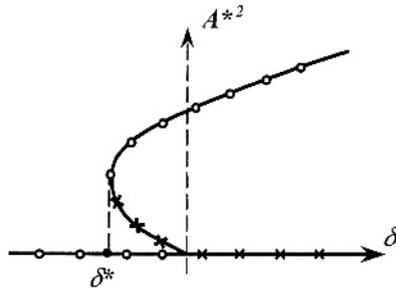
Fig. 17.4. A bifurcation diagram of hard excitation of oscillations. The circles and daggers correspond to stable and unstable equilibrium states and periodic oscillations.

I would like to finish my narration on soft and hard excitations of auto-oscillations with some mental experiment. Suppose you are sitting at a computer and, with some rolling knob, you can regulate (i.e. increase or decrease) a parameter of a dynamical system. Simultaneously, you can observe on a computer monitor how a state of the system (one of its components) is changing. Then, in case of the soft excitation regime, your rolling the knob slowly will, at first, bring you   no oscillations. Then, they will be arising and their amplitude will be growing slowly; the slower you are rolling your knob, the   slower they are growing (Fig. 17.5). If you roll the knob in a reverse direction,   all you saw previously will be repeated in a reverse order.
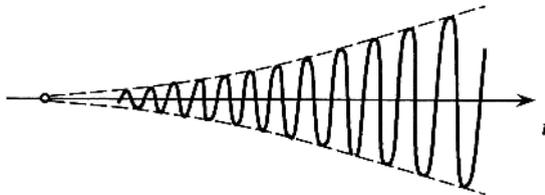
Fig. 17.5. An oscillogram of soft excitation of oscillations with smoothly changing the parameter.

For the hard excitation regime, at first, there will be no oscillations either; but then they will be arising at once, and a fastness of this transition will not decrease, if you are even slowing down your knob rotation (Fig. 17. 6).
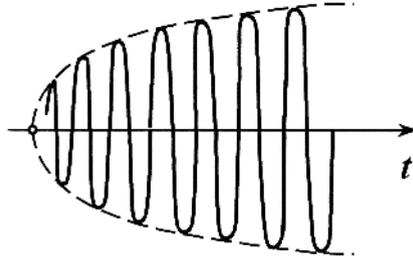
Fig. 17.6. An oscillogram of hard excitation of oscillations with smoothly changing the parameter.

With a reverse rolling of the knob, the picture will not be changed in reverse. The auto-oscillations will vanish and by a leap as well, if you roll the knob inversely for a longer time.

18 Stochastic oscillator (the "contrary clock")

An oscillator of an unpredictable behaviour. A mathematical model, a phase portrait and a point mapping. Unpredictability and randomness. Two kinds of behaviour of dynamical systems. Stability and instability, predictability and unpredictability. An oscillator (a dynamical system) as a generator of stochastic oscillations.

In the previous chapters we have analysed the Galileo-Huygens clock. The basis of this clock is a pendulum or any another oscillator with oscillations of a sufficiently stable period being sustained by any energy source. A stability of the period of oscillations means an accuracy of the clock. We have studied the possible reasons for instability and the ways to reduce it. A very important point for stability was a smallness of damping, which implies both a small dependence of the period upon this damping and a sufficiency of very small impacts and a small instability being caused by the impacts.

However, an oscillator can be both a high-precision time measurer with a rather stable period and can possess chaotic oscillations of sufficiently variable periods which may become so much greatly different that speaking about an oscillating period becomes senseless. Moreover, these oscillations obtain a property of randomnicity; they become unpredictable and may have a probabilistic description. This surprising opportunity has become a scientific sensation in the last years. It has changed conventional ideas concerning randomness. It turned out so that randomness can be generated by a dynamical system being described by differential equations. This takes place in spite of the fact that the uniqueness theorem retains still valid and a solution of these equations is still determined uniquely by their initial conditions (by the initial state). How can it be? It seems obvious that it is impossible, but nevertheless it can be so and is implemented around us not less frequently than the determinate processes being described by differential equations. It turns out that real solutions of the differential equations can be deterministic and predictable, but they can also be stochastic and unpredictable.

The purpose of my further narration is to explain how it can be by means of a very simple example of a stochastic oscillator or the "contrary clock".

A clock is a damping oscillator being pushed from time to time. The "contrary clock" is a self-swinging oscillator whose oscillations are restrained from time to time.

The mathematical model of the clock was of the form

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = 0 \qquad \text{for } x \neq 0 \text{ or } x = 0, \text{ and } \dot{x} < 0,$$

$$\dot{x}_+ = \dot{x}_- + p \qquad \text{for } x = 0 \text{ and } \dot{x} > 0. \tag{18.1}$$

The mathematical model of the "contrary clock" ($p \leq a$) is as follows

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = 0 \qquad \text{for } x \neq 0, \text{ or } x = 0 \text{ and } \dot{x} < a,$$

$$\dot{x}_+ = \dot{x}_- - p \qquad \text{for } x = 0 \text{ and } \dot{x} \geq a. \tag{18.2}$$

The equations (18.2) are obtained from (18.1) by substituting $-\delta$ for δ and $-p$ for p. Besides, it is also zero that is replaced by the number $a > 0$.

We need to investigate the model (18.2). First of all, let us imagine the form of its phase trajectory (Fig. 18.1).
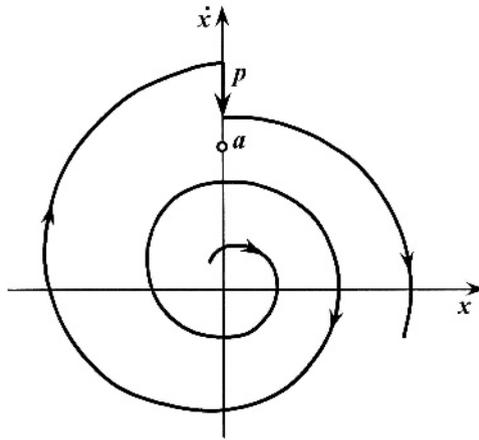
Fig. 18.1. A phase portrait of the "contrary clock" (a stochastic oscillator).

Then, the investigation of the dynamics of the model (18.2) will be reduced to a point mapping, to the Keniks-Lamerey diagram. To this end, let us consider the subsequent departures of the phase point from the semi-axis $x = 0$, $\dot{x} > 0$.

Suppose that upon the first intersection we have $\dot{x} = u$, and upon the second $\dot{x} = \Gamma\bar{u}$. Let us determine the connection between u and $\bar{u}$. Let the trajectory

be departing from the point $M(0, u)$. It will arrive at the half-line $x = 0, \dot{x} > 0$ again at some point $M_1(0, u_1)$, where

$$u_1 = u e^{2\pi\delta/\Omega}.$$

It is not restrained, if $u_1 < a$ (Fig. 18.2a) and then $\bar{u} = u_1$, or is restrained, if $u_1 \geq a$ and then $\bar{u} = u - p$ (Fig. 18.2b).
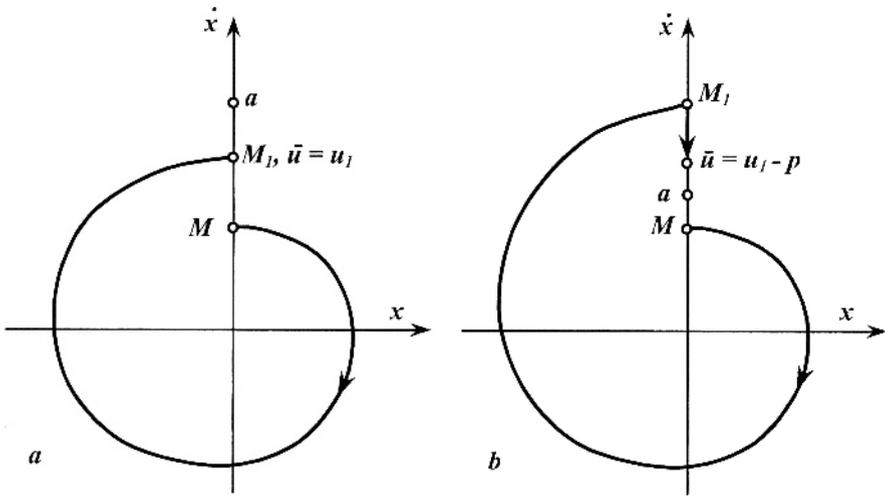
Fig. 18.2. Possible shapes of phase trajectories for the "contrary clock": **a** for $u_1 < a$ and **b** for $u_1 > a$. For an oscillation cycle the point $M(0, u)$ passes to the point $M_1(0, u_1)$, at $u_1 > a$ this passage being performed with a jump from u_1 to $\bar{u}$.

Therefore, we obtain

$$\bar{u} = \begin{cases} qu & for \quad u < a/q \\ qu - p & for \quad u \geq a/q \end{cases}, \tag{18.3}$$

where, for shortening the notation,

$$q = e^{2\pi\delta/\Omega} > 1$$

was introduced.

Let us depict the Keniks-Lamerey diagram for the point mapping obtained. For this, we shall construct a plot of   dependence (Fig. 18.3) and shall draw a bisecting line.
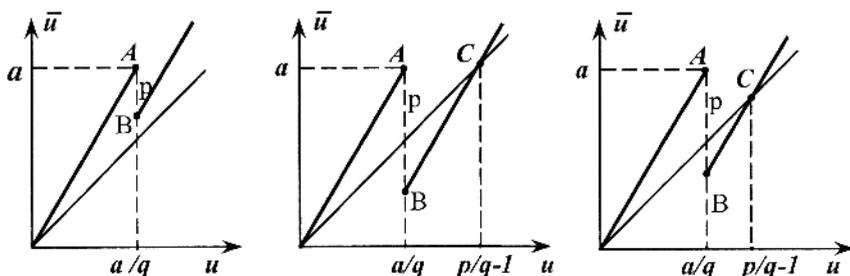
Fig. 18.3. Possible shapes of the Keniks-Lamerey diagrams for the "contrary clock".

In spite of its simplicity, studying the plot is sufficiently complicated. However, it is worth to spend efforts. Even upon spending a lot of effort, we shall not be able to prove everything, and so you will have to take something on trust.

The point mapping diagrams constructed in this way can be of three different forms. These are shown in figure 18.3 a, b, c.

The first case: both points A and B lie above the bisecting line. For this case, $a - p > a / q$ or $p / a < 1 - 1 / q$. The second case: the point B is below the bisecting line and the point C is above the point A. Here, $a - p < a / q$ and $a < p /(q - 1)$, or $p / a > 1 - 1 / q$ and $p / a > q - 1$. The third case: the point B is below the bisecting line and the point C is below than the point A. For this case, $a - p < q - 1$ and $a > p /(q - 1)$, or $p / a > 1 - 1 / q$ and $p / a < q - 1$. As seen above, what from these cases will take place depends upon the magnitudes of q and p / a, provided that $q < 1$, $0 < p / a \leq 1$ are taken into account.

To each of the above listed cases there corresponds its own domain on the plane of magnitudes of q and p / a . These domains indicated as 1,2 and 3, respectively, are shown in figure 18.4.

In the cases 1 and 3, almost all the subsequent transformations extend at infinity. This corresponds to an unbounded swing of oscillations. The case 2 covers the two possible different behaviours being dependent upon the initial value of u. For $u < p /(q - 1)$, the subsequent transformations of u will be

bounded for all the time. For $u > p/(q-1)$, they, on the contrary, will increase unboundedly.
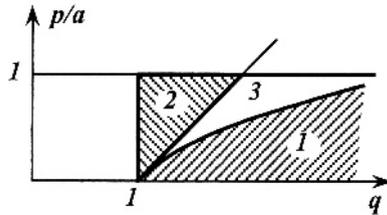
Fig. 18.4. The parametric portrait corresponding to the different shapes of the Keniks-Lamerey diagrams shown in figure 18.3.

Of the most interest for us is the case 2, when values of transformations retain bounded all the time. For any initial $u_0 \in (0, p/(q-1))$, the subsequent transformations $u_0, u_1, u_2, \ldots$ from some number fall onto the segment $[a-p, a]$ and remain on it. Any such phase trajectory $u_0, u_1, u_2, \ldots$ is unstable, and the trajectories close to it are running away. They may come close to each other, in view of the discontinuity in the mapping (18.3), but further they are again dispersing, etc.

Let us track how this is occurring through comparing the two trajectories $u_0, u_1, u_2, \ldots$ and $v_0, v_1, v_2, \ldots$ that have the close initial points u_0 and v_0. If the discontinuity point a/q of the mapping (18.3) is not located between the points u_0 and v_0, then in accordance with (18.3), one has

$$u_1 - v_1 = q(u_0 - v_0) \tag{18.4}$$

and, therefore, the distance between them will be increasing $q > 1$ times. If the point a/q is located between the points u_0 and v_0, then

$$u_1 - v_1 = q(u_0 - v_0) \pm p . \tag{18.5}$$

The same alternative takes place in all subsequent transformations as well. Therefore, in any case after some number of transformations, however close could be the initial points, they will diverge at a finite distance being not less than $p/2q$. Besides, this definitely occurs during the number of steps not more than n, for which

$$q^n \mid u_0 - v_0 \mid > \frac{p}{2q}. \qquad (18.6)$$

Therefore, however accurately though with some error we would have calculated the subsequent transformations (18.3), their errors upon some of their number will   exceed the magnitude $p/2q$. An exception here will be only the case when the calculation is absolutely precise, but this is not real at all. Since these small errors, as well as small disturbances, are unpredictable in their physical application and are not subject to any account, then we shall   actually encounter the challenge of unpredictability of a phase trajectory for the motions of the "contrary clock". The motions of the "contrary clock" are unpredictable, and this unpredictability reveals itself very soon. Indeed, for $q = 1.76$, $p = 4$ and for the very small error of 10^{-8} order in the initial conditions, the inequalities (18.6) will hold, since $n \geq 32$. For the error of the atom size order, they will hold for $n \geq 48$.

The running of the "contrary clock" is also unpredictable because of the fact that we cannot assign or know of their initial conditions absolutely precisely. Therefore, any error in an initial value will, upon some time, lead to   large differences, i.e. will bring unpredictability.

The idea that stochasticity can be generated by the vanishing   neglected unknown uncertainties with their further exponential amplification produced by the instability –this idea had been yet expressed by J. H. Poincare and later on, in a more detailed form, by N.S. Krylov and M. Born. With reference to a long-range weather forecast, the effect of the surprisingly sensitive dependence upon small disturbances was yet noted by E. Lorenz as the "butterfly effect". A butterfly by its flight can change the atmosphere state and the future weather in the way similar to that through which it has influenced the events unfolding in one of Ray D. Bradbury's science fiction stories.

Through comprehending the above said we are   naturally coming   to the understanding of the fact that instability gives birth to unpredictability, though this does not imply stochasticity. For stochasticity to appear, it is also required to have a probability description . Where can this description be taken from? As I think, to-day it is still difficult to give an exhaustive   unique answer. This question needs a further deep studying. But, nevertheless, something can be said right now.

The dynamical system described by the mapping (18.3) turns out to possess some surprising properties which allow its unpredictability to be assigned with the probability being determined by the system's dynamics only and not by the stochastic disturbances that are causing this unpredictability. What is needed for this is only the availability of these negligibly small randomnesses. It is sufficient to have   those randomnesses which are generated by a thermal motion of molecules or even by a quantum probabilistic nature of the microworld.

Let us make it   clearer in the example with the "contrary clock" with the parameters from the domain 2 (Fig. 18.4). Here, we take   the point with $p = a$

and $q = 2$. The diagram of the point mapping (18.3), that we shall denote by T, will assume the form shown in figure 18.5, and this mapping itself will be written as

$$\bar{u} = \begin{cases} 2u & for \quad u < a/2 \\ 2u - a & for \quad u \geq a/2 \end{cases} \tag{18.7}$$

The mapping inverse to (18.7) is two-valued. Let T_1^{-1} and T_2^{-1} be its unique branches. In accordance with (18.7), we have

$$T_1^{-1} : \quad u = \frac{\bar{u}}{2} ; \quad T_2^{-1} : \quad u = \frac{\bar{u} + a}{2} . \tag{18.8}$$
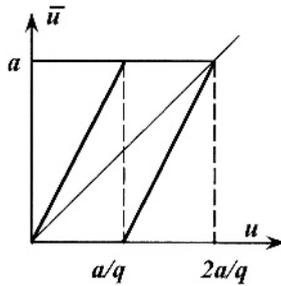
Fig. 18.5. The Keniks-Lamerey diagram for the "contrary clock" at $p = a$ and $q = 2$.

Let us assign the initial point u_0 by supposing it to be random with some probability density $p_0(u)$. A small uncertainty in choosing u_0 implies the probability density $p_0(u)$ to be similar to the δ-function and to be of the form shown in figure 18.6. The narrower and higher the peak of this function, the more precise the initial value u_0 is assigned.

Now let us perform the transformation T under which every u_0 passes into the new u_1. If u_0 is random, then u_1 will be random as well. Through the probability density $p_0(u)$ of the random point u_0, one may find the probability density for the transformed point u_1. Let it be $p_1(u)$. Proceeding the

transformations T further, one may find $p_2(u), p_3(u)$ and so on. In order to calculate them actually, let us find the relation between $p_s(u)$ and $p_{s+1}(u)$.
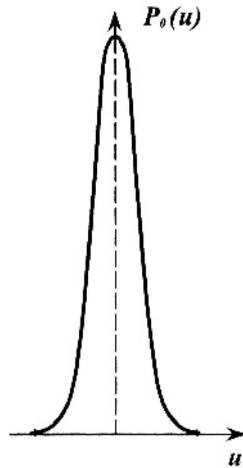
Fig. 18.6. Probability density of the initial value u .

The point $u_{s+1} \in (u, u + du)$, if $u_s \in (T_1^{-1}(u), T_1^{-1}(u + du))$ or $u_s \in (T_2^{-1}(u), T_2^{-1}(u + du))$. Therefore,

$$P(u_{s+1} \in (u, u + du)) =$$
$$= P(u_s \in (T_1^{-1}(u), T_1^{-1}(u + du))) + P(u_s \in (T_2^{-1}(u), T_2^{-1}(u + du)))$$

(18.9)

where $P(\ )$ denotes the probability of the corresponding event. Thus, the relation (18.9) means that the probability of the fact that the point u_{s+1} being an image of the point u_s belongs to the interval $(u, u + du)$ is equal to the sum of the probabilities of the fact that u_s belongs to $(T_1^{-1}(u), T_1^{-1}(u + du))$ or $(T_2^{-1}(u), T_2^{-1}(u + du))$, respectively.

In view of (18.8), let us rewrite (18.9) as

$$P(u_{s+1} \in (u, u + du)) =$$

$$= P(u_s \in (\frac{u}{2}, \frac{u + du}{2})) + P(u_s \in (\frac{u + a}{2}, \frac{u + a + du}{2}))$$

or

$$p_{s+1}(u)du = p_s(\frac{u}{2})\frac{du}{2} + p_s(\frac{u + a}{2})\frac{du}{2}$$

or, cancelling du,

$$p_{s+1}(u) = \frac{1}{2}[p_s(\frac{u}{2}) + p_s(\frac{u + a}{2})]. \qquad (18.10)$$

The desired relation (18.10) allows us, starting with the known probability density $p_0(u)$, to successively calculate the probability densities $p_1(u), p_2(u)$, etc. Here a natural question then arises: is there any limiting probability density and what is it? One may establish a convergence and find a limit as follows.

Let the initial probability density $p_0(u)$ have a derivative. Then, according to (18.10), all the subsequent $p_s(u)$ are differentiable as well. Differentiating (18.10), we find that

$$p'_{s+1}(u) = \frac{1}{4}\{p'_s(\frac{u}{2}) + p'_s(\frac{u + a}{2})\}. \qquad (18.11)$$

It is evident from (18.11) that

$$\lim_{s \to \infty} p'_s(u) = 0,$$

and, therefore, the limiting function $p_\infty(u)$ is constant. Taking into account the normalization, we come immediately to

$$p_\infty(u) = \frac{1}{a}.$$

Thus, upon some time, the magnitudes of the post-impact velocities u_s at the instants of passing by the pendulum of the "contrary clock" through the equilibrium will be the uniformly distributed random variables with the interval of values being equal to $(0, a)$. These random variables are not independent. On the contrary, the magnitudes close by number are connected with each other more closely, but when departing from each other, this connection is attenuating, and the very distant values will be practically independent.

We have substantiated the availability of probability only with use of some special example, and at this making use of the particular case $p = a$ and $q = 2$. However, this proof can be also done for a wider case. In the above said the consideration adopted concerning the relations between the probability densities $p_s(u)$ and $p_{s+1}(u)$ carries a general nature. The difficulties here arise only in proving the existence of the limit of $p_s(u)$ as $s \rightarrow \infty$. That is why we restrict ourselves by a study of the particular case $p = a$ and $q = 2$, though the present investigations make it possible to formulate far more general statements. They allow us to state that the existence of the limit is not something extraordinary, but, on the contrary, is one of general possibilities.

Thus, the situation revealed in the very simple example with the "contrary clock" is of a general nature. It takes place in a wide range of turbulence phenomena in fluids and gases. It is because of it that a long-range weather forecast becomes very difficult. Stochasticity is also observed in various electrical and mechanical systems, and in motions of gas molecules. One may say that to stability and instability there correspond the two fundamentally different basic types of behaviour of dynamical systems. The first case pertains to predictability and determinism, and the second to unpredictability and stochasticity.

The simplest typical example of the first type of dynamical systems is the Galileo-Huygens clock, while the "contrary clock" or the simplest stochastic oscillator constitutes a typical example for the second case.

A clock serves for measuring time and its merit lies in the regularity and stability of its running, whereas a stochastic oscillator produces macro-randomness from micro-randomness. The Galileo-Huygens clock generates a uniform series of time instants, while the "contrary clock" generates a random sequence.

Phase trajectories in the first-type dynamical systems are coming close, tending to equilibrium states or periodic motions; in the second-type systems, phase trajectories are diverging all the time, forming thus a complex intricate ball.

Earlier, when discussing auto-oscillations, there has been revealed some naturalness of their appearance in dynamical systems and has been suggested a geometrical image of auto-oscillations, i.e. a closed phase curve, which all the neighbouring trajectories are converging to. With unpredictable chaotic motions, this can be done with greater difficulties, since these motions are more many-sided. Their general specificity lies in their irregularity and unpredictability that can be born by very negligible disturbances. Nevertheless, for such motions, one may distinguish some attractive set of unstable phase trajectories. This is when each separate phase trajectory is unstable and its neighbouring trajectories are running away from it, but, taken as a whole, all the trajectories are attractive. This is schematically shown in figure 18.7.

The presence of such a property is enough for unpredictability. As for stochasticity, it, at least, additionally demands the existence of the limiting probability density discussed above. Generally speaking, this subject is still waiting for its researchers and further investigations still ahead, and these

investigations will be able to modify and enrich our phenomenological and postulationally constructed theory of probability.
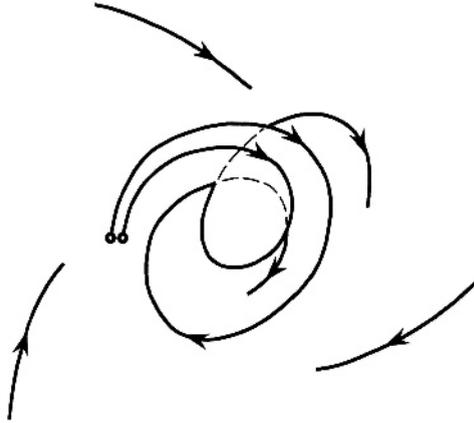
Fig. 18.7. A schematic representation of the attractive set for an unstable unpredictable motion.

Here, I could have finished, but still I would like to acquaint you with the following remarkable system of differential equations by Lorenz:

$$\dot{x} = -\sigma x + \sigma y$$

$$\dot{y} = -rx - y - xz$$

$$\dot{z} = -bz + xy \,.$$

It describes, in an approximate way, the thermal convection between a heated lower and a colder upper horizontal plates. For $b = 0$, these equations describe a convective fluid circulation in a closed circular tube being heated from below. If the tube is not heated from below, this system has a stable equilibrium. With heating being increased, this equilibrium becomes unstable and the two stable fluid circulations, clockwise and counterclockwise, are being born. With heating being increased further, these circulations become unstable and chaos is appearing, i.e. the equilibrium is broken and a circulation is appearing, then it, in its turn, is broken and another circulation is arising, and then the last, in its turn, is broken and the former is appearing again, etc; therefore, these events proceed without any regularity, without any hints about   periodicity and only with a

uniquely regular replacement of directions of the circulation; the replacement of one-side circulation by another-side circulation is unpredictable and is performed after different random time intervals. However, the most remarkable thing here is that the sequence of these circulation replacements turns out to be able to be described   through   the point mapping similar to one considered by us in the example with the "contrary clock". And now right here we are ready to put a point.

19   Instability and auto-oscillations caused by friction

A classical Coulomb model of dry friction and a necessity to improve it. Oscillation excitation by friction and an appearance of auto-oscillations.

The below narrative is about a mechanism and simplest mathematical models describing a birth of instability and auto-oscillations due to friction.

In the fact that friction brings instability and auto-oscillations there is hidden something paradoxical, since we got accustomed to friction as retarding and stopping a motion. However, this turns out sometimes not so, and the paradox ceases as friction plays a role of a distributor of energy of some other energy source, not related to friction. As for friction itself, it actually can only retard a motion.

A schematic simplified model, from which we start our study of the systems with a paradoxical role of the Coulomb friction, consists of a mass being pulled by a spring along a horizontal plane. Here, a massive body is moving under action of the force of a stretchable spring (Fig. 19.1a). In this model, it may be considered that a plane is moving with some constant velocity and the spring terminal attached to the body is motionless.

Let us consider this last variant of the model (Fig. 19.1b). This model can also represents the so-called Froude pendulum (Fig. 19.1c), which is a usual pendulum mounted through a muff on an axis rotating at a constant velocity. The axis is rotating in the muff with some, not very large, friction so that the pendulum might swing on it, but not rotate together with the axis. As real prototypes of this schematically described model there may serve brake blocks in cars, trams, railroad wagons and in other various mechanically braking devices as well. With some portion of reserve, to this area of applications there may be brought a violin string along which a bow is sliding with some friction.

Let us now return to the schematic model adopted. Denoting a magnitude of the displacement of the mass m from its position (under this displacement the spring is not extended) through x and the force of friction between the mass and the moving plane through F, we may write the Newton equation as

$$m\ddot{x} = -kx + F \qquad (19.1)$$
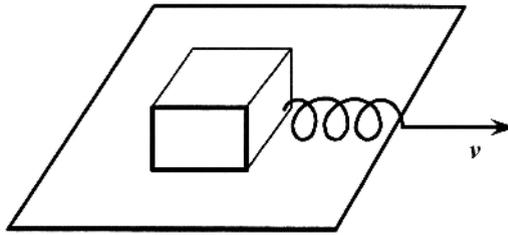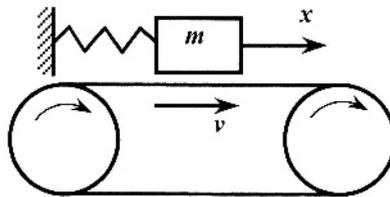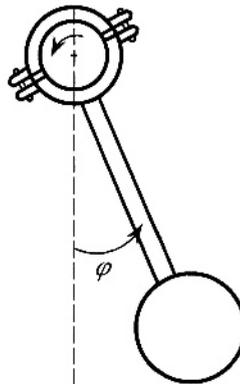
Figs. 19.1. Different specific realizations of the mathematical model (19.3).

The nature of the dry   friction force is very complicated. The first mathematical model of the dry friction force was suggested yet in the   18-th century by Ch. Coulomb. This force was called after him. You were told about it at the school as about the Coulomb friction laws, according to which the friction force does not depend upon the mutual velocity of the rubbing surfaces; it is directed against the velocity and in its value is proportional to the normal pressure so that

$$F = F_0 sign(v - \dot{x}), \tag{19.2}$$

where $sign\,\xi$ is a sign of the value ξ. This formula is valid for $\dot{x} \neq v$. For $\dot{x} = v$, i.e. when the mass is moving together with the plane (and the mass seems fastened to the plane) , the friction force, i.e. a static friction force, may be of any value within the limits $(-F_0, F_0)$. The last implies that any force being less in value than F_0 cannot displace the mass being at rest with respect to the moving plane. The movement will start when an acting force in its value is more than F_0.

These last statements about a static friction force can be included into the formula (19.2) by adopting that $sign\,0$ will indicate any value between -1 and 1. Upon that, the relation (19.2) will be a mathematical model for the Coulomb friction, and the motion equation will be written as

$$m\ddot{x} + kx = F_0 sign(v - \dot{x}). \tag{19.3}$$

The differential equation (19.3) is a mathematical model for the system under study. Revealing its dynamics is reduced to a study of the phase portrait for the dynamical system (19.3).

The phase space of the dynamical system defined by the equation (19.3) is a plane of the phase variables x and $\dot{x} = y$. In the half-planes $y < v$ and $y > v$, motions of the phase points are subjected, according to (19.3), to the equations

$$m\ddot{x} + kx = \mp F_0, \tag{19.4}$$

each of which represents a harmonic oscillator with an equilibrium state at $x = -F_0/k$, $\dot{x} = y = 0$ or at $x = F_0/k$, $\dot{x} = y = 0$, respectively. This allows us to immediately depict the phase trajectories separately in each of the half-planes $y < v$ and $y > v$, as shown in figure 19.2.

The motion of the phase point in this figure is defined everywhere except the points of the line $\dot{x} = y = v$. This line is divided into three parts: the half-line $(-\infty, O_1)$, with respect to which the phase trajectories come from below and go away from above; the closed interval $[O_1, O_2]$, on which the phase trajectories come from below and above; and the half-line (O_2, ∞), where they come from above and go away from below.

It is natural to expect, due to the continuity of the variation of the velocity $y = \dot{x}$ and that of the coordinate x (a phase point cannot perform jumps), that in its travelling the phase point will transverse the half-lines $(-\infty, O_1)$ and (O_2, ∞) from below to above and from above to below, respectively; falling into the interval $[O_1, O_2]$ from below or from above, the phase point will remain on

it, moving along it in accordance with $y = \dot{x} = v$ lying on the half-line; and then, having arrived to the point O_2 , the phase point will leave it passing on to the below half-plane.
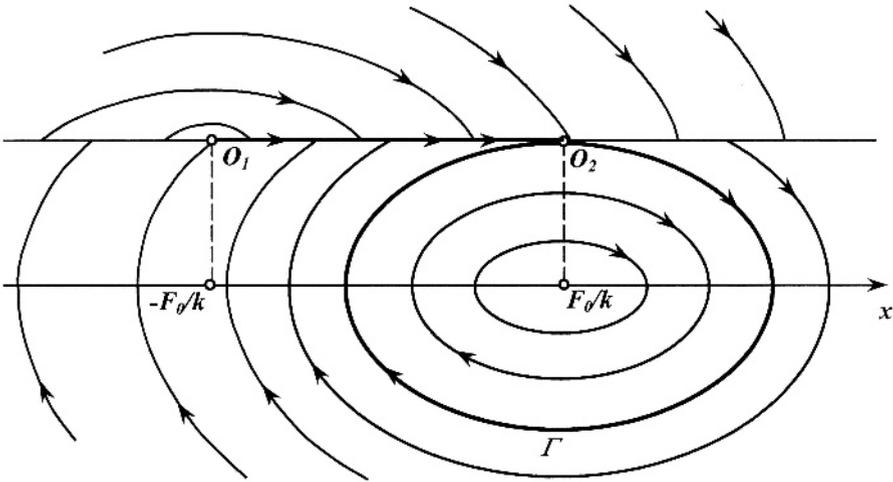
Fig. 19.2. The phase portrait of the linearized Froude-pendulum-like system under the Coulomb idealization for friction.

   The description done completely defines the motion of phase points within the entire phase plane. However, it may seem not proved completely or, at any case, may seem to need a more detailed description.
   The mass m can move along the supporting plane with some slipping or without it, i.e. together with the plane. Within the half-planes $y < v$ and $y > v$, the motion will be everywhere with a slipping. On the half-lines $(-\infty, O_1)$ and (O_2, ∞) , there will be instantaneous halts with respect to the plane caused by the replacements of the slipping   directions being performed at the velocity $v - \dot{x}$. Within the interval $[O_1, O_2]$, no slipping will occur and here the mass will move together with the plane at the velocity v , i.e. here $\dot{x} = y = v$. In the interval $[O_1, O_2]$, the friction force is a static friction force, and here, during some finite time, we shall have $y = v$ and $\dot{y} = 0$, and hence, in view of the equation (19. 3), we shall get

$$kx = F_0 sign0 . \tag{19.5}$$

Since on the interval $[O_1, O_2]$ we obtain $-F_0/k < x < F_0/k$, then the tighting force kx of the spring will not exceed the value F_0 of the static friction and, therefore, no slipping will appear. On the contrary, on the half-lines $(-\infty, O_1)$ and (O_2, ∞), there will occur instantaneous halts, because here the tighting force kx will exceed the limit F_0 of the static friction force.

Now, we are turning to a general description for the phase portrait shown in figure 19.2 and for admissible motions of the mass m.

There is an equilibrium O $(x = F_0/k, y = 0)$. It is stable, a centre-type equilibrium. Around it, there may be an infinite set of harmonic oscillations, up to some limiting amplitude corresponding to the curve Γ. These harmonic oscillations do not damp, though friction is available. All the other motions, upon some oscillations with a degressive swinging amplitude, lead to a temporary movement of the mass m together with the moving plane, i.e. to a movement of the phase point along the interval $[O_1, O_2]$, whereupon there arise harmonic oscillations described by the closed phase curve Γ.

We have done a theoretical research of the motions of the spring-attached mass, lying on a moving plane, on the basis of the Coulomb friction model and have revealed not only a stable equilibrium (that looks rather natural), but also, to our astonishment, some set of different periodic oscillations. However, these theoretical results are not confirmed experimentally: during some small finite time intervals the theoretic and experimental data come close to each other, but with time the divergence between them is growing. This drastic growth of differences takes place only within the phase space domain confined by the curve Γ and implies that the equilibrium may be unstable and instead of an infinite number of periodic motions there may be either no one motion or one, that is most often. At the same time, the Coulomb laws are observed very precisely. Here for us there is no way out, and so even negligible divergences from the Coulomb law may seemingly yield such significant differences. This is really so; and you have already encountered this challenge when you were studying the Volterra-Lotka "predator-prey" model. There the reason was hidden in the incorrectness of the mathematical model and that incorrectness was brought about by the presence of the centre-type equilibrium. In our case, a centre is available as well. Therefore, everything possibly depends upon it; and, hence, the mathematical model for the Coulomb friction should be improved in such a way as to make the mathematical model of the system under question correct. It is desirable to be done so that the differences between real and ideal Coulomb friction characteristics be accounted and simultaneously the advanced mathematical model should not complicate our investigation. We note here that to the Coulomb law there corresponds the plot in figure 19.3, showing a dependence of the friction force F upon the mutual velocity u of the rubbing surfaces.
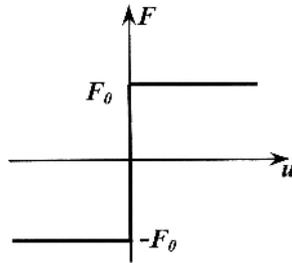
Fig. 19.3. The Coulomb idealization for dry friction.

An experiment suggests somewhat other plots. In general, they   are of the
following three types: the friction force is increasing as velocity increases; the
friction force is decreasing as velocity decreases; at first, is decreasing and then
increasing. Possible also   are   some other variants, but we shall   leave them
beyond our attention. The three listed types are shown in figure 19.4.
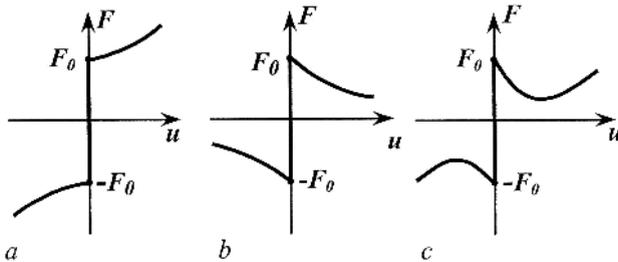
Fig. 19.4. Possible deviations from the Coulomb characteristic of friction.

For the first two types, the following piecewise-linear approximation is natural,
under which the friction force will be equal to:

$$F = F_0 sign\, u + hu ,$$ (19.6)

where u is a mutual velocity of the rubbing surfaces; in our case it will be
$u = v - \dot{x}$. In the third case, the friction force is approximated by the third-order
polynomial

$$F = F_0 sign\, u + \beta u + \gamma u^3 .$$ (19.7)

Naturally, these approximations can be good only for bounded values of u .

In the case (19.6), a mathematical model of the system under consideration will be as follows

$$m\ddot{x} + h\dot{x} + kx = F_0 sign(v - \dot{x}) + hv ,\qquad (19.8)$$

and in the case (19.7), as follows

$$m\ddot{x} + h\dot{x} + kx = F_0 sign(v - \dot{x}) + \beta u + \gamma u^3 ,\qquad (19.9)$$

where h, β, γ are small.

As for the general case, the equation for improved mathematical model can be written as

$$m\ddot{x} + kx = F(u) ,\quad u = v - \dot{x} ,\qquad (19.10)$$

where $F(u)$ is a piecewise smooth function having a discontinuity point at $u = 0$. It is clear that the mathematical model (19.10) involves the models (19.8) and (19.9) as particular cases.

We start our investigation from the very general model, but we will only deal with the equilibrium state $x = F(v)/k,\ \dot{x} = 0$, when the friction force and the spring tighting force are balanced. Of interest for us will be only a stability of this equilibrium.

For stability analysis, we need to consider the motions occurring close to the equilibrium $x = F(v)/k$. Let us introduce a new variable

$$\xi = x - \frac{F(v)}{k}\qquad (19.11)$$

and write the equation (19.10) as follows

$$m\ddot{\xi} + k\xi = F(u) - F(v) = -\frac{\partial F}{\partial u}\Big|_{\dot{x}=0} \dot{\xi} + \ldots .$$

Neglecting the terms of the second and subsequent orders of smallness with respect to the small amount of ξ, denoted above by dots, we will consider the linearized equation

$$m\ddot{\xi} + \frac{\partial F}{\partial u}\Big|_{\dot{x}=0} \dot{\xi} + k\xi = 0 .\qquad (19.12)$$

This is a well-known oscillator. Its equilibrium is stable if

$$\frac{\partial F}{\partial u}\Big|_{u=v} > 0\qquad (19.13)$$

and unstable under the otherwise inequality.

Thus, if the velocity v is on an ascending branch of the friction characteristic, then the equilibrium will be stable, and, if it lies on a descending branch, then the equilibrium will be unstable. This general statement derived from the general model will retain also true, of course, for the models (19.6) and (19.7). A qualitative shape of the friction characteristic for the model (19.6) is given in figure 19.4. Here,

$$\frac{\partial F}{\partial u}\bigg|_{u=v} = h .$$
(19.14)

For the model (19.7), we have

$$\frac{\partial F}{\partial u}\bigg|_{u=v} = \beta + 3\gamma v^2 .$$
(19.15)

Therefore, the equilibrium state $x = F(v)/k,\ \dot{x} = 0$ of these models is stable for the positive h and $\beta + 3\gamma v^2$, and unstable for negative.

Without specifying the shape of the friction characteristic, to say something else concerning the model (19.10) is hardly possible; therefore, let us address to the more specific models (19.8) and (19.9).

Let us rewrite the equation (19.8)

$$m\ddot{x} + h\dot{x} + kx = F_0 sign(v - \dot{x}) + hv .$$

In the half-planes $\dot{x} > v$ and $\dot{x} < v$ of its phase plane $x,\ \dot{x} = y$, the equation will take the form

$$m\ddot{x} + h\dot{x} + kx = \mp F_0 + hv ,$$
(19.16)

i.e. in each of them, it will be a linear oscillator with the equilibrium state removed to the point $x = \dfrac{h}{k} v \mp \dfrac{F_0}{k},\ \dot{x} = 0$ (with respect to the common case when the equilibrium is placed in the point $x = \dot{x} = 0$). The quantity h is assumed to be small, and therefore, their phase portraits will be a stable focus for $h > 0$ and an unstable focus for $h < 0$. Here, a twisting and untwisting of the phase trajectories occurs slowly because of smallness of h ; therefore, a drastic change of the phase trajectories for the equation (19.16) with respect to those for the earlier considered equation (19.4) will happen only with the closed phase trajectories covering the point $(F_0/k, 0)$ (Fig. 19.2). For $h > 0$ and $h < 0$ these changes will be different ; they are represented in figure 19.5 a and b. The case considered earlier refers to $h = 0$. In the first case, for $h > 0$, any motion terminates with   the stable equilibrium state O with the coordinates

$x = \dfrac{F_0}{k}$, $\dot{x} = 0$. In the second case, when $h < 0$, this equilibrium is unstable, and any motions will terminate with a stable periodic motion, i.e. with an auto-oscillation depicted in figure 19.5b by the closed curve Γ (a thick line).

These qualitative variations occur for the indefinitely small h, and it confirms and proves the above made statement about incorrectness of the mathematical model (19.8), being coincident with the model at hand for $h = 0$. For $h \neq 0$, the mathematical model (19.8) is correct. This follows from the general theory of robustness, i.e. from structural stability, for second-order dynamical systems. Therefore, it may be expected that the derived result is of general nature, i.e. it takes place for any monotonous – not very steep – friction characteristics, increasing and decreasing. Of course, before making such a supposition it is not bad to look at the computer display. Besides, one should keep in mind that the above approximations for the friction force presuppose $\dot{x}$ and v to be not very large.

Let us turn to the model (19.9). Its friction characteristic consists of a descending and a subsequent ascending parts (Fig. 19.4c). Let us take as a sample the investigation made earlier for dynamics of a lamp tube described by the Van der Pol equation.

Let $v > 0$ and $\dot{x}$ be small. Then we shall write the equation (19.9) as follows

$$m\ddot{x} + kx = F_0 + \beta v + \gamma v^3 + (-\beta - 3\gamma v^2)\dot{x} + 3\gamma v \dot{x}^2 - \gamma \dot{x}^3 =$$

$$= a_0 + a_1 \dot{x} + a_2 \dot{x}^2 + a_3 \dot{x}^3 .$$

After replacing the variables $\xi = x - a_0 / k$, we reduce this differential equation to

$$\ddot{\xi} - 2\delta(1 - a\dot{\xi} - b\dot{\xi}^2)\dot{\xi} + \omega^2 \xi = 0 , \qquad (19.17)$$

where $2\delta = a_1 / m$, $a = a_2 / ma_1$, $b = a_3 / ma_1$, and $\omega^2 = k / m$. For $a = 0$, it is very similar to the Van der Pol equation. As it will be apparent in the sequel, the term $a\dot{\xi}$ is inessential. The equation (19.17) with $a = 0$ and the Van der Pol equation are the simplest mathematical models for auto-oscillation systems.
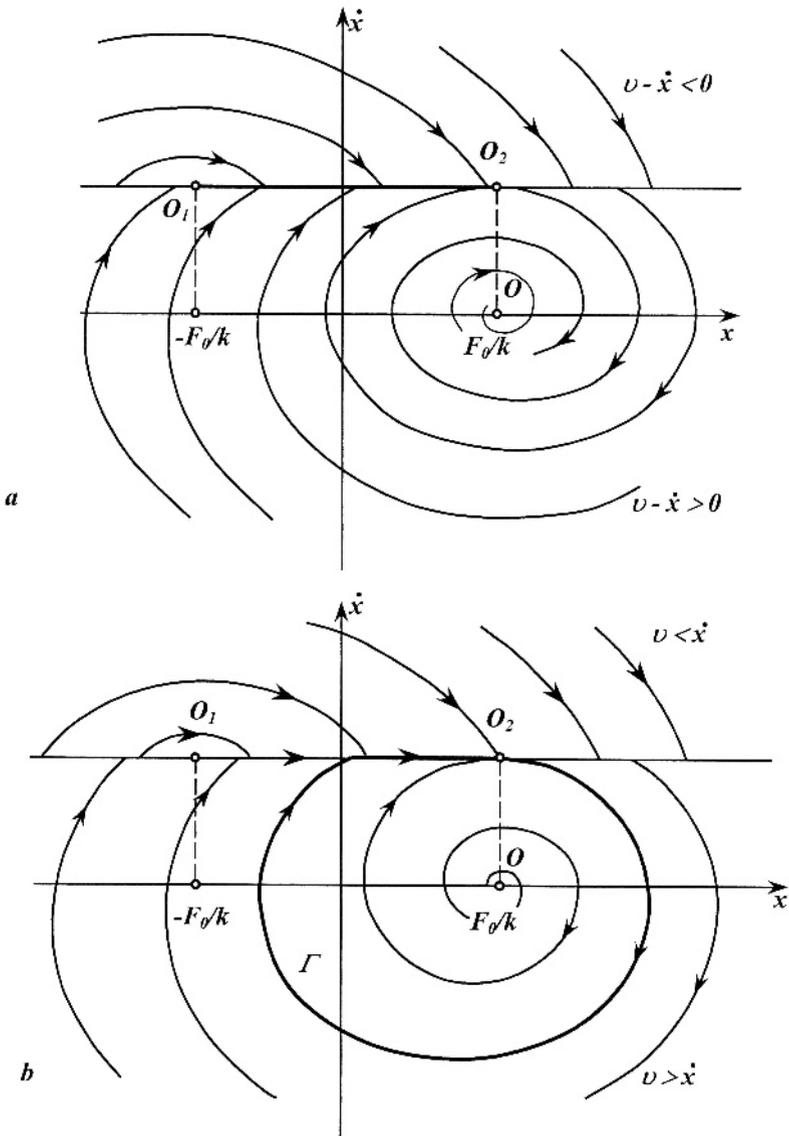
Fig. 19.5. The phase portraits of the Froude-pendulum-like systems for the friction characteristic of the form shown in the cases **a** and **b** of figure 19.4.

Let us start with an investigation of the equation (19.17). As earlier, we assume that

$$\xi = A\cos(\omega t + \varphi) \qquad (19.18)$$

and determine the time derivative of the energy $E = (\dot{\xi}^2 + \omega^2 \xi^2)/2$ for the corresponding linear oscillator as

$$\frac{dE}{dt} = \frac{d}{d\xi}\frac{\omega^2 A^2}{2} = \dot{\xi}\ddot{\xi} + \omega^2 \xi\dot{\xi} =$$

$$= [2\delta(1 - a\xi - b\dot{\xi}^2)\dot{\xi}]\dot{\xi} = \qquad (19.19)$$

$$= 2\delta[1 + a\omega A\sin(\omega t + \varphi) - b\omega^2 A^2 \sin^2(\omega t + \varphi)]\omega^2 A^2 \sin^2(\omega t + \varphi).$$

Recall here that these last calculations are carried out with the terms with $\dot{A}$ and $\dot{\varphi}$ being neglected, i.e. when $\dot{A} = \dot{\varphi} = 0$.

By integrating (19.19) over the period $2\pi/\omega$, we find that

$$\Delta \frac{\omega^2 A^2}{2} = (\delta\omega^2 A^2 - \frac{1}{2}\delta b\omega^4 A^4)\frac{2\pi}{\omega}$$

or

$$\Delta A^2 = \frac{4\pi\delta A^2}{\omega} - 2\pi\delta b\omega\ A^4. \qquad (19.20)$$

From (19.20), there follows the recurrent relation for the subsequent amplitudes A_n and A_{n+1}

$$A_{n+1}^2 = (1 + \frac{4\pi\delta}{\omega})A_n^2 - 2\pi\delta b\omega A_n^4, \qquad (19.21)$$

which can be regarded as a point mapping of a half-plane into itself. To its fixed point $A = 0$ there will correspond an equilibrium state that for $\delta < 0$ will be stable and for $\delta > 0$ unstable. For $b > 0$, there will be also another fixed point

$$A^* = \frac{b\omega^2}{2}, \qquad (19.22)$$

which for $\delta > 0$ will be stable, i.e. when the equilibrium is unstable.

For the friction characteristic of the form in figure 19.4c, which is approximated by the polynomial (19.7), if the velocity v lies on its descending part, then $\delta > 0$ and $b > 0$ and, therefore, an equilibrium state will be unstable, and there will be born the auto-oscillations of the amplitude being approximately determined by the formula (19.22) (Fig. 19.6).
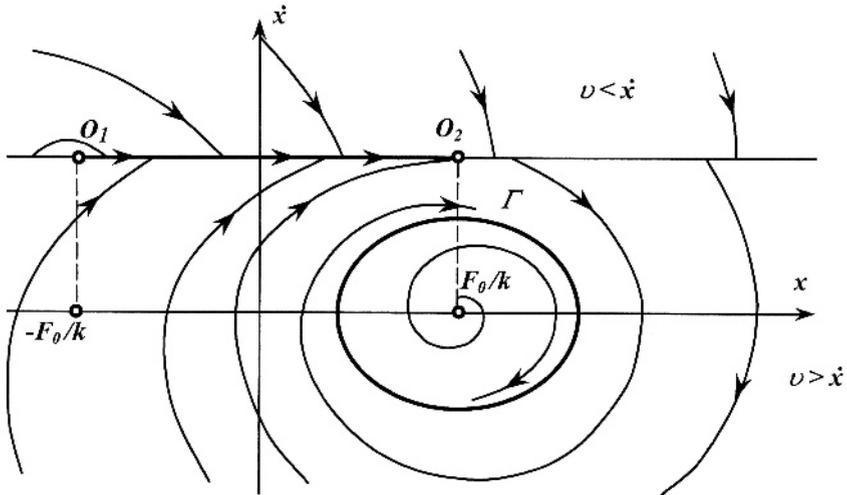
Fig. 19.6. A phase portrait of the Froude-pendulum-like systems when a descending part is available for the friction characteristic shown in figure 19.4c.

One may expect, with some portion of optimism, that the similar situation will occur for any characteristic of the form shown in figure 19.4c (but for small δ).

20 Forced oscillations of a linear oscillator

An amplitude-phase frequency characteristic. Phenomena of a resonance and a phase shift. Examples of a ship pitching and a dynamical oscillations damper. A spectral approach and the superposition principle. The generalized δ-function and the superposition principle.

We have got acquainted with such simplest physical implementations of a linear oscillator as a spring-attached mass (if there is a viscous friction available) and an electric circuit represented by a self-induction, a capacitance and a resistance. Now, we are going to study how these objects are affected by the harmonic force and by voltage, respectively (Fig. 20.1).
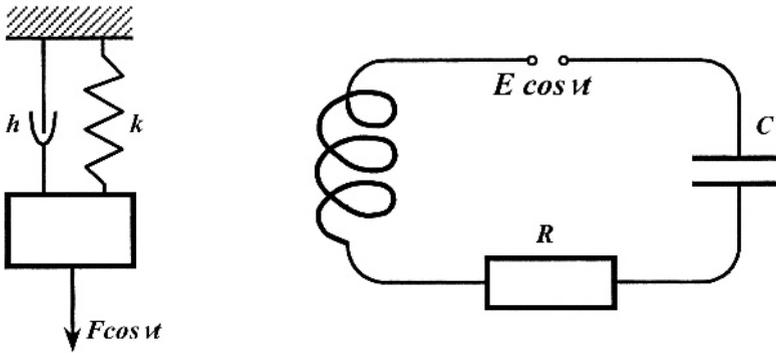
Fig. 20.1. Mechanical and electrical oscillators under external harmonic forces.

In both cases, a mathematical model will be one and the same and of the form

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = a \cos \nu t .\tag{20.1}$$

For the sake of simplicity in this and similar situations, we shall use a complex form of notations, through replacing the equation (20.1) by the equation

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = ae^{i\varkappa t} . \tag{20.2}$$

This equation should be understood in the following way: its real part is the equation under our study (20.1); accordingly, $ae^{i\varkappa t}$ is a complex notation for the real action $a\cos\varkappa t$ and x in (20.1) is a real part of x in (20.2).

A solution of the non-homogeneous equation (20.2) includes a general solution of a homogeneous equation (well-known for us) and any solution of the non-homogeneous equation we are trying to find. We shall search for the solution in the form $x = Ae^{i\varkappa t}$, i.e. as oscillations of the same frequency as the external force $ae^{i\varkappa t}$ has. A direct substitution will give

$$(-\nu^2 + 2\delta\nu i + \omega^2)A = a ,$$

from which

$$A = \frac{a}{\omega^2 - \nu^2 + 2\delta\nu i} . \tag{20.3}$$

If the general solution to the homogeneous equation is presupposed to be disappearing as $t \to \infty$ (this is the case for $\delta > 0$ and $\omega^2 > 0$), then the oscillation, remaining with the unbounded increase of time, will coincide with the particular solution derived of the non-homogeneous equation

$$x = \frac{a}{\omega^2 - \nu^2 + 2\delta\nu i}e^{i\varkappa t} . \tag{20.4}$$

Let us clarify what this solution is. We shall start with a general consideration. Let

$$x = K(i\nu)e^{i\varkappa t} , \tag{20.5}$$

where $K(i\nu)$ is a complex function of the frequency ν. Only the real part of (20.5) has an actual sense. Let us determine it, by preserving its denotation x,

$$x = \mathrm{Re}(K(i\nu)e^{i\varkappa t}) = \mathrm{Re}[|K(i\nu)|e^{i(\varkappa t + \arg K(i\nu))}] =$$

$$= |K(i\nu)|\cos(\varkappa t + \varphi) ,$$

where $\varphi = \arg K(i\nu)$.

Thus, the derived particular solution of (20.4), which will be set with t being increased for the asymptotically stable oscillator, is a harmonic oscillation with the frequency of the external force ν, the amplitude $|K(i\nu)|$ and the phase $\varphi = \arg K(i\nu)$. Therefore, this oscillation is called a forced oscillation.

We invite your attention to the fact that the characteristics of a harmonic oscillator, i.e. its amplitude and phase, are determined as functions of the external force frequency ν by one and the same complex number $K(i\nu)$. Due to this fact, $K(i\nu)$ was called a gain-phase (amplitude-phase) frequency characteristic, i.e. the characteristic which simultaneously unifies the frequency dependencies for amplitude and phase.

The amplitude-phase frequency characteristic may be made geometrically visual. For this purpose, let us draw a locus , i.e. a curve being run on the complex plane W by the complex point $w = K(i\nu)$ as ν is varying from 0 to ∞ (Fig. 20.2).
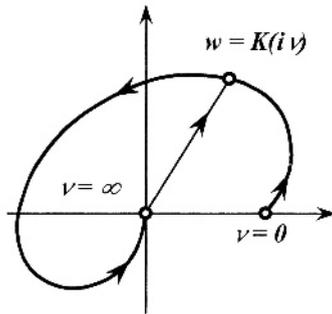
Fig. 20.2. An amplitude-phase frequency characteristic.

The length of the radius-vector for the point w of this curve, corresponding to the frequency ν, and the angle included between this vector and the real axis will be, respectively, the amplitude and the phase for the respective forced oscillation.

The locus relevant to the formula (20.4) can be easily depicted by its three specific points for which $\nu = 0$, $\nu = \omega$ and $\nu = \infty$ (Fig. 20.3).For a small δ, the locus will be strongly elongated along the negative direction of the imaginary axis and for $\nu = \omega$ will take the value $-ia/2\delta\omega$. The locus begins in the point $w = a/\omega^2$ and terminates in the point $w = 0$.

With the amplitude-phase frequency characteristic (Fig. 20.3) being known, one can easily plot an amplitude dependence, i.e. a resonance curve, and a phase dependence upon the external disturbance frequency ν. These dependencies for different $\delta > 0$ are qualitatively presented in figure 20.4. From them there follows a well-known phenomenon of the amplitude resonance, i.e. a large growth of the amplitude near $\nu = \omega$ for $\delta \ll 1$, and a less known, but not less remarkable, phenomenon of the antiphase for forced oscillations with respect to the external disturbance for $\nu > \omega$. For $\delta = 0$, the oscillation phase depends

discontinuously upon the frequency v: for $v < \omega$, it is equal to zero, and for $v > \omega$, it is equal to $-\pi$. For not large $\delta > 0$, this dependence is already continuous, but with a fast change of phase φ close to $v = \omega$ (Fig. 20.5).
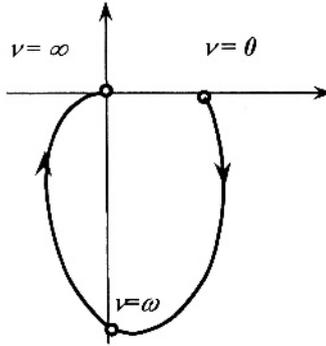
Fig. 20.3. An amplitude-phase frequency characteristic for the linear oscillator.
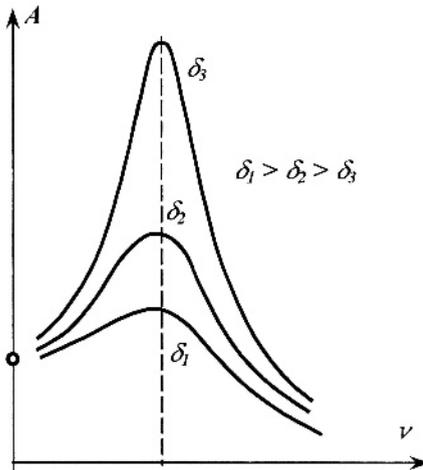
Fig. 20. 4. Amplitude frequency characteristics (resonance curves) for three decreasing values of the damping coefficient δ (viscous friction or ohmic resistance).

The resonance is distinguished the more vividly, the less damping δ is. For small δ, the maximal amplitude is attained at $v = \omega$ approximately, and, more

precisely, at $\nu = \sqrt{\omega^2 - 2\delta^2}$, because it is at this very value of the external disturbance frequency that the maximum of the module of A , defined by the formula (20.3), is attained (Fig. 20.4).
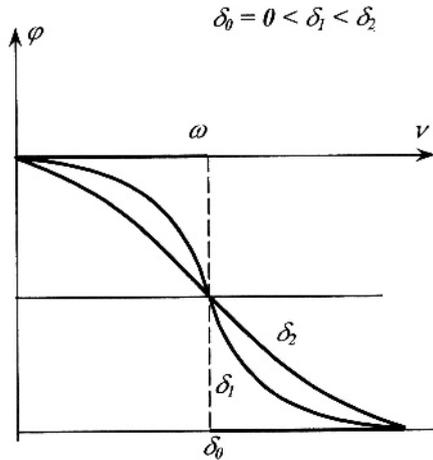
Fig. 20.5. Phase frequency characteristics for three decreasing values of the damping coefficient δ (viscous friction or ohmic resistance) .

These both phenomena can be brightly illustrated by ship's rolling and pitching when meeting the waves. In both cases, as a approximate model for ship's rolling or pitching there serves an oscillator subject to the action of some external harmonic force; A.N. Krylov was the first to pay attention to this fact. Of a special interest here is ship's pitching , when for a sufficiently large amplitude of disturbing waves, the frequency of pitching oscillations depends upon the velocity and the course of the ship. If the course and the velocity are unsuccessful, then the onrunning waves can be of the frequency somewhat exceeding the proper frequency of the pitching oscillations; then the resonance phenomenon will be accompanied with the fact that the ship will be not climbing upon the onrunning waves, but will be falling down and burying itself into them.

Let us find the conditions under which there occurs a birth of such an undesirable and dangerous pitching. Let θ be an angle formed between directions of a wave motion of the velocity u and the ship moving at the velocity υ against the wave. Here, the velocity of the onrunning waves is equal to $u + \upsilon \cos \theta$ (Fig. 20.6), and the frequency of their run-ons is equal to $\nu = 2\pi\lambda^{-1}(u + \upsilon \cos \theta)$, where λ is a length of the onrunning waves.

Let ω be a proper frequency of free ship oscillations. Then especially dangerous will be the situation when v and ω are close to each other and $v > \omega$, since it is the very case that can bring the resonance, and the oscillations of the bow will be in antiphase with respect to the onrunning waves.
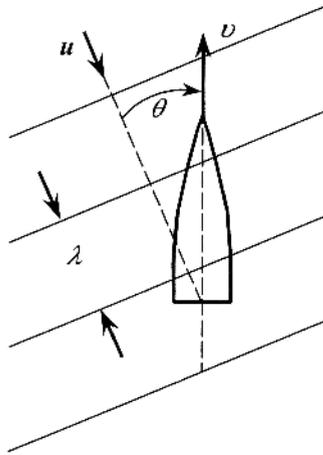
Fig. 20. 6. A course of a ship moving at the velocity v at the angle θ to the wave onrunning at the velocity u .

The following example of forced oscillations of a system is somewhat more complicated than an oscillator. Its main part constitutes a mechanism mounted on some elastic base and subject to a harmonic force arising during its work. The mathematical model of this system coincides with (20.1). The periodic disturbance in this system can be caused by, say, eccentrically revolving components or parts moving to and fro at some frequency v. Under the action of this force, the mechanism begins to perform very undesirable vibrations.

How to suppress or at least decrease these vibrations. One of the effective means for suppressing them is a dynamic damper that represents a mass with a spring attached to a basic large mass of the vibrating mechanism (Fig. 20.7).

Let x and y be displacements of the mechanism and the damper mass, respectively, from their equilibrium positions, and M and m be masses of the mechanism and the damper, K and k be a stiffness of the mechanism base (fastening) and a damper spring stiffness. A mathematical model for the mechanism on an elastic base and with a dynamic damper (Fig. 20.7) can be written in the form

$$M\ddot{x} = -Kx - k(x-y) + Fe^{i\mu t},$$

$$m\ddot{y} = k(x-y). \tag{20.6}$$
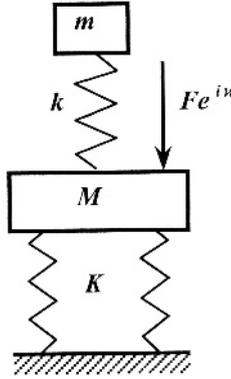
Fig. 20.7. A scheme for a mechanical resonance damper of forced oscillations.

For the sake of simplicity, frictions are not taken into account by this model. Of course, for a real calculation, the frictions should be considered, though for our demonstrative purposes they are not essential and will just make our analysis more complicated.

The equation (20.6) can be written on the basis of either Newton laws immediately or the Lagrange function that, in the given case, is equal to

$$L = \frac{M\dot{x}^2}{2} + \frac{m\dot{y}^2}{2} - \frac{Kx^2}{2} - \frac{k(x-y)^2}{2}$$

and, besides, there are the forces

$$Q_x = Fe^{i\mu t}, \quad Q_y = 0.$$

Let us find the forced oscillations of the model (20.6). In accordance with the preceding, assume that

$$x = Ae^{i\mu t}, \quad y = Be^{i\mu t} \tag{20.7}$$

and insert them into (20.6). Thus, we obtain

$$-M\nu^2 A = -KA - k(A-B) + F,$$

$$-m\nu^2 B = k(A-B), \tag{20.8}$$

from which it follows that

$$A = \frac{F(k-m\nu^2)}{(K+k-M\nu^2)(k-m\nu^2)-k^2},$$

$$B = \frac{kF}{(K+k-M\nu^2)(k-m\nu^2)-k^2}. \tag{20.9}$$

From these formulae it follows that the forced oscillations of a base-mounted mechanism will have the amplitude A, being nullified for

$$k-m\nu^2 = 0. \tag{20.10}$$

Under this condition, the mass of the damper oscillates with the amplitude $B = -F/k$, so that the springs of the damper act upon the mechanism mass with the force $kBe^{i\nu t} = -Fe^{i\nu t}$, compensating exactly the effect $Fe^{i\nu t}$. Since, according to (20.10), $k = m\nu^2$ and the mass m is small, the amplitude of the damper oscillations can be large and sufficient to neutralize the influence of the harmonic force upon the mechanism.

This is a remarkable and surprising result: the damper is oscillating, but the mechanism is still not vibrating. However, how did the damper begin vibrating? Apparently, there can be a unique answer here: the mechanism was vibrating in the beginning and caused oscillations of the damper; then the damper oscillations suppressed the mechanism oscillations and, as a result, they disappeared; upon it only damper oscillations remained. This looks truthfully, but it is not the case, since one may speak about forced oscillations as steady-state oscillations only upon the proper oscillations having damped, i.e. when friction is available. Therefore, this can happen only if vanishingly small frictions are present. They exists always, of course, and all the above is true ultima analysi. However, if the frictions are very small, the mechanism vibrations are not fully absent, but retain very small.

We have thus considered the action exerted by a harmonic force upon a linear oscillator, determined the forced oscillations caused by it and have revealed the phenomena of synchronism, resonance and a phase shift.

Synchronism becomes apparent because of the fact that a harmonic action of the frequency ν induces harmonic forced oscillations of the same frequency. The phenomena of resonance manifest themselves the more heavily, the smaller the friction parameter $\delta > 0$ is, and, therefore, the amplitude of the forced oscillations is growing unboundedly as $\delta \to 0$.

The phenomenon of a phase shift is its dependence upon the action frequency ν. This dependence is such that, for small frequencies, the phase shift is near to zero, and for large frequencies, it is near to $-\pi$, i.e. to the antiphase. For $\delta \to 0$, this dependence assumes a jump-like nature: for $\nu < \omega$, the phase shift is equal to zero, and for $\nu > \omega$, it is equal to $-\pi$.

We have also introduced   a notion of the amplitude-phase frequency characteristic $K(i\nu)$ in such a way as to have the relationship between the harmonic disturbance $ae^{i\nu t}$ and the forced oscillations $Ae^{i\nu t}$ written in a very simple fashion

$$A = K(i\nu)a \ , \tag{20.11}$$

that reminds the lever law. Of course, from this notation the phenomena of synchronism, resonance and a phase shift follow. Such is the success of a good theory: the complicated becomes simple.

Now I would like to do one important addition to this brief story of our research. The formula (20.8) relates to a harmonic disturbance, and would not it be possible to generalize it to an arbitrary periodic force? This turns out to be possible. This becomes possible thanks to the superposition principle and the representability of any periodic disturbance as a sum of harmonic disturbances. You know this representability from the mathematical analysis, where it is called the Fourier series.

According to the theory of Fourier series, the periodic function $f(t)$ of the period $T = 2\pi/\nu$, under some non-limiting for us assumptions, can be represented as follows

$$f(t) = \sum a_s \cos s\,\nu t + b_s \sin s\,\nu t \ , \tag{20.12}$$

where

$$a_0 = \frac{\nu}{2\pi}\int_0^T f(t)dt \ , \qquad a_s = \frac{\nu}{\pi}\int_0^T f(t)\cos s\,\nu t \, dt \ ,$$

$$b_s = \frac{\nu}{\pi}\int_0^T f(t)\sin s\,\nu t dt \ , \qquad (s = 1,2,...) \ ,$$

or as follows

$$f(t) = \mathrm{Re}\sum_{s=0}^{\infty}(a_s - ib_s)e^{is\nu t} \ . \tag{20.13}$$

The last formula (20.13) implies that a periodic disturbance is a sum of harmonic disturbances (in the complex notation)

$$f(t) = \operatorname{Re} \sum_{s=0}^{\infty} c_s e^{is\nu t}, \quad c_s = a_s - ib_s. \tag{20.14}$$

Such a representation of the function $f(t)$ is called spectral, and the frequencies of its harmonic components constitute its spectrum. The constant disturbance $c_0 = a_0$ is of the zero frequency. The harmonic component of the frequency ν is called fundamental, and the rest disturbances, named as harmonics, have the frequencies $2\nu, 3\nu, \dots$. The amplitudes of these harmonic components, the fundamental and the harmonics, will be equal to $|c_1|, |c_2|, \dots$, respectively.

Now, in order to find the forced oscillations corresponding to an arbitrary periodic disturbance $f(t)$, it is sufficient to apply the superposition principle. To a separate harmonic disturbance involved in (20.14) there will correspond the below forced oscillation

$$c_s K(is\nu) e^{is\nu t}, \tag{20.15}$$

and, therefore, to the sum of disturbances (20.14), i.e. to the disturbance $f(t)$, there will correspond a sum of the forced harmonic oscillations (20.15), i.e.

$$\sum_{s=0}^{\infty} c_s K(is\nu) e^{is\nu t}. \tag{20.16}$$

Corresponding to (20.16) are the harmonic oscillations with frequencies $s\nu$, amplitudes $|c_s K(is\nu)|$ and phases $\arg[c_s K(is\nu)]$. Here, the phenomenon of cynchronism remains   in the sense that the oscillator response is of the same period as the disturbance has; and the phenomenon of a phase shift under the periodic disturbance (20.14) becomes senseless; whereas the phenomenon of resonance retains. However, a resonance becomes now possible not only at $\nu = \sqrt{\omega^2 - 2\delta^2}$, but also at $s\nu = \sqrt{\omega^2 - 2\delta^2}$ for any integer s, i.e. not only at the fundamental frequency, but at its harmonics as well. However, since $c_s \to 0$ as $s \to \infty$, the phenomenon of resonance can arise only at not very distant harmonics in the case $\delta > 0$. For $\delta = 0$, it is possible at any harmonic.

The above said concerning an external periodic disturbance of the oscillator can be generalized to an arbitrary disturbance $f(t)$. This further generalization is also based on the superposition principle, but follows from the more general techniques of representing the disturbances $f(t)$ as a superposition of simpler disturbances. Here, under simpler disturbances there are implied those for which a response of the oscillator can be easily found. In the above case, such a simple

disturbance was the harmonic signal $e^{i\nu t}$. The response to this disturbance was also harmonic and equal to $K(i\nu)e^{i\nu t}$, where $K(i\nu)$ is the amplitude-phase frequency characteristic of the oscillator. We used then the fact that the periodic disturbance $f(t)$ of the period T can be represented as a Fourier series, i.e. as a superposition of the harmonic oscillations of the frequencies $2\pi n / T$, $n = 1,2,3,...$ and some constant disturbance. But except the Fourier series, there also exists the Fourier integral that allows to represent any function $f(t)$, decreasing sufficiently fast as $t \rightarrow \pm\infty$, in the form

$$f(t) = \int_{-\infty}^{\infty} g(\nu)e^{i\nu t}d\nu, \qquad (20.17)$$

i.e. also as a superposition of the harmonic disturbances $g(\nu)e^{i\nu t}d\nu$, having frequency ν, amplitude $|g(\nu)d\nu|$ and phase $\arg[g(\nu)d\nu]$. Here, the function $g(\nu)$, called the Fourier transformation for the function $f(t)$, is known to be determined by the formula

$$g(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)e^{-i\nu t}dt. \qquad (20.18)$$

The response of the oscillator to each harmonic component $g(\nu)e^{i\nu t}d\nu$ of the integral sum (20.17) will be $K(i\nu)g(i\nu)e^{i\nu t}d\nu$, and the response to the entire disturbance $f(t)$ will be

$$x(t) = \int_{-\infty}^{\infty} K(i\nu)g(\nu)e^{i\nu t}d\nu. \qquad (20.19)$$

Alongside with the spectral representation of the external disturbance $f(t)$, there may also exist an impulse representation, i.e. representing the disturbance $f(t)$ as a sequence of instantaneous impulse disturbances. At the instant t, for the time interval dt, the magnitude of this impulse is equal to $f(t)dt$, where the duration dt is vanishingly small. It is impossible to assign such an impulse through common functions, and this is attained with use of a generalized function that was called the delta function. An instantaneous unit impulse at the instant τ is written as $\delta(t - \tau)$. To conceive this description, i.e. this mathematical model for an instantaneous unit impulse, is possible, if we imagine it as a limit of a pick-like function $\varphi(t - \tau, \varepsilon)$, shown in figure 20.8, as $\varepsilon \rightarrow 0$. It is also assumed here that, in spite of the fact that the domain of the width ε near the

point τ, in which the function $\varphi(t - \tau, \varepsilon)$ differs from zero, is narrowing unboundedly, the integral

$$\int_{-\infty}^{\infty} \varphi(t - \tau, \varepsilon)dt = \int_{\tau - \varepsilon/2}^{\tau + \varepsilon/2} \varphi(t - \tau, \varepsilon)dt \qquad (20.20)$$

will remain all the time equal to unity. It is clear that such a requirement leads to an unbounded growth of the function   as $\varepsilon \to 0$ and to the absence of a limiting function as an ordinary notion. Let us bypass this difficulty by having called the limiting function a generalized function and by writing it in the form $\delta(t - \tau)$. With the help of this generalized function, any disturbance f(t) can be then written as follows

$$f(t) = \int_{-\infty}^{\infty} f(\tau)\delta(t - \tau)d\tau. \qquad (20.21)$$
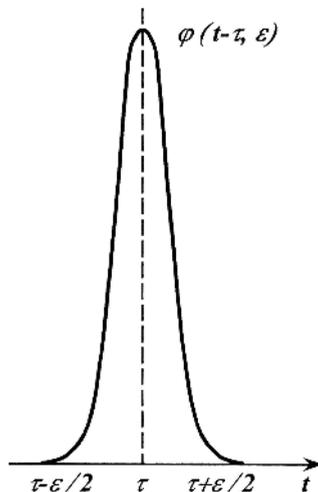
Fig. 20.8. A pike-like disturbance by a unit impulse transferring, as $\varepsilon \to 0$, to the generalized δ-function $\delta(t - \tau)$.

Indeed, according to the mean value theorem, one has

$$\int\limits_{-\infty}^{\infty} f(\tau)\delta(t-\tau)d\tau = \lim_{\varepsilon \to 0} \int\limits_{\tau-\varepsilon/2}^{\tau+\varepsilon/2} f(\tau)\varphi(t-\tau,\varepsilon)d\tau =$$

$$= \lim_{\varepsilon \to 0} f(\overline{\tau}) \int\limits_{\tau-\varepsilon/2}^{\tau+\varepsilon/2} \kappa(t-\tau)d\tau = f(t).$$

Recall that $\overline{\tau}$ is a value lying within the narrowing interval $(t-\varepsilon/2, t+\varepsilon/2)$.

Through the formula (20.21) an external disturbance is represented as a superposition of impulse disturbances being described by the generalized function $\delta(t-\tau)$. Let us denote a response of the oscillator to a unit impulse through $\psi(t-\tau)$. Then, by the superposition principle, the response $x(t)$ of the oscillator to any disturbance $f(t)$ can be written as

$$x(t) = \int\limits_{-\infty}^{\infty} f(\tau)\psi(t-\tau)d\tau. \tag{20.22}$$

Now, we need only to find the response $\psi(t-\tau)$ of the oscillator to the unit impulse $\delta(t-\tau)$, i.e. to the disturbance $f(t)=\delta(t-\tau)$. Before starting a search for this response, let us specify the conditions for its action and also what is understood under a response of the oscillator to the disturbance $f(t)$.

The matter is that a motion of the oscillator is determined by its initial state and by a subsequent external disturbance. Hence, under a response to an external disturbance it is natural to understand only its component which corresponds to this very external disturbance, but not to the initial disturbance of the oscillator born by the nonzero initial conditions. Thus, under $\psi(t-\tau)$ we understand the result of the action exerted by the impulse $\delta(t-\tau)$ under the zero initial conditions, i.e. we want to solve the following problem: to find a solution of the differential equation

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = \delta(t-\tau) \tag{20.23}$$

such that

$$x(\tau-\xi) = 0, \quad \dot{x}(\tau-\xi) = 0, \quad (\xi > 0). \tag{20.24}$$

Let us find this solution informally and then its correctness may be checked by means of a direct substitution.

An instantaneous impulse yields a jump of the velocity $\dot{x}$, but the coordinate x is not changed, since an impulse duration is infinitely small and the velocity $\dot{x}$ is finite. Therefore, by integrating (20.23) from $\tau - \varepsilon/2$ to $\tau + \varepsilon/2$ and by letting $\varepsilon \to 0$, we have

$$\Delta\dot{x} = \dot{x}(\tau+0) - \dot{x}(\tau-0) = 1,$$

from which, in view of $\dot{x}(\tau-0) = 0$, it follows that $\dot{x}(\tau+0) = 1$. In addition, we have $x(\tau+0) = x(\tau-0) = 0$.

Thus, as a response to the disturbance $\delta(t-\tau)$ there will be a solution of the equation

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = 0$$

which is equal to zero for $t < \tau$ and, for $t \ge \tau$, is determined by the initial conditions

$$x(\tau) = 0, \quad \dot{x}(\tau) = 1.$$

According to (13.7), this solution for $\omega^2 - \delta^2 > 0$ is written as

$$x(t) = \psi(t-\tau) = \begin{cases} 0 & for \quad t \le \tau \\ e^{-\delta(t-\tau)}\dfrac{1}{\Omega}\sin\Omega(t-\tau) & for \quad t \ge \tau \end{cases} \qquad (20.25)$$

In conclusion, I would like to note that the response formula (20.22) derived (with (20.25) accounted) is a direct consequence of the employment of the new, unusual for you, generalized function $\delta(t)$ possessing the following properties

$$\delta(t) = \begin{cases} 0 & for \quad t \ne 0 \\ \infty & for \quad t = 0 \end{cases}, \qquad (20.26)$$

$$\int_{-\infty}^{\infty} \delta(t)dt = 1.$$

In spite of uncommonness of the function $\delta(t)$, it is nothing, but a natural mathematical model for an instantaneous impulse disturbance. Simultaneously, it is also a mathematical model for density of a unit point mass on a line or that of a unit point charge. For the first time this function was introduced by Heavyside, in connection with the operational calculus constructed by him, and later on this function was introduced by physicist Dirac into the quantum physics developed by him. Physicists called it the Dirac function and mathematicians the generalized δ − function or simply the δ − function.

21 Parametric excitation and stabilization

Phenomena of a parametric excitation and a resonance. Peculiarities of the parametric resonance distinguishing it from a usual resonance.

In the previous chapters we have considered a linear oscillator under some external harmonic force. It was described by the below differential equation

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = a\cos \varkappa t . \tag{21.1}$$

However, a time-dependent disturbance can also yield other changes of the linear oscillator equation. For instance, these changes may cover the harmonic variations of the frequency ω^2 in time, and then the equation of the oscillator will be as

$$\ddot{x} + 2\delta\dot{x} + (\omega^2 + \varepsilon\cos \varkappa t)x = 0 . \tag{21.2}$$

As an immediate physical model for the equation (21.2) there serves a spring-attached mass where a spring has a harmonically changing stiffness and a viscous friction. Besides, the equation (21.2) describes, with some reasonable approximation, small oscillations of a pendulum where a point of suspension is vibrating vertically under the harmonic law. If the suspending point is vibrating horizontally, then we come to the equation (21.1) with a similar approximation. Let us now consider these both cases by supposing the pendulum suspending point to be displaced in the horizontal and vertical directions under the laws $u(t)$ and $v(t)$. The pendulum motion equations are constructed with use of the Lagrange equation. Let us express the coordinates x and y of the pendulum mass m through the displacements $u(t)$, $v(t)$, and the angle φ of the pendulum deviation from the vertical direction (Fig. 21.1):

$$x = u + l\sin\varphi , \quad y = \upsilon + l\cos\varphi$$

and write the Lagrange function

$$L = \frac{m}{2}\{(\dot{u} + l\dot{\varphi}\cos\varphi)^2 + (\dot{\upsilon} - l\dot{\varphi}\sin\varphi)^2\} + mg(l\cos\varphi + \upsilon) .$$

After not so complicated computations we shall find the desired Lagrange equation in the form

$$\ddot{\varphi} + \frac{g}{l}\sin\varphi = -\frac{1}{l}\ddot{u}\cos\varphi + \frac{1}{l}\ddot{\upsilon}\sin\varphi$$

or, taking into account the viscous friction as well and replacing $\sin\varphi$ by φ and $\cos\varphi$ by 1, in the form

$$\ddot{\varphi} + 2\delta\dot{\varphi} + \omega^2\varphi = -\frac{1}{l}\ddot{u} + \frac{1}{l}\ddot{\upsilon}\varphi \qquad (21.3)$$

The equation (21.3) is reduced to (21.1) and (21.2), respectively, for the harmonic horizontal $(u = al\nu^{-2}\cos\varkappa, \upsilon = 0)$ and vertical $(u = o, \upsilon = \varepsilon l\nu^{-2}\cos\varkappa)$ oscillations of the suspending point.
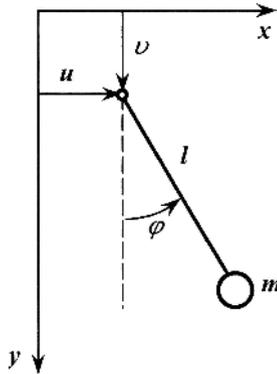
Fig. 21.1. A pendulum with a displaceable suspending point.

Thus, if the pendulum suspending point is harmonically moved   in the horizontal direction, then this will be an equivalent to the action exerted upon the pendulum by a harmonic force. And if the suspending point is harmonically displaced in the vertical direction, then this will be equivalent to a harmonic variation of the pendulum frequency. This will make it possible for us to visually observe what is happening to the pendulum in either cases described by the differential equations (21.1) and (21.2). Let us take this occasion in order to demonstrate once more the phenomena of synchronism, phase shift and resonance (they were described in the previous chapter), and then, in the end of this chapter, we will derive and study the differential equation (21.2). For demonstrating I

would like to use one of your handbags. I am holding the bag by my hand and now start replacing my hand periodically to the right and to the left. The handbag begins to oscillate in time with the hand exhibiting the phenomenon of resonance. Resonance appears at the time when the handbag is swinging very much whereas the displacements of the hand are practically not distinguished. When the frequency of a hand motion is larger than the proper frequency of handbag oscillations, the handbag is oscillating to the left, whereas the hand is displaced to the right; and with the hand motions being made at greater frequency the hand and the bag will be oscillating simultaneously to the right and to the left.

Let us return to the equation (21.2). By changing the variables $x = e^{-\delta t} y$, one may cancel the term of the viscous friction and so reduce this equation to the form

$$\ddot{y} + (\Omega^2 + \varepsilon \cos \varkappa t) y = 0,$$

where $\Omega^2 = \omega^2 - \delta^2$. In its turn, a replacement of the time scale $\tau = \varkappa t$, allows to retain only two parameters in this equation, by reducing it to the well-known Mathieu equation

$$\ddot{y} + (v^{-2} \Omega^2 + \varepsilon v^{-2} \cos \tau) y = 0,$$

which we write in the initial variables as follows

$$\ddot{x} + (\omega^2 + \varepsilon \cos t) x = 0. \tag{21.4}$$

To study the Mathieu equation (21.4) is not very simple, and it may serve as an example to show how a good idea generates a technique of research; the research itself results in calculations and the calculations bring us to clear-cut and surprising results .

This idea is coming yet from great mathematician J.H. Poincare and consists in reducing a study of the Mathieu equation to a point mapping. The technique and the calculations consist in constructing, studying and numerically computing this point mapping. Let x and $\dot{x} = y$ be the initial conditions for $t = 0$, $\overline{x}$ and $\overline{y}$ be the values of the same variables upon the time 2π. The values $\overline{x}$ and $\overline{y}$ are uniquely determined by x and y so that

$$(\overline{x}, \overline{y}) = T(x, y), \tag{21.5}$$

where T is some unknown unique operator transferring x and y into $\overline{x}$ and $\overline{y}$. The relation (21.5) can be also interpreted as a mapping of the point (x, y) into the point $(\overline{x}, \overline{y})$. We have already encountered such an interpretation while discussing the dynamics of the Galileo-Huygens clock and the "contrary clock" (a stochastic oscillator). These were the Keniks-Lamerey diagrams, which geometrically represented one-dimensional point mappings. Now, we are having a two-dimensional point mapping. Unfortunately, the geometrical representation being also so much convenient is not familiar for this two-dimensional mapping.

But even without it, the application of the point mapping turns out to be very useful. Its usefulness consists in the fact that the continuation of the solution to the Mathieu equation (21.2) with the initial conditions x, y up to the times $2\pi, 4\pi.6\pi,\ldots$ will be reduced to a single-, double-, three-time transformations (21.5), i.e. for any positive integer n and for any solution $x(t), y(t) = \dot{x}(t)$ of the Mathieu equation (21.2), the following will take place:

$$(x(2n\pi), y(2n\pi)) = T^{n}(x(0), y(0)).\qquad(21.6)$$

This trivial fact helps us to reduce a study of behaviour of solutions of the Mathieu equation (21.2) for unboundedly increasing time to a study of iterations (successive transformations) for the point (x, y).

    Thus, the idea consists in studying the point mapping T of the variable plane x, y into itself, instead of studying the Mathieu equation (21.2) and its three-dimensional phase space. Let us find the mapping T. For this purpose, one needs to integrate the Mathieu differential equation (21.4). Suppose we have done it for the initial conditions $x = 1$ and $y = 0$, and for $x = 0$ and $y = 1$; these solutions being denoted by $\varphi_1(t)$ and $\varphi_2(t)$. Then, due to linearity of the Mathieu equation, its solution for arbitrary initial conditions x, y is written as

$$x(t) = x\varphi_1(t) + y\varphi_2(t)$$

and

$$y(t) = \dot{x}(t) = x\dot{\varphi}_1(t) + y\dot{\varphi}_2(t).$$

Hence,

$$\bar{x} = x(2\pi) = \varphi_1(2\pi)x + \varphi_2(2\pi)y$$

$$\bar{y} = \dot{x}(2\pi) = \dot{\varphi}_1(2\pi)x + \dot{\varphi}_2(2\pi)y,\qquad(21.7)$$

i.e. we have found the mapping T and it turned out to be linear. For it, we have used only two particular solutions, $\varphi_1(t)$ and $\varphi_2(t)$, or, more precisely, only their values and values of their derivatives at $t = 2\pi$.

    How to research a linear mapping is familiar for us and we know about its structure too. Let us apply this knowledge. Most important in this study are eigenvalues of the matrix of the linear mapping T, i.e. roots of the quadratic equation

$$\begin{vmatrix} \varphi_1(2\pi) - \lambda & \varphi_2(2\pi) \\ \dot{\varphi}_1(2\pi) & \dot{\varphi}_2(2\pi) - \lambda \end{vmatrix} = \qquad(21.8)$$

$$= \lambda^2 - (\varphi_1(2\pi) + \dot{\varphi}_2(2\pi))\lambda + \varphi_1(2\pi)\dot{\varphi}_2(2\pi) - \dot{\varphi}_1(2\pi)\varphi_2(2\pi) = 0 .$$

This equation is not of a general form, since its free term is equal to unity. Indeed, let

$$\varphi_1(t)\dot{\varphi}_2(t) - \dot{\varphi}_1(t)\varphi_2(t) = W(t) .$$

For $t = 0$, we obtain $W(t) = 1$. Further, it is easy to check that $\dot{W}(t) = 0$ and, hence, $W(2\pi) = 1$, i.e. the free term is equal to unity. Thus, the characteristic equation (21.8) is of the form

$$\lambda^2 - 2A\lambda + 1 = 0 , \tag{21.9}$$

where

$$2A = \varphi_1(2\pi) + \dot{\varphi}_2(2\pi) .$$

The roots λ_1 and λ_2 of this equation equal to

$$\lambda_{1,2} = A \pm \sqrt{A^2 - 1}$$

are real and different for $|A| > 1$, one of them being larger and another less than unity (since its product is equal to unity); for $|A| < 1$, these roots are complex conjugate and in absolute value equal to unity. The transition from the former case to the last one occurs for $|A| = 1$: for $A = 1$, through the double root equal to unity, and for $A = -1$, through the double root equal to minus unity.

   As known, the linear mapping T , in case of different roots, i.e. for $|A| \neq 1$, may be represented, after an uniform (possibly, complex) linear change of variables, in a diagonal form

$$\bar{u} = \lambda_1 u , \quad \bar{v} = \lambda_2 v . \tag{21.10}$$

For $|A| > 1$, the roots λ_1 and λ_2 are real, and the change of variables is also real, and, according to (21.10), the mapping T will be a contracting one in one direction and an expanding one in another direction. For $A < -1$, the expansion and compression are accompanied with a symmetric mapping with respect to the fixed point $x = y = 0$, coinciding with the point $u = v = 0$. It is clear that, in this case, almost all the points x, y under subsequent transformations will be receding to infinity. The only exception will constitute the points of the line $u = 0$ and $v = 0$, along which a compression occurs (corresponding to $|\lambda| < 1$).

Therefore, for $|A| > 1$, the equilibrium state $x = y = 0$ of the Mathieu equation is unstable, and almost all its solutions are increasing unboundedly in time. This is a so-called parametric excitation of oscillations under which oscillations are increasing unboundedly.

For $|A| < 1$, the situation is quite another. In this case, λ_1 and λ_2, as well as the variables u and v, are complex conjugate. In order to pass to real variables, one should change the variables:

$$\xi = \frac{u+v}{2}, \quad \eta = \frac{u-v}{2i} .$$

Since $|\lambda_1| = |\lambda_2| = 1$ and, hence, $\lambda_1 = \cos\varphi - i\sin\varphi$ and $\lambda_2 = \cos\varphi + i\sin\varphi$ for some φ, the mapping (21.10) and the initial mapping (21.7) will be of the following form in the new variables:

$$\overline{\xi} = \xi\cos\varphi - \eta\sin\varphi, \quad \overline{\eta} = \xi\sin\varphi + \eta\cos\varphi,$$

i.e. this is a rotation around the point $x = y = 0$ at the angle φ. It means that, with time being unboundedly increasing, any solution $x(t), y(t) = \dot{x}(t)$ of the Mathieu equation remains bounded and its equilibrium $x = y = 0$ is stable.

Thus, for $|A| > 1$ there occurs instability of exponential increase, and for $|A| < 1$ a stability without approaching the equilibrium. Here we allowed ourselves to leave the boundary case $|A| = 1$ beyond our attention.

The value of A is determined by the parameters ω^2 and ε of the Mathieu equation. To find this dependence analytically is difficult, but for any ω^2 and ε one may compute a corresponding value of A. This can be easily done on a computer. Imagine that we have made it and found the domains in the plane of parameters ω^2 and ε where $|A| > 1$ and where $|A| < 1$. They are shown in figure 21.2. There are some details here that cannot be revealed in a computational way, and namely, this is a behaviour of bounding curves in the vicinity of the ε-axis. Do they arrive to it or terminate very close to it? And how exactly do they behave near it? These are the very questions that can be studied analytically, since in this study it is possible to restrict ourselves to small ε. This analytical investigation has been done, and figure 21.2 shows its result: the "noses" of the instability domains set their sharpened ends against the ω^2-axis in the points for which $\omega^2 = n / 4$ ($n = 1, 2, \ldots$).

What follows from it concerning the motions of the oscillator described by the Mathieu equation? From here there follows not only a surprising phenomenon of

the parametric excitation arising under the influence of a harmonically varying parameter but also an ability of this excitation at indefinitely small amplitudes to change the parameter, provided that the frequency ω of the initial reduced unperturbed oscillator is sufficiently close to one of the values $n/2$ (this is certainly true for $\omega^2 > 0$).
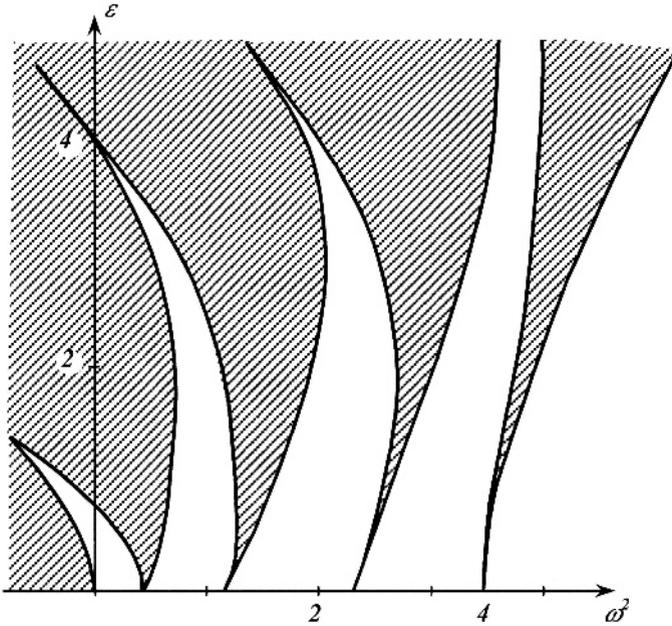
Fig. 21.2. Domains in parameters ω^2 and ε for the parametric resonance (shaded) and stability of the parametrically excitable oscillator described by the Mathieu equation (21.4).

The parametric excitation is a surprising phenomenon, because it can occur also at the time when the oscillator is stable for each fixed value of the varying parameter. Here, a more surprising thing is that an excitation can occur for indefinitely small changes of the parameter. However, it occurs only under sharply defined conditions for the oscillator frequency, and more precisely, for the relation between the frequency ω of the initial oscillator (21.9) and the frequency ν of varying the parameter ε. This last phenomenon is of the same type as a phenomenon of resonance under forced oscillations, and that is why it was called a parametric resonance. But, whereas a usual resonance (for $\delta = 0$) occurs with the oscillating frequencies and an external disturbance being equal, the parametric resonance, instead, occurs under countable series of relations between frequencies:

$$\omega = \frac{n}{2}\nu \ , \ (n = 1,2,...) .$$ (21.11)

Most expressively the parametric resonance is exposed when $n = 1$, i.e. when the ratio between the frequencies ω of the oscillator and ν of the parametric action is equal to $1/2$. It is due to this very resonance that we are swaying to and fro on a swing, by squatting and standing up two times more frequently than oscillations of the swing. These squats and standups bring changes to a pendulum length, which is approximately described by the Mathieu equation as well. "To apprehend" this basic parametric resonance is possible from the energy point of view, since a periodic change of stiffness or a variation of the pendulum length needs some consumption of work which will transfer to an oscillating energy of the pendulum. To perform this energy "pump-up", you will need to stand up for a moment when you are mostly pressed down and to squat when this pressure becomes less, i.e. in the bottom to stand up and in your upper position to squat. It is clear here that, in doing so, we are performing a work and, therefore, swinging the pendulum.

The parametric excitation for a main resonance can be treated as a transformation of the oscillations of the frequency ν into the oscillations of the frequency $\nu/2$. The scaling of frequency does not happen for a usual resonance, where the forced oscillations have the same frequency as an external disturbance. This frequency transformation is widely used in engineering, in the so-called parametric transformations of frequency. Sometimes, the parametric resonance, as well as a usual one, leads to the extremely undesirable and even dangerous oscillations, for example, when the oscillations of dual wheels in electric locomotives are excited, because a stiffness of their connection depends upon the rotating angle of the wheels.

While discussing the results of the study of the Mathieu equation, it was supposed that the dissipation is absent and $\delta = 0$. If a usual resonance takes place and dissipation is present, the amplitude of oscillations becomes finite irrespective of the amount of disturbance. As for the parametric resonance, this happens not so. Here, the oscillating amplitude can be growing unboundedly even with the dissipation being available. For this, the amplitude of varying ε should only exceed some limit. This limit for some $\delta > 0$ is shown in figure 21.2 by a dotted line. The parametric excitation takes place over this dotted line. Naturally, with δ being increased, this dotted line is displacing higher and higher.

Let us look at figure 21.2 carefully. Here, the domains of stability and instability are depicted for both $\omega^2 > 0$ and $\omega^2 < 0$. A shaded one is the domain of stability. The absence of shading for $\omega^2 > 0$ means that the parametric resonance is available there. For $\omega^2 < 0$, the oscillator becomes unstable at $\varepsilon = 0$ as well and, hence, the instability domain becomes contiguous with the

abscissa axis. For $\omega^2 > 0$, the "noses" of the instability domains indicate the phenomenon of the parametric resonance, while the presence of stability domains for $\omega^2 < 0$, on the contrary, exhibits a phenomenon of the parametric stabilization of a statically unstable oscillator, the inverted pendulum. The presence of these stability domains demonstrates that the inverted pendulum can be made stable, if its point of support is oscillated at suitable frequency and amplitude.

In this place, our narrative about the oscillator with a parametric harmonic disturbance could be finished. However, why does even small dissipation make the amplitude of the forced oscillations to be finite under a usual resonance, while, under the parametric resonance, it is unboundedly growing?

Let us consider the process of growth of amplitude under a usual resonance in the absence of dissipation. This case is represented by the differential equation

$$\ddot{x} + \omega^2 x = a \cos \omega t \tag{21.12}$$

with zero initial conditions $x = \dot{x} = 0$ at $t = 0$. Its solution is easily found and is of the form

$$x = At \sin \omega t , \tag{21.13}$$

where after the direct substitution, it becomes clear that $A = a / 2\omega$.

Thus, a rise of the amplitude under a usual resonance is proportional to time, while a fall of the amplitude caused by $\delta > 0$ will be exponential. Thus, a rise of the amplitude will be obligatorily halted . With the parametric resonance present, the entire picture will be quite different. Under instability $(A > 1)$, the amplitude will grow exponentially, and its exponent can be more than the magnitude $-\delta$ of exponential dissipation. Therefore, a usual resonance can be suppressed by dissipation only partially, by decreasing the amplitude. For the case of parametric resonance, this suppression will be absolute, since it will transform an unstable equilibrium into a stable one. However, if this suppression is insufficient, the amplitude will keep on growing up exponentially.

In the conclusion, I am demonstrating, as promised, the parametric excitation, but only for the main resonance, when $\nu = 2\omega$, with the help of the same handbag with a small-length strap, being familiar to you from the earlier case. As you see, I am raising the bag now and by letting it down two times more frequently than it can swing, it will start swaying to and fro. Upon some time, this oscillating will grow, but, of course, not to infinity, since with oscillations being increased, our model will cease to be admissible.

22   Normal oscillations and beatings

A mathematical model of two interacting oscillators. Normal oscillations and an energy pump-up causing a beating phenomenon.

We have earlier got acquainted with the famous mathematical model, a linear oscillator. Its generalizations are systems of interactively connected oscillators. In the subsequent narrative you will find some new phenomena and cases relating to these generalizations of the linear oscillator. Their presentations will not be of a general character. They will be restricted by a simple example, upon understanding which you will have some idea of what new things may be in the multidimensional dynamical systems.

In the example below there are described the two masses whose displacements along a straight line are constrained by three springs (Fig. 22.1).
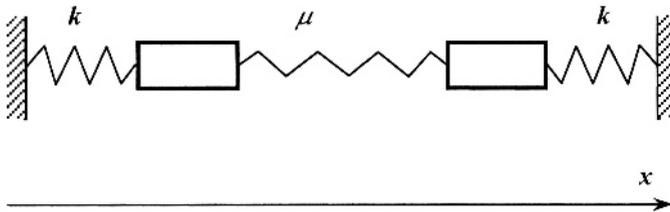
Fig. 22.1. Two harmonic oscillators interacting through an elastic linkage of the stiffness μ .

The masses are identical and equal to m , the extreme springs are of equal stiffnesses k . In between these masses there is a spring of the stiffness μ . If x_1 and x_2 are displacements of the masses from their equilibrium positions, then the Newton differential equations are immediately written in the form

$$m\ddot{x}_1 = -kx_1 - \mu(x_1 - x_2)$$

$$m\ddot{x}_2 = -kx_2 + \mu(x_1 - x_2)$$

or

$$m\ddot{x}_1 + (k + \mu)x_1 = \mu x_2$$

$$m\ddot{x}_2 + (k + \mu)x_2 = \mu x_1 .$$

(22.1)

Setting $\mu = 0$ in the right-hand sides of these equations, we come to the two not mutually interconnected linear oscillators. That both oscillators are of the identical frequency $\omega = \sqrt{(k + \mu)/m}$ is inessential but it simplifies the subsequent calculations. The equation for each of these oscillators will be derived, if the position of another is fixed by accepting x_1 or x_2, respectively, to be equal to zero. This, naturally, encourages one to consider the system (22.1) as two interactive oscillators, the first is acting upon the second with the force μx_1 and the second upon the first by the force μx_2. What is a result of this interaction? To answer this question one needs to solve the system (22.1). This can be done in the same way as for one linear oscillator, i.e. by searching for a solution of the form

$$x_1 = a_1 e^{\lambda t}, \quad x_2 = a_2 e^{\lambda t} ,$$

(22.2)

where the both variables x_1 and x_2 are varying proportionally to $e^{\lambda t}$. Such a solution, and more precisely, such real solutions associated with it, are called normal oscillations, i.e. the oscillations under which all parts of the system are oscillating with the same frequency, but, possibly, at different amplitudes. After the substitution of (22.2) into (22.1), we come to the two equations

$$(m\lambda^2 + k + \mu)a_1 - \mu a_2 = 0$$

$$- \mu a_1 + (m\lambda^2 + k + \mu)a_2 = 0 .$$

(22.3)

These are the two equations with respect to the three unknowns a_1, a_2 and λ. It is clear that we cannot find them uniquely. It was clear yet earlier, for together with (22.2) there also exists any other solution derived through multiplying (22.2) by one and the same factor. Via excluding the variables a_1 and a_2 or via employing the general theorem from linear algebra concerning a nonzero solution for a linear uniform system of equations, we come to the characteristic equation for all possible values of λ

$$\chi(\lambda) = \begin{vmatrix} m\lambda^2 + k + \mu & -\mu \\ -\mu & m\lambda^2 + k + \mu \end{vmatrix} = 0 .$$

(22.4)

This equation is easily solved (the example has been chosen for this purpose), and it turns out that

$$\lambda_{1,2} = \pm i\sqrt{\frac{k}{m}} = \pm i\omega_1, \qquad \lambda_{3,4} = \pm i\sqrt{\frac{k + 2\mu}{m}} = \pm i\omega_2. \qquad (22.5)$$

Now, by sequentially inserting the values (22.5) into (22.3), we find the constants a_1 and a_2 with accuracy of up to a factor. In doing so, one should insert the obtained values of λ only into one of the equations (22.3), since they are linearly dependent.

For $\lambda = \lambda_{1,2}$, we have $\mu a_1 - \mu a_2 = 0$ or

$$a_2 = a_1.$$

For $\lambda = \lambda_{3,4}$, we have

$$-\mu a_1 - \mu a_2 = 0$$

or

$$a_2 = -a_1.$$

Thus, we have found the particular solutions of the form

$$x_1 = ae^{i\omega_1 t}, \qquad x_2 = ae^{i\omega_1 t}$$

and of the form

$$x_1 = ae^{i\omega_2 t}, \qquad x_2 = -ae^{i\omega_2 t}.$$

The corresponding solutions to them are the following real solutions

$$x_1 = a \frac{\cos}{\sin} \omega_1 t, \qquad x_2 = a \frac{\cos}{\sin} \omega_1 t \qquad (22.6)$$

and

$$x_1 = a \frac{\cos}{\sin} \omega_2 t, \qquad x_2 = -a \frac{\cos}{\sin} \omega_2 t. \qquad (22.7)$$

These solutions, as noted above, are called normal oscillations. Their frequencies are different (ω_1 and ω_2, respectively); and, besides, $\omega_2 > \omega_1$.

For the first normal oscillation, the amplitudes of both masses are identical, and they are oscillating in phase. For the second one, the amplitudes are equal as well, but the oscillations are performed in anti-phase. This means that, for the first normal oscillation, both masses are oscillating together to the right and to the left. For the second one, on the contrary, they are oscillating towards each other at larger frequency. Any other motion of the system of two masses or of mutually connected oscillators under consideration will be a superposition of the normal oscillations (22.6) and (22.7). Through combining the oscillations (22.6) and (22.7), it becomes possible to construct various plots, but seemingly, the most interesting appears for ω_1 close to ω_2 , since, in this case, we can observe an unexpected phenomenon of a successive energy pump-over from one oscillator to another; in this situation the plots of x_1 and x_2 are of the form given in figure 22.2.
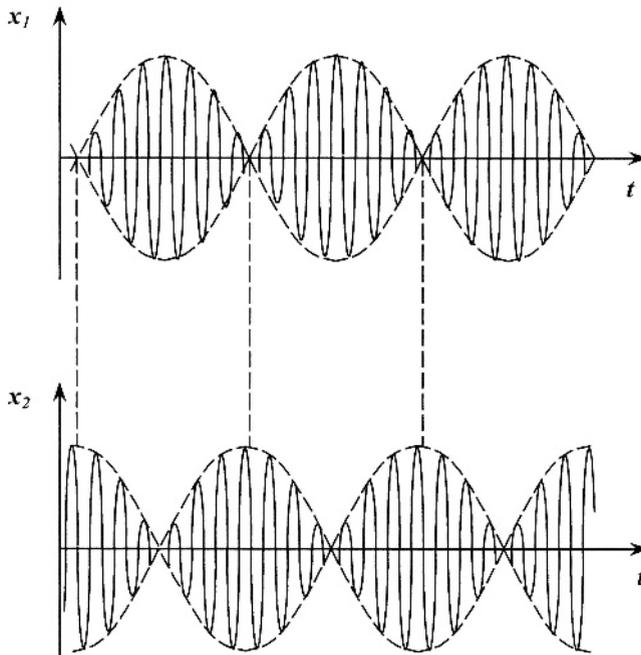
Fig. 22.2. A beating phenomenon in oscillations of the two connected oscillators.

According to (22.5), the frequencies ω_1 and ω_2 are close to each other, if μ is small; and this case will be assumed in what follows. Here, by making use of the complex notation of oscillations, we obtain

$$x_1 = C_1 e^{i\omega_1 t} + C_2 e^{i\omega_2 t},$$

$$x_2 = C_1 e^{i\omega_1 t} - C_2 e^{i\omega_2 t}$$

or, after some manipulations made in supposing $|C_1| \geq |C_2|$,

$$x_1 = C_1 e^{i\omega_1 t} (1 + \frac{C_2}{C_1} e^{i(\omega_2 - \omega_1)t}),$$

(22.8)

$$x_2 = C_1 e^{i\omega_1 t} (1 - \frac{C_2}{C_1} e^{i(\omega_2 - \omega_1)t}).$$

Each of these oscillations, in view of the notation (22.8) and $\omega_1 \approx \omega_2$, can be considered as an oscillation of the frequency ω_1 and of the amplitudes equal to

$$\left| 1 \pm \frac{C_2}{C_1} e^{i(\omega_2 - \omega_1)t} \right|,$$

which are varying slowly in the range from $1 - |C_2/C_1|$ to $1 + |C_2/C_1|$.

The frequency of amplitude variations is small and equal to $\omega_2 - \omega_1$. We have supposed that $|C_1| \geq |C_2|$; had we supposed that $|C_1| \leq |C_2|$, we would have analogously come to the oscillations of the frequency ω_2 and of the amplitude varying slowly within the range   from $1 - |C_1/C_2|$ to $1 + |C_1/C_2|$. It is confirmed by the plot in figure 22.2. The completeness of the energy pump-over from one oscillator to another depends upon the relation $|C_2/C_1|$. For $|C_1| = |C_2|$, the pump-over is executed fully, while for $C_2 = 0$ or $C_1 = 0$ it is absent, but this occurs only for the normal oscillation of the frequency ω_1 or ω_2, i.e. in the cases when the masses are oscillating together or towards each other.

This simple example should be thought to be some kind of introduction to a world of various oscillations of connected oscillators and linear systems; this world is governed by   the superposition principle.   In particular, this world embraces the processes of forming the chains, the crystal lattices or separate molecules having an intricate spectrum of normal oscillations and a variety of forms and superpositions. The qualitatively new, probably, will be the normal oscillations in the shape of standing and travelling waves that brightly manifest themselves at a great number of   oscillators. These wave phenomena build a

bridge from an unpretentious linear oscillator to the oscillations of distributed mediums corresponding to their unboundedly large quantity.

Let us now pass to a visual description of the revealed phenomenon of beatings and energy pump-over from one oscillator to another and conversely. Note that this takes place in spite of smallness of oscillator connections, when the stiffness μ is small, but the frequencies of oscillators are equal for $\mu = 0$. In other words, a pump-over from one oscillator to another and conversely occurs when their frequencies are equal in resonance. This creates a possibility to understand why a small connection brings about great consequences, i.e. a complete energy pump-over from one oscillator to another.

This energy pump-over can be observed visually by means of two identical pendulums mounted on a single axis and connected by a weak spring or by an elastic thread (Fig.22.3).
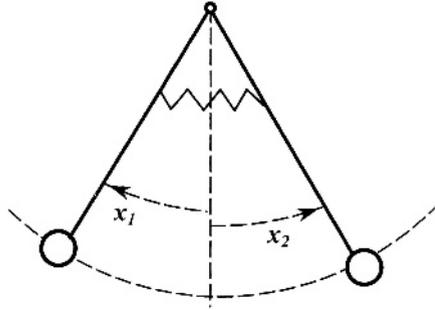
Fig. 22.3. Elastically connected oscillators as pendulums.

If one of the pendulums has been pushed, then it starts to sway the other until it itself stops; and upon this it will be in its turn swayed by another, etc. This mutual swaying is caused by their small force, but resonance, interaction. A similar phenomenon of a resonance pump-over of the energy can be observed when one of the connections between oscillators is a force connection and another is parametric. Imagine some weight suspended by a long spring. It can swing up and down and can also oscillate as a pendulum (Fig. 22.4).

The frequencies of these oscillations are different, and let everything be chosen in such a way that the oscillating frequency is two times more than the swinging frequency. Then, when the weight is swaying, the length of this pendulum will be periodically varying, and its sways become parametrically excited, naturally, at the expense of its oscillations. Further, these sways of the pendulum excite the oscillations of the weight, because the inertia force has the frequency being doubled with respect to the frequency of pendulum sways. All this brings us to a very interesting and beautiful sequence of passages from oscillations to sways, and, vice versa, from sways to oscillations.
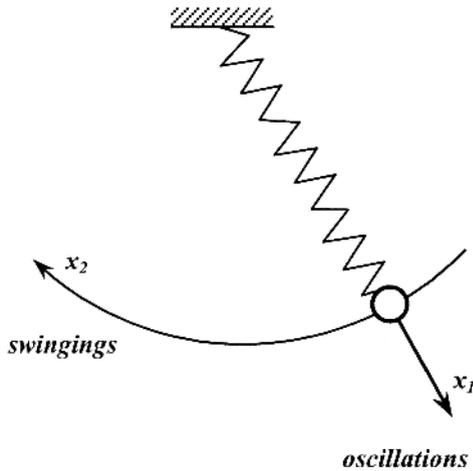
Fig. 22.4. The elastically-suspended pendulum as two connected oscillators; the connection from x_1 to x_2 is parametric and from x_2 to x_1 a force one.

In conclusion, I would like to emphasize that a formal mathematical aspect of this narration is mainly familiar to you from the course of differential equations. You were taught there how systems of linear differential equations with constant coefficients are solved. In this aspect, you have not heard here anything new. However, we have tried to bring here some picturesqueness to what we were doing by introducing a concept of a normal oscillation, its frequencies and forms and by representing any motion as a superposition of normal oscillations. In order to feel a usefulness of these general visual representations you are invited to try to reveal what possible oscillations of three equal masses will be if these masses mounted in a friction-free fashion on a ring, provided that the attractive force of each pair of masses, due to the connecting spring, is proportional to the angle between them. By general rules this problem will be solved very long and in a boring way. And instead, with a use of descriptive representation for the normal oscillation, this solution will need not more than a single line.

23 Stabilizing an inverted pendulum

Control as a powerful facility for changing behaviour and properties of dynamical objects and systems. A juggling by a stick vertically standing on a support and by two sticks standing on each other. The mathematical models. A control strategy, conditions for stabilization, a role of delay.

Now I would like to tell you about a universal method allowing to fundamentally change properties of various objects. This method is demonstrated on a simple example of transforming some unstable vertically standing and always falling stick into a stable independently standing one. This universal method widely used by wild nature and humans is called a control. A vertically installed stick cannot keep standing and it will fall necessarily. To make it not falling can be done only in circus. This is performed by jugglers with the help of hardly visible displacements of its supporting point. They change a position of its supporting point so that a stick or some other thing keeps standing vertically and does not fall. How are they doing it? Let us try to understand it by having constructed a mathematical model for stabilizing a vertical pendulum.

Well, imagine that we have a plane pendulum with a displaceable point of support (Fig. 23.1). Let an angle of pendulum deviation from a vertical line be φ and a horizontal displacement of its point of support within a swinging plane be u . Let the length of the pendulum be l and its mass be m . Let us find the Lagrange function and with its help construct the motion equations.

At first, we immediately determine the coordinates x, y of the pendulum mass m

$$x = u - l \sin \varphi , \quad y = l \cos \varphi .$$

Next, we find the kinetic and potential energies

$$T = \frac{m}{2} \left[(\dot{u} - l\dot{\varphi} \cos \varphi)^2 + l^2 \dot{\varphi}^2 \sin^2 \varphi \right] =$$

$$= \frac{m}{2} (\dot{u}^2 - 2u\dot{\varphi} \cos \varphi + l^2 \dot{\varphi}^2),$$

$$V = mgy = mgl \cos \varphi$$

and the Lagrange function

$$L = T - V = \frac{m}{2}(\dot{u}^2 - 2lu\dot{\varphi} + l^2\dot{\varphi}^2) - mgl\cos\varphi. \qquad (23.1)$$
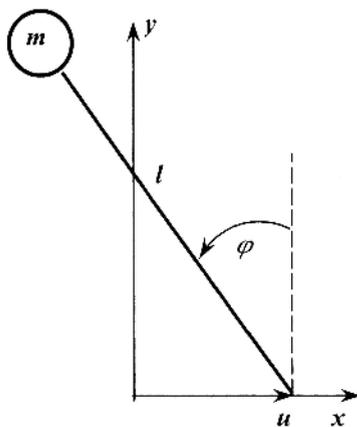
Fig. 23.1. An inverted pendulum with the angle φ of deviation from the upper unstable position.

We construct the Lagrange equation

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\varphi}} - \frac{\partial L}{\partial \varphi} = 0. \qquad (23.2)$$

Since in the case under consideration we have

$$\frac{\partial L}{\partial \dot{\varphi}} = -ml\dot{u}\cos\varphi + ml^2\dot{\varphi}, \quad \frac{\partial L}{\partial \varphi} = mlu\dot{\varphi}\sin\varphi + mgl\sin\varphi,$$

then (23.2) will be as follows

$$-ml\ddot{u}\cos\varphi + mlu\dot{\varphi}\sin\varphi + ml^2\ddot{\varphi} - mlu\dot{\varphi}\sin\varphi - mgl\sin\varphi = 0$$

or

$$\ddot{\varphi} - \frac{g}{l}\sin\varphi = \frac{\ddot{u}}{l}\cos\varphi.$$

Let us restrict ourselves by considering only small angles φ and then simplify this equation by writing it in the form

$$\ddot{\varphi} - \frac{g}{l}\varphi = \frac{1}{l}\ddot{u}. \tag{23.3}$$

For $u = 0$, i.e. when the supporting point is fixed, the equation (23.3) turns into the well-known pendulum equation linearized about the upper unstable equilibrium position. This unstable equilibrium is of the saddle type. However, we want it to become a stable equilibrium of the node or focus type. How can it be reached? Let us use a possibility of choosing the displacement u for the point of support by controlling its position. How it should be done we do not know, but we want the equation (23.3) to turn into the equation

$$\ddot{\varphi} + 2\delta\dot{\varphi} + \omega^2\varphi = 0 \tag{23.4}$$

for the oscillator with the stable equilibrium position $\varphi = 0$, i.e. we want the pendulum to be standing vertically up and to behave in the way as if it was hanging downward and its oscillations would be damping $(\delta > 0)$.

Comparing (23.3) and (23.4), we find that to reach the above aim, it is necessary to have the following

$$\ddot{u} = -(g + l\omega^2)\varphi - 2l\delta\dot{\varphi}. \tag{23.5}$$

Such u may be implemented through observing the pendulum and measuring its deviation φ and the deviating velocity $\dot{\varphi}$, and by providing the point of support with the acceleration $\ddot{u}$, according to the formula (23.5).

The   controlling technique given by the formula (23.5) is called a control strategy. As was already said, to implement this strategy it is necessary to measure φ and $\dot{\varphi}$, and, naturally, to apply to the supporting point the force   that could impart to it the required acceleration $\ddot{u}$. Thus, for implementing the control it is necessary to have a measuring and an executive systems . In the circus, these both functions are successfully executed by a juggler.

The described process of control may be depicted by the scheme shown in figure 23.2. The scheme involves an object under control; an inverted pendulum to be acted upon   by displacing its support; a measuring system to measure the angle φ and the angle velocity $\dot{\varphi}$; a control system implementing some control strategy; and a force impact displacing a base of the inverted pendulum and providing it with the acceleration $\ddot{u}$ required in accordance with the control strategy. What does the control strategy (23.5) represent itself? It is linear in φ and $\dot{\varphi}$, so that

$$\ddot{u} = -a\varphi - b\dot{\varphi},$$

where $a > g$ and $b > 0$, as required for stability. For $b = 0$ the inverted pendulum is stable as well, but its oscillations about the equilibrium are not damping, since they are described by the oscillator (23.4), where $\delta = 0$.
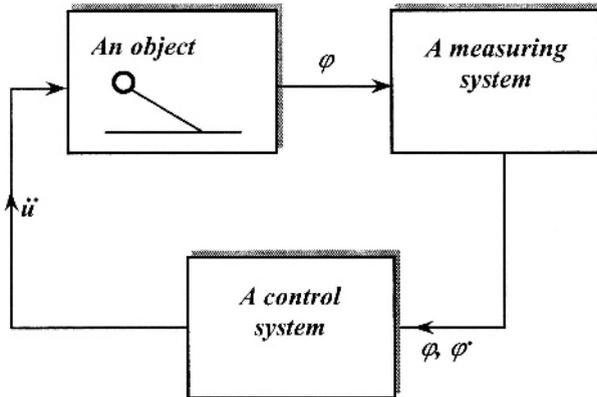
Fig. 23.2. A schematic diagram for stabilizing a vertically standing pendulum, which represents a basic principle of control, feedback .

The first term $a\varphi$ of the control strategy is a control by the pendulum deviation φ. According to this term, the pendulum support should be moved in an accelerated way to the same direction as the pendulum derivation is, i.e. for $\varphi > 0$, one has $\ddot{u} < 0$ and for $\varphi < 0$, one has $\ddot{u} > 0$.

The next term is responsible for correcting a value of the acceleration $\ddot{u}$ by accounting an angle velocity of the pendulum rotation. If the pendulum is moving to the same side as its deviation, then this results in increasing the required acceleration $\ddot{u}$ as compared with the term $a\varphi$; and on the contrary, if it is moving to the inverse side with respect to its deviation, then this acceleration is decreasing, so becoming less than $a\varphi$.

Let us now pass from an jugglery with a stick to a control of a ship. The target of this control is keeping a certain course of its movement by means of a rudder. If a rudder is not controlled, then a ship may move into any direction or will be changing its course all the time.

Let us construct a mathematical model of a ship moving at the velocity v and that of its control, i.e. a model for the ship under control.

Suppose that a ship is moving with the constant velocity v, so that a resistance force to its movement is compensated by a driving force of oars or a marine propeller. If an onrunning flow is exactly directed along the ship hull, then no

moment of force appears there. The moment of force (let it be with respect to its centre of gravity) will arise only if an onrunning   stream deviates from a longitudinal axis of the ship, i.e. from the direction of its keel. In the rough approximation, a divergence angle between the velocity of the gravity centre and the longitudinal axis is determined approximately by the velocity $\dot{\varphi}$ of the ship rotation and is equal to $-h\dot{\varphi}$, where the coefficient h can be positive or negative, depending upon a ship construction. Here, a variation of the angle φ of the ship turning with respect to its gravity centre, if an influence of a rudder not considered, will be described by the equation

$$J\ddot{\varphi} = -h\dot{\varphi}, \qquad\qquad (23.6)$$

where J is a moment of inertia of the ship. As we see, a ship as an object to be controlled represents itself an oscillator of a zero stiffness.

Integrating the equation (23.6), we determine that

$$J\dot{\varphi} + h\varphi = const. \qquad\qquad (23.7)$$

That the first integral (23.7) is available allows to create immediately phase portraits for the ship moving without a rudder; figure 23.3a describes the case $h > 0$ and figure 23.3b the case $h < 0$.
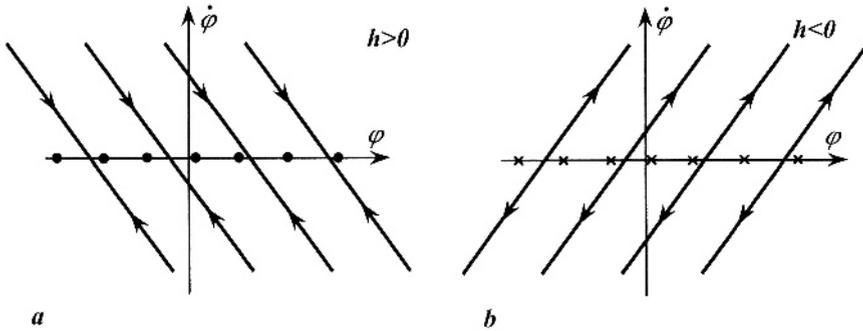
Fig. 23.3. Phase portraits for an uncontrolled **a** stable and **b** unstable ship.

In both cases, the ship possesses a set of equilibriums corresponding to the phase points of the axis $\dot{\varphi} = 0$, under which the ship is moving with constant velocity at some constant   angle φ. For $h > 0$, these equilibriums are stable and any movement of the ship turns to a movement in some constant direction. On the contrary, for $h < 0$, the equilibrium states are unstable and the ship tends to

rotate faster and faster (Fig. 23.3b). As our linearized simplified equation (23.6) is valid only for not very large $\dot{\varphi}$, the said should be understood in the way that the ship cannot maintain a certain course and it will necessarily rotate. For $h > 0$, the equilibrium states are stable, but this stability is not asymptotic; a smallest disturbance will produce an unremovable change in the direction of the ship motion.

Assume that $\varphi = 0$ corresponds to a desired course of the ship. From the above it follows that neither for $h > 0$ nor for $h < 0$ especially the ship cannot stably maintain the given course $\varphi = 0$.

To have this course maintained, the ship should be controllable, say, through a rudder. Let ψ be an angle of rudder deviation (Fig. 23.4).
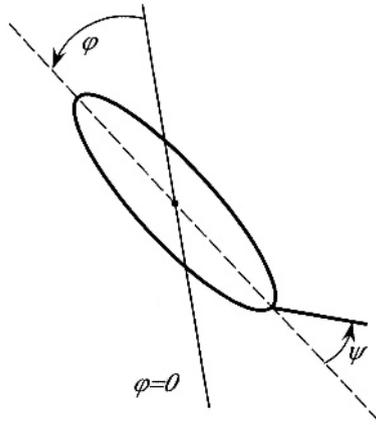
Fig. 23.4. A ship deviating from the desired course at the angle φ and with the rudder angle ψ.

For $\psi \neq 0$, the pressure of the water running upon the rudder will create a turning moment and the equation (23.6) will be replaced by the equation

$$J\ddot{\varphi} + h\dot{\varphi} = -k\psi. \qquad (23.8)$$

Such an equation is reasonable, of course, only for not very large ψ. The equation (23.8) presupposes $\dot{\varphi}$ and ψ to be small and φ can be arbitrary. To provide the ship with a new property of asymptotically stable movement along the desired course $\varphi = 0$, let us apply here the same control strategy as that for juggling a stick, and namely, assume that

$$\psi = a\varphi + b\dot{\varphi}\,. \tag{23.9}$$

Inserting (23.9) into (23.8), we find that

$$J\ddot{\varphi} + (h + b)\dot{\varphi} + ka\varphi = 0\,. \tag{23.10}$$

The equilibrium state $\varphi = 0$ of this differential equation, a linear oscillator, will be stable, if all the coefficients of its characteristic equation

$$J\lambda^2 + (h + b)\lambda + ka = 0 \tag{23.11}$$

are positive, i.e. when $h + b > 0,\ ka > 0$, which are easy to be satisfied by choosing the appropriate a and b so that

$$a > 0\,,\ b > -h\,. \tag{23.12}$$

Thus, the ship under the control strategy (23.9), if the conditions (23.12) are met, will be asymptotically stable along its course. The idea of this strategy is very simple. It is the same as that used for juggling a stick, but, if earlier it was required to have a sufficiently large value of the parameter a, then now, for $h < 0$, this requirement is imposed upon the parameter b. From this fact it follows that to control the ship for $h > 0$ is comparatively easy. On the contrary, for $h < 0$ it becomes a sufficiently complicated problem, because there will be needed a sufficiently large and correct response to the rotating velocity of the ship .

For this correct reaction there exists even its own name – in the proper time you should energetically "restrain" your ship leaving its course. If in itself the ship is stable $(h > 0)$, then the control of it looks like the following: the ship is to the left, the rudder is to the right, the ship is to the right, the rudder is to the left. For $h < 0$, the controlling strategy will be more complicated and the main idea in it will be in "restraining" the ship, i.e. this restraining is done even for $\varphi = 0$, if $\dot{\varphi} \neq 0$.

Now, let us return again to juggling a stick. Before proceeding our discussion, we notice here that a circus juggling has much in common with a stabilization of the vertical take-off and speed-up of a rocket. Just like a stick, a rocket is unstably sticking up and is supported from below by its jet engines. Of course, a rocket needs more precise and complete models, but, in principle, everything is the same as for a stick. More precise models are needed , at least, to take into account such things as an   elasticity of the rocket body, an   influence of its oscillations upon readings of the   measuring systems and upon the stabilizing process, delays in implementing the   observational results and the control actions, and, possibly, many other things.

Yet let us return to stabilizing an inverted pendulum. The strategy (23.5) sustains a verticality of the inverted pendulum, but, in this process, its supporting point   may be displaced considerably having no tendency to come back. Here there arises a question as whether it can be possible to stabilize an upper

position of the pendulum through simultaneously making an initial position of its supporting point asymptotically stable.

Let us try to apply the above strategy to this more complicated problem, i.e. a linear strategy in phase variables of the object. Thus, we want to stabilize a pendulum not only in φ but in the coordinate u of the supporting point as well. Now the phase variables of the pendulum with a displacing point of support will be the variables $\varphi, \dot\varphi$ and $u, \dot u$, for which

$$\ddot\varphi - \frac{g}{l}\varphi = \frac{\ddot u}{l},$$

(23.13)

$$\ddot u = a\varphi + b\dot\varphi + cu + d\dot u,$$

where the control strategy, as before, will be a linear function of all phase variables.

The problem has now become more complicated. The characteristic equation whose roots define a stability is now not of the second, but of the fourth order. Let us find it. From (23.13) immediately follows the characteristic equation in the form of the determinant

$$\begin{vmatrix} \lambda^2 - g/l & -\lambda^2/l \\ a+b\lambda & c+d\lambda - \lambda^2 \end{vmatrix} = 0$$

or, by uncovering the determinant,

$$\lambda^4 - (d + \frac{b}{l})\lambda^3 - (c + \frac{g}{l} + \frac{a}{l})\lambda^2 + d\frac{g}{l}\lambda + c\frac{g}{l} = 0.$$

(23.14)

Through choosing the parameters a, b, c, d, the roots of this equation can be done arbitrary. Indeed, a free term can be made arbitrary due to c, the coefficient before λ will be made arbitrary due to d, the coefficient before λ^2 due to a, and, finally, the coefficient before λ^3 due to b.

Thus, an appropriate strategy of the form

$$u = a\varphi + b\dot\varphi + cu + d\dot u,$$

under which all the roots of the equation (23.14) are from the left of the imaginary axis, will provide the desired stabilization of the inverted pendulum. We emphasize here that for $c = d = 0$, i.e. under the control strategy $\ddot u = -a\varphi - b\dot\varphi$, the characteristic equation will have two zero roots and the equilibrium $\varphi = 0$, $u = 0$ will be unstable, which will yield a deviation of u from zero.

Let us also touch the point concerning a so-called levitation, i.e. a non-contact stable "suspension" of a body in some place. According to a legend, the coffin of Mohammed is hanging in the air in some grotto. At once, there comes an idea to do it with the help of magnets , i.e. to magnetically suspend an iron thing or a body involving a ferrous. However, there exists a general theorem that forbids it. At the same time, a ferromagnetic small-size ball can be stably suspended by the same control strategy as before. To do it is very tempting , since a contact-free magnetic suspension allows   to create many new high-precision measuring instruments and other devices now free from inaccuracies and dissipation caused by friction, not speaking about the absence of undesirable contacts with other bodies.

During your childhood, you probably tried fruitlessly to "suspend" a school pen (or some small iron thing) with the help of a magnet: as a result, this pen habitually used to fall down or got stuck to the magnet. A position of a magnet may be controlled, but it is much simpler to replace a magnet by an electromagnet and to control its current I . As earlier, we will need here some system for measuring a position and a velocity of a suspended body; as for a control strategy, it may consist in varying the electrical current as a linear function of these measurements. With the parameters chosen properly, the body is hanging freely and asymptotically stably. This hanging body may be touched with a finger and one may feel that the body is hanging firmly not willing to change its position. One may also start turning it and it will be rotating for a long time, since friction against air is insignificant. Friction can be decreased by pumping the air out and then the body will be rotating longer.

We have thus considered a juggling (stabilizing) of a single inverted pendulum, but here comes a question as whether it is possible in a similar way to juggle some pendulums placed upon each other and to control the supporting point of the most lower pendulum. To do it seems impossible, because the upper pendulums will fall down. However, let us see what will follow from our mathematical model. Let us take only two pendulums of equal lengths and masses, which we assume, without loss of generality, unitary. By adopting the notations of figure 23.5, we immediately find that the   Lagrange function L, in view of the subsequent linearization, i.e. retaining only linear terms   regarding $\varphi_1, \varphi_2, \dot{\varphi}_1, \dot{\varphi}_2$ in the Lagrange differential equations, will be equal to

$$L = \frac{1}{2}(2\dot{u}^2 - 4\dot{u}\dot{\varphi}_1 - 2\dot{u}\dot{\varphi}_2 + 2\dot{\varphi}_1^{\,2} \ \ + \dot{\varphi}_2^{\,2} \ \ + 2\dot{\varphi}_1\dot{\varphi}_2) + g\varphi_1^{\,2} + \frac{g}{2}\varphi_2^{\,2}$$

Therefore, we obtain the following equations:

$$2\ddot{\varphi}_1 + \ddot{\varphi}_2 - 2\ddot{u} - 2g\varphi_1 = 0$$

$$\ddot{\varphi}_1 + \ddot{\varphi}_2 - \ddot{u} - g\varphi_2 = 0$$

or

$$-\ddot{\varphi}_2 - 2g\varphi_1 + 2g\varphi_2 = 0$$

$$\ddot{\varphi}_1 + \ddot{\varphi}_2 - \ddot{u} - g\varphi_2 = 0.$$

(23.15)

As before, we assume a linear function of all the phase variables to be a controlling strategy, i.e.

$$\ddot{u} = a_1\varphi_1 + b_1\dot{\varphi}_1 + a_2\varphi_2 + b_2\dot{\varphi}_2.$$

(23.16)

Let us write the characteristic equation of the system (23.15), (23.16) in the form

$$\begin{vmatrix} -2g & -\lambda^2 + 2g \\ \lambda^2 - a_1 - b_1\lambda & \lambda^2 - g - a_2 - b_2\lambda \end{vmatrix} = 0$$

or

$$\lambda^4 - b_1\lambda^3 - (a_1 + 4g)\lambda + 2g(b_1 + b_2)\lambda + 2g(g - a_1 + a_2) = 0.$$ (23.17)

From the form of the characteristic equation (23.17) it follows that the coefficients in front of $\lambda^3, \lambda^2, \lambda$ and λ^0 may be arbitrary for the appropriate choice of the parameters b_1, a_1, b_2 and a_2 in the control strategy.
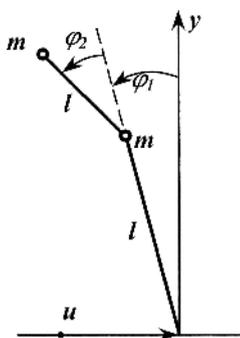
Fig. 23.5. A double pendulum with the angle deviations φ_1 and φ_2 from the vertical.

Thus, stabilization of two sticks placed upon each other is possible, and for it it is sufficient to properly displace the lower supporting point in accordance with the control strategy (23.16).

At the same time, one should remember that this and the preceding models were constructed under the assumption that both a measuring and an executive components of the controlling system are ideal, i.e. a measuring system determines values of the phase variables sufficiently accurately and fast, and the executive system implements a desired control accurately and quickly (accelerating a supporting point and regulating a turning angle of the rudder). Besides, from the mathematical models considered by us it is not clear to what extent one may deviate from these ideal requirements and in what way one may implement them.

It is hardly advisable to consider these questions, important in themselves, in detail here; this should be better done in special courses. Therefore we restrict ourselves here by the above said and by an example on how to consider unavoidable delays (delays in measuring φ and $\dot{\varphi}$, and in implementing the controlling acceleration $\ddot{u}$) while stabilizing a vertical position of the pendulum. If these delays were absent, the equation of the closed-loop system was of the form (the equations (23.3), (23.5))

$$\ddot{\varphi} - \frac{g}{l}\varphi = -a - b\dot{\varphi}.$$

Taking into account the delay τ in both the measurements of φ, $\dot{\varphi}$ and a control implementation, we shall replace this equation by the following

$$\ddot{\varphi}(t) - \frac{g}{l}\varphi(t) = -a\varphi(t-\tau) - b\dot{\varphi}(t-\tau). \tag{23.18}$$

Its characteristic equation, derived in a usual way by assigning $\varphi = e^{\lambda t}$, will be as follows

$$\lambda^2 - \frac{g}{l} + ae^{-\tau\lambda} + b\lambda e^{-\tau\lambda} = 0. \tag{23.19}$$

In order to have the control system, described by the linear differential equation with delays (23.18), stable, i.e. to have its equilibrium state $\varphi = 0$ asymptotically stable, it is necessary, as in the case of an ordinary differential equation, to have the roots of its characteristic equation (23.19) lying to the left from the imaginary axis, i.e. to have them satisfying the condition $\mathrm{Re}\,\lambda < 0$. A disposition of the roots of the characteristic equation (23.19) depends upon the parameters a and b of the controlling strategy. If no delay occurs, i.e. when $\tau = 0$, then the stability takes place for

$$a > \frac{g}{l}, \quad b > 0. \tag{23.20}$$

As it has been revealed in the appropriate study whose technique will be explained to you during your academic course on the control theory, for small τ's, less than $\sqrt{2l/g}$, the stability domain in the parameter plane a and b will exist and will have the form shown in figure 23.6. As $\tau \to 0$, this stability domain will tend to the domain D_0, determined by the inequalities (23.20). With τ being increased, the stability domain in the parameters a and b of figure 23.6 is decreasing, and for $\tau = \tau^* = \sqrt{2l/g}$ it will vanish. For $\tau > \tau^*$, it is entirely absent.
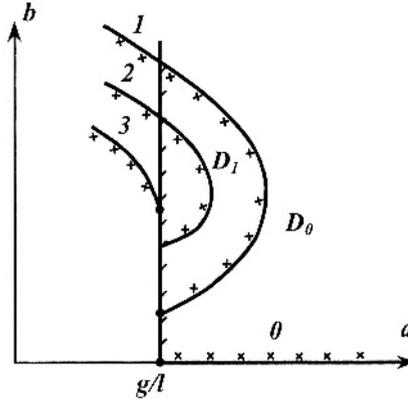
Fig. 23.6. The stability domain of the inverted pendulum and its dependence upon the delay τ of the control strategy implementation. The boundary curves $0,1,2,3$ correspond to a delay increase. For $\tau = 0$, the stability domain will be $a > g/l,\ b > 0$; for τ corresponding to the boundary curve 3, the stability domain will disappear.

From the said it follows that, in order to juggle successfully, one needs not only to choose the parameters a and b of the controlling strategy in a correct way, but also to implement measurements and controls sufficiently fast. This fastness must be the more, the less a length of the stick is. When juggling with a very short stick, you will inevitably fail. And the least length of the stick in your successful juggling speaks about how quickly you manage to implement a desired control and, of course, speaks simultaneously about your skills in choosing a required controlling strategy.

With a stick (a ball-point pen) of 16 cm in length, I failed to juggle always. To do juggling seemed hopeless to me. The stick used to fall down before I managed to do something against. For the stick of this length, a crucial value of the delay makes up $\tau \approx 0.17$ sec. For $\tau^* = 0.3$, the appropriate length is equal to $l = 45$ cm. With such a stick we may manage to juggle. I mean myself, and

what about you? If you wish, you may compare the τ^* 's, obtained for you from the juggling experiments, with direct measurements of the time of your reactions; this may be exercised by your friends with a help of some ruler. It is interesting that the last value is calculated by the same formula

$$\tau = \sqrt{2l / g} \, ,$$

where l is a least length of the ruler that is falling down between your thumb and forefinger and which you, upon letting it go, manage to clutch and prevent from dropping. In my experiment, I obtained   the length of about 20 cm that corresponds to $\tau \approx 0.2$ sec.

Through some simple examples we have thus become acquainted with a greatest invention of nature and of human beings, and this invention is a control. Without it, life would be impossible. Man performs a control of his behaviour; via controlling through enzymes   he performs biochemical processes in his organism; by switching on light in a dark room he is regulating electric current by pressing a button; by turning his steering wheel he is controlling his bicycle and, pedalling, he makes it go; through controlling by his hands and foots he is driving a car; through pronouncing commands and through other people he is performing a management of industrial processes; via his nervous system he is controlling the contractions and the enfeeblements of his muscles and, through it, any motions of his body, the heart beating, breathing, the contractions and the peristalsis of his bowels, the stream of substances moving into a stomach and duodenum. Just look how beautifully a horse is running or galloping, how gracefully a cat is displacing and moving – all these things are happening thanks to a controlling. And a mimicry of our face reflecting our intrinsic psychological processes and enriching our intercourse; and our speech itself is possibly existing only thanks to a control of vocal cords, a throat and articulation.

Man has created and developed a new science, the control theory, and it prompted him how to construct an auto-pilot and an automatic steering unit, and how to control a nuclear reactor, an aircraft flight, a running of machine-tools and mechanisms, and how to establish a communication system and information transmission systems, and how to control a missile and his own life on the Earth. To control the latter is most difficult and still far from perfection. And our life itself is possible only due to ferments, governing biochemical reactions in our organism, and nervous system that is controlling our heart, organs, motions and actions.

We shall speak about the control theory later on and what has been told is merely an introduction to it.

24 Controlled pendulum and a two-legged pacing

An inverted controlled pendulum as a mathematical model for a two-legged pacing.

In the previous chapters I have told you about a control and used the examples of juggling, stabilizing a single stick and two vertically standing sticks, and also controlling a course of a ship. You usually meet with these phenomena, though, perhaps, comparatively seldom. And now I will tell you about a controlling phenomenon that is also specific and you encountered it all the time and employ it every day. With juggling you also deal every day when maintaining a vertical position of your body, but the models of this kind of juggling are different from those we have considered earlier. A simplest kind of juggling with your body is your standing by resting upon your both legs. You are standing steady though slightly swinging, but this is hardly visible. If you try to listen to yourself, you may notice how the efforts of the supporting legs are being varied. This prompts an idea of simulating your juggling through an inverted pendulum at the fixed supporting point of which there is some controlling moment M. In each of two perpendicular planes there take place linearized equations of the form

$$\ddot{\varphi} - \frac{g}{l}\varphi = \frac{1}{ml^2}M.$$

According to the preceding, a strategy for changing the moment M, providing stabilization of the vertical position, can be in the form

$$M = a\varphi + b\dot{\varphi}.$$

However, standing vertically is possible not only by resting upon the soles of your both feet, but also through standing on one foot or even on a toe of one foot, i.e. by resting upon such a small area that it becomes impossible to create the controlling moment M. Meantime, you keep standing and this is attained by you at the expense of motions by your arms, body and a free leg. It turns out to be sufficient to change only the positions of the body, with the help of the moment between the body and the supporting leg. As a simplified model for such a juggling let us take the double pendulum considered earlier, but, this time, subject to regulation will be not a position of its supporting point, but the moment M in their coupling. Its mathematical model will be written with use of the linearized

equations of the double pendulum, which immediately precede the equations (23.15), assuming there $u = 0$ and adding on the right-hand sides the moments M and $-M$, respectively. Then, we obtain

$$2\ddot{\varphi}_1 + \ddot{\varphi}_2 - 2g\varphi_1 = M$$

$$\ddot{\varphi}_1 + \ddot{\varphi}_2 - g\varphi_2 = -M .$$

Let us choose a control strategy for the moment M, as before, in the form of a linear function of all phase variables

$$M = a_1\varphi_1 + b_1\dot{\varphi}_1 + a_2\varphi_2 + b_2\dot{\varphi}_2 .$$

This will lead us to the characteristic equation

$$\lambda^4 + (3b_2 - 2b_1)\lambda^3 + (3a_2 - 2a_1 - 3g)\lambda^2 +$$

$$+ (b_1 - 2b_2)g\lambda + g(2g - a_1 - a_2) = 0 ,$$

whose form shows that its four coefficients may be chosen arbitrarily thanks to an appropriate choice of the controlling strategy parameters a_1, b_1, a_2, b_2 .

Thus, a stabilization of two inverted pendulums, placed one upon another, with a fixed supporting point is possible through a suitable control of the moment M in their coupling. Such strategies have been also found by nature that develops them partly during a training period after your birth, on the basis of some genetic heredity.

Well, we are able to be standing. However, we are also capable of pacing. A two-legged pacing of a human is a very complicated process: many dozens of our muscles are consistently participating in the implementation of this pacing. At the same time, a pacing phenomenon in its simplest form is nothing, but some peri-odically repeated process, to be more exact, an auto-oscillation, if no human bones, but a man as a whole is chosen to be an object of control. To generate oscillations is possible with use of a control and a feedback. In the above examples, it was shown how to use a control and a feedback to implement it. In these examples, a control and a feedback were interpreted as some general technique for changing the properties and behaviour of various objects. Perhaps, a control and a feedback should be estimated at their true worth and be promoted to a prin-ciple of feedback. It is in this very way that N. Wiener did when proclaiming cy-bernetics as a new science. Now, we return to a two-legged pacing. My story about this phenomenon will be from afar.

We have considered the case and got an idea of how a juggler forces a verti-cally standing stick not to fall down: he is controlling it . He is doing it by dis-placing its supporting point in accordance with its motion. It is clear that this is not a unique technique of controlling an inverted pendulum. A pendulum may be also controlled in another way. For example, you may connect to the end of the

pendulum one more section and maintain its direction constant. Do not ponder over why it is needed; instead, let us better see what will follow from it.

Thus, we have a double pendulum (Fig. 24.1). Its sections have lengths l_1 and l_2 ; the masses m_1 and m_2 that, for simplicity, are supposed to be concentrated on the terminals of the sticks; φ_1 and φ_2 are angles of sections with a vertical line. In the coupling of the sections there has been mounted a controlling device, which with the help of the moment M between the adjoined section and the pendulum is responsible for maintaining the angle φ_2 that should be always constant and equal to α . The moment M is unknown for us, but we know that the control system is coping with its task and, hence, the angle φ_2 is constantly equal to α. Let us construct a mathematical model for this inverted pendulum being controlled in such a fanciful way.
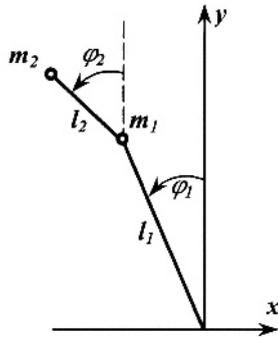
Fig. 24.1. A double pendulum.

The pendulum with a section attached is a double pendulum. You have already met it when we were discussing a juggling by two vertical sticks standing on each other. Now we are interested in some other thing, and namely, how the pendulum will be moving under the control described.

Let x_1, y_1 and x_2, y_2 be the coordinates of the point masses m_1 and m_2. From figure 24.1 it is apparent that

$$x_1 = -l_1 \sin \varphi_1, \quad y_1 = l_1 \cos \varphi_1,$$

$$x_2 = -l_1 \sin \varphi_1 - l_2 \sin \varphi_2, \quad y_2 = l_1 \cos \varphi_1 + l_2 \cos \varphi_2.$$

Further, we find the kinetic energy T and the potential function V for gravity:

$$T = \frac{1}{2}(m_1 + m_2)l_1^2\dot{\varphi}_1^2 + m_2 l_1 l_2 \dot{\varphi}_1 \dot{\varphi}_2 \cos(\varphi_1 - \varphi_2) + \frac{1}{2}m_2 l_2^2 \dot{\varphi}_2^2 \quad ,$$

$$V = (m_1 + m_2)gl_1 \cos\varphi_1 + m_2 g l_2 \cos\varphi_2 .$$

The generalized forces Q_1 and Q_2 corresponding to the moment M of the controlling mechanism are equal to $Q_1 = M$, $Q_2 = -M$.

Now we know the Lagrange function and the additional generalized forces and may write the Lagrange equations:

$$(m_1 + m_2)l_1^2\ddot{\varphi}_1 + m_2 l_1 l_2 \ddot{\varphi}_2 \cos(\varphi_1 - \varphi_2) -$$

$$- m_2 l_1 l_2 \dot{\varphi}_2 (\dot{\varphi}_1 - \dot{\varphi}_2)\sin(\varphi_1 - \varphi_2) +$$

$$+ m_2 l_1 l_2 \dot{\varphi}_1 \dot{\varphi}_2 \sin(\varphi_1 - \varphi_2) - (m_1 + m_2)l_1 g \sin\varphi_1 = M ,$$

$$m_2 l_2^2 \ddot{\varphi}_2 + m_2 l_1 l_2 \ddot{\varphi}_1 \cos(\varphi_1 - \varphi_2) -$$

$$- m_2 l_1 l_2 (\dot{\varphi}_1 - \dot{\varphi}_2)\dot{\varphi}_1 \sin(\varphi_1 - \varphi_2) -$$

(24.1)

$$- m_2 l_1 l_2 \dot{\varphi}_1 \dot{\varphi}_2 \sin(\varphi_1 - \varphi_2) - m_2 l_2 g \sin\varphi_2 = -M$$

or, taking into account that $\varphi_2 = \alpha$ $(\dot{\varphi}_2 = \ddot{\varphi}_2 = 0)$,

$$(m_1 + m_2)l_1^2\ddot{\varphi}_1 - (m_1 + m_2)l_1 g \sin\varphi_1 = M ,$$

$$m_2 l_1 l_2 \ddot{\varphi}_1 \cos(\varphi_1 - \alpha) - m_2 l_1 l_2 \dot{\varphi}_1^2 \sin(\varphi_1 - \alpha) -$$

(24.2)

$$- m_2 l_2 g \sin\alpha = -M .$$

This system of equations describes the variations of   the pendulum angle φ_1 with respect to the vertical line and the controlling moment M . Excluding   the unknown moment M from these equations, we come to the equation for the angle φ_1 . Eliminating an index from the angle φ_1 , let us write the obtained second-order differential equation as follows

$$[(m_1 + m_2)l_1^2 + m_2 l_1 l_2 \cos(\varphi - \alpha)]\ddot{\varphi} -$$

$$- m_2 l_1 l_2 \dot{\varphi}^2 \sin(\varphi - \alpha) - (m_1 + m_2)gl_1 \sin \varphi - \qquad (24.3)$$

$$- m_2 gl_2 \sin \alpha = 0.$$

To find the moment M for any solution of the equation (24.3), one may use each of the equations (24.2).

Let us begin studying the solutions of the equation (24.3), i.e. its phase portrait. First of all, reduce it to the form

$$[1 + \kappa \cos(\varphi - \alpha)]\ddot{\varphi} - \kappa \dot{\varphi}^2 \sin(\varphi - \alpha) - \omega^2 \sin \varphi - \kappa \omega^2 \sin \alpha = 0 \quad (24.4)$$

involving only two parameters, κ and ω

$$\kappa = \frac{m_2}{m_1 + m_2} \frac{l_2}{l_1}, \quad \omega^2 = \frac{g}{l_1}.$$

For $\kappa = 0$ the equation (24.4) is converted into the equation for the pendulum, which is standing up vertically for $\varphi = 0$. It turns out that the equation (24.4) can be integrated. For this purpose, we multiply it by $[1 + \kappa \cos(\varphi - \alpha)]\dot{\varphi}$:

$$[1 + \kappa \cos(\varphi - \alpha)]^2 \dot{\varphi}\ddot{\varphi} - \kappa[1 + \kappa \cos(\varphi - \alpha)]\dot{\varphi}^3 \sin(\varphi - \alpha) -$$

$$- \omega^2[1 + \kappa \cos(\varphi - \alpha)]\dot{\varphi} \sin \varphi - \kappa \omega^2[1 + \kappa \cos(\varphi - \alpha)]\dot{\varphi} \sin \alpha = 0$$

and transfer it to the form

$$\frac{1}{2}\frac{d}{dt}\{[1 + \kappa \cos(\varphi - \alpha)]^2 \dot{\varphi}^2\} -$$

$$- [1 + \kappa \cos(\varphi - \alpha)]\omega^2 (\sin \varphi + \kappa \sin \alpha)\dot{\varphi} = 0.$$

Integrating this equation, we find that

$$[1 + \kappa \cos(\varphi - \alpha)]^2 \dot{\varphi}^2 + F(\varphi) = h, \qquad (24.5)$$

where h is an integrating constant and

$$F(\varphi) = -2\omega^2 \int [1 + \kappa \cos(\varphi - \alpha)](\sin \varphi + \kappa \sin \alpha) d\varphi =$$

$$= 2\omega^2 [\cos \varphi - \frac{3\kappa}{2} \varphi \sin \alpha + \frac{\kappa}{4} \cos(2\varphi - \alpha) - \kappa^2 \sin \alpha \sin(\varphi - \alpha)].$$

Solving (24.5) regarding $\dot{\varphi}$, we get

$$\dot{\varphi} = \frac{1}{1 + \kappa \cos(\varphi - \alpha)} \sqrt{h - F(\varphi)}. \qquad (24.6)$$

As you already know, constructing a phase portrait according to (24.6) is reduced to finding a graph of the function $F(\varphi)$. However, for $\kappa \geq 1$, when the angle φ is approaching the value under which

$$1 + \kappa \cos(\varphi - \alpha) = 0, \qquad (24.7)$$

there may be an unbounded increase of the angular velocity.

This sudden phenomenon arises from the fact that, in order to implement a desired control, the controlling moment M will unboundedly grow, since, after excepting $\ddot{\varphi}$ from the equations (24.2), it will follow that

$$M = \frac{m_2 l_1 l_2}{1 + \kappa \cos(\varphi - \alpha)} \times$$

$$\times [\dot{\varphi}^2 \sin(\varphi - \alpha) + \omega^2 \sin \alpha - \omega^2 \sin \varphi \cos(\varphi - \alpha)], \qquad (24.8)$$

where φ_1 was replaced by φ.

An implementation of the unbounded moment is impossible and, therefore, there is no possibility to satisfy the condition $\varphi_2 = \alpha$, and we have to put some corrections into the mathematical model under consideration, if we want to study this case. These improvements can consist, for example, in assigning $M = M^* signM$, when $|M|$ has reached some limiting value, and in retaining this value as long as $|M|$, being determined from (24.8), turns out to be less than M^*. Here, the angles φ_1 and φ_2 are defined by the equations (24.1) as before. However, we restrain ourselves by the case when the described phenomenon does not occur. Only notice here that this phenomenon is similar to an impact and results in very fast changes of the velocities $\dot{\varphi}_1$ and $\dot{\varphi}_2$.

As it has been noted above, for constructing a phase portrait it is sufficient to have a graph of the function $F(\varphi)$ defined by the formula (24.5). The shape of

the graph for this function will depend upon two parameters, κ and α. Its shape is sufficiently apparent for small κ and α; it is shown in figure 24.2 for $\alpha > 0$.
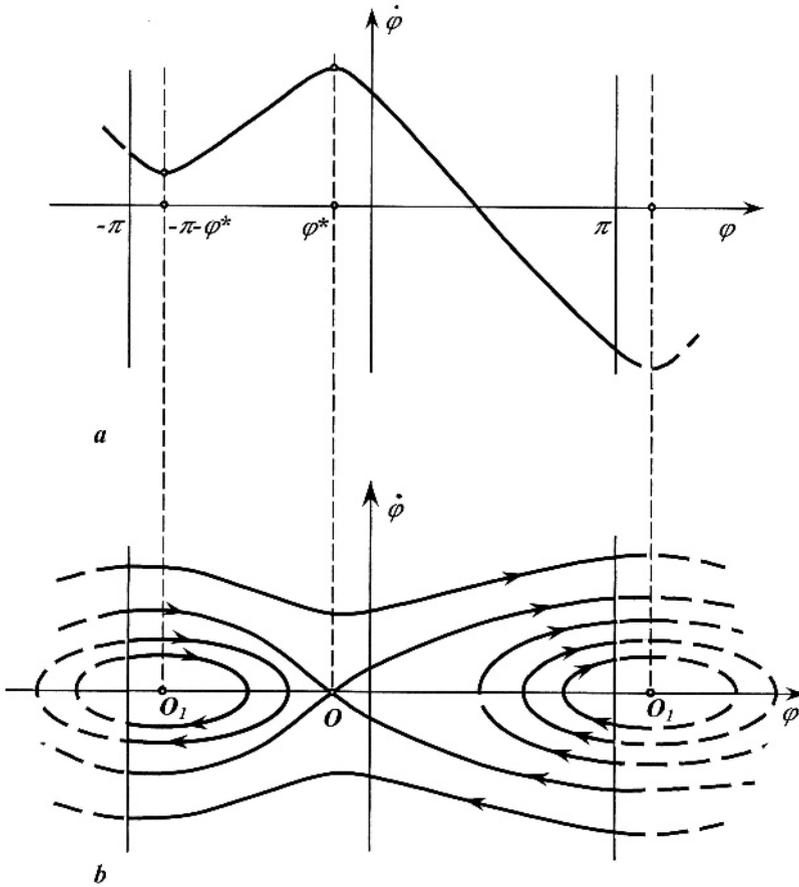
Fig. 24.2. The phase portrait in the cylinder development for the controlled double-section pendulum, obtained with the help of the graph of the function $F(\varphi)$.

According to this graph of the function $F(\varphi)$, there have been constructed the phase trajectories determined by the equation (24.6) for different h. A totality of these trajectories constitutes a phase portrait in a cylindrical phase space (Fig. 24.2). There are two equilibrium states O and O_1 on this phase portrait.

The   equilibrium $O(\varphi = \varphi^*)$,   where φ^* satisfies   the   condition $\sin \varphi^* = -\kappa \sin \alpha$,   is unstable and corresponds to the pendulum position shown in figure 24.3, i.e. to an inverted pendulum.
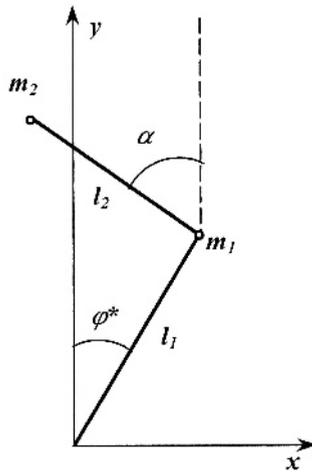
Fig. 24.3. A position of the double-link pendulum corresponding to the unstable equilibrium (O on the phase portrait in figure 24.2).

The equilibrium O_1, corresponding to the same condition, is associated with the pendulum suspended downward. This equilibrium is stable. Near this stable equilibrium, oscillations are possible. They relate to the closed phase trajectories in a region in figure 24.3 holding the equilibrium O_1.

All the rest phase trajectories, except the boundary trajectories and the separatrices of the saddle O, correspond to the motions with an unboundedly increasing velocity of the pendulum rotation ($\dot{\varphi} \rightarrow \infty$). This is a sudden effect of controlling the second section of the pendulum, i.e. the tendency to maintaining the angle between the second section and the vertical line equal to the same α leads to a rotation of the main pendulum section.

We have got used to the fact that one may start rotating a pendulum by pushing it to one and the same side, or turn it strongly by hand immediately or with the help of belt and gear transmissions.   In these cases, the force is of an external na-ture. However, it turns out so that the same effect can be obtained in some other way, with the help of attaching to this main pendulum some supplementary con-trolled pendulum. As a result, the main pendulum begins rotating in a faster way. Of course, such an unbounded untwisting presupposes an absence of dissipation (viscous and dry frictions). If dissipation, caused, for example, by a small viscous friction, is taken into account, then there appears a stable rotation Γ with some velocity, as it is apparent from figure 24.4.
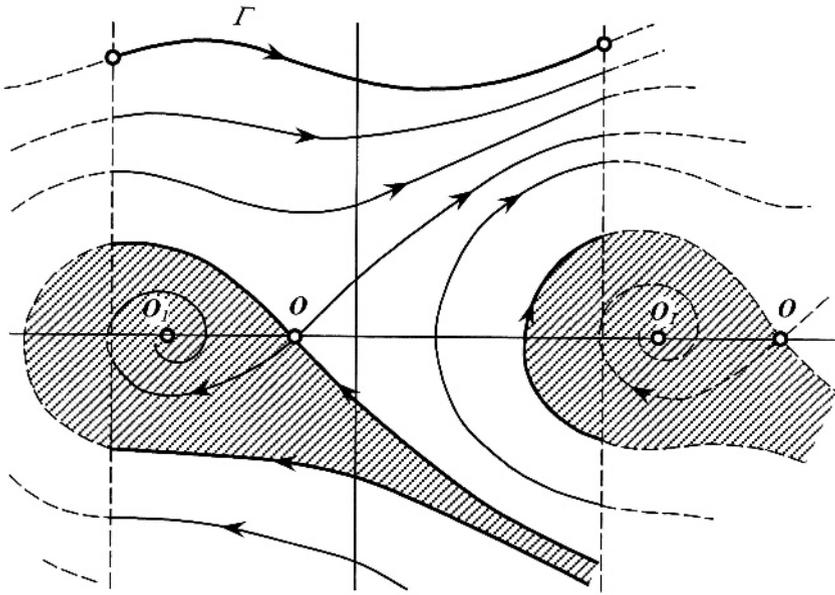
Fig. 24.4. A phase portrait of the inverted controlled double-section pendulum with dissipation taken into account.

In the domain, surrounded by the curve composed of the separatrices of the sad-dle O, all the phase trajectories converge to the stable equilibrium O_1. The equi-librium O, as before, is an unstable saddle. With the friction coefficient being further increased, all the motions will converge to the stable equilibrium O_1.

Of most interest for us will be the case when there arises the stable rotation Γ that relates to a phase portrait in figure 24.4.

Thus, if real energy losses are available, then the described control will yield a rotation of the pendulum. We have already noted previously an uncommonness of this way for generating and maintaining a rotation. In addition to this, we notice that this effect is unexpectedly exploited by humans in performing their two-legged pacing. Hardly ever you thought about how you are walking. But even a very superficial observation exposes the fact that when walking there is no need for us to push ourselves forward by legs. When walking you rather feel that some "force" is pulling you forward and you are only putting your legs in front of you, not to fall. You can feel this force drawing you forward not only while de-scending but also when walking along a flat road and even when ascending a small hill. This force, drawing you forward, can be amplified if you bend your body slightly ahead. To realise where this drawing force is coming from is possi-ble, if you look at the phase portrait in figure 24.4. Here, our legs we are resting upon may be treated as some section of an inverted pendulum and our body, at-tached to this leg as some attached section which we maintain to be at some con-stant angle to the vertical line: when we are standing the body is vertical, and when walking the body is slightly inclined ahead. Our supporting leg cannot per-form a complete rotation around its point of support; it only turns at some angle, from $-\beta$ to β. Whereupon, the leg H_1 is replaced and another leg H_2 will turn from $-\beta$ to β and so on all the time, if you do not vary a length of your pace (Fig. 24.5).
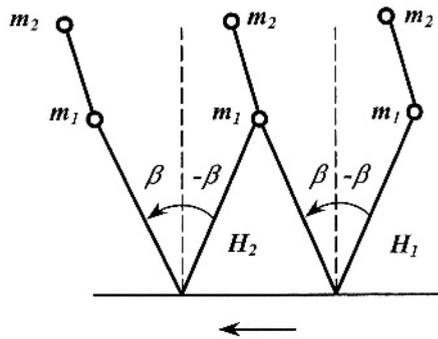
Fig. 24.5. A scheme of a two-legged pacing as a cyclically repeated motions of the con-trolled double-section pendulum, an auto-oscillating pacing .

Of course, here we are to some extent simplifying a moving manner of our sup-porting leg; the leg does not rest upon one point; supporting is performed by the

entire leg. In general, this pacing through changing legs looks similar to a rolling of a wheel with spokes, but without a rim.

According to the above said, let us draw two vertical lines $\varphi = -\beta$ and $\varphi = \beta$ on the phase portrait of figure 24.4 and analyse figure 24.6.
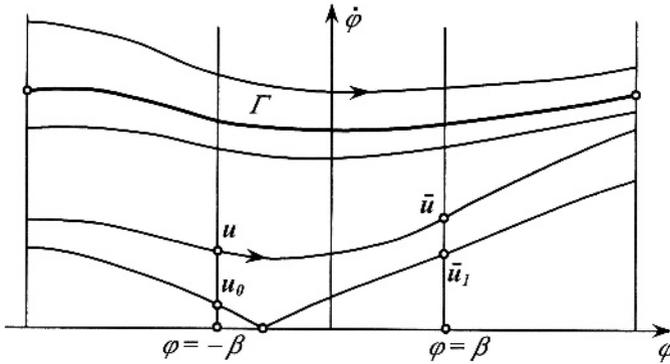
Fig. 24.6. A fragment of a phase portrait for the double-section pendulum with dissipation taken into account.

The phase trajectories leaving the points of the line $\varphi = -\beta$ can cross in their course the line $\varphi = \beta$. Suppose that, on the initial line, $\dot{\varphi} = u$ and on the final line, $\dot{\varphi} = \bar{u}$. The mapping of u into $\bar{u}$ of figure 24.6 has a diagram shown in figure 24.7. A stable fixed point of this diagram corresponds to the uniform pacing, where each new pace begins at the velocity u equal to the coordinate u^* of this fixed point. Besides, if you have used figure 24.2, then no dissipation is occurring at all, and if figure 24.4, then dissipation is occurring for the entire pacing period, from $\varphi = -\beta$ to $\varphi = \beta$.

During a real walk, dissipation takes place in the time of rolling on a supporting leg and, especially, when a replacing of a supporting leg creates some "jerk" against a support, and this is accompanied with some energy loss. Specificities of the energy loosing mechanism produce no effects upon a qualitatively correct shape of the diagram in figure 24.7, where the fixed point relates to a uniform periodic pacing .
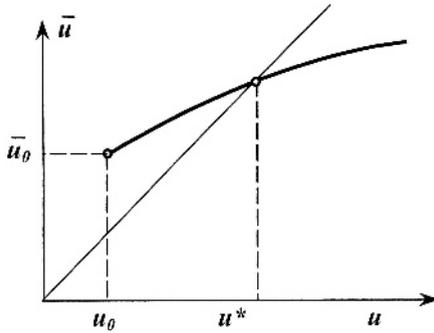
Fig. 24.7. The Keniks-Lamerey diagram proving an auto-oscillating pacing.

    Summarizing the things discussed above, I would like to notice that, as it fol-
lows from the preceding, not only humans but nature also has thought out its
wheel, and not only a wheel but also original techniques to rotate it. Among hu-
mans this rotation is based upon controlling a body's position and upon timely
replacements of legs, preventing a fall. It may be also noticed here that for dino-
saurs, for whom $\kappa > 1$, this technique needed additional supports, and one of
them was a huge tail. However, this is through a magic crystal only, since a de-
tailed analysis of the case $\kappa > 1$ has not been done by us.

25 Dynamical models for games, teaching and rational behaviour

Deterministic and stochastic automata models. Constructing and studying the dynamical models for automata games.

In the following chapters I will narrate the simplest dynamical models of rational behaviour, games and teaching. The development of these problems started recently and in its initial stage it was accompanied with a well-known public discussion on the subject "Is a computer able to think?". This discussion was fierce and, seemingly, turned out formally useless, since each of the arguing parties retained its opinion. One party asserted a computer to be capable of thinking so far as it is able to solve complicated mathematical problems, sometimes beyond human capabilities, and it can prove geometric theorems and find a way out from a labyrinth, can play chess and control a nuclear reactor or a missile, can "read" a draught and in compliance with it can supervise a production of a corresponding metal piece, etc. The opposite party objected and stated that a computer is incapable of thinking but only able to perform actions just outlined for it by humans.

It is clear that each party was right in its own way and this dispute seems senseless until there is made an agreement on what «thinking» is. For the time being, each party understands it in its own way.

Where is then a watershed between a thoughtless execution of computer instructions, statements and an intelligent search for a decision and its implementation? In what does "a thinking" differ from the execution of some program of actions? One should realize that any intelligent approach is also performed with use of a certain hidden algorithm not observable by us. This algorithm, perhaps, is not deterministic and carries in itself some elements of chance. May be in this very chance and in a later filtration there lies an essence of our thinking? Or perhaps, this essence lies in associations and in our skills to distinguish analogies and exploit the knowledge already gained, or, at last, in our ability to extract conclusions from unsuccessful and successful trials? But if all these things had really happened, they, nevertheless, happen in compliance with certain rules.

It is hardly possible, nowadays, to give a profound answer to all these questions, i.e. to define exactly what "a thinking" implies. And actually, how can it be done when we in fact know nothing about how our brain, being allotted by us a thinking capability, is functioning?

But we are well aware about how man-made computers operate, and with respect to them we can try to spot where arises the difference between a "thoughtless" execution of man-assigned instructions and where a "computer's thinking" starts. That we are adopting a computer thinking capability should not embarrass us, for this capability was granted to computers by nobody but a man; and this man himself was also granted this thinking capability – by nature. This thinking capability was granted to humans by the evolutionary process lasting for milliard years and this process was reduced to a severe selection of rational random changes.

Even within the frames of this specific and narrow-oriented statement of the problem it is hardly advisable to strive right now to some single-meaning and clear-cut answers. Instead, it is better to discuss these answers with you, having suggested you some priming for your meditation. This priming will be held in my story about simplest models in which there are already exposed the elements of something that is, perhaps, perceived to be a thinking.

Let us start from describing the automaton models for rational behaviour and learning. In our case, a decision-making subject will perform a "thinking" process in a manner like a certain automaton possessing an internal state upon whom a decision to be made depends. A state itself, in its turn, is determined by the success of the previous decisions. An idea and a study of such models was initially suggested by M.L. Tsetlin.

Let our automaton perform only two actions; then all its possible states are decomposed into two classes, in one class one decision is made and in the other another one. Let us denote these decisions as "+1" and "-1". As an example of such situation there may serve a question about in what hand of a referee a coin was hidden. The coin is used as a prize. The situation considered also covers the case when a rat is choosing between two feeding troughs as shown in figure 25.1.
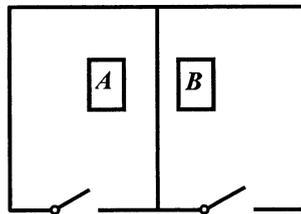
Fig. 25.1. A schematic position of the two feeding troughs for the experiment of teaching rats or other animals or birds.

Troughs were used to investigate learning capabilities of rats. A rat is released from time to time from a room through a passage leading to one of two troughs, to the left one, A, or to the right, B. Food is placed from time to time into one trough or another; and the rat, depending upon its choice, can satisfy its hunger or remain

hungry till other time. The first situation (with a coin) pertains to an outguessing game, and the second reminds a life-like situation of where to go for food.

Let us start from the simplest automaton descriptions for different types of players – a simpleton, a bull-head, a sly-boots and a mystic. Each player is assumed to be in one of two states, the state x_{-1} or the state x_{+1}; in one state he makes the decision – 1 and in the other the decision + 1 (Fig. 25.2).
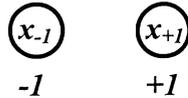
Fig. 25.2. The two different states of the automaton-player.

The (-1)-decision and the (+1)-decision imply respectively a player choosing the left or the right hand. A referee hides a coin into one of his hands and an outguesser outguesses its place. What current state a player has and how he changes his states in accordance with the results gained in the previous games will depend upon the type of this player, i.e. whether he is a simpleton, a bull-head, a sly-boots or a mystic. A simpleton always uses the following way of thinking: if he has won, then he should not change his opinion concerning a coin-keeping hand. On the contrary, if he has lost, he decides that his state has to be changed. Let the magnitudes $\xi = +1$ and $\xi = -1$ respectively stand for a win or a loss in the previous outguessing. Then, for $\xi = +1$ a simpleton will not change his state and for $\xi = -1$ he, on the contrary, will, as shown in figure 25.3.
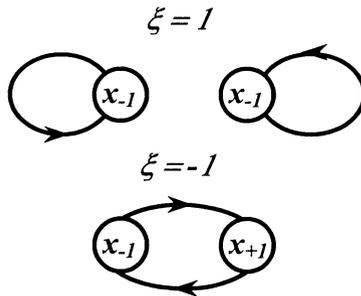
Fig. 25.3. The graph of state changes for the "simpleton" at the win ($\xi = 1$) or the loss ($\xi = -1$).

As for a sly-boots, his thinking is that upon his win his opponent will change the position of the coin and, therefore, this player, upon winning, decides to change his state. Conversely, upon losing he does not change his state, for he thinks that his opponent will change it himself. The strategy of a sly-boots is shown in figure 25.4.
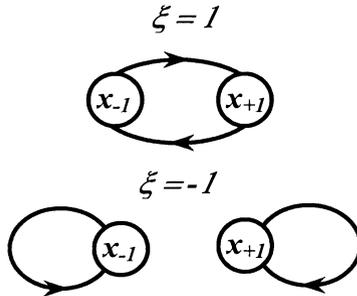
Fig. 25.4. The graph of state changes for the sly-boots at the win ($\xi = 1$) or the loss ($\xi = -1$).

As for a bull-head, he will not change his decision and all the time will remain in the state x_{-1} or x_{+1}, independently of the previous results. His state transition graph will not depend upon the result ξ of the previous game (Fig. 25.5).
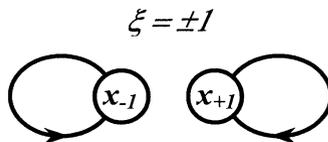
Fig. 25.5. The graph of state changes for the bull-head at the win ($\xi = 1$) or the loss ($\xi = -1$).

A mystic will leave his decision to chance, through casting some (not obligatorily symmetric) cubic and on this basis will independently take his state x_{-1} or x_{+1} with the probabilities p and $q = 1 - p$. Here, he pays no attention to his wins and losses, absolutely relying upon his talisman, i.e. his cubic. This behaviour may be also modelled by a two-state automaton. Now, the states of this automaton are changed not in strict accordance with its loss or win but rather

according to arrows and with the probabilities p and q shown on them (see the state transition graph in figure 25.6).
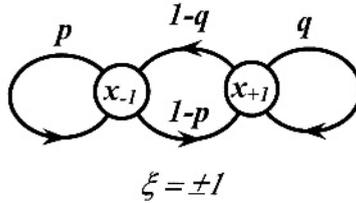
Fig. 25.6. The graph of the transition probability for the "mystic".

Thus, we have described an outguessing simpleton, a sly-boots, a bull-head and a mystic. Similar strategies may be used by a simpleton, a sly-boots, a bull-head and a mystic referee. Similar strategies may be also employed by a rat searching for food; though, for a rat the strategies of a sly-boots, a mystic or a bull-head will be hardly efficient. Though, here everything will depend upon the strategies used to fill the troughs with food.

In our story about possible strategies in the outguessing game we restricted ourselves to the two states only and assumed, except the mystic, these states to be deterministically changed after each game, depending upon its result. The number of states can be increased and their changes may be governed stochastically. For example, the simpleton may be introduced as equipped with some memory and whose opinion is changed not immediately. Upon his first failure, he will only start doubting and only upon his repeated failure he will change his opinion. How such a retentive simpleton changes his states is shown in figure 25.7. Here, you see four states: the extreme right and left one relate to assurance in the decision to be adopted; the middle states, on the contrary, pertain to the situation when the decision is made, but with some portion of doubt. The number of states may be, certainly, not only four, but any. To each of the decisions there may correspond not necessarily equal numbers of states, as it was in our previous case where each player had two states.
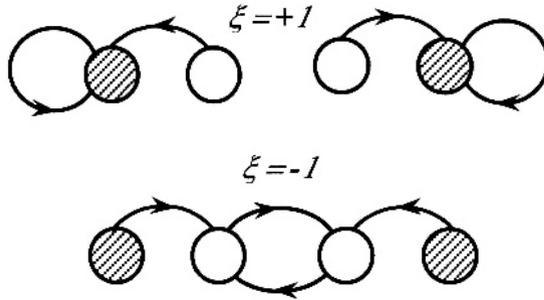
Fig. 25.7. The graph of the state changes for the simpleton with a memory (the so-called linear playing strategy).

Other automaton models may be also constructed, for example, with the state transition graph presented in figure 25.8.
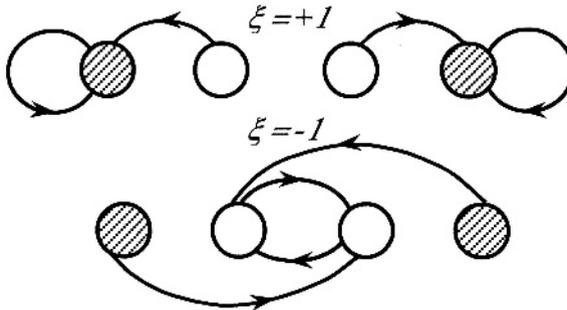
Fig. 25.8. The automaton with a memory but its state change graph is different from that in figure 25.7.

Figure 25.9 shows the state transition graph for a two-state automaton in the case when the states are changed randomly, with the probability $1-p$, if it wins (when $\xi=+1$), and accordingly with the probability $1-q$, if it loses (when $\xi=-1$). The player with the stochastic strategy, depicted in figure 25.9, for $p=1$, $q=0$ will be a simpleton; for $p=0$, $q=1$ will be a sly-boots; and for $p=1$, $q=1$ a bull-head.

We have thus described the simplest strategies in the outguessing game against another player or against "nature". Nature may also be considered to be a player retaining indifferent to its loss or win.
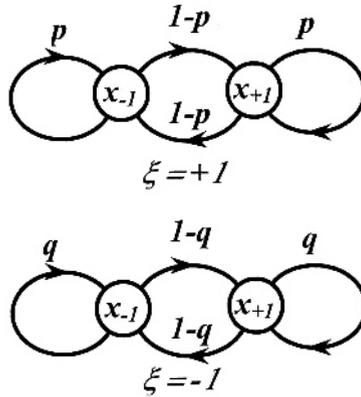
Fig. 25.9. The graph of transition probability for the "stochastist".

The simplest models for the strategies of stochastic nature can be represented by a sequence of independent decisions with the probabilities p and $q\,(p+q=1)$, as the mystic has, or by a markovian sequence determined by a matrix of state transitions. For example, there are the two states x_{-1} and x_{+1} in which the decisions -1 and $+1$ are made respectively. The states x_{-1} and x_{+1} are changed randomly, with some fixed probability, say, δ (Fig. 25.10).
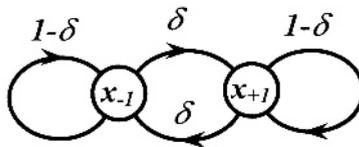
Fig. 25.10. The graph of transition probabilities for the "stochastist" retaining indifferent to a win or a loss.

Now, upon getting familiar with the simplest automaton models of players, let us consider an outguessing game between two players, a simpleton and a sly-boots. Let us see who will win. A simpleton's state will be denoted by x_{-1} or x_{+1}, and

that of a sly-boots by y_{-1} or y_{+1}. The two players will possess four states, (x_{-1}, y_{-1}), (x_{+1}, y_{-1}), (x_{-1}, y_{+1}) and (x_{+1}, y_{+1}). In the states x_{-1} and y_{-1}, the outguesser chooses the left hand and the referee hides a coin in his left hand. In the states x_{+1} and y_{+1} the outguesser chooses the right hand and the referee puts a coin into his right hand. In the states (x_{-1}, y_{-1}) and (x_{+1}, y_{+1}), the outguesser will win, and in the states (x_{-1}, y_{+1}) and (x_{+1}, y_{-1}) the referee will. Now, we are able to depict the state transition graph for the dynamical system "referee-outguesser" shown in figure 25.11. The result will be unexpected to some extent. If the first game was won by the outguessing simpleton, then all the remaining games will be won by him as well.
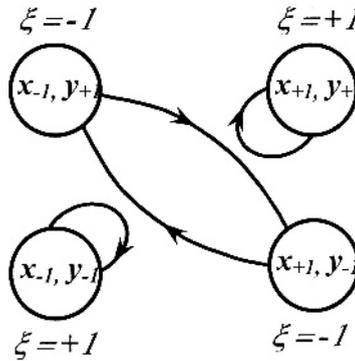
Fig. 25.11. The phase portrait for the "simpleton – slyboots" game.

Conversely, if the first game was won by the referee - slyboots, then he will win further.

A similar graph for two simpletons is shown in figure 25.12. Their game runs periodically, with a four-game period – two games are won by the referee-simpleton and the other two by the outguesser. Thus, the total result of their games will be a draw.

Now, let us consider a two-player game with stochastic strategies. This case covers both all the specific variants of the games like a sly-boots, a simpleton, a mystic or nature with $p = q$ (Fig. 25.10). Let one player have the associated probabilities p, q and the other the probabilities r, s. The players may possess four combinations of their states, (x_{-1}, y_{-1}), (x_{+1}, y_{-1}), (x_{-1}, y_{+1}) and (x_{+1}, y_{+1}). In the first and the last states, the outguesser will win, and in the two
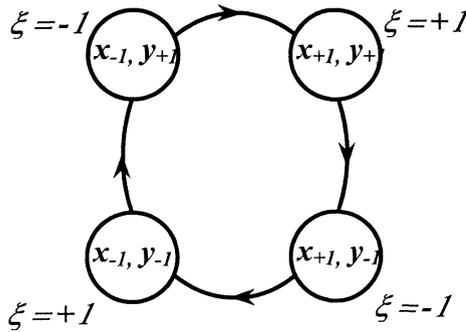
Fig. 25.12. The phase portrait for the game of two "simpletons".

middle states the referee will. Both players may change their states in any case –
in case of a win, with the probabilities $1-p$, and in case of a loss, $1-r$, with
the probabilities $1-q$ and $1-s$, respectively. Let these players sequentially
play some games and $p_n(x_{-1}, y_{-1})$, $p_n(x_{+1}, y_{-1})$, $p_n(x_{-1}, y_{+1})$ and
$p_n(x_{+1}, y_{+1})$ be the probabilities of the associated state combinations x_i and
y_j in the n-th game. In the initial moment, there will take place a certain
combination of states x_i, y_j, so that only the single probability $p(x_i, y_j)$ will
be equal to 1, whereas others will be equal to 0.

It is apparent that the combinations (x_i, y_j) are no states any more, since the
knowledge of x_i and y_j is not sufficient to predict what x_i and y_j will be in
the next game. Though, the probabilities for these states can be predicted. Thus,
this situation brings us to the idea that as a state there may be employed the vector
$\overline{p}$ whose components are the probabilities $p(x_i, y_j)$.

We assume now that in the n-th game a state vector will be known and equal
to $\overline{p}^n$. Let us find the vector $\overline{p}^{n+1}$, i.e. its components $p^{n+1}(x_{-1}, y_{-1})$,
$p^{n+1}(x_{+1}, y_{-1})$, $p^{n+1}(x_{-1}, y_{+1})$ and $p^{n+1}(x_{+1}, y_{+1})$, through
$p^n(x_{-1}, y_{-1})$, $p^n(x_{+1}, y_{-1})$, $p^n(x_{-1}, y_{+1})$ and $p^n(x_{+1}, y_{+1})$. It turns
out that

$$\overline{p}^{n+1} = \overline{p}^n P \ , \tag{25.1}$$

where the transition probability matrix P is equal to

$$\left\|\begin{array}{cccc} ps & (1-p)s & p(1-s) & (1-s)(1-p) \\ (1-q)r & qr & (1-q)(1-r) & q(1-r) \\ q(1-r) & (1-q)(1-r) & qr & (1-q)r \\ (1-p)(1-s) & (1-s)p & s(1-p) & ps \end{array}\right\| \quad (25.2)$$

and $\overline{p}^{\,n}$ and $\overline{p}^{\,n+1}$ are row vectors.

If the state combinations (x_{-1}, y_{-1}), (x_{-1}, y_{+1}), (x_{+1}, y_{-1}), (x_{+1}, y_{+1}) are denoted with digits 1,2,3,4 and the elements of the matrix P with p_{ij}, where i is a number of the row and j is a number of the column, then p_{ij} will be a probability of the transition from the state combination with the number i to the state combination with the number j. For example, the element p_{12} will be the probability of the transition from the state combination (x_{-1}, y_{-1}) to (x_{+1}, y_{-1}) being equal to the product of the probability that the winning outguesser will change its state, i.e. $1-p$, by the probability that the losing referee will retain its state, i.e. s. All the remaining elements of the matrix are found similarly.

The validity of the formula (25.1) follows from the relation

$$p_i^{\,n+1} = p_1^{\,n} p_{1i} + p_2^{\,n} p_{21} + p_3^{\,n} p_{3i} + p_4^{\,n} p_{4i}$$

that implies that the i-th combination of automaton states may appear, with the probability p_{1i}, from the first combination; from the second combination, with p_{2i}; from the third combination, with p_{3i}; and from the fourth one, with the probability p_{4i}; whereas the probabilities of the first, second, third and fourth combinations will be $p_1^{\,n}, p_2^{\,n}, p_3^{\,n}$ and $p_4^{\,n}$, respectively.

Thus, the knowledge of the probability vector $\overline{p}^{\,n}$ makes it possible to find its next value $\overline{p}^{\,n+1}$ after playing one more game. This allows us to consider this game of two stochastic automata to be a dynamical system with the state described by the vector $\overline{p}$.

A phase space of this dynamical system will be all possible vectors $\overline{p}(p_1, p_2, p_3, p_4)$ whose all p_i are not negative and their sum is equal to unity, because the system under consideration will be obligatorily in one of the incompatible states (1, 2,3 or 4). This phase space is three-dimensional and represents itself a three-dimensional simplex. The vertices of this simplex Σ_3,

depicted in figure 25.13, correspond to unity vectors $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ and $(0,0,0,1)$. On its four edges, we shall have $p_1 = 0$, $p_2 = 0$, $p_3 = 0$ and $p_4 = 0$, respectively.

The dynamical system with the phase space like the simplex Σ_3 (and in a general case, with a simplex of any dimension) and with the operator described by the matrix P, according to the formula (25.1), will be called a discrete markovian system.

For a general case, the phase portrait of such a system will be very simple: all its phase trajectories will converge to the single equilibrium $\overline{p}*$ determined by the condition

$$\overline{p}^{n+1} = \overline{p}^{n} = \overline{p}*$$

or, according to (25.1), by

$$\overline{p}* = \overline{p}* P \quad . \tag{25.3}$$

The equation (25.3) in a general case has the single normed nonzero solution $p*$. The markovian system is defined by the matrix P, i.e. by its elements $p_{ij} \geq 0$. Therefore, the case with no one of p_{ij} being equal to zero is understood to be a general case. In our specific case, when the matrix P is of the type (25.2), the general case occurs for p, q, r and s being different from zero and unity. When this general condition is fulfilled, the equation (25.3) has the single solution $\overline{p}*(p_1*, p_2*, p_3*, p_4*)$.

This is a sufficient condition for the existence of the single globally stable equilibrium to which all phase space points tend asymptotically. The necessary and sufficient condition is that some degree of the matrix P, i.e. P^k must satisfy this sufficient condition of positiveness of all its elements.

In the situations (x_{-1}, y_{-1}) and (x_{+1}, y_{+1}), i.e. in the first and the fourth, the outguesser will win, and in the rest (in the second and the third situations) he will lose as much as he wins (a five-kopeck coin). Therefore, in a long game the mathematical expectation of his win f will be close to its value in the equilibrium state

$$Mf = 5(p_1* + p_4*) - 5(p_2* + p_3*) \quad . \tag{25.4}$$

The form of the transition probability matrix (25.2) will be used for studying the games of the simpleton, the sly-boots and the mystic against the stochastist, with the probabilities $0 < r < 1$ and $0 < s < 1$.
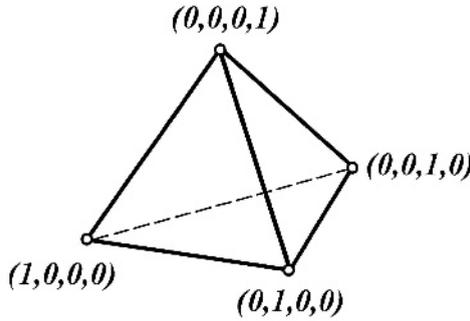
Fig. 25.13. The phase space in the form of a three-dimensional simplex for the case when at least one player changes his state randomly with the probabilities dependent upon the result of the previous game.

For the simpleton we have $p = 1$ and $q = 0$. Thus, in accordance with (25.1), the matrix P is written in the form

$$P = \begin{Vmatrix} s & 0 & 1-s & 0 \\ r & 0 & 1-r & 0 \\ 0 & 1-r & 0 & r \\ 0 & 1-s & 0 & s \end{Vmatrix} \tag{25.5}$$

that brings us to the equilibrium equations

$$p_1{}^* = sp_1{}^* + rp_2{}^*$$

$$p_2{}^* = (1-r)p_3{}^* + (1-s)p_4{}^*$$

$$p_3{}^* = (1-s)p_1{}^* + (1-r)p_2{}^* \tag{25.6}$$

$$p_4{}^* = rp_3{}^* + sp_4{}^* ,$$

to which there should be added the below rating conditions:

$$p_1{}^* + p_2{}^* + p_3{}^* + p_4{}^* = 1 . \tag{25.7}$$

The equations (25.6) and (25.7) are easily solved if at first we find the needed $p_1{}^*, p_2{}^*, p_3{}^*$ and $p_4{}^*$ from (25.6) with accuracy of up to a common factor

$$p_2 = p_3 = \frac{1-s}{r} p_1, \qquad p_4 = p_1; \qquad (25.8)$$

then, this factor p_1 is found from (25.7). Therefore, we find that

$$p_1{}^* = p_4{}^* = \frac{r}{2(1+r-s)}, \qquad p_2{}^* = p_3{}^* = \frac{1-s}{2(1+r-s)}. \qquad (25.9)$$

Thus, the mathematical expectation (25.4) for the simpleton's outguessing win will be equal to

$$Mf = 5\frac{-1+r+s}{1+r-s}. \qquad (25.10)$$

Now, it becomes easy to study how a win in the globally stable equilibrium will depend upon the probabilities r and s. Assuming $Mf = 5\mu$, we, by virtue of (25.10), shall easily arrive at the case that on the line

$$s + \frac{r(1-\mu)}{1+\mu} = 1 \qquad (-1 \le \mu \le 1), \qquad (25.11)$$

inside the square $0 < r < 1$, $0 < s < 1$, the mathematical expectation (25.10) will be equal to 5μ (Fig. 25.14).

Therefore, depending upon the probabilities r and s, the simpleton can both win and lose and also obtain any intermediate result. The simpleton is close to a maximally possible win at $\mu \approx +1$; vice versa, at $\mu \approx -1$ to a maximal loss. In the point $r = 0$, $s = 1$ (Fig. 25.14), corresponding to the sly-boots, all the lines of the equal wins 5μ will intersect. Depending upon their initial states, the sly-boots will always win or lose against the simpleton. The stochastic players being close to the sly-boots can obtain any win from 5 to -5, depending upon the values $r \approx 0$ and $s \approx 1$.

Such a drastic dependence of the win 5μ upon the values r and s in the neighbourhood of the point $r = 0$, $s = 1$ and the absence of the certain value μ in this point is explained by the fact that for $r = 0$, $s = 1$ the existence of the globally stable equilibrium is broken.
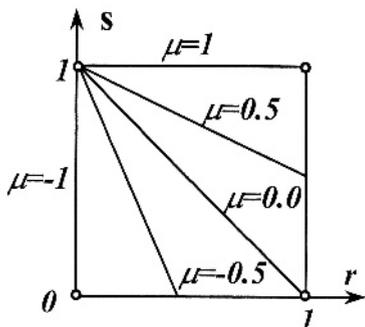
Fig. 25.14. The geometrically represented dependence of the win μ upon the state changing probabilities r and s in the "simpleton – stochastist" game.

For the "slyboots – stochastist" game we find that (for $p = 0, q = 1$ in the formula (25.2)) the matrix P will be of the form

$$P = \begin{Vmatrix} 0 & s & 0 & 1-s \\ 0 & r & 0 & 1-r \\ 1-r & 0 & r & 0 \\ 1-s & 0 & s & 0 \end{Vmatrix}. \qquad (25.12)$$

Therefore, an equilibrium state is found from the equations

$$p_1{}^* = (1-r)p_3{}^* + (1-s)p_4{}^*$$

$$p_2{}^* = sp_1{}^* + rp_2{}^*$$

$$p_3{}^* = rp_3{}^* + sp_4{}^* \qquad (25.13)$$

$$p_4{}^* = (1-s)p_1{}^* + (1-r)p_2{}^* \ ,$$

from the rating condition (25.7). From (25.13) and (25.7), we find that

$$p_1{}^* = p_4{}^* = \frac{1-r}{2(1+s-r)} \ , \qquad p_2{}^* = p_3{}^* = \frac{s}{2(1+s-r)} \ . \qquad (25.14)$$

Further, it turns out that the mathematical expectation of the sly-boots' win in a globally stable equilibrium state will be equal to

$$Mf = 5\frac{1-r-s}{1+s-r} \; . \tag{25.15}$$

The expectation Mf found will be equal to $-Mf$ from the formula (25.10), if in this formula s and r exchange their places. Here, the situation is represented by the same graph from figure 25.13, provided that μ is replaced by $-\mu$ and the axis r by s and s by r.

Now, let us look at the game of the mystic, for whom $p = q = \alpha$, against the stochastist. In this case, the transition probability matrix P, in accordance with (25.2), will be of the form

$$\left\| \begin{array}{cccc} \alpha s & (1-\alpha)s & \alpha(1-s) & (1-\alpha)(1-s) \\ (1-\alpha)r & \alpha r & (1-\alpha)(1-r) & \alpha(1-r) \\ \alpha(1-r) & (1-\alpha)(1-r) & \alpha r & (1-\alpha)r \\ (1-\alpha)(1-s) & \alpha(1-s)0 & (1-\alpha)s & \alpha s \end{array} \right\| \tag{25.16}$$

According to (25.16), the equations of the stable equilibrium will be written as

$$(\alpha s - 1)p_1{}^* + (1-\alpha)rp_2{}^* + \alpha(1-r)p_3{}^* + (1-\alpha)(1-s)p_4{}^* = 0$$

$$(1-\alpha)sp_1{}^* + (\alpha r - 1)p_2{}^* + (1-\alpha)(1-r)p_3{}^* + \alpha(1-s)p_4{}^* = 0$$
$$\tag{25.17}$$
$$\alpha(1-s)p_1{}^* + (1-\alpha)(1-r)p_2{}^* + (\alpha r - 1)p_3{}^* + (1-\alpha)sp_4{}^* = 0$$

$$(1-\alpha)(1-s)p_1{}^* + \alpha(1-r)p_2{}^* + (1-\alpha)rp_3{}^* + (\alpha s - 1)p_4{}^* = 0 \; .$$

The equations (25.17) are linearly dependent and, therefore, admit the nonzero solution $p_1{}^*, p_2{}^*, p_3{}^*, p_4{}^*$, which then has to be normalized. Through manipulating this system of equations (25.17) (i.e. by adding the first equation with the third one and the third with the fourth) we shall reduce this system to the below form (omitting the fourth)

$$p_1{}^* - p_2{}^* + p_3{}^* - p_4{}^* = 0$$

$$(1-s)p_1{}^* + (1-r)p_2{}^* + (r-1)p_3{}^* + (s-1)p_4{}^* = 0$$

$$\alpha(1-s)p_1*+(1-\alpha)(1-r)p_2*+(\alpha s-1)p_3*+(1-\alpha)sp_4*=0 \quad .$$

We directly find that the needed normalized solution will be as follows

$$p_1* = p_4* = \Delta^{-1}(2-s-r)(r+\alpha-2\alpha r)$$

$$p_2* = p_3* = \Delta^{-1}(2-\alpha-s)(\alpha+s-2\alpha s)$$

$$\Delta = 2(\alpha-s-r)[2\alpha+s+r-2\alpha(r+s)] \quad .$$

The mathematical expectation of a win in the equilibrium state will be

$$Mf = 5(p_1*+p_4*-p_2*-p_3*) = \frac{5(r-s)(1-2\alpha)}{2\alpha+r+s-2\alpha(r+s)} \quad . \qquad (25.18)$$

From the expression (25.18), it follows that for $\alpha = 0.5$ the mystic will neither lose nor win from any automaton (for any r and s). On the contrary, for $\alpha \neq 0.5$ he may both lose and win.

For $\alpha < 0.5$, he will win for $r > s$ and lose for $r < s$. Here, his maximum win from the simpleton will be for $r = 1$, $s = 0$; his maximum loss to the sly-boots will be for $r = 0$, $s = 1$.

For $\alpha > 0.5$, everything will be vice versa. The mystic will maximally win for $r < s$ from the sly-boots and for $r > s$ will maximally lose to the simpleton. (In our conclusions on a maximum win and a maximum loss, α is assumed to be different from 0 and 1, which is again explained by the violation of conditions for the global stability of the equilibrium).

Thus, we have considered both the simpleton – sly-boots game and the games of the simpleton, the sly-boots and the mystic against the stochastist as well. We have found the probabilities of various playing situations in the steady-state equilibrium and also the wins and the losses of players in it. In order to get a more complete idea concerning these games, I would like to remind you the associated state transition graphs for the deterministic simpleton - sly-boots game and the simpleton - simpleton game (see Figs 25.11 and 25.12). Here, the situation transition graphs are simultaneously phase portraits, because a playing situation here is nothing but a state. For stochastic games, a situation ceases to be a state any longer and a situation transition graphs is no more a phase portrait. A phase portrait will be all possible phase trajectories of the point $\overline{p}(p_1, p_2, p_3, p_4)$ of the three-dimensional simplex

$$p_s \geq 0 , \qquad p_1 + p_2 + p_3 + p_4 = 1 \quad .$$

For a general case of positive values of all transition probabilities p_{ij}, this phase portrait will be simple and will hold phase trajectories contracting to the single stable equilibrium $\overline{p}*(p_1*, p_2*, p_3*, p_4*)$ (Fig. 25.15).
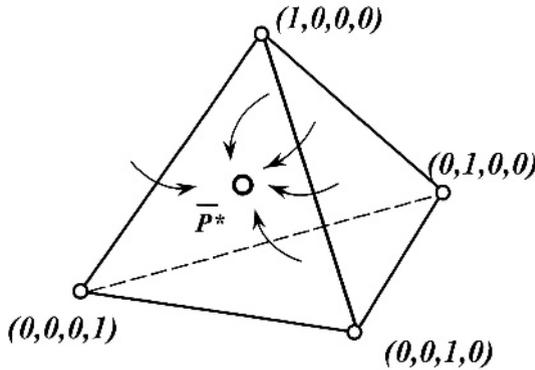
Fig. 25.15. The illustration of presence of the globally stable equilibrium in the game of randomly-changed states.

The situation transition graph may be immediately constructed with use of the transition probability matrix P (25.2). The matrices of state changes in the game of the simpleton or the sly-boots against the stochastist and in the game of the stochastic mystic against the stochastist will constitute particular cases of this matrix P. We may, therefore, restrict ourselves only to demonstrating figure 26.16 for the transition graph of the matrix P of the type (25.2).

We have, thus, considered very simple examples of the outguessing games. Where is then a teaching in them? Teaching in such games lies in the fact that the players (the simpleton, the sly-boots and the stochastist) were playing not under a certain preassigned schedule but in strict accordance with the results of the previous games, since they were making a decision for each next game on the basis of the result obtained. The bull-head was not trained; he was acting according to a previously compiled plan. He could be also of another nature, implementing, say, some other arbitrary preassigned sequence of decisions. As for the mystic, this player had not been trained either; though, having no previously given plan, his decisions, nevertheless, did not depend upon the previously gained results and, instead, were governed by chance. The playing strategies for the bull-head and the mystic imply no teaching. Instead, the strategies of the simpleton, the sly-boots and the stochastist are such that teaching takes place.

The simplest strategies in the outguessing game discussed above can be also applied to rat's decision making as to what feeding trough to go to eat. This case relates to the game of the outguesser against nature, since the food-carrying troughs are determined by the reasons independent of the rat's actions. Nature's

strategy also differs from the referee's strategy in the following: each time there obligatorily exist not only a single food-filled trough and an empty trough but also possible is the situation of both troughs being filled with food or both being empty.

In interpreting an outguessing game as well as in the case with the rat searching for food via learning itself by the results of the previous searches – in all these cases a learning component is evident. This component is incorporated into a player itself and is very simple, even primitive. That is why its principle difference and absolutely new capabilities, matched against the strictly assigned behavioural strategies not exploiting the previously gained experience, have not been still realized at present. I would like to emphasize here that these new learning capabilities imply not only other algorithms for actions but also some new information about the results of the previous actions used by the algorithm.

In the above examples, this information was exploited for decision-making about the automaton's transition to this or that state. There exists a more general and deeper employment of such information for modifying the automaton itself; in particular, for changing the probabilities p and q of the stochastic automaton.

The rat's behavioural strategy with a learning component incorporated seems
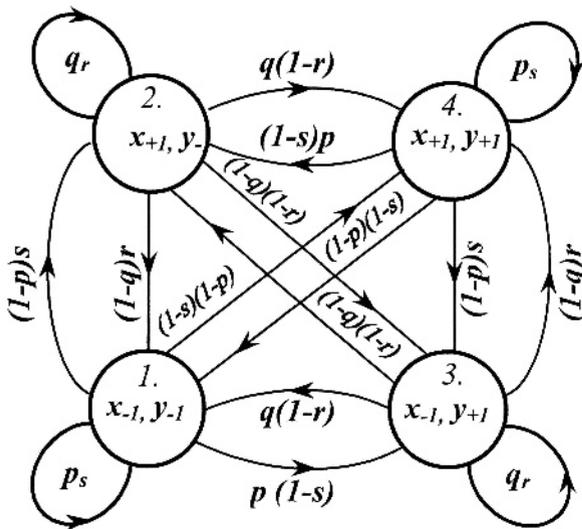
Fig. 25.16. The transition probabilities graph for the game of two "stochastists".

expedient. Nevertheless, why it is so is not so evident, because it is not clear at all against what it must be compared and against what this strategy looks better. This question is not so simple. Its discussion here could lead us far away from the subject. Therefore, let us restrict ourselves to only comparing this strategy against

the random behavioural strategy and which is better than a random strategy will be considered as more rational. The random behaviour with independent choice has equal probabilities for a win and a loss. The rational strategy will be the one for which a win will exceed 0.5. Such an understanding seemingly agrees with our intuition. A more complete analysis of this question will be, probably, given to you in your special courses.

In conclusion, I would like to touch the two general points not related directly to our subject, viz. to explain why the stochastist with $p = q = 0.5$ wins from nobody and loses to nobody and why the markovian system with the matrix P, with all its elements p_{ij} being positive, has the single stable equilibrium state $\overline{p}*$. That he neither wins from nor loses to the simpleton, the sly-boots and the stochastist was revealed by the specific analysis of the mathematical models of their games. Though, here naturally arises the following question: is it possible to generate a successful strategy against the stochastist who with equal probabilities and independently chooses -1 and $+1$? (This kind of strategy is a dream of fanatics in all gambling houses.) A simple analysis reveals that this strategy is impossible. It is impossible because the stochastist's realizations -1 and $+1$ are independent either upon his previous actions or his opponent's actions. This very independence makes conditional probabilities of his wins and losses constantly similar, equal to 0.5.

In our time, D.M. Tchertok and I tried to devise the more complicated learning strategies than the above described and to overplay a human with their help. A human is not an equal-probability stochastist, i.e. he inevitably takes into account the results of his previous games. If this very account of his has been recognized, then a human's loss will be inevitable. In order to implement the strategy based upon recognizing his strategy, it is necessary to construct a mathematical model for a playing human. For this, rather efficient becomes the assumption that a human strategy is random in its nature, but the probabilities of choosing -1 and $+1$ depend upon finite, though not so great, number of previous games. What a human chooses in the final end, -1 and $+1$, will not be equiprobable and not independent.

Alongside with it, a computer overplaying a human in such a way puts itself into a risky position. It may lose to man, provided that man has thought out a strategy superior to the computer's. This human strategy is based upon recognizing the computer's strategy. Mutual perfection of these playing strategies may last long, like wars among humans. This analogy is not so direct, since in any war not only its strategies are subjected to perfection but also its fighting facilities. These facilities may be also considered to be a strategy. At present, mankind is facing a vitally important danger – a breakage of the chain of evolutional perfections which was incorporated into the instincts of humans as representatives of living creatures. This breakage is seemingly possible only through apprehending the destruction of the "natural" path of human civilization and understanding the fact that in the time of coming ecological catastrophe the collective behavioural strategies should be crucially improved.

In this connection, of special interest becomes the problem of creating the coexistence models based not only upon Darwinian struggle for existence but also upon the principle of human coexistence under new living conditions. This is some kind of a jump to the new evolutionary strategies resting not upon open struggle for existence but upon compromising strategies and agreements. Such strategies and agreements are mutually profitable for a short and a long periods of time and are based upon the comprehension that except primitive and instinctive strategies there can exist substantially more efficient and rational strategies.

That these strategies exist is demonstrated by Ju. B. Germeyer's example which very clearly explains a possible existence of strategies being more profitable than a direct maximization of one's own win. We imply here the n-player game, where the s-th player's win is equal to

$$f_s = 1 - x_s + \mu \sum_{i \neq s} x_i \qquad (s = 1,2,...,n) \ . \tag{25.19}$$

In this game, any player exploits his own way of choosing his variable $0 \leq x_s \leq 1$.

It is clear that the immediate approximation by the s-th player of his win f_s will lead us to the choice $x_s = 0$. If all the players do this, a win of each of them will be then equal to 1. But if it is agreed between them that they will not tend to an individual win and, instead, all will choose $x_s = 1$, then everybody's wins will not be equal to unity, but, perhaps, to much more, $\mu(n-1)$. In its nature such an agreement is not stable, because for each individual player it becomes favourable to violate the agreement and obtain a far greater win, $\mu(n-1)+1$. Though, if m players behave in this way, then their win will be $\mu(n-m-1)+1$ and the win of others will be $\mu(n-m-1)$. It is clear that if $\mu(n-m-1)$ turns out to be less than unity, then the above agreement will be not profitable for players. Hence, such "a union" will be broken – and this will be unfavourable for all of them. It is possible to establish arrange a more stable agreement, namely, through arranging a certain common cash box. That is, they may agree between themselves to keep all wins in this common box and then, for example, distribute them equally among the members. In this case, a total win will be equal to

$$F = \sum f_s = n - x_1 - ... - x_n + \mu(n-1)(x_1 + ... + x_n) =$$
$$= n + (\mu(n-1)-1)(x_1 + x_2 + ... + x_n) \ .$$

Let we have $\mu(n-1)-1 > 1$, then not for all together but for each player it becomes advantageous to choose $x_s = 1$.

The second general question has been, perhaps, explained to you at your lectures on the theory of probabilities. This surprising and important general fact

needs a heuristically visual understanding that it is given below. Here, I just wish to recall you a single thing, i.e. that the conditions $p_{ij} > 0$ are sufficient but not necessary. Though, necessary and sufficient conditions are known as well. But let us not touch this point here though these conditions are also beautiful and immediately checked through a shape of the situation transition graph. Thus, let all $p_{ij} > 0$ be positive.

In this case, the linear transformation with the matrix P converts the simplex Σ_n again into the simplex $\overline{\Sigma}_n$ lying strictly inside the simplex Σ_n. It lies strictly inside because all the vertices of the simplex Σ_n (1,0, ...,0), (0,1, ...), ..., (0,0, ...,1) are converted, due to $p_{ij} > 0$, strictly inside the simplex Σ_n. It means that the simplex Σ_n is transformed into the inside of the geometrically similar simplex $\widetilde{\Sigma}_n$ which is $q < 1$ times less than the simplex Σ_n. By the next transformation, the simplex $\widetilde{\Sigma}_n$ will analogously be converted into the q times lesser simplex $\widetilde{\widetilde{\Sigma}}_n$ remaining similar to it, etc. Hence, a sequence of such similar and infinitely lessening simplexes will contract to some internal point $p *$.

26   Perceptron and pattern recognition

The Rosenblatt perceptron as a dynamical system. Mathematical models for an object, an image, recognition and teaching a recognition. The theorem about the finiteness of the number of errors. A relation between the perceptron teaching algorithm and the stochastic approximation.

Now let us come to the model that has come into being comparatively not long ago, namely, in the middle of the previous century. It became already widely popular in the 60s. This was a model of the human's intellectual activity. It was implemented on a computer and this thus proves that a computer may have a thinking capability. The model has shocked contemporaries and opened a new branch in cybernetics – pattern recognition. This mathematical model was preceded by a wonderful device designed in 1957 by Frank Rosenblatt. He gave his device a very sounding name, a "perceptron". This name was chosen on the basis of the then conceptions on the structure of brain, eye retina and the conceptions about processing retina's signals in the nerve cell layers. By profession F. Rosenblatt was a physiologist. His device was able to perceive, distinguish and recognize visual patterns. It could distinguish triangles and ovals, recognize letters and other images. The most amazing thing was that these capabilities had not been previously built into the computer and rather came into being through teaching the perceptron. Through training the perceptron became capable of distinguishing new geometrical figures and visual patterns. Earlier these skills usually betrayed a human intelligence; and this capability of computers was beyond their capabilities. The surprise caused by these capabilities was enlarged by the fact that it was absolutely incomprehensible how the perceptron was performing these actions – since everybody could see only a board of photosensitive cells imitating the eye retina, the random connections of these cells with the A-layer, B-layer, an adder, a formal neuron, various feedbacks. All these things were somehow interrelated and were running in a certain way, in accordance with unpredictable principles of neurodynamics. Possibly, if this device had not been a real one but merely its description and if there had been the same talking about its capabilities, then nobody would have paid any attention to it. But here there was a really existing and functioning device, and it was recognizing patterns and learning itself how to recognize.

F. Rosenblatt wrote a thick book, "Principles of neurodynamics", that was translated into Russian. However, most likely it was read by nobody, for it was hardly understandable to anybody except the author. Nevertheless, a human thinking was in progress. These principles were understood. They turned out to lie in the amazingly simple, far from neurophysiology, mathematical model for the feature space and in the model of a pattern being some domain in this space. This

mathematical model exploits some abstract fantastic concept of the multidimensional space born by the boundless imagination of mathematicians. Then, any mystery, confusion and incomprehensibility have vanished. Everything has become very simple, even trivial. This powerful strength of simplicity has brought to life a fast development of a new branch of cybernetics, "pattern recognition", and its numerous applications. Nowadays, a robot is capable of playing a piano at sight, i.e. looking at notes placed before it; a machine-tool is able to manufacture metal pieces according to a designer's draft; a computer can type an article or a book from a manuscript or at voice; a robot can also diagnoze a disease, reveal faults in an engine, reveal a deposit of mineral resources and can perform other numerous intelligent actions.

Here, we again are returning to the problem of perception. I will not block up your mind with small details of the perceptron design, though they might be greatly useful for you. Now, I will try to describe the perceptron's functioning from the point of view of its contemporary understanding.

With the help of its board of 100 x 100 photocells (elements), the perceptron perceived some visual pattern as a combination of 10^4 signals $x_1, x_2, ..., x_{10000}$. These signals formed a vector of features or a description x of this pattern. To each incoming signal x, the perceptron responded with "yes" or "no", or with "+ 1" or "– 1". How it responds in each specific case to a given input signal x depends upon its internal state which is characterized by the values of the parameters $\gamma_1, \gamma_2, ..., \gamma_m$ constituting the vector γ. These parameters γ may vary under the actions of the teacher who in each case informs whether the perceptron's response was correct or not. All the objects to be demonstrated to the perceptron are supposed to be divided by the teacher in two classes (class A and class B). The teacher wants to teach the perceptron to classify these objects so that the perceptron would respond with "$+1$" if it is shown an A-class object and with "-1" when shown a B-class object. The teacher teaches the perceptron this recognizing skill through showing it different patterns, thus causing the corresponding changes in the internal state γ of the perceptron. The teacher hopes that upon this teaching the perceptron will correctly respond (like a teacher) to the new shows which had not been, perhaps, suggested to it during the previous teaching session.

According to the above description, an internal structure of the perceptron, whatever it may be, can be represented by the model

$$y = f(x, \gamma) \tag{26.1}$$

where $f(x, \gamma)$ is a function of the two vectors x and γ. This function can take only the two values, $+1$ and -1. In the simplest variant of the Rozenblatt perceptron, this function is of the form

$$f(x, \gamma) = sign \sum_{s=1}^{m} \gamma_s \varphi_s(x_1, ..., x_n) = sign \gamma \varphi(x) \ . \tag{26.2}$$

The above model simulates only the perceptron's response to the signal x; here, the process of teaching the perceptron is yet left beyond description.

In general, teaching may be thought as exploiting the teacher's information by the perceptron for determining its state vector γ , i.e.

$$\gamma = g(x^1, z^1; \ x^2, z^2; ..., x^P, z^P) \ . \tag{26.3}$$

Here, x^i, z^i is a description of the i-th object being demonstrated and the teacher's indication to what pattern this description belongs: to the pattern A, if $z = -1$; and to B, if $z = 1$. The sequence $x^1, z^1; ...; x^P, z^P$ makes up the teaching information. The perceptron employs this information to determine its state γ. When a new show is done, the perceptron's response (26.1) to the show of the object with the description x will be determined by x and the state γ. The teacher's response z will be determined by some function $f_{teach}(x) = z$ unknown for us.

Thus, the teaching purpose may be thought as bringing the perceptron's function (26.1) closer to the unknown teacher's function $f_{teach}(x)$, through using the teaching information $I(x^1, z^1; ...; x^P, z^P)$. Now, after this explanation, let us describe the teaching algorithm (26.3) used for the simplest version of the perceptron.

Let X be a set of all possible objects x. These objects are finite or infinite in number. Each object $x \in X$ is referred by the teacher to one of the classes, A or B. This capability of the teacher may be represented by the certain function $f_{teach}(x)$ that will assume for any object x only one of the two values, $+1$ or -1. An ideal purpose for teaching the perceptron is to get the pupil (i.e. the perceptron) doing what its teacher is doing , i.e. for any $x \in X$ to have

$$y = f(x, \gamma) = sign \ \gamma \varphi(x) = f_{teach}(x) \ . \tag{26.4}$$

It is clear here that the principle possibility of teaching is connected with the existence of such a number $\gamma = \gamma *$ for which at all $x \in X$ we could get (26.4). Then comes the question as what the teaching (26.3) should be in order to find such $\gamma *$, if any. We recall here that in the case under consideration the function $f(x, \gamma)$ is determined by the formula (26.2).

Rozenblatt has thought out the following algorithm for choosing the needed $\gamma *$ during the sequential shows of the objects $x^1, x^2, ...$ and the teacher's responses $f_{teach}(x^1)$, $f_{teach}(x^2), ...$ informing whether the perceptron has responded correctly or not.

Let us take the arbitrary γ. If the perseptron's response is correct, then this fact is confirmed by the teacher. The perceptron, being satisfied, will change nothing, i.e. it will leave γ without any change. If the perceptron's response does not co-incide with the teacher's, i.e. the response is wrong, then γ will be changed, thus, bringing the perceptron closer to the correct response. Namely, let the teacher's response be $+1$ and the perceptron's -1; then we have

$$f(x,\gamma) = sign\ \gamma\varphi(x) = -1 \ .$$

In order to bring the perceptron closer to the true response $+1$, it is possible to add the vector $\varepsilon\varphi(x)$ $(\varepsilon > 0)$ to the vector γ, i.e. to substitute γ by $\bar{\gamma}$ according to the formula

$$\bar{\gamma} = \gamma + \varepsilon\ \varphi(x) \ .$$

Here, we obtain

$$\bar{\gamma}\varphi(x) = \gamma\varphi(x) + \varepsilon\varphi^2(x) \quad (\varepsilon > 0) \ .$$

Hence, $f(x,\bar{\gamma}) > f(x,\gamma)$, and in this way the perceptron has come nearer to the right response.

Now vice versa, let the teacher's response be -1 and the perceptron's $+1$. Then, the magnitude $\gamma\varphi(x)$ should be reduced, i.e. γ should be substituted by $\bar{\gamma}$ in accordance with the formula

$$\bar{\gamma} = \gamma - \varepsilon\ \varphi(x) \ ,$$

since $f(x,\gamma) = \gamma\varphi(x)$ will be decreased by $-\xi\varphi^2(x)$.

Let the teacher's response $f_{teach}(x)$ and the perceptron's response be not co-incident, then, in accordance with the above said, γ will be changed according to the formula

$$\bar{\gamma} = \gamma + \mathscr{E}f_{teach}(x)\varphi(x) \ . \tag{26.5}$$

This change will take place only when the perceptron has performed an error. If the perceptron is responding correctly, then instead of (26.5) there will take place

$$\bar{\gamma} = \gamma \ . \tag{26.6}$$

The equations (26.5) and (26.6) may be easily written as a single formula

$$\bar{\gamma} = \begin{cases} \gamma & for \quad f_{teach}(x) = f \\ \gamma + \mathscr{E}f_{teach}(x)\varphi(x) & for \quad f_{teach}(x) \neq f \end{cases} \tag{26.7}$$

or

$$\bar{\gamma} = \gamma - 0.5\varepsilon(f - f_{teach})\varphi(x) \ ,$$

where f is the perceptron's response determined by the formula (26.2) and f_{teach} is the teacher's response to the show of one and the same object $x \in X$.

Prior to studying the teaching algorithm (26.7), let us describe briefly and schematically how this was implemented in the Rozenblat's perceptron.

As said before, in designing his perceptron Rozenblat proceeded from his understanding the visual receptor structure and from his understanding the technique of how the eye retina and the brain perform a further processing of this visual image. This processing was thought by him to be done on a layer-by-layer basis, viz. a layer of receptors; a layer of constructing the functions $\varphi(x_1,...,x_n)$, a layer for multiplying the functions φ_s by the coefficients γ_s; at last, a neuron implementing the threshold function $sign$. Accordingly, a schematic representation of this process is given in figure 26.1.
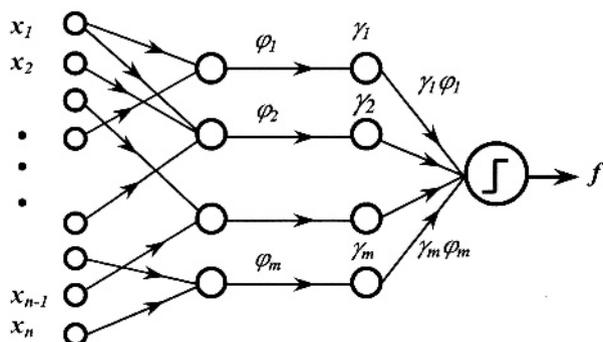
Fig. 26.1. The scheme of the three-layer perceptron with a formal neuron (a threshold element) in the output (the feedbacks of the phase variables $\gamma_1,...,\gamma_m$ in the third layer not shown).

The teaching mechanism was implemented through matching a neuron's response against a teacher's response; the result of this comparison was then used to correct, through the feedbacks, each of the multipliers γ_s, in accordance with the formula (26.7). Such is a schematic description of the perceptron's operation.

In addition to this functional and schematic description, the perceptron may be described geometrically. The pattern x, appearing at the input of the perceptron, may be interpreted as an appearance of the point x in some multidimensional

feature space R_n. In this space, the function (26.2) implemented by the perceptron may be geometrically represented as the fact on what side of the surface

$$\gamma\varphi(x) = 0 \tag{26.8}$$

this point x is lying. If it lies on one side from this surface $\gamma\varphi(x) > 0$, then this will correspond to the response $+1$. If on another side from the surface $\gamma\varphi(x) < 0$, then this will indicate the response -1. The set of points X is divided into the A-class points and B-class points which must lie on a different side from the surface (26.8).

According to the above said, a set of points x relating to the pattern A may be treated as a model for the pattern A and a set of points of the pattern B accordingly as a model for the pattern B. In this case, a teaching process is thought to be a constructing of the surface (26.8), splitting the sets of points of the patterns A and B, in accordance with the formula

$$x \in \begin{cases} A \ , & if \quad \gamma\varphi(x) > 0 \\ B \ , & if \quad \gamma\varphi(x) < 0 \end{cases}.$$

Thus, the intelligent pattern recognition may be geometrically interpreted as constructing a surface, which will separate the sets of points (models) representing the patterns A and B.

This geometrical interpretation of the visual image recognition has given rise to enormous consequences. Under patterns people now started to imply not only visual patterns but any other represented numerically, by somewhat numerical features $x_1, x_2, ..., x_n$. This kind of understanding has made it possible to recognize various classes of patterns – handwriting, various sounds, oil fields, text reading, human voice and words pronounced, medical and engineering objects, etc.

At last, also important and interesting is the fact that the perceptron is a dynamical system and that a process of its teaching and recognizing itself is also dynamical.

A state of the perceptron is an m-dimensional vector γ. An operator of this dynamical system is given by the formula (26.2) which determines a change of the state γ. A perceptron's response to the external action x is described by the formula (26.2). Gaining by a perceptron its recognizing skill will imply that a perceptron has arrived to its equilibrium state in which the perceptron's state γ retains unchanged at any show x.

Now let us prove a possibility of teaching the perceptron with use of the teaching algorithm (26.7). Here, we shall consider the perceptron as a dynamical system. The perceptron state vector γ will be interpreted as a point in the multidimensional phase space Γ_m. The formula (26.7) will define a law of state changes and may be considered as an operator of the dynamical system. The perceptron

states $\gamma *$, in which the perceptron has been taught and ways responds like the teacher does, will not (due to 26.7) change and will constitute, therefore, its equilibrium states.

It is easily seen that the set of equilibrium states, if any, will form a convex cone. Indeed, the equilibrium $\gamma *$ will meet the requirement

$$sign\, \gamma * \varphi(x) = f_{teach}(x) \qquad (26.9)$$

for all $x \in X$. Though, if (26.9) was fulfilled for some $\gamma *$, then it will be also fulfilled for $\lambda \gamma *$ at any $\lambda > 0$. Analogously, if $\gamma_1 *$ and $\gamma_2 *$ are equilibriums, then $\lambda_1 \gamma_1 * + \lambda_2 \gamma_2 *$ will be also an equilibrium at positive λ_1 and λ_2 (in particular, for $\lambda_2 = 1 - \lambda_1$ and $0 \le \lambda_1 \le 1$).

Now, let us prove that if the perceptron has a cone of equilibriums containing the internal point $\gamma *$, then it will be able to make not more than some finite number of errors and upon it the perceptron will cease to make them. This stunning theorem is proved in a comparatively simple way. Its idea is that any discrete phase trajectory of the perceptron at any preset sequence of shows cannot hold more than some finite number of points. From this, it does not follow that this trajectory should obligatorily terminate in an equilibrium; upon some finite number of errors the perceptron will simply cease to make errors at any preassigned sequence of shows. Though, if we sequentially show all the points of the set X some times, then the phase trajectory will obligatorily terminate in an equilibrium and the perceptron will learn itself to correctly recognize all the points from the set X.

To prove this wonderful theorem by Novikov is not difficult, as was said above.

Let for $x \in X$ we have $0 < r \le \|\varphi(x)\| \le R < \infty$ and let $\gamma *$ be an internal point of the equilibrium set which corresponds to the trained perceptron. Then, we shall obtain

$$\left| \gamma * \varphi(x) \right| \ge \rho \|\gamma *\| \|\varphi(x)\| \,, \qquad \rho > 0 \qquad (26.10)$$

independently of the length of the vector $\gamma *$ that may be chosen arbitrarily.

Actually, if $\gamma *$ is an internal point of the equilibrium cone, then for some $\rho > 0$ and for any vector α, for which $\|\alpha\| \le \rho$, we shall get the below inequality

$$\left| (\gamma * + \alpha \|\gamma *\|) \varphi(x) \right| > 0$$

for all $x \in X$, or , taking $\alpha = v\varphi(x) / \|\varphi(x)\|$, the inequality

$$\left| \gamma * \varphi(x) + v \right\| \gamma * \| \| \varphi(x) \| \| > 0$$

for any $\left| v \right| \le \rho$, from which (26.10) will follow. Also, let γ^0 be an arbitrary initial state of the perceptron and $\gamma^1, \gamma^2, \gamma^3, \ldots$ be a sequence of states of the perceptron in the course of its training by showing it some arbitrary points $x^1, x^2, x^3, \ldots$ from the set X . Let us see how the distance $V(\gamma^s) = (\gamma^s - \gamma*)^2$ of the point γ^s from the point $\gamma*$ will vary with growth of s . We find directly that

$$V(\gamma^{s+1}) - V(\gamma^s) = (\gamma^s + \varepsilon f_{teach} \varphi(x) - \gamma*)^2 - (\gamma^s - \gamma*)^2 =$$

$$= 2\varepsilon f_{teach} \gamma^s \varphi(x) - 2\varepsilon f_{teach} \gamma* \varphi(x) + \varepsilon^2 f^2_{teach} \varphi^2 (x) \quad .$$

Since

$$f_{teach} \gamma^s \varphi(x^s) < 0 \qquad f_{teach} \gamma* \varphi(x) > 0 ,$$

then, taking into account (26.10), we get

$$V(\gamma^{s+1}) - V(\gamma^s) < -2\varepsilon \left| \gamma * \varphi(x) \right| + \varepsilon^2 \varphi^2 (x) \le$$

$$\le -2\varepsilon \rho \| \gamma * \| \| \varphi(x) \| + \varepsilon^2 \varphi^2 (x) \quad .$$

Since the length of the vector $\gamma*$ is arbitrary, let us choose it such that

$$- 2\varepsilon \rho \| \gamma * \| \| \varphi(x) \| + \varepsilon^2 \varphi^2 (x) \le -q < 0 \quad .$$

Then, with each transition from γ^s to γ^{s+1} the distance to $\gamma*$ will decrease by not less than $q > 0$ and, therefore, the number of such transitions cannot exceed $N = \dfrac{V(\gamma^0)}{q}$. Hence, the statement has been proved.

Let us find a direct expression of the estimate N for the upper limit of the perceptron's errors. At $\| \gamma * \| = \dfrac{\varepsilon R}{\rho}$ and $\gamma^0 = 0$ we get

$$N = \frac{\varepsilon^2 R^2}{\rho^2} \max \left[2\varepsilon \rho \frac{\varepsilon R}{\rho} \| \varphi(x) \| - \varepsilon^2 \varphi^2 (x) \right]^{-1} =$$

$$= \frac{R^2}{\rho^2} \max_{r \le \xi \le R} \left[2R\xi - \xi^2 \right]^{-1} = \frac{1}{\rho^2} \quad ,$$

and this is natural, since a necessary decrease of ρ implies a greater difficulty of splitting the subsets A and B of the set X.

Thus, the number of perceptron's errors will not exceed some finite number if the perceptron is able in principle to recognize the patterns A and B. This capability will retain even when the set X is infinite. If the set X is finite, then by showing this set repeatedly the perceptron will be finally trained of correct recognition during the number of shows not exceeding N. On the contrary, if the set X is infinite, then any length of the teaching session will not guarantee a complete teaching of the perceptron. In order to speed up the perceptron teaching, it is desirable for the perceptron to perform during this teaching as many errors as possible, in order to exhaust the admissible number of them. When this number has been exhausted, the perceptron will correctly classify all the remaining points , though they were not shown by the teacher. In this extrapolating predicting capability, there lies its most essential importance and practical weight. This capability of the perceptron, in particular, enables, upon demonstrating some patients of some disease , to successfully diagnoze this disease among new patients. Through showing the letter A, it becomes possible to recognize this letter in other writings, etc.

Above, alongside with the description of perceptron's model there has been also described a general model of teaching to recognize two classes of objects (that only two classes were presented is not a substantial restriction, since via a dichotomy it becomes possible to split objects into any number of classes). This model includes the decision rule $f(x, \gamma)$ and a teaching algorithm in the form (26.3) which is an algorithm for changing the state γ. The contemporary knowledge does not permit to indicate all possible successful algorithms for any given decision rule $f(x, \gamma)$. Neither can be shown the best of the decision rules. But there already exist numerous teaching algorithms. Now we will show you one common technique for their construction. It is based upon minimizing a stochastic functional and was called the method of stochastic approximation.

Let we have some functional $G(x, \gamma)$ being dependent upon some random vector $x \in X$ and the vector of parameters λ. The random vector has some distribution function $\rho(x)$. The stochastic approximation provides sequential approximations to the value $\lambda *$, which will minimize the mathematical expectation of the functional

$$MG(x, \lambda) = \int_x G(x, \lambda) \rho(x) dx \qquad (26.11)$$

over the parameter λ. These subsequent approximations are described by the formula

$$\lambda^{p+1} = \lambda^p - \varepsilon^p \nabla_\lambda G(x^{\rho+1}, \lambda^\rho) \quad , \tag{26.12}$$

where points $x^1, x^2, \ldots$ will appear randomly, according to the probability density $p(x)$.

We are not going to formulate the conditions for the convergence of the sequential approximations $\lambda^1, \lambda^2, \ldots$ to the minimum of the functional (26.11). This, if needed, may be found out by you from the associated lecturing courses or books. Here, we restrict ourselves to showing only how this general technique makes it possible to arrive at the above teaching algorithm and to other numerous algorithms, as was shown by Ja. Z. Tsypkin.

It is clear that a teaching purpose may be formulated as the attempt to perform possibly less number of errors, i.e. to search for such $\lambda *$ at which the decision rule $f(x, \lambda)$ will admit a minimum of errors. Now let us introduce the penalty function $G(x, \lambda)$ instead of number of errors and let us demand that the mathematical expectation of this penalty function be minimal. The required λ may be found in accordance with (26.12).

Assume that

$$f(x, \lambda) = sign\lambda \, \varphi(x)$$

$$G(x, \lambda) = (f(x, \lambda) - f_{teach}(x))\lambda\varphi(x)$$

and, by virtue of (26.12), we shall find

$$\lambda^{p+1} = \lambda^p - \varepsilon^p \nabla_\lambda (f(x, \lambda) - f_{teach}(x))\lambda\varphi(x) =$$
$$= \lambda^p - \varepsilon^p (f(x, \lambda) - f_{teach}(x))\varphi(x) \quad ,$$

which, as easily seen, coincides with (26.7) for $\varepsilon^\rho = 0.5\varepsilon$.

The penalty function $G(x, \lambda)$ was chosen so as to turn to zero if the response is true $(f = f_{teach})$. It will become equal to $\left| \lambda\varphi(x) \right|$ if the response was wrong $(f \neq f_{teach})$. Via choosing other penalty functions, it is possible to construct other teaching algorithms.

27 Kepler laws and the two-body problem solved by Newton

The two-body problem and its relations with some astronomic problems like black holes, universe extension and solar system evolution.

Here I am starting a narration about a dynamical system whose description and mathematical simulation have passed across the entire history of mankind. From Egyptians, Babylonians, Arabs and ancient Greeks to Europeans and to the entire contemporary mankind there has stretched a chain of these investigations performed for twenty five centuries and reflecting people's variational thinking about the environmental world and their place in this world, and also about their capabilities to implement these investigations in their life. This chain started from primitive descriptions of the immediately observed, then it passed through a comprehension of the magnificent intentions of gods and the world harmony created by them, through describing this harmony by perfectly ideal mathematical images (by circles, spheres, regular polyhedrons, the golden section, etc). Then, this chain was extended to the contemporary laconic purely mathematical abstract modelling and investigations, and to understanding a unity of the worlds, terrestrial and celestial, a unity of their laws expressed in the mathematical language. This system embraces the starry sky with the Sun, the Moon, the planets, and the Earth we are living on.

A day is replaced with a night. This happens for centuries and centuries. Winter is replaced with spring, spring with summer, summer again with winter, and all these events occur all the time. Lives of humans are immediately connected with this eternal replacement of year seasons, days and nights. The Moon appears and vanishes, its phases are changed; the planets on the sky queerly move on the firmament. These mysterious metamorphoses could not help exciting a human; his life seemed to him wholly dependent upon their mercy, being at the mercy of this magnificent divine sky and heavenly bodies.

From our school days we are aware of the geocentric system of the universe created by ancient Greeks and the people in the Middle Ages. We know about the Ptolemaic system (a descriptive kinematic model of the universe), and the geocentric system by Copernicus that replaced the Ptolemaic system. We are also familiar with the Kepler laws and the universal gravitation. But how from these laws of I. Newton the three laws of J. Kepler follow remains mysterious for us so far. And this very mystery we will try to shatter. This point will be the subject of the coming story.

The Kepler laws describe the movements of celestial bodies in an incomparably simpler way. They are able to predict the positions of the planets, in contrast to the very complicated kinetic model by Ptolemy and the simpler Copernican model in which travelling planets were represented through a principally similar approach, and namely, by a superposition of uniformly circular revolutions. The 77 circles held in the Ptolemaic system were reduced in the Copernican system to only 34. Both systems, Ptolemaic and Copernican, suggested a very large predicting accuracy, up to an order of angle minutes. Though, these systems required very complicated calculations. All these things were surprisingly simplified by Kepler, though his description retained still kinematic. The three laws by Kepler were not interrelated and had no unique explanation either. It was I. Newton who explained them from a unique position, in a unique way, not only them but also all mechanical phenomena occurring in the sky and on the Earth. Everything has become as a corollary of one and the same laws, the laws of mechanical movements. Now let us recall the Kepler laws.

His first law considers an orbit of each planet to be an ellipse, and in one of the foci of the ellipse there lies the Sun.

His second law describes how and at what velocity each planet travels along its ellipse. Its velocity is variable but its so-called sector velocity, i.e. the velocity at which the radius-vector connecting the Sun and the planet circumscribes an ellipse, retains constant

The third law, according to Kepler, is a general harmony of celestial spheres being manifested by a constant ratio of the squared circulating period of the planet to the cubed semi-axis of its ellipse.

As a whole, the three laws by Kepler provide a full kinematic description of the regularities in the planet movements within the solar system. But are there any more general foundations for these regularities? Cannot they be a corollary of any more general laws of nature? In order to answer these questions, Newton had to overcome tremendous difficulties. Through a titanic effort of his human genius he laid a foundation for a number of new scientific directions – mechanics, theory of gravitation, differential and integral calculus. Moreover, the calculation of the Kepler laws, according to Newton, was for that time an unimaginably difficult problem but Newton, nevertheless, "calculated" them.

These calculations will be repeated by us now but in a contemporary way. Though, prior to this let us cite one extract from Newton's writings where his methodology is seen to be very close to ours. Here, we see the place of the subject of our narration in his general program for the universe comprehension. His extract speaks: "The entire difficulty of physics … lies in the fact that through the phenomena of movements it should recognize natural forces and then through these forces explain the remaining phenomena".

In his annotation to his narration on the mysterious gravitation acting through emptiness he wrote: "…hypotheses I do not invent. It is enough that gravitation exists in reality and acts in accordance with the laws described by us; and it is completely enough to explain all movements of the heavenly bodies and the sea".

Perhaps, it is worth to recall here that in his approach great Newton followed great G. Galileo.

Now , as we did before, we have to construct a mathematical model and study it. The objects to be simulated are the two point masses being mutually attracted by the force equal to

$$\gamma \frac{mM}{\rho^2} \quad ,$$

where m and M are masses of the attracting bodies; ρ is a distance between them; and γ is an experimentally exposed coefficient, a gravitational constant. The attracting force is directed along the line connecting the two bodies.

Let $\bar{r}$ and $\bar{R}$ be radius-vectors of these masses in some inertial coordinate system (Fig. 27.1); then, in accordance with the second law of Newton
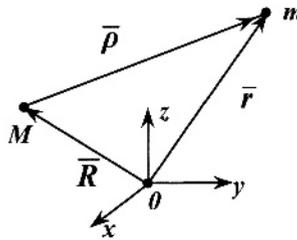
Fig. 27.1. The two-body problem graphically illustrated: m and M are masses mutually attracted under the Newton law.

and the law of gravity, we obtain

$$M\ddot{\bar{R}} = \gamma \frac{Mm}{(\bar{R} - \bar{r})^2} \frac{\bar{r} - \bar{R}}{|\bar{r} - \bar{R}|}$$

$$\tag{27.1}$$

$$m\ddot{\bar{r}} = -\gamma \frac{Mm}{(\bar{R} - \bar{r})^2} \frac{\bar{r} - \bar{R}}{|\bar{r} - \bar{R}|} \quad .$$

In these differential equations, $(\bar{r} - \bar{R})|\bar{r} - \bar{R}|^{-1}$ makes up a unity vector directed from the body M to the body m. If written in a scalar fashion, as was naturally done by Newton, they will constitute a system of six second-order equations, i.e. a twelve-order system.

Now, let us come to computations with use of this system. They are extremely easier because of the vector notation (27.1). First, let us add these equations together and find that

$$M\ddot{\overline{R}} + m\ddot{\overline{r}} = 0$$

or

$$\frac{d}{dt} \frac{M\overline{R} + m\overline{r}}{M + m} = \overline{V}_c = const \tag{27.2}$$

It implies that the gravitational centre of the two bodies is moving uniformly and along a straight line at the constant velocity $\overline{v}_c$.

Let us write the equations (27.1) as

$$\ddot{\overline{R}} = \gamma \frac{m(\overline{r} - \overline{R})}{\left|\overline{r} - \overline{R}\right|^3}, \quad \ddot{\overline{r}} = -\gamma \frac{M(\overline{r} - \overline{R})}{\left|\overline{r} - \overline{R}\right|^3}$$

and, subtracting one from another, find that

$$\ddot{\overline{\rho}} = -\gamma(M + m)\frac{\overline{\rho}}{\left|\overline{\rho}\right|^3}, \tag{27.3}$$

where $\overline{\rho} = \overline{r} - \overline{R}$ is a radius-vector of the mass m (of the planet) with respect to the mass M (of the Sun), i.e. these are the differential equations of their relative movement. How the gravitational centre moves is already known by us.

Through vector-multiplying the both sides of the equation (27.3) by $\overline{\rho}$ we shall come to

$$\ddot{\overline{\rho}} \times \rho = 0$$

or to

$$\frac{d}{dt}\left(\dot{\overline{\rho}} \times \overline{\rho}\right) = 0$$

or to

$$\dot{\overline{\rho}} \times \overline{\rho} = \overline{g} = const. \tag{27.4}$$

Let us clear out the sense of the relation (27.4) obtained. From it, it follows that both masses, M and m, lie all the time within the plane passing through a gravitational centre perpendicular to the vector $\overline{g}$. Indeed, the radius-vectors $\overline{\rho}_m$ and $\overline{\rho}_M$ of the masses m and M (of the planet and the Sun) with respect to their gravitational centre will be respectively equal to

$$\overline{\rho}_m = \overline{r} - \frac{M\overline{R} + m\overline{r}}{M + m} = -\frac{m}{M + m}\overline{\rho}$$

$$\overline{\rho}_M = \overline{R} - \frac{M\overline{R} + m\overline{r}}{M + m} = \frac{M}{M + m}\overline{\rho} \quad ,$$

where the vector $\overline{\rho}$, according to (27.4), will be always perpendicular to the vector $\overline{g}$.

Further, let us write (27.4) in a scalar form choosing the origin in the beginning of the vector $\overline{\rho}$, i.e. in the centre of the Sun, and prolong the axis Oz along the vector $\overline{g}$ of the length g.

Thus, we shall obtain

$$\dot{\overline{\rho}} \times \overline{\rho} = \begin{vmatrix} \overline{i} & \overline{j} & \overline{k} \\ \dot{x} & \dot{y} & \dot{z} \\ x & y & z \end{vmatrix} = g\overline{k} \quad , \tag{27.5}$$

where $\overline{i}$, $\overline{j}$ and $\overline{k}$ are the basis vectors of the axes Ox, Oy and Oz respectively, and x, y and z are the projections of the vector $\overline{\rho}$ onto these basis vectors. Besides, the vectors $\overline{\rho}$ and $\dot{\overline{\rho}}$ will be orthogonal to the basis vector $\overline{k}$, according to its choice, and hence we shall have $z = \dot{z} = 0$. Therefore, from (27.5) there will follow only the below single ratio

$$\dot{x}y - \dot{y}x = g \quad . \tag{27.6}$$

Now, let us pass to polar coordinates

$$x = \rho\cos\varphi, \quad y = \rho\sin\varphi$$

and reduce (27.6) to the shape

$$\rho^2\dot{\varphi} = -g \quad . \tag{27.7}$$

It is not difficult to notice here that the relation (27.7) obtained expresses the second law of Kepler. Indeed, the increment of the square $d\sigma$, being circumscribed by the vector $\overline{\rho}$, will be equal to

$$g\sigma = \frac{1}{2}\rho^2 d\varphi$$

and, therefore,

$$d\sigma = \frac{1}{2}\rho^2 \dot\varphi \, dt = -\frac{g}{2}dt$$

or

$$\frac{d\sigma}{dt} = -\frac{g}{2} = const \quad.$$

This was just required to prove (Fig. 27.2).
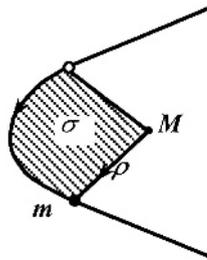
Fig. 27.2. The illustration to the second Kepler law: σ is the square circumscribed by the vector $\overline{\rho}$ of the mass m travelling around the Sun M .

We are passing now to the search for the planet's orbit with respect to the Sun. For this, we have already the equation (27.7) that will be written in the form

$$\frac{d\varphi}{dt} = -\frac{g}{\rho^2} \quad . \tag{27.8}$$

The second equation for $d\rho / dt$ can be also obtained from the vector equation (27.3), through scalar-multiplying it by the velocity vector $\dot{\overline{\rho}}$

$$\ddot{\overline{\rho}}\,\dot{\overline{\rho}} = -\gamma(M + m)\frac{\overline{\rho}\,\dot{\overline{\rho}}}{\rho^3}$$

or

$$\frac{d}{dt}\left(\frac{\dot{\overline{\rho}}^2}{2}\right) = -\gamma(M + m)\frac{1}{\rho^3}\frac{d}{dt}\frac{\overline{\rho}^2}{2} = -\gamma(M + m)\frac{1}{\rho^3}\frac{d}{dt}\left(\frac{\rho^2}{2}\right) =$$

$$= -\gamma(M + m)\frac{\dot{\rho}}{\rho^2} = \gamma(M + m)\frac{d}{dt}\left(\frac{1}{\rho}\right) \ .$$

From here, through integration, we shall find that

$$\frac{\dot{\overline{\rho}}^2}{2} - \frac{\gamma(M + m)}{\rho} = h = const \ . \tag{27.9}$$

Let us express $\dot{\overline{\rho}}^2$ in (27.9) through the polar coordinates ρ and φ. We shall have

$$\dot{\overline{\rho}}^2 = \left(\dot{x}^2 + \dot{y}^2\right) = \left(\dot{\rho}\cos\varphi - \rho\dot{\varphi}\sin\varphi\right)^2 + \left(\dot{\rho}\sin\varphi + \rho\dot{\varphi}\cos\varphi\right)^2 =$$

$$= \dot{\rho}^2 + \rho^2\dot{\varphi}^2 \ .$$

Upon substitution into (27.9), we shall obtain $\dot{\rho}^2 + \rho^2\dot{\varphi}^2 - \frac{2\gamma(M + m)}{\rho} = 2h$

or, by expressing $\dot{\varphi}$ through ρ in accordance with (27.8), we shall obtain

$$\dot{\rho}^2 + \frac{g^2}{\rho^2} - \frac{2\gamma(M + m)}{\rho} = 2h \ .$$

Solving the last equation with respect to $\dot{\rho}$ we shall find that

$$\frac{d\rho}{dt} = \sqrt{2h + \frac{2\gamma(M + m)}{\rho} - \frac{g^2}{\rho^2}} \tag{27.10}$$

The equation (27.10), together with the earlier derived equation (27.8), will be differential equations of the relative motion of the planet in the plane OXY. If the time t is excluded from them through termwise dividing the equation (27.10) by the equation (27.8), then we shall arrive at the below single equation

for a planet orbit

$$\frac{d\rho}{d\varphi} = -\frac{\rho^2}{g}\sqrt{2h + \frac{2\gamma(M+m)}{\rho} - \frac{g^2}{\rho^2}} .$$

To integrate this first-order differential equation and find an orbit equation will be possible in the following way. Let us substitute the variables introducing the new variable $u = 1/\rho$; this will yield the equation

$$\frac{du}{d\varphi} = \frac{1}{g}\sqrt{2h - 2\gamma(M+m)u - g^2 u^2} = \frac{1}{g}\sqrt{f} .$$

Then, through differentiating it in φ, we shall increase its order and get

$$\frac{d^2 u}{d\varphi^2} = -\frac{1}{g}\frac{1}{2}\frac{1}{\sqrt{f}}\frac{df}{du}\frac{du}{d\varphi} = -\frac{1}{2g}\frac{1}{\sqrt{f}}\frac{df}{du}\frac{1}{g}\sqrt{f} =$$

$$= -\frac{1}{2g^2}\frac{df}{du} = -\frac{1}{2g^2}\left(-2\gamma(M+m) - 2g^2 u\right)$$

or

$$\frac{d^2 u}{d\varphi^2} + u = \frac{\mu}{g^2} \qquad (\mu = \gamma(M+m)) . \tag{27.11}$$

We have thus arrived at the well known nonhomogeneous differential equation for a harmonic oscillator whose general solution is well familiar for us and is of the form

$$u = \frac{\mu}{g^2} + c_1 \cos\varphi + c_2 \sin\varphi = \frac{\mu}{g^2}\{1 + e\cos(\varphi - \varphi_0)\} ,$$

where e and φ_0 are the integration constants. The constant φ_0 carries an idea of the origin for the variable φ on the orbit and may be chosen equal to 0 or π in order to make the constant e nonnegative. Upon it, the planet orbit equation will take the form

$$\rho = \frac{g^2}{\mu}\frac{1}{1 + e\cos\varphi}, \quad e \geq 0 . \tag{27.12}$$

For $e = 0$, this orbit will be circular, for $0 < e < 1$ ellipsoidal, for $e = 1$ parabolic, and for $e > 1$ it will be hyperbolic (Fig. 27.3).
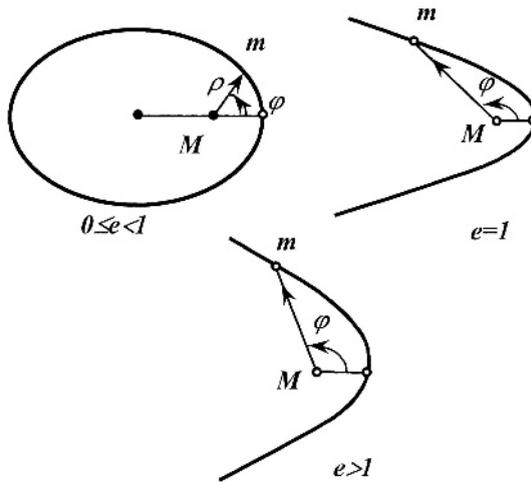
Fig. 27.3. The elliptical, parabolic and hyperbolic orbits of the mass motion with respect to the inertial system of the origin in the mass centre.

The focus lies in the point $\rho = 0$. The orbit point, the nearest to the focus, will be the point

$$\varphi = 0 \ , \qquad \rho = \frac{g^2}{M(1 + e)} \ .$$

The parameter e has a sense of eccentricity[1].

_____

[1] This follows immediately from the definitions of the conic cross sections listed. For example, for an ellipse, in accordance with figure 27.4 and its definition, we shall obtain

$$(2a - \rho)^2 = \rho^2 + 4c^2 + 2\rho c \, cos \, \varphi \ .$$

From this we shall get $\rho = \dfrac{a^2 - c^2}{a} \ \dfrac{1}{1 + c/a \, cos \, \varphi} \ .$

With use of the notation $e = c/a$, $g^2 / \mu = a^2 - c^2 / a$, it will coincide with (27.12).
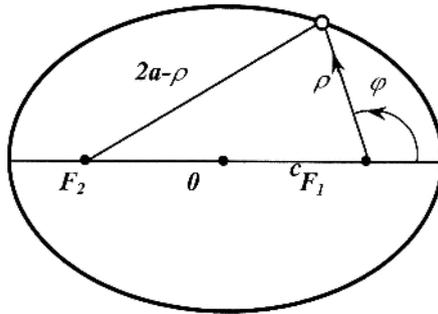
Fig. 27.4. The ellipse with the foci F_1 and F_2.

Between the magnitudes a and b of the major and minor semiaxes of the ellipse and between the distance $c = \sqrt{a^2 - b^2}$ from the centre of the ellipse to its foci F_1 and F_2 and the parameters of the formula (27.12) there exist the below relations

$$e = \frac{c}{a}, \quad a - c = \frac{g^2}{\mu(1+e)}, \quad a + e = \frac{g^2}{\mu(1-e)} \quad . \tag{27.13}$$

We will now pass to deriving the Kepler third law from the ratios found. First, integrating the formula of the planet sector velocity

$$\frac{d\sigma}{dt} = -\frac{g}{2}$$

we shall find that

$$\int_0^T \frac{d\sigma}{dt} dt = -\frac{g}{2} T \quad .$$

Making use of the ellipse square being equal to πab we shall find that

$$\pi ab = -\frac{g}{2} T$$

or

$$T = -\frac{2\pi ab}{g} \ . \tag{27.14}$$

The relations (27.14), (27.13) and $c = \sqrt{a^2 - b^2}$ are 5 equations for the magnitudes a, b, c, e, g, T. This will allow, in particular, to find the relation between a and T. This relation may be calculated in the following way

$$T = -\frac{2\pi ab}{g} = -\frac{2\pi a}{g}\sqrt{a^2 - c^2} = -\frac{2\pi a^2}{g}\sqrt{1 - e^2} \ . \tag{27.15}$$

Further, from the relations (27.14) it follows that

$$2a = \frac{g^2}{\mu}\frac{2}{1 - e}$$

or

$$\sqrt{1 - e^2} = \frac{g}{\sqrt{\mu a}} \ .$$

Then, by putting the found expression into (27.15), we find that

$$T = -\frac{2\pi a^2}{g}\frac{g}{\mu a} = \frac{2\pi a^{3/2}}{\sqrt{\mu}}$$

or

$$\frac{T^2}{a^3} = \frac{4\pi^2}{\mu} \ . \tag{27.16}$$

This will be the very relation to express the third law of Kepler. The magnitude

$$\mu = \gamma(M + m) = \gamma M\left(1 + \frac{m}{M}\right)$$

is somewhat different for different planets, but these difference could not be distinguished in the time of Kepler since for the solar system the ratio $\frac{m}{M}$ is of a very small value.

Now, let us apply the obtained results concerning the motions of the two mutually gravitationally attracted bodies to the movement of the body pushed with the velocity v from the mountain on the Earth in the horizontal direction, with air

resistance being not accounted. First, we would like to note here that the ball-like Earth, with its spherical homogeneity, will attract the pushed body in the same way as the material point possessing the mass of the Earth and lying in its centre. This nontrivial fact was revealed by Newton. He also stated and solved the above formulated problem on the travelling trajectory of the body launched from the Earth. He found that at some sufficient velocity this body will become a satellite of the Earth, like the Moon.

According to the above, the body will travel along the below curve described by the equation (27.12)

$$\rho = \frac{g^2}{\mu} \frac{1}{1 + e\cos\varphi} \quad , \tag{27.17}$$

and in one of its foci there will lie the centre of the Earth. At $\varphi = 0$, we shall obtain $\rho = R_E$, where R_E is the radius of the Earth; or, to be more exact, the distance from the centre of the Earth to the place from which the body has been launched. Thus, we shall have

$$R_E = \frac{g^2}{\mu} \frac{1}{1 + e} \quad . \tag{27.18}$$

One more thing should be accounted here. It is that the body is launched at the velocity υ. From the ratio (27.8), it will follow that

$$\dot{\varphi} = -\frac{g}{R_E^2} \quad ,$$

where

$$\dot{\varphi} = \frac{\upsilon}{R_E} \quad .$$

Thus, the equation (27.18) may be written as

$$R_E = \frac{(\upsilon R_E)^2}{\mu} \frac{1}{1 + e} \quad .$$

From it, we shall get

$$1 + e = \frac{\upsilon^2 R_E}{\mu} \quad ,$$

or, neglecting the mass of the launched body, we shall obtain

$$e = \frac{\upsilon R_E}{\gamma M_E} - 1 \ . \tag{27.19}$$

The relation (27.19) makes it possible to find through the velocity v the eccentricity e of the orbit of the thrown body. This, together with (27.18) determinating the ratio $\dfrac{g^2}{\mu}$, will help us to find the parameters of the orbit equation (27.17).

With use of the formula (27.19), we can find the first and the second cosmic velocities, υ_1 and υ_2. The first velocity is a minimal one; at this velocity the thrown body will not fall down to the Earth and will become its satellite. The second velocity is the minimal one at which the thrown body will cease to be a satellite and will leave the Earth. The first case will take place at $e = 0$, with the orbit being circular; he second at $e = 1$, with the orbit becoming parabolic. From (27.9), we shall find that

$$\upsilon_1 = \sqrt{\frac{\gamma M_E}{R_E}} = \sqrt{\frac{\gamma M_E}{R_E} R_E} =$$

$$= \sqrt{9,8 \times 6,10^{\,6}} \approx 8 \ km \ / \sec \tag{27.20}$$

$$\upsilon_2 = \sqrt{\frac{2 M_E}{R_E}} \approx 11 \quad km / sec \ .$$

These velocities are enormous; they are sufficient to provide a flight round our planet for less than two hours. Man has achieved these velocities and performed the flight only in the 20-th century, with the help of the missile we were speaking about earlier. Such velocities are also obtainable by gas molecules.

On the Earth, these velocities are achieved by few molecules only. On the Moon, where the second cosmic velocity is much less, these molecules are so numerous that the Moon has no atmosphere. As for the Sun, the second cosmic velocity there is much greater than on the Earth, but the temperature on the Sun is so extremely high that a stream of charged particles is always emitted by the Sun.

In addition to these vivid remarks on the derived formulae (27.20), let us also note here about their two corollaries. The first covers the situation when a "star" becomes a black hole and nothing, even light, can escape from this black hole. The black hole does not shine and is recognized only through its attracting effect exerted upon its nearest really visible stars. Therefore, the answer to the question raised is simple: a "star" will be a black hole if its second cosmic velocity exceeds

the light velocity – faster than light c nothing can move. Thus, the condition for existence of a black hole will assume the form

$$\sqrt{\frac{2\gamma M}{R}} > c \ ,$$ (27.21)

where M is its mass and R its radius. In order to realize what this condition may be, let us notice here that the Sun would have been a black hole if it had been less than 3 km in diameter, i.e. if it could have the gigantic density of millions of tons per a cubic centimeter. Such a gigantic density is nevertheless possible, and so, black holes exist. Within the vicinity of a black hole, the geometry of the space is non-Euclidean and in this place a watch will run slower (according to the Einstein gravitational theory, the gravitational effect, scarcely detectable within the gravitational field of the Sun and planets, will become substantial in the gigantic gravitational fields of black holes).

The second question is about whether celestial bodies are coming together under universal gravitation or, vice versa, are scattering. This question is not so simple. Instead, I would say it is so complicated that the answer to it is not so evident and clear-cut.

Let us start from afar. In his time, Newton already knew that the force f of attracting a unity mass at the distance r from the centre of the a massive homogeneous ball of the radius R and of the mass M is determined by the formula

$$f = \begin{cases} \dfrac{\gamma M}{r^2} & for \quad r \geq R \\[2ex] \dfrac{\gamma M}{r^2}\dfrac{r^3}{R^3} & for \quad r \leq R \end{cases} \ ,$$ (27.22)

i.e. for $r \geq R$ the ball may be replaced with the point mass M in its centre; and for $r < R$ the external part of the ball will exert no attracting action upon the unit mass (since all its attractions, directed into different sides, will be balanced).

The above conclusions drawn from the formula (27.22) are generalized for the nonhomogeneous ball whose density ρ is dependent upon the distance from the centre only. Now, let us apply these considerations to the universe. Within this approach and the measuring accuracy, its density will be constant. Under the density there should be certainly understood the averaged density within a sufficiently large ball, though being not too large. The size of the ball is very much less than the size of the observable universe.

It may be also thought that somewhere farther this density will decrease and will even vanish and that our solar system lies somewhere in the vicinity of the universe centre, i.e. the centre with respect to this very distant decrease. This is revealed by an absence of attraction of our solar system by the universe system, with the accuracy being equal to the local fluctuations of density.

Further, to answer whether the universe is contracting or scattering one needs to have some understanding of the velocities of the masses which the universe is composed of. These velocities are described by the Habble law under which chaotic velocities of stars are augmented by a systematic radial component. Stars are radially scattering at the below velocity

$$V = HR \ \text{km/sec} \ , \tag{27.23}$$

where H is the Habble constant approximately equal to 75 km/sec; R is a distance to the star measured in Mic (a distance unit equal to the distance covered by the running light for a three-year period of time).

We will now choose a large ball of the radius R with the centre where we are standing and will ask ourselves whether its stars or other bodies are able to escape from the ball. They can leave it, provided that the velocities on the surface of this ball exceed the second cosmic velocity of the chosen ball, i.e. if

$$V = HR > \sqrt{\frac{2\gamma M}{R}} \ ; \tag{27.24}$$

they will not be able to leave the ball if the opposite inequality holds. By expressing the mass M of the chosen ball through the averaged density ρ, we can then find that

$$M = \frac{4}{3}\pi\rho R^3 \ .$$

By inserting the value of M into (27.24), we shall find that the critical density ρ_{cr}, separating the scattering and contracting cases, will not depend upon R and will be equal to

$$p_{cr} = \frac{3H}{8\pi\gamma} \approx 10^{-29} \ \text{gr/cm}^3 \ . \tag{27.25}$$

Thus, if the mass density in the universe is more than ρ_{cr}, then the universe will be contracting; and if less , then it will be scattering . The contemporary calculations of ρ_{cr} for the universe demand an account of the interstellar dark substance. The calculations reveal the magnitude equal to ρ_{cr}. Therefore, the question retains still open.

We have rather thoroughly performed all the calculations which had been successfully done by Newton. It is clear that in his time these calculations needed titanic effort and encountered numerous and sudden difficulties. You know that in the time of Newton neither differential nor integral calculus existed. There were no differential equations either.

Further, some trials were made to extend Newton's results upon greater number of mutually attracting bodies, upon the so-called n -body problem. Though, nobody managed to solve this problem. After a long period of time it became clear that the solution to this problem is impossible, since the n -body problem for $n > 2$ is nonintegrable. Then, there have come into being some complicated theories of approximated solutions and the methods of the so-called perturbation theory. This theory has greatly advanced and played a great role not only in celestial mechanics. Through the calculations based on the perturbation theory it has become possible to foretell very much ahead not only the positions of the planets but also to open the new planet, Neptune. Though, nobody could manage to integrate even the simplified three-body problem when one of the bodies has a vanishingly small mass and its attraction of the two other bodies may be neglected.

The perturbation theory is based upon constructing the sequential approximations, for example, in the form of a power series in some small parameter. The first approximation takes into account a basic interaction between each planet and the Sun. Here, all planets are assumed to be travelling along the Kepler unchangeable orbits. The planes of these orbits lie close to the plane of the ecliptic. The position of each orbit with respect to the ecliptic plane is described by somewhat parameters. In the first approximation, they are constant. Then, there is constructed the second approximation which takes into account the effect of all or of a portion of the remaining planets and also the fact of their travel along the unchangeable elliptic orbits. This procedure is then repeated, with due account of the corrections in the planet orbits, etc. As a result of these procedures, a change in orbit parameters and the motion of planets are represented in the form of a series in the small parameter, which is a quotient of the masses of planets to the mass of the Sun. To derive these series is too complicated. On the one hand, this technique was thoroughly developed; on the other hand, numerous difficulties were exposed. Among them, the most persistent and insurmountable problem is the problem of small denominators.

The qualitative theory of differential equations by J.H. Poincaré was a response to these difficulties. Initially, this theory was presented by him in his famous writings "New methods of celestial mechanics" and "About the curves defined by differential equations". But it happened so that this new wonderful theory turned out to be also incapable to solve not only the n -body problem but even the simplified 3-body problem. Nowadays, the trajectories of planets and of man-made satellites are automatically calculated by computational methods and high-speed computers. Computations are performed fast and with great accuracy. Though, the computers' abilities are not endless. Until now, there exists no answer as to what the Sun will be in milliards of years, or whether it is stable or may undergo substantial changes. Immediate computations here are of no help, for whatever insignificant errors may arise for milliards of years they will, nevertheless, accumulate to become very large and unpredictable.

This is caused by the fact that the differential equations for the n -body problem are the conservative Hamiltonians not admitting the exponentially stable motions. Any of their solutions will be either exponentially unstable (then the error in

approximate calculations will inevitably grow exponentially) or neutral (then the computational error will also grow, though greatly slower). Moreover, with in-definitely small and independent random errors, the solar system may arrive, upon some sufficient time, at any state reachable, if the basic conservation laws – of energy, momentum and angular momentum – hold.

Meanwhile, the solar system demonstrates surprising structural specificities and regularities. Especially amazing were the surprising synchronisms between the periods of planet revolutions, orbital and own , a thin structure of the Saturn rings, etc. For a long time, neither the theory nor computations could explain these things. They really failed to do it, since that time there existed a supremacy of the conservative Hamiltonian model of the solar system.

These regularities were born not by the conservative forces, not only by the Newton attracting forces but also by the dissipative forces, though being strikingly small. The solar system has been existing for milliards of years and these are the seemingly negligible forces that have namely given birth to this surprising struc-ture and regularities in the movement of the solar system.

These structural specificities in the long-evolving solar system cannot be either investigated without a greater account of gravitational interactions between its bodies. If the homogeneous spherical bodies are interacting as the material points placed in their centres (this was known yet by Newton), then for the non-spherical and non-homogeneous bodies this will be not so. Not only the nonsphericity of the solar system bodies should be taken into account, but also their deformability, which gives rise to a tidal friction and the forces caused by internal deformations. This account brings about a substantially qualitative change in the phase portrait and the modelled motion of the solar system, since the conservative Hamiltonian system of differential equations for its movement will turn to the dissipative sys-tem admitting the exponentially stable motions. These motions may be already found computationally; whereas the exponentially stable steady-state motions may be found theoretically, through the contemporary theory of the systems of fast-revolving phases.

A presence of a small dissipation has simplified the phase portrait of the solar system; in it, there already exist no previous, possibly infinite, number of periodic motions. The portrait has become simpler and more accessible for theoretical and computational studies. This was provided through taking into account the seem-ingly absolutely negligible additions not found in the first approximations.

28 Distributed dynamical models in mechanics and physics

The notion of a destributed system. The distributed models in classic mathematical physics and quantum mechanics. The Euler equation, the Navier-Stoke equation, the Maxwell equations, the Schrödinger equation.

Until now we considered the mathematical models, i.e. the dynamical systems whose state was determined by a finite number of scalars. The states of these systems are guided by an ordinary differential equation

$$\frac{dx}{dt} = X(x) \quad , \tag{28.1}$$

where x is a finite dimension vector of the state; $X(x)$ is a vector-function of x.

Now, we are coming to another type of dynamical systems whose state is determined by the functions of some variables; the variations of the state in time, i.e. a change of the state describing functions, are described by the equation in partial derivatives of the type

$$\frac{\partial u}{\partial t} = F\left(u, \frac{\partial u}{\partial x_1}, ..., \frac{\partial u}{\partial x_n}, ... \right) \quad , \tag{28.2}$$

where $u(x_1, ..., x_n; t)$ is a function or a vector-function describing the state at the instant t; the right-hand side of the equation (28.2) depends not only upon the function u, but also upon its partial derivatives in the variables $x_1, ..., x_n$ up to some order.

Here, the earlier given general definition for a dynamical system retains fully. Subjected to changes was only the way of describing its state and accordingly the form of its operator. As earlier, a change of the state in time, i.e. the derivative $\frac{\partial u}{\partial t}$, is determined through the state, i.e. through the function $u(x_1, ..., x_n; t)$ and its partial derivatives in $x_1, ..., x_n$ determined by it. As earlier, this makes it possible to approximately find the state at the next instant $t + \Delta t$ through the state of the previous moment of the time t

$$u(x_1,...,x_n;t+\Delta t) \approx u(x_1,...,x_n;t) + \widetilde{F}(x_1,...,x_n)\Delta t$$

and to hope that it may be done infinitely exactly through decreasing Δt. If a smooth solution is supposed to be existent, then an error for one step will be of the order $(\Delta t)^2$; for the finite interval T it will be $(\Delta t)^2 \dfrac{T}{\Delta t} = T\Delta t$, and it will tend to zero as Δt tends to zero.

Describing a state through functions is in many cases caused by a spatial distribution of the real system under question. For example, a thermal state of the environment is described through the temperature $T(x, y, z, t)$ dependent upon the spatial variables x, y, z and varying with the time t. As for the pressure in fluid (or in gas), it is determined in each of its point (x, y, z) by its magnitude $p(x, y, z, t)$. Analogously, the descriptions $\rho(x, y, z, t)$ are constructed for the density of the substance or for the spatial distribution of an electrical charge.

All these examples touch the scalar fields and show how these fields are described by one function of spatial variables and time.

The fields may be vectorial. Such is the field of the velocities $\overline{v}(x, y, z, t)$ of a fluid or gas flow, or the field of the displacements $\overline{r}(x, y, z, t)$ of the points in an elastodeformed body, or the electrical and magnetic fields described by the vectors of electrical and magnetic fields.

The above described systems, with the state represented by a finite number of scalar variables, are called discrete systems. The systems we are speaking about right now and whose state is described by functions are called distributed.

A distributed system may be interpreted as a discrete system with an infinite number of parameters. Indeed, the function, say, of three variables $f(x, y, z)$ may be expanded into a series in some functions $\varphi_1(x, y, z)$, $\varphi_2(x, y, z)$... so that we obtain

$$f(x, y, z) = \sum_{j=1}^{\infty} c_j \varphi_j(x, y, z) \ .$$

Thus, the function $f(x, y, z)$ is determined by an infinite number of the scalar variables $c_1, c_2,...$.

Through such a substitution, the equation (28.2), with the functions $\varphi_1, \varphi_2,...$ being independent, will yield an infinite system of the ordinary differential equations

$$\frac{dc_j}{dt} = F(c_1, c_2,...) \quad (j = 1,2,3,...)$$

being equivalent to it.

The said is rather evident but needs to be proved. This proof is not so simple; it will be, rather, very complicated, if we try to decrease the constraints for the functions f and F. But right now this talk falls beyond our scope.

A world of the distributed systems seems much more diverse and wider than that of the discrete systems. Moreover, the discrete systems may be considered as a special finite-dimensional case of the distributed systems. This way of their interpretation is not exploited only because the existing methods for studying the equations with partial derivatives differ significantly from the methods in the theory of ordinary differential equations. These are actually different scientific disciplines; they make up different branches of mathematics being studied by different researchers.

At the same time, these and those differential equations, with use of general concepts of functional analysis, are written similarly, as in (28.1), where x is now an element of some functional space and $X(x)$ is not a vector-function of x, but the operator X of x.

Our further narration will be some kind of introduction into the science on the distributed dynamical systems. More profound acquaintance with them you will have at your courses on mathematical physics, hydrodynamics, theory of elasticity, electrodynamics, statistical physics, theory of probability, random processes and others. As far as my narration is only an introduction, most things in it will be only descriptive, not conclusive. First, it refers to the basic models of the classic mathematical physics and the great equations with partial derivatives for mechanics and physics, i.e. the equations by Euler, Navier and Stokes, and Maxwell. You will have to take them on trust, without demanding any proof and derivation. This will do no harm for your understanding these equations and applying them to specific cases. In details there will be explained only the equation for the thermal conductivity or diffusion. It will be done so to give you a possibility to see how and where from these equations with partial derivatives appear. I mean the equation

$$\frac{\partial u}{\partial t} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) , \qquad (28.3)$$

which simultaneously describe the events of thermal conductivity, diffusion and random walk. In the first case, $u(x, y, z, t)$ is the temperature in the point (x, y, z) at the instant t; in the second, it is the density of a diffusing substance; in the third case, it is the density of the probabilities, with which a randomly walking point at the instant t will occur in the point (x, y, z).

Let us start from heat conduction within some homogeneous environment. In this case, the thermal field is described by the function $T(x, t)$. It will vary because of the natural thermal flow from the more heated places to the colder

places. For many cases, this flow may be accepted to be proportional to a temperature gradient. So, we have

$$dQ = -k \frac{\partial T}{\partial x} dt \quad , \tag{28.4}$$

where dQ is the quantity of the heat flowing through the section x in the direction of the axis Ox for the time dt (Fig.28.1). The sign "minus" indicates the heat conduction occurring in the direction reverse to the temperature growth. The environmental homogeneity is expressed by the heat conductivity coefficient k being constant.

Let us take some arbitrary sector $[x_1, x_2]$. Assuming the thermal conductivity of the environment to be constant and equal to c, we shall find that at the instant t this sector will hold the quantity of heat equal to

$$\int_{x_1}^{x_2} cT(x,t) dx \quad .$$

The gain to this quantity of heat, if no emission, for the time dt will be determined by its penetration through the sections $x = x_1$ and $x = x_2$ ($x_2 > x_1$), i.e. we shall have

$$\frac{\partial}{\partial t}\left(\int_{x_1}^{x_2} cT(x,t) dx \right) dt = -k \frac{\partial T}{\partial x}\Big|_{x=x_1} dt + k \frac{\partial T}{\partial x}\Big|_{x=x_2} dt ,$$

or, through reducing by dt and differentiating in t, we shall get

$$c \int_{x_1}^{x_2} \frac{\partial T}{\partial t} dx = k \left(\frac{\partial T}{\partial x}\Big|_{x=x_2} - \frac{\partial T}{\partial x}\Big|_{x=x_1} \right) . \tag{28.5}$$

Now, we should take the limit for the unbounded convergence of the sections x_1 and x_2 to the arbitrary x. For this, let us use the mean-value theorem for the integral (28.5) and write this relation as

$$c(x_2 - x_1)\frac{\partial T}{\partial t}\Big|_{x=\xi} = k\left(\frac{\partial T}{\partial x}\Big|_{x=x_2} - \frac{\partial T}{\partial x}\Big|_{x=x_1}\right) ,$$

where ξ is some intermediate point between x_1 and x_2. We shall divide the obtained equality by $x_2 - x_1$ and take the limit as this difference will tend to zero (since the points x_1 and x_2 converge to the point x). As a result, we shall arrive at the needed equation of thermal conductivity

$$c\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial x^2},$$

which is usually written as

$$\frac{\partial T}{\partial t} = a^2\frac{\partial^2 T}{\partial x^2} \qquad (a^2 = \frac{k}{c}) . \tag{28.6}$$

The same equation describes a diffusion of one substance with the concentration $\rho(x,t)$ into another substance. It follows from the fact that during the diffusion the substance will displace from the places of greater concentration to the places of less concentration. The quantity of the displaced substance dm for the time dt will be described by the same equation that determines the heat conduction, i.e. by

$$dm = -D\frac{\partial p}{\partial x}dt , \tag{28.7}$$

where D is a diffusion coefficient. Therefore, the similar considerations and computations will lead us to the equation (28.6), in which $a^2 = D$.

The same equation will be obtained when we solve the problem about the particle randomly walking along the line Ox. Let this problem be called the problem of an absolutely drunken who knows nothing as where to go and with equal probabilities is making a step forward or backward. Let him make a step for the time τ and a length of his step be a. Also let $p(sa,n\tau)$ be the probability of his being at the point $x = sa$ (s is an integer) at the instant $t = n\tau$, upon making n steps.

The relation between the probabilities $p(sa,n\tau)$ and $p(sa,(n+1)\tau)$ is not difficult to be found. Indeed, the drunken may find himself in the point sa only upon making previously a forward step from the point $(s-1)a$ or a backward

step from the point $(s+1)a$. This or that he is making with the equal probability $1/2$. Thus, we have

$$p\big(sa,(n+1)\tau\big) = \frac{1}{2}\,p\big((s-1)a,n\,\tau\big) + \frac{1}{2}\,p\big((s+1)a,n\,\tau\big). \quad (28.8)$$

If the function $p(sa,n\,\tau)$ known and $n=0$, the equation (28.8) will allow us to sequentially find this function for any n. Therefore, if for $n=0$ the drunken (or the walking particle) finds himself at the point $x=0$, then the function $p(sa,0)$ will be known and equal to

$$p(sa,0) = \begin{cases} 0 & for \quad s \ne 0 \\ 1 & for \quad s = 0 \end{cases}.$$

Through it, there may be found any of the probabilities $p(sa,n\,\tau)$ for any s and n.

Now, let us assume that the drunken's steps are more and more quickened and their length is lessened. Let τ and a tend to zero in such a way that

$$\lim \frac{a^2}{2\tau} = D \quad (28.9)$$

Let us write (28.8) in the form

$$\frac{p\big(sa,(n+1)\tau\big) - p\big(sa,n\,\tau\big)}{\tau} =$$

$$= \frac{a^2}{2\tau}\,\frac{p\big((s+1)a,n\,\tau\big) - 2p\big(sa,n\,\tau\big) + p\big(sa,n\,\tau\big)}{a^2}\;.$$

In it, passing to the limit for $\tau \to 0$, $a \to 0$ and taking into account the condition (28.9) we shall come to the below partial derivative equation with respect to the function $p(x,t)$

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2}\;. \quad (28.10)$$

The form of this equation will be similar to the form of the thermal conductivity equation (28.6) $(a^2 = D)$.

The equation (28.10) should not be thought, as this would be natural, to hold a limit for the discrete function $p(sa,n\,\tau)$. This limit, as not difficult to see, will be equal to zero, since all the probabilities $p(sa,n\,\tau)$, as $a \to 0$, will tend to zero. Certainly, zero satisfies the equation (28.10), but us does not. For the

equation (28.10) we shall need a nonzero solution satisfying the normalization requirement

$$\int\limits_{-\infty}^{+\infty} p(x,t)dx = 1 \quad ,$$

which is interpreted for the discrete functions $p(sa,n\tau)$ as

$$\sum_{s=-\infty}^{+\infty} p(sa,n\tau) = 1 \quad .$$

It may be written as

$$\sum \frac{p(sa,n\tau)}{a} a = 1 \quad .$$

This form brings us to the idea that $p(x,t)$ should be considered as a limit for the discrete function $\dfrac{p(sa,n\tau)}{a}$.

The above derivations are not difficult to be repeated under the wider assumptions concerning the two- or three-dimensional nature of the space. Here, in all three cases we shall arrive at the equation of the form

$$\frac{\partial T}{\partial t} = a^2 \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) \tag{28.11}$$

or, respectively, of the form

$$\frac{\partial T}{\partial t} = a^2 \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) \quad . \tag{28.12}$$

The differential equation in partial derivatives (28.12) (and its special cases (28.6) and (28.11)) constitutes one of the fundamental equations of the classic mathematical physics. Its second fundamental equation is the wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \tag{28.13}$$

that describes the oscillations of the elastic environment. For the audio-oscillations of gas, the function $u(x, y, z, t)$ implies a pressure in the point (x, y, z) at the instant t. In the one-dimensional case, when

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \quad,$$

the function $u(x, t)$ may imply a deviation of the strained string from its equilibrium state. The same equation describes the longitudinal elastic oscillations of a homogeneous rod. In this case, $u(x, t)$ is an elastic longitudinal displacement of the rod material in the cross-section x at the instant t. A two-dimensional variant of the equation (28.13) can describe the transversal oscillations of a membrane.

The thermal conductivity equation and the wave equation of the classic mathematical physics are linear. It means that the superposition principle may be applied to them, i.e. the same principle used for a linear oscillator or any other linear system. In its turn, the derived solutions may be used to construct new solutions or even any particular solution needed. As particular solutions there will be the so-called fundamental solutions for the thermal conductivity equation and harmonic running waves. With both of them you will become acquainted below.

The dynamical systems described by the differential equations in partial derivatives substantially differ from those described by the ordinary differential equations in that that the equilibrium states of the discrete dynamical systems are found in a comparatively simple way, just through nullifying their right-hand sides, i.e. from some nonlinear equations; whereas for the distributed dynamical systems the equations, describing an equilibrium state, turn out to be differential. Thus, by assuming

$$T(x, y, z, t) = T^*(x, y, z) \text{ and } u(x, y, z, t) = u^*(x, y, z)$$

in (28.12) or (28.13) to be equilibrium states, fixed with time, we shall conclude that they satisfy the equations in partial derivatives

$$\frac{\partial^2 T^*}{\partial x^2} + \frac{\partial^2 T^*}{\partial y^2} + \frac{\partial^2 T^*}{\partial z^2} = 0 \quad, \tag{28.14}$$

and, respectively,

$$\frac{\partial^2 u^*}{\partial x^2} + \frac{\partial^2 u^*}{\partial y^2} + \frac{\partial^2 u^*}{\partial z^2} = 0 \quad. \tag{28.15}$$

To solve them is so difficult that they also fall into the list of the equations of classic mathematical physics.

Now, we are going to narrate other mathematical models for the distributed dynamical systems.

A flow of viscous fluid is described by the Navier-Stokes nonlinear differential equations in partial derivatives

$$\frac{\partial \upsilon}{\partial t} + (\upsilon \nabla)\upsilon = -\frac{1}{\rho}\nabla p + \nu \nabla \upsilon \ldots\ldots\ldots\ldots(\upsilon = (\upsilon_x, \upsilon_y, \upsilon_z)) \ . \quad (28.16)$$

Its particular case for $\nu = 0$ is the Euler equation for the flow of ideal (nonviscous) fluid. In this equation, $\upsilon(x, y, z, t)$ is a vector of the fluid velocity in the point x, y, z at the instant t; $p(x, y, z, t)$ is a pressure; ρ is a density; ν is a parameter, the so-called kinematic viscosity; ∇ and Δ are the widely-used symbols for the gradient operator and the Laplace operator. Namely, we have

$$\nabla = i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z} \ ,$$

where i, j, k are the basis vectors in the rectangular system of coordinates; the Laplace operator is of the form

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \ .$$

In case of the incompressible fluid, the Navier-Stokes equation or the Euler equation should be supplemented with the incompressibility condition. This condition will imply the velocities of the fluid particles to be such as to have their volume unchangeable. This condition is also written with use of the equation in partial derivatives, namely, as

$$div\ \upsilon = \frac{\partial \upsilon_x}{\partial x} + \frac{\partial \upsilon_y}{\partial y} + \frac{\partial \upsilon_z}{\partial z} = 0 \ ; \quad (28.17)$$

for the compressible fluid, as

$$div\ \ p\upsilon = 0 \ , \quad (28.18)$$

expressing, thus, a conservation of mass but not volume.

Here, I would like you to become acquainted also with the great equations by Maxwell. This is a rather complicated system of equations in partial derivatives. It includes various functions: the vector-functions $E(x, y, z, t)$, $D(x, y, z, t)$, $H(x, y, z, t)$, $B(x, y, z, t)$ describing electrical and magnetic fields (the fields of the vector of electrical intensity and induction, magnetic intensity and induction), the scalar function of the distributed electrical charge $\rho(x, y, z, t)$ and the vector field of electrical currents $j(x, y, z, t)$. The Maxwell equations

interconnect these functions. In the vector form and through the well-known notation they may be written in the form of the following basic equations[1]

$$\frac{\partial D}{\partial t} = -4\pi\, j + c\; rotH \quad, \qquad \frac{\partial B}{\partial t} = -c\; rotE \quad.$$ (28.19)

To these equations there should be also added the other two equations

$$div\, B = 0, \vdots \quad div\, D = -4\pi\,\rho \quad ;$$ (28.20)

for the simplest, specific case, there should be added three more algebraic equations

$$D = eE \ , \quad B = \mu H \ , \quad j = \sigma E \quad.$$ (28.21)

This system of equations is very complicated. This is rather clear, since all the phenomena of electrostatics, magnetostatics, electrodynamics, electromagnetic emission and waves follow from this system. With ε and μ being independent of the fields E and H, these equations will be linear; thus, they will admit a use of the superposition principle. In this sense, the equations of hydrodynamics, being simpler in their shape, are far worse, for they are nonlinear and the superposition principle is not admitted for them.

You have, thus, seen some wonderful – and even great – distributed mathematical models from mechanics and physics. Perhaps, you are not aware of them and do not understand them. You should not feel embarrassed, since they are nothing else but mathematical descriptions, mathematical models. Right here you are only required to accept them in this capacity, you should not be confused either by their complexity or your misunderstanding about how to solve and exploit them.

In all the models discussed, a need to describe a state via functions arose because our talk was touching these or those physical spatial fields: the field of

---

[1] $rot\, A$ is a deferential operation over the vector $A(A_x, A_y, A_z)$ determined by the determinant of the form

$$rot\, A = \begin{vmatrix} i & j & k \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} = i\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)$$

$$+ j\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) + k\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)$$

$div\, A$ is defined in the equation (28.17).

temperatures, density, velocity, displacements, pressure, electrical and magnetic intensities and induction, and probability density. However, there may also arise completely different reasons to describe a state through functions. You will encounter this case later when we shall consider the mathematical models of microworld. A state of the dynamical systems for microworld is described by the function of the coordinates and spines of all the particles held in the system. The equation to be satisfied by this function is the below famous equation by Schrödinger for the state function ψ

$$ih\frac{\partial \psi}{\partial t} = H\psi \quad .$$
(28.22)

This equation, though written in a very brief form, depicts the entire world. Indeed, this equation is of a very complicated nature and is solved with a great difficulty: the function ψ is complex and may depend upon a very large number of variables; by a sign of the Hamilton operator H there is denoted the complicated differential operator having second-order partial derivatives. All these things should not surprise you because this is the very equation which in the final end describes all happenings in the living and inorganic worlds.

In addition to the above said concerning the distributed models of dynamical systems (the classic equations of mathematical physics, equations by Euler and Navier and Stokes, by Maxwell and Schrödinger), it should be also emphasized here that they themselves are not able to fully describe the changes of the functions they hold. These equations also need some supplementary conditions. They are numerous and determined by the specificity of the objects under study. For example, for the equations of mathematical physics with an unbounded space of variables x, y, z, these conditions are the boundedness in infinity. Also, with the space of the variables x, y, z being bounded, there will be needed some conditions on the boundary (called the edge conditions). The same refers to the equations of hydrodynamics, electrodynamics and the Schrödinger equations. Only together with these supplementary conditions, the functions of the spatial variables, involved in these equations, will determine a state, i.e. given at $t = t_0$, they may be found at any $t > t_0$. With some of the specific edge conditions, you will encounter later. Right now, let us restrict our narration by a simple example about the strained string with fixed end points.

In this case, the string small oscillation equation will be the wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \quad ,$$
(28.23)

where $u(x,t)$ is a string deviation from the equilibrium strained position at the instant t. This equation is valid throughout the entire string, i.e. for $0 \le x \le l$, where $x = 0$ and $x = l$ are the coordinates of the fixed string terminals. On

these terminals, the deviation u will be absent and, hence, there will occur the edge conditions

$$u\big|_{x=0} = 0, u\big|_{x=l} = 0 \quad .\qquad (28.24)$$

The equations in partial derivatives (28.23), together with the edge conditions (28.24), will admit the single solution $u(x,t)$ for the given initial position and velocities

$$u(x,t)\big|_{t=0} = \varphi(x), \quad \frac{\partial u}{\partial t}\Big|_{t=0} = \psi(x) \quad .\qquad (28.25)$$

In this case, the state at the initial instant will be determined by the functions $\varphi(x)$ and $\psi(x)$ satisfying the edge conditions

$$\varphi(0) = \varphi(l) = 0, \quad \psi(0) = \psi(l) = 0 .$$

At any t, the state will be described by the functions of $x - u(x,t)$ and $\dfrac{\partial u(x,t)}{\partial t}$, which satisfy the edge conditions. You should not be surprised that alongside with the function $u(x,t)$ the state description will also hold the function $\dfrac{\partial u(x,t)}{\partial t}$. This is explained by the wave equation (28.13) being, when written as (28.2), of the form

$$\frac{\partial}{\partial t}(u,\upsilon) = (\upsilon, a^2 \Delta u) \quad .\qquad (28.26)$$

As a state for this equation there will the vector composed of the two functions, $u(x,t)$ and $\upsilon(x,t)$, as the functions of x for the fixed time t.

29 Fundamental solution of the thermal conductivity equation

Fuzziness of the thermal field, the diffused substance or the density of probabilities for a randomly walking particle initially concentrated in some point. The solution of the semispace warm-up problem.

Heat conduction within the unbounded homogeneous one-dimensional heat-conducting environment is described by the differential equation in partial derivatives of the form

$$\frac{\partial T}{\partial t} = a^2 \frac{\partial^2 T}{\partial x^2} \quad , \quad T = T(x,t) \quad . \tag{29.1}$$

A diffusion or a random walk of a particle is described by the equation similar in form. Let at the initial time $t_0 = 0$ we have some arbitrary thermal field $T_0(x)$ such that

$$T(x,t)\big|_{t=0} = T_0(x) \quad . \tag{29.2}$$

With the initial thermal distribution $T_0(x)$ given, the thermal conductivity equation (29.1) will determine the thermal field $T(x,t)$ for any $t > 0$.

The equation (29.1) and the initial condition (29.2) describe not only the fluctuation of the initial thermal field, but also the substance diffusion and the probability density variations for a randomly walking particle.

This problem (29.1), (29.2), being general for all the above listed phenomena, will not be immediately solved by us for the arbitrary initial field $T_0(x)$.

Instead, we will try to find its solution for the point initial field, when the environment is heated only in the single point x_0 or only in this point there lies the diffusing substance or the random particle.

Here, there immediately arises the question concerning the model, through which one can describe the point concentration of heat, the substance or the probabilistic density. From the physical point of view, the idealizations through mass or heat condensation are usual. Though, how may they be described mathematically? You know that in these cases temperature, density of the diffusing substance or, accordingly, probabilistic density will become infinite and

everywhere beyond this point they will be equal to zero. Let us be not embarrassed by this and write that for the describing function $T_0(x)$ we have

$$T_0(x) = \begin{cases} \infty \\ 0 \end{cases} \quad \text{at} \quad \begin{matrix} x = x_0 \\ x \neq x_0 \end{matrix} \quad . \tag{29.3}$$

Let us also write that the point $x = x_0$ will hold c thermal units (c is the thermal capacity of the substance), a single mass or the probability equal to unity

$$\int\limits_{-\infty}^{\infty} T_0(x)dx = 1 \quad . \tag{29.4}$$

The function $T_0(x)$ defined by (29.3) will be, certainly, not a usual function; the integral (29.4) will be not a usual integral either. Let us consider this function to be some new symbol δ -function or the generalized function. This unusual δ - function was introduced into mathematics by O. Heaviside in connection with the operational calculus created by him. Later, it was introduced into quantum mechanics by physicist P. Dirac.

The generalized function $\delta(x)$ is defined by the conditions

$$\delta(x) = \begin{cases} \dfrac{\infty}{} \\ 0 \end{cases} \quad \text{at} \quad \begin{matrix} x = 0 \\ x \neq 0 \end{matrix} \tag{29.5}$$

$$\int\limits_{-\infty}^{\infty} \delta(x)dx = 1$$

From (29.5), it formally follows that for any continuous function $f(x)$ we have

$$\int\limits_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = f(x_0) \quad . \tag{29.6}$$

Actually, for any indefinitely small $\varepsilon > 0$ we get

$$\int\limits_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = \int\limits_{x_0 - \varepsilon}^{x_0 + \varepsilon} f(x)\delta(x - x_0)dx \quad .$$

According to the well-known mean-value theorem, this last integral will be equal to

$$f(\xi) \int_{x_0-\varepsilon}^{x_0+\varepsilon} \delta(x-x_0)dx = f(\xi),$$

where ξ will lie within the interval from $x_0 - \varepsilon$ to $x_0 + \varepsilon$. Since ξ may be indefinitely small and the function $f(x)$ is continuous, we shall obtain $\xi = x_0$.

Let we have $x_0 = 0$ and let the variable thermal field $\varphi(x,t)$ correspond to the initial thermal distribution $T_0(x) = \delta(x)$. The function $\varphi(x,t)$ shows how c thermal units of heat concentrated in the point $x = 0$ will spread along the entire line. The function $\varphi(x,t)$ will describe a diffusion of the substance initially concentrated in the point $x = 0$ and for a randomly walking particle this function will represent a density of the probability of its occurrence in the point x after the time t, if the particle initially occupied the point $x = 0$.

To find the function $\varphi(x,t)$ through logical reasoning is not simple. One can outguess it but it is not easy. We will do neither this nor that and shall write it simply as

$$\varphi(x,t) = \frac{1}{2a\sqrt{\pi t}} \exp\left(-\frac{x^2}{4a^2 t}\right). \tag{29.7}$$

Then, we shall examine whether this function $\varphi(x,t)$ at all x and $t > 0$ satisfies the thermal conductivity equation (29.1) and whether it satisfies the initial conditions (29.2), if $T_0(x) = \delta(x)$. The latter should not be understood literally: if $t = 0$ is inserted into (29.7), then we shall have to divide by 0 twice. Hence, we shall not insert $t = 0$ and shall take a limit as $t \to 0$, i.e. we shall show that

$$\lim_{t \to 0} \varphi(x,t) = \begin{cases} 0 & x \neq 0 \\ \infty & x = 0 \end{cases} \quad \text{at}$$

and that

$$\lim_{t \to 0} \int_{-\infty}^{\infty} \varphi(x,t)dx = 1.$$

The first is evident; the second is found by showing that at all $t > 0$ we obtain

$$\int_{-\infty}^{\infty} \varphi(x,t)dx = 1$$

Indeed,

$$\int_{-\infty}^{\infty} \frac{1}{2a\sqrt{\pi t}} exp\left(-\frac{x^2}{4a^2 t}\right)dx =$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} exp\left(-\frac{x^2}{4a^2 t}\right)d\frac{x}{2a\sqrt{t}} = \qquad (29.8)$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} exp\left(-\xi^2\right)d\xi = 1 .$$

The improper definite integral obtained is well known and is equal to $\sqrt{\pi}$. You may persuade yourself in it in the following way. Denote this integral through A Then, the following is obtained:

$$I^2 = \int_{-\infty}^{\infty} exp\left(-\xi^2\right)d\xi \int_{-\infty}^{\infty} exp\left(-\eta^2\right)d\eta =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} exp\left(-\xi^2 - \eta^2\right)d\xi d\eta,$$

or , passing to the polar coordinates r and φ, we get

$$I^2 = \int_{0}^{\infty} dr \int_{-\pi}^{\pi} exp\left(-r^2\right)dr\varphi = \pi \int_{0}^{\infty} exp\left(-r^2\right)dr^2 = \pi .$$

That the function $\varphi(x,t)$ satisfies the thermal conductivity equation (29.1) for all $t = 0$ may be verified by immediate substitution. Thus, the function $\varphi(x,t)$ is the one we were looking for. Here, we remind again that $\varphi(x,t)$, being determined by the formula (29.7), corresponds to the situation when at the initial instant $t = 0$ in the point $x = 0$ there have been placed c thermal units. As for the case of diffusion, at the initial instant $t = 0$ in the point $x = 0$ there has been placed a single mass of the diffusing substance. As for the walking particle, at the instant $t = 0$, it will be placed in the point $x = 0$. For $t \approx 0$, the function

$\varphi(x,t)$ has the form of a high and thin peak (its height is $\dfrac{1}{2a\sqrt{\pi t}}$). As t grows, this peak will be extending in the way shown in figure 29.1.
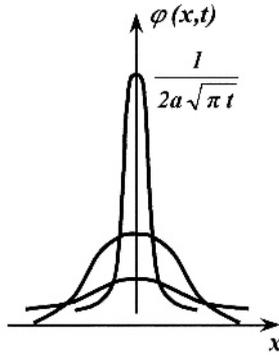
Fig. 29.1. Time changes of the response function $\varphi(x,t)$.

The square constrained by this spreading bell-like curve is, due to (29.8), will be all the time equal to unity at all t. This implies heat or substance conservation, or, respectively, a normalization of the probabilistic density.

If one knows the function $\varphi(x,t)$ describing a spread of single-point-concentrated heat, then through this function it becomes possible to construct similar functions for a two- and three-dimensional cases. For a three-dimensional case, such function will assume the form

$$\varphi(x,y,z,t,) = \varphi(x,t,)\varphi(y,t,)\varphi(z,t,) =$$
$$= \frac{1}{8a^3\left(\sqrt{\pi t}\right)^3}\exp\left(-\frac{x^2+y^2+z^2}{4a^2t}\right) \quad . \qquad (29.9)$$

To verify it is very simple. It is immediately found that

$$\lim_{t\to 0}\varphi(x,y,z,t) = \begin{cases} \infty \\ 0 \end{cases} \quad \text{at} \quad \begin{matrix} x=y=z=0 \\ x^2+y^2+z^2 \neq 0 \end{matrix} \quad . \qquad (29.10)$$

Now, let us write

$$\int\limits_{-\infty}^{\infty}\varphi(x,y,z,t)dxdydz =$$

$$= \int\limits_{-\infty}^{\infty}\varphi(x,t)dx \int\limits_{-\infty}^{\infty}\varphi(y,t)dx \int\limits_{-\infty}^{\infty}\varphi(z,t)dx = 1 .$$

Further, the equation

$$\frac{\partial\varphi(x,y,z,t)}{\partial t} = \frac{\partial^2\varphi(x,y,z,t)}{\partial x^2} + \frac{\partial^2\varphi(x,y,z,t)}{\partial y^2} +$$

$$+\frac{\partial^2\varphi(x,y,z,t)}{\partial z^2}$$

will be written in the form

$$\varphi(y,t)\varphi(z,t)\frac{\partial\varphi(x,t)}{\partial t} + \varphi(x,t)\varphi(z,t)\frac{\partial\varphi(y,t)}{\partial t} +$$

$$+ \varphi(x,t)\varphi(y,t)\frac{\partial\varphi(z,t)}{\partial t} = \varphi(y,t)\varphi(z,t)\frac{\partial^2\varphi(x,t)}{\partial x^2} +$$

$$+ \varphi(x,t)\varphi(z,t)\frac{\partial^2\varphi(y,t)}{\partial y^2} + \varphi(x,t)\varphi(y,t)\frac{\partial^2\varphi(z,t)}{\partial z^2} ,$$

from which there follows its solvability.

The function $\varphi(x,t)$ describes the thermal field that arises as a result of the spreading of c thermal units held in the point $x = 0$ at the instant $t = 0$. If the same quantity of heat is placed into the point x_0 at the instant τ, then the resulting thermal field will be described by the function $\varphi(x,t)$, where x will be substituted by $x - x_0$ and t by $t - \tau$, i.e. by the function $\varphi(x - x_0, t - \tau)$.

It turns out so that via this function $\varphi(x,t)$ it becomes possible to find the thermal field $T(x,t)$ for any initial thermal distribution $T_0(x)$. That is why the special solution of the function $\varphi(x,t)$ was called fundamental. In order to prove it, let us apply the superposition principle. But prior to it, let us consider a random walk being described by it.

The probabilistic density of a randomly wandering particle (or an absolutely drunken) varies in time in the fashion shown in figure 29.1. Here, a question may

arise about how far this absolutely drunken will move off during the period of time t from his initial point $x = 0$.

The average distance (the mathematical expectation of x) is equal to zero, because possible departures to the left and right will balance each other; hence, it is better to apply the mathematical expectation of $|x|$ or $\sqrt{Mx^2}$. This and that magnitudes are easily calculated and are equal to

$$M|x| = a\sqrt{\frac{t}{\pi}}, \sqrt{Mx^2} = a\sqrt{2t} \quad , \tag{29.11}$$

i.e. in both cases the distance is proportional to $\sqrt{t}$ while in case of the uniform movement this move off will be proportional to t. Thus, even a very slowly travelling particle will move off farther than an endlessly fast randomly wandering particle. Alongside with it, upon any time (very short as well) this particle, with very small probability density, may move off indefinitely far from its initial location.

Thermal spreading and substance diffusion occur in the way similar to a spreading of the probabilistic density of a randomly wandering particle. It is worth emphasizing here that this spreading is unusual to some extent. During some finite and even small time, the temperature will increase arbitrary far. This increase is negligible but existent, i.e. heat is spreading infinitely fast. The same refers to diffusion. This absurdity being insignificant because of negligible smallness of the distant values of the fundamental solution $\varphi(x,t)$ constitutes a somewhat principal drawback of our models. The essence of this drawback is seen from (28.9), according to which τ and a tend to zero, but the magnitude $\dfrac{a^2}{2\tau} = \dfrac{a}{\tau}\dfrac{a}{2}$ tends to the finite number D. Therefore, the displacing velocity of the random particle a/τ grows unlimitedly.

Let us return to discussing the role played by the fundamental solution $\varphi(x,t)$ and the superposition principle in solving the general problem of the thermal field $T(x,t)$ with its initial value $T_0(x)$. It follows from the superposition principle that if for initial thermal fields $T_{01}(x)$ and $T_{02}(x)$ we obtain the thermal fields $T_1(x,t)$ and $T_2(x,t)$, respectively, then for the initial thermal field $c_1 T_{01}(x) + c_2 T_{02}(x)$ we obtain the subsequent thermal field $c_1 T_1(x,t) + c_2 T_2(x,t)$.

The initial thermal field $T_0(x)$ may be written as

$$T_0(x) = \int\limits_{-\infty}^{\infty} T_0(\xi)\delta(x - \xi)d\xi \qquad (29.12)$$

and may be thought as a sum of fields $\delta(x - \xi)T_0(\xi)d\xi$ for all possible ξ ; to each of the components of this sum will (upon the time t) correspond the thermal field $\varphi(x - \xi)T_0(\xi)d\xi$.

Accordingly, under the superposition principle, we obtain

$$T(x,t) = \int\limits_{-\infty}^{\infty} T_0(\xi)\varphi(x - \xi, t)d\xi \quad . \qquad (29.13)$$

This is the very solution of the problem (29.1), (29.2). If necessary, you may – and it will be useful for you – verify it by substituting (29.13) into (29.1) and (29.2).

Let us again write this solution (29.13) showing a specific form of the fundamental solution as

$$T(x,t) = \frac{1}{2a\sqrt{\pi t}} \int\limits_{-\infty}^{\infty} T_0(\xi)\exp\left(-\frac{(x - \xi)^2}{4a^2 t}\right)d\xi \quad . \quad (29.14)$$

In the two- and three-dimensional cases the wanted solutions are written similarly. So, in the three-dimensional case we get

$$T(x, y, z, t) =$$

$$= \int\limits_{-\infty}^{\infty} T_0(\xi, \eta, \zeta)\varphi(x - \xi, t)\varphi(y - \eta, t)\varphi(z - \zeta, t)d\xi d\eta d\zeta =$$

$$= \frac{1}{8a^3\left(\sqrt{\pi t}\right)^3} \int\limits_{-\infty}^{\infty} T_0(\xi, \eta, \zeta) \times$$

$$\times \exp\left(-\frac{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2}{4a^2 t}\right)d\xi d\eta d\zeta \quad ,$$

where, as earlier, $T_0(x, y, z)$ is a three-dimensional initial thermal field for $t = 0$.

Here, the capabilities of the fundamental solution are not over. It may be also employed to solve a more general problem, when together with the initial thermal

field there exists a distributed thermal emission or thermal absorption. The first may be imagined as a burning or an exothermic chemical reaction in a heat conducting environment; the second case as a melting of small-size pieces of ice or an endothermic chemical reaction in this environment. This new problem may be solved by us even without constructing the necessary equation in partial derivatives.

Let the emitting heat for the time $d\tau$ in any instant τ create some additional thermal distribution $cf(x,\tau)d\tau$. As we know, this distribution is represented by the solution (29.13) or (29.14), where $T_0(\xi)$ should be replaced by $cf(\xi,\tau)d\tau$ and t by $t-\tau$, i.e. we obtain

$$\int_{-\infty}^{\infty} f(\xi,\tau)d\tau\varphi(x-\xi,t-\tau)d\xi \ .$$

By the time t, all these thermal fields will be summed up, i.e. we get

$$T(x,t) = \int_0^t \int_{-\infty}^{\infty} f(\xi,\tau)\varphi(x-\xi,t-\tau)d\xi d\tau \ . \tag{29.15}$$

It occurs, certainly, for the initial thermal field $T_0(x) = 0$. If there is $T_0(x) \neq 0$, then to the solution (29.15) there should be also added the solution (29.13); thus, as a result, we shall have

$$T(x,t) = \int_{-\infty}^{\infty} T_0(\xi)\varphi(x-\xi,t)d\xi +$$

$$\tag{29.16}$$

$$+ \int_0^t \int_{-\infty}^{\infty} f(\xi,\tau)\varphi(x-\xi,t-\tau)d\xi d\tau \ .$$

This is a solution of the equation (29.1), whose right-hand side is augmented with the component $f(x,t)$ standing for the associated continuous heat emission (if (29.2) obeyed). In such a beautiful way the fundamental solution is exploited for deriving the solutions on the thermal conductivity in the unlimited heat-conducting environment. The beauty here lies in the fact that the solutions, though complicated, are found easily, without calculations, but only on the basis of the superposition principle. But here one had to outguess that the solution $\varphi(x,t)$ corresponding to a point heat distribution should be found.

Now, let us solve the problem of warming up a half-space with use of the fundamental solution $\varphi(x,t)$. This problem is formulated as follows. From one side of the space, temperature is always equal to T_1; from another there is a heat-conducting environment having initially the temperature T_0. The question is how the temperature in this last half-space will vary.

Let the origin O be chosen on the environmental boundary and the axis Ox be directed into the depth of the heat-conducting environment (Fig. 29.2). Its thermal field will, then, depend only upon x and, therefore, will be described by the function $T(x,t)$.
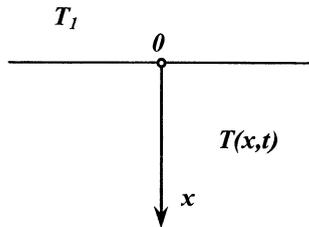
Fig. 29.2. The illustration to the problem of heating the finitely thick wall.

The function $T(x,t)$ satisfies the below thermal conductivity equation

$$\frac{\partial T}{\partial t} = a^2 \frac{\partial^2 T}{\partial x^2} \quad , \tag{29.17}$$

the initial conditions

$$T(x,t)\big|_{t=0} = T_0 \tag{29.18}$$

and, besides, the boundary condition

$$T(x,t)\big|_{x=0} = T_1 \quad , \tag{29.19}$$

since, in accordance with the conditions of the problem, the temperature T_1 in the half-space $x \leq 0$ will be sustained. Here, in contrast to the previous case, there is a boundary of the area, within which the thermal field is considered. Accordingly, a supplementary condition, i.e. a boundary condition (29.19), is introduced.

Note that the temperature T_0 may be assumed equal to zero. Here T_1 is substituted by $T_1 - T_0$. Though, it is possible not to do it and simply to find a thermal field in the form

$$T(x,t) = T_0 + \overline{T}(x,t) \; ,$$

where $\overline{T}(x,t)$ meets the conditions (29.17), (29.18) and (29.19), in which T_0 and T_1 are replaced with zero and $T_1 - T_0$ respectively.

This problem may, perhaps, be solved on the basis of purely mathematical reasoning. Though here, I would like to demonstrate you on this example how purely physical concepts can help and prompt you a solution.

To make things definite, let we have $T_1 > T_0$ and, therefore, there occurs a warming up of the half-space $x > 0$. This warming up occurs at the expense of heat transition occurring through the boundary point $x = 0$ (through a square unit of the boundary of the segment $x = 0$) with the intensity $q(t)$. The function $q(t)$ is not known to us, but through it we can, with use of the fundamental solution, write $T(x,t)$ in the form

$$T(x,t) = T_0 + \int_0^t q(\tau)\varphi(x, t - \tau)d\tau \; . \qquad (29.20)$$

The function $T(x,t)$ determined by the formula (29.20) satisfies (29.17) and (29.18) for any function $q(\tau)$. Hence, we need to satisfy only the requirement (29.19), i.e.

$$T_1 = T_0 + \int_0^t q(\tau)\varphi(0, t - \tau)d\tau$$

or

$$T_1 = T_0 + \int_0^t \frac{q(\tau)}{2a\sqrt{\pi(t - \tau)}}d\tau \; . \qquad (29.21)$$

Introducing the new variable $u = \dfrac{\tau}{t}$ we write (29.21) as

$$T_1 - T_0 = \frac{1}{2a\sqrt{\pi}} \int_0^t \frac{q(tu)\sqrt{t}}{\sqrt{1 - u}}du \; . \qquad (29.22)$$

To meet this last condition, the dependence of the integrand upon t should be, at least, eliminated. For this purpose it is necessary to choose

$$q(tu) = \frac{c}{\sqrt{tu}} \quad , \quad (29.23)$$

after which (29.22) takes the form

$$T_1 - T_0 = \frac{1}{2a\sqrt{\pi}} \int_0^t \frac{du}{\sqrt{u(1-u)}} = \frac{\sqrt{\pi}}{2a} c \quad .$$

From this, we shall find that

$$c = \frac{2\alpha(T_1 - T_0)}{\sqrt{\pi}} \quad ; \quad (29.24)$$

so, the needed solution, due to (29.22), (29.23) and (29.24), will take the form

$$T(x,t) = T_0 + \frac{T_1 - T_0}{\pi} \int_0^t \frac{d\tau}{\sqrt{\tau(1-\tau)}} \exp\left(-\frac{x^2}{4a^2(t-\tau)}\right) \quad . \quad (29.25)$$

The analysis of this solution reveals it to be a function of $\dfrac{t}{x^2}$. It implies in particular that in order to achieve at the depth $2x$ the temperature similar to that already available at the depth x it is necessary to have the time as much as four times (not two times only).

The general statement given above is proved through reducing the expression (29.25), upon substituting the variable τ by $\eta = \dfrac{\tau}{x^2}$, to the form

$$T(x,t) = T_0 +$$
$$+ \frac{T_1 - T_0}{\pi} \int_0^{t/x^2} \frac{d\eta}{\sqrt{\eta(t/x^2 - \eta)}} \exp\left(\frac{1}{4a^2(t/x^2 - \eta)}\right) = \quad (29.26)$$
$$= T_0 + F(t/x^2).$$

This solution, as the function of $\dfrac{t}{x^2}$, is graphically shown in figure 29.3.
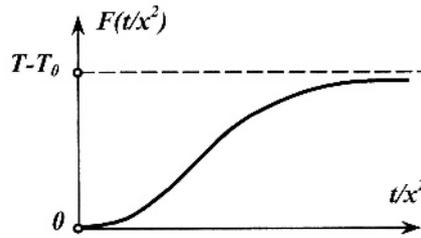
Fig. 29.3. The temperature changes in the thick wall under heating.

Such a shape is a result of F growing monotonously as a function of the argument $\dfrac{t}{x^2}$, as a result of $T(x,t) = T_0$ for $\dfrac{t}{x^2} = 0$ and $T(x,t) = T_1$ for $\dfrac{t}{x^2} = \infty$.

30 Running waves and the dispersion equation

The running harmonic waves. The earth surface warm-up under influence of daily and yearly temperature fluctuations. The problem of the rate of ice freeze over the water surface.

The linear oscillator equation

$$\ddot{x} + 2\delta\dot{x} + \omega^2 x = 0$$

has the specific solution of the type $e^{\lambda t}$, where the values of λ are found from the characteristic equation

$$X(\lambda) = \lambda^2 + 2\delta\lambda + \omega^2 = 0 \ .$$

Analogously, for the thermal conductivity equation

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} \tag{30.1}$$

and for the wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \tag{30.2}$$

there exist the solutions of the type $e^{i(\omega t - kx)}$, where the possible values of ω and k are found from the dispersion equation

$$X(\omega, k) = 0 \ . \tag{30.3}$$

Inserting into (29.1) the assumed solution

$$u(x,t) = e^{i(\omega t - kx)} \tag{30.4}$$

we shall come to the dispersion equation

$$i\omega + a^2 k^2 = 0 \ . \tag{30.5}$$

For the equation (30.2), the similar substitution will bring

$$\omega^2 - a^2 k^2 = 0 \ . \tag{30.6}$$

By expressing k through ω, in accordance with (30.5), and in accordance with (30.6), we find the equation (30.1) to have all possible solutions of the type

$$u(x,t) = \exp i \left(\omega t \pm \frac{\sqrt{\omega}}{a} \frac{1-i}{\sqrt{2}} x \right) \ . \tag{30.7}$$

The equation (30.2) will have the solutions of the type

$$u(x,t) = \exp i \left(\omega t \pm \frac{\omega}{a} x \right) \ . \tag{30.8}$$

Let us clarify what the solutions (30.7) and (30.8) will represent, i.e. the solutions of the type

$$u(x,t) = \exp i (\omega t + kx) \ , \tag{30.9}$$

where ω is real and k may be complex.

Let ω and k be real initially. Then,

$$\exp i (\omega t + kx) = \cos (\omega t + kx) + i \sin (\omega t + kx) \ .$$

Therefore, the functions

$$\cos (\omega t + kx), \sin (\omega t + kx) \tag{30.10}$$

will be the real solutions.

For $t = 0$, to these solutions there will correspond the variable harmonically varying fields of the wave length $\lambda = \dfrac{2\pi}{|k|}$ (Fig. 30.1). When t grows, this field as a whole will travel along the axis x at the velocity $v = -\dfrac{\omega}{k}$. Therefore, the functions (30.10) are, and (30.9) may be interpreted to be a harmonic wave of the length $\lambda = \dfrac{2\pi}{|k|}$ running at the velocity $v = -\dfrac{\omega}{k}$. Here, in each point of x, as seen from (30.10), the field, as a function of t, will be also vary, in accordance with the harmonic law, with the frequency ω and the unity amplitude.
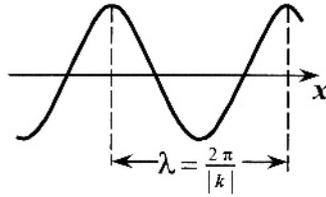
Fig. 30.1. The harmonically varied one-dimensional scalar field.

That is why the magnitudes ω and k are called a frequency and a wave number of the running harmonic wave (30.9).

Let now the wave number k be complex, i.e.

$$k = k_1 + ik_2 \ .$$

Then,

$$\exp i\,(\omega t + kx) = e^{-k_2 x}\cos(\omega t + k_1 x) + ie^{-k_2 x}\sin(\omega t + kx) \ .$$

Therefore, the real solutions will be

$$e^{-k_2 x}\cos(\omega t + k_1 x) \ , \ \ e^{-k_2 x}\sin(\omega t + kx) \ .$$

If we had had $k_2 = 0$, then it would be a running harmonic wave. The presence of the multiplier $e^{-k_2 x}$ makes this wave decrease, for $k_2 > 0$, exponentially in the direction of the positive x, or, vice versa, increase, for $k_2 < 0$, exponentially. The wave itself will displace to the right when $k_1 < 0$ and to the left when $k_1 > 0$. The displacement rate v, as well as in the previous case, will be equal to $\dfrac{2\pi}{|k|}$. Thus, for $k_1 < 0$ and $k_2 > 0$ this wave will be a descending right-running wave (Fig. 30.2). For $k_1 < 0$ and $k_2 < 0$, it will be an ascending right-running wave (Fig. 30.3).
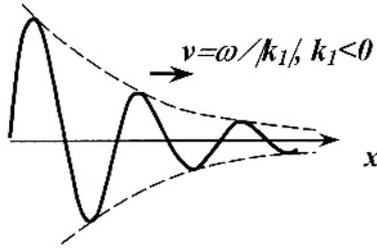
Figs. 30.2. The running descending harmonic wave.
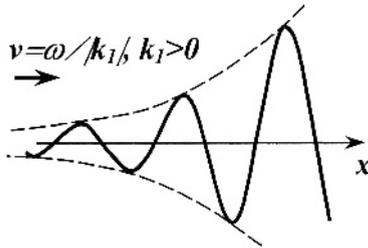
Fig. 30.3. The running ascending harmonic wave.

Analogously, for $k_1 > 0$ we shall get the left-running wave descending at $k_2 > 0$ and ascending at $k_2 < 0$. In all these cases, the field in each point of x will vary under the harmonic law with the frequency ω and the amplitude $e^{-k_2 x}$.

In accordance with (30.8), the wave equation (30.2) admits all possible travelling harmonic waves of all possible frequencies and wave numbers, respectively, w and $\pm w / a$. All these waves have similar propagating velocity, $-a$ and, respectively, $+a$. The lengths of the waves will be different and equal to $2\pi a / w$, as well as their corresponding frequencies.

According to the superposition principle, alongside with all possible travelling harmonic waves (30.8), there will also take place all possible running waves of the type

$$\sum_s c_s \cos\left[\omega_s\left(t \pm \frac{x}{a}\right) + \varphi_s\right]$$

and of the type

$$\int\limits_{0}^{\infty}c(\omega)\cos\left[\omega\left(t\pm\frac{x}{a}\right)+\varphi\right]d\omega \ .$$

This brings us to the idea that the wave equation admits the arbitrary (travelling at the velocity $\pm\alpha$) waves

$$u(x,t)= f\left(t\pm\frac{1}{a}x\right) \tag{30.11}$$

of any form. Indeed, it occurs really so. It is easily checked that for any function f the equation (30.11) meets the wave equation (30.2).

If all running waves of the wave equation spread in both directions at the same velocity a and neither damping nor growing occur, then the wave solutions (30.7) of the thermal conductivity equation will have different spreading velocities and will diminish exponentially. In accordance with (30.7), the real wave solutions will be the variable fields of the below type

$$u(x,t)= exp\frac{\sqrt{\omega}}{a\sqrt{2}}x\cos\left(\omega t+\frac{\sqrt{\omega}}{a\sqrt{2}}x\right) \tag{30.12}$$

and the fields of the type

$$u(x,t)= \exp\left(-\frac{\sqrt{\omega}}{a\sqrt{2}}x\right)\cos\left(\omega t+\frac{\sqrt{\omega}}{a\sqrt{2}}x\right) , \tag{30.13}$$

for different nonnegative ω.

The field described by the formula (30.12) is a damped left-running wave; the field (30.13) a damped right-running wave. Here, both the damping rate and the velocity of the wave motion will depend upon its frequency ω. The wave displacement rate will be equal to $a\sqrt{2\omega}$; it will be the greater, the greater will be its frequency ω or the less will be its length $\lambda=\dfrac{2\sqrt{2}\pi a}{\sqrt{\omega}}$. The shape of these running waves is shown in figure 30.4.
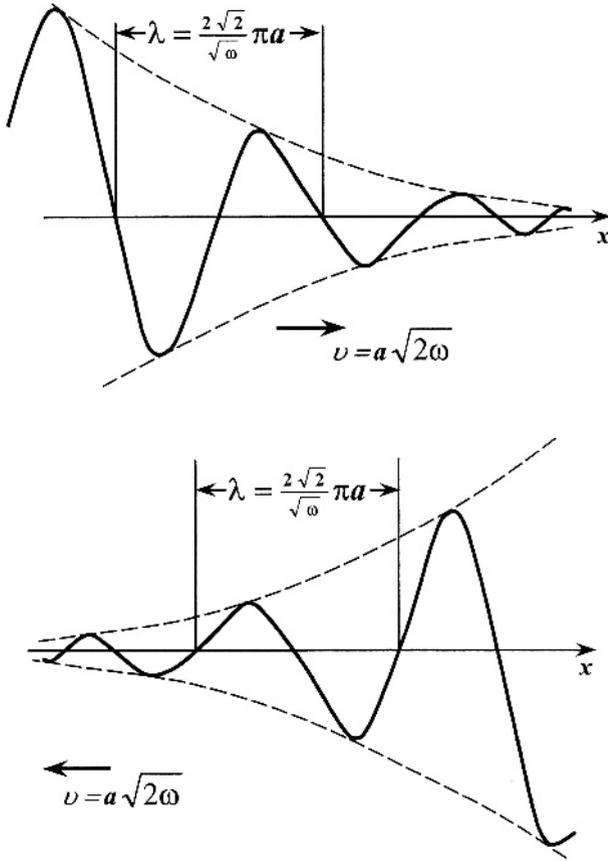
Fig. 30.4. The running harmonic thermal waves; ω is a time frequency of the wave, λ is its special length, ν is its propagating velocity.

The figure shows the wave spreading velocity, the length and how fast the wave is damping.

Let us now apply the above information to the next two similar problems for the wave equation and for the thermal conductivity equation. As a physical model for the first problem, we will take a semi-infinite rope whose tail is being moved up and down according to the law $A \cos \Omega t$. As a physical model for the second problem there will be used a half-space; on its boundary, the temperature will vary under the same law.

I think that in your childhood you enjoyed very often creating waves along a rope by shaking its tail manually. In the same way our nature changes the temperature above the surface of the planet through changing days by nights or

winters by summers. Here, we shall arrive at the above formulated problems, if we accept the movements of your hand and variations of temperatures to vary under the harmonic law. This is, certainly, a rough approximation. But it is sufficient for understanding the occurring phenomena.

Let us construct the mathematical models for the first and second cases. We assume one to have started to oscillate the rope long ago, as very long ago the temperature above the surface of our planet started to vary.

In the first case, the oscillations of the semi-infinite rope is described by the wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \quad , \tag{30.14}$$

for $-\infty < t < +\infty$ and $x \geq 0$. For $x = 0$ and all t, there will take place

$$u(x,t)\Big|_{x=0} = A \cos \Omega t \quad . \tag{30.15}$$

In the second case, the thermal field is governed by the thermal conductivity equation

$$\frac{\partial T}{\partial t} = a^2 \frac{\partial^2 T}{\partial x^2} \quad , \tag{30.16}$$

for $-\infty < t < +\infty$ and $x \geq 0$; for $x = 0$ and all t, we also obtain

$$T(x,t)\Big|_{x=0} = A \cos \Omega t \quad . \tag{30.17}$$

As you see here, the statements of both problems are very similar. In the first case you know very well what will happen. This allows us, combining the seen by you with your knowledge you have now obtained concerning the harmonic waves, to outguess the solution. It will be the Ω-frequency harmonic wave running from your hand along the rope, i.e.

$$u(x,t) = A \cos \Omega \left(t - \frac{x}{a} \right) \quad . \tag{30.18}$$

This wave will satisfy the equation (30.14) and the edge condition (30.15).

Analogously, the solution of the second problem is written in the form of the thermal wave running from the earth's surface into the depth and having the frequency Ω, i.e. by virtue of (30.13) we shall obtain

$$T(x,t) = A \exp\left(-\frac{\sqrt{\Omega}}{a\sqrt{2}} x \right) \cos\left(\Omega t - \frac{\sqrt{\Omega}}{a\sqrt{2}} x \right) \quad . \tag{30.19}$$

It is clear that the function $T(x,t)$ determined by (30.19) satisfies the thermal conductivity equation (30.16) and the edge condition (30.17).

The solution (30.18), I think, is well familiar to you from observations. Though, if the rope had been sufficiently long, then you could have noticed that the running wave is damping. This is already a defect of our idealization, since both the internal friction in the rope during its deformations and its friction against air were left beyond our consideration. Though, when needed, they may be accounted; say, the friction force might have been considered proportional to the velocity of the rope movement, but directed contrariwise. This could have brought us to the equation of the shape

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} - h \frac{\partial u}{\partial t} \quad . \tag{30.20}$$

Further, it becomes possible to derive the wave solutions of the form (30.4) and it may be used for improving the obtained solution (30.18). Indeed, by inserting (30.4) into (30.20) we shall find that

$$\omega^2 - a^2 k^2 - ih\omega = 0$$

or

$$k = \frac{\omega}{a}\sqrt{1 - i\frac{h}{\omega}} \quad .$$

For $h/\omega \ll 1$, we shall approximately find that

$$k = \pm\frac{\omega}{a}\left(1 - i\frac{h}{2\omega}\right) \quad .$$

We are interested in the wave running from the hand; so, in the obtained value of the wave number k we will take the "minus" sign. The corresponding improved solution will be as

$$u(x,t) = A\exp\left(-\frac{hx}{2a}\right)\cos\left(\Omega t - \frac{\Omega}{a}x\right) \quad . \tag{30.21}$$

With x growing, this wave will damp really; for $h = 0$, it will transfer, as it should, to (30.18).

The wave running along the rope may produce a sudden effect, if the rope is made thinner. This thinning brings a growth of the wave spreading velocity a, because $a^2 = \dfrac{T}{\rho}$, where T is a tension of the rope and ρ its linear density.

Here, subjected to growth is not the velocity only but the oscillating amplitude as

well. The wave will be growing. As a result, there may be produced a strong clap caused by the fact that the travelling velocity near the rope terminal will exceed the sound velocity. To understand this phenomenon, let us take into account that the wave conserves the energy and the oscillating mass decreases. This phenomenon may be observed on a pasture when shepherds are cracking long whips made thinner in the end.

The solution (30.19) derived by us describes in an approximate qualitative fashion a warming up of the earth surface at the daily and yearly fluctuations of temperature. Here, the external temperature is assumed to fluctuate under (30.17) oscillating about the average value equal to zero. If the average temperature is assumed different from zero and equal to T_0, then the solution (30.19) will be augmented with T_0 and will take the shape

$$T(x,t) = T_0 + A \exp\left(-\frac{\sqrt{\omega}}{a\sqrt{2}} x \right) \cos\left(\Omega t - \frac{\sqrt{\Omega}}{a\sqrt{2}} x \right) . \qquad (30.22)$$

Let us analyse this solution. It includes the constant component T_0 and an exponentially damped running wave. The exponential decrease of daily fluctuations greatly exceeds the yearly fluctuations. The amplitude of thermal oscillations decreases e times at the depth $\dfrac{a\sqrt{2}}{\sqrt{\Omega}}$ and, therefore, if for the yearly thermal fluctuations it makes up about one meter, then for the daily fluctuations it will be $\dfrac{\sqrt{\Omega_{year}}}{\sqrt{\Omega_{day}}} \approx 20$ times less, i.e. of the order of centimeters. It follows that at the depth of some meters the temperature will be practically constant and equal to the average temperature. This explains the existence of the permafrost in the depth, while the external average temperature is below zero.

The thermal wave length λ is equal to $\dfrac{a\pi\sqrt{2}}{\sqrt{\Omega}}$ and for the yearly fluctuations, it will approximately exceed the daily fluctuations twenty times. The minimal and maximal surface temperature displaces into depth one after another at the wave spreading velocity v being equal to $a\sqrt{2\Omega}$. Thus, at the depth x the maximum and minimum of the temperature occur $\dfrac{x}{a\sqrt{2\Omega}}$ time later than on the surface. This phenomenon explains the fact that when the deeply dug water tubes are frozen in severe winter it should be kept in mind that this freezing may happen not necessarily in very severe frosts, but even much later, possibly, in spring, i.e. when it gets warmer significantly. The daily thermal waves spread into depth

much faster than the yearly waves (approximately 20 times faster) and damp significantly faster, exponentially, also approximately 20 times faster.

We have, thus, considered the thermal conductivity problems, which cover a warm-up of the earth surface, a warm-up of a thick wall in the building, the thermal fluctuations in the depth of the earth surface at daily and yearly thermal fluctuations on the surface.

Now, let us also consider the problem on how ice is frozen over the water surface. Water starts to freeze when it is cooled down to zero and its cooling goes on from its surface. Figure 30.5 schematically shows ice thickness and the unfrozen water at $0°$ C under the ice floe.
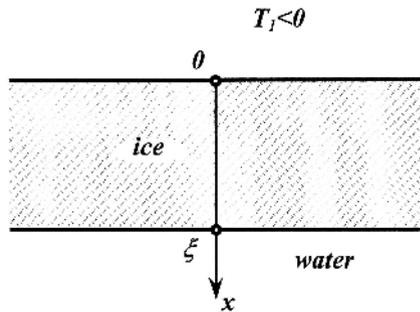
Fig. 30.5. The illustration to the ice freezing over the water surface.

On the surface of the ice floe the temperature T_1 is negative; the ice thickness is ξ. We would like to find out how ξ varies in time. Therefore, let the ice temperature be $T(x,t)$. This temperature is described by the below thermal conductivity equation

$$\frac{\partial T}{\partial t} = a^2 \frac{\partial^2 T}{\partial x^2} \tag{30.23}$$

and obeys the edge conditions

$$T(x,t)\big|_{x=0} = T_1 \quad , \quad T(x,t)\big|_{x=\xi} = 0 \quad , \tag{30.24}$$

which indicate that on the top surface and the bottom surface of the ice floe the temperature is T_1 and 0, respectively.

The heat flow from the warmer water to the cooler ice through the bottom surface of the ice will be equal to

$$Q \doteq k \frac{\partial T}{\partial x}\bigg|_{x=\xi} \quad . \tag{30.25}$$

This heat outflow gives birth to a fresh ice in such a way that

$$k \frac{\partial T(x,t)}{\partial x}\bigg|_{x=\xi} = \rho c \frac{d\xi}{dt} \quad , \tag{30.26}$$

where c is the heat emitted by a freezing unity of the ice mass, and ρ an ice density.

Such is our initial mathematical model. It is rather complicated. Its complexity is stipulated by the edge condition (30.25) holding the unknown variable ξ that is described by the equation (30.26), which itself is dependent upon the unknown ξ.

Let us take into consideration the specificity of the problem and simplify its mathematical model. The ice is being formed very slowly. Thus, the temperature in the depth of the ice may be assumed to be quasistationary, i.e. such that will appear at the constant ξ.

With ξ being constant, the steady-state time-independent thermal field $T(x,t) = u^*(x)$ will satisfy the equation

$$\frac{d^2 u^*}{dx^2} = 0 \tag{30.27}$$

following from (30.23) and will also satisfy the edge conditions

$$u^*\big|_{x=0} = T_1, \quad u^*\big|_{x=\xi} = 0 \tag{30.28}$$

following from (30.24). From (30.27), (30.28), we immediately obtain

$$u^* = -\frac{T_1}{\xi} x + T_1 \quad .$$

Therefore, the equation (30.25) will take the below form

$$\rho c \frac{d\xi}{dt} = -k \frac{T_1}{\xi} \quad . \tag{30.29}$$

From it, taking into account that for $t=0$ we shall get $\xi=0$, we shall obtain

$$\xi = \sqrt{-\frac{2kT_1}{\rho c} t} \quad . \tag{30.30}$$

This implies that the ice thickness will increase under the law $\sqrt{t}$, i.e. the growth of the thickness will get slower with time. Thus, if one centimeter of ice appears per a day, then a 6-cm thickness (enough to allow a walk on) will appear not for 6 but 36 days.

The dependence upon external temperature is also decelerated, being proportional to $\sqrt{-T_1}$. Therefore with the external temperature being 2 times lower, the ice formation velocity will increase $\sqrt{2}$ times only.

Actually everything certainly does not go in this way because the temperature T_1 is not constant and varies with time. If these temperature fluctuations are not too fast and allow for the thermal field to be established in the ice depth according to the temperature $T_1(t)$, then the same equation (30.29) may be used as earlier; though, T_1 in this equation has to be taken as a function of time. From this equation (30.29) one may find that

$$\rho c \xi \, d\xi = -kT_1(t) \, dt$$

or

$$\frac{\rho c}{2} \xi^2 = -k \int_0^t T_1(\tau) d\tau \ .$$

From which we obtain

$$\xi = \sqrt{\frac{2k}{\rho c}\left(-\int_0^t T_1(\tau)d\tau\right)^{1/2}} \ . \tag{30.31}$$

From the approximate formula (30.31) it follows that within its errors the thickness of the ice frozen over during the time t is described by the same formula (30.30), where T_1 is already an average temperature over the ice surface for the time t . Thus, we obtain that the ice thickness frozen over during the winter will be proportional to the squared root from the average winter temperature and to the squared root from the winter duration.

31 Faraday-Maxwell theory of electromagnetism and the Maxwell-Hertz electromagnetic waves

It is shown that the Maxwell equations possess the solution in the form of the harmonic wave running at the light speed.

Earlier you became acquainted with the great equations of J. C. Maxwell. They are rather complicated and not easy to understand. They were not understood by the contemporaries and only few accepted them seriously. To comprehend them even nowadays when we have a clearer and simpler vector form of writing and an advanced level of thinking needs a lot of thinking effort and a comprehension of new notions and ideas. In the initial Maxwell's notation, the equations were written in a quaternion fashion and held some supplementary equations making their idea vague. Prior to Maxwell, the electrical and magnetic fields were studied separately. They were supposed to be born instantly, respectively, by charges and currents, which were considered as the first sources of electromagnetic phenomena. In contrast to them, the Maxwell equations described something different and new. They described electromagnetic fields, which, varying and interacting each other, can exist and spread across space by themselves at some finite velocity. From the Maxwell equations there could be derived the wave equation manifesting an existence of a running harmonic electromagnetic wave. Its velocity turned out to be near to the light velocity and this made it possible for Maxwell to suggest the hypothesis on electromagnetic nature of light waves.

The theory of J.C. Maxwell, a Scottish physicist, was opposing the then existent European theory. The latter, from our contemporary viewpoint, referred to quasistationary electrodynamics, well developed mathematically and seemed completely consistent to the experiment. H. R. Hertz, being a supporter of the European theory, decided to disprove this "unintelligent" theory by Maxwell and arranged some fundamental and well-thought experiments. In essence, these experiments pursued the objective of upholding the then European theory of electromagnetism successfully developed by such giants as A.M. Ampere, P.S. Laplace, K.F. Gauss, C. G. Neumann, W.E. Weber, H. L. F. Helmholtz , against the "absurd" theory by M. Faraday and J.C. Maxwell. The first jumped from charges and currents, from the description, on this basis, of the electromagnetic phenomena. As we know at present, they were moving along a blind alley. Charges and currents are not able in all cases to construct a state of the electrodynamical system; hence, they do not allow to understand, investigate and predict its behaviour. To construct a state, one greatly needed a fundamentally new understanding of the field. There was a need in an idea of an electromagnetic

field, for which charges and currents were only its particularities. In the Faraday-Maxwell theory, electromagnetic phenomena can occur without charges and currents. In Faraday-Maxwell electrodynamics, it is the electromagnetic fields that constitute a state of the dynamical system. If in the mechanics of continua the displacement fields and velocity fields arose naturally and were not generated by new flights of thinking, then, in contrast to this, the picture in electrodynamics was absolutely another. This branch of science constructs absolutely strange fields difficult for human comprehension.

This leap in human thinking was performed by Faraday. It was he who thought out electrical and magnetic fields and through them described mysterious electromagnetic phenomena. Maxwell came to believe in the genius of self-taught Faraday and he, as he wrote himself, only described Faraday's ideas mathematically. But this very reckless leap of thinking was needed for constructing an adequate mathematical model.

The experimental results obtained by Hertz were shocking; they opposed the theory that was supported by Hertz and was so dear to him. The results fully corresponded to the theory by Maxwell. In spite of this, Hertz published his experimental results and via them he developed the Maxwell theory further – he revealed an electromagnetic emitter that was called the Hertz vibrator. In this way, together with Maxwell, he was standing close to the source of opening the radio, the greatest technical discovery in the beginning of the 20-th century. Hertz's experiments proved Maxwell's theoretical concepts; from this time on, nothing was standing in the way of triumphally stalking equations of Maxwell.

Nowadays, it is not so easy to repeat Maxwell's research. If done, they would be hardly understandable. Therefore, within the frames of our narration on distributed dynamical systems and travelling harmonic waves let us demonstrate how the existence of electromagnetic waves follows from the Maxwell equations. In other words, our objective is to make it clear for ourselves whether the Maxwell equations admit the solutions in the form of the travelling waves, when the vectors E and H are only the functions of the two variables, t and, say, the variable x, i.e. in the form of the electromagnetic wave independently existing in an empty space, without charges and currents, i.e. when $\varepsilon = \mu = 1$ and $\sigma = 0$. Under these conditions, the Maxwell equations (31.19), (31.20) and (31.21) are simplified and assume the form

$$\frac{\partial E}{\partial t} = c\ rot\ H \qquad \frac{\partial H}{\partial t} = -c\ rot\ E \qquad (31.1)$$

$$div\ E = 0 \qquad div\ H = 0 \quad ,$$

where $E = E(x,t)$ and $H = H(x,t)$. Note here that without simplifying the problem by the immediate search for a plane wave we can reveal that the equations (31.1) yield the wave equations for the fields E and H. Indeed, let us differentiate the first equation (31.1) in t and apply the operation rot to the second equation

$$\frac{\partial^2 E}{\partial t^2} = c \quad rot \frac{\partial H}{\partial t} \quad , \qquad rot \frac{\partial H}{\partial t} = -c \quad rot \, rot \, E \quad .$$

Upon these operations, we, using the well-known formula in vector analysis $rot \, rot = \nabla div - \Delta$, shall obtain

$$\frac{\partial^2 E}{\partial t^2} = -c \, rot \, rot \, E = -c^2 \left(\nabla div E - \Delta E \right) \quad ;$$

by virtue of the third equation (31.1) we shall arrive at the below wave equation for the vector field

$$\frac{\partial^2 E}{\partial t^2} = c^2 \Delta E \quad .$$

The wave equation for the magnetic field vector is derived similarly. Further, it becomes already possible to derive the solutions of these equations (obeying the requirements of (31.1)). We drop here pursuing this general problem and forget this efficient application of vector analysis. Instead, we shall modestly proceed our search for the solution of the equations (31.1) in the form of the plane wave travelling along the axis x.

Therefore, in the equations (31.1) the vectors E and H will be the functions of x and t only. From the last two equations (31.1), it immediately follows that

$$\frac{\partial E_x}{\partial x} = 0 \qquad \frac{\partial H_x}{\partial x} = 0 \quad ;$$

therefore, the components E_x and H_x of the vectors E and H will be independent of x. That is why they are not interesting for us and so we will take $E_x = H_x = 0$. Then, we represent the first two equations (31.1) through the components

$$\frac{\partial E_y}{\partial t} = -c \frac{\partial H_z}{\partial x} \qquad\qquad \frac{\partial E_z}{\partial t} = c \frac{\partial H_y}{\partial x}$$

$$\frac{\partial H_y}{\partial t} = c \frac{\partial E_z}{\partial x} \qquad\qquad \frac{\partial H_z}{\partial t} = -c \frac{\partial E_y}{\partial x} \quad .$$

These equations are divided into the two independent pairs. One of them is written as

$$\frac{\partial E_y}{\partial t} = -c \frac{\partial H_z}{\partial x}, \qquad \frac{\partial H_z}{\partial t} = -c \frac{\partial E_y}{\partial x} \quad . \tag{31.2}$$

Differentiating the first equation in t and the second in x and, then, comparing them we achieve the wave equation

$$\frac{\partial^2 E_y}{\partial t} = c^2 \frac{\partial^2 E_y}{\partial x^2} \quad . \tag{31.3}$$

As known for us, it possesses the solution in the form of the running harmonic wave

$$E_y = A \sin\left(\omega t - \frac{\omega x}{c}\right) \quad . \tag{31.4}$$

Further, from any of the equations (31.2), we find that

$$H_z = A \sin\left(\omega t - \frac{\omega x}{c}\right) \quad . \tag{31.5}$$

The particular solution (31.4), (31.5) derived meets the initial equations (31.1) and represents itself the electromagnetic wave consisting of the two mutually perpendicular harmonic waves of the electrical and magnetic fields travelling along the axis x at the velocity c (Fig. 31.1). For $\varepsilon = \mu = 1$ and $\sigma = 0$, the Maxwell equations have, therefore, the solution (31.4), (31.5). All other components, E_x, E_z and H_x, H_z, will be equal to zero.
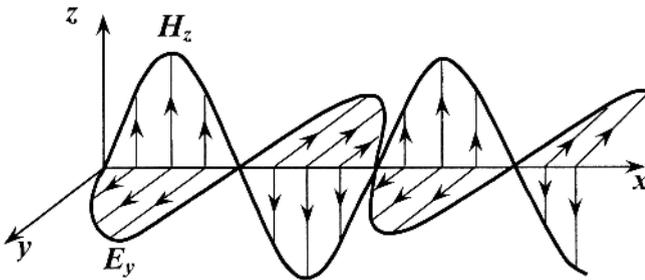
Fig. 31.1. The harmonic electromagnetic wave.

This solution is a plane electromagnetic harmonic wave having the arbitrary frequency ω, the wave-length $\dfrac{2\pi c}{\omega}$ and the propagating velocity c. Such different waves are numerous. This type of waves embraces light waves, radio

waves, ultraviolet and infrared waves, X-ray waves and gamma emission waves. All these waves are of similar electromagnetic nature and differ only by frequencies and wave lengths, at the similar propagating velocity.

32 Wave reflection and refraction

The reflected and refracted waves arising when a propagating harmonic wave is running along an environmental boundary.

We have, thus, become acquainted with the distributed dynamical systems described by fundamental equations of classic mathematical physics. These equations are: the thermal conductivity equation, the diffusion equation, the equation of random walk, the wave equation simulating the oscillations of a strained string or membrane, the longitudinal and transversal oscillations in an elastic environment, audio oscillations and electromagnetic waves. In contrast to discrete systems, new will be the waves; the thermal damped waves whose spreading velocity depends upon their frequency and length; undamped waves propagating across the elastic environments; and the electromagnetic waves running at the fixed velocity. Waves are associated with such amazing phenomena as interference; diffraction; a queer evolution of a wave package, this evolution entails distortions in speech and music when they are transmitted across long distances; and also wave reflection and refraction, etc. These phenomena are well familiar to you from your school course on general physics. Though, that course gave you only a descriptive narration holding no mathematical models. Right now we can not only describe these models but also study them exposing thus new quantitative and qualitative data. As another simple example let us consider a phenomenon of waves reflection and refraction caused by an existing boundary between the elastic environments of different characteristics. Let us look more specifically at the wave of the below shape

$$Ae^{i(\omega t - kx)} \tag{32.1}$$

travelling along some strained string and running against the spasmodic fluctuations of string densities. Let us direct the axis Ox along the string and let for $x < 0$ and $x > 0$ the string have various densities, i.e. the point $x = 0$ will be a point of environments partition. We need to find the alternating field $u(x,t)$ of string displacements; the field is induced by the wave (32.1) arriving at the partition boundary from the left. The mathematical model should include the wave equations for the string segments $x < 0$ and $x > 0$, the conditions of coordinating their solutions at $x = 0$. As well, the model will take into account the fact that the wave of the form (32.1) travels within the string segment $x < 0$ from left to right.

In accordance with all this, let we have $u(x,t) = u_1(x,t)$ for $x < 0$ and $u(x,t) = u_2(x,t)$ for $x > 0$. For $u_1(x,t)$ and $u_2(x,t)$ there will take place the below equations

$$\frac{\partial^2 u_1(x,t)}{\partial t^2} = a_1^2 \frac{\partial^2 u_1}{\partial x^2} \quad (x < 0) \tag{32.2}$$

and

$$\frac{\partial^2 u_2(x,t)}{\partial t^2} = a_1^2 \frac{\partial^2 u_2}{\partial x^2} \quad (x > 0) \tag{32.3}$$

where between $u_1(x,t)$ and $u_2(x,t)$ for $x = 0$ there will occur the conjugation conditions[1]

$$u_1(x,t)\big|_{x=0} = u_2(x,t)\big|_{x=0}$$

$$\frac{\partial u_1(x,t)}{\partial x}\bigg|_{x=0} = \frac{\partial u_2(x,t)}{\partial x}\bigg|_{x=0}.$$

$$\tag{32.4}$$

Here, it should be also said that at the environmental boundary there will arrive from the left the wave of the shape (32.1), with the wave number $k = \omega/a_1$.

Despite the simplicity of this problem, to solve it without the physical prompt given seems not easy. The physical prompt implies the following: this case is known to give birth to a reflected and a refracted waves. What these waves are is a question, but they may be naturally assumed, as well as an incident wave, to be similar to the running harmonic waves.

---

[1] The first relation is evident since in the point $x = 0$ the string is not broken. The second condition is not evident so much. This condition follows from the identity of the string tension N to the left and right from the point $x = 0$. From the left, the string junction point will be acted perpendicular to the axis Ox by the force $- N \sin\alpha_1$ where α_1 is an angle formed by the string lying to the left from the point $x = 0$ and the axis Ox. Analogously, from the right there will act the force $+ N \sin\alpha_2$ (Fig. 32.1). The string point corresponding to $x = 0$ will carry a zero mass and therefore the forces $- N \sin\alpha_1$ and $+ N \sin\alpha_2$ will equalize each other, i.e. $\alpha_1 = \alpha_2$. Therefore, in the point $x = 0$ the tangents to the strings from left and right will coincide. This will actually mean a fulfillment of the second relation in (32.4).
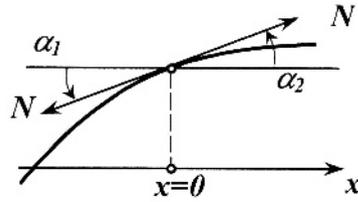
Fig. 32.1. The refraction and reflection of a running harmonic wave on the boundary $x = 0$ of different one-dimensional environments.

Accordingly, it may be assumed that

$$u_1(x,t) = A e^{i(\omega t - \omega x/a_1)} + A_2 e^{i(\omega_1 t + \omega_1 x/a_1)}$$

$$\tag{32.5}$$

$$u_2(x,t) = A_2 e^{i(\omega_2 t - \omega_2 x/a_2)}$$

where

A, A_1 and A_2 are amplitudes of the incident, reflected and refracted waves;

a_1 and a_2 are given in the equations (32.2) and (32.3); and

A_1, A_2 and ω_1, ω_2 are the unknown subjected to be determined.

From the conjugating conditions we derive

$$A e^{i\omega t} + A_1 e^{i\omega_1 t} = A_2 e^{i\omega_2 t}$$

$$-i\frac{\omega}{a_1} A e^{i\omega t} + i\frac{\omega_1}{a_1} A_1 e^{i\omega_1 t} = -i\frac{\omega_2}{a_2} A_2 e^{i\omega_2 t} \quad ,$$

where both ratios will be true for any t.

From these ratios, it follows that $\omega_1 = \omega_2 = \omega$. One may guess about it immediately because the conditions (32.4) must be met for all t. Taking $\omega_1 = \omega_2 = \omega$ we come to the conditions

$$A + A_1 = A_2 ,$$

$$-\frac{\omega}{a_1} A + \frac{\omega}{a_1} A_1 = -\frac{\omega}{a_2} A_2 \quad .$$

From which we then have

$$A_1 = \frac{1 - a_1/a_2}{1 + a_1/a_2} A$$

(32.6)

$$A_2 = \frac{2}{1 + a_1/a_2} A \; .$$

The needed solutions $u_1(x,t)$ and $u_2(x,t)$ have been thus found. Now, let us analyze a dependence of the amplitudes A_1 and A_2 of the reflected and refracted waves upon a_1 and a_2. These dependences $\dfrac{A_1}{A}$ and $\dfrac{A_2}{A}$ of $\dfrac{a_1}{a_2}$ are graphically shown in figure 32.2.
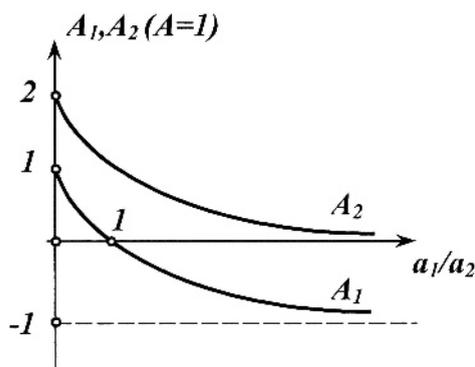
Fig. 32.2. The plots of the amplitudes A_1 and A_2 of the reflected and refracted harmonic waves.

Recall here that a_1^2 and a_2^2 in the wave equations (32.2) and (32.3) make up a ratio of a general tightening force to the string densities for $x < 0$ and $x > 0$. From the graph (Fig. 32.2) it follows that the homogeneous string with $\dfrac{a_1}{a_2} = 1$ will produce no reflected wave (as it must be), and the refracted wave will be a continuation of the incident wave. With the vanishingly small or very great density of the string within the segment $x > 0$, the ratio $\dfrac{a_1}{a_2}$ will be equal to 0

and, accordingly, to ∞. In the first case, the waves will be reflected with the same amplitude and phase; in the second, with the same amplitude, but with the reverse phase. As for the refracted wave amplitude, it will be either doubled respectively or will become equal to zero. These extreme cases may be regarded as a reflection from the free and the fixed end ($x = 0$) of the string for $x < 0$. In the first case, the incident wave will be reflected changing its direction only, whereas in the second case, it will be overturned (Fig. 32.3).

Fig. 32.3. The turnover phenomenon when the running harmonic wave is reflected from a "fastening" (no refracted wave is present).

33 Standing waves and oscillations of a bounded string

The Fourier method for studying the string oscillations caused by a concentrated impact. A spectrum and its relation to a sound shade.

In this chapter you will become acquainted with a new type of oscillations in distributed dynamical systems describing a continuum or electromagnetic fields – standing waves. They arise in 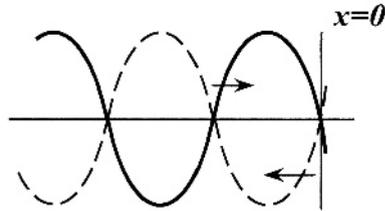bounded elastic bodies or volumes. A standing wave may also appear as a result of superposition (interference) of the travelling harmonic waves $Ae^{i(wt-kx)}$ and $Ae^{i(wt+kx)}$ that are running towards each other, since

$$Ae^{i(wt-kx)} + Ae^{i(wt+kx)} = 2Ae^{iwt}\cos kx \quad . \tag{33.1}$$

The wave (33.1) oscillates in each point x with one and the same frequency ω and phase, but its amplitude depends upon x in accordance with the law $2A\cos kx$. According to (33.1), a general form of a standing one-dimensional wave may be written as

$$u(x,t) = X(x)T(t) \quad , \tag{33.2}$$

where $X(x)$ is a function of x only; and $T(t)$ of the time t only.

In numerous cases, oscillations in a distributed system may be represented as a superposition of standing waves. As an example, let us study arbitrary oscillations of some bounded string with fixed terminal points and their special case, when these oscillations were born by some concentrated impact exerted on some small-length segment of the string.

A general mathematical model will be as follows. Given are the below wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \quad (0 \le x \le l) \tag{33.3}$$

and also the initial and boundary conditions for the function $u(x,t)$.

The initial conditions are

$$u(x,t)\big|_{t=0} = \varphi(x) \qquad \frac{\partial u}{\partial t}\big|_{t=0} = \psi(x) \tag{33.4}$$

and the boundary conditions are

$$u(x,t)\big|_{x=0} = 0 \ , \qquad u(x,t)\big|_{x=l} = 0 \ . \tag{33.5}$$

The method of solving this problem was first suggested by J.B.J. Fourier and was called his name. This method includes the following steps. First, the solution of the equation (33.3) with the boundary conditions (33.5) is found in the form of standing waves (33.2). The number of such solutions turns out countable. Second, the solution of the problem (33.3) – (33.5) is found as a superposition of the obtained solutions of the form (33.2). Here, the superposition may be chosen in such a way as to satisfy the initial conditions (33.4). The boundary conditions as well as the equation (33.3) itself will be satisfied automatically, since they are met by each component of the superposition.

Inserting (33.2) and (33.3) we find that

$$\frac{T''}{T} = a^2 \frac{X''}{X} \ .$$

Since the left-hand part of this equality is independent of x and the right-hand part independent of t, they will be equal to a constant, say, $-\lambda$. This will then lead us to the equations

$$T'' + \lambda T = 0 \qquad X'' + a^2 \lambda X = 0 \ . \tag{33.6}$$

We are interested only in the solutions of (33.6), for which the function $u(x,t)$ determined in accordance with (33.2) satisfies the boundary conditions (33.5). It is possible only for positive λ. For $\lambda > 0$, the general solutions of (33.6) will be of the form

$$T = \sin\sqrt{\lambda}t + B\cos\sqrt{\lambda}t \qquad X = C\sin a\sqrt{\lambda}x + D\cos a\sqrt{\lambda}x \ ;$$

therefore, we obtain

$$u(x,t) = TX =$$

$$= (A\sin\sqrt{\lambda}t + B\cos\sqrt{\lambda}t)(C\sin a\sqrt{\lambda}x + D\cos a\sqrt{\lambda}x) \ . \tag{33.7}$$

This solution must satisfy the boundary conditions (33.6). From them, it will follow that

$$D = 0 \qquad C\sin a\sqrt{\lambda}l = 0 \ . \tag{33.8}$$

To assume $C = 0$ here is impossible, because as a result we come to a trivial zero solution. Thus, from (33.8) the below will be obtained

$$\sqrt{\lambda} = \frac{\pi}{al} n \qquad (n = 1,2,3,...) \ .$$

Hence, we arrive at the below form of the equation

$$(A \sin \frac{\pi n}{al} t + B \cos \frac{\pi n}{al} t) \sin \frac{\pi n}{l} x \ . \tag{33.9}$$

These are standing waves with the discrete wave lengths $\dfrac{2l}{n}$ and the respective frequencies $\dfrac{\pi n}{al}$. In these waves, all their points oscillate synchronously with the time frequency $\omega_n = \dfrac{\pi n}{al}$ and the amplitude harmonically changeable (with x changing) proportionally to $\sin \dfrac{\pi n}{a} x$. Some standing waves for $n = 1,2,3$ are shown in figure 33.1.
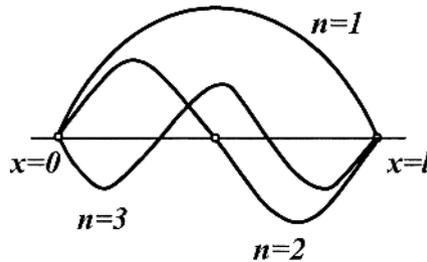
Fig. 33.1. The harmonic standing waves; **1** – of the basic tone; **2** and **3** – of the first and second harmonics.

Further, a solution for the initial problems (33.3), (33.4), (33.5) will be derived by us as an infinite sum of all standing waves (33.9)

$$u(x,t) = \sum \left(A_n \sin \frac{\pi n}{al} t + B_n \cos \frac{\pi n}{al} t \right) \sin \frac{\pi n}{l} x \ , \qquad (33.10)$$

where A_n and B_n are arbitrary so far. The initial conditions (33.4) will yield the conditions

$$\sum B_n \sin \frac{\pi n}{l} x = \varphi(x)$$

$$\qquad , \qquad (33.11)$$

$$\sum \frac{\pi n}{al} A_n \sin \frac{\pi n}{l} x = \psi(x)$$

which may be thought to be Fourier expansions of the functions $\varphi(x)$ and $\psi(x)$, respectively. Since for $n \neq m$ we obtain

$$\int_0^l \sin \frac{\pi n}{l} x \sin \frac{\pi m}{l} x dx = 0$$

and for $n = m$ the same integral is equal to $\dfrac{l}{2\pi}$, through multiplying each of the equalities (33.11) by $\sin \dfrac{\pi m}{l} x$ and by integrating in x from 0 to l we shall find that

$$B_m = \frac{2\pi}{l} \int_0^l \varphi(x) \sin \frac{\pi m}{l} x dx$$

$$\qquad (33.12)$$

$$A_m = \frac{2a}{m} \int_0^l \psi(x) \sin \frac{\pi m}{l} x dx \ .$$

In such a nice and simple way the general problem (33.3), (33.4), (33.5) is solved. The solution has been derived in the form of the superposition of a discrete sequence of standing waves. Each standing wave is characterized by the frequency $\omega_n = \dfrac{n\pi}{al}$ and by the length $\dfrac{2al}{n}$. Its form is described by the formula (33.9).

The frequencies ω_n constitute a frequency spectrum of possible oscillations in a bounded string. In it, the least frequency ω_1 represents a basic tone, the frequency ω_2 represents the first harmonic, ω_3 the second harmonic, etc. (Fig. 33.1). The string oscillations are described by a frequency spectrum and appropriate amplitudes.

As for a piano string, the basic tone of the sound and the sound colour, depending upon harmonics, are determined by a spectrum of its oscillations. With the basic tone retaining similar, the sound colour may be absolutely different. For example, a similar musical sound played on the piano, fiddle and guitar will sound in absolutely different ways. This will happen with similar basic tones. Therefore, a basic tone is determined by length, tightness and density of the string. As for the spectrum, it depends upon the excitation of oscillations, i.e. where, say, the string was stricken by the piano hammer, where and how a fiddle stick is touching a string and what a fiddle-stick is itself, i.e. whether it has been well rubbed with rosin or not, how it has been tightened and what it is made of. Certainly, the point is not only in what the oscillating spectrum of the string itself is, but also in how these oscillations are converted into sounds. We would like to put aside these complex questions; instead, first, we wish to consider only how the string oscillating spectrum depends upon the stricken segment ($x = \xi$). The Bechstein and Steinway pianos produce different sounds firstly due to the difference in hammer-striking segments.

In constructing a mathematical model for the oscillations of the initially fixed string, which has received a concentrated impact within a very small-length segment, thus, imparting this string the impulse p, we shall first start from the general model (33.3), (33.4), (33.5). In this general model, we need only to specify the shapes of the functions $\varphi(x)$ and $\psi(x)$.

The impact will be assumed to be instant and pointwise. Accordingly, during this instant impact no displacement of the string will occur and, hence, $\varphi(x) = 0$. Since the string, during the impact, accepts the impulse p, we obtain

$$\int_0^l \rho\psi(x)dx = p \quad , \tag{33.13}$$

where ρ is a constant density of the string; $\psi(x)$ is a string velocity in the point ξ gained upon the impact. The impact is supposed to be concentrated in the vanishingly small neighbourhood of the point $x = \xi$. Beyond this vanishingly small neighbourhood, the velocity $\psi(x)$, immediately after the impact, will be, therefore, equal to zero. Hence, the condition (33.13) may be written as

$$\int_{\xi-\varepsilon}^{\xi+\varepsilon} \psi(x)dx = \frac{p}{\rho} \quad (\varepsilon > 0) \; ; \tag{33.14}$$

here, we have $\psi(x) = 0$ for $x \notin [x - \varepsilon, x + \varepsilon]$, where ε is vanishingly small. With the string segment (being stricken by the hammer) decreasing infinitely, the function $\psi(x)$ will infinitely increase within this small-length segment, but the condition (33.13) will remain satisfied all the time. So, we shall obtain $\varphi(x) = 0$, while the function $\psi(x)$ will everywhere remain equal to zero, except the vanishingly small neighbourhood of the point $x = \xi$, and will satisfy the condition (33.13). Let this information be employed for calculating the constant A_m and B_m according to the formulae (33.12). According to the first of them, we have $B_m = 0$; according to the second, we shall obtain

$$A_m = \frac{2a}{m} \int_{\xi-\varepsilon}^{\xi+\varepsilon} \psi(x) \sin\frac{\pi m}{l} x\, d\, x \approx \frac{2ap}{m\rho} \sin\frac{\pi m}{l} \xi \quad , \tag{33.15}$$

where the formulae will be the more exact, the less $\varepsilon > 0$ is.

The application of the general solution (33.10) to our search for specific values of the coefficients A_n and B_n ($n = 1,2,...$) may be formalized and simplified via the already known generalized δ-function. An instant impact by the hammer will impart the string, and more exactly, its point $x = \xi$ the impulse p. The mathematical model for the concentrated impulse p in this point will be $p\delta(x - \xi)$, i.e.

$$\psi(x) = \frac{p}{\rho}\delta(x - \xi) \quad . \tag{33.16}$$

By substituting (33.16) into the formulae (33.12) for A_m, we formally arrive at (33.15), where the approximate sign is replaced by equality.

In both cases we arrive at the situation when the oscillations of the string, receiving the instantaneous impulse p concentrated in the point $x = \xi$, will be of the form

$$u(x,t) = \sum \frac{2ap}{n\rho} \sin\frac{\pi n}{l} \xi \sin\frac{\pi n}{l} x \sin\frac{\pi n}{al} t \quad . \tag{33.17}$$

In accordance with (33.17), the maximal amplitudes of the basic frequency $\dfrac{\pi}{al}$

and of harmonics $\dfrac{2\pi}{al}$, $\dfrac{3\pi}{al}$, will be respectively equal to

$$\frac{2ap}{\rho}\sin\frac{\pi}{l}\xi \quad , \quad \frac{ap}{\rho}\sin\frac{2\pi}{l}\xi \quad , \quad \frac{2ap}{3\rho}\sin\frac{3\pi}{l}\xi \quad .$$

This is vividly represented below in figure 33.2 for the basic tone $n = 1$ and the first three harmonics, $n = 2$, $n = 3$ and $n = 4$.
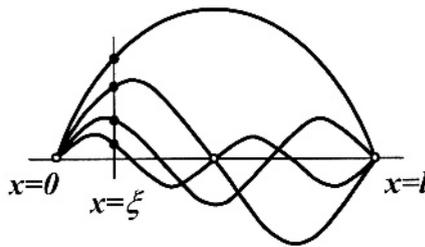
Fig. 33.2. The graphically illustrated dependence of the string amplitude upon the order of harmonics at the impulse impact in the point $x = \xi$.

In particular, figure 33.2 demonstrates a basic tone to be better represented by striking upon the string centre, but here no second and fourth harmonics are available. In order to have some first harmonics well represented, the strike must be exerted not far from the string terminals. This is so done on the piano. Besides, the less this distance to the terminals is, the more harmonics will arise. At the Steinway piano, this distance is less than that at the Bechstein piano; that is why its sound is "louder", but, instead, the Bechstein sound is more mellow (or richer) and "more soft".

Sounds produced by a fiddle string and a piano string of the similar basic tone will differ enormously; this difference is also explained by differences in the numbers of harmonics born.

A human ear is constructed so as to perceive a sound in the form of its spectrum. Each musical sound like do, re, mi, … possesses a spectrum consisting of a basic tone and harmonics. Musical chords have a spectrum composed of the spectrums of sounds held in chords. Noise has numerous spectral frequencies; such a spectrum is called continuous, in contrast to a discrete spectrum when frequencies are not so many.

Why some certain sequences of sounds bring us very strong emotions is, perhaps, one of paramount mysteries in nature. Though, some necessary requirements to these sounds are already known and they constitute a subject of the musical theory.

34 Microparticles

The particularities of the microworld. A necessity to revise the habitual concepts of classic physics. The formalism of quantum physics and examples of its application. The notions of a quantum state, an operator of its changes and its relation with measured and observed physic magnitudes. The behaviour of a free quantum particle in the potential well and in the radially symmetric electrical field of a charge.

Until now, we were discussing various mathematical models of our environmental world, at times the world is misunderstandable and mysterious, but it is the world we are living in. Besides it, there exist some other worlds hidden from us and inaccessible to; in this sense, they are existing in our imagination. Such is the microworld, a world of molecules, atoms and elementary microparticles. This is the world which a human guessed out long ago, but which became accessible for studying only in the last century, upon appearance of new neoclassic physics – quantum mechanics.

The mathematical models for the objects of the microworld – of quantum mechanics – are more mathematical than those of classic physics; this mathematical nature of them is not only far from our immediate every-day perceptions and intuition but is even rejected by them. These disagreements also took place earlier, in the times when the laws of mechanics and electrodynamics were formulated. In mechanics, the every-day observations seemingly should undoubtedly prove that a movement is caused by forces, whereas these forces are only responsible for the changes of this movement. More complicated electrical phenomena seemed to be born by electrical charges and currents; the idea about the existence of electromagnetic waves was absolutely unacceptable, since no charges and currents are here present. As for quantum mechanics, such inacceptibility was expressed even more fiercely. Never and in no way the microworld was accepted by our sensible organs; it turned out to be unimaginably different from our usual macroworld. At the same time, in the end of all ends, it is this microworld that lies in the foundation of all the manifestations of the macroworld. Therefore, the microworld can be exposed via man-made devices and with a use of special experimental conditions, viz. through observing the atom radiation and absorption spectrums, radioactivity, diffraction of electrons, photoelectrical emission (photoeffect) and other phenomena.

New data obtained in this way and its theoretical interpretation contradicted not only multi-century experience but also the existing theoretical conceptions about the environmental physical world. Later, these conceptions were named classic

physics. This, certainly, covers not only the microworld but also the other sections of the new physics of our century – the physics of the velocities close to the light velocity, enormous distances of millions of the light years, gigantic masses and great densities, superhigh temperatures, unbelievable magnetic fields, etc.

Let us now show some vivid distinctions in manifestations of the microworld; these distinctions declare the necessity to cardinally revise the current conceptions in classic physics and to create principally new mathematical models.

Long ago, perhaps from ancient Greeks, the mankind guessed that the material world is composed of the smallest microparticles (i.e. atoms) and all the events occurring in it are generated by their movements and interactions. These microparticles, though extremely small, were allotted with the properties similar to those existent in ordinary bodies, i.e. the microparticles can move, clash with each other and catch on each other, etc. The initially founded kinetic theory of gases and the then conceptions concerning fluid and solid bodies and various chemical transformations did not contradict these naive interpretations. Though, on the threshold of 20-th century these discrepancies started to arise as completely incompatible with classic physics.

For example, it was revealed that in some cases it is useless to talk about travelling trajectories of the microparticles, that any conceptions on the presence of such trajectories vividly contradict experimental results. In the microworld, the classic law of energy conservation can be not observed;also, an electrical charge (i.e. an electron) oscillating in an atom cannot, for some reason, emit (though it should due to the classic electrodynamics) and if it emits, then it emits only wave packages of certain frequencies and energies. You have evidently heard a lot about these and other peculiarities of the microworld, viz. about the electrons diffraction, radioactive decay, atom-emitting and atom-absorbing spectrums, quantum generators-lasers, photoeffect, nuclear reactions and reactors, atom bomb, the excessively great energy being emitted by the Sun, unexplainable regularities in the Mendeleev's table of chemical elements, and, possibly, about many other things.

To prove the groundlessness of scientific interpretations and theories, there is no need to obtain numerous discrepencies; instead, a lot of correspondences will be needed, whereas a single discrepency will be sufficient. One of the discrepencies will be discussed below. This is a diffraction of electrons.

According to the initial interpretation, an electron is the tiniest charged microparticle with very small charge and mass. Its travelling can be deviated by a magnetic field, in full accordance with the Lorentz force acting upon the electrical charge moving in the magnetic field. An electron can be detected through its striking a photoplate; on this photoplate when developed one can see a black point. The electron's travel within the electrical field is completely governed by the Newton laws. Now, let us put a two-hole barrier on the way of the electronic beam of the constant intensity and behind the barrier a screen, on which the strikes by electrons will be registered. It is natural to suppose that the screen will keep only the traces of the electrons that have passed through the holes. The remaining electrons will be blocked by the barrier and will not get to the screen. In this experiment, some electrons are supposed to fly through one hole and the other

electrons through the other. Having flown through a hole they approach the screen and leave their traces on it. From the above said it is evident that if the electrons are allowed to pass for some time T through a single hole (with another closed) and then during the same time to pass through another hole (the first closed), then this procedure should bring us one and the same screen hitting picture – as if the electrons were allowed to simultaneously pass through the two holes during the same time T. The difference may be observed only in that that the electrons, having passed through the two holes, are interacting each other. Though, this Coulomb interaction is very small. Besides, this kind of interaction, as any other, is excluded, if a time gap between electrons (i.e. the electronic beam is of rather small intensity) exceeds the time of their travel from the barrier to the screen. For the described case of a separate travel of electrons, first, through one hole and, second, through another or at once through the two holes but during the time twice less, we should obtain similar pictures of the electron hits. It should be so, if electrons fly along the definite trajectories as small particles usually do. What will then this experiment bring us? It will bring us the absolutely unexpected things shown in figure 34.1. The pictures obtained are absolutely different, absolutely not similar. The second picture looks like the picture showing the flight through the two holes not by electronic microparticles but light (i.e. the electromagnetic waves), with its diffraction picture being so characteristic for light. So, what is then a flying electron? Is it a flying microparticle or a wave having a specific diffracting phenomenon? Sooner, this is neither this nor that. Electrons are microparticles; these are objects of the nature unknown for us. It is true that in one case they reveal themselves like particles, and in the other case behave like waves, but in some other cases they are neither waves nor particles, but microparticles.

The history of synthesizing the mathematical models for the microworld constitute a complicated, inconsistent and intricate process being very far from a logical inference. This history is full of numerous fantastic hypotheses, unimaginable conjectures and aha factors, which later were subjected to filtration, specification and mutual adjustments. Even to physicists, the quantum mechanics avoids to suggest a somehow complete description of this process. For mathematicians, to avoid this complete description is more natural, starting it immediately from the description of a final mathematical model, from its formalism, which will be then illustrated by examples.

The application of unusual mathematical models of the microworld demands new interpretations, concepts and new intuition. All this is developed on the basis of comprehending the mathematical model by penetrating into its logic and harmony issuing from purely mathematical considerations. This way of gaining truth is also accessible for a human intelligence and this very capability of a human intelligence manifests its power and strength, thus bringing about a gigantic leap in the natural science of the last centuries.

These unusual mathematical models appeared due to the inability of classic physics to interpret observable phenomena of the microworld. Truthfulness of these models is proved by the qualitative and quantitative coincidences of the conclusions and results (obtained with their help) with various observational and experimental results (diffraction of electrons on crystal lattices, spectrums of

atoms and molecules, photoeffect, spectrums of equilibrium emission, chemical properties of substances and the Mendeleev's table, physical phenomena of solids and semiconductors, quantum generators and many other things). Here, it should be also said that, nevertheless, the unusual quantum models transfer to classic ones, when the values of the microworld parameters transfer to the values of the macroworld parameters. This fact makes the compatibility between the laws of the macroworld and the microworld possible.

The model of the microworld written below is better to be accepted as the game "Life" by J. Conway you have read earlier. Though now, subject to changes are not the counters on a chess board but some function ψ and the physical magnitudes found with its help. This approach is supposed to help you to easily accept the rules of this game and to master the game, if you are laborious. Upon your training in how to play this game, the microworld model will be familiar to you and, with time, understandable.

Whereas in classic physics the basic mathematical tools are differential equations, in quantum mechanics there are additionally exploited some absolutely different tools, viz. the mathematical techniques for linear operators. Mathematicians had created these tools long before they became needed by physicists; these tools were brought to life by the internal mathematical needs and internal motives of mathematics itself. You are familiar with this theory of linear operators from the academic course on functional analysis, but, of course, for you to deal with linear operators is less convenient than with differential equations.

Mathematical models in classic physics and quantum mechanics differ greatly, but the principle of determinism – in its sufficiently wide interpretation –has retained in them. As well as in classic physics, objects of the microworld are described through a state; though a state is now represented as some complex function ψ of the microparticle coordinates and, possibly, of their spines. In this case, a state of the microsystem is understood as this very complex function ψ of the listed variables, and not the values of the variables included into it.

The function ψ changes in time in accordance with the Schrödinger differential equation

$$ih\frac{\partial \psi}{\partial t} = H\psi \ ,$$ (34.1)

where H is an operator of the system under study and h is the so-called Plank constant equal to the very small magnitude 1.05×10^{-27} erg/sec.

The function ψ, as was already noted, is a function of the vector $\overline{x}$, coordinates of microparticles, spines and of the time t. The operator H is the so-called Hamilton operator over the vector $\overline{x}$. This operator is so called, since it is calculated by some rules from the Hamilton function of the microsystem considered as a classic mechanical system.

This function as a description of a state was also employed by classic mechanics for describing a continuum and in electrodynamics for describing

electrical and magnetic fields. It is not surprising that this function satisfies some equation in partial derivatives. Also there is no surprise in the function ψ being complex: this happens when the theory of functions of complex variables
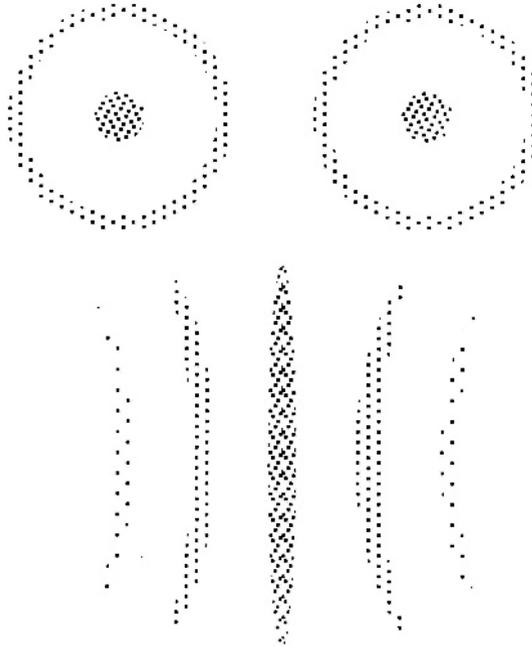
Fig. 34.1. Difraction of electrons during their sequential and simultaneous travel through the two holes: the black points correspond to 'impacts' of separate electrons against a photoemulsional layer.

is applied to the problems of hydrodynamics and electrodynamics. Though, the essential difference lies not in this. It lies in how through this state function ψ one can find the possible values of the observed physical magnitudes characterizing the microsystem under study, and how the function ψ is connected with the observable physical magnitudes, i.e. with coordinates, impulses, moments, energies, spines, frequencies, wave lengths, etc.

34.1.
Mathematical formalism in quantum mechanics

In quantum mechanics, it is postulated that to each physical magnitude there corresponds a Hermitian, i.e. a complex self-conjugate linear operator, whose eigenvalues are single possible values of this physical magnitude.

We recall here that the operator L is Hermitian if for any functions ψ_1 and ψ_2 the below takes place (a star stands for the complex conjugation):

$$\int \psi_1 * L \psi_2 \, d\overline{x} = \int \psi_2 (L \psi_1) * d\overline{x} \quad . \tag{34.2}$$

Accordingly, the eigenvalues λ of any Hermitian operator L are real. The numbers λ and the associated eigenfunction φ are found from the condition

$$L\varphi = \lambda\varphi \qquad (\varphi \neq 0) \ . \tag{34.3}$$

The set of all eigenfunctions of the Hermitian operator is complete and the eigenfunctions φ_1 and φ_2 corresponding to different eigenvalues $\lambda_1 \neq \lambda_2$ are orthogonal

$$\int \varphi_1 * \varphi_2 \, d\overline{x} = 0 \quad . \tag{34.4}$$

Thus, as was said above, to each physical magnitude l there is put into correspondence, according to the below rules, the Hermitian operator L. Possible values of the physical magnitude l will be its eigenvalues only. These values may be observed during measurements. Besides, in identical conditions the measuring results may differ and only their average value (mathematical expectation) carries an actual macroscopic physical sense. This average value $\langle l \rangle$, according to the quantum mechanical formalism, is calculated with a use of the state ψ of the quantum system as follows

$$\langle l \rangle = \int \psi * L \psi d\overline{x} \quad , \tag{34.5}$$

where the state function ψ is supposed to be normalized, i.e.

$$\int \psi * \psi d\overline{x} = 1 \quad .$$

The above said completes a formal description of the mathematical model for the quantum system. The mathematical model includes a concept of a state (the function ψ); the Schrödinger equation describing a change of the state with time; and some formalism for calculating the possible values of any physical magnitude

and its statistically averaged value through the state function and the operators of these physical variables, in accordance with (34.5).

To actually create a mathematical model for the quantum system one needs a writing of the function ψ and H, the Schrödinger equation, and also the knowledge of the operators relative to our physical variables.

Within the above said, studying a mathematical model includes the solution of the Schrödinger equation, perhaps, with a use of some supplementary information concerning the function ψ, finding eigenvalues and eigenfunctions of the operators of the necessary physical magnitudes and calculating through the derived function ψ their average values, by the formula (34.5).

Let us show the operators which correspond to the coordinates x, y and z, the impulses p_x, p_y and p_z, the kinetic T and potential V energies, and also to the Hamilton operator H, for the particle of the mass m in the potential field $V(x, y, z)$ (the spine-free case).

The operators corresponding to the coordinates x, y and z are multiplications by x, y and, respectively, z. Thus, the operator (relative to the coordinate x) of the function ψ will be $x\psi$, and for y and z will be $y\psi$ and $z\psi$.

The operator of ψ corresponding to the function $V(x, y, z)$ will be $V(x, y, z)\psi$.

To the impulses p_x, p_y and p_z there will correspond the differential operators

$$-ih\frac{\partial}{\partial x} \quad , \quad -ih\frac{\partial}{\partial y} \quad , \quad -ih\frac{\partial}{\partial z} \quad . \tag{34.6}$$

To the impulse vector $p(p_x, p_y, p_z)$ there will correspond the vector-operator

$$-ih\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) = -ih\nabla \quad , \tag{34.7}$$

where ∇ is a sign for the gradient operator.

To the kinetic energy T of the particle of the mass m there will correspond the operator

$$-\frac{h^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) = -\frac{h^2}{2m}\nabla^2 = -\frac{h^2}{2m}\Delta \quad , \tag{34.8}$$

where Δ is a sign for the Laplace operator.

The potential energy $V(x, y, z)$ is a function of the coordinates and, therefore, the relative operator will be a multiplication by $V(x, y, z)$.

Now, only the Hamilton operator for the m-mass particle is left to be indicated. According to the so-called correspondence principle between the classic Hamilton function

$$\frac{1}{2m}\left(p_x^2 + p_y^2 + p_z^2\right) + V(x, y, z)$$

and the Hamilton quantum operator, this operator is written as

$$H = -\frac{h^2}{2m}\Delta + V(x, y, z) \ . \tag{34.9}$$

Now, we can write the Schrödinger equation for the microparticle of the mass m (spine-free). The state function will be

$$\psi = \psi(\bar{x}, t) = \psi(x, y, z, t) \ ; \tag{34.10}$$

the Schrödinger equation, due to the above (34.1), will have the form

$$ih\frac{\partial \psi}{\partial t} = -\frac{h^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + V(x, y, z)\psi \ . \tag{34.11}$$

Prior to proceeding to some specific examples, let us note here about two important corollaries following from the quantum system mathematical model described above. One corollary pertains to the possibility of indicating not only values of this or that physical magnitude and not only its average value but also the probabilities of these values.

In order to expose it, let us decompose the function ψ in (34.5) into the normalized eigenfunctions of the operator L. For the sake of simplicity, let the spectrum of the eigenvalues of the operator L be discrete. Then, the decomposition of the normalized function ψ into the eigenfunctions $\varphi_1, \varphi_2, \ldots$ with the eigenvalues $l_1, l_2, \ldots$, which are the possible values of the magnitude l, will be of the form

$$\psi = c_1\varphi_1 + c_2\varphi_2 + \ldots, \tag{34.12}$$

where

$$c_s = \int \psi(\bar{x}) \varphi_s^*(\bar{x}) d\bar{x} \ .$$

By inserting (34.12) into (34.5), we shall immediately find that

$$\langle l \rangle = \sum |c_s|^2 l_s \quad , \tag{34.13}$$

where, due to the normability of the function ψ, we shall obtain

$$\sum |c_s|^2 = 1.$$

As it is known, the mathematical expectation of the magnitude l, which can take the values l_1, l_2, ... with the probabilities p_1, p_2, ..., will be equal to

$$Ml = \sum l_s p_s \quad . \tag{34.14}$$

Comparing (34.14) against (34.13) we shall find that the probability p_s of taking the value l_s is equal to

$$p_s = |c_s|^2 \quad . \tag{34.15}$$

In particular, from here it follows that if the function ψ coincides with one of the eigenfunctions of the operator L, let it be with φ_s, then $p_s = 1$. Therefore, in this case the physical magnitude may take only the single value l_s and the probabilities of other values l_j ($j \neq s$) will be equal to zero.

Thus, in a general case of the function ψ the magnitude l may take any value from the values l_1, l_2, ... with the probabilities, respectively, p_1, p_2. This does not exclude the possibility for the magnitude l to take in specific cases (i.e. specific types of the function ψ) only some values or even a single value. Only in the last case, a result of measuring the magnitude l may be predicted.

This is the first important conclusion from the mathematical model of the quantum system stated.

Let us demonstrate it in a simple example touching the values of the coordinates x, y and z of a microparticle. To these magnitudes there will correspond the operators of the multiplication by x, y and z. Thus, due to the formula (34.5), we have

$$\langle x \rangle = \int \psi * x \psi dx dy dz = \int x |\psi|^2 dx dy dz =$$

$$= \int x \left(\int |\psi|^2 dy dz \right) dx \quad ,$$

and the analogous formulae for $\langle y \rangle$ and $\langle z \rangle$. From it, it will immediately follow that

$$\iint |\psi|^2 \, dydz$$

is a probability density for the values of the coordinate x, and $|\psi|^2$ is a probability density of x, y and z.

In this way there have been found the following statistical sense of the quantum state function ψ: the square of its module $|\psi|^2$ is a probability density for the coordinates to take the values x, y and z.

However, according to the above, the function ψ statistically describes not only the values of the coordinates x, y and z but also any other physical magnitudes whose operators are known.

The second conclusion possible to be derived from the definition of the mathematical model relates to the so-called Geizenberg indetermining principle manifesting the possible results of the simultaneous measurements of the physical magnitudes (l and k with the operators L and K) and stating that a single-valued result of measuring the two physical magnitudes is only possible if the operators L and K are commutative, i.e. when

$$LK - KL = 0.$$

Otherwise, either the two magnitudes, l and k, have no unique values or if one is unique, then the other is indefinite. More fully, for the dispersions of the magnitudes l and k there will occur the inequality

$$DlDk \geq \frac{1}{4} \int \psi * \left(\frac{LK - KL}{i} \right)^2 \psi d\overline{x} \quad . \qquad (34.16)$$

This important inequality needs some explanation. In it, l and k are arbitrary physical measurable magnitudes, L and K being their respective operators. This inequality is a corollary of the trivial inequality

$$\int \left| L\psi + \frac{\lambda}{i} K\psi \right|^2 d\overline{x} \geq 0$$

retaining true for any real λ.

In the above formula (34.5), there have been determined the average values of the physical magnitudes l and k. Their dispersions may be analogously determined as average values of the magnitudes $\left(l - \langle l \rangle \right)^2$ and $\left(k - \langle k \rangle \right)^2$. Thus, we have

$$Dl = \int \psi * \left(L - \langle l \rangle \right)^2 \psi d\overline{x} \quad ,$$

where the function $\psi(\overline{x}, t)$, being a state of the microparticle, is supposed, as well as in the formula (34.5), to be normalized.

For example, the operators of the coordinate x and of the impulse p_x are not commuted, since

$$\left(-ih\frac{\partial}{\partial x}\right)x + xih\frac{\partial}{\partial x} =$$

$$= -ih - ihx\frac{\partial}{\partial x} + xih\frac{\partial}{\partial x} = -ih \quad .$$

Therefore, due to the formula (34.16), we shall obtain

$$DxDp_x \geq \frac{h^2}{4} \quad , \tag{34.17}$$

i.e. a product of the dispersions of the measurement results of the coordinate x and the associated impulse p_x cannot be less than $h^2/4$.

This significantly differs from what occurs in the classic mechanics, where, on the contrary, to determine a state needs to have the coordinate x and the impulse p_x assigned. As for the quantum mechanics, they cannot be simultaneously assigned in principal. This is possible only in the case of the commutativity of the respective operators. For example, since the operators of the coordinates x, y and z and also the operators of the impulses p_x, p_y and p_z are commuted, it becomes possible to simultaneously and uniquely measure only all the coordinates or only all the impulses. But if the three coordinates or the three impulses are determined simultaneously, then all the impulses and accordingly all the coordinates will be not determined.

Thus, a simultaneous determination of all the physical magnitudes, habitually being done in classic mechanics, is not performed in quantum mechanics. Besides, we recall here that except this substantial difference with classic science there exists some other substantial difference mentioned above. Namely, physical magnitudes are discrete and are of the quantum nature, because the values of a physical magnitude are described by the spectrum of the eigenvalues of its operator; and this spectrum may be discrete.

Now, we start describing a travel of some free quantum particle within some simple one-dimensional fields. Also will be considered a harmonic oscillator and a quantum variant of the Newton problem – an atom of hydrogen and a hydrogen-like ionized atom of helium or lithium.

In all these cases, the Schrödinger equation will, due to (34.10), be of the form

$$ih\frac{\partial\psi}{\partial t}=-\frac{h^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}\right)+V(x,y,z)\psi \quad. \tag{34.18}$$

Let us find its solution as

$$\psi(x,y,z,t)=A(t)\overline{\psi}(x,y,z) \quad. \tag{34.19}$$

Upon inserting (34.19) into (34.18) and separating the variables t and x, y, z, we get

$$\frac{ih}{A}\frac{\partial A}{\partial t}=\frac{1}{\overline{\psi}}\left\{-\frac{h^2}{2m}\Delta\overline{\psi}+V\overline{\psi}\right\}=E \quad, \tag{34.20}$$

where E is a constant. Thus, we have derived two equations, the first with respect to $A(t)$ and the second to $\overline{\psi}(x,y,z)$. The first equation is solved easily and its solution is of the form

$$A(t)=e^{-\frac{iE}{h}t} \quad. \tag{34.21}$$

To solve the second equation

$$-\frac{h^2}{2m}\Delta\overline{\psi}+V\overline{\psi}=E\overline{\psi} \quad, \tag{34.22}$$

we shall need to specify the function V. If we have managed to derive the solution of the equation (34.22) satisfying the natural conditions imposed upon the state function ψ (continuity, boundedness and some other conditions following from the problem statement), then the solution (more exactly, one of the solutions) of the initial equation (34.20) will be

$$\psi(x,y,z,t)=e^{-\frac{iE}{h}t}\overline{\psi}(x,y,z) \quad. \tag{34.23}$$

This solution ψ at any t satisfies the equation (34.22) (if to substitute ψ for $\overline{\psi}$) and, therefore, is an eigenfunction of the energy operator

$$-\frac{h^2}{2m}\Delta+V \quad. \tag{34.24}$$

Hence, in the state ψ determined, the energy of the system will have the definite value E, since E, due to (34.20), is an eigenvalue of the energy operator respective to the eigenfunction ψ.

The quantum system states respective to the certain values of the energy E are called stationary. Thus, ψ is a stationary state with the energy E. Note here that

$$|\psi|^2 = |\overline{\psi}|^2 \quad , \quad \left(\left| e^{-\frac{iE}{h}t} \right| = 1 \right) \; ; \qquad (34.25)$$

therefore, the probability densities of coordinate values in a stationary state are time-independent (that is why this state was called stationary).

The quantum system satisfies the superposition principle. This follows from the linearity of the Schrödinger equation with respect to the state function. Therefore, a superposition of the stationary states will be also the state satisfying the Schrödinger equation. Also valid will be the reverse: any admissible solution of the Schrödinger equation can be represented by a superposition of the stationary states. Thus, a study of possible states of a quantum system is, in this sense, reduced to a search for all its stationary states.

A spectrum of the energy operator eigenvalues may be discrete, continuous or mixed.

To each non-degenerate value of the spectrum there corresponds a unique normalized eigenfunction. With the eigenvalue being degenerate, there may be some and even infinitely many eigenfunctions. However, in any case all the eigenfunctions may be orthogonalized and normalized.

The condition of orthogonality of the two different eigenfunctions ψ_1 and ψ_2 corresponding to a discrete or a continuous spectrum is that

$$\int \psi_1 \psi_2 * d\overline{x} = 0 \; . \qquad (34.26)$$

For the eigenvalue corresponding to a discrete or continuous spectrum, the normalizing condition is written differently.

For the case of a discrete spectrum, the normalizing condition for the function ψ_s corresponding to the eigenvalue E_s has been already written and is of the form

$$\int \psi_s \psi_s * d\overline{x} = 1 \; . \qquad (34.27)$$

For a continuous spectrum, whose eigenfunctions are dependent upon the parameter E, i.e. $\psi = \psi_E$, the normalizing condition may be written as

$$\int \psi_E \psi_{E^1} * d\overline{x} = \delta (E - E') \; . \qquad (34.28)$$

This notation implies that for $E \neq E'$ we have

$$\int \psi_E \psi_{E^1} * d\overline{x} = 0 \qquad (34.29)$$

that corresponds to orthogonality. For $E = E'$, the integral (34.29) turns to infinity, but so that there still exists an integral from it by E being equal to unity, i.e.

$$\int \psi_E \, \psi_{E^1} * \, d\overline{x} = 1 \ . \tag{34.30}$$

Here, integrating over E is performed along the continuous portion of the spectrum.

In order to make the notations (34.26 – 34.30) close, we shall write them as

$$\int \psi_i \psi_j * d\overline{x} = \delta_{ij} \qquad \int \psi_E \psi_{E^1} d\overline{x} = \delta(E - E') \ , \tag{34.31}$$

where the symbols δ_{ij} and $\delta(E - E')$ are nullified for $i \neq j$ and $E \neq E'$, respectively; for $i = j$ and $E = E'$, the first symbol is equal to 1 and for the second the below integral is equal to 1

$$\int \delta(E - E') dE = 1 \ .$$

We recall here that the symbol δ_{ij} (the function from the integer values i and j) was introduced by L. Kronecker and is called the Kronecker symbol; the symbol $\delta(E - E')$ was introduced into theoretical physics by P. Dirac and is called the Dirac function. Initially, this function was employed by O. Heaviside in the operational calculus created by him and not understandable by his contemporaries.

Now, upon this prolonged but necessary introduction, let us come, at last, to considering the quantum systems mentioned above.

A microparticle (for example, an electron) is described by operators of coordinates, impulses, impulse moments, energy and spine; besides, a microparticle has a mass and a charge. If a micro particle has a charge, it is acted upon by the electrical field, and a travelling micriparticle by the magnetic field as well. Such a rich variety of descriptions and possible actions create a gigantic variety of behaviours of the microparticle. Below, you will see only some of them; they are simple and bring about no mathematical difficulties; nevertheless, they are sufficiently representative to make you feel the difference between the microworld and our usual macroworld. In the below narration, the examples will pertain only to the microparticle, for which a spine is absent or not significant and no magnetic field available.

34.2
Free microparticle

For a free microparticle we have $V = 0$; therefore, the equation (34.22) will assume the form

$$-\frac{h^2}{2m}\Delta\overline{\psi} = E\overline{\psi} \quad . \tag{34.32}$$

By examining it immediately we may get assured that it has a solution of the form

$$\overline{\psi} = exp\frac{i}{h}\left(p_x x + p_y y + p_z z\right) , \tag{34.33}$$

where p_x, p_y, p_z are arbitrary constants meeting the condition

$$\frac{1}{2m}\left(p_x{}^2 + p_y{}^2 + p_z{}^2\right) = E \quad . \tag{34.34}$$

Therefore, due to (34.23), the related solution of the Schrödinger equation for a free microparticle will be written as

$$\psi = \exp\frac{i}{h}\left(p_x x + p_y y + p_z z - Et\right) . \tag{34.35}$$

This is a famous function. It was called der Broil wave and from it a wave interpretation of microparticles started. In order to understand this, we note here that, on the one hand, the function ψ describes the complex plane wave of the frequency $\omega = \dfrac{E}{h}$ and of the wave number $k = \dfrac{|\overline{p}|}{h} = \sqrt{2m\dfrac{E}{h}}$ propagating along the vector $\overline{p}$, and, on the other hand, this function represents a microparticle of the energy E and with the impulse $\overline{p}$. The last follows from the function ψ being an eigenfunction of the impulse operators

$$-ih\frac{\partial}{\partial x} \quad , \quad -ih\frac{\partial}{\partial y} \quad , \quad -ih\frac{\partial}{\partial z}$$

and simultaneously of the kinetic energy operator $-h^2\Delta/2m$ with the characteristic numbers p_x, p_y, p_z and E, respectively. According to the idea of the state function ψ, the probability density of its location is proportional to

$\left|\psi\right|^{2}$, which, as seen from (34.35), for all x, y, z and t will be equal to unity. In other words, der Broil wave as a state function will describe the microparticle of some certain impulse and energy, but this microparticle is located within the entire space uniformly, strictly in accordance with the indetermining principle.

34.3
Microparticle in a potential well

Let the microparticle be held in a rectangular potential well with infinite walls. In this case, in the Schrödinger equation

$$ih\frac{\partial\psi}{\partial t} = -\frac{h^{2}}{2m}\Delta\psi + V\psi \tag{34.36}$$

the function V will depend only upon the single coordinate x :

$$V = \begin{cases} 0 & for \quad x \in (0,a) \\ \\ \infty & for \quad x \notin (0,a) \end{cases} ; \tag{34.37}$$

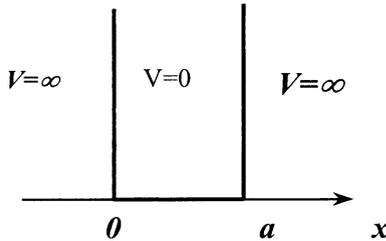its plot will be of the form shown in figure 34.2.

Fig. 34.2. The behaviour of electrons in the infinitely deep potential well.

The solution of the Schrödinger equation may be found in the form

$$\psi = \psi_{1}(x)\psi_{2}(y,z) . \tag{34.38}$$

Upon inserting (34.38) into (34.36), the equation obtained may be written as

$$\frac{1}{\psi_1}\left(ih\frac{\partial\psi_1}{\partial t}+\frac{h^2}{2m}\frac{\partial^2\psi_1}{\partial x^2}-V\psi_1\right)=$$

$$=-\frac{1}{\psi_2}\left(ih\frac{\partial\psi_2}{\partial t}+\frac{h^2}{2m}\left(\frac{\partial^2\psi_2}{\partial y^2}+\frac{\partial^2\psi_2}{\partial z^2}\right)\right) ,$$

(34.39)

with separable variables. This allows to reduce the below consideration to the search for the solution of the one-dimensional Schrödinger equation

$$ih\frac{\partial\psi_1}{\partial t}=-\frac{h^2}{2m}\frac{\partial^2\psi_1}{\partial x^2}+V\psi_1 . $$

(34.40)

This reduction technique will be also employed in the subsequent problems.

For $x \notin (0,a)$, it follows from the equation (34.40) that $\psi_1 = 0$, since here we have $V(x) = \infty$. For $x \in (0,a)$, it will assume the form of the below one-dimensional equation for a free microparticle (here, ψ_1 is substituted by ψ)

$$ih\frac{\partial\psi}{\partial t}=-\frac{h^2}{2m}\frac{\partial^2\psi}{\partial x^2} ,$$

whose solution is easily derived and written as a superposition of the functions of the form

$$\exp i\left(\frac{p_x}{h}x-\frac{E}{h}t\right) \quad , \quad E=\frac{p_x^2}{2m} ,$$

which should satisfy the edge conditions, i.e. to be equal to zero for $x = 0$ and $x = a$. These edge conditions will be satisfied only by the functions of the form

$$c \exp i\left(\frac{p_x}{h}x-\frac{E}{h}t\right)-c \exp i\left(-\frac{p_x}{h}x-\frac{E}{h}t\right)=$$

(34.41)

$$=2ic \exp\left(-\frac{iEh}{t}\right)\sin\frac{p_x}{h}x ,$$

for which

$$\frac{p_x a}{h}=\pi k \quad (k=\pm1,\pm2,...) ,$$

i.e.

$$p_x = \frac{\pi h}{a} k \;, \qquad E_k = \frac{\pi^2 h^2}{2ma^2} k^2 \;. \tag{34.42}$$

We know already that the functions (34.41) are the eigenfunctions of the kinetic energy operator $-\dfrac{h}{2m} \dfrac{\partial^2}{\partial x^2}$ with eigenvalues E_k being determined by the formula (34.42). The function (34.41) is composed of the superposition of two eigenfunctions for the impulse operator $-ih\dfrac{\partial}{\partial x}$ with eigenvalues equal to $p_k = \pm \pi k / a$.

Therefore, to the state function ψ of the formula (34.41) there will correspond the quantum microparticle of the energy E_k and of the impulses p_k and $-p_k$ acceptable equiprobably.

Let us also find the probability density of the coordinate x of the quantum microparticle – located in the potential well with infinite walls – for the functions ψ of the form (34.41).

For this purpose it is necessary to normalize the function of the formula (34.41) and to find a square power of its module. Denote this function as ψ_k . Immediate calculations reveal that

$$\int \psi_k \psi_k \, *dx = \frac{1}{\sqrt{2a}} \left(\psi_k = e^{-i\frac{E}{h}t} \sin \frac{p_k}{h} x \right) \;;$$

therefore, for the normalized function ψ_k we shall obtain

$$|\psi_k|^2 = \frac{1}{2a} \sin^2 \frac{k\pi}{a} x \;. \tag{34.43}$$

The associated plots of the probability density are given in figure 34.3 (for $k = 1,2$ and 3).

The possible state functions ψ are all possible superpositions with complex coefficients of the found functions ψ_k , i.e.

$$\psi(x,t) = \frac{1}{2a} \sum c_k \exp\left(-i \frac{\pi^2 h k^2}{2ma^2} t \right) \sin \frac{k\pi}{a} x \;,$$

for which the normalizing condition will imply the sum of squares of modules of the coefficients c_k being equal to unity. Here, the quantum microparticle has the energy E_k and the impulse p_k or $-p_k$ with the probabilities ½ $\left| c_k \right|^2$.

What then makes a travel of the quantum microparticle in the potential well different from a travel of the classic microparticle? The classic microparticle within the well will travel from one wall to another refracting from them ideally. It may have any energy E. One half of the time it will alternately and periodically have the impulse $\sqrt{2mE}$ and another half the impulse $-\sqrt{2mE}$. No superposition of travels can surely occur. In essence, between the energy and the impulse there exists little common, only the relation $p_x^{\ 2} = 2mE$.
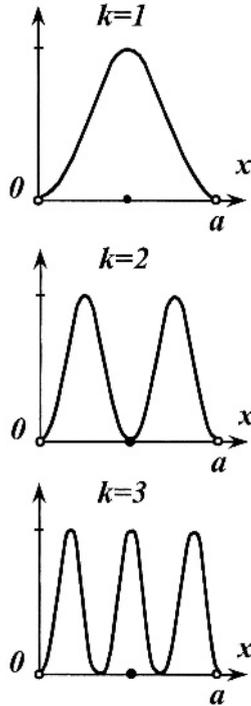

Fig. 34.3. The behaviour of the microparticle in the infinitely deep potential well. The figure shows the first three harmonics of the standing harmonic waves composing the function $\psi(x,t)$.

We have, thus, solved the problem of a one-dimensional movement of the quantum particle in the infinite rectangular potential well determined by the function (34.37). The function will be denoted as $V_a(x)$. Having this solution

$$\psi_k(x) = \frac{1}{\sqrt{2a}} \exp\left(-i\frac{E_k}{n}t\right) \sin\frac{\pi k}{a}x$$

it becomes possible to derive also the solution for the particle located in a three-dimensional rectangular well (a rectangular box). For this solution, the potential function will be of the form

$$V = V_a(x) + V_b(y) + V_c(z) \quad .$$

Namely, the needed functions $\psi(x, y, z)$ will be

$$\psi_{k_1 k_2 k_3}(x, y, z) = \frac{c^{-iEt/h}}{\sqrt{8abc}} \sin\frac{\pi k_1}{a}x \sin\frac{\pi k_2}{b}y \sin\frac{\pi k_3}{c}z \quad ,$$

where k_1, k_2, k_3 are the integers exceeding zero, and

$$E_{k_1 k_2 k_3} = \frac{\pi^2 h^2}{2m}\left(\frac{k_1^2}{a^2} + \frac{k_2^2}{b^2} + \frac{k_3^2}{c^2}\right) \quad .$$

34.4
Diffusion of a microparticle through a potential barrier

Let us assume microparticles to be travelling in the direction of the axis Ox. Let the energy of each of them be E; hence, they will be described by der Broil wave derived above. Let us put on their way some potential barrier of the energetic level V_0 (Fig. 34.4).

In the classic case ($E > V_0$), the microparticles will pass through the potential barrier and will proceed their travel at the previous velocity along the axis Ox. Though, if $E < V_0$, the microparticles will refract fully and no microparticle will penetrate the barrier.

For the macroparticles, the picture will be absolutely different. For $E < V_0$, only some portion of them will "break through" the barrier; the remaining macroparticles will be refracted. A partial refraction will also occur at $E > V_0$. The remaining macroparticles will successfully come through the barrier, though

their energy E is less than the energetic height V_0 of the barrier. If to think in terms of the wave dualism, then such a behaviour is rather natural – der Boil wave falling onto the potential barrier is partially reflected and partially, after being refracted, will penetrate it.
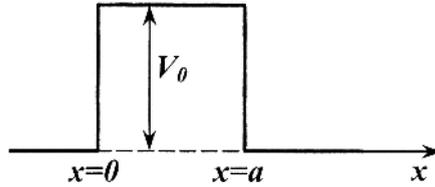

Fig. 34.4. The case when the microparticle is passing through the potential barrier of the width a; in particular, when the microparticle energy is less than the barrier energetic height V_0.

Now, let us see what will be achieved by us if a microparticle is represented via a formal mathematical model. Let us try to find a stationary solution of the Schrödinger equation with the energy E. Taking into consideration the description of the potential energy $V(x)$ (Fig. 34.4) by the formula

$$V(x) = \begin{cases} 0 & for & x < 0 \\ V_0 & for & 0 \le x \le a \\ 0 & for & x > a \end{cases}$$

we shall find that the equation (34.22), following from the Schrödinger equation, will be, for different x, of different forms. Thus, we shall obtain

$$-\frac{h^2}{2m}\frac{\partial^2 \overline{\psi}}{\partial x^2} = E\overline{\psi} \qquad \text{for } x < 0$$

$$-\frac{h^2}{2m}\frac{\partial^2 \overline{\psi}}{\partial x^2} = (E - V_0)\overline{\psi} \qquad \text{for } 0 \le x \le a \qquad (34.44)$$

$$-\frac{h^2}{2m}\frac{\partial^2 \overline{\psi}}{\partial x^2}= E\overline{\psi} \qquad \text{for } x > a \quad .$$

The solution $\overline{\psi}$ of these equations and its derivative must be bounded and continuous.

Introducing, for brevity, the below notation

$$\frac{2mE}{h^2}= k^2 \quad , \qquad \frac{2m(E-V_0)}{h^2}= k_1{}^2$$

we shall immediately obtain from the equation (34.44) that

$$Ae^{ikx} + Be^{-ikx} \quad \text{for } x < 0$$

$$Ce^{ik_1 x} + De^{-ik_1 x} \quad \text{for } 0 \le x \le a \qquad (34.45)$$

$$Ee^{ikx} + Fe^{-ikx} \quad \text{for } x > a \quad .$$

The smoothness conditions for the function $\overline{\psi}(x)$ at $x = 0$ and $x = a$ will lead us to the four conditions

$$A + B = C + D$$

$$ikA - ikB = ik_1 C - ik_1 D$$

$$(34.46)$$

$$Ce^{ik_1 a} + De^{-k_1 a} = Ee^{ika} + Fe^{-ika}$$

$$ik_1 Ce^{ik_1 a} - ik_1 De^{-ik_1 a} = ikEe^{ika} - ikFe^{-ika}$$

for the six unknown complex constants, A, B, C, D, E and F. Since the equations are only four, their solution will not be unique. This is understandable, since the function $\overline{\psi}$ may be determined to within the constant multiplier. Also, the physical sense of the problem tells us that the microparticles cannot run from infinity from right to left; this has not yet been accounted. Hence, this makes it possible to assume that $A = 1$ and $F = 0$. Upon this, the equations (34.46) are solved uniquely and we get

$$B = \frac{\left(k^2 - k_1{}^2\right)\left[exp\left(ik_1 a\right) - exp\left(-ik_1 a\right)\right]}{\left(k - k_1\right)^2 exp\left(ik_1 a\right) - \left(k + k_1\right)^2 exp\left(-ik_1 a\right)}$$

$$E = \frac{-4kk_1}{\left(k - k_1\right)^2 \exp\left(ik_1 a\right) - \left(k + k_1\right)^2 \exp\left(-ik_1 a\right)}$$

$$C = \frac{k + k_1 + (k_1 - k)B}{2k_1}$$

$$D = \frac{k_1 - k + (k + k_1)B}{2k_1} \quad .$$

From these formulas, it follows that for the reflected and refracted Broil wave we have

$$|B|^2 = \frac{4(k^2 - k_1^2)\sin^2 k_1 a}{4(k^2 + k_1^2)^2 \sin^2 k_1 a + 16k^2 k_1^2 \cos^2 k_1 a}$$

$$|E|^2 = \frac{16k^2 k_1^2}{4(k^2 + k_1^2)^2 \sin^2 k_1 a + 16k^2 k_1^2 \cos^2 k_1 a} \quad .$$

Therefore, alongside with the reflected particles there exist the particles passing the barrier independently of whether greater or less than the barrier energetic height V_0 is the energy of the particle. This is an absolutely new phenomenon of the "ïnfiltration" or diffusion of the microparticle through some energetic barrier. This has no classic analogue and contradicts the classic science.

34.5
Atom of hydrogen

An atom of hydrogen is an electron in the central Coulomb field of a positively-charged nucleus. A well-known two-body problem solved by Newton is a classic analogue for a hydrogen atom. In the classic case, a movement of the body is described by the three Kepler laws – concerning the orbit ellipticity, the sector velocity constancy and the relation of cerculating periods and large orbit semiaxes. For a microparticle, everything becomes different – it has no orbit, its energy and impulse are quantified and simultaneously have, undoubtedly, no definite values. Its possible values of energy

$$E_n = \frac{me^4}{2h^2}\frac{1}{n^2} \quad (n = 1,2,3,...) \tag{34.47}$$

turn out to be dependent upon the integer n .

From mathematical viewpoint, finding the quantum values of energy E_n is an effective solution for the complicated and difficult problem concerning eigenvalues and functions of a linear differential operator.

The results being obtained are also vividly confirmed by the observations over the frequences of spectral lines in a hydrogen atom. We recall here that the frequency ω of the atom-emitting photon (der Broil wave), when the electron is passing from the energetic level E_n to E_m, is determined by the relation

$$h\omega = E_n - E_m \qquad (n < m) \ .$$

This yields the observable series of the spectral lines

of Liman

$$\omega_{1m} = E_1 - \frac{E_m}{h} \quad (m = 2,3,4,...) ,$$

of Balmer

$$\omega_{2m} = \frac{E_2 - E_m}{h} \quad (m = 3,4,5,...) ,$$

of Fitz-Paschen

$$\omega_{3m} = \frac{E_3 - E_m}{h} \quad (m = 4,5,6,...) ,$$

and of others.

The purpose of our further narration is describing a statement of the problem and describing its solution for hydrogen atoms or hydrogen-like ion with the nucleus charge z and a single electron.

In our case, the potential energy is equal to

$$V(r) = -\frac{ze^2}{r}$$

and the Schrödinger equation is better to be written with a use of the following spherical coordinates $0 \le r \le \infty$, $0 \le \vartheta \le \pi$, $0 \le \varphi \le 2\pi$:

$$-ih\frac{\partial \psi}{\partial t} = -\frac{h^2}{2m}\Delta \psi + V(r)\psi =$$

$$= -\frac{h^2}{2m}\left\{\frac{1}{r^2}\frac{\partial}{\partial r\left(r^2\frac{\partial}{\partial r}\right)} + \frac{1}{r^2}\left[\frac{1}{sin\,\vartheta}\frac{\partial}{\partial \vartheta}\left(sin\,\vartheta\frac{\partial}{\partial \vartheta}\right) + \frac{1}{sin^2\,\vartheta}\frac{\partial^2}{\partial \varphi^2}\right]\right\}\psi +$$

$$+ V(r)\psi \quad.$$

A search for the stationary solutions for this Schrödinger equation will lead us to the equation (to the problem of eigenfunctions and eigenvalues of E).

To find for all r, ϑ and φ (including $r = \infty$) the continuous finite solutions of this equation will be a rather complicated and beautifully solved problem. Its solution is a result of the effort of numerous well-known scientists and is at present described in textbooks on quantum mechanics and mathematical physics. The basic results of this research include the following. For $E > 0$, the spectrum is continuous and the associated eigenfunctions do not turn to zero at infinity. This corresponds to a possibility of unlimited escape of a microparticle from atom nucleus. On the contrary, for $E < 0$ the spectrum is discrete and at infinity the eigenfunctions are nullified. This means that the electron will remain at some finite distance from the nucleus. The discrete eigenvalues will be equal to

$$E_n = -\frac{z^2 me^4}{2h^2}n^{-2} \quad (\,n = 1,2,3,...\,)\quad.$$

To each of their eigenvalues there corresponds some finite number of eigenfunctions of the form

$$\overline{\psi}_{nlm} = R_{nl}(r)Y_{lm}(\vartheta,\psi)$$

$$\left(l = 0,1,2,...,n-1; m = 0,\pm 1,...,\pm l\right)\quad.$$

As mentioned above, this fully corresponds to the observable hydrogen atom spectrums.

34.6
Quantum linear oscillator

In the previous chapters, a great portion of our attention was dedicated to the classic linear harmonic oscillator described by the equation

$$m\ddot{x} + kx = 0$$

and by the Hamilton function

$$H = \frac{p_x^{2}}{2m} + \frac{m\omega^2 x^2}{2} \quad \left(p_x = m\dot{x}, \omega^2 = \frac{k}{m} \right) \; .$$

The associated quantum oscillator is described by the Schrödinger equation

$$ih \frac{\partial \psi}{\partial t} = -\frac{h^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + \frac{m\omega^2}{2} x^2 \psi \qquad (\psi = \psi(x,t))$$

and the search for stationary solutions with the energy E of this equation leads us to the equation

$$-\frac{h^2}{2m} \frac{d^2 \overline{\psi}}{dx^2} + \frac{m\omega^2}{2} x^2 \overline{\psi} = E \overline{\psi} \; .$$

To find its eigenvalues and eigenfunctions is not so simple. As a result, the following is obtained

$$E_n = h\omega \left(n + \frac{1}{2} \right) \quad (n = 0,1,2,...) \; ;$$

the corresponding normalized eigenfunctions will be equal to

$$\overline{\psi}_n (x) = \frac{1}{\sqrt{x_0}} H_n \left(\frac{x}{x_0} \right) \exp \left(-\frac{x^2}{2x_0^{2}} \right) \; ,$$

where

$$x_0 = \sqrt{\frac{h}{m\omega}} \; , \qquad H_n(y) = \frac{(-1)^n}{\sqrt{2^n n! \sqrt{\pi}}} e^{y^2} \frac{d^n e^{-y^2}}{dy^n} \; .$$

The first three functions $\overline{\psi}_0$, $\overline{\psi}_1$ and $\overline{\psi}_2$ are depicted in figure 34.5.
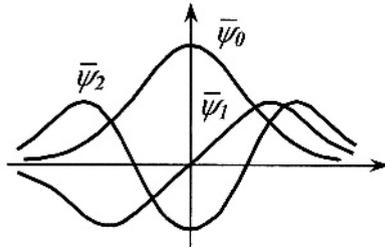

Fig. 34.5. The types of the first three functions $\overline{\psi}$ for the stationary solutions of the Schrödinger equation of the quantum linear oscillator.

In accordance with the above said, the values of energy are equidistantly quantized and the minimally possible energy is more than zero and equal to $h\omega/2$. The diagrams $|\psi_n|^2$ (Fig. 34.5) are also in no way similar to the diagram for the probability of finding, at random instant, the classic oscillator in the place of the coordinate x. This last diagram is held in figure 34.6.
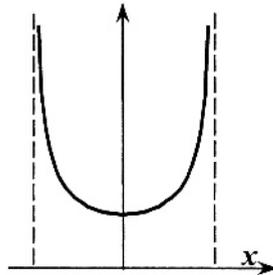

Fig. 34.6. The probability density of x for the classic linear oscillator (see Fig. 34.5 for comparison); the vertical asymptotes correspond to the extremal possible values of x ($x = \pm a$, where a is an oscillating amplitude).

34.7
Newton quantum equation

The travel of a microparticle is described through the changes of its state function and in no way resembles the travel along a trajectory. Even if a location of the microparticle is already known, its velocity will be still indefinite, i.e. no trajectory can be constructed. Nevertheless, a fast flying microparticle in a bubble chamber will leave some trace. What is then this trace itself?

The matter is that in this case the dispersions of coordinates and velocity (impulse) are so small that a trace of the "random walk" from one formation of a drop to another will be interpreted by us as some trajectory, because it is rather thin. A width of this "trajectory" will be gigantic as compared against an atom, but in our scale of the macroworld it will be, nevertheless, thin.

Let us agree here that some indefiniteness in its position and velocity will be very small all the time, but why then does this movement correspond in general to our classic interpretations? In his time, this question was answered

by P. Ehrenfest. He found that for the average value $< x >$ of the coordinate x may be described by some analogue of the Newton equation, which can be called the Newton quantum equation:

$$ m \frac{d^2}{dx^2} < x >= -\left\langle \frac{\partial V}{\partial x} \right\rangle \approx -\frac{\partial V(\langle x \rangle)}{\partial x} \quad . $$

To derive this equation is not difficult; its derivation is given in the textbooks on quantum mechanics and makes use of the microparticle model described earlier. In case of the macrobodies, $< x >$ will be indistinguishable from x, since the dispersion of x will be very small.

35 Space and time

The Galileo and Lorentz transformations. The invariance of the Newton equations and the noninvariance of the Maxwell equations with respect to the Galileo transformations. The relation between the events observed in various coordinate systems: shortening of distances, a clock slowdown and addition of velocities. The four-dimensional space – time. The idea of simultaneity and the causality principle. The improvement of the Newton equations for the case of very fast velocities close to the velocity of light.

In our previous chapters, various mathematical models have been discussed. Nearly all of them contained spatial variables and time. What they are was not explained by us, for their idea seemed self-evident. I. Kant considered the ideas of space and time to be inherent. Though, it turned out so that all of them need explanations and deserve some special attention. Now, it is the time to ponder over what spatial variables and time were exploited in the great mathematical models in mechanics and physics, and, consequently, in all the mathematical models that were founded on them. Please just think that the question here is not in what space and time generally mean. Here, we are putting a specific question about what specific variables are exploited by the mathematical models in natural science and engineering. In our case, this question will be even put in a more specific way: how are these variables interrelated in various inertial systems, travelling at constant velocities respective to each other?

I. Newton believed that there exists some distinguished absolute reference system. In it, coordinates are measured through some assigned scale and time by an absolute clock. According to him, there exists some absolute space and absolute time.

Yet by Euclid the space surrounding us was subjected to mathematical investigation and the result of that was the Euclidean geometry and our Euclidean space. Though, it was already K.F. Gauss who doubted our space to be Euclidean and wanted to measure the angles of a large triangle formed by mountain tops. If the sum of the angles of that triangle had been different from π, then this would have manifested the non-Euclideanism of our space. The later appearing geometries by N.I. Lobatchevsky and G.F. Riemann laid a theoretical foundation under these Gaussian doubts, but the measurements of the entire Earth have brought no authentic deviations from the Euclideanism.

Doubts about the nature of geometry of our space were originated comparatively long ago. What that time seemed certain was that space is space and time is time; they were thought to exist independently and not interrelated. That this way of thinking was wrong was revealed in the beginning of the 20-th century. We shall touch this very subject in the below narration. It was made clear that one should speak not about a three-dimensional space and one-dimensional time but rather about a unified four-dimensional space-time world.

The coordinate systems where the Newton laws are true are called inertial. The inertial systems are the systems travelling with respect to each other at the constant velocity. If the axes Ox and $O'x'$ are directed along this velocity of the relative movement of the systems $Oxyz$ and $O'x'y'z'$, then the coordinates and the time of the same point in these different reference systems will be interrelated through the ratios

$$x' = x - \upsilon t, \ y' = y, \ z' = z, \ t' = t \tag{35.1}$$

called the Galileo transformations.

The fact of the validity of the Newton equations in all inertial systems and, in particular, in the systems $Oxyz$ and $O'x'y'z'$ implies the Newton equations to be invariant (i.e. they retain their form) when the variables x, y, z, t are replaced by x', y', z', t', according to (35.1). Indeed, the equations

$$m\ddot{x} = F_x, \qquad m\ddot{y} = F_y, \qquad m\ddot{z} = F_z, \tag{35.2}$$

upon the variables being replaced in accordance with (35.1), are transferred for x', y', z' and t' to the similar equations. The similarity of the forces F_x, F_y and F_z in both coordinate systems may be substantiated by the fact that this similarity retains valid for interactions dependent upon distances or deformations of elastic bodies. Thus, the Newton equations and everything following from them, i.e. all equations in mathematical models of mechanical systems, are invariant with respect to the Galileo transformations. This invariance is the consequence of the validity of the Newton laws for all inertial coordinate systems interconnected through the Galileo transformations.

Now, we make a jump from mechanics to electrodynamics. Our experience and the specially arranged experiments manifest the Maxwell equations to hold in inertial systems, i.e. they should be also invariant with respect to the Galileo transformations. We shall not touch the entire system of Maxwell equations. Instead, we shall take one of its corollaries – the wave equation for the plane electromagnetic wave travelling along the axis Ox and the electrical field directed along the axis Oy. This equation, due to (35.3), will be of the form

$$\frac{\partial^2 E_y}{\partial t^2} = c^2 \frac{\partial^2 E_y}{\partial x^2}. \tag{35.3}$$

This wave equation written in terms of the system $Oxyz$, upon the Galileo transformations (35.1), should again transfer to a wave equation. What E_y should transfer to is not known for us and we shall not ponder upon it. It is clear that the operator

$$\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} \tag{35.4}$$

should transfer to the operator

$$\frac{\partial^2}{\partial t'^2} - c^2 \frac{\partial^2}{\partial x'^2} \quad . \tag{35.5}$$

The first of them is nullified for the function $E_y(x,t)$ and the second for the unknown function $E'_{y'}(x',t')$.

We immediately find that

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial x'} \cdot \frac{\partial x'}{\partial x} + \frac{\partial}{\partial t'} \cdot \frac{\partial t'}{\partial x} = \frac{\partial}{\partial x'}, \qquad \frac{\partial^2}{\partial x^2} \rightarrow \frac{\partial^2}{\partial x'^2}$$

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial x'} \frac{\partial x'}{\partial t} + \frac{\partial}{\partial x'} \frac{\partial x'}{\partial t} = -\upsilon \frac{\partial}{\partial x'} + \frac{\partial}{\partial t'}$$

$$\frac{\partial^2}{\partial t^2} = -\upsilon \frac{\partial^2}{\partial x'^2} \frac{\partial x'}{\partial t} - \upsilon \frac{\partial^2}{\partial x'\partial t'} \frac{\partial t'}{\partial t} + \frac{\partial^2}{\partial t\partial x'} \frac{\partial x'}{\partial t} + \frac{\partial^2}{\partial t'^2} \frac{\partial t'}{\partial t} =$$

$$= \upsilon^2 \frac{\partial^2}{\partial x'^2} - 2\upsilon \frac{\partial^2}{\partial x'\partial t'} + \frac{\partial^2}{\partial t'^2} \quad ;$$

therefore, the operator (35.4) will transfer, upon replacing the variables (35.1) (the Galileo transformations), to the operator

$$\frac{\partial^2}{\partial t'^2} - 2\upsilon \frac{\partial^2}{\partial x'\partial t'} - (c^2 - \upsilon^2) \frac{\partial^2}{\partial x'^2} \quad , \tag{35.6}$$

being different from the wave operator (35.5).

On the one hand, this is sudden and implies either the incorrectness of the statements on the validity of the Newton and Maxwell equations in the inertial systems or the incorrectness of the Galileo transformation (then the Newton equations are not true) or, at last, the incorrectness of the Maxwell equations. Which of these three possibilities should be chosen?

On the other hand, that the operator (35.4) under the Galileo transformation has not transferred to the similar operator (35.5) is rather evident – since the system $O'x'y'z'$ moves at the velocity υ with respect to the system $Oxyz$ and , therefore, under this transformation the velocity c of the electromagnetic waves should change and accordingly a form of the wave should change as well. Let us get assured in it through finding the harmonic waves $e^{i(wt'-kx')}$ nullifying the operator (35.6)

$$\left\{\frac{\partial^2}{\partial t'^2} - 2\upsilon\frac{\partial^2}{\partial x'\partial t'} + (\upsilon^2 - c^2)\frac{\partial^2}{\partial x'^2}\right\} e^{i(wt'-kx')} =$$

$$= [-w^2 - 2\upsilon wk - (\upsilon^2 - c^2)k^2] e^{i(wt'-kx')} = 0 .$$

From which

$$(w + \upsilon k)^2 = c^2 k^2$$

or

$$w/k = -\upsilon \pm c .$$

It means that in the system $Ox'y'z'$ a wave, for $\upsilon < c$, will travell into one side at the velocity $c - \upsilon$ and into the other at the velocity $-c - \upsilon$.

In order to reveal whether the velocity of light in various inertial systems is different and whether it depends upon the movement of the source of light, special very accurate and well-founded experiments have been conducted. One of them is the famous Mickelson–Murray experiment. It revealed the light velocity in all systems to be similar and equal to $c \approx 300000$ km/sec. This proves that the Galileo transformations are not true and, hence, another transformation is needed to retain the type of the wave operator. Such transformations were first suggested by H.A. Lorentz and were called his name though he did not expand them to cover the general problem of space and time.

The Lorentz transformations had to replace the Galileo transformations and they are able to prompt how the Newton equations are improved. These improvements are very small and, therefore, in the mechanical observations and experiments with $\upsilon \ll c$ they did not expose themselves.

The same slight difference is also observed between the Lorentz transformations and the Galileo transformations. However, at the velocities close to the light velocity – such velocities are obtainable in the contemporary accelerators and observable among cosmic microparticles – this slight difference may not be already neglected.

Let us write the famous Lorentz equations

$$x' = \frac{x - \upsilon t}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}} \qquad y = y' \qquad z = z' \qquad t' = \frac{t - \dfrac{\upsilon}{c^2}x}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}} \tag{35.7}$$

and then derive them. To derive them is not so difficult. More difficult is to comprehend what will follow from them. We shall consider some simplest corollaries, but now only note here that one and the same event occurring at some point and at some instant of time has in different inertial systems not only different coordinates but times as well. Two events being simultaneous within one inertial system will not be simultaneous in another; their difference in times will be dependent both upon their relative velocity υ and the distance r between the places where these events occur.

Indeed, let the events occur at one and the same time t in the points x and $x + r$. This happens in one reference system called by us K. Then, in the other reference system K' travelling with respect to K at the velocity υ, the events, due to (35.7), have occurred at the times

$$t_1' = \frac{t - \dfrac{\upsilon}{c^2}x}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}} \qquad \text{and} \qquad t_2' = \frac{t - \dfrac{\upsilon}{c^2}(x + r)}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}}$$

differing by the magnitude

$$t_1' - t_2' = \frac{\upsilon r}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}} \quad .$$

The Lorentz equations (35.7) will be derived by us from the properties of space homogeneity and space isotropy, time homogeneity and wave operator invariance (the invariance of the propagating velocities of electromagnetic waves) – existent in all inertial systems. As was shown by J. H. Poincaré, with respect to the Lorentz transformations there will be invariant not only the wave equation derived from the Maxwell equations but also the Maxwell equations themselves.

We shall assume that one and the same event in different inertial systems of coordinates K and K' has a general type of relations between the coordinates and the time x, t and x', t'; these systems travel respective to each other at the constant velocity υ: the system K' with respect to K (its origin) at the velocity υ; the system K with respect to K' at the velocity $-\upsilon$. Their mutually

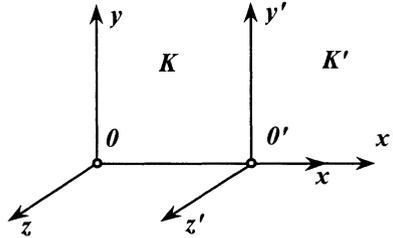respective travel is performed along the coinciding axes Ox and $O'x'$ (Fig. 35.1).

Fig. 35.1. The two inertial systems K and K' travelling repectively to each other at the constant velocity along the axis Ox coincident with $O'x'$.

At the time $t = t' = 0$, the origins, O and O', of both coordinate systems will coincide. We shall assume the coordinates and the time of systems K and K' to possess a general type of relations:

$$x' = ax + bt$$

$$y' = y$$

$$z' = z \qquad\qquad (35.8)$$

$$t' = \alpha x + \beta t \quad .$$

To be more exact, we will try to meet all the above conditions using the above transformations and choosing the parameters a, b, α and β. The Galileo transformation corresponds to $a = 1$, $b = -\upsilon$, $\alpha = 0$, $\beta = 1$. As we already know, this transformation does not fit. It meets the homogeneity and isotropy requirements but does not retain the wave operator. Hence, we shall have to find the values of the four parameters, a, b, α and β.

We shall start from the requirement of the wave operator preservation. Repeating the calculations already known to you we shall obtain

$$\frac{\partial^2}{\partial x^2} = a^2 \frac{\partial^2}{\partial x'^2} + 2a\alpha \frac{\partial^2}{\partial x'\partial t'} + \alpha^2 \frac{\partial^2}{\partial t'^2}$$

$$\frac{\partial^2}{\partial t^2} = b^2 \frac{\partial^2}{\partial x'^2} + 2b\beta \frac{\partial^2}{\partial x'\partial t'} + \beta^2 \frac{\partial^2}{\partial t'^2} \quad .$$

Therefore, after a change of variables (35.8) the wave operator (35.4) will transfer to the operator

$$\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} = \left(b^2 - c^2 a^2\right)\frac{\partial^2}{\partial x'^2} +$$

$$+ \left(2b\beta - 2a\alpha c^2\right)\frac{\partial^2}{\partial x'\partial t'} + \left(\beta^2 - c^2\alpha^2\right)\frac{\partial^2}{\partial t'^2} \quad .$$

The wave equation transformed to new variables and corresponding to the operator (35.4) will now coincide with the wave equation corresponding to the operator (35.5), provided that their operators differ in a multiplier, i.e. if they meet the conditions

$$\frac{b^2 - c^2 a^2}{-c^2} = \frac{2(b\beta - a\alpha c^2)}{0} = \frac{\beta^2 - c^2\alpha^2}{1} \quad ,$$

which give rise to the below two conditions

$$b\beta - a\alpha c^2 = 0 \quad , \quad c^2 a^2 - b^2 = c^2(\beta^2 - c^2\alpha^2) \quad . \tag{35.9}$$

The other two conditions for the parameters a, b, α and β follow from the fact that the origin O' travels with respect to the origin O at the velocity v; O with the respect to O' at the velocity $-\upsilon$. In the system K', the point O' has the coordinate $x' = 0$. Hence, we have $ax + bt = 0$, where x is its coordinate in the system K upon the time t, i.e. the velocity of the point O' in the system K will be $\upsilon = -b/a$. Similarly, the point O in the system K' will have, at the instant $t' = \beta t$, the coordinate $x' = bt$. Therefore, its velocity in the system K' will be equal to b/β and this will lead us to one more ratio $b/\beta = -\upsilon$.

Therefore, the two ratios (35.9) will be added by the other two:

$$-b/a = \upsilon \qquad b/\beta = -\upsilon \quad . \tag{35.10}$$

These four ratios turn out to be interdependent and do not allow to find all four parameters. Though, they may be expressed via a single parameter, for example, a, as bellow

$$b = -a\upsilon \quad , \quad \beta = \alpha \quad , \quad \alpha = -\frac{a\upsilon}{c^2} \quad . \quad (35.11)$$

In accordance with the relations (35.11) the needed transformation (35.8) will be written as

$$x' = ax - a\upsilon t \quad , \quad t' = -\frac{a\upsilon}{c^2} + at \quad . \quad (35.12)$$

While looking at (35.12) one feels tempting to assign $\upsilon = 0$ and find that $a = 1$. This is right but only in the sense that for $\upsilon = 0$ we get $a = 1$. But what will happen for other $\upsilon \neq 0$?

Let us use the space isotropy and the fact that the K'-to-K transition should be of the same form (35.12) but with υ being substituted by $-\upsilon$. According to this, let us solve the equations (35.12) with respect to x, t and find

$$x' = \frac{x' + \upsilon t'}{a(1 - \frac{\upsilon^2}{c^2})} \qquad t = \frac{x'\frac{\upsilon}{c^2} + t'}{a(1 - \frac{\upsilon^2}{c^2})} \qquad (35.13)$$

The transformation (35.12) should coincide with (35.13) if in the last we substitute υ by $-\upsilon$ and x', t' by x, t; then again x, t by x', t'. The similarity gives rise to the conditions

$$a = \frac{1}{a(1 - \frac{\upsilon^2}{c^2})} \qquad a\upsilon = \frac{\upsilon}{a(1 - \frac{\upsilon^2}{c^2})}$$

$$\frac{a\upsilon}{c^2} = \frac{\upsilon/c^2}{a(1 - \frac{\upsilon^2}{c^2})} \quad ,$$

which are reduced to a single condition

$$a^2 (1 - \frac{\upsilon^2}{c^2}) = 1 \quad .$$

$$a = \pm \left(1 - \frac{\upsilon^2}{c^2} \right)^{-1/2} .$$

But it is already known by us that for $\upsilon = 0$ we have $a = 1$. Therefore,

$$a = \pm \left(1 - \frac{\upsilon^2}{c^2} \right)^{-1/2}$$

and the transformation (35.8) will have the form of the Lorentz transformations (35.7).

Let us rewrite it and its reverse variant with a use of the abbreviated notation

$$\gamma(\upsilon^2) = \left(1 - \frac{\upsilon^2}{c^2} \right)^{-1/2} ;$$

then, we obtain

$$x' = \gamma(\upsilon^2)(x - \upsilon t) \qquad t' = \gamma(\upsilon^2)(t - \frac{\upsilon}{c^2} x) \qquad (35.14)$$

and

$$x = \gamma(\upsilon^2)(x' + \upsilon t') \qquad t = \gamma(\upsilon^2)\left(t' + \frac{\upsilon}{c^2} x' \right) . \qquad (35.15)$$

We recall here that υ is a velocity of the system K' with respect to the system K and $-\upsilon$ is a velocity of K with respect to K'.

It must be noted again that from the Lorentz transformations it follows that any event occurring in some place and at some time has in different inertial coordinate systems not only different coordinates, that is habitual for us, but different times also. Indeed, suppose that in the coordinate system K there has occurred some event in the point x at the moment t. Then, due to (35.7), in the system K' this event will occur in the point $x' = \gamma(\upsilon^2)(x - \upsilon t)$ at the moment $t' = \gamma(\upsilon^2)\left(t - \frac{\upsilon x}{c^2} \right)$.

This means that there exist neither absolute time nor absolute coordinates.. Each inertial system possesses its proper coordinate system and its proper clock.

Now, it is not so sudden and is understandable that the events being simultaneous in one system will be not simultaneous in the other. The notion of simultaneity is not absolute but relative, since each coordinate system has its own clock. Now, how do they relate to each other? They relate in a strange way: the clock in the system K' is slow with respect to K, but the K-system clock

observable in the system K' is slow as well. This inference being natural from the viewpoint of logic (with due account of symmetry of the systems K and K') will be absolutely unfit from the viewpoint of our habitual interpretation of the running of different clocks.

Let us be watching during the time τ from the system K the K'-system clock located in the point x' with a use of the K-system clock, i.e. from the moment t_1 to t_2, where $t_2 - t_1 = \tau$.

Let the readings on the K'-system clock (located in the point x') corresponding to the times t_1 and t_2 be the times t_1' and t_2'. Then, due to the second of the equations (35.15), we shall obtain

$$t_1 = \gamma(v^2)\left(t_1' + \frac{vx'}{c^2}\right) \qquad t_2 = \gamma(v^2)\left(t_2' + \frac{vx'}{c^2}\right)$$

and, consequently,

$$r = t_2 - t_1 = \gamma(v^2)\left(t_2' - t_1'\right) = \gamma(v^2)\tau' , \qquad (35.16)$$

i.e. the K'-system clock in the system K will run $\gamma^{-1}(v^2)$ times slower. Analogously, the K-system clock in the system K' will also run slower so much.

A slow down of the travelling clock respective to a fixed clock is directly proved by experiments. This fact is supported by the measurements of the short-life fast-travelling microparticles. A fast-travelling microparticle by the fixed clock has a longer life and this life prolongation agrees with the deductions derived from the Lorentz transformations, i.e. the particle lives $\gamma(v^2)$ times longer. Life of a flying microparticle is the time on the clock of the coordinate system travelling together with this microparticle. This type of time is called proper time of the microparticle. From the equality of the status of all inertial systems it will follow that the proper life time of a microparticle will not depend upon its velocity; in all the inertial systems this time will be similar.

The above assertion also retains somewhat valid for the clock travelling on a man-made satellite; with respect to the Earth-located clock, this travelling clock will be slow.

At the same time, this experimentally proved slow-down of the travelling clock gives rise to the paradox of twins that was widely discussed in its time. The idea of the paradox lies in the following: from the two twins (A and B) born on the Earth, one (A) had been left on the Earth and another (B) had been sent travelling there and back at the velocity being close to the velocity of light. The travelling twin (B) returned to the Earth retaining young while his brother (B) by this time became already very old. According to physicists, everything in this case should be exactly so. Though, the paradox lies in that that the similar consideration can be

performed by the travelling twin (B), with respect to whom the twin on the Earth (A) will become a traveller. Then, an old man will be the twin B and young will remain the twin A. This paradox is solved by the fact that there exists no full symmetry between the twins: a travelling twin was retarding and on his way back he was speeding up; here, what was happening with his own (proper) clock cannot be yet explained by our theory. As for the twin left on the Earth, he was doing nothing of the kind.

Let us now consider another phenomenon – the contraction of the sizes of a moving segment. Let in the system K' along the axis $O'x'$ lie the segment $\left[x_1', x_2' \right]$ of the length $l' = x_2' - x_1'$. In the coordinate system K', it is of the length l'. This segment, we call it L', is moving at the velocity v relatively to the system K. We adopt here the following natural procedure for measuring its length in the system K.

Let us locate in the system K the coordinates x_1 and x_2 of the terminals of the segment L' at some similar instant t. In the system K, the length of the segment will be assumed to be equal to $l = x_2 - x_1$. We will find the ratio between l and l'.

In accordance with the first relation in (35.15) and the second in (35.14), we obtain

$$x_1 = \gamma\left(v^2\right)\left(x_1' + vt_1' \right) \qquad x_2 = \gamma\left(v^2\right)\left(x_2' + vt_2' \right)$$

$$t_1' = \gamma\left(v^2\right)\left(t - \frac{v}{c^2} x_1 \right) \qquad t_2' = \gamma\left(v^2\right)\left(t - \frac{v}{c^2} x_2 \right) .$$

From it, we get

$$l = x_2 - x_1 = \gamma\left(v^2\right)\left[x_2' - x_1' + v\left(t_2' - t_1' \right) \right] =$$

$$= \gamma\left(v^2\right)\left[l' - \gamma\left(v^2\right)\frac{v^2}{c^2}(x_2 - x_1) \right] = \gamma\left(v^2\right)\left[l' - \gamma\left(v^2\right)\frac{v^2}{c^2} l \right] ;$$

therefore,

$$(1 + \gamma^2 \frac{v^2}{c^2}) l = \gamma\left(v^2\right) l'$$

or

$$l = \gamma^{-1}\left(\upsilon^2\right)l' \quad . \tag{35.17}$$

This last relation implies that with the K'-system segment of length l' being measured in the system K the length of this segment will be reduced as much as $\gamma^{-1}\left(\upsilon^2\right)$ times.

The previous corollaries from the Lorentz transformations explained us the specificities of physical space and time. Certainly, it is simultaneously both space and time we are living in, but in order to reveal them we have to deal with the velocities similar to the velocily of light or should be able to register the effects negligible in size.

The next sudden, but also for the velocities much less than the light velocity, slightly noticeable effect is the incorrectness of the rule of adding the transferable and the relative velocities. In our time, we have derived a formula for a missile speed-up and found that with the limitless reduction of its mass the missile will increase its velocity infinitely. This conclusion followed from the fact that with each ejection of mass dm at the velocity c with respect to the missile, the missile's velocity is augmented by cdm/m. This increment to the missile velocity was in the coordinate system related to the missile. From the Lorentz transformations it follows that in the coordinate system related to the Earth from which the missile was launched the missile velocity will not be augmented in this way; the increment will be less. Hence, due to this reduction the missile velocity will not surpass the light velocity though can come close to it.

Therefore, let the K'-system be travelling, as earlier, respectively to the K-system with the velocity υ and within the K'-system let some point be travelling, respectively to it, at the velocity u'. Then, the velocity u of the same point with respect to the K-system will be $\upsilon + u'$. This is so in accordance with the classic interpretation of space and time. In mechanics, the velocity υ is called transferable and u relative; the movement of a point in the K-system is called absolute; and the movement of the same point in the K'-system relative. In this terminology, an absolute velocity is equal to a sum of the transferable and relative velocities. Let us now see what will happen if we follow this thinking about space and time. Let the moving point M in the systems K and K' be in the points x and x' at the instants t and t'. Now, t should not necessarily coincide with t'. The values of x, t and of x', t' are interconnected by the Lorentz transformations (35.14) and (35.15). The first transformation provides a transition from the system K to K'; the second, vice versa, from K' to K. The K'-system moves with respect to the K-system at the velocity υ; the point M in the K'-system moves with the velocity $u' = \dfrac{dx'}{dt'}$. In the K-system, the velocity of the same point M will be equal to $u = \dfrac{dx}{dt}$. We need to find a

relation between these velocities (u and u') knowing that x, t and x', t' are connected by the ratios (35.14) and (35.15).

Immediately we obtain

$$u = \frac{dx}{dt} = \frac{d\left\{\gamma\left(v^2\right)\left(x' + vt'\right)\right\}}{d\left\{\gamma\left(v^2\right)\left(t' + \dfrac{v}{c^2}x'\right)\right\}} = \frac{dx' + vdt'}{dt' + \dfrac{v}{c^2}dx'} = \frac{u' + v}{1 + \dfrac{u'v}{c^2}} \quad .$$

Thus, the velocities v and u' are not added actually. There takes place a more complicated dependence

$$u = \frac{v + u'}{1 + \dfrac{u'v}{c^2}} \quad , \tag{35.18}$$

which for $v \ll c$ and $u' \ll c$ comes close to a simple addition; though, for $v = c$ and $u' = c$ we have $u = c$, but not $u = 2c$; here, this difference is rather great.

It may be noted here that for $v < c$ and $u' < c$ one has $u < c$; therefore, any number of the velocity augmentations being less than the light velocity will never let the missile surpass the light velocity.

A natural consequence of the change in comprehensions concerning space and time is their unification, i.e. understanding the necessity to consider them together. Earlier, a three-dimensional space and one-dimensional time existed separately; nowadays, there exists a four-dimensional space of space – time. The points in this four-dimensional space will be events. One and the same event is differently described in different reference systems. In the system K, the coordinates will be x, y, z, t; in the system K' the coordinates x', y', z', t'. Besides, different may be not only x and x', y and y', z and z' but t and t' as well. The Lorentz transformations constitute some particular case, when the axes of the systems K and K' are parallel and moving respective to each other along the axis Ox coincident with $O'x'$. To this particular case, there may be reduced any two systems $\widetilde{K}$ and $\widetilde{K}'$ travelling with respect to each other at a constant velocity, since the transition from $\widetilde{K}$ to K, as well as from K' to $\widetilde{K}'$`, will be nothing but some transitions from one space system to another. Here, the transformation from $\widetilde{K}$ to $\widetilde{K}'$ may be thought as a sequence of transitions from $\widetilde{K}$ to K, from K to K' and from K' to $\widetilde{K}'$ (Fig. 35.2).

When transforming the coordinates from one system to another being fixed with respect to the former, the distances
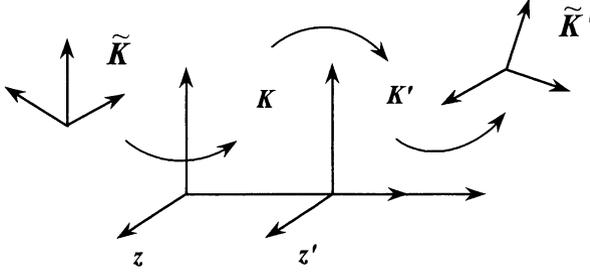
$$dr^2 = dx^2 + dy^2 + dz^2 \tag{35.19}$$

will retain.

Fig. 35.2. The subsequent transformations from one inertial system to any other: the systems $\widetilde{K}$, K and K', $\widetilde{K}'$ are stationary respective to each other; K and K' are travelling with respect to each other at the constant velocity.

Analogously, in the four-dimensional space-time the magnitude

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2 \qquad (35.20)$$

will retain.

To be sure in this, it is sufficient to show that this takes place for the Lorentz transformations (35.15) from the system K to the system K', since for the transformations from $\widetilde{K}$ to K and from K' to $\widetilde{K}'$ it takes place with evidence. Using (35.15), we immediately find

$$dx^2 + dy^2 + dz^2 - c^2 dt^2 =$$

$$= \left[d\left(\gamma\left(v^2 \right)\left(x' + vt' \right) \right) \right]^2 + dy'^2 + dz'^2 -$$

$$- c^2 \left[d\left(\gamma\left(v^2 \right)\left(t' + \frac{v}{c^2} x' \right) \right) \right]^2 =$$

$$= \gamma^2\left(v \right)\left(dx'^2 + 2v\,dx'\,dt' + v^2\,dt'^2 \right) + dy'^2 + dz'^2 -$$

$$- c^2 \gamma^2\left(v \right)\left(dt'^2 + 2\frac{v}{c^2}\,dt'\,dx' + \frac{v^2}{c^4}\,dx'^2 \right) =$$

$$= dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2 \ .$$

Therefore,

$$ds^2 = ds'^2 \ . \tag{35.21}$$

The magnitude ds may be called a distance in the four-dimensional space of events. The transformations relating to the transition from one system $\widetilde{K}$ to another $\widetilde{K}'$ will retain this distance and in this sense they are the analogues of the movements in the three-dimensional space. This analogy may be done vivid, if the time t is replaced by $i\tau/c$ and υ by $-ic\upsilon$, respectively. Here, the quadratic form ds^2 will take the form

$$ds^2 = dx^2 + dy^2 + dz^2 + dr^2 \ , \tag{35.22}$$

and the Lorentz transformation will be of the form

$$x' = \frac{x + \upsilon r}{\sqrt{1 + \upsilon^2}} \qquad\qquad r' = \frac{r - \upsilon x}{\sqrt{1 + \upsilon^2}} \ ,$$

or, assuming $\left(1 + \upsilon^2\right)^{-1/2} = \cos\theta$ and respectively $\upsilon\left(1 + \upsilon^2\right)^{-1/2} = \sin\theta$, it will be of the form

$$x' = x\cos\theta + r\sin\theta$$

$$\tag{35.23}$$

$$r' = -x\sin\theta + r\cos\theta$$

representing itself the transformation from the system $Ox\tau$ to the system $Ox'\tau'$ turned relative to it by the angle θ (Fig. 35.3).
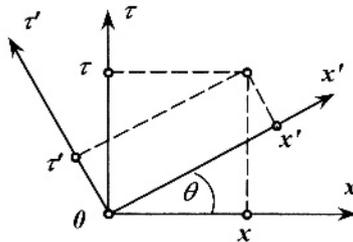
Fig. 35.3. The geometrically interpreted Lorentz transformation in the variables x, τ.

Thus, a transition from one inertial system to another may be represented as a usual transformation in the Euclidian space from one orthogonal coordinate system to another.

In connection with the invariance of ds^2 expressed by (35.21) I would like to tell you some words about causality in a new spatial and temporal world. Let us ask ourselves what events $S_1(x_1, y_1, z_1, t_1)$ and $S_2(x_2, y_2, z_2, t_2)$ in different points x_1, y_1, z_1 and x_2, y_2, z_2 and at different times t_1 and t_2 may be related causatively and what not. According to the idea of the casuality principal itself, the event S_2 may serve as an effect of the event S_1 only in the case $t_2 \geq t_1$. With this inequality being reverse, this is impossible. With one inertial system being replaced by another, the times t_2 and t_1 may vary and the inequality $t_2 \geq t_1$ may not hold; whereas the magnitude

$$ds^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 - c^2(t_2 - t_1)^2$$

will not change. Therefore, if $ds^2 < 0$, then $t_2 - t_1$ cannot be nullified and ,hence, the initial inequality $t_2 > t_1$ holds. Therefore, the cause-and-effect relations between such events can exist in all reference systems. This case is characterized by the fact that the time the light covers the distance

$$r = \left[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2\right]^{1/2}$$

will be less than $t_2 - t_1$. On the contrary, for $ds^2 > 0$ this time will exceed $|t_2 - t_1|$. Here, $t_2 - t_1$, when a transition to another coordinate system was made, may also change its sign, i.e. such events cannot be connected casually. The impossibility of casual relations between the events with $ds^2 > 0$ also follows from the fact that the action assigning this casual relation had to be propagated faster than the velocity of light. So, we are arriving here at the conclusion that the following will contradict an application of the casuality principle to all inertial systems. The Maxwell equations eliminated the instantaneousness of the Coulomb interaction between electrical charges; the same was done by the contemporary theory of gravitation and gravitational waves with the Newton interaction between masses. Neither this nor that can implement themselves faster than the velocity of light.

Hence, the causality principle holds also in the absence of absolute time.

The Maxwell equations are invariant with respect to the Lorentz transformations and not invariant respectively to the Galileo transformations. As for the Newton equations, they are, on the contrary, invariant respectively to the Galileo transformations and not invariant with respect to the Lorentz transformations. The difference between the Galileo and Lorentz transformations

makes up about υ^2/c^2 and is very small for the velocities observable in the time of the Newton mechanical systems. Nowadays, this is already not so. The observable cosmic microparticles have the velocities close to the light velocity c; also comparable with the light velocity are the velocities of the microparticles in the contemporary powerful accelerators. The observations of these running microparticles reveal a divergence from the Newton equations. The Newton equations need to be corrected in order to retain them true for the large velocities υ as well. To find out what corrections are needed is possible on the basis of the invariance principle with respect to the Lorentz transformations. Omitting all details, let us show only a final result. It includes the following.

The equation

$$\frac{d\overline{p}}{dt} = \overline{F} \ , \tag{35.24}$$

with $\overline{p}$ being a quantity of movement or an impulse, will retain, but the impulse itself, earlier equal to $m\overline{\upsilon}$, will be now equal to

$$\overline{p} = \frac{m_0\overline{\upsilon}}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}} \ , \tag{35.25}$$

where m_0 is rest-mass. The complete energy E of the particle is equal to

$$E = \frac{m_0 c^2}{\sqrt{1 - \dfrac{\upsilon^2}{c^2}}} \ ; \tag{35.26}$$

the rest-energy E_0 and the kinetic energy will be respectively equal to

$$E_0 = m_0 c^2 \qquad T = E - E_0 \ . \tag{35.27}$$

Besides, the kinetic energy T

$$T = E - E_0 = \frac{m_0 \upsilon^2}{2} + \frac{3}{8}m_0\frac{\upsilon^4}{c^2} + ... ,$$

even for $\upsilon \ll c$, will be close to its classic value $m\upsilon^2/2$.

Thus, the changes happened in the Newton equation will be reduced to the mass being substituted by the mass

$$m = \frac{m_0}{\sqrt{1 - \dfrac{v^2}{c^2}}} = m_0 \left(1 + \frac{1}{2}\frac{v^2}{c^2} + ...\right) \tag{35.28}$$

dependent upon the velocity and different from the previous mass by the magnitude of the order v^2 / c^2. The Newton mechanics specified to be applied to the large velocities of the travelling bodies was called the relativistic mechanics. A microparticle at the velocity of the order of the light velocity was called relativistic.

36   Speeding up relativistic microparticles in a cyclotron

The model of a relativistic particle moving in a cyclotron. A quasistationary stable resonance speed-up.

A cyclotron is a device for generating a bunch of charged microparticles of the velocities coming close to the extreme velocity, light velocity. The idea of the cyclotron is very simple. A charged microparticle in a homogeneous constant magnetic field, having the initial velocity perpendicular to the field, moves circlewise with the frequency ω in the classic approximation retaining independent of its velocity. That the rotating frequency of a microparticle is independent of its velocity makes it possible to create its resonance speed-up by the alternating electric field of the constant frequency ω. An actual speed-up of the charged microparticle is exercised not along the entire way, but only during its passage through the narrow aperture of the potential deference $V \sin \omega t$ (Fig. 36.1). Passing at some moment t the aperture, the e-charge microparticle will
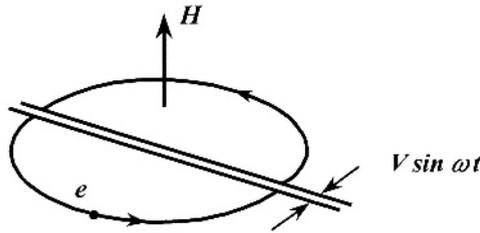
Fig. 36.1. The schematic diagram of a cyclotron.

receive some complementary energy $eV \sin \omega t$. Having covered a half-circle for the time π / ω, it will again cross the aperture and will again accept the augmentation equal to $-eV \sin(\omega t + \pi) = eV \sin \omega t$. This speed-up for $eV \sin \omega t > 0$ could proceed until the particle emission and energy loss during its movement become equal to the magnitude of the pumped up energy and the constant velocity circulation occurs. Though, this is hampered by the relativistic effect of the microparticle mass increase that will be caused by the growth of the velocity and by the angle velocity drop following from it. Here, the resonance

conditions will become disturbed and the energy pump-up will fall down sharply; as a result, a further speed-up of the microparticle will cease.

This is a brief story concerning how a microparticle speed-up in a cyclotron is exerted and the difficulties arising in its application if we wish to obtain the velocities close to the light velocity.

A search for the techniques to overcome these difficulties should be naturally accompanied with the construction and the study of mathematical models of the cyclotron. Here, our attention has to be focused only upon the mechanism of the charged microparticle resonance speed-up, with due account of the relativistic effect.

We start here from exposing a dependence of the microparticle rotation frequency upon the charge e, the mass m and upon the magnetic field H. The charged particle travelling at the velocity υ perpendicular to the magnetic field is acted upon by the Lorenz force being equal to $\dfrac{e\upsilon H}{c}$. This force perpendicular to the magnetic field and to the microparticle velocity does not change its velocity and gives rise to the centripetal acceleration corresponding to the travel of the microparticle along the circle of some radius R with the frequency ν. So, we obtain (Fig. 36.2)

$$R\,\nu^2 = \upsilon\nu = \frac{e\,\upsilon H}{mc}$$

or

$$\nu = \frac{eH}{mc} \quad .\tag{36.1}$$

Certainly, the same result may be obtained immediately from the Newton equation, with due account of the dependence of the mass upon the velocity υ

$$\frac{d}{dt}(m\overline{\upsilon}) = \frac{e}{c}\left[\overline{\upsilon}, H\right] \ , \quad m = m_0(1 - \frac{\upsilon^2}{c^2})^{-1/2} \ , \tag{36.2}$$

under the assumption that the initial velocity $\overline{\upsilon}$ is orthogonal to the constant homogeneous magnetic field $\overline{H}$. In the formulae (36.1) and (36.2), the mass m is dependent upon the velocity. Taking into consideration the energy E of the microparticle being equal to mc^2, the ratio (36.1) may be written as

$$\nu = \frac{eHc}{E} \quad .\tag{36.3}$$

This vividly demonstrates a dependence of the frequency ν upon the velocity υ. In the classical case, with the mass m being constant, this dependence will be not

available. Thus, if we want the equality $\omega = \nu$ to be retained with the variations of the microparticle velocity ν , then the frequency ω or the magnetic field H must be changed accordingly. To implement these dependences artificially is not an easy thing. Though, it turns out so that the nonlinear effects of the resonance speed-up create a possibility to automatically implement a needed relation between ω and H . This possibility is created by the mathematical model to be constructed by us.

Let E and $\overline{E}$ be values of the energy of one and the same microparticle upon its subsequent passes through the aperture; then we shall obtain

$$\overline{E} = E - f(E) + eV \sin \omega t \quad , \tag{36.4}$$

where t is an instant of the microparticle passing through the aperture (upon this passage the energy of the microparticle becomes equal to $\overline{E}$); $f(E)$ are the energy losses of the microparticle during its travel through the aperture; ω (assumed constant) is a frequency of the potential changes in the aperture of the negligibly small width. Let us further write the ratio between the time t and the time $\overline{t}$ of the next passage through the aperture

$$\overline{t} = t + \pi \frac{\widetilde{\overline{E}}}{eHc} \quad . \tag{36.5}$$

If in the interval from the instant t to the instant $\overline{t}$ the energy of the particle had remained constant, then in the formula (36.5) the magnitude $\widetilde{\overline{E}}$ would have coincided with $\overline{E}$. During its travel from one aperture to another the particle will lose its energy by $f(\overline{E})$, as was already assumed in (36.4). Thus, it is natural to approximately take that

$$\widetilde{\overline{E}} = \overline{E} - \frac{1}{2} f(\overline{E}) \quad .$$

When the particle passes through the aperture at the instants t and $\overline{t}$, the values of energy pump-up will, respectively, be $eV \sin \omega t$ and $-eV \sin \omega \overline{t} = eV \sin(\omega \overline{t} + \pi)$. In order to exert a resonance speed up, one should have for some integer $n > 0$ the following :

$$\omega \overline{t} = \omega t + (2n - 1)\pi \quad . \tag{36.6}$$

In the simplest case, $n = 1$ will take place. Here, we shall restrain ourselves by it. The ratio (36.6) indicates that the phases $\varphi = \omega t$ and $\overline{\varphi} = \omega \overline{t} + \pi$ of the harmonically oscillating voltage in the aperture at the instant of passing it by the particles will coincide or differ by the integer 2π .

From this requirement (36.6) for $n = 1$, due to (36.5) and the assumed magnitude $\widetilde{\widetilde{E}}$, we shall have the following

$$\overline{E} - \frac{1}{2} f(\overline{E}) = \frac{1}{\omega} eHc \ . \qquad (36.7)$$

The condition (36.7) defines the energy of the particle $E*$ as a root in the equation (36.7). Further, from (36.4) for $\overline{E} = E$ we shall arrive at the ratio

$$f(E*) < eV \sin \omega t \ , \qquad (36.8)$$

which can be fulfilled only for

$$f(E*) < eV \ . \qquad (36.9)$$

With the conditions (36.9) being met, the equation (36.8) will have two solutions, $0 < t_1 * < \pi/2$ and $\pi/2 < t_2 * = \pi - t_1 * < \pi$. To the found values of $E*$, $t_1 *$ and $E*$, $t_2 *$ there will correspond the stationary movements of the charged particle at which the consequent values of the particle energy E and $\overline{E}$ and, accordingly, the phases φ and $\overline{\varphi}$ will coincide and be equal to $E*$ and $\varphi_1 *$ or $\varphi_2 *$ (the phases will coincide if they are different by the integer 2π). In these stationary states, the charged particle will move periodically at the frequency $V = \omega$, i.e. a resonance will occur.

According to the equation (36.7) describing the energy $E*$ of the microparticle in the stationary movement, the energy will increase with the increase of the magnetic field H, but only until the existence condition of the stationary regime is obeyed (36.9). It will occur so because the energetic losses for the microparticle movement cannot in the stationary regime surpass the maximally possible energy pump-up from the side of the aperture electrical field.

In order to establish this stationary regime and have this dependence of $E*$ upon H, this stationary regime must be stable. This condition is not only necessary but also sufficient when H undergoes slow quasistationary changes. We have, thus, arrived at the necessity of studying the stability of the revealed stationary motions.

For this, let dE and dt be small disturbances of the stationary values of $E*$ and $\varphi*$, and $d\overline{E}$ and $d\bar{t}$ be the corresponding changes of $\overline{E}$ and $\bar{t}$ during a next passage through the aperture by the microparticle. In accordance with (36.4) and (36.5), we find that

$$dE = dE - f'(E^*)dE + eV\omega \cos \omega t^* \, dt$$

(36.10)

$$d\bar{t} = dt + \frac{\pi}{eHc}(d\bar{E} - \frac{1}{2} f'(E^*) d\bar{E}) .$$

Assuming $d\bar{E} = zdE$ and $d\bar{t} = zdt$ we arrive at the characteristic equation

$$\xi = \begin{vmatrix} 1 - f'(E^*) - z & eV\omega \cos \omega t^* \\ \left[1 - \frac{1}{2} f'(E^*)\right] \dfrac{\pi}{eHc} z & 1 - z \end{vmatrix} = 0$$

or

(36.11)

$$z^2 + \left[-2 + f' - (1 - \frac{1}{2} f') \frac{\pi V\omega}{Hc} \cos \omega t^* \right] z + 1 - f' = 0$$

determining the possible values of z .

   The stability conditions imply that both roots of the equation (36.11) should be in module less than unity. For this, the coefficients in the characteristic equation (36.11) τ and ν at the first and zero powers of z should meet the inequalities

$$\nu < 1 \quad , \quad 1 - \tau + \nu > 0 \quad , \quad 1 + \tau + \nu > 0 .$$

The corresponding substitution will yield the inequality

$$1 - f'(E^*) < 1$$

$$1 + 2 - f' + (1 - \frac{1}{2} f') \frac{\pi V\omega}{Hc} \cos \omega t^* + 1 - f' > 0 \qquad (36.12)$$

$$1 - 2 + f' - (1 - \frac{1}{2} f') \frac{\pi V\omega}{Hc} \cos \omega t^* + 1 - f' > 0 .$$

   The first and second inequalities are fulfilled due to the positiveness of f' and the smallness of $\dfrac{\pi \omega V}{Hc}$. The last inequality leads to the condition

$$\cos \omega t^* < 0 , \qquad (36.13)$$

which determines what of the two stationary regimes will be stable. This stable stationary regime will group the particles and will allow, via slowly increasing H , to increase their energy. With E^* coming close to the extreme limiting

value assigned by the equation $f(E_{ext}) = eV$, there will occur a disturbance of the stability and a disruption of the quasistationary speed-up.

The idea of the above described technique for speeding up the relativistic charged particles was initially suggested by V.I. Vecksler in 1944; its theoretical investigation presented above was performed by A.A. Andronov and G.S. Gorelik (1945). The basic result of their investigation was represented by them in the form of a resonance curve with a slant and with the phenomena of a breakdown and a hysteresis being specific for the nonlinear resonance.

37 Mathematics as a language and as an operating system and models

Mathematics as a special language including an operating system and models. Isomorphism and a commutativity diagram as a basis of mathematical modelling.

Below is the subject touched in the introduction. Here we shall speak about mathematics and its structure and why its role is so great and decisive in exact natural science and in exact sciences. However, this subject is suggested here only upon your having got already acquainted with various mathematical models and upon your having realized how the surrounding world can be recognized through these models, sometimes through very simple models and simple methods of their study.

So, what is mathematics? Or, more precisely, what does mathematics exist for? What is its role and meaning? An answer to these questions has been already given in the Introduction: mathematics is a language, in its assignment and role, i.e. from the point of view of its functions, it is very similar to such natural languages as Russian, English, German, French and others. Yet G. Galileo said: "The great book of nature has been written by mathematical symbols". From this it follows that only one who knows and possesses this symbolic language can hope to read this book of nature and understand his surrounding world, i.e. nature, engineering and society. Later on, the same idea was very laconically expressed by the great physicist W. Gibbs in his remark "mathematics is a language too". Therefore, its weight in studying the surrounding world is as great as that of usual languages in our ordinary life. We cannot help agreeing with it. In general, it is so. Though, the mathematical language is specific to some extent and this specificity will be discussed by us below.

A usual language includes a vocabulary, sound and graphical techniques for coding words, a grammar, all possible phrases and their combinations in the form of descriptions, statements, orders, narrations, instructions, stories, novels and so on. Something similar exists in the mathematical language. In it, as analogues to words and the grammar assigning the rules for word combinations, there may be regarded numbers, vectors, matrices, functions and operations over them: addition, subtraction, division, differentiation, integration, etc; as an analogue to word descriptions there can serve mathematical models. The analogy of word descriptions with mathematical models is sufficiently direct and apparent; however, as for the analogy between a word combination and mathematical operations and rules, it is not so clear. Here, the specificity of the mathematical language declares about

itself, since its "grammar" incorporates not only rules of element combinations, i.e. combining of words, but also the rules for transforming the words into other words.

The first part of mathematics, i.e. elements and operations over them, can be naturally called an operational system and the second one, mathematical models.

To understand the assignment and general principles for constructing the operational system is possible, if one follows the history of how this operational system was established and developing. We emphasize here that one should realize the resulting logic of its establishment, and not the actual specific history being very knotty sometimes and inconsequent.

The development of the operational system started from positive integers. How they had come into being was lost in the far past. However, Roman and Arabic numerals were already well-known that time. Practical needs of those old times brought out the inventions of operations over integers: addition, subtraction, multiplication and division. They appeared in order to replace prolonged and labour-consuming recalculations for quantities of stuffs, goods, money, incoming from different places in different times and into different places. These numerical operations were not always executable; to eliminate this there were invented negative and fractional numbers loaded with some definite sense. For example, if a positive number stands for a cash, then a negative number implies a debt. As for irrational numbers there was a long delay with them: the Greek did not assign a number to the length of a hypotenuse in an isosceles rectangular triangle with unit legs; this length was not measured by a number. This lack of numerical measure was proved by the Greek in a very original way: they stated that a hypotenuse and legs possess no general measure, i.e. they have no segment that could be packed within each of them integer number of times. This obstacle being insurmountable for the Greek's thinking was later ignored for a long time, and only in the 19th century it was surmounted by G. Cantor, K. Weierstrass and R. Dedekind. In essence, rational numbers were augmented with real numbers, through the operation of passing to a limit, as a completeness requirement for this operation.

Complex numbers were invented on the same reasons, viz. in order to eliminate unsolvability of square equations, which existed because of the need to extract a root from a negative number.

William R. Hamilton found a generalization of complex numbers and invented quaternions. For quaternions there remained true all the rules for operations over real and complex numbers, except the property of commutativity. Then, it was proved that the generalization of such a kind is unique, although in future there have been invented a lot of hypercomplex numbers being similar to quaternions.

A quaternion is a hypercomplex number of the form

$$a + \alpha\, i + \beta\, j + \gamma\, k\,, \qquad\qquad (37.1)$$

where a, α, β, γ are usual real numbers, and i, j, k are unit numbers of a new nature, i.e. some symbols or unit elements for which the following multiplication table is valid:

	1	i	j	k
1	1	i	j	k
i	i	-1	k	-j
j	j	-k	-1	i
k	k	j	-i	-1

For quaternions, the operations of addition, subtraction and multiplication by a real number are the same as for complex numbers, but multiplication for quaternions is governed by the above table. From this table it is apparent that the quaternions, for which $\beta = \gamma = 0$ or $\alpha = \beta = 0$ or $\alpha = \gamma = 0$, are usual complex numbers and that multiplication for them is not commutative. Dividing the quaternions is possible due to the fact that a product of the quaternion (37.1) by the quaternion $a - \alpha i - \beta j - \gamma k$ makes up a real number; therefore, the operation of dividing the number (37.1) by the number $a_1 + \alpha_1 i + \beta_1 j + \gamma_1 k$, after multiplying the numerator and the denominator by the conjugate quaternion $a_1 - \alpha_1 i - \beta_1 j - \gamma_1 k$, is reduced to dividing this quaternion by a real number, and this is reduced to dividing by it the numbers a, α, β and γ.

The quaternions were used by James C. Maxwell for writing the equations suggested by him for an electromagnetic field. The so familiar vector form for writing the Maxwell equations was suggested only when a notion of vector and vector calculus had been generated by the quaternion calculus. It took place approximately as follows. Similar to a complex number consisting of a real and an imaginary parts, the quaternion $a + \alpha i + \beta j + \gamma k$ was decomposed into the real part a and the vector part $\alpha i + \beta j + \gamma k$; then, the product of two vectors

$$(\alpha i + \beta j + \gamma k)(\alpha_1 i + \beta_1 j + \gamma_1 k) = -\alpha\alpha_1 - \beta\beta_1 - \gamma\gamma_1 + \begin{vmatrix} i & j & k \\ \alpha & \beta & \gamma \\ \alpha_1 & \beta_1 & \gamma_1 \end{vmatrix}$$

was decomposed into a real part with minus and a vector part; these parts were called the scalar and vector products, respectively. This very fact has given birth to the vector calculus with operations of addition, subtraction, inner and vector products, and, of course, multiplication and division by a real number. Further, a notion of a vector has given birth to linear algebra, linear transformations, matrices and matrix calculus.

It is worth recalling here that quaternions are no uniquely invented hypercomplex numbers; but they are unique in their own way, i.e. such numbers in

which, as compared against real and complex numbers, only commutativity of a product is violated and associativity and distributivity retain.

A need in describing the evolutionary processes brought out an invention of functions and operations over   them, including differentiation, integration and operator transformations, i.e. led to a birth of mathematical and functional analyses, at first, for functions of a real variable only, then for many variables and a complex variable. Operators and transformations started to be interpreted, in their turn, as elements over which various operations are possible.

This was not a halt in the expansion of the mathematical operational system. This process is still proceeding and proceeding, but the above said is enough to formulate a basic stimulus for the development of the system and the requirements assigned to it. The developing stimulus is a trial to entirely describe everything that is needed and possible; the requirements imply a completeness of operations and their fundamental feasibility. A completeness makes up an essential specificity of the very mathematical language.

These principles of constructing an operational system – the possibility of a wider description   through   natural numbers and operations over them and the completeness requirement for newly invented operations – provide, to some extent, an answer to the perplexing question by A. Einstein: "How can it be that mathematics, being, after all, a product of people's thinking, independent of experience, is so nicely adapted to the objects of reality?"

The above said does not cover an entire concept of the mathematical operational system; there also exist sets, operations and relations between them. Operations include unification, intersection and complementation. Relations include belonging and non-belonging, inclusion and non-inclusion. The relations "larger" and "smaller" take place for real numbers and modules of numbers, functions and operators. Of course, for all of them there are relations of equality and identity.

Sets and operations over them can be attached to the earlier described operational system, though they may be thought to be another mathematical operational system. There exists, nevertheless, some distinction between them, and this allows to distinguish a specificity in each of them. The like specificities   and essential distinctions are available even among isomorphic operational systems being indistinguishable from the point of view of their preceding descriptions.

Two operational systems A and B are isomorphic, if between their elements and operations it is possible to establish the one-to-one mappings T and D, respectively, such that for any element $x \in A$ and any admissible operation $f \in A$ there takes place the below commutativity diagram
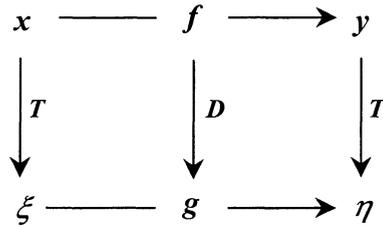
where $y = fx,\ \eta = g\xi,\ Tx = \xi,\ Ty = \eta,\ \mathrm{D}f = g.$

Due to this commutativity diagram, the isomorphic systems differ from each other, as it were, by their notations only. However, in reality this is not so, because the preceding narration did not describe the execution of operations. And this turns out to be of great importance. The operations can be assigned by some abstract agreement, in the form of multiplication table, but this is actually possible only for the finite numbers of elements, whereas their number is infinite. Hence, there must exist some other procedure, simple as much as possible, for implementing this abstract relation laid in the table.

Here, we will make it clearer by means of the example with Roman and Arabic notations for the similar integers. The numbers 156 and 267 in Roman notation are $CLVI$ and $CCLXVII$, respectively. To multiply 156 by 267 is not a difficult task for any junior class pupil, while to multiply the numbers $CLVI$ and $CCLXVII$ in the Roman notation is very difficult, and in case of four-digit numbers practically impossible. The Greek were usually trained in this for many years.

The above said touches not only multiplication of two numbers; more vividly it is seen on the operations of differentiation and integration. Even integrating the elementary functions is not always possible. But these operations can be very much simplified and always easily executed in the operational system created by Oliver Heaviside. In this system, it is very simple to differentiate or integrate the function $f(t)$, and the results will be, respectively, $pF(p) - f(0)$ and $F(p)/p$, where $F(p)$ is a notation of the function $f(t)$ in the Heaviside system, and p is a complex number.

The Heaviside operational system is isomorphic to our usual system and this isomorphism is established by the well-known Laplace transformation. Though in this system things are not good with multiplication – it is performed in a very complicated way, while in our usual system very simply.

One more thing left to be added, it is that the operational systems similar to Heaviside's are rather numerous. There exists another, absolutely different, isomorphic operational system, where almost all operations are performed easily and fast, although only approximately. This system is the operating system in a contemporary computer having powerful software, memory and high speed.

Why do we then need other isomorphic operational systems different from the customary mathematical one? An answer can be formulated as follows: in a number of cases it becomes impossible to solve some problem within a habitual mathematical operational system, but, if the problem is transferred to another operational system being isomorphic to the initial one, the solution may be then found easily. Having solved the problem in the isomorphic system, we may return to the initial system and obtain there a wanted solution. This opportunity is provided by the commutativity diagram. Indeed, let the operation f over the element x be unrealizable in the operational system A but possible in the operational system B. According to the commutativity diagram, to the operation f in the isomorphic system there corresponds the operation g and, as it is apparent immediately from this diagram, we have

$$y = fx = T^{-1}\eta = T^{-1}g\xi = T^{-1}gTx,$$

where all the operations T, g and T^{-1} are feasible, with the operation T corresponding to a transfer from the initial system A to the system B and the operation T^{-1} corresponding to a return to the initial system A.

Here, we stop our brief description of mathematical operational systems and pass to the analogues existing in the mathematical language – to phrases, descriptions, figures, etc of a natural language. As mentioned above, these analogues are mathematical models, i.e. isomorphic mappings of some real or imaginary objects obtained with the help of a mathematical operational system. To make the picture clearer, let us here give an abstract and formal definition of a mathematical model. Let the model be some set of elements and connections between them. Under elements and connections here we mean the elements of a mathematical operational system and the connections being established by its operations and relations. In accordance with this definition, the equation $x^2 + px + q = 0$ will be a model, the relation $y = ax + b$ will be a model as well, and x will be a model too. $dx / dt = z + x$ will be also a model and $x < y$ too. However, the notion of a model carries also some portion of informal sense. In order to distinguish what this sense is, we note here that the notion of isomorphism for operational systems retains valid for models as well.

Any two models are isomorphic if between their elements, operations and relations between them there may be established an one-to-one mapping for which the commutativity diagram is valid. Here, we speak about models belonging to similar or different operational systems.

Now, it remains to tell you the most important things (unfortunately not yet formalizable) concerning a mathematical model and a mathematical modelling. Our nature as the world surrounding us is also nothing but some kind of an operational system, since we manage to represent it, with this or that completeness, in the form of various elements and connections between them and current time. They are called natural laws and principles of nature organization. From this

point of view, natural objects are models of nature and these models can be iso-morphic to mathematical models. More precisely, the humanity has created the mathematical language in such a way as to have it true. And as soon as we manage to construct a mathematical model being isomorphic to a natural object, we obtain a possibility to comprehend – via studying this model – this natural object itself. In the above description, an informal thing is revealing the elements in the natural objects, sometimes imaginable and invented, and the connections between the elements and the current time, i.e. the informal thing here is how we comprehend an organization of the world and the operational system of nature and how we comprehend the mathematical operational system, being isomorphic to it, the models of which can be studied, investigated and understood. Hence, the science as a whole and the mathematical modelling as its part is not only an exact science but also an unformalizable art. In the history of science and separate investigations, formalizable and unformolizable, the exact science and the art become closely interrelated and difficult to be separated from each other. However, this separation, for example, became very vivid and clear-cut, when M. Faraday revealed electromagnetic phenomena and suggested an idea of a magnetic and an electric fields and J. C. Maxwell constructed a corresponding mathematical model, i.e. the famous Maxwell equations. In mechanics, there were devised forces and the material points possessing mass, being located in some places and having motion velocity. For their motions I. Newton suggested a mathematical model, the Newton equations; then J. L. Lagrange derived the Lagrange equations for systems of material points with holonomic ideal connections between them, but without necessity to know the forces being caused by these connections.

Our conscious   as well as unconscious comprehension of the world by our sensitive organs and brain are evidently resting upon constructing the isomorphic models. How these models are organized, constructed and comprehended remains one of the most important and so far unexposed mysteries of nature. With the help of this intuitive thinking, a human is consciously inventing much more powerful and consciously substantiated   mathematical models and the associated operational system necessary for their construction and study.

38 Geometrical, physical, analogous, mathematical and imitative types of modelling

Isomorphism and a commutativity diagram as a basis of geometrical, physical, analogous and imitative types of modelling. Similarity criteria and rules for modelling.

All kinds of modelling listed in the title are based on isomorphisms. Geometrical and physical types of modelling are based on a geometrical and, respectively, a physical similarities caused by scale length changes, and, accordingly, by changes in scales of such magnitudes as, for example, length, time and mass. An analogous modelling is possible due to identity of some kinds of regularities for magnitudes of different physical nature, such as mechanical and electrical magnitudes, processes of thermal conductivity, diffusion and formation of an electric field. Earlier, we were already discussing the isomorphism of real objects and mathematical models. A need in imitative modelling appears, because the unknown magnitude to be determined turns out to be coincident with the statistical expectation of a specially-arranged random process.

A geometrical modelling is used from ancient times: the laws of geometrical similarity were already familiar to ancient Greeks. A physical modelling came into being more than a century ago because of large catastrophes with sea ships; then, it was also stimulated by a progress in aviation. As for the analogous modelling, more precisely, the lying in its basis identity of natural laws for objects of different physical nature, this modelling was already clear to J. Rayleigh, and, in particular, in the form of analogies between the Thomson electric oscillatory contour and the mechanical harmonic oscillator. These very analogies and some other lesser analogies gave birth to the theory of oscillations as a science of general regularities in various evolutionary processes of mechanics, physics, chemistry, biology and the society science. Contemporary mathematical modelling has become advanced and widely applied, and, to a considerable extent, not only due to its theoretical success but also due to appearance of universal high-speed computers.

Some classes of problems successfully exploit the imitative modelling being also called the Monte Carlo method or the statistical modelling.

All these kinds of modelling are widely described in a numerous literature. Here, they are listed in order to emphasize the idea that all of them hold a common basis, an isomorphism. This is the isomorphism between different real objects, processes, on the one hand, and imaginable objects, processes expressed and represented on paper or some other information carrier, on the other hand.

Now, upon some general words, we will describe you in more details the physical and imitative types of modelling, and thereupon, separately, the mathematical modelling as a science and art.

38.1
Physical modelling

Physical modelling does not imply an actual availability of the mathematical model; instead, it is sufficient to rely only upon some small information and understanding of the object under study; namely, it should be only known what physical magnitudes completely determine a behaviour of the object in question, all necessary magnitudes should be present and no one superfluous. However, this kind of modelling is based on the fact that there exists some mathematical model and the physical magnitudes involved have some dimension; and this model is such that changing all basic scales in the physical magnitudes retains this model invariant – the scales are contracted. This is equivalent to the natural observable independence of natural laws from the acceptable scales for measuring   physical magnitudes. However, in principle, this might be not so, since the non-Euclidean geometry holds no usual laws of geometrical similarity.

Now, let the determinative magnitudes be $f_1,\ldots,f_m$ and the physical dimension of the magnitude f_i be $L^{\alpha_i},S^{\beta_i},M^{\gamma_i}$, where L,S and M are scales for length, time and mass. Let this data be written in the tabular form:

	L	S	M
f_i	α_1	β_1	γ_1
f_2	α_2	β_2	γ_2
.	.	.	.
.	.	.	.
.	.	.	.
f_m	α_m	β_m	γ_m

Changing the scales of L,S and M , respectively, K_L,K_S and K_M times will convert each determinative magnitude f_i into the below magnitude

$$\bar{f_i} = K_L^{\alpha_i}\ K_S^{\beta_i}\ K_M^{\gamma_i}\ f_i, \quad i = \overline{1,m} \ . \tag{38.1}$$

The new physical determinative magnitudes $\bar{f_1},\ldots,\bar{f_m}$ in fact will correspond to the same object and will be assigned by the same mathematical model. From this, it follows that the object characterized by the magnitudes $f_1,\ldots,f_m$ is similar to the object characterized by the magnitudes $\bar{f_1},\ldots,\bar{f_m}$, if and only if the relation (38.1) holds for some K_L, K_S and K_M. The same formulae (38.1) also make it possible to recalculate determinative magnitudes of one object into another object represented by the same determinative magnitudes and the same mathematical models, which are not obligatory to be known. In addition to the determinative magnitudes $f_1,\ldots,f_m$, the object can be also described by other magnitudes which must be determined by the above magnitudes, and, hence, must be again converted according to (38.1).

Let us now find the conditions for physical similarity, i.e. reveal when there exist such K_L, K_S and K_M for which the formulae (38.1) will hold.

Finding the logarithm of (38.1), one can write these formulae as follows

$$\alpha_i \ln K_L + \beta_i \ln K_S + \gamma_i \ln K_M = \ln \frac{\bar{f_i}}{f_i}, \quad i = \overline{1,m}. \tag{38.2}$$

The system of linear equations (38.2) in the unknowns $\ln K_L$, $\ln K_S$ and $\ln K_M$ has a solution, if the rank of the matrix

$$\begin{pmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \cdot & \cdot & \cdot \\ \alpha_m & \beta_m & \gamma_m \end{pmatrix} \tag{38.3}$$

is equal to the rank of the extended matrix

$$\begin{pmatrix} \alpha_1 & \beta_1 & \gamma_1 & \ln\dfrac{\bar{f_1}}{f_1} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_m & \beta_m & \gamma_m & \ln\dfrac{\bar{f_m}}{f_m} \end{pmatrix}. \tag{38.4}$$

This will be the case, if from the linear dependence for the rows of the matrix (38.3)

$$x_1 \alpha_1 + \ldots + x_m \alpha_m = 0$$
$$x_1 \beta_1 + \ldots + x_m \beta_m = 0 \tag{38.5}$$
$$x_1 \gamma_1 + \ldots + x_m \gamma_m = 0,$$

where $x_1^2 + x_2^2 + \ldots + x_m^2 \neq 0,$ it will follow that

$$x_1 \ln \frac{\bar{f}_1}{f_1} + \ldots + x_m \ln \frac{\bar{f}_m}{f_m} = 0$$

or

$$f_1^{x_1} \cdots f_m^{x_m} = \bar{f}_1^{x_1} \cdots \bar{f}_m^{x_m}, \tag{38.6}$$

where all magnitudes of this equality are of the zero dimension. A fulfillment of these equalities for any solution $x_1, \ldots, x_m$ of the system (38.5) will be, thus, a necessary and sufficient condition for a physical similarity. It is clear here that obeying these conditions for independent solutions $x_1, \ldots, x_m$ of the system (38.5) will be sufficient to obtain this similarity. To each independent solution $x_1, \ldots, x_m$ there corresponds its condition of similarity; they are so many as the number of the independent solutions of the system (38.5). Such are the similarity conditions. If they are satisfied, the conversion formulae are derived from solving the linear equations (38.2) in the unknowns K_L, K_S and K_M. These formulae are ambiguous to within the similarity conditions being similar for an object and its model.

Let us pass on to examples.

Example 1. Let a statically determinated core structure to be manufactured be composed of rods of the lengths $l_1, l_2, \ldots, l_n$ joined in the nodal points where it is loaded with some forces $F_1, \ldots, F_n$; and we want to estimate the loads $G_1, \ldots, G_n$ upon all rods, to check their sufficient strength. We want to solve this problem by creating some reduced model and through estimating loads in rods experimentally. First of all, it is evident that the structure to be constructed and its model must be geometrically similar, since, otherwise, the ratios l_i / l_j, $i \neq j$ for them would be different. For the same reason, proportionality for the loads applied to the structure must be obeyed. When these conditions fulfilled, it is thought sufficiently proved that in this case the determinative magnitudes will be a distinctive size and an applied load, for example, l_1 and F_1. For this case, the table will then be as follows:

		L	S	M
x_1	l_1	1	0	0
x_2	F_1	1	-2	1

For this table, the equations (38.5) will be of the form

$$x_1 + x_2 = 0, \qquad -2x_2 = 0, \qquad x_2 = 0$$

and will have a zero solution only. Therefore, physically similar will be any described structures with proportional external loads; the loads $G_1,\ldots,G_n$ will be estimated through the analogous loads $\bar{G}_1,\ldots,\bar{G}_n$ in the model by the formula

$$G_s = \frac{F_1}{\bar{F}_1}\bar{G}_s, \qquad s = \overline{1,m}. \tag{38.7}$$

Example 2. Let in the same structure, not being necessarily statically described, the rods be elastic and with the elasticity coefficients $k_1,\ldots,k_p$. The above considerations imply that the structure should be geometrically similar to the model, and the relations between external loads and, respectively, between the elasticity coefficients should be similar as well. Now, the basic things will be the distinctive size l, the external load F and the elasticity coefficient k; for example, l_1, F_1, k_1. As before, let us form a table

		L	S	M
x_1	1	1	0	0
x_2	F	1	-2	1
x_3	k	0	-2	1

and write the equations for searching for similarity criteria

$$x_1 + x_2 = 0, \qquad -2x_2 - 2x_3 = 0, \qquad x_2 + x_3 = 0.$$

Now the equations have a nonzero solution $x_1 = 1, \ x_2 = 1, \ x_3 = 1$; and to it there corresponds the similarity criterion $lF^{-1}k$. For similarity purposes, these

magnitudes in the structure under design and in the model must be identical, and the loads upon the rods will be again, as before, governed by the formulae (38.7).

Example 3. As a next example we take a pendulum. A pendulum is a mechanical system described by its length l , mass m and the gravitational acceleration g . This system is presented by the below table

		L	S	M
x_1	l	1	0	0
x_2	m	0	0	1
x_3	g	1	-2	0

and by the equations

$$x_1 + x_3 = 0, \qquad -2x_3 = 0, \qquad x_2 = 0,$$

having no nonzero solution; therefore, all pendulums are similar between themselves. This conclusion immediately follows from their mathematical model

$$J\ddot{\varphi} + lmg \cdot \sin \varphi = 0,$$

since after a change of the time scale, $t \to \sqrt{lmg / J}\,\tau,$ this equation is reduced to the form

$$\frac{d^2 \varphi}{d\tau^2} + \sin \varphi = 0.$$

It means that all pendulums are physically similar.

Further, suppose that we want to find the oscillation period T by using this physical similarity. To do it immediately is impossible, since the period T depends not only upon the magnitudes l , m and g but also upon the amplitude a of pendulum oscillations; therefore, it should be added into the above table. After this addition, there will appear one more variable x_4 ; and the corresponding equations will then be as

$$x_1 + x_3 + x_4 = 0, \qquad -2x_3 = 0, \qquad x_2 = 0.$$

Now, we have already obtained a nonzero solution: $x_1 = 1,$ $x_2 = 0,$ $x_3 = 1,$ and to it there corresponds the physical similarity criterion a / l . If it is satisfied, the ratio between the oscillating periods T and $\overline{T}$ will be determined

by the quotient of magnitudes $\sqrt{\dfrac{lmg}{J}}$ and $\sqrt{\dfrac{\overline{lmg}}{\overline{J}}}$ being similarly dimen-

sional to these periods, i.e. we obtain

$$T = \overline{T} \cdot (lmg / J)^{1/2} \cdot (\overline{lmg} / \overline{J})^{-1/2}$$

or

$$T = (lmg / J)^{1/2} \cdot f(a / l),$$

because $\overline{T} \cdot (\overline{lmg} / \overline{J})^{-1/2}$ has no dimension and depends upon the quotient

$\overline{a} / \overline{l} = a / l$ only. As known, for $a / l \ll 1$, we have $f(a / l) = 2\pi$.

   Example 4. In this final example let us find a similarity criterion for some stationary flow of viscous fluid in a gravitational field, when the flow of fluid is represented by the distinctive size l, the distinctive velocity υ, the density ρ, the viscosity ν and the gravitational acceleration g. Under these assumptions, the appropriate table and the equations, describing the physical similarity criterion, will be of the form

		L	S	M
x_1	1	1	0	0
x_2	υ	1	-1	0
x_3	ρ	-3	0	1
x_4	ν	2	-1	0
x_5	g	1	-2	0

$$x_1 + x_2 - 3x_3 + 2x_4 + x_5 = 0, \quad -x_2 - x_4 - 2x_5 = 0, \quad x_3 = 0.$$

The equations for $x_1, \ldots, x_5$ have two independent solutions – $x_1 = 1$, $x_2 = 1$, $x_4 = -1$ and $x_1 = -1$, $x_2 = 2$, $x_5 = -1$, to which there correspond the famous numbers of Reynolds and Froude

$$\mathrm{Re} = \frac{l\upsilon}{\nu}, \quad Fr = \frac{\upsilon^2}{lg}.$$

   As an example of comparatively recent sensational application of modelling with use of the dimensions of physical magnitudes and general simplified representations of the phenomenon under simulation, there may be taken a determination of the security-restricted energy of the first nuclear explosion through ana-

lysing the film demonstrated in public. The explosion was assumed point-wise and an arising shock wave spherical. In these simplified representations, the radius R of the shock wave is a time function depending upon the energy E of the explosion and the initial density ρ of the surrounding air. The corresponding table for dimensions of the energy E, the density ρ , the time and the radius R of the shock wave is of the form

	E	ρ	t	R
L	2	-3	0	1
M	1	1	0	0
T	-2	0	1	0

A search for a dimensionless combination of the determinative magnitudes E, ρ and t results in the following system of equations

$$2x - 3y = 0, \quad x + y = 0, \quad -2x + z = 0,$$

having only a zero solution $x = y = z = 0$; hence, all the point-wise explosions with a spherical shock wave will be similar to each other. Thus, according to the above table, we find that the radius R is expressed through the determinative physical magnitudes E, ρ and t by the formula

$$R = CE^{1/3} \rho^{-1/3} t^{2/3},$$

where C is some constant able to be determined by a single specific experiment or by a theoretical analysis having shown that $C = 1$. Now, the derived formula allows us to determine the energy E of the nuclear explosion, provided that the radius R of the shock wave at some t is known. However, E may be estimated in a more reliable way, with a help of a series of time values t and the associated values of R . By finding an logarithm of the formula for the dependence of R upon E, ρ, t and taking into account that $C = 1$, we find that

$$\frac{5}{2} \ln R = \frac{1}{2} \ln \frac{E}{\rho} + \ln t.$$

Accordingly, the graph of dependence of $\dfrac{5}{2}\ln R$ upon $\ln t$ is depicted by the

straight line cutting off the segment of the length $\dfrac{1}{2}\ln\dfrac{E}{\rho}$ on the ordinate axis.

Onto this very line there should be laid the data from the film on the nuclear explosion; on this basis the energy of the nuclear explosion can be found.

38.2
Imitative modelling

Here we are beginning a short narration about imitative modelling. This narration starts with a historical example of searching for the number π by means of randomly tossing a needle onto a sheet of paper divided into identical stripes. In the past, even well-known scientists were fond of this fun by tossing a needle thousand times – the more times you will toss, the more exactly you will determine the number π.

Under a random tossing it was understood that a needle, upon reaching a sheet and stopping, shows its position regarding the parallel lines drawn on the sheet in such a way as to show that the nearest distance x of needle's centre and the angle φ with respect to these lines will be independent random variables of constant probability densities $1/l$, where $2l$ is a width of stripes, and $1/\pi$, respectively. The needle of the length $2l$ will intersect with one of the lines, if

$$l\cos\varphi \geq x .\qquad\qquad(38.1)$$
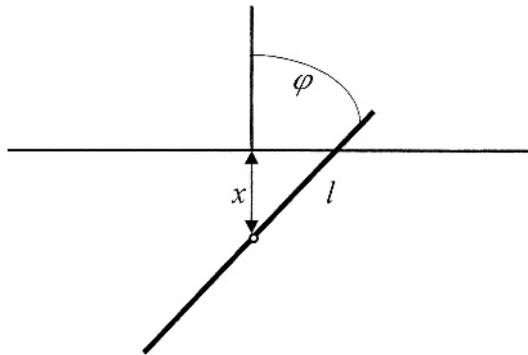
Fig. 38.1. The needle intersecting with one of the lines.

In accordance with the notations adopted, the probability that the needle will intersect one of the lines is equal to

$$\iint_D \frac{1}{l\pi} dx d\varphi , \tag{38.2}$$

where D is a domain described by the inequalities $0 \le x \le l$, $-\dfrac{\pi}{2} \le \varphi \le \dfrac{\pi}{2}$

and (38.1). The integral (38.2) is taken easily and equal to $2/\pi$.

Now, let the needle have been tossed N times and have intersected one of the lines n times. In accordance with the law of large numbers, the frequency of intersections n/N approaches the found probability $2/\pi$, so that we approximately obtain

$$\pi \approx \frac{2N}{n} ;$$

here, the more the tossing number N is, the more exact the probability will be. The well-known Tchebyshev inequality makes it possible to estimate the probability of the error not exceeding some small ε. According to this inequality, we shall have

$$P\left(\left| \frac{n}{N} - M\frac{n}{N} \right| > \varepsilon \right) < \frac{D\dfrac{n}{N}}{\varepsilon^2} , \tag{38.3}$$

where Mn/N and Dn/N are the expectation and the dispersion, respectively, for the intersecting frequency n/N. The frequency will be

$$\frac{n}{N} = \frac{1}{N}(\xi_1 + ... + \xi_N), \tag{38.4}$$

where $\xi_i = 1$, if in the i-th toss the needle has intersected one of the lines, and if not, then $\xi_i = 0$. The random variables ξ_i are independent; for each of them the probability to be unity will be equal to $p = 2/\pi$. Further, according to (38.3) and (38.4), we determine immediately that

$$P\left(\left| \frac{n}{N} - p \right| > \varepsilon \right) < \frac{p(1-p)}{N\varepsilon^2} , \tag{38.5}$$

i.e. the probability that the frequency n/N will deviate from the value p more than by ε will not exceed $\dfrac{p(1-p)}{N\varepsilon^2}$ and will converge to zero as $N \to \infty$.

Thus, by randomly tossing a needle, π can be determined approximately and with whatever wanted arbitrary large accuracy. However, the number of these tosses should be large enough. For instance, to find three exact digits of π, i.e. for $\varepsilon < 10^{-3}$, with the probability 0.99, one will need to toss a needle about 10^8 times. Of course, to-day to calculate the number π in such a way is unreasonable, at least. However, there are examples where this is not only possible, but, perhaps, uniquely acceptable. For example, this is the case in calculating multidimensional definite integrals of functions of many (say, ten) variables of the form

$$\int_D f(x_1,...,x_n) \cdot dx_1 \cdots dx_n \qquad (38.6)$$

over some restricted domain D determined by some inequalities, which allow to reveal whether the point $(x_1,...,x_n)$ belongs to this domain or not.

Let $a_1 \le x_1 \le b_1,...,a_n \le x_n \le b_n$ and $A \le f(x_1,...,x_n) \le B$. The first n inequalities and the inequality $A \le z \le B$ determine the $(n+1)$-dimensional parallelepiped Π in the $(n+1)$-dimensional space R_{n+1} of the variables $x_1,...,x_n$ and z. The value of the integral is nothing else than a $(n+1)$-dimensional volume of the domain

$$0 \le z \le f(x_1,...,x_n), \quad (x_1,...,x_n) \in D \qquad (38.7)$$

in the space R_{n+1}. Denote it by $\overline{D}$. The domain $\overline{D}$ lies inside the parallelepiped Π. Let the volumes of the domains $\overline{D}$ and Π be denoted through $\overline{V}$ and V. Then, $\overline{V}$ will be a needed value of the integral (38.6) and V will be equal to $(B-A)\prod_i(b_i - a_i)$. The ratio $\overline{V}/V$ is nothing but the probability of the random point $M(x_1,...,x_n,z)$ falling within the domain $\overline{D}$, provided that this point has the constant probability density $[(B-A)\prod_i(b_i - a_i)]^{-1}$ inside this parallelepiped Π. It means that the magnitude $\overline{V}/V$ can be determined via independent random "tosses" of the point M in an absolutely similar way as it was done with a needle for estimating the number π. Of course, here a lot of "tosses" will be needed. However, the most remarkable thing here is that the necessary number of "tosses" does not depend upon the dimension n of the integral. Random tosses may be performed by random number generators. Upon

each "toss", it is immediately found – in accordance with the inequalities (38.7) determining the domain   – whether the "tossed" point has fallen into the domain   or not.

39   General scheme of mathematical modelling

Basic stages in constructing a mathematical model.

Mathematical modelling involves a solution of two problems – constructing a mathematical model and its investigation . These problems are very different and require various knowledge, habits and intuition, but in spite of this they are closely interconnected, since when constructing a model one should take into account difficulties and possibilities of its study ; an investigation of this model can reveal a necessity of correcting the model constructed. Besides, when investigating and making conclusions flowing from the investigation, one should understand what should be accounted in the model and what not, i.e. to what questions the model should be able to give answers and to what not .

The mathematical models considered in this book can be called simple. They are simple in writing and for investigating, though they are not always simple from the point of view of the results obtained. The results of investigating the object in question may be unexpected and may need some explanation. The models are simple at the expense of simulating only some sides of the object and not approaching an all-round and complete adequacy with its real object; therefore, they, naturally, can explain only what has been put and taken into account; sometimes these explanations are only qualitative and their quantitative explanations are very approximate. In spite of this, and may be due to this, the role of simple models in science is very high, for they are simple and bring to understanding an essence of the existent processes. This understanding is very important both for a scientist, an engineer, a designer and an inventor. However, this necessary understanding is not always sufficient and in many cases more exact quantitative data is needed.   This, in its turn, demands additional improvements and complications in the model that brings about a need in numerical research being, as a rule, more difficult. In spite of its importance, this numerical side of the mathematical modelling has not been actually touched in this book. The book sooner touches a problem of how to escape from these computations, if possible and admissible.

Constructing a model and investigating it can require from you not only sufficient knowledge, but also some art. This art, in some sense, lies beyond the science, since it is a result of some individual experience and that of training by means of showing examples. Art can be learned as well. Indeed, man is trained to write poems, music, to draw and to play violin, and to dance. It has to be admitted here that a contemporary   traditional training in mathematics, theoretical and

partly applied, does not promote this art; s a result, only some personalities gifted by nature are able to timely overcome a barrier of the mathematical formalism. More successfully they overcome this obstacle in scientific schools thanks to picturesque examples, intercourse, talks, discussions and demonstrations. You have possibly already guessed and recalled that the immediate aim of this book is to make it easier for a reader to penetrate through the barrier between the theoretical mathematics and the applied mathematical modelling.

Thus, suppose we have to construct and investigate a mathematical model, i.e. you are facing a problem of mathematical modelling. What should you at first start with? First of all, do not save your effort for thinking it over, improving and specifying it, and try to look at your problem from different points of view. As precisely and completely as possible you will need to answer the following questions: What do I want to find out? What questions should I answer? What object should be taken for this purpose? What things in the object should be obligatorily taken into consideration? And what things in it must or may be neglected? This most important stage of mathematical modelling can be called "an initial statement of the problem and a determination of the object to be investigated". You should not think that in future you will not deal with this stage. Conversely, it is highly possible that you will have to deal with this stage again, and may be even at the time when you will think that your problem has been already solved. After this stage, you can begin with the second one that covers a verbal or a conceivable description of the object and that of natural laws governing this object. Upon it, again you have to think it over whether it is possible for you to neglect something or, vice versa, to take into account something additionally. As a result of these steps, you will finish the conceivable or verbal description of the object in question and the clarification of the natural laws governing the object's behaviour and properties. Further, this verbal and conceivable description should be isomorphically mapped into a mathematical model. This mapping should be better done in the most simple and convenient way, for subsequent investigation. At this step, do not neglect any possible opportunities for simplifications or decompositions leading to splitting the problem of your research into sequential stages. All these actions will lead you to an initial mathematical model. Further, you will have to think about how to study and investigate this model. Perhaps, upon this thinking you will somehow modify your model, thus increasing its simplicity, convenience and efficiency of its investigation. Then, you should study the model and make conclusions about the properties of the object, its behaviour and possible phenomena in it; then, all this data should be matched against your knowledge of the object. If something does not coincide and doubts raise or something does not suit you completely, you should return to one of the preceding stages, or even to the initial problem statement. If everything suits you, then you may pass to formulating conclusions and recommendations. Of course, the above said should be thought to be a general scheme only, and you are not invited to follow it literally.

Now, finishing this short and general description of the complicated and many-sided mathematical modelling, illustrated many times above and to be also

illustrated below, let us make a list of its main stages and then recommend some general considerations useful for modelling:

   1) stating a problem and determining the object under study;

   2) performing idealization and simplification of the object;

   3) creating a conceivable and verbal description of the object and describing the natural laws defining object's behaviour and properties;

   4) simplifying further the object's model or, on the contrary, taking into account   something new;

   5) constructing an initial mathematical model;

   6) correcting the mathematical model;

   7) choosing and implementing the investigating techniques for the mathematical model;

   8) comparing the results of the model investigation against your knowledge of the object;

   9) making conclusions and recommendations.

Note here that returns from each stage to any previous one are possible. And now below there are some useful general considerations:

   1) the simpler a model, the less possibilities for erroneous conclusions;

   2) a model must be simple, but not simpler than possible;

   3) you may neglect anything you like, but only know how this will influence on your solution.

   4) a model must be robust, and small and admissible improvements for it should not change its behaviour cardinally and essentially;

   5) a model and a calculation should not be more exact than the initial data. Conclusions should not be essentially changed within the bounds of initial data errors and calculation errors. If this cannot be managed at the expense of model modifications, more exact initial data is then needed;

   6) while analysing the results of the model study, it is important to know not only specific numerical results, but to understand also why and how everything

happens and how all this depends upon the parameters. At least, you need to strive
to it, although sometimes success will come far not immediately.

40 Models of vibratory tongue driving

Vibratory pile driving and its modelling.

Once I saw myself how a vibratory 18- metre thin metallic tongue was being driven into a frozen ground, and the creator of this wonder, engineer D.D. Barkan, having striken with a pencil across the tongue, showed me how the tongue was vibrating and sinking (Fig. 40.1).
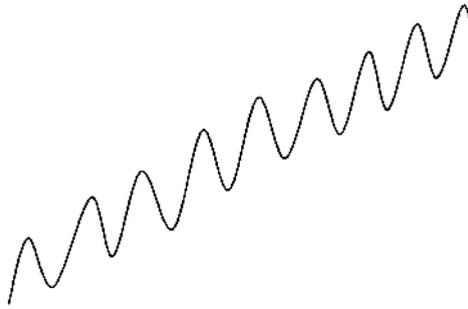
Fig. 40.1. The oscillogram of vibratory tongue driving, obtained by tracing with a pencil across the tongue.

According to Barkan's thinking, the reason of tongue sinking lied in the phenomenon experimentally observed by him, and namely, in fluidifying the ground under influence of vibrations. The problem he was constantly fighting with was that the vibrators usually got disabled very fast and a huge heap of them, broken and useless, was lying not far from the site. It happened at the construction site of the dam for the Gorky hydroelectric power station in winter of 1961. I was sent there by the Gorky Research Institute for Physics and Engineering (GIFTI) of the Gorky State University after N.I. Lobatchevsky for assisting them to fight these troubles. Naturally, I could not assist them directly and immediately, but soon found out the reason responsible for such a wonderful sinking of the tongue into a frozen ground, as if it was water. The secret there was in averaging the resisting

forces under the influence of high-frequency vibrations (50Hz). This idea was used as a basis for constructing the below models and for determining the dependences of the driving velocity upon their parameters. These results were also used to create still more efficient designs of silent pile drivers, where, in particular, vibrators did not break down so fast.

Now, I would like to narrate in brief how this unusual phenomenon was thought over and how its admissible mathematical model was constructed. The results of investigating this model were then exploited as a basis for an engineering design.

At first, I got acquainted with the publications of D.D. Barkan on fluidifying the ground by vibrations and on applying this phenomenon to vibratory pile driving. The author also wrote about experimental sinking of a heavy small-size ball into the vibrating ground, but this news was not consistent with the theoretical considerations and with the experimental data about a real vibratory driving of piles and tongues. Moreover, it seemed apparent for me that this fluidifying cannot take place in the frozen ground, while the tongue was penetrating through it rather successfully. The first idea was that the tongue is sinking into the ground like a snake, at the expense of longitudinal waves running along it. However, this idea had been then rejected, because a very high velocity of wave propagation along a small-sized tongue could drive it as a whole. Then the following very natural model was accepted by me – the tongue from its sides is subjected to the Coulomb friction and by its butt-end is tamping down the ground and pulling it aside. Schematically, this model is represented in figure 40.2; its relevant mathematical model is of the below form

$$m\ddot{x} = P + F \cdot \cos \omega t - F_\text{s} - F_\text{f} \ , \tag{40.1}$$

where $F \cos \omega t$ is a vibrator harmonic force acting upon the tongue; P is a tongue weight, together with the vibrator; F_s and F_f are the side forces and frontal resistances; x is a tongue replacement being counted down. The ground from the sides of the tongue is assumed to be motionless; therefore,

$$F_\text{s} = Q \, \text{sgn}(\dot{x}) . \tag{40.2}$$

The level of the ground under the butt-end of the tongue has the coordinate y, and an increase of the coordinate will require some effort R. If x is a coordinate of the lower butt-end of the tongue and y is a level of the ground punched by the tongue, then we shall obtain

$$F_\text{f} = \begin{cases} R , \ \text{for} \ x = y \ \text{and} \ \dot{x} > 0 \\ 0 , \ \text{for} \ x < y \ \text{or} \ x = y \ \text{and} \ \dot{x} \le 0 \end{cases} \tag{40.3}$$

and

$$\dot{y} = \begin{cases} \dot{x}, & \text{for } x = y \text{ and } \dot{x} > 0 \\ 0, & \text{for } x < y \text{ or } x = y \text{ and } \dot{x} \leq 0. \end{cases} \tag{40.4}$$
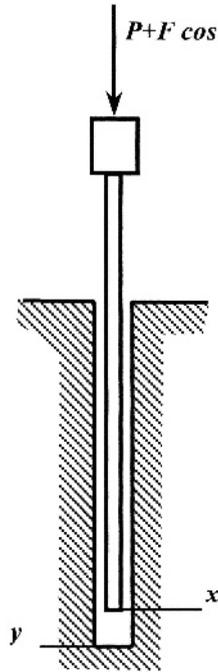
Fig. 40.2. The simplified model of the vibrator-carrying tongue being driven into the ground, with its side (40.2) and frontal (40.3) resistances only accounted.

The contemporary solution of the equations (40.1)-(40.4) by computers is not very difficult; but at that time this was a complicated and time-consuming calculating problem. However, it became possible to find its approximate solution analytically , by supposing , according to the tongue motion oscillogram shown to me by Barkan, that

$$x = a \cos \omega t + vt, \tag{40.5}$$

where a is an amplitude of tongue oscillations, ω is a frequency, and v a sinking velocity. By this supposition, integrating the equation (40.1) along the period and considering that the relation (40.5) takes place, we immediately determine that

$$
\int_0^{2\pi/\omega} m\ddot{x}\,dt = P\frac{2\pi}{\omega} - \int_0^{2\pi/\omega} F_s\,dt - \int_0^{2\pi/\omega} F_f =
$$

$$
= \frac{2\pi}{\omega}P + \left(t_1 - t_0 + t_1 - t_0 - \frac{2\pi}{\omega}\right)Q + \left(t_0 + \frac{2\pi}{\omega} - t_3\right)R = 0.
$$

(40.6)

The values of the time instants t_0, t_1, t_3 are indicated on the oscillogram of the tongue movement (40.5) presented in figure 40.3.
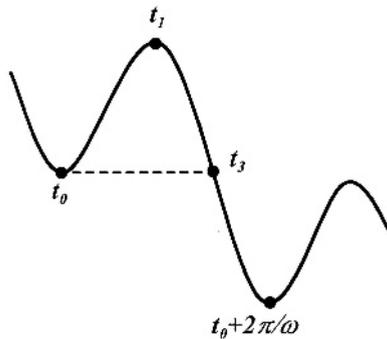
Fig. 40.3. The experimentally obtained graph of the tongue oscillations with indication of specific time instants t_0, t_1, t_3 and $t_0 + 2\pi/\omega$.

They are computed from the obvious equations and then the equation (40.6) will be approximately as follows

$$
\frac{2\pi}{\omega}P - \frac{2v}{a\omega^2}Q - \frac{2v}{a\omega^2}R = 0,
$$

from which

$$
v = \frac{\pi a\omega}{Q+R}P.
$$

(40.7)

This "surprising" result may be interpreted as the fact of averaging the forces F_s and F_f under influence of vibrations and that these forces turn out to be equivalent to the viscous friction with the coefficient $\dfrac{Q+R}{\pi a \omega}$. This coefficient of viscous friction decreases as the amplitude a and the frequency ω increases. Roughly estimating $F >> Q + R$ and $a = \dfrac{F}{M\omega^2}$, we obtain the viscous friction coefficient equal to

$$h = \frac{Q+R}{\pi\, a \omega} = \frac{(Q+R)M\omega}{\pi\, F} \quad . \tag{40.8}$$

Thus, while the vibratory pile is being driven, the resistant forces of the ground are, indeed, similar to the viscous friction, but this happens so not due to the ground fluidification; and here no fluidifying is needed.

This model is very simple and explains fully the essence of the driving phenomenon, but it does not bring good coincidence with experiments. The matter is that the ground possesses also some elasticity, both from the sides of the tongue and in the area of its butt-end. This elasticity shows itself in the form of some clearly or unclearly expressed resonance phenomena: the tongue in the ground has some proper frequency and it is similar to an oscillator. A more complete model embracing also the ground elasticity is depicted in figure 40.4.

This model is already described by three variables, x, y and z. Here, x is a position of the tongue, y is a replacement of the side ground, and z is a replacement of the bottom of a dimple in the ground. As earlier, a slipping motion of the tongue and forcing through the ground take place under the forces Q and R, respectively, but now the side and the bottom of the ground have elasticity k_s and k_b, respectively. Then, the mathematical model will consist, as above, of the equation (40.1) for x, the equations for y and z, and the following formulae assigning the forces F_s and F_f:

$$\dot{y} = \begin{cases} \dot{x}, & \text{for } \left|k_s y\right| < Q \\ 0, & \text{for } \left|k_s y\right| = Q \end{cases}$$

$$\dot{z} = \begin{cases} \dot{x}, & \text{for } k_b(x-z) = R \text{ and } \dot{x} \geq 0 \\ 0, & \text{for } k_b(x-z) < R \end{cases} \tag{40.9}$$

$$F_s = k_s y, \quad F_f = \begin{cases} k_b(x - z), & \text{for } x > z \\ 0, & for \quad x \leq z. \end{cases}$$
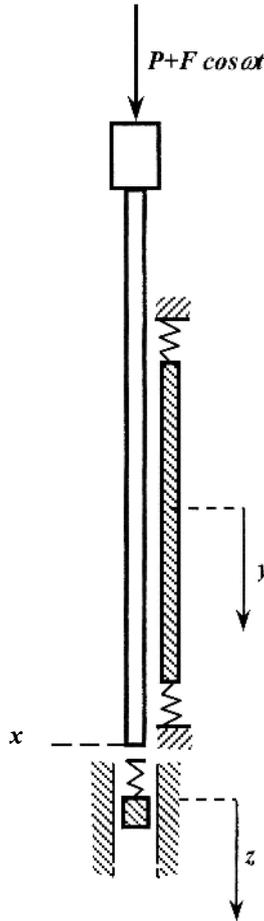
Fig. 40.4. The improved model, described by (40.9), for vibratory tongue driving with a simplified account of elasto-plastic properties of the ground under the side and frontal resistances; the mass of the ground is not taken into account.

To-day, to approximately study the model (40.1), (40.9) by the same very simple method is hardly useful, since the formulae will be cumbersome and not always acceptable. Instead, computer calculations may be done easily and fast.

The essentially new here is a substantially different dependence of the tongue oscillating amplitude upon the frequency and the arising opportunity for it to grow essentially, if the frequency is arriving at the resonance frequency of

tongue's oscillations. In this advanced model, the effect of the transfer of resistance forces (during vibrations) to viscous friction remains the same, whereas with the growth of amplitude, the "viscous friction", as before, will decrease, and, hence, to work in the resonance regime becomes profitable. This is, probably, all. Perhaps, it is worth noticing here that a further perfection of this vibratory driving method has brought a jump from the centrifugal vibrators (the generators of the force $F\cos\omega t$) to the vibroimpact mechanisms, which, to some extent, have turned out more preferable. The simplest model of a vibroimpactor was studied by the point mapping method in [1,2] and the above narration was published in [3,4]. A more complete mathematical study of the models for vibratory driving by the point mapping method was done in [5,6].

References

1. Neimark, Ju. I. (1972): Method of point mappings in the theory of nonlinear oscillations . Nauka, Moscow (in Russian).
2. Bespalova, L.V. (1957): To a theory of vibroimpact mechanism. News of Russian Academy of Sciences. 5 (in Russian).
3. Neimark, Ju. I. (1952): The theory of a vibratory pile driving. Hydroelectric Development. 4, 24-27 (in Russian).
4. Neimark, Ju. I. (1953): The theory of a vibratory pile drive and pull out. Engineering Transactions. 16, 13-48 (in Russian).
5. Batalova, Z.S. (1966): Studying a vibratory pile driving under a head resistance of ground. Mechanics of a Rigid Body. 2 (in Russian).
6. Batalova, Z.S., Kantorovitch, E.B., Trousenkova, A.I. (1966): Studying a vibratory pile pull out. Mechanics of a Rigid Body. 5 (in Russian).

41 Fundamental mathematical model of the contemporary science and the theory of oscillations

An all-embracing role of the oscillations theory and a model of dynamical system. Auto-oscillations, regular and chaotic motions, homoclinic structures. A.A. Andronov as a founder of the theory of oscillations.

In its essense, this chapter is a modified version of my reports at the international conferences dedicated to the memory of Ya. Z. Tsypkin (Moscow, 1998) and of his and my teacher A.A. Andronov (Nizhny Novgorod, 2001), the personalities being known worldwide. Both reports of mine are interrelated thematically and lie within the scope of this book, viz. in the basis of the oscillations theory there lies the mathematical model "dynamical system", which is the very fundamental model of the contemporary science and a basic model in this book. Alexandr Alexandrovitch Andronov, being a pupil of L.I. Mandel'schtam, is a worldwide distinguished founder of the new science, theory of nonlinear oscillations. That is why it is natural to connect the presentation of the matter, role and place of the theory of oscillations in science with a talk about him .

41.1
A dynamical system as a basic mathematical model of the contemporary science

A scientific comprehension of the world came to humanity from heaven. It is impossible to imagine any other object but the Sun, Moon and planets against the background of a majestic starry night or daily sky simultaneously attracting our attention so much and being so significant, mysterious and meanwhile possessing simple and bright regularities. This object is not weather coming down to us also from heaven, because its caprices are difficult to catch but the Sun, the Moon, planets and stars. What is predictable is closely connected with the Sun and stars , revealing an invariable causal connection. Ptolemey, Copernicus, Galileo, Newton are the habitual landmarks in the history of studying the sky; and , at last, Laplace who carried the causal conditionality of the celestial phenomena to the earthly phenomena and proclaimed the causal conditionality and predictability for all phenomena of the world and its changes, i.e. for everything happening in the world. This was done by that very Laplace who was developing the probability theory and introduced into this science the most-known random variable with the distribution named after him.

The Laplace determinism was based on the tremendously impressing successes in the astronomy and in the mechanistic interpretation of the world. The daily observable randomness did not confuse him . It was only later on that the entire world phenomena were classified as predictable and unpredictable (random), being sharply opposed to each other. However, to-day the determinism principle carries a universal sense. In its extended interpretation, it retains also valid for the markovian stochastic processes, statistical physics and quantum mechanics. This principle lies in the basis of our understanding of the world and is embodied in the dynamical system model involving the state (a description) and the single-valued operator to predict the state. The state may be not immediately observable and measurable, but it is somehow related to observable and measurable magnitudes.

Another historical landmark is J.H. Poincare. The future follows the present and is defined by it. This principle of determinism is implemented in the descriptions of natural phenomena through differential equations. Prior to Poincare, the efforts of researchers were directed to obtaining their solutions – analytical, approximate and numerical. Poincare started to investigate   an entire totality of solutions, through their geometrical interpretation in a phase space. This created a new topological representation for a whole totality of all solutions as a decomposition of the phase space into phase trajectories. As a result, this approach gave birth to the qualitative theory of differential equations and the theory of dynamical systems. A phase portrait and a bifurcation portrait of the dynamical system made it possible to construct a full qualitative representation of dynamics, properties and phenomena in the object under study, which may occur with time and with change of parameters.

Constructing the mathematical models for evolutionary processes in the form of dynamical systems is a search , a devising, an inventing of some self-sufficient description – state – and an operator of its change in time. In its search for this self-sufficient description, the humanity has demonstrated a great deal of inventiveness. In mechanics, under a state there are implied the positions and velocities of all material points; in electrodynamics, Faraday's fantastic electric and magnetic fields; in quantum physics, the famous function ψ ; the state of stochastic markovian processes and of physical phenomena described by them is a probability density. A state is constructed by man as a self-sufficient description on the basis of which it becomes possible to predict this description. Biological and social studies will, probably, reveal absolutely new and unexpected constructions of the state.

As far as control systems are concerned, a state for them is not only a self-sufficient description for prediction but also a complete information for organizing a control. The notions of controllability and observability are based upon a state, its controllability and observability.

A dynamical system as a mathematical model is studied for more than a hundred years already. On the one hand, results of this study are enormous and very much has been learnt, but, on the other hand, a contemporary sufficiently full study of the specific multidimensional system still remains often a very difficult and unsolvable problem. A regular success here has been attained only for the

dimension $n \leq 2$; for $n = 3$, to obtain a result is difficult and possible only with use of new computers; for $n \geq 4$, it is very difficult and often impossible.

Until recently, there existed a gap between well-known specific systems, having mainly a simple-structured phase portrait and simple motions, and complicated structures and Poisson-stable motions, exposed theoretically by J.H. Poincare and G. D. Birkhoff. Poisson-stable motions are those which to-day are called chaotic, stochastic and strange attractors. The new that was brought by 60s-80s is not in revealing new types of motions and strange attractors, but rather in the apprehension and vivid demonstration of how these complicated motions arise in the dynamical systems described by ?structurally simple differential equations and how and why these motions turn out unpredictable and random, despite a validity of the uniqueness theorem for the solution of differential equations.

To my mind, the first strike on habitual limited representations of a behaviour of dynamical systems that served as a trigger was made by the works of mathematician S. Smale in the beginning of 60s. Before him, the ideas about dissipative dynamical systems were mainly restricted by what was observed in two-dimensional structurally stable dynamical systems [1,2,3]. Smale showed that in the multi-dimensional case, $n > 2$, there exist structurally stable dynamical systems of very complicated motions; the structurally stable systems may admit an infinite number of periodic motions; within a space of dynamical systems there may exist domains of structurally unstable systems; and that a structural instability is not thus an uncommonness [4,5,6].

These surprising facts shortly created a boom that in 70s – 80s was widely spreading among new people, mainly physicists and mechanicians knowing neither the classical theory of dynamical systems by Poincare and Birkhoff nor investigations of geodesic lines (the trajectories of a material freely travelling point) on surfaces of negative crookedness, initiated yet by J. S. Hadamard nor symbolic dynamics nor later works on the theory of nonlinear oscillations and specific systems. To such people, these facts seemed to be happening for the first time. And for them, this was really so.

When speaking about arbitrary dynamical systems, one should know that among them there are undoubtedly very many (and, possibly, the overwhelming majority) such that we can never study. Though, most likely, this study is not so necessary. What from the formally mathematical point of view of a space of dynamical systems is general and specific does not always coincide with what is interesting and what occurs in real systems. It seems to me that as a good example of this difference, there may serve the problem of commonness and frequency of occurrence of stable equilibriums . Once A.A. Andronov told me (that time I was dealing with a space of polynomial coefficients) that from the viewpoint of this large dimension space stability is a very rare thing. I agreed with it but did not respond anyway to the suggestion to attack this problem . Later on, this subject attracted the attention of other researchers who showed that this is actually so [7]. At the same time, to my mind, this problem has no relation to the frequency of occurrence of stable equilibriums in nature, and, especially, in engineering. The matter here is how a parameter space has been chosen in which a stability domain

is found, what parameters are chosen by nature and what are constructed by a human. The influence of variable parameters upon a system is defined by its physical structure. In his time, there have been performed very interesting and partly forgotten studies by I.I. Gal'perin concerning the structural schemes of linear control systems being stable for any values of parameters, i.e. stable due to their structure [8]. In the contemporary world, engineers are persistently looking for control systems which are less sensitive with respect to variations of their parameters, i.e. structurally stable systems by Andronov-Pontryagin terminology or robust systems. It seems to me that in this problem, to find out the systems being robust in their structure is of great importance .

Apparently, a behaviour of our world is essentially different from the dynamical systems possessing a compact phase space. The property of recurrence of steady motions in dynamical systems is not inherent to our world. Most likely, everything in our world happens for the first time. Ekleziast was not right saying that all what was in the past will be in the future. It was Democritus who was right. Though, within the limited time intervals and within the restricted scales the simplifying idealization adopted by the theory of dynamical systems is quite admissible.

The simplest effective and practically significant classification for motions of dynamical systems is apparently splitting them into stable and unstable, Lyapunov-stable and Lyapunov-unstable, and orbitally stable and unstable. The difference between the Lyapunov-stability and the orbital stability is extremely large. The Lyapunov-stability implies a closeness of state changing laws for undisturbed and disturbed motions. The orbital stability is an unbounded staying of a disturbed motion within a small neighbourhood of an undisturbed motion. In this case, the state changing laws for a disturbed and an undisturbed motions can differ significantly. From the said it follows that the Lyapunov-stability always implies the orbital stability, but the inverse implication does not hold.

Simple steady-state motions in dynamical systems are stable equilibrium states and periodic motions. Complicated motions are Poisson-stable motions and different from simple. They possess a weakened property of periodicity; this property implies that any of the previous states is repeated in time with any degree of accuracy; such motions are orbitally stable, but generally Lyapunov-unstable.

In a general case and from the point of view of Lyapunov exponent values, the Lyapunov stability and instability are an exponential stability and an instability. Supposing this commonness, complicated motions are orbitally stable motions which are exponentially Lyapunov-unstable, i.e. those from which almost any disturbed motion is locally exponentially moved off, but all the time is walking within some domain, an attractor. However, a Poisson-stable and exponentially stable or asymptotically stable motion is necessarily simply periodical.

Almost all the rest motions asymptotically converge to simple or complicated steady-state motions, equilibrium states, periodic motions and Poisson-stable motions.

In addition to the above said, I would like to touch briefly the point concerning predictability and unpredictability for dynamical system motions and how unusual, complicated, unpredictable and random motions are born. We have already

dealt with these motions before narrating a stochastic oscillator, i.e. the "contrary clock". Below is about the same but in a more general fashion.

If no unknown disturbances and no miscalculations, then all motions will be predictable. An actual unpredictability arises from, on the one hand, unavoidable random or unknown disturbances and, on the other hand, from the inevitable calculating errors as well. The difference between these two factors is in the fact that the first is repeated, while the second is not, if identical digital computers are used. However, the resulting action of these unavoidable factors can be substantially different.

It is possible that with the decrease of the unavoidable inaccuracies , the attainable prediction accuracy will also grow unrestrictedly for all $t > 0$. In this case, we are inclined to admit the presence of predictability. However, here an absolutely different case is possible, i.e. for infinitely small unavoidable inaccuracies a prediction error may be not less than some finite magnitude that can essentially exceed an admissible one. In this case we have to admit the presence of unpredictability . It is the very situation that takes place for complicated motions, due to their exponential Lyapunov-instability.

A dynamical system can be also considered from the viewpoint familiar to researchers in the control theory, and namely, as a link with some input and output. However, here one should assume that in this link there occur autonomous dynamical process disturbed by some input random perturbation. According to this approach, the link may be a transformer, an amplifier or a generator of stochasticity [9]. A transformer of stochasticity is well familiar to you and was studied hard by you. An amplifier and generator of stochasticity are the things absolutely surprising to you and earlier not accounted in any way.

A dispersion of a stochasticity amplifier output can significantly exceed the dispersion of the input perturbation; it takes place so even with the local amplification coefficient tending to infinity, as a random input decreases. Here, similar to the case of the stochasticity transformer, the stochastic properties of the output will be defined by the vanishingly small random input. For the stochasticity generator, for an infinitely small random input the output dispersion will remain larger than some finite magnitude and a statistical description of the output will not depend upon the input. It will be completely defined by the unperturbed system itself, by its deterministic description.

Finally, concerning the generation of complicated, Poisson-stable motions of amplifiers and generators of stochasticity. In his time, a birth of the theory of auto-oscillations was inseparably linked with Andronov's discovery of the relations existent between auto-oscillations and Poincare limiting cycles. This was repeated, to some extent, with complicated motions as well; however, this time not due to the limiting cycles introduced and studied by Poincare, but due to his homoclinic and heteroclinic curves [10]. These curves were revealed by him in connection with the restricted three-body problem he was hopelessly studying for a long time. Homoclinic and heteroclinic curves or motions of a dynamical system are its doubly asymptotic motions to one and the same or different unstable saddle periodic motions as $t \rightarrow \infty$ and $t \rightarrow -\infty$, respectively. These very homoclinic and heteroclinic curves (motions), arising naturally and generally, being asymp-

totic to seemingly nonexistent unstable saddle periodic motions, have turned out to be able to generate complicated, orbitally stable and Lyapunov-unstable motions [9, 11-14].

Complexity of these motions was yet emphasized by Poincare, though he seemingly did not suspect their ability to generate complicated steady-state motions. It turned out that the complicated motions can generate the motions being doubly asymptotic not only to periodic motions but also to equilibrium states or to both of them (we shall call them homoclinic and heteroclinic as well).

A stable periodic motion is widely known to be able to be born from an equilibrium state losing its stability (the Andronov bifurcation). Analogously, a complicated motion being orbitally stable and Lyapunov-unstable can be born from closed contours consisting of homoclinic and heteroclinic curves [15].

Above we were talking about description and behaviour of an isolated system not subjected to any external disturbances; such a system is usually called autonomous. It is only in this case, that a state of a system and changes of a state are determined by the system itself. In a non-isolated system subjected to disturbances external with respect to this system, this is not the case any longer, since a change of a state depends not only upon the state itself but also upon external disturbances. The dynamical externally disturbed system is also called an automaton. With this notion we usually encounter in describing games, rational behaviour and recognition (a perceptron). An automaton is usually supposed to be equipped with an output determined either by automaton's state or an external disturbance. These features bring the notion of an automaton close to the notion of an input-output link, and, if it is linear, with the gain and the gain-phase frequency characteristics. If no information about the internal structure of the link and only its input and output may be registered, then such a link is called a black box. A non-isolated non-autonomous system is undoubtedly more general than an isolated and autonomous one, but despite it, this very isolated autonomous system remains to be a basic model, because, first, it describes fundamental natural laws and, second, via expanding the object in question it becomes possible, as we think, to describe this object through the isolated and autonomous system. From this point of view, a non-autonomy may be thought to be a result of distinguishing some portion within an isolated autonomous system.

41.2
Andronov and the oscillations theory as the science about evolutionary processes and phenomena

Alexandr Alexandrovitch Andronov has created a new science, the theory of nonlinear oscillations and a world-wide known scientific school. He has trained his pupils who are proceeding his research. The theory of oscillations of to-day is a wide universal science dealing with evolutionary processes in nature, engineering and society, mechanics, physics, astronomy, chemistry, biology and in everything we are surrounded with and also inside us ourselves. Literally and in perceptions of many people the name "oscillations theory" is not adequate to what I have said.

However, historically it happened so   that this term has originated from pendulum oscillations, mechanical and electrical oscillators. As well, it was also the very history that also confirmed the power and significance of the oscillations theory as an all-embracing science implementing in itself a great idea of isomorphism and a unity of nature and world.

Quite recently – and also at present – the theoretic and applied mechanics involves the oscillations theory , stability theory and partly the control theory even. In any case, these sections were included in the tetra-volume book "Mechanics in USSR for 50 years" and in scientific programs of traditional congresses on mechanics. What is more, the founder of the modern stability theory, A.M. Lyapunov, was an academecian of the mechanics division of the Russian Academy of Sciences. J. L. Lagrange , a creator of the theory of small oscillations used mechanics as a basis. The famous treatise by J. W. Strutt and J. W. S. Rayleigh, from which the science on oscillations is supposed to have been originated, was called "Theory of sound", i.e. it related to mechanics. The well-known book "A treatise on the analytical dynamics of particles and rigid bodies" by E.T. Whittaker involves not only the oscillations theory and stability theory but the theory of dynamical systems as well. The famous work by Poincare having laid a foundation for the theory of dynamical systems was called "New methods of celestial mechanics". The widely known engineering books of Den Gartog and Timoshenko deal with oscillations in mechanical systems. If we stop our talk here , then there may appear an impression of the oscillation theory being related to mechanics entirely. However, the said only touches the significant role of mechanics in the birth and development of the oscillations theory . As early as in the lectures of L.I. Mandel'schtam in 1930-1931 and 1931-1932 at the physics faculty of the Moscow university (published in 1955), the oscillations theory appeared as an independent science. In these lectures, you can find much about oscillations not only in mechanics but in many other sections of physics. Due to the successful research by Mandel'shtam's scientific school, the first All-Russian conference on the theory of oscillations was held in 1931. It was at this conference that A.A. Andronov rendered a wide and detailed report concerning mathematical problems of the oscillations theory. The next decisive landmark in developing the oscillations theory as an independent science is the book "Theory of oscillations" by A.A. Andronov, A.A. Vitt and S.E. Khaykin, published in 1937. It is also worth noticing the remarkable book by G.S. Gorelic "Oscillations and waves" dated 1950th. The former book had laid a foundation for the oscillations theory, while the last one expanded a range of physical phenomena embraced by the oscillations theory.

A considerable contribution to the development of the oscillations theory   as an independent science has also brought the Krylov-Bogolubov-Mitropol'sky school. It was this scientific school that initiated and arranged a long series of international conferences and symposiums devoted to the following section of the non-linear oscillations theory – "analytic and qualitative research methods and their applications". Later, since 1972, when nobody already doubted the independence of the oscillations theory, scientific conferences were organized in Gorky (Nizhny Novgorod), in the Research Institute for Applied Physics and Research

Institute for Applied Mathematics and Cybernetics of the Nizhny Novgorod State University after N.I. Lobatchevsky. The boom having taken place at these conferences and caused by the "discovery" of chaotic and stochastic oscillations swept then across the entire country. The oscillations theory began to be associated with the non-linear physics, non-linear dynamics and synergetics. The oscillations theory has become a science of general regularities in evolutionary processes in mechanics, physics, radio engineering, radio physics, hydrodynamics, control theory, chemistry, biology, society science, etc. Simultaneously, this theory found a strong mathematical basis in the qualitative theory of differential equations and the theory of dynamical systems, not only exploiting them but also defining essentially their problems and scientific directions.

From above it is seen how various are the links of the oscillations theory with other sciences and that other sciences are not only an object for applications of its methods but also an inexhaustible productive source of new questions, tasks and problems. A core of the oscillations theory is mathematical models for the laws of nature and the processes occurring in it, including engineering and society. The notions of a state and an operator of its changes lie in the basis of these models. A state is invented, whereas an operator is revealed. M. Faraday has invented electrical and magnetical fields, i.e. the state, whereas J. K. Maxwell has derived the equations determining the operator.

In the birth and in the development of any science the facts (exposed in this or that way), research methods and also a general understanding play a fundamental role. In the oscillations theory , the facts imply both some accumulated knowledge of mathematical models and the knowledge about the specificity of the simulated processes and phenomena accompanying these processes ; the methods are the techniques of studying these mathematical models represented by ordinary differential equations or equations in partial derivatives or by dynamical systems of this or that form. As for the general understanding it implies, first of all, a vivid comprehension by people of possible evolutionary processes and their regularities and the phenomena accompanying these regularities; in particular, understanding the ideas of a phase space, an equilibrium state, a periodic motion, an auto-oscillation, chaos, a strange attractor, a wave, wave and mode interactions, a dispersion, external and parametric disturbances, a resonance, a synchronization, bifurcations and a bifurcation portrait, local and global stabilities, an attracting domain, a soft and a hard excitations, a hysteresis, a superposition principle, sliding motions and discontinuities, etc.

In a rough and general fashion, main types of the evolutionary processes may be listed as follows:

1) order, synchronization, predictability on the basis of the Lyapunov asymptotic stability;

2) chaos, unpredictability, stochasticity on the basis of the exponential instability under the orbital and global stability;

3) homeostasis, implementation of an objective or a program, self-organization on the basis of directed feedbacks and controls.

This short list holds centuries-old fallacies surmounted in the 20th century, new vast knowledge, a new scientific world outlook and an understanding of the evolutionary processes in nature, engineering and society. Here lies not only the yesterday and the to-day of the oscillations theory but its predictable tomorrow tool. One of the most important differences between the yesterday and the to-day of this science is that this science of yesterday mainly studied and mastered the simplest steady-state motions, i.e. equilibrium states and periodic motions, although the non-vivid complicated chaotic motions happened to implicitly expose themselves in the statistic thermodynamics and physics, the turbulence of fluids, gases and plasma; but to-day it has become possible for this theory to consider all these phenomena from some unified positions, as a manifestation of the same dynamical system. The difference here lies not only in this. To-day, this science started to study the processes based on the directed connections and controls – this is the approach already lying beyond the the frames of the rapidly developing theory of automatic control. This way of study is essential not only for understanding any living organizm but also for constructing intelligent machines and comprehending how a human and a human society will evolve.

A lot of very important scientific discoveries are usually performed twice: at first, in an abstract theoretical way, by individuals; these theoretical discoveries are not usually brought to public attention, nobody gets excited and touched by them and nobody notices them; at second, these discoveries are performed practically, in this case they attract public attention, astonish many people opening for them something previously unveiled , sudden and very important. In the oscillations theory , this happened so with the discoveries of auto-oscillations , chaotic and stochastic oscillations. Great discoveries may also have a third phase in their development – an earlier amazing discovery becomes commonly known and then, as it were, it "is vanishing". The same happened with such a great idea as that the Earth is a sphere and nothing supports it, that the Earth revolves round the Sun, that two masses or two charges are pulling each other, in spite of emptiness between them. The same happened with auto-oscillations. At present, the same is taking place with chaotic and stochastic motions. A similar thing took place even with the special theory of relativity – nowadays, it surprises no one any more and is accepted by students as a due and habitual thing.

For the first time, stable limit cycles were discovered by J.H. Poincare. Poincare, a great geometrician, depicted the solutions of differential equations through curves in the space of their variables. This way made it possible for him to easily reveal limit cycles. They were natural, since in some common situations to close a curve was inevitable; this inevitability – at least, on a plane – was made by him vivid , through introducing a section segment and a consecutive function on it (Fig. 41.1). He made evident not only the inevitability of closing a curve, i.e. a periodicity of the solution of the differential equation, but also a possibility for its stability. This very naturalness and simplicity, down to primitivity, promoted the situation that no one paid special attention to this idea ; it came to no head that this idea is interesting, very important and brings a discovery to humanity.

For the second time, upon a considerable period of time, stable limit cycles in the form of auto-oscillations were discovered by A.A. Andronov [16]. He discovered them not as abstract limit cycles but as practically vitally important phenomena in mechanics, electricity, astronomy, chemistry and biology. They were exposed by him as phenomena in nature and engineering: in distant Cepheids, in the radio-transmisson world-wide, in the sounds of musical instruments, in the shimmy of a car and in air-plane wing flutter capable of destroying them, in the undesired vibrations of tools in a mechanical processing, and in other numerous fields.
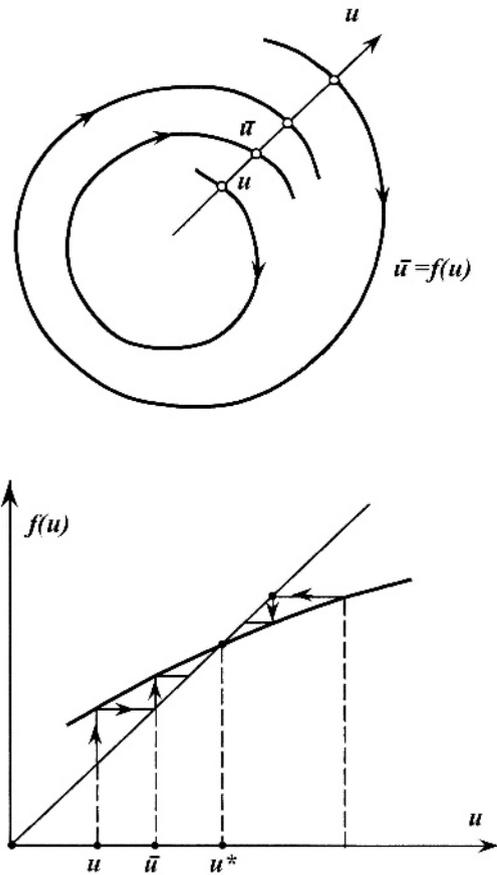
Fig. 41.1. A section segment, a consecutive function and a Lamerey diagram; u^* is a fixed point corresponding to a stable limit cycle.

This second discovery brought a revolution in the thinking of engineers and scientists. Sometimes it was accepted with distrust and difficulties, quite often it was even vigorously rejected, as this happened, for example, in chemistry with the well-known Belousov's chemical reaction, even despite its demonstration in a transparent bottle with the chemical solution changing its colour periodically. Even in mechanics, where everything might seem sufficiently simple, there existed a long aversion for this discovery; for example, a car shimmy was interpreted as a resonance caused by road surface irregularities, and vibrations in whetting were treated as the same resonance born by separations of chips, etc. Auto-oscillations were accepted quietly only in incipient radio engineering, since there was no way out, because everything in this field was resting on auto-oscillations.

A.A. Andronov showed not only the identity of the Poincare's limit cycles with auto-oscillations in real objects and systems [17] but found and studied their behaviour in numerous real and specific systems; he revealed and investigated the phenomena of catching and tightening (to-day, synchronization) under a periodic action onto a self-oscillating system and in the interaction of its modes. In this, he employed, at first, the Poincare perturbation theory and then, in the very nonlinear systems, the method of point mappings [18]. He also studied how auto-oscillations arise and vanish as parameters change (Fig. 41.2a,b,c,d) – now, this phenomenon is widely known as a bifurcation of generating auto-oscillations (a limit cycle) from a stability-losing equilibrium and other (from a closed separatrix of the saddle-type equilibrium when stable and unstable equilibrium states merge and vanish) [1,17,19,20]. These discoveries were made applicable to smooth second-order systems. Their extension to a multidimensional case and more general systems (piecewise-smooth and singular perturbed) as well as a use of new opportunities of computers was done by subsequent generations (Fig. 41.3a,b,c).

All these discoveries resulted in the fact that, instead of mythical mechanisms of oscillation birth, humanity had learnt to unveil real causes of oscillations and to eliminate them where they are harmful and dangerous and also to produce them where they are necessary and useful. In the contemporary engineering, such cases are very numerous. Thus, a barrier of aversion to auto-oscillations, arising, as it were, from nothing, was penetrated . Earlier, there existed an understanding of the damping and undamping (in idealization) oscillations of a pendulum, a weight attached to a spring and an elastic medium, etc. Also understandable were the forced oscillations caused by a variable force or voltage, a surprising phenomenon of resonance, but how oscillations can arise from "nothing", without any exciting action , was absolutely mysterious and unacceptable. This psychological barrier was got over and auto-oscillations of to-day are usual and understandable for everybody.

Such is a comparatively simple and clear-cut history of discovering auto-oscillations and the associated notions of stability, a phase portrait, a bifurcation, a bifurcation portrait, synchronization, structural stability and other notions. Something similar, but on a large scale because of the increase in the number of people involved in science and because of the difficult nature of the problem, also happened with the discovery of the complicated steady-state motions – chaotic and stochastic. But now, the number of the immediate participants is enormous

and everything is twisted and knotty, and the reason for aversion is ambiguous. To describe this case here completely is impossible. The history of this discovery seems contradictory: on the one hand, the chaotic unpredictable motions seemed to be rejected, and, on the other hand, they were exploited as a basis for statistical thermodynamics and physics, with turbulence not denied either. So, in the end of all ends, the point here lies, most likely, not in a lack of common admission but sooner in the incomprehension of how the determinism is able to give birth to un-predictability and randomness. However, in this place of our narration the picture remains still uncompleted. A general explanation of all these things was given by Poincare, Krylov and others. Probably, the most essential thing here is that there was no clear and acceptable explanation of how this phenomenon occurs and how often it must occur. From this point of view, Poincare was again the first to ex-plain it [10] and again his idea was not noticed by anybody and no respective conclusions were drawn from it.
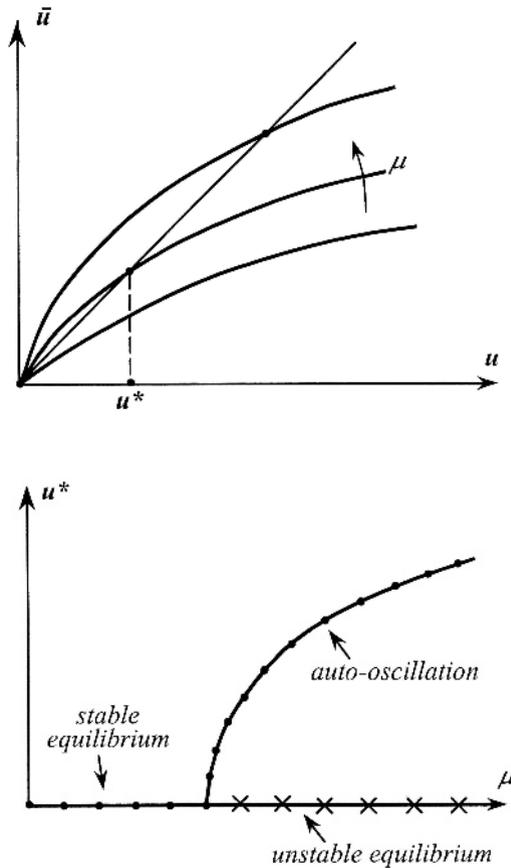
Fig. 41.2a. The bifurcation of generating an auto-oscillation (soft excitation) with the change of the parameter μ ; the Lamerey diagram and bifurcation diagram.
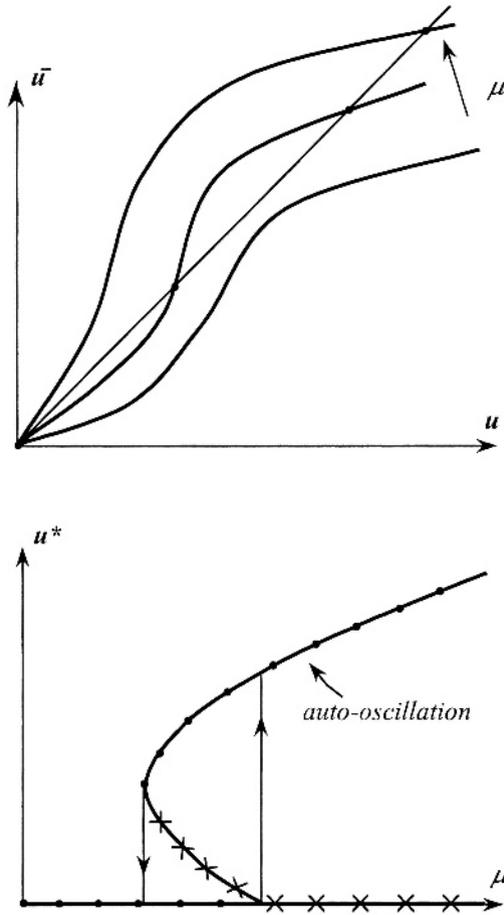
Fig. 41.2b. The hard excitation of oscillations.
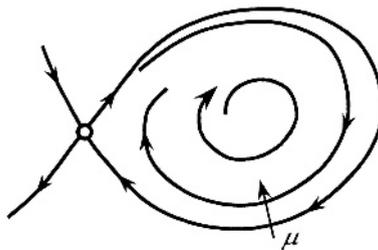
Fig. 41.2c. A birth of an oscillation from a separatrix of the saddle.
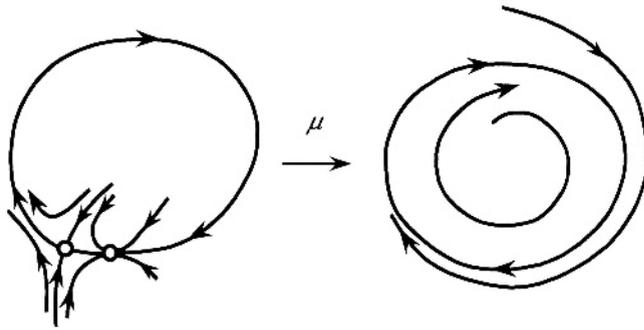
Fig. 41.2d. A birth of an oscillation with a saddle and a node <u>being merged</u>.
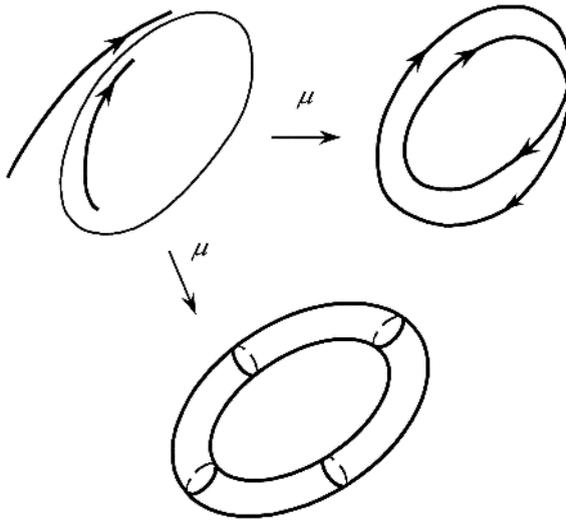
Fig. 41.3a. New bifurcations of an auto-oscillation in a multidimensional case (doubling a period, a birth of a tore).

The discovery of Poincare is the homoclinic and heteroclinic curves and rela-tive surprisingly complicated and intricate behaviour of neighbouring curves, which Poincare did not even try to depict. These curves were revealed by him when he was stubbornly and hopelessly solving the restricted three-body problem. It is in this incomprehensible complexity that Poincare has noticed a stumbling block lying on the way to the solution. However, he ceased his study. Why he did so is difficult to say; probably, he was attracted by other problems and the time to solve this problem has not yet come.
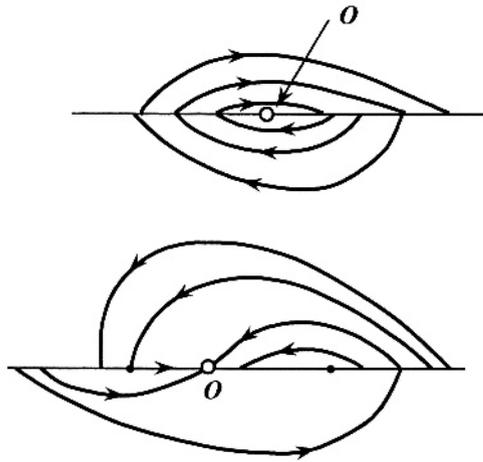
Fig. 41.3b. New possible equilibriums O in discontinuous piecewise-smooth systems (the right-hand sides of the equations in the equilibrium point are not reduced to zero).
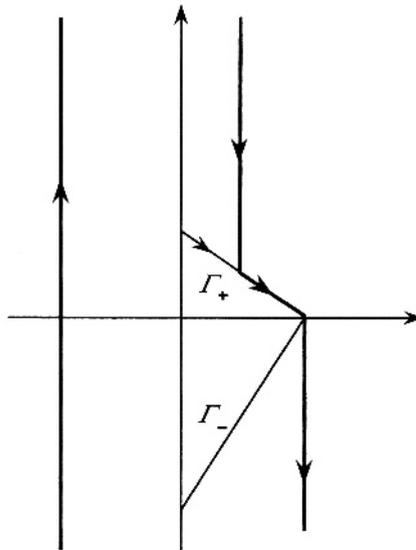
Fig. 41.3c. Discontinuous auto-oscillations in a limit singularly perturbed system.

   Homoclinic and heteroclinic curves are not already so simple as limit cycles, a section segment and a consecutive function are. Instead of a section segment, Poincare resorted to a transversal plane and the point mapping T generated on it.

   Figure 41.4a shows the saddle periodic motion Γ and the appropriate homoclinic curve γ. Figure 4b shows their cutting by a two-dimensional plane, on which O is a saddle fixed point of the intersection of Γ with the cutting plane, and the points M_{-1}, M_0, M_1 are consecutive points of intersecting the cutting plane with the homoclinic curve γ and simultaneously the intersection points of the invariant curves S^- and S^+ of the saddle point O. The complexity Poincare was speaking about is a behaviour of the curves S^- and S^+ : no one of them intersects itself, but they intersect each other in the infinite number of points and prolong unboundedly. Poincare did not begin to describe how they do it.
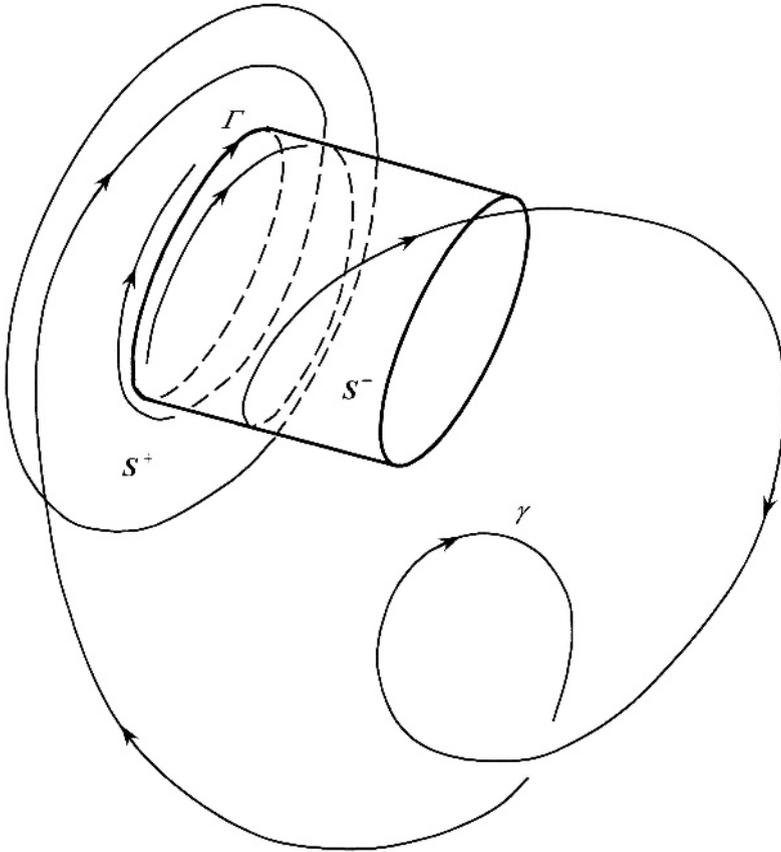
Fig. 41.4a. The saddle motion Γ and the Poincare homoclinic curve γ.

One may notice in figure 41.4b that the domains $..., \sigma_{-1}, \sigma_0, \sigma_1, ...$ and the domains $... \nu_{-1}, \nu_0, \nu_1, ...$ will consecutively transform into each other according to the point mapping T and, thus, will form one flow of the points incoming the domain D and another flow of the points leaving it. In order to understand the entire complexity of what is happening with them inside the domain D, we note here that in the domain G composed of a neighbourhood of the saddle closed curve Γ and a neighbourhood of the homoclinic curve γ (Fig. 41.4c), there exists an infinite set of different saddle curves, so that to any infinite-in-both-directions sequence of integers of the form $..., j_{-2}, j_{-1}, j_0, j_1, j_2, ...$ $(j_s > N)$ there will correspond a single saddle curve that will consecutively turn round the curve Γ $..., j_{-2}, j_{-1}, j_0, j_1, j_2, ...$ times, and, between them, it will return to Γ by passing along the homoclinic curve γ within its neighbourhood [11,12]. From this it follows that except the homoclinic curve γ there will exist an infinite set of other homoclinic curves and near each of them there will exist, in their turn, similar curves and so on.

Now, as if looking through a magic crystal, a general mechanism of forming complicated motions can be represented in the form of phase trajectories running in a successive parallel way close to the homoclinic and heteroclinic curves [13]. Here, their separation occurs in a complicated and thin way, which is similar to how it happens with the incoming flow of the domains $..., \sigma_{-1}, \sigma_0, \sigma_1, ...$. If the domain σ_i is intersected with σ_{-j}, then the phase point is thrown out (out of the domain D) as a point of the domains $\nu_{-j+1}, ..., \nu_0, \nu_1, \nu_2, ...$ and this flow of ejected points will come back again and again. There are very many variants here, but the mechanism of generating a chaotization is similar to what is happening with the peas within the Halton board when the peas are picked up below the funnel and loaded into the upper funnel again and again.

The challenge of complicated motions was also faced by J. S. Hadamard when he was studying geodesic lines on a surface of negative curvature in the dynamical interpretation of a free motion of a material point on it [21]. This discovery was accepted by mathematicians and led to the symbolic dynamics and its statistical description [22]. Here, we would like to emphasize that in that time no relations to motions in a homoclinic structure were yet revealed.

Complicated motions were also present in the G. D. Birkhoff's theory of central motions [23]. However, this is rather "what" can be and not "how" it can be.

All the said pertains to the first stage of the discovery of the complicated motions. The second stage in the numerous general studies of complicated chaotic and stochastic motions in deterministic dynamical systems was caused by , probably, not theoretical discoveries of this first stage but, rather, by real possibilities to obtain and observe them in many specific systems with the help of computers. A computer has become a decisive factor for their existence and reality; this has

brought about a specific computer understanding and interpretation for their origin and classification. Now, I would like to describe A.A. Andronov's viewpoint on the complicated motions.
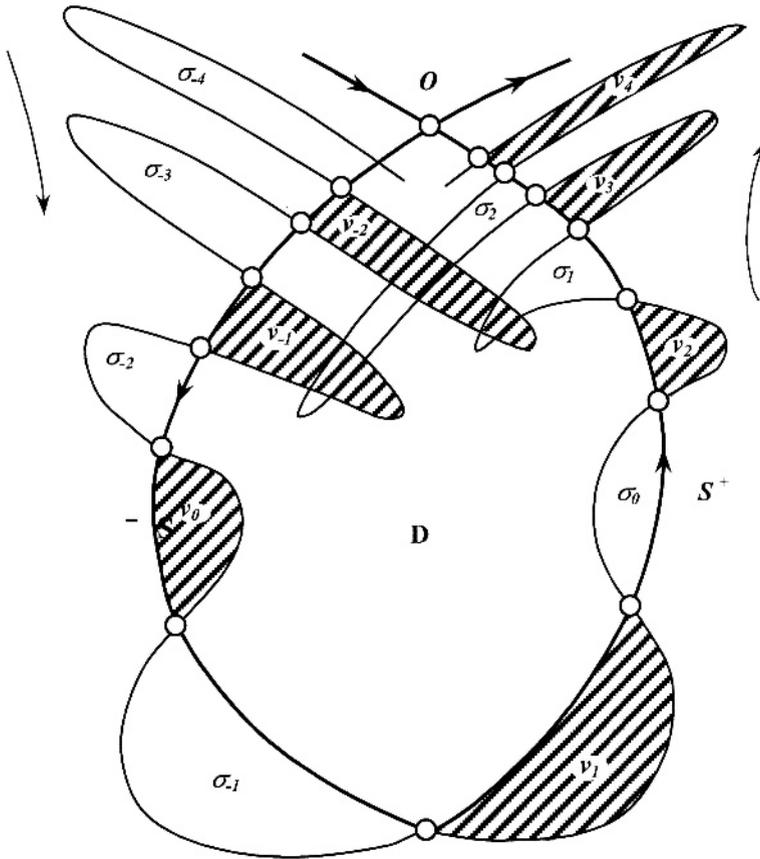
Fig. 41.4b. The homoclinic structure on the intersecting surface.

A.A. Andronov was aware of Birkhoff's complicated central motions and he was speaking about them in his report made at the first All-Russia conference on oscillations in 1931 [17]. He was speaking about a general classification of mo-tions in dynamical systems, Birkhoff's central motions including the recurrent and Poisson-stable motions – to-day called chaotic and stochastic – and about a mod-est place taken among them by auto-oscillations. But, having raised a number of vital problems for the oscillations theory, he did not declare a need to essen-tially extend   the range of the motions under question. It was proved by history

that this would have been untimely, since this extension needed more powerful approaches and facilities than those the mathematics and science of that time had available. These investigations were expanded much later, in 60-80s. In his report he was touching only the time when the main mastered tool for researchers was a combination of the Poincare perturbed method and the Lyapunov stability theory. To reveal complicated motions, their capabilities were not sufficient. Besides, there were also some other arguments that forced A.A. Andronov not to touch this point.
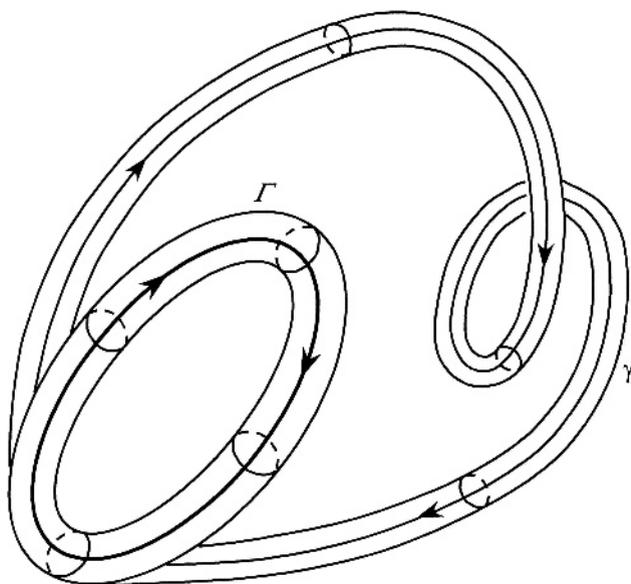
Fig. 41.4c. Neighbourhoods of the homoclinic structure.

He was attracted by a new general branch of science, dynamics of machines; he thought that the oscillation theory should bring to it and would allow to create it. He left after himself some sketches concerning a content of this intended book, but he was not fated to write it. His thinking was that a machine, for its functioning, must possess a property of robustness and its working motion must be asymptoti- cally Lyapunov-stable [24]. From this point of view, the single steady-state mo- tions to be subjected for study should be only asymptotic stable equilibrium states and periodic motions. Perhaps, thanks to this very fact, he have written the fol- lowing in the footnote of the introduction to his well-known book "Theory of oscillations" [1] : "However, the requirement of robustness for a system seemingly forbids all other recurrent motions except equilibrium states and periodic mo- tions". Note that the talk here was not about the Poisson-stable periodic motions

but about the recurrent motions only, since the Poisson-stable motions were considered by him to be not appropriate and not needed for machines. Whereas the notion of robustness (this was prior to his collaboration with Pontryagin [20]) was understood very wide and had some aspects – physical implementation, mathematical correctness of a model, approximate computability and even commonness.

These things refer to 30s. Later on, in the second half of 40s, A. Andronov showed some interest to complicated motions in specific systems. His postgraduate students, N.A. Fufaev and A.C. Alekseev, discovered and studied them in a magnetic relay timer and in a temperature regulator with outstripping. Further, these investigations were continued by his pupils and pupils of his pupils [9]. However, this is already beyond the subject of this book. I note here only that they have made a valuable and essential contribution both to the first stage and also to the second stage of general studies of complicated chaotic and stochastic motions. To-day, these motions are revealed everywhere and known by everybody. Perhaps, it should be also noted here that the most considerable results in the general theory of dynamical systems were obtained in 60s by S. Smale,   as he wrote himself, under influence of the A. . Andronov's papers.
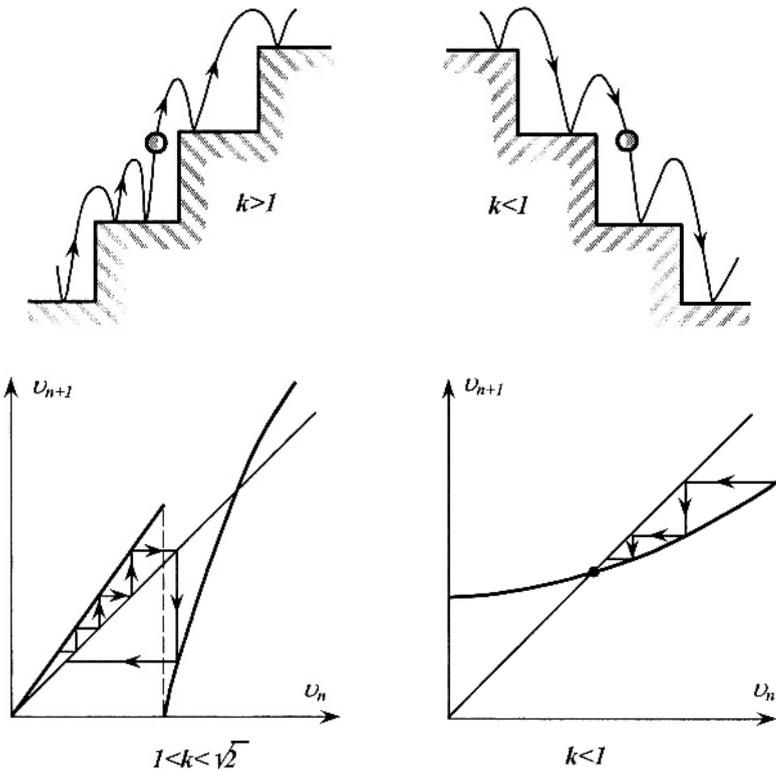
Fig. 41.5. A small jumping ball with $k > 1$ and $k < 1$.

In conclusion, let me give you a joke example concerning the two most outstanding discoveries of the 20th century that helps you to understand the dynamics of evolutionary processes. Imagine that our life is in the form of a small ball jumping at first upstairs and then downstairs. Upstars, because it is young and its recoil coefficient is more than unity; downstairs, because it has got older and its recoil coefficient is already less than unity. The Lamerey diagrams for the successive values of recoil velocities are given in figure 41.5. Ascending here is stochastic, while descending is regular.

References

1. Andronov, A.A., Vitt, A.A., Khaykin, S.E. (1959,1981): Theory of Oscillations. Nauka, Moscow.
2. Andronov, A.A., , Leontovich, E.A., Gordon, I.I., Mayer, A.G. (1966): Qualitative Theory of the Second Order Dynamical Systems. Nauka (Science), Moscow (In Russian).
3. Andronov, A.A., , Leontovich, E.A., Gordon, I.I., Mayer, A.G. (1967): Theory of Bifurcations of Dynamical Systems on a Plane. Science, Moscow.
4. Smale, S. (1963): Structurally stable homeomorphisms with the infinite number of periodic points. In: Proceedings of International symposium on nonlinear oscillations. Kiev, pp. 365-366 (in Russian).
5. Smale, S. (1966): Structually stable are not dense. Amer. J. Math., 88, 491-496.
6. Smale, S. (1970): Differential dynamical systems. Uspehi matematicheskih nauk. 25(1), 113-185.
7. Nemirovskii, A.S., Polyak, B.T. (1994): Necessary conditions for stability of polynomials and their application. Automation and Remote Control. 11, 113-119 (in Russian).
8. Neimark, Ju.I. (1953): On Gal'perin's papers devoted to structural stability of dynamical systems. Automation and Remote Control. 14(1), 88-92.
9. Neimark, Ju.I., Landa, P.S. (1992): Stochastic and Chaotic Oscillations. Kluwer Academic Publishers, Dordrecht-Boston-London.
10. Poincare, J.H. (1899): Les Methodes Nouvelles. Paris, Gauthier-Villars.
11. Neimark, Ju.I. (1967): On motions close to a doubly-asymptotic motion. Reports of Russian Academy of Sciences (DAN SSSR). 172(5), 1021-1024.
12. Neimark, Ju.I. (1967): Structure of motions of a dynamical system in a neighbourhood of a homoclinic curve. In: Procedings of 5-th summer mathematical school, Uzhgorod. Kiev, pp. 400-435 (in Russian).
13. Neimark, Ju.I. (1970): On one class of dynamical systems. In: Proceedings of International symposium on nonlinear oscillations. Kiev, 2, pp. 363-376 (in Russian).
14. Neimark, Ju.I. (1972): Method of point mappings. Nauka, Moscow (in Russian).
15. Neimark Ju.I. (1977): Invariant manifolds and stochastic motions of dynamical systems. In: Problems of asymptotic theory of nonlinear oscillations. Naukova Dumka, Kiev (in Russian).

16. Andronov, A.A. (1956) The limit cycles of Poincare and the oscillations the-
    ory (Reports of the 6-th Congress of Russian physicists, 1928). In: Selected
    works of A.A. Andronov. AN SSSR, Moscow (in Russian).
17. Andronov, A.A. (1956): Mathematical problems of the theory of auto-
    oscillations. In: Selected Works of A.A. Andronov, AN SSSR, Moscow , pp.
    84-124 (in Russian).
18. Andronov, A.A. (1944): Theory of Poincare-Brouwer-Birkhoff point trans-
    formations and the theory of nonlinear oscillations. Vestnik AN SSSR. 6 (in
    Russsian).
19. Andronov, A.A., Leontovich, E.A. (1939): Some cases of dependence of
    limiting cycles upon parameters. Uchenyye zapiski Gorky State University.
    6(3), 3-33 (in Russian).
20. Andronov, A.A., Pontryagin, L.S. (1937): Structurally stable systems. DAN
    SSSR. 14(5), 247-250 (in Russian).
21. Hadamar, J.(1901): Bull. Sos. Math. France. 29, 224-228.
22. Mors, E., Hedlund, G.A. (1938) Amer. J. Math.. 60, 813-866; (1940) 62.
23. Birkhoff, G. (1927): Dynamical Systems. Am. Math. Soc.
24. Andronov, A.A., Vitt, A.A. (1933): On Lyapunov stability. ZHETF. 3(5).

42 Mathematical model as a fruitful idea of research. The D-partition

A role of a mathematical model in research. The development of applied problem of stability (the D-partition).

As far as this book concerns mathematical models, it is natural to notice and pay attention to all aspects of their usage. These aspects are actually wide and irreplaceable, since only through a model one can find an answer to many questions, provided that these questions can be formulated in the mathematical language. This answer can be derived only on the basis of the known information; this knowledge stated in the mathematical language constitutes the very mathematical model. Therefore, all the internal mathematical investigations as well as investigating all the mathematically-stated problems from the natural science, engineering and other fields are nothing but a mathematical modelling as well. For example, the theory of integer positive numbers is nothing but investigating the mathematical model consisting of the elements $a_0, a_1, a_2, \ldots$ connected by the relations

$$a_n + a_0 = a_n, \ a_{n+1} = a_n + a_1, \ n = 0, 1, 2, \ldots.$$

Among very effective internal mathematical models constructed comparatively long ago and widely applied, the following models are worth mentioning: models of geometries and topologies, models of various groups, models of linear and non-linear equations, differential equations and many other models. This list of models, of course, also includes the basic mathematical model from this book, i.e. a dynamical system that may be also interpreted both as a semi-group and, if its operator is one-to-one, as a one-dimensional Lie group. Earlier, when narrating the models for pattern recognition and perceptron, we already showed how simple mathematical models make it possible to solve very complicated and seemingly inaccessible problems. In the field of recognition, such simple models were the models of an object, a pattern and a pattern recognition, presented, respectively, by a point in a multidimensional space, by a set of points within this space and by a distinguishing process of whether a new object belongs to this set or not. Now, we will describe another example touching the stability of movement. This description will hold two parts. The first will describe how this simple model has been created in response to urgent demands of various applications; the second part will involve new scientific results of its application.

The theory of control as a science was born in the 20-th century, but initially it originated yet as far back as in second half of the 19-th century, in the works of genius   physicist   J. C. Maxwell (1868) and   outstanding engineer I.A. Vyschnegradsky (1877). These works were brought to existence by crucial practical demands, viz. to provide stability in the rotating velocity of steam-engines, which in the then starting industrial revolution were becoming a basic source of mechanical energy. It was the time when James Watt created his steam-engine equipped with a centrifugal regulator for rotative   velocity. However, it happened so that all these regulators started to demonstrate their unruly manners – the seemingly better they were designed and the more the people wanted to get from them, the worse they were working; sometime their poor functioning led to emergencies. This arising engineering problem became widely spread out – in the then England alone there were exploited about hundred thousand steam-engines and this amount was later increased by steam turbines as well. An unruly behaviour of regulators, i.e. their instability, gave birth to the mathematical theory of stability. This theory was first discussed in the works of   Maxwell and Vyschnegradsky;   then, it borrowed the ideas and techniques of great J. H. Poincare and A.M. Lyapunov, and later on, it has become deeper and wider, due to the intensive promotion from numerous outstanding researchers in different countries of the world.

The history of the theory of automatic control may be thought as a history of eliminating the conflict arising between stability and control quality. This very conflict yet arose in the Watt's centrifugal regulators; this very conflict also led to the development of an isodromic regulator, relay systems, adaptive and robust controls. Choosing a control strategy is restricted by the requirement for the dynamics of the system implementing this strategy; in many cases, this entails a necessity of preliminary theoretical research where to determine stability conditions is the most essential component. For linear systems, the stability problem , at least from the formal point of view, is completely solved by the well-known Routh-Hurwitz criteria, originated soon under the influence of the works by Maxwell and Vyschnegradsky.

In the beginning, these Routh-Hurwitz   algebraic stability criteria seemed exhaustive. Then, they turned out not liable to generalization for non-algebraic equations and not good to entirely meet applicational needs, especially the needs in the design of new control systems, machines and devices. Engineers needed a method to study stability. This method should be able to help to understand within what bounds they could choose and modify parameters of the system under construction, without disturbing its stability and what freedom they could have to fulfil the remaining requirements. It is this very reason that in the first half of the 20-th century gave birth to the   new stability criteria, among which the graphical criteria by Mikhailov and   Nyquist won a wide recognition. Note here that Vyschnegradsky, having by that time derived the stability criterion in the form of an algebraic inequality, then represented this criterion as a stability domain on a plane of two parameters – an irregularity and a specially introduced viscous friction. The diagram shows that the geometrically interpreted   conflict between stability and stabilizing quality for stream-engine rotation velocity hold the idea

that their determinative magnitudes must fall into some domain D, whose shape must be such that a decrease of one parameter demands an increase of another. From the said it follows that there was needed a new statement of the mathematical problem being more general and adequate to applications than the Routh-Hurwitz problem. This was apprehended by mathematicians and implied a search for necessary and sufficient conditions for all polynomial roots to be in the left half–plane, i.e. $\mathrm{Re}\ z_\nu < 0$ for all ν. Besides an astonishing mystery hidden in the not quite proved, but valid, Nyquist criterion, the central question in this new approach was the question about how it may be that the polynomial roots are impossible to be derived, whereas the conditions of their presence in the left half-plane or inside a unitary circle can be derived.

This natural question was not replied by the criteria by Routh, Hurwitz, Nyquist, Mikhailov and others being well-known in the middle of 40s. Though, the answer to this question is simple and rewards you at once [1-3]. The matter is that, though the roots cannot actually be found, it is easy to find when they escape from the left half-plane or any other assigned domain G in a complex plane. In order to distinguish the instant when this escape occurs, it is sufficient to have a parametric equation for the boundary of the domain G in the form $z = f(\omega),\ \omega \in \Omega$. It is implied here that z is running along the boundary of the domain G, while the real variable ω is running across the set Ω. Then, the root of the characteristic equation $\chi(z) = 0$ will intersect the boundary of the domain G, if and only if for some $\omega \in \Omega$ we have

$$\chi(f(\omega)) = 0 . \tag{42.1}$$

Suppose now that the system, whose equilibrium state stability with respect to parameters u is subject to study, has the below characteristic equation

$$\chi(z, u) = 0 ; \tag{42.2}$$

the stability here implies all the roots of this equation to lie within the domain G of the complex plane Z. Further, let the point z belong to the boundary of the domain G, if and only if for some $\omega \in \Omega$ one has

$$z = f(\omega) . \tag{42.3}$$

From this, it follows that the stability domain boundary holds the points for which

$$\chi(f(\omega)), u) = 0, \qquad \omega \in \Omega . \tag{42.4}$$

The set N_ω of the points satisfying the condition (42.4) involves all the boundary points of the stability domain, but can also involve other points. In general, this set N_ω defines completely a partition of the parameter space into the domains $D(j)$ with different numbers j of roots of the equation (42.2) in the domain G. Let us call this partitioning technique the D- partition. We are interested in

the stability domain $D(0)$ only, but it turns out so that to find the domain $D(0)$ it is useful to consider the entire D-partition. Here lies the new idea suggested for studying stability. In the basis of this idea, there lies the model defined by the condition (42.4). As for the stability model, it will be whether the point belongs to the domain $D(0)$. The model described does not suppose the characteristic equation (42.2) to be polynomial in z. This can be any function. The domain G can be not only a left half-plane and the system in question can be described not only by ordinary differential equations. The possibilities of this approach are much wider.

Now, let us sequentially use this mathematical model of stability for the following problems:

1) constructing a stability domain with respect to the complex parameter w;

2) revealing the fact that the stability domain for the polynomial $z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n$ with respect to the parameters a_1, a_2, $\ldots$, a_n is a simply connected domain and is a convex domain in each of them;

3) proving and generalizing the well-known Nyquist criterion;

4) constructing the stability domain with respect to the two real parameters for a supercentrifuge;

5) estimating a measure of robust stability.

The above list does not cover all useful applications of the stability model described. What is more, a notion of stability can be replaced by another one carrying similar general properties.

1. Let, for simplicity, the complex parameter w be involved in the characteristic equation linearly so that it can be written as follows

$$f(z) + w\,g(z) = 0. \tag{42.5}$$

The boundary N_ω of the stability domain $D(0)$ is a mapping of the boundary of the domain $\operatorname{Re} z \le 0$, i.e. the imaginary axis of the plane Z, for whose points we have $z = i\,\omega$, $-\infty < \omega < \infty$; therefore, in the plane W, we obtain the following equation

$$w = -\frac{f(i\omega)}{g(i\omega)}, \qquad -\infty < \omega < \infty\ ; \tag{42.6}$$

here, a left-hand side of the imaginary axis is mapped into a left-hand side of the curve N_ω, described by the equation (42.6). Let us shade this left side in the assumption that the curve N_ω is running in the direction of the increase of the parameter ω. Then, when passing on the complex plane W from the shaded side of the curve N_ω to the unshaded side, one root of the equation (42.5) will become purely imaginary and will pass from the left half-plane $\mathrm{Re}\ z_\nu < 0$ onto the right one. The curve N_ω will partition the plane W into some domains among which there may be some stability domains as well. If the number s of the roots located to the right from the imaginary axis is known at some point w^*, then in all the points of the domain, which this point belongs to, the number s of the roots will be the same. Moreover, if from the point w^* we are running along the curve γ to any other point $\overline{w}^*$, then the number s in this point will increase by the number of the intersections of the curve N_ω from the shaded side and will decrease by the number of the intersections against the shaded side. To the stability domain $D(0)$ there corresponds $s = 0$. Let the remaining domains with $s > 0$ be denoted through $D(s)$. Through this notation the above said is illustrated in figure 42.1, where the arrow on the curve N_ω indicates a direction we should follow, according to (42.4), with growth of ω; the intersecting arrow shows a passage from one domain $D(s)$ to another $D(s+1)$. In the course of such a passage, the number s increases by unity; during an inverse passage, s decreases by unity. The conclusion drawn from this situation is that, to search for the stability domain $D(0)$ with the respect to the complex parameter w, it is sufficient to depict the curve N_ω and shade its left side, and, additionally, we should be aware of the number of roots to the right from the imaginary axis for the equation (42.5) at some value $w = w^*$.
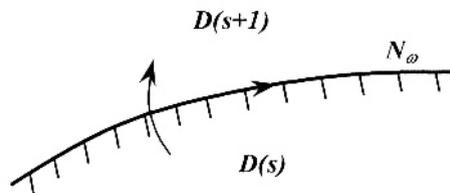
Fig. 42.1. The shading rule of the curve N_ω.

2. As an example and in order to apply the above said, let us now construct a stability domain with respect to the parameter w for the characteristic equation of the form

$$z^n + a_1 z^{n-1} + \ldots + a_n - w z^m = 0, \quad 0 \le m < n, \tag{42.7}$$

where $a_1, \ldots, a_n$ are the given real numbers for which the equation (42.7) at $w = 0$ has all its roots lying to the left from an imaginary axis, i.e. the point $w = 0$ belongs to the stability domain $D(0)$. The curve N_ω is described by the following equation

$$w = \frac{1}{(i\omega)^m} \left(z^n + a_1 z^{n-1} + \ldots + a_n \right)_{z=i\omega} =$$

$$\tag{42.8}$$

$$= \frac{1}{(i\omega)^m} (i\omega - z_1) \cdots (i\omega - z_n),$$

where $z_1, \ldots, z_n$ are the roots of the equation (42.7) for $w = 0$. They all lie to the left from the imaginary axis. From (42.8), it is immediately seen that with ω varying from $-\infty$ to ∞, $\arg w$ will increase by the magnitude πn; $|w|$ will be all the time different from zero, and at $\omega = 0$ it will reduce to infinity. At this, the left side and, hence, the shading of the curve N_ω, will always face the point $w = 0$ and any beam emerging from the point $w = 0$ will intersect the curve N_ω from the shading side only. Therefore, the stability domain with respect to the real parameter $\mathrm{Re}\, w$ or $a_m + \mathrm{Re}\, w$ will be the single section holding the point $w = 0$. From this it follows that the stability domain $D(0)$ with respect to the parameters $a_1, a_2, \ldots, a_n$ in the space of these parameters will be convex with respect to each of these parameters, i.e. this domain will either intersect any axis Oa_s of this space along some unique segment or it will not intersect it at all.

3. The famous Nyquist criterion states that, if an open-loop linear link with the input x and the output y is stable(i.e. if from some moment we have $x = 0$ and then $y \to 0$ as $t \to \infty$), then the closed-loop system obtained from the open-loop one (when the output y is also supplied to the input x, i.e. with $x = y$) will be also stable, if the point $w = 1$ of the complex plane W lies beyond the curve

$$w = K(i\omega),\qquad(42.9)$$

representing a gain-phase frequency characteristic of the open-loop link. This criterion turned out to be very convenient and applicable to many cases but also somewhat mysterious, since it had not been proved convincingly and fully. In view of the above stated positions, the Nyquist criterion and its generalizations immediately follow from the consideration of the stability domain with respect to the complex parameter w in the closed-loop system characteristic equation, which can be written in the form

$$P(z) - wQ(z) = 0,\qquad(42.10)$$

where $P(z)$ and $Q(z)$ are some entire functions having no common zeros, such that

$$K(z) = \frac{P(z)}{Q(z)}.$$

The equation (42.10) at $w = 1$ is a characteristic one for a closed-loop system and at $w = \infty$ for an open-loop system. The stability domain boundary N_ω for (42.10) is defined by the equation (42.9) and the point $w = \infty$ belongs to a stability domain; therefore, if the point $w = 1$ is not separated from the point $w = \infty$ of the curve N_ω , then the closed-loop system is stable. This is the very Nyquist criterion. From the above brief description, it follows that the Nyquist criterion can be immediately generalized for the cases when the point $w = 1$ is not embraced by the curve N_ω given in (42.9) or embraced by this curve but in such a way that, escaping from it and moving to the point $w = \infty$, the numbers of intersections from a shaded or an unshaded sides will be equal. Here, it seems also evident to derive a generalization for the case when an open-loop link has a known number of instability degrees, i.e. the number s of the roots to the right from the imaginary axis y of the equation (42.10) at $w = \infty$ is known.

    4. An ultracentrifuge is a very fast-rotating body which cannot be fixed hard, because   the slightest disbalance results in enormous inertial forces. So, a fastening must be soft and be self-centred, but then it will not provide a stable rotation of the centrifuge. This stability can be provided by introducing special mechanisms for damping possible oscillations. As a solid body, the centrifuge has four freedom degrees, two of them are progressive and two rotational; besides, its damper is of two degrees of freedom, i.e. the differential equations for its motion are of the twelfth order. The same is the degree of the characteristic equation whose roots govern the stability. Instability should be eliminated by choosing the two parameters k and h for the damper. The twelfth order of the real characteristic equation can be reduced (thanks to a circle symmetry) to the six

order, but here the characteristic polynomial will be already complex. This complex characteristic equation is written in the form

$$\chi(z) = (k + v z^2 + \varepsilon + hz)(P_4(z) - iP_3(z)) = 0,$$

where $P_3(z)$ and $P_4(z)$ are real 3-degree and 4-degree polynomials. The boundary of the stability domain $D(0)$ on the plane of the parameters k and h is described, as before, by the parametric equations

$$(k - \omega^2 + \varepsilon + ihz)(P_4(i\omega) - iP_3(i\omega)) = 0,$$

which can be solved with respect to k and h and written as

$$k = f(\omega), \quad h = g(\omega), \qquad -\infty < \omega < \infty. \qquad (42.11)$$

This makes it possible to quickly construct the curve N_ω at a computer, put shading on it and distinguish the stability domain $D(0)$. One of the possible pictures of the boundary curve N_ω defined by the equations (42.11) is given in figure 42.2.
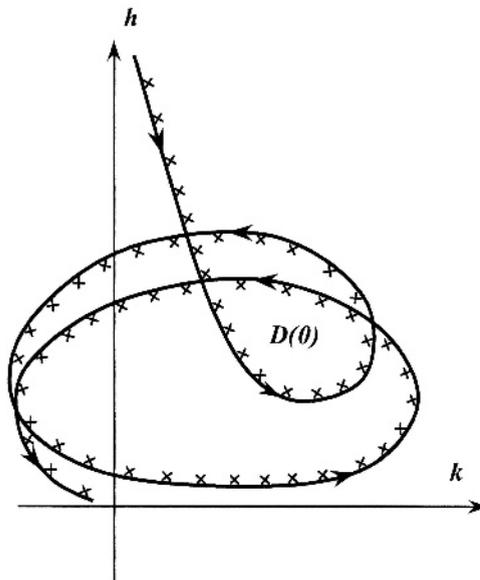
Fig. 42.2. The D-partition and the stability domain for the ultracentrifuge.

From this picture it follows that the stability domain $D(0)$ is restricted by the two parts of this curve N_ω : the upper stands for a high-frequency disturbance of stability and the lower a low-frequency disturbance, i.e. they correspond to large and small values of the frequency ω, respectively. From the same figure it is seen that choosing and providing the needed values of parameters is a rather delicate operation;   without this picture it is hardly possible. You know that every unsuccessful choice implies an emergency and a damage of the centrifuge.

   5. Now comes our last item, viz. a robust stability and a robust stability measure. The problems of robust stability were discussed in a stormy way and studied for a long time. A robust stability implies   the stability retaining under sufficiently large changes of parameters. As it follows from the above, this problem is fully solved, if there exist a single complex parameter or two real parameters; in this case, to construct the stability domain $D(0)$ is not very difficult. However, the situation becomes absolutely different when the parameters are many and to visually represent a stability domain becomes very difficult or even absolutely impossible. In this case, there arises a natural question as when a domain of a simple form lies within a stability domain. Simple domains are a sphere, an ellipsoid, a parallelepiped and others. Furthermore, how can this domain be chosen as large as possible? These questions led to the following general mathematical problem: find the minimum $\rho = \rho^*$ for which all the functions of the complex variable in the form

$$\chi(z) = a_1 \varphi_1(z) + ... + a_n \varphi_n(z), \qquad (42.12)$$

where

$$\rho = k_1 (a_1 - a_1^*)^m + \ ... \ + k_n (a_n - a_n^*)^m \le \rho^*, \qquad (42.13)$$

have their roots belonging only to the given domain G in the complex plane Z . Here, $\varphi_1(z),...,\varphi_n(z)$ are the given functions of z ; $a_1^*,...,a_n^*$ are some nominal values of the real parameters $a_1,...,a_n$, for which all the roots of the function (42.12) lie in the domain G ; $k_1,...,k_n$ are arbitrary positive numbers; and $m \ge 1$. Notice that for $m = 2$ and $m = \infty$ the inequality (42.13) will define simple domains, i.e. an ellipsoid and a parallelepiped.

   According to the above accepted model of stability, to find the measure ρ^* for a robust stability is comparatively simple – it will be equal to a minimum of the magnitude ρ determined by (42.13) over all the parameters $a_1,...,a_n$ lying on the stability domain boundary, i.e. satisfying for any $\omega \in \Omega$ the condition

$$\chi(z(\omega); a_1, ..., a_n) = 0,$$

where $z(\omega)$ is any point of the G domain boundary for the suitable $\omega \in \Omega$. Here, we are not going to solve this problem of conditional optimisation, for it is beyond the basic subject of this book. Note here only that its solution is possible and not very difficult [4-6]. For example, for the polynomial or the quasi-polynomial written in the form

$$\chi(z) = \sum \left(a_{j2s} z^s sh(\tau_j z) + a_{j2s} z^s ch(\tau_j z) \right),$$

when the domain G is a left-hand half-plane, and for the polynomial of the form

$$\chi(z) = \sum a_{2s} \left(z^{n-s} + z^{-n+s} \right) + a_{2s+1} \left(z^{n-s} - z^{-n+s} \right),$$

when the domain G is a unitary circle, it turns out that

$$\rho^* = \min_{\omega \in \Omega} \left\{ \frac{|\chi_1|^m}{\left(\Phi_1^2(\gamma) \right)^{m+1}} + \frac{|\chi_2|^m}{\left(\Phi_2^2(\gamma) \right)^{m+1}} \right\}$$

where Φ_k are vectors with the components $(k_j \varphi_{kj})^{\gamma/2}$, $\gamma = \frac{m}{m-1}$,

$$\varphi_{kj} = \frac{\partial \chi_k}{\partial a_j}.$$

References

1. Neimark, Ju.I. (1949): Stability of linearized systems (discrete and distributed). LKVVIA, Leningrad (in Russian).
2. Neimark, Ju.I. (1978): Dynamical systems and controllable processes. Nauka, Moscow (in Russian).
3. Neimark, Ju.I. (2001): Stability problems of control theory in the works of the Andronov scientific school. Bulletin of the Niznhy Novgorod State University. Mathematical modelling and optimal control. 1(23), 168-197.
4. Neimark, Ju.I. (1991): Robust stability of linear systems. Reports of the USSR Academy of Sciences (DAN SSSR). 319(3), 578-580 (in Russian).
5. Neimark, Ju.I. (1992): Measures for robust stability and modality of linear systems. Reports of the USSR Academy of Sciences (DAN SSSR). 325(2), 247-249 (in Russian).
6. Neimark, Ju.I. (1992): A domain of robust stability and robustness with respect to nonlinear parameters. Reports of the USSR Academy of Sciences (DAN SSSR). 325(3), 438-440 (in Russian).

43 Idealization, mathematical correctness and reality

A role of idealization in constructing an adequate model for an object and mathematical correctness of the model in the examples of a centrifugal frictional regulator of rotating velocity and the well-known Panleve's problem.

This subject or problem is immensely wide and diverse. It is faced with by anybody constructing and investigating a new or a partly new model that has no tested analogues or prototypes. Our below narration will be restricted by two specific examples: a centrifugal frictional regulator of rotating velocity and the well-known Panleve's paradox. What unifies them is that, in either cases, we speak about idealizing real elastic bodies and compliant connections as ideally absolutely solid and rigid.

43.1
Frictional regulator of rotating velocity

Figure 43.1 schematically shows a very simple frictional regulator of rotating velocity. This unit is designed to stabilize a rotating velocity, since a growth of the rotating velocity yields an increase of friction force against a restriction ring and a decrease of the velocity brings a fall of the friction. The centrifugal force pressing a ball close to a ring is equal to $(n\omega)^2 r$ and, therefore,

$$J\dot{\omega} = M - k(n\omega)^2 r \cdot rn, \tag{43.1}$$

where the notations introduced are clear from the figure. From (43.1) it immediately follows that the rotating velocity has the below stable equilibrium value

$$\omega^* = \left(\frac{M}{k\, r^2 n^3} \right)^{1/2}. \tag{43.2}$$

From (43.2) it follows that the differential irregularity of the equilibrium rotation velocity ω^* with respect to the moment M is equal to

$$\frac{d\omega^*}{dM} = \frac{1}{2\omega^* kr^2 n^3} \tag{43.3}$$

and it decreases with an increase of the transmission gear coefficient n. The regulator described was used, in particular, for astronomical observations, to compensate the rotation of the Earth; in these observations, the magnitude M of the moment depended also upon the moment needed to rotate astronomic devices (refractors, reflectors, etc.). As it is seen from (43.3), a needed rotational precision could be provided by a choice of the sufficiently large n. Though, it was revealed that the centrifugal regulator may be unstable and fails in its operation. Nowadays, the reason for regulator instability would be naturally found in a falling characteristic of the friction, i.e. in the fact that the friction coefficient is dependent upon the rotating velocity and that for the velocity $\omega = \omega^*$ we obtain $\dfrac{dk}{d\omega} < 0$. This instability mechanism was considered earlier. Of course, this situation is possible, but the instability may be also born by the case $\dfrac{dk}{d\omega} \geq 0$. In order to understand its origin and eliminate it, one should account a real compliance of the ring, i.e. a change of r with a change of the regulator ball pressure against a restricting ring. This is very surprising and is not felt intuitively: a compliance of the ring here seemingly plays no role, because the ring may be assumed absolutely solid. However, this is not so and this is discovered only upon its account.

Let us take into account the fact that the restriction imposed by the ring upon the centrifugal regulator ball motion is not absolutely rigid. As generalized coordinates there will be taken the turning angle φ of the vertical rod of the centrifugal regulator and the angle ϑ of the ball deviation from the rod. The Lagrange function, if friction forces and ring pressure upon the ball are not accounted, but the gravitational force is accounted, is written as follows

$$L = \frac{1}{2}\left\{ J_0\dot{\varphi}^2 + J_1(\vartheta)\, n^2 \dot{\varphi}^2 \right\} + \frac{1}{2} B\dot{\vartheta}^2 - V(\vartheta), \tag{43.4}$$

and the additional generalized forces are equal to

$$Q_\varphi = M(\dot{\varphi}) - N(\vartheta)K(\dot{\varphi}), \qquad Q_\vartheta = -N(\vartheta),$$

where $N(\vartheta)$ is a moment of the ring normal pressure upon the ball. The rest of the notation is evident.

Now, we immediately write the Lagrange nonlinear equations of motion in the form

$$\{J_0 + n^2 J_1(\vartheta)\}\ddot{\varphi} + n^2 J_1'(\vartheta)\dot{\vartheta}\,\dot{\varphi} = M(\dot{\varphi}) - K(\dot{\varphi})N(\vartheta)$$
$$B\ddot{\vartheta} - \frac{1}{2}n^2 J_1'(\vartheta)\dot{\varphi}^2 + V' = -N(\vartheta).$$
(43.5)
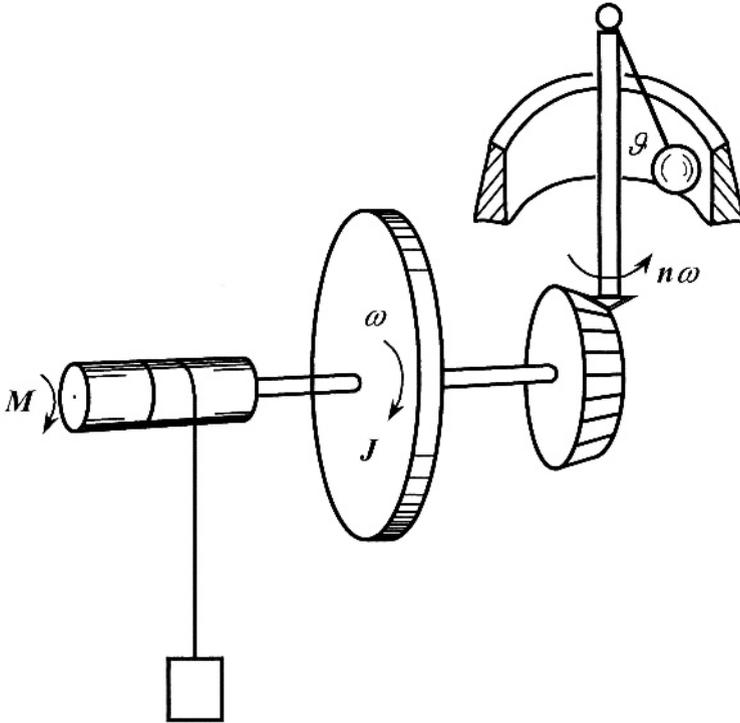
Fig. 43.1. The frictional regulator of rotating velocity.

From them, the uniform rotation $\dot{\varphi} = \omega^*$, $\vartheta = \vartheta^*$ we are interested in is found from the equations

$$M(\omega^*) - K(\omega^*)N(\vartheta^*) = 0$$
$$N(\vartheta^*) - \frac{1}{2}n^2 J_1'(\vartheta^*)\omega^{*2} + V'(\vartheta^*) = 0.$$
(43.6)

Let its solution ω^*, ϑ^* be known for us. Notice that from (43.6) one may determine the differential irregularity for regulation

$$\frac{d\omega^*}{dM} = \frac{N' - \dfrac{n^2}{2}J_1'' \omega^{*2} + V''}{(k'N+M')\left(N' - \dfrac{n^2}{2}J_1'' \omega^{*2} + V''\right) + kN' J_1' \omega^*}.$$

As $N' \to \infty$, this irregularity decreases infinitely with a growth of n. There is nothing new in it. However, investigating the stability for the uniform rotation now brings us to the linearized equations ($\omega = \omega^* + \xi,\ \vartheta = \vartheta^* + \eta$)

$$(J_0 + n^2 J_1)\dot{\xi} + nJ_1' \omega^* \eta = -k\,N'\eta - N\,k'\xi - M'\xi,$$

$$B\ddot{\eta} - nJ_1' \omega^* \xi - \frac{1}{2}n^2 \omega^{*2}\eta - V''\eta = -N'\eta$$

and to completely new and unexpected conditions of stability

$$N\,k' + M' > 0,$$

$$(N\,k' + M')\left(N' + V'' - \frac{n^2}{2}J_1'' \omega^{*2}\right) + n\,J_1' N'\omega^*, \qquad (43.7)$$

$$(N\,k' + M')\,n^2\,J_1' \omega^{*2} - k\,N' > 0.$$

From these conditions it is seen (see the last inequality) that, as $N' \to \infty$, instability is born inevitably. Thus, on the one hand, an account of even negligibly small elasticity with respect to ϑ in a ring restriction will result in an instability, and, on the other hand, the inequalities (43.7) show that, if a not very large rigidity is artificially introduced into the ball-ring contact, then the centrifugal regulator may obtain stability.

The mathematical incorrectness of the model (43.1) lies in that that the infinitely small compliance of the ring brings a sharp qualitative change in model behaviour. This incorrectness is not revealed so immediately as it was before, for example, in the Volterra-Lotck "predator-prey" model. Now, the incorrectness is attributed to a non-obvious increase of the model dimension and to a transfer from the first-order differential equations to the third-order ones. In the "predator-prey" model, this incorrectness was exposed within the frames of the initial model, for small changes of the right-hand parts in the second-order differential equations. Alongside with it, what actually happens with the frictional regulator in the unstable state because of too large rigidity of the contact with the restrictive ring produces some prompt to what else should be taken into account. Violation of stability brings a jumping of the ball along the ring; during these jumps the ball becomes separated from the ring and this fact prompts to account one more degree

of its freedom with respect to the angle ϑ. This very thing has been taken into account.

43.2
Panleve paradox and auto-oscillations under the Coulomb friction

At the very end of the 19-th century French mathematician P. Panleve, delivering a course in friction for Parisian students and trying to develop a general theory, found out that for some simple mechanical systems with a friction, the motion equations, derived with use of general rules of mechanics, are unsolvable. He published the book on friction, where those surprising examples were given. Ten years later these examples caused a stormy discussion, with participation of such prominent scientists of that time as R. Misec, L. Prandtle, Ph. Klain, G. Gamel, L. Lekornu and others, and, of course, of P. Panleve himself. The extraordinary occurrence discovered by P. Panleve was called the Panleve paradox. As of to-day, this paradox still has no solution which could afford us to predict with confidence the behaviour of very simple mechanical systems with the Coulomb friction. This impossibility of a precise prediction is accounted for the fact that the systems with the Panleve paradox demonstrate an essential dependence of their behaviour upon small and not easily determinable parameters and regularities, being very unusual in such problems.

   Below, I would like you to get acquainted with the essence of the Panleve paradox and to emphasize here that it gives rise to an unexpected and earlier unknown possibility for creating instability and auto-oscillations, though the friction characteristic has no falling part. In terms of a   natural limit description these arising oscillations are discontinuous, or, in accordance with the contemporary mathematical terminology, they make up contrasting structures with skipping and splashes, which from the physical point of view can be called "internal" impacts. Discontinuous oscillations in the form of periodic contrasting structures are born, because the mathematical model of the Panleve paradox systems is a singularly perturbed system of differential equations [1,2].

   Figure 43.2 shows some well-known examples of mechanical systems with the Panleve paradox:

   –   the Panleve-Klain system describing a motion of a rod between parallel guiding lines L_1 and L_2, one of which (L_1) is smooth while another (L_2) is rough, under action of the constant force F_0 (Fig. 43.2a);

   –   a brake block being pressed by the force F_0 to a wheel rotatable by the moment M (Fig. 43.2b);

–   a circle (cylinder), rotatable by the moment M and being pressed into the angle by the force F_0 (Fig. 43.2c), one side of the angle being rough and another smooth.
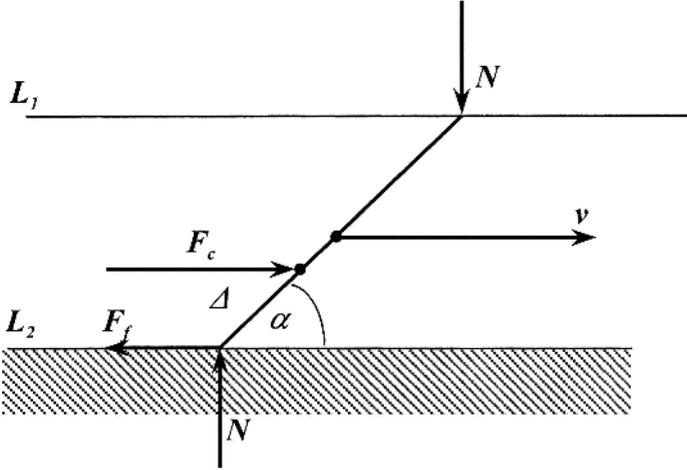
Fig. 43.2a. A model-like representation for the Panleve-Klain dry friction system with the Panleve paradox available.
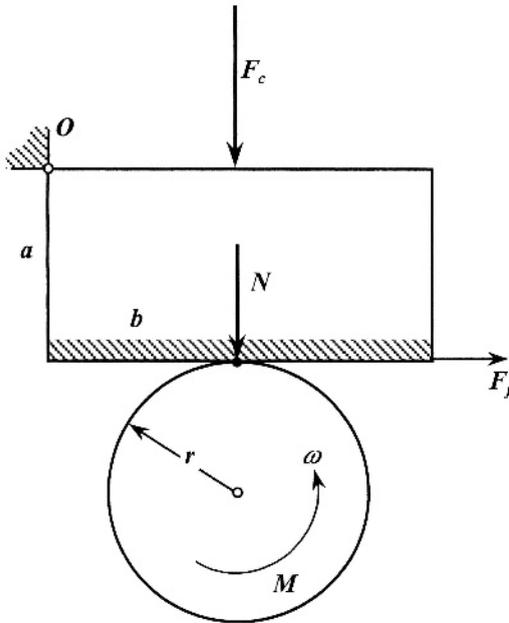
Fig. 43.2b. A model-like representation for a brake block.
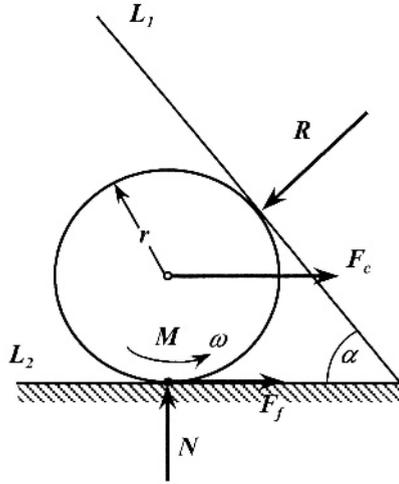
Fig. 43.2c. A model-like representation for the torsion shaft pressed into a constriction.

The equations of motions in these three cases are reduced to the similar form:

$$\dot{v} = 1 - F, \quad F = |N| \operatorname{sgn} v, \quad \chi N - \lambda + F = 0, \tag{43.1}$$

where v is a linear or an angular velocity; F and N are friction and normal pressure forces; χ and λ are parameters ($\chi > 0, \quad -1 \le \lambda \le 1$).

As a dynamical system, the equations (43.1) involve the single phase variable v . To obtain the solutions $v(t)$, $F(t)$, $N(t)$ for the given initial condition $v = v_0$, it is necessary to find F from the last two equations (43.1). However, for $\chi < 1$ this very operation is impossible: there may be no or two solutions. More precisely, in the case $\chi < 1$, the equations

$$F = |N| \operatorname{sgn} v, \quad \chi N - \lambda + F = 0 \tag{43.2}$$

for $\operatorname{sgn} v = \operatorname{sgn} \lambda$ have two solutions; otherwise, there exist no solutions. This is immediately seen, if on the plane of the variables N and F one depicts a graph of the functions corresponding to the equations (43.2) (Fig. 43.3).

Here, the broken lines 1 and 1$'$ stand for the first of the equations (43.2) for $v > 0$ and $v < 0$, respectively; the straight line 2 corresponds to the second equations in (43.2) and it is supposed that $0 < \chi < 1, \quad \lambda > 0$. According to

figure 43.2, in this case for $v > 0$ there will exist two solutions, and for $v < 0$ there will be no solutions. For $0 < \chi < 1$ and $\lambda < 0$, the situation will be similar. In this impossibility to find a necessary unique solution for the system (43.2) there lies the Panleve paradox. It points out the fact that the classic description (43.1) for $0 < \chi < 1$ is imperfect.
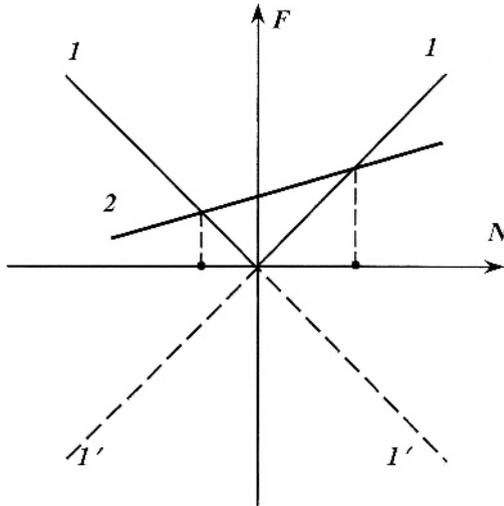
Fig. 43. 3.   The Panleve paradox explained graphically; the desired value of the velocity v is found as points of intersection of the lines 1 and 2 ($v > 0$), $1'$ and 2 ($v < 0$).

During the discussion mentioned above it was advised to take into account a real elasticity of solid bodies. In the Panleve-Klain problems, various ways of accounting elasticity were used; this really helped to eliminate the paradox, but, nevertheless, brought different results. To somewhat extent, these ways were done blindly, i.e. without understanding the deep essence of the Panleve paradox, though this or that way of accounting the elasticity helped to eliminate the paradox. However, a real reason remained existent; to clarify it, there were needed a new understanding of interactions between forces in the case of dry friction, a more full treatment of the Newton law about action and counteraction, and also there was needed a disclosure of the directedness for the action. There was needed an understanding of the fact that in the Coulomb law, the normal force N, causing the friction force F, is directed from N to F and is effected not at the expense of the force N, but at the expense of some external source of energy. Understanding this fact gave a new treatment of the equations (43.2) as a closed-loop cycle of the directed forces from N to F (the first relation in

(43.2)) and from F to N (the second relation in (43.2)). This directed cycle of interactions is depicted in figure 43.4.

a) F= /N /sign v

b) zN-λ+F=0
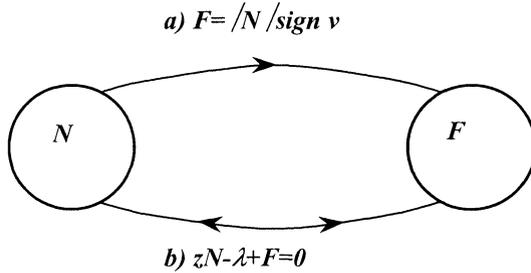
Fig. 43.4. The closed directed cycle of the instantaneous actions by the force generating the Panleve paradox. As a directed action in the cycle there serves the pressure N generating the friction force F .

The case $\chi < 1$ brings about the Panleve paradox, since in this cycle the interactions are instantaneous and its amplification factor, equal to $1 / \chi$, for $\chi < 1$ will be larger than unity. However, in a real system for all $\chi > 0$ there will be no paradoxicality, since the connections from N to F and from F to N are not instant, thanks to an actual elastico-viscous compliance of solid bodies and connections between them.

Such interpretation of the Panleve paradox makes it possible to suggest the simplest way for the account of elastico-viscous inertia compliances of real solid bodies and the connections between them in the form of time relaxation delays, by writing the equations (43.1) as follows [3]

$$\dot{v} = 1 - F, \quad \tau_1 \dot{F} + F = |N| \operatorname{sgn} v, \quad \tau_2 \dot{N} + N = \chi^{-1}(\lambda - F), \quad (43.3)$$

where $\tau_1 > 0$ and $\tau_2 > 0$ are small time delays.

Introduction of these delays and writing the singular equations (43.1) in the form of singularly perturbed differential equations (43.3) eliminates the Panleve paradox; but now instead of the first-order system (43.1) there appears a third-order system, and its studying is far from simple. This study may be simplified by presupposing τ_1 and τ_2 to tend to zero (for they are actually small), but under this presupposition the result turns out to depend upon the ratio τ_1 / τ_2 , and, when it is not fixed, the system (43.3) will have no limiting behaviour at all.

Certainly, the model (43.3) is not accurate; it is phenomenological. In some cases, a real nature of delays may be more complicated, but, nevertheless, this

model, though roughly, but specifies the paradoxical equations (43.1) and creates, at least, a possibility to understand what may happen in the Panleve paradox system. Notice here that similar equations and results may be obtained by some particular model-based techniques for accounting the elastico-viscous properties of solid bodies. However, against these particular techniques for specifying the equations (43.1) the equations (43.3) have some certain advantages, viz. the equations (43.3) are more general and involve two essentially small parameters, τ_1 and τ_2, while the model-based constructions take into account only one of them. This explains, in particular, why definite specifications yield different results.

Let us now study the singularly perturbed equations (43.3) and, more precisely, describe some fragments of this study, allowing to represent possible motions and phenomena generated by these equations. We shall start with the extreme particular cases of the equations (43.3): $\tau_1 = 0$ and $\tau_2 = 0$. Here, the order of the equations (43.3) will be decreased. Excluding the variable F, we shall rewrite them, for the purpose of comparison, in terms of the similar phase variables v and N:

$$\begin{aligned} \dot{v} &= 1 - \lambda + \chi N \\ \tau \dot{N} &= -N\left[1 + \chi^{-1} \operatorname{sgn}(v - N)\right] + \chi^{-1}\lambda \end{aligned}$$

(43.4)

or, respectively,

$$\begin{aligned} \dot{v} &= 1 - |N|\operatorname{sgn} v \\ \tau \dot{N} &= -N\left[1 + \chi^{-1}\operatorname{sgn}(vN)\right] + \chi^{-1}\lambda. \end{aligned}$$

(43.5)

As $\tau \to 0$, both models, (43.4) and (43.5), will be reduced to the paradoxical model (43.1) representing an absolute rigidity of solid bodies. These two models contain not only similar phase variables, v and N, but also depend upon the same two parameters χ and λ, and, as it turns out, undergo bifurcations on one and the same lines. Due to this, they will have identical partitions of the parameter plane into domains with different phase portraits, which are not always identical. The parameter plane is divided into six domains, A, B, C, D, E and G (Fig. 43.5).

The first three domains relate to paradoxicality $(\chi < 1)$ and the rest are not paradoxical. The phase portraits of the models (43.4) and (43.5) in the domain A are represented in figures 43.6a and 43.6b. In both cases we come to the uniformly accelerated speed-up Γ_+. In the first case, as $\tau \to 0$, a final motion can be preceded by one or two impacts followed by an infinite value of the force N and by a finite jump of the velocity v. Here, the coefficient of the velocity jump is equal to $\mu = (\chi - 1)(\chi + 1)^{-1}$. In the second case, instantaneous halts, i.e. the impacts with the coefficient $\mu = 0$, are possible. The impacts may be visually

represented by depicting the limit phase portraits for $\tau \to 0$ (Figs. 43.7a and 43.7b), corresponding to figures 43.6a and 43.6b.

In figure 43.7a the point M passes instantly into the point M' , and the point M' into M'' ,...., and, finally, into $\overline{M}$, whereupon there arises a uniformly accelerated speed-up. Analogously, in figure 43.7b the point M passes into M', M' into $\overline{M}$, the point K into K' , K' into $\overline{K} = \overline{M}$. From the post-impact point $\overline{M} = \overline{K}$ there starts a uniformly accelerated speed-up with the initial zero velocity.
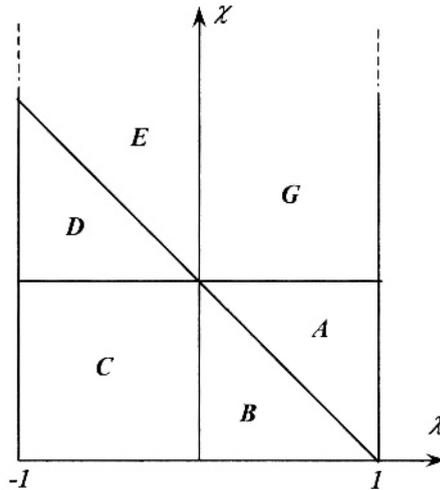
Fig. 43.5. Partitioning the parameter plane λ, χ into domains of different dynamical behaviours in the systems described by the equations (43.4) and (43.5).

Within the domain B , phase portraits for the models (43.4) and (43.5) are given in figures 43.8a and 43.8b, respectively.

Each of them includes two steady limit motions – the stable equilibrium state O and the uniformly accelerated speed-up Γ_+ . Between the stable equilibrium states O there exists some essential difference – in the first case, as $\tau \to 0$, the attracting domain (shaded in figure 43.7) infinitely decreases and vanishes ; in the second case (Fig. 43. 8), it retains. As before, for $\tau \to 0$, there may occur the impacts producing the earlier given velocity jump coefficients and yielding an infinite value of N . Impacts and limiting motions in the form of a stable equilibrium are possible in the paradoxical domain C as well. In the non-paradoxical

domains D, E and G, as $\tau \to 0$, no impacts will arise, and the limiting steady motions will be identical and unique. We emphasize here that the paradoxical case B produces a non-uniqueness of the limit motion, and as $\tau \to 0$, this motion might be different in different models (43.3) and (43.4).

Such are the specificities in the behaviour of the simplified models with the Panleve paradox. Specifying the model with due account of both time delays makes it possible to expose a greater variety of possible motions and a sudden possibility of auto-oscillations, which as $\tau \to 0$, may assume a form of the so-called contrast structures or infinitely increase their frequency, with the oscillation amplitudes of the forces N and F being retained finite. Without going into details concerning a rather complicated and, to a great extent, numerical computation, we will shortly describe now the basic facts. They will touch only the model (43.3), where the viscous friction is also taken into account, i.e. the model will be of the form

$$\dot{v} = 1 - F, \quad \tau_1 \dot{F} + F = |N| \operatorname{sgn} v + hv, \tag{43.6}$$

$$\tau_2 \dot{N} + N = \chi^{-1}(\lambda - F).$$

Dry friction in mechanical systems is well known to yield instability and auto-oscillations, but this occurs, if the friction characteristic has a falling component. The latter turns out to be not obligatory, if the mechanical system with dry friction is Panleve-paradoxical. In a paradoxical system, instability and auto-oscillations are possible due to a cycle of directed forces with an energy pump-up available in it; this role of the cycle cannot even be violated by a supplement of the viscous friction; on the contrary, it, in some sense, will strengthen this role.

Below we would like to present a new result. For small, but finite delays τ_1 and τ_2, there occur instability and a birth of auto-oscillations, according to the Andronov bifurcation. For τ_1 and τ_2 approaching zero, the auto-oscillations take a form of contrasting structures, i.e. they are accompanied with discontinuous jumps of the velocity v ("internal" impacts), at which the forces F and N turn to infinity.

The equilibriums of the system (43.6) are described by the equations

$$\lambda - F = 0, \quad \chi N - \lambda + F = 0, \quad F = |N| \operatorname{sgn} v + hv. \tag{43.7}$$

From them it follows, in view of $\lambda \leq 1$, that an equilibrium is available only for $h > 0$, $\chi > 1 - \lambda$, $v > 0$; for this we must have

$$F^* = 1, \quad N^* = \chi^{-1}(\lambda - 1), \quad v^* = h^{-1}[1 - \chi^{-1}(1 - \lambda)]. \tag{43.8}$$
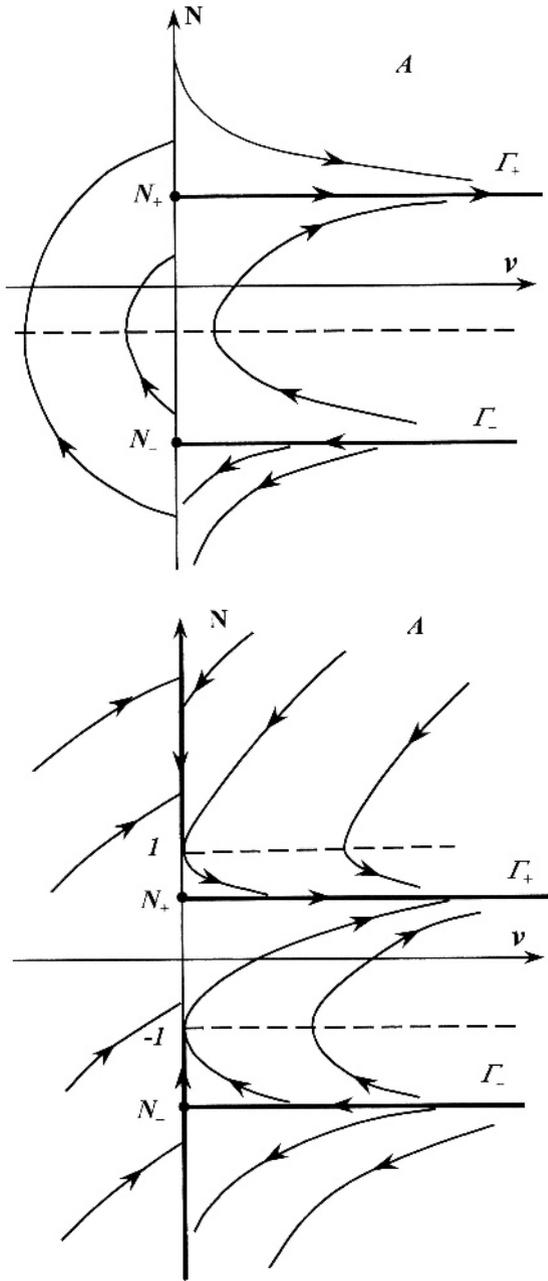
Fig. 43.6. The phase portraits for the systems described by the equations **a** (43.4) and **b** (43.4) in the domain A of figure 43.5.
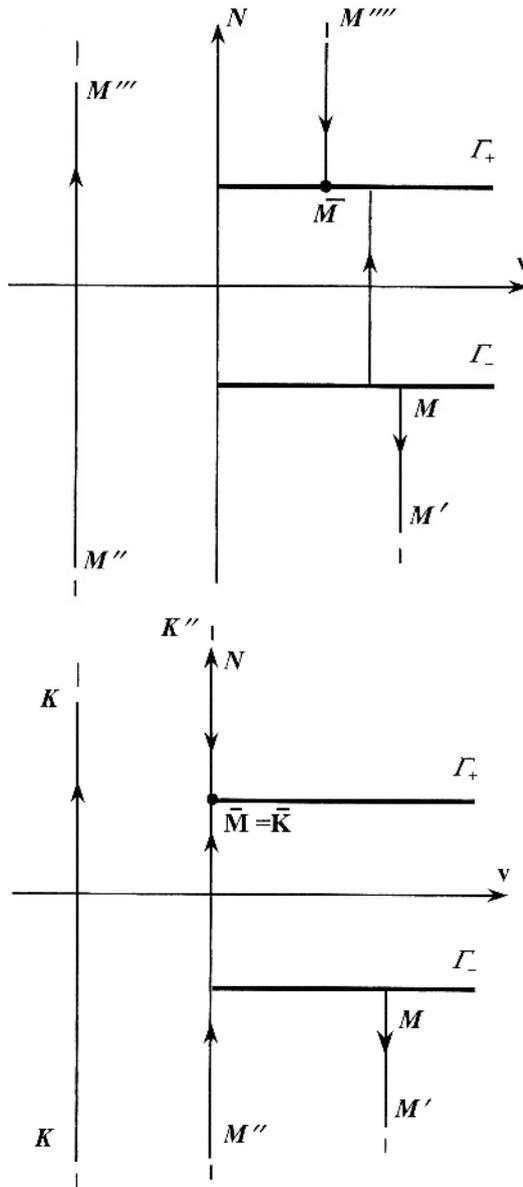
Fig. 43.7. The limiting phase portraits (as $\tau \to 0$) of the systems **a** (43.4) and **b** (43.5); in the case **a,** the coefficient of the internal impact changes the velocity $\mu = (\chi - 1)(\chi + 1)^{-1}$ times; and in the case **b,** it reduces the velocity to zero ($\mu = 0$).
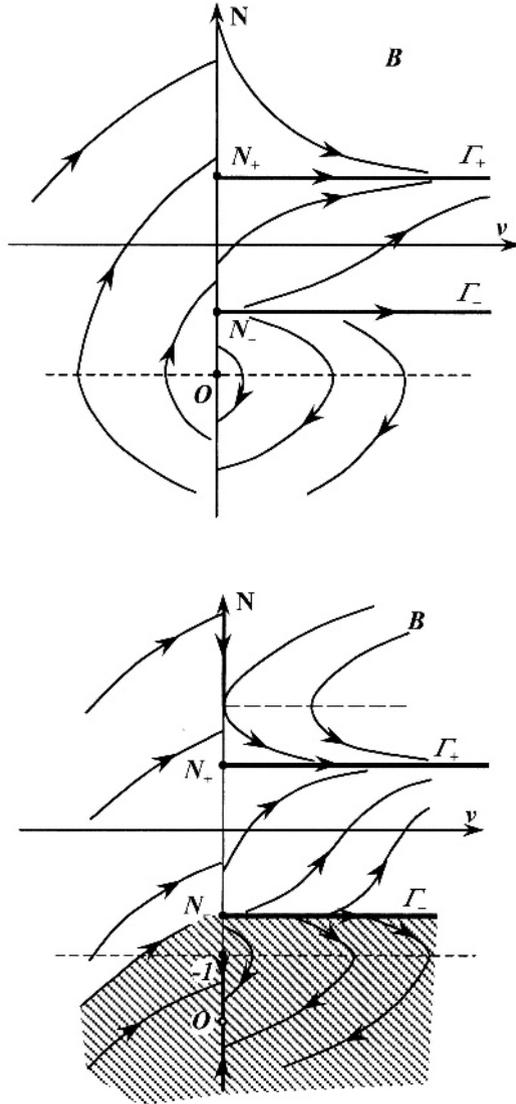
Fig. 43.8. The phase portraits for the systems **a** (43.4) and **b** (43.5) in the domain B of the parameters λ, χ ; in contrast to those in the domain A, the steady state motions of the systems (43.4) and (43.5) are now different.

According to (43.6), the characteristic equation is of the form

$$\tau_1\tau_2 z^3 + (\tau_1 + \tau_2)z^2 + (1 + \sigma\chi^{-1} + h\tau_2)z + h = 0,$$

where $\sigma = \text{sgn}\, N^* < 0$, and the unique condition for stability implies a fulfillment of the inequality

$$(\tau_1 + \tau_2)(1 + \sigma\chi^{-1} + h\tau_2) - \tau_1\tau_2 h > 0,$$

which for vanishingly small positive τ_1 and τ_2 is reduced to the unique requirement $\chi > 1$. Thus, the unique equilibrium (43.8) exists for $\chi > 1 - \lambda$; it is stable for $\chi > 1$ and unstable for $\chi < 1$; it loses its stability through a focus point.

Computer simulations show that with stability being lost there is born a stable periodic motion, i.e. auto-oscillations appear. With a decrease of τ_1 and τ_2, this limiting cycle is stretching, moving off, partly, to infinity; it assumes a form of a contrast structure with skips and splashes (Figs. 43.9 a ,b,c). The described phenomenon occurs within the parameter domain $1 > \chi > 1 - \lambda$, $\lambda > 0$, $h > 0$.

In another parameter domain, where $\chi < 1 - \lambda$, $\lambda > 0$, $h > 0$, for τ_1 and τ_2 approaching zero, there will also arise auto-oscillations in the form of the periodic contrast structures in figure 43.9c. The mechanism of their birth is different from the above described and consists in availability in the singularly perturbed system (43.6) such a structure of the phase space that provides the following:

1) there exist at least two manifolds of slow motions and only one of them is stable;
2) there exists a domain of initial values at which a phase trajectory, as $\tau_1 \to 0$, $\tau_2 \to 0$; can escape to infinity;
3) there exists such a peculiarity of the structure that provides a return of the phase trajectories from infinity;
4) there exists a mechanism for the move-off of the phase point from the stable manifold.

In a simpler case, $\tau_2 = 0$, when the system (43.6) becomes two-dimensional, the auto-oscillations of the contrast structure type arise as a result of satisfying the conditions 1 – 4 ; an example of this is given in figure 43.10. Here, Γ_+ and Γ_- are the stable and unstable manifolds of the slow motions, defined by the singular equations (43.6) ($\tau_1 = \tau_2 = 0$) for $\chi < 1 - \lambda$, $\lambda > 0$, $h > 0$; KC is a periodic motion of the contrast structure type with splashes, as $\tau_1 \to 0$.
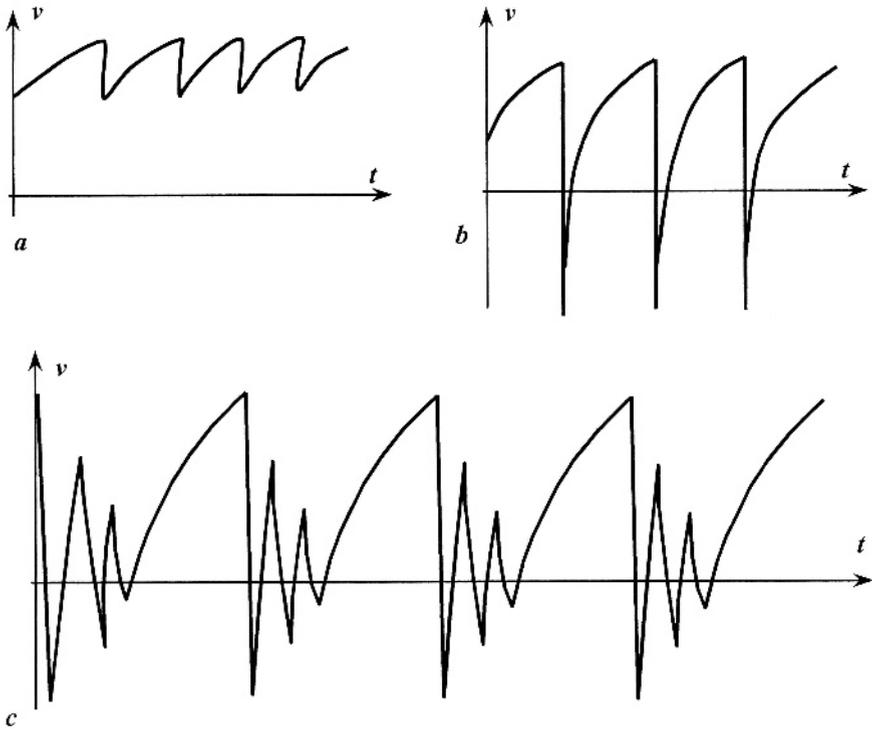
Fig. 43.9. The oscillograms of the periodic contrast structures (discontinuous auto-oscillations) for the system (43.6).

For $h = 0$, the periodic contrast structures, as τ_1 and τ_2 approach zero, do not appear, but for finite τ_1 and τ_2 there arise the auto-oscillations which, with τ_1 and τ_2 approaching zero, unrestrictedly increase their frequency. Here, the amplitude of the oscillating velocity v tends to zero while the amplitudes of the oscillating forces F and N remain finite.

Thus, the dry Coulomb friction in itself or in a combination with the viscous friction ($h > 0$) can, if the Panleve paradox available, yield auto-oscillations, which, in the case $h > 0$, with the rigidity of the solids involved in the system and with the connections between them being increased, will pass to various periodic contrast structures, discontinuous periodic auto-oscillations. Their existence is caused by the closed-loop cycle of the directed connections with an external

energy source. This is similar to what takes place in relay automatic control systems where the relay has the same characteristic as the Coulomb friction. We emphasize here that the average velocity of auto-oscillations retains always positive, and the external force F_0 and the moment M perform a certain work in each period. Specificity of these auto-oscillations lies in that that they may be close to discontinuous oscillations and may take the form of various contrast structures with skips and splashes. A peculiarity of the systems with the Panleve paradox lies also in their behaviour being essentially dependent upon small parameters and relations between them. In the models (43.3) and (43.6) under question , the small parameters are the parameters τ_1 and τ_2 ; the relation between them is expressed by the ratio τ_1/τ_2 .
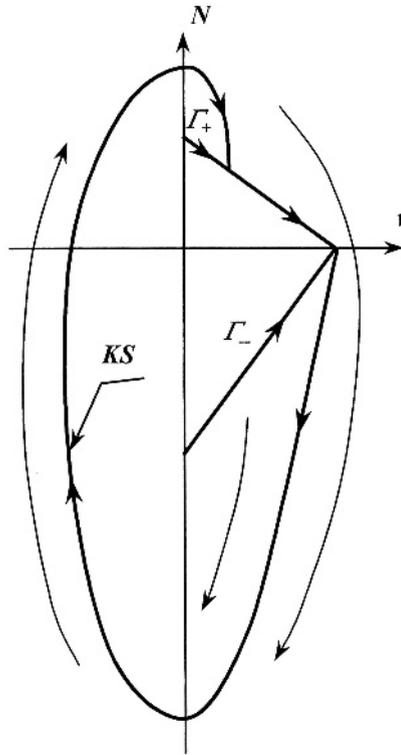
Fig. 43.10. The periodic contrast structures (discontinuous auto-oscillations) of the two-dimensional system (43.6) for $\tau_2 = 0$.

References

1. Neimark, Ju.I., Smirnova, V.N. (2000): Singular perturbed problems and the Panleve paradox. Differential equations. 36(11), 1-8 (In Russian).
2. Neimark, Ju.I., Smirnova, V.N. (2001): Contrasting structures, limiting dy-namics and the Panleve paradox. Differential equations. 37(11), 1510-1517 (In Russian).
3. Neimark, Ju.I., Smirnova, V.N. (1999): Idealization and the Panleve paradox. In: Bulletin of the Nyzhny Novgorod State University. Mathematical model-ling and optimization. Nyzhny Novgorod, 2, 53-66 (In Russian).

44 Dynamical interpretation of the least-square method and global search optimization with use of an adaptive model

The least-square method and the global optimization algorithm as an illustration of applying the dynamical system model .

As a basis in describing evolutionary processes and phenomena in nature and engineering and in the construction and study of their mathematical models there is used a dynamical system. All the models in this book are dynamical systems.

The below narration pursues a purpose to attract your attention to opportunity, naturalness and purposefulness of the dynamical treatment of some computational, searching and algorithmic procedures. This is done with a use of examples of the well-known and widely-used least-square (LS) method and the global search optimization with the adaptive stochastic model of the optimizable function. The first example demonstrates an increase of possibilities and efficiency of LS method applications as a result of its dynamical representation; the second example exhibits a width of the notion of the dynamical system and a width of the possibilities of description through a dynamical system.

44.1
A universal recurrent form of the LS method

The LS method is a method of estimating unknown parameters of theoretical models through indirect measurements. Its authors are considered to be K. F. Gauss and A. Legendre. The classical "method of equating" by Gauss was developed in 1795 and successfully used by him in 1801 for calculating a trajectory of the asteroid Ceres. However, firstly the elements of this method and its name "least-square method" were presented by Legendre in 1806 in his work devoted to calculations of comet orbits. Though, a probabilistic proof of the method in 1809-1810 and a profound development of calculating problems belong to Gauss. Further development of this method was given in the works of many well-known mathematicians – Laplace, Chebyshev, Neiman, Rao and others. An especially large contribution was brought to by A.A. Markov and A .N. Kolmogorov, who gave a rigorous proof of the LS method and suggested the boundaries of its substantive applicability.

Main advantages of performing estimations by this method include a priori solvability of the problem and such remarkable properties of the obtained estimates as their non-bias, efficiency and consistency. Thanks to these features, the LS method has become one of the most known and widely applied mathematical techniques for processing observations and experimental data. On the basis of acceptable hypotheses and mathematical models it allows to determine the unknown parameters and regularities not only in direct data processing but also in filtering, identification, pattern recognition, description compression (encoding), autocorrelated analysis and in other problems. However, there exist two difficulties in the application of the classical LS method. The first pertains to the inversion of a matrix (sometimes, of a large dimension), without any assurance in its correctness. The second difficulty comes from the fact that the data analysis, where the "initial data" is transformed to a "result", usually needs to change both the data samples and the hypotheses assumed with respect to processes and phenomena under study. The LS method in its classic form demands a full repetition of the entire calculations if any changes take place in the data used, in the models adopted and in the hypothesis assumed. That is why, it was as early as in 1821 that Gauss suggested a recurrent version of the procedure for the case of scalar observations. This version allowed to correct an earlier calculated estimate with due account of the results of new additional observations; in this case, any necessity in repeating the preceding calculations was eliminated. In 1950, this idea was generalized by Placett for the case of vectorial magnitudes. The recurrent version of the LS method has found a very wide application , especially in identification theory, adaptive control and filtration theory. Further expansion of the LS recurrent version to find a unique solution of a minimal norm through a matrix pseudoinversion was carried out by A. Albert. The universal recurrent form for the LS method was derived in [1-3]. In this version, the method is recurrent not only with respect to additional and eliminated data but also to the models used – basis functions, parameters to be searched, and linear restrictions on them.

The basis of the LS method constitutes a minimization of the quadratic functional

$$J(a_1, a_2, ..., a_m) = \sum_{j=1}^{N} \left(\sum_{i=1}^{m} a_i \varphi_i(x^j) - b_j \right)^2 , \qquad (44.1)$$

subjected to the linear restrictions

$$L_k = \sum_{i=1}^{m} c_{ki} a_i - d_k = 0 \quad (k = 1, 2, ..., l) , \qquad (44.2)$$

over the needed values of the parameters $a_1, a_2, ..., a_m$. In (44.1), $x^1, b_1; x^2, b_2; ...; x^N, b_N$ is the data to be used. The hypothesis lying in the basis of the LS application says that for the needed values of the parameters $a_1, a_2, ..., a_m$ and for the error-free data there will take place the depen-

dences that will zero the functional J and   convert all L_k to zero. These as-sumptions, which may be a precise or an admitted approximation, are defined by the chosen basic functions $\varphi_1(x), \varphi_2(x), ..., \varphi_m(x)$ and by the linear restric-tions (44.2) and all together will constitute the model assumed to be a basis for the specific application of the LS method.

Through the techniques of the linear algebra and optimization (the Lagrange multiplier method for conditional optimization), one can express analytically the parameter values minimizing the functional (44.1) under the conditions (44.2). This is the LS method in its original classic non-recurrent form. The difficulties and inconveniences of its direct application were described above.

The LS-based transformation of input data and the researcher's controlling commands to the output results can be represented schematically in the way shown in figure 44.1. Here, a rectangle stands for a LS calculating procedure; A , B and C are inputs for the data necessary for calculations (A is the initial data, B the basic functions, and C linear restrictions); D is the output supplying the values of the needed parameters. The classical form of the LS method, described above, presupposes all input data to be firstly   entered,   then calculations are carried out and the required output data is delivered.
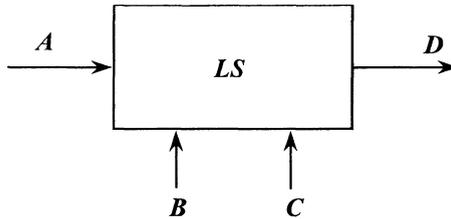
Fig. 44.1. Data processing by the LS method.

The universal recurrent form of the LS method implies that new data is con-stantly delivered to the inputs, A , B and C, or outdated data is eliminated; ac-cordingly, the output D will generate   the associated values of the needed pa-rameters. This process is exercised not through multiplex repetition of the LS calculating procedure but recurrently, on the basis of the state Θ introduced as a collection of five matrices containing, in particular, the values of the parameters a_s and of the functional J. So, we obtain

$$\Theta^{k+1} = T(u^{k+1})\Theta^k$$
$$a^{k+1} = f\,\Theta^{k+1}, \tag{44.3}$$

where Θ^k and Θ^{k+1} are the states in the k-th and the following $\underline{k+1}$-th cal-culating steps; a^{k+1} is a vector of the unknown values of the parameters in the $k+1$-th step; u^{k+1} is a vector of the input data and the researcher's controlling commands arriving after the k–th step. The vector u^{k+1} represents one of the following six possible input commands:

– introducing the new input data $\mathbf{x}^j, b_j$ or eliminating the outdated data;

– introducing the new function $\varphi_i(x)$ or eliminating any previous one, with re-spectively introducing the new parameter a_i or eliminating a previous one;

– imposing a new linear constraint or eliminating an outdated one .

   The operator for the controlled dynamical system (44.3) is implemented with-out matrix inversion; instead, only their addition and multiplication of matrices are needed. Though, a description of the operator $T(u)$ is rather cumbersome and so here it is omitted.

   The universal recurrent form described and its representation in the shape of a controllable dynamical system bring to the LS method new wider applications and provide necessary adaptability for the problems to be solved by this method. In particular, no matrix inversion is needed; the problem of weak conditionality for the information matrix is also removed; an error of the initial approximation, being inevitable under the previous partial recurrency, is now eliminated easily; a search for an appropriate model and elimination of errors in the incorrect or ran-domly strongly deviated data become much more effective. Thus, there appears a good possibility to construct a suitable basis, to compress a description and a coding, and to provide a current data processing, etc.

44.2
Global search optimization with a use of the adaptive model

The algorithm of the search for a global optimization, the minimization of the function $Q(x)$ over the set X, is presented by the rules for assigning the fol-lowing point

$$x^n = f(x^1, Q^1; ...; x^{n-1}, Q^{n-1}), \qquad (Q^i = Q(x^i)) \qquad (44.4)$$

and by the search interruption

$$g(x^1, Q^1; ...; x^n, Q^n) > 0. \qquad (44.5)$$

After interruption, as a "minimum" of the function $Q(x)$ over X there is ac-cepted the value

$$Q_{\min} = \min\{Q^1,...,Q^n\} = Q^s , \qquad (44.6)$$

being taken at the point $x_{\min} = x^s$. Here, the function $Q(x)$ is supposed to be given in the sense that for any $x \in X$ it is possible to find the value $Q(x)$. The next point x^{n+1}, after the points $x^1, x^2,...,x^n$, may be chosen under various assumptions. Below, we shall assume most probable the fact that in the new chosen point x^{n+1}, the probability of satisfying the inequality

$$Q_{\min} - Q(x^{n+1}) > \eta > 0 \qquad (44.7)$$

will be maximal. In order to have this choice in reality, the stochastic model of the function $Q(x)$ is assumed to be constructed on the basis of the already known values of the function $Q(x)$ at the points $x^1, x^2,...,x^n$ [4]. According to this model, the probability of obtaining $Q(x) < \overline{Q}$ is equal to

$$P(Q(x)<\overline{Q} / x^1, Q^1;...;x^n Q^n)=\sum C_S(x)\varphi\left(\frac{\overline{Q}-Q^s}{\gamma_S(x)}\right) . \qquad (44.8)$$

In this model, the functions $\varphi(\xi)$, $C_s(x)$ and $\gamma_s(x)$ are chosen from the available general representations for the functions $Q(x)$. In accordance with them, $\varphi(\xi)$ is a nondecreasing function of ξ, varying from 0 to 1 with ξ being changed from $-\infty$ to $+\infty$; $C_s(x)$ are such functions of x that $C_s(x) \to 0$ as the distance ρ_s between the points x and x^s increases; $C_s(x^k) = \delta_{sk}$ (the Kronecker symbol); $\gamma_s(x)$ is the increasing functions of the distance ρ_s, for example, being equal to $D_s\rho_s$, where D_s is the coefficient estimating the rate of changes of the function $Q(x)$ from the point x to the point x^s. As specific forms of the functions $\varphi(\xi)$ and $C_s(x)$ there may be taken, for example, the following

$$C_s(x) = \rho_s^{-1}\left(\sum_{i=1}^{n}\rho_i^{-1}\right)^{-1}, \qquad \varphi(\xi) = \begin{cases} \frac{1}{2}e^{\xi} & \text{for } \xi \leq 0, \\ 1-\frac{1}{2}e^{-\xi} & \text{for } \xi \geq 0. \end{cases} \qquad (44.9)$$

We emphasize here that the stochastic model (44.7) of the function $Q(x)$ covers not only a priori representations of the function $Q(x)$ but also the information obtained about this function upon n search steps. With each current step, this model is being improved. That is why it was called the adaptive stochastic model

of the optimizable function. The following point x^{n+1} is searched as a point of maximum for the probability of the inequality (44.7), being found as a maximum (44.8) for $\overline{Q} = Q_{\min} - \eta$. Therefore, the optimizing technique described above may be practically exploited only in the case, if the function (44.8) is optimized much simpler than the maximization of the function $Q(x)$. This is the case when the arising difficulties become significant or the computation of the function $Q(x)$ needs a lot of effort, whereas this computational effort should be reduced if possible .

At each search step, to choose a current search point requires to have the magnitude $\eta > 0$ already assigned in the inequality (44.7). This magnitude η may be considered as a controlling parameter of the search. It is intuitively clear that at the beginning of the search we should try as fast as possible to improve the model and only upon it, a search of the minimum itself is performed. These steps are advised to be combined during the search. In the beginning of the search, one may hope for a significant decrease of the values of Q, while in the end of it this decrease is less probable. From this it follows that, at first, one should choose large values of η and then decrease them. For large η it may be presupposed that the model will be improved and it will be confirmed by experiments; while for small η no large decreases of the minimum will occur. However, all these things are not more than general heuristic assumptions to be proved   experimentally. A more proved choice of η in each step can be done if there is available an approximate graph of the dependence

$$P\big(Q(x) < Q_{\min}\eta\big)$$

upon the number of η . This probability does not increase with η and, in general, decreases with η; this decrease may be slow or fast. For a slow decrease, it is natural to increase η till a reasonable limit; for a very fast decrease, the magnitude η must be chosen smaller, because a small, but reachable, result is more preferable to a greater, but very improbable. I think that here we encounter the very case when human intuition and human experience play a decisive role; and it is hardly possible that any attempts to improve the theory might be useful, at least, within the frames of the accepted model and within the searching technique.

However, here we will not deviate from our main subject. Now, we need to bring the above described search for the minimum to a controllable dynamical system, i.e. we should define a state, a state space and a state changing operator dependent upon the controlling variable η . For determining   the following search point x^{n+1} we have to know all previous sequences of   values $x^1 Q^1; x^2 Q^2; ..., x^n Q^n$ and also we should find the maximum over x of the

function (44.8) at the given $\overline{Q} = Q_{\min} - \eta$. We recall here that $x_{\min}$ is a value of x^s, for which Q^s is a maximum among all known values $Q^1, Q^2, \ldots, Q^n$. From the said it follows that as a state there should be taken the entire sequence of values $x^1 Q^1; \ldots, x^n Q^n$. As a state space there will be taken a set of such all possible sequences for all $n \geq 1$ with $x^i \in X$ and $Q^i = Q(x^i)$. An operator will be defined by the fact that x^{n+1} is a maximum of the function (44.8) for $\overline{Q} = Q_{\min} - \eta$, where $\eta > 0$ is a controlling variable being assigned at each searching step.

References

1. Neimark, Ju.I., Teklina L.G. (1996): A recurrent form for the least-square method over determinate parameters. Reports of Russian Academy of Sciences. 349(5), 608-609 (in Russian).
2. Neimark, Ju.I., Teklina L.G. (1995): An extended recurrent form for the least-square method in applying the pattern recognition problem. In:Dynamics of Systems (Nizhny Novgorod State University after N.I. Lobatchevsky), 29-45 (in Russian).
3. Neimark, Ju.I., Teklina L.G. (2001): Recurrent procedures of the least-square method under restrictions on parameters in coding and recognition problems. Pattern recognition and image analysis. 11(5), 228-230.
4. Neimark, Ju.I. (1978): Dynamical systems and controllable processes. Nauka, Moscow (in Russian).

45 Theoretical game model of the human society

> Basic functioning, structural and management principles of the human so-
> ciety. An ideal public game in the society. The problem of authority in the
> society.

Mathematical models we have considered up to now are though sometimes very
simplified but the models for the existing or certainly feasible objects and proc-
esses. When we were speaking about modelling, it was supposed by us to be the
modelling of real objects. However, mathematical models may also describe the
non-existent, imaginary, hypothetical and even fantastic objects. In one case, a
feasibility of such models does not give rise to any doubts, and in another cases it
does. When simulating a flight to the Moon, we are able to realize whether this
flight is feasible and what is needed for it. But there may be the models whose
implementation is hardly possible, though possible in principle. The similar
situation takes place in modelling a structure of the future human society or possi-
ble future attempts to simulate new genetic structures for living organisms and
man. This kind of modelling, let us call it fantastic, has also the right to live, and,
what is more, is interesting and may happen to be useful.

　These introductory words precede my final narration on a fantastic theoretical
game model of the human society [1-4]. Perhaps, this model is not only fantastic
but also very unusual, since constructing it relies, if necessary, not only upon
exact knowledge and exact sciences but upon general humanitarian and natural
science considerations and arguments; however, the conclusions drawn from them
are sudden and interesting and agree with some determinative tendencies observed
in the contemporary human society. Even in studying the human society mathe-
matically, the humanitarian considerations are inevitable, and the designation of
mathematics here lies not in escaping from them (that is impossible) but only in
assisting to make them more clear-cut and to neatly split what is accepted from
what will be inferred from it; and then to deduce what is needed and interesting.

　What follows below represents itself an attempt to comprehend a structure
and a functioning of the contemporary society (its tendencies and evolutionary
perspectives) in terms of such exact sciences as mathematical modelling, control
theory and oscillations theory.

　Mathematical modelling is needed for constructing the mathematical model,
which could serve as a basis for our further conclusions. The contemporary soci-
ety and its activity are supposed to be very complicated; to understand them we

use   different opinions. Here, the exact sciences, and mathematics in particular, seem of no help for us.

But actually this is not so. Not so, because such a phenomenon as   simple mathematical models can provide a general description of the object under study, by reflecting only its most essential specificities and features needed by us.

The control theory is here also needed, because any society is functioning on the basis of some purposeful synchronization of human actions; this synchronization is performed and maintained through management as well.

The theory of oscillations is also needed here, for this is an all-embracing science about evolutionary processes; and the functioning of the human society is some kind of the evolutionary process.

In their organization and functioning, the society and its constituents, people, look similar to a living cellular organism. Life to the entire organism is provided by an association of cells; a living activity of cells is provided through a homeostasis of the environment of cells in the organism. Something similar is available in a human society too. Life of the society needs some appropriate environment to be provided. Let this environment be called here a basic level of the society. Alongside with the perfection of the multi-cell living organism, there appeared, in addition to the environmental homeostasis, humoral and nervous controls, as a necessary condition for life. In the contemporary human society, they are analogous to the activities exerted by governmental and public organizations. In a living organism, everything happens as if on its own, under natural laws, in accordance with the tendencies of nature self-organization. These laws and tendencies are still insufficiently understood by us. Instead, in the human society, in contrast to wild nature, many things are arranged   conscientiously and with a purposeful interference of people into their life.

The above   analogy of a living organism with a human society is evident, though the differences between them are also astonishing, not to say more. With respect to the living organism, the society lies on the higher evolutionary level , when a cell   becomes intelligent and gets an opportunity to directly influence upon society's organization and activity. The cells of the cellular organism lack such an effective, direct and fast influence upon it. . This fundamental difference creates a gigantic acceleration for the life evolution, perfection and development of the human society, in contrast to a living cellular organism.

Man started to think about the functioning of the human society long ago. But only at present, to my mind, there has appeared a necessary scientific basis and this problem has reached its critical point. Up to some time, the social and economical structures of the society were formed and changed, mainly, in the way similar to the development of all the living on the Earth, i.e. they evolved biologically, not without trials, mistakes, dead ends and a selection. Alongside with it, this process was influenced by traditions, religious beliefs, general comprehensions and natural-science views. From the times of ancient Greece, there were outlined the two approaches in the study of the society. They can be denoted as "from the man" and "from the State". The State approach had triumphed both in describing historical events and the current life of the society. Under this ap-

proach, the State and its authorities, their interests and objectives were put into the foundation of all happening events and all fates of the society . Only recently there has appeared a sudden change in this approach; the interests of all people in the society as a whole received a partial account. All this is rather natural, since the State and the authorities executed the defending function , necessary for the entire society, against the external aggression, enslavement and annihilation. The development of material and productive basis of the society has stepped over these limits and assumed the form of general scientific, technical and informational revolutions. These revolutions , nowadays, largely determine the structure of the contemporary society. I guess that to-day something similar is also possible in social, economical and public spheres. If you like, time has come and, I think you will agree with me, at present there is no more important and urgent object for the scientific study than the human society. This study must cover not only the local problems of economical efficiency, profitability, revenues and the efficiency of enterprises, productions, agriculture, circulation of goods and money (that in itself is also very important) but, rather, the functioning mechanisms of the society, its general organizational principles and management.

There exists a great need in a global mathematical model and the theory of functioning, organizing and managing the society, the theory resting on this model. They could help the humanity in surmounting the imminent social, ecological, economical crises and catastrophes and could promote the social revolution similar to the industrial and informational. Though, as it seems to me, these changes will not occur by themselves, as this took place and takes place in the engineering and informational revolutions. There are at least the two reasons for this: the first is an absence of immediate personal incentive among those whom it depends upon, and the second reason is an insufficient competence of those whom it depends upon. In order to succeed in the sphere of material production, greatly needed are the both things, personal incentive and competence, whereas these things are not obligatory for the authorities. The incentive, if any, has a very vague foundation not arising from the fundamental interests of the authorities; as for their competence, it is of no necessity at all; more exactly, for a social revolution, no competence is needed. Though, to become a shoemaker one should learn, the personnel for the shoe-making industry should be trained in special colleges and institutions. In order to become a builder, you should learn as well. To become a doctor, you need to take a five-year course of special training and then have a clinical practice; upon this, you will also have to study and study continuously. Further, to become, say, a senior researcher you should not finish only an institute or a university but also have a PhD degree. Only upon all these things, you may become capable of winning a competition that will demand your personal incentive and purposefulness.

And now, what should you study, what skills should you develop and what should be obligatorily known by you in order to become a member of the Duma (the Russian parliament), a member of the government or a member of the city administration? No severe requirements for this. No requirements, since the authority is an owner of itself and no lows are prescribed for it. No requirements

are needed either, because the person willing to join the authority lines is usually demanded to possess, first of all, absolutely other distinguishing features. These features are quite different from the features thanks to which the people have performed the scientific and technical revolution, are performing the contemporary informational revolution and could have done the scientific and social economical revolution.

In order to have the scientific social and economical revolution really executed on the basis of knowledge and intelligence, it is absolutely necessary to have the authoritative activities absolutely open, an availability of the most severe criteria for estimating the authoritative actions, the mechanism to cut off incompetent actions, lies and trickery. As well, greatly needed is a constant inflow of energetic and talented people to the authoritative lines.

What can make these demands implemented? My thinking is that they may be implemented only through understanding the inevitability of the deepest and hardest crisis, understanding the coming catastrophe that is threatening everybody's existence including the authority's. Also, through understanding our own interests and profits both in the nearest and distant future, understanding the integrity of all the people as a species and their fate; and finally, through understanding what should be done and what man should strive to, in order to escape from the catastrophe.

The greatest temptation of any authority is gaining a forced supremacy over the society and the world. However, to perform management of a super-complicated and, above all, active object (and such is the contemporary human society) by forceful and even all-crushing methods is impossible without destroying it, without depriving it of its absolutely necessary self-organizational abilities bringing to the society efficiency and prosperity. The forced management will, in the final end , lead to an actual destruction of the human society. The authoritative management should not be dictatorial but promotional to the active nature of the people and to the most efficient employment of them. The management of the human society cannot be completely centralized. This management should rest upon its ability for self-organization; this self-organization should be promoted and supported by the authority. The collective intelligence and human activities cannot be replaced by any centralized authority – this intelligence constitutes a basis for life and evolution of the society.

Much of the above said concerning our desire to construct and discuss a global mathematical model for the human society basic level (the model helps to understand the functional, organizational and management principles of the society) was given here without proof and expresses the author's point of view. This viewpoint is expressed here not to make you obligatorily agree with it. Not all disclosed things will be even used. This, if you wish, is some kind of invitation to you for a discussion, to make you ponder on them.

Alongside with these things, any mathematical description presupposes the acceptance of some initial statements. Usually, these are the generally accepted representations of exact sciences. They are, so to say, hidden in the background: they lie beyond the discussion and are obeyed when a further theory is constructed. Nowadays, one cannot behave in such a way, since formulating the

initial foundations for the society implies their discussion and reasoning, since these foundations are not simple and not evident. In the given case, these initial foundations will be as follows: a game-like perception of life, human aspirations and what is wanted by humans from the human society. Only upon discussing and accepting all these things, one can pass to the conclusions that may be called the organizational and management principles of the society.

An initial proposition will be the idea that the society is created and exists for a human being, for all humans; hence, the society should correspond to the nature of a human and to his aspirations and wishes, which are compatible and accept-able for all people. Naturally, this initial proposition cannot be proved, it can be only reasoned and not more, since always there may happen a person who may say: "No, the society has been created only for me or for us, and the State is me" or "The world and all people exist for my people only and everybody must obey and serve to them". Similar viewpoints also hold powerful hidden reasons de-termined by human's nature and by the life being until now improved through a severe struggle for existence. This cannot be disputed, since that was so in reality. Though, in the contemporary stage of the life evolution this is already not so. This evolutionary mechanism has become outdated, it must be replaced, because it is ceasing to properly serve to the life progress on the Earth. The struggle for existence with its extreme manifestations in the form of armed conflicts and wars, terrorism, competitive, political and ideological murders, must assume other shapes. At present, in this way think most of the people, but not everybody. Nev-ertheless, to look into these fundamentals is much easier than into the human society. Right here, of possible help may be the exact sciences, since they are able to reduce complicated and very complicated questions to less complicated and even simple.

For us, the below idea that the society exists for all people equally will be an axiom. The struggle for existence as an evolutionary mechanism and the life improvement should be softened and replaced by something other. The necessity of this replacement is caused by the fact that the period of purely physiological and physical perfection of a human has been supplemented, if not replaced, with the period of his intellectual perfection. The intellectual perfection uncovers vast possibilities and ways for evolution and progress of life, which are faster and more effective. In a highly intelligent society, the struggle for existence brings the undesirable consequences, which cannot promote life perfection and successful activity of people. The struggle for existence has become not only unacceptable but – and this is the most essential – there appears an actual possibility to elimi-nate and replace it. This possibility was born by the success in science and engi-neering and by the technological innovations based upon it. Nowadays, a principal opportunity appears to provide worthy life for all people and to cease the exces-sive reproduction that feeds the struggle for existence.

There is no logic in formulating the initial postulates, but there are arguments and considerations. In order to protect humans against undesired consequences of the struggle for existence, the living nature has created restrictive instincts.

A human is capable of restricting himself and must do it consciously. Initially he did it on the basis of his religious ideas and beliefs, which were coming down as if from heavens, they were highly respected and awesome. To-day, this should be done on the basis of intelligence, natural and exact sciences.

The necessary correction of the struggle for existence implies a transference of this struggle to a competition, on the basis of the aspiration to achieve the preset objective and within the rigidly obeyed rules. This transformation of the struggle for existence may be called a game. The game where all participants are equal in their possibilities, nobody violates the rules and the winners appreciate the losers, because without the losers there would have been neither victory nor joy from it, and no game itself. Without the losers, a game mechanism of the competitive perfection of the humanity will not work.

Thus, we have a game, instead of the struggle for existence, for equal rights and equal opportunities for all people; finally, all this is for the sake of people. Such is the result of preliminary considerations concerning the postulates to be laid into the foundation of further, already logical, conclusions.

I foresee here and have already heard weighty objections:" all this is so, but people, due to their nature, instincts and egoism, will never obey this, even though they will accept it". Unfortunately, this is true. Almost true, since I see, nevertheless, the two objections: the first objection is that nature has, nevertheless, gifted man with morals, ethical ideas and standards, and with soul; the second objection is that humans are still able to realize that straightforward egoism and their yearning to catch a momentary profit will inevitably bring them to a crisis and catastrophe and that changes and admission of behavioural constraints by them will be, in the final end, more profitable and beneficial for them.

The description of the really existing threatening situation and possible ways to overcome it can help people to improve their behaviour. It is just here that the exact sciences, the convincingness of their conclusions and logic may be of great help. Of course, different nations and different peoples are at different stages of their capabilities to comprehend, think over and improve their behaviour. This will happen not instantaneously, but this must happen, since it is hardly possible that the gigantic self-organization of nature will make a failure. What is more, the appropriate tendencies in civilized countries are undoubtedly available.

45.1
Game-like perception of life and a theoretical game model of the society

You undoubtedly enjoy playing with somebody and something. A human life is also a game; more exactly, one may apprehend and describe it as a game, some-times, comic, tragical, interesting, boring, fascinating , odious,..., but almost al-ways, desirable. Probably, you all remember the words from the aria "What is our

life?… A game...". This phrase opens the popular Russian TV game "What? Where? When?" in the elite club of fans, where everybody can earn a good penny by his intelligence.

The well-known Russian writer M. Prischvin has once written: "The desired life is a game, and all who can are playing, but who cannot are working with the hope to play some other time".

Nature has endowed a human many features that promote his active and sometimes even selfless participation in a game: self-love, ambition, desire to be a leader and to gain the wanted.

A win in a game can cause emotions being unforgettable in pleasure and at-traction. People are able to torture themselves for long years with training for the sake of a unique instant of feeling delight and triumph of their victory in Olympic games or other competitions.

In addition, I will cite here the words of A. Einstein: "Success is a tireless la-bour plus an ability to treat life as a game...".

Thus, a game is not only desirable and loved, but it is efficient and promotes to win success in life. A game is an acceptable, desired and efficient form and an image of human life.

You have already got accustomed that in the play someone wins and someone loses. All cannot win simultaneously. This seems indisputable; but this is not so. To be more exact, there exist many games for which this does not hold. First of all, this is not so for the "man – nature " games; this is not so in social games either, i.e. in the games of people in the society.

Moreover, you have got used that when playing you should hide your playing "cards" and conceal the moves you are going to do. But sometimes this also turns out to be not so. It may be absolutely vice versa; sometimes it becomes even profitable to open your "cards" and even to arrange an agreement with your partners concerning the future joint actions.

I admit here that you may be surprised at this and even do not trust me. I will try to explain why the above said may occur and may be natural for numerous games. The matter is that games in mathematics are understood in a far wider way than those in which we play between ourselves for the sake of pleasure. Mathe-matics describes a game, and more precisely, a simplest single-move game as the situation, where every player makes his move and obtains his win depending upon his move and the moves of the remaining players. The above said can be visually represented as a table, where the first line holds all the players $I_1, I_2, ..., I_n$; and the second all the possible moves by them. Here, every player I_s chooses any move x_s from his set Ω_s of the allowed moves. In the last line are given the wins obtained by all the players upon their moves: the player I_s upon his move $x_s \in \Omega_s$ will get the win $f_s(x_1, x_2, ..., x_n)$. This is a simplest single-move game of n players, but this game is already sufficient to demonstrate the above said. For this purpose, let us use the example of Yu. B. Germeyer discussed

above. In the game invented by him every player I_s chooses any number $x_s (s = 1,2,...n), \ n \geq 3,$ between 0 and 1 and obtains the win

$$f_s(x_1,...,x_n) = x_s + \sum_{j \neq s} (1 - x_j). \qquad (45.1)$$

Let every player know only his payoff function (45.1). Looking at it, he sees that in order to maximally contribute to his win, he should choose $x_s = 1$. All players behave also in this way, i.e. they choose $x_1 = x_2 = ... = x_n = 1$, and as a result, according to (45.1), every player will receive a unity win. If all players expose their payoff functions and every player sees these functions, then upon pondering a little they will realize that they will win much more, if they all agree to choose $x_s = 0 \ (s = 1,2,...n)$, though "to their detriment". After that, everybody will win $n - 1$, which is more than the former win equal to $1 \ (n \geq 3)$. At the same time, every player has a temptation not to follow the treaty by choosing unity instead of zero and thus to win unity more, i.e. n instead of $n - 1$. Here, all the remaining players will receive unity less, and most likely this will break down the agreement and all, as a result, will be at a loss having obtained unity instead of $n - 1$.

It is not difficult to outguess why everything in this game does not look like in our habitual games. In this game, a sum of all wins is positive while in usual games it is equal to zero and a payoff of some players is obtained at the expense of losses of others only. For a two-player game with the zero payoff sum, the game will be antagonistic, with all its consequences so habitual to us. Whereas in the game against nature, the sum of all the wins of all players can be positive, at the expense of nature, its resources and gifts.

Now, we are able to formulate the game we play living in the society and participating in general labouring or other activities.

Each man I_s, by his will or according to these or those circumstances and ideas, chooses a desired game from the set O of possible games. Having chosen the game O_s, he simultaneously chooses the set Ω_s of his possible moves and the payoff function $f_s(x_1,...,x_n)$. The game implies every player to successively choose the set Ω_s, the move x_s and receive the corresponding win $f_s(x_1,...,x_n)$ dependent both upon his move x_s and the moves of remaining players. During his life, every player performs many moves and may change games, but the latter usually happens rather rarely.

It is worth noticing here that the moves to be chosen by the player I_s from the set Ω_s are defined not only by the set Ω_s but also by actual opportunities of the player. This circumstance could be described in the following way: the move to be chosen must belong to some set ω_s being inherent to the given player and

being determined by his entire prehistory, in particular, by his preceding game and his results. In principle, the payoff function can also vary with time.

45.2
Organizational and management principles of the society

We shall try to formulate what every man, living in the society, is willing to get and wants and what people want from the society. This seems very difficult to do and even impossible: the people are so many and they are all so different. But we will not go into details and restrict ourselves by what is general for all or almost for all. What we want to do is to take only a simple model as a basis and discuss it; we shall not deal with the model sufficiently and adequately describing the society in all its manifestations. The things general for all people will be the following:

1) an aspiration to satisfy their needs as much as possible;
2) a desire, in spite of natural competition, to unite with others in the a society, for more efficient production and greater satisfaction of their needs.

   The question about what people want from the society, besides the things indicated in the item 2, can be stated, in accordance with the preceding discussion, as follows:
3) aspirations of individuals to maximally satisfy their needs must bring, in the final end, to a possibly greater satisfaction of the needs for all members of the society; and, as a result, an individual success will be simultaneously a general success.

   Let us discuss the conditions of practicability for the postulates 1-3 (the aspirations of people and their desires) within the frames of the theoretical game model of the society; to be more exact, the necessary conditions for their implementation. Let us call them the organizational and management principles of the society.

   First of all, we notice here that equality of rights and opportunities for all people of the society were laid into the foundation of the very theoretical game model. Nevertheless, there retains some kind of inequality because of a disparity of the sets ω_s ; but this inequality is as inevitable as a difference in features is inherited by a human from nature, and also it is determined by what he was born and how he has shown his worth. Perhaps, this natural inequality should be somehow compensated by the society; this is observed, indeed.

   Further, according to the points 1-3 it is absolutely necessary to observe the rules of the game; and in particular, only in case of their observance the equality of rights and opportunities will take place. Only with these rules being obeyed, the game becomes attractive, desirable and efficient. From the contemporary point of view, the observance of the playing rules can be treated as an observance of human rights, and more exactly, the rights and responsibilities of each person-

ality. For observance of these rules and success of the game, which requires, possibly, more information about it, there should be open both the functioning of the society and the decision-making processes in this society. Upon them, there will depend the payoff functions of each player. We shall not specify what and how much must be known or how one should be able to forecast. What is important for us now is the thing that the society should be open for this. Thus, the openness of the society is one of the necessary guarantees for observing the human rights and for providing a successful game for all people of the society. Perhaps, this many-sided and essential role of the society's openness was highly estimated by George Soros and was chosen by him to be a basic characteristic for the desired society, he is striving to create. Also, one more necessary condition arises from the fact that the wider is the set of admissible moves, the wider will be the opportunities for optimizing the wins. This requirement of maximally admissible size of the sets Ω_s for the admissible moves may be interpreted as a principle of minimal compulsion, or a minimality of constraints for human moves . So, we have come to the three organizational and management principles of the society: observance of human rights, openness of the society and minimality of compulsion.

Now we will discuss how to approach the objective introduced and stated by the item 3. Earlier, consumption and products could be treated very widely. Below, they imply only money, themselves or their equivalent. Let us denote the monetary part of the win f_s by $\bar{f}_s$ and introduce additionally the actual monetary win φ_s for every player. It is clear that the general win F of the entire society will be equal to $F = \sum_s \bar{f}_s = \sum_s \varphi_s$. A win of everybody will promote to the win of all remaining players and to the general win, if we assume that $\bar{f}_s = \alpha_s F$ $(\alpha_s > 0, \ \sum_s \alpha_s = 1)$. However, in this way all players are deprived of the possibility to estimate the actual effect of their moves and, moreover, they lose interest in the game, because a win of the player will not actually depend upon his moves. This   can be avoided through the assumption $\bar{f}_s = \beta_s \varphi_s$ $(\beta_s > 0,, \sum_s \beta_s \varphi_s = \sum_s \varphi_s)$; then, the condition that an individual win should promote the general win of all players will be violated.

Both requirements can be observed, if we assume that

$$\bar{f}_s = \alpha_s F + \beta_s \varphi_s, \quad (\alpha_s > 0, \ \beta_s > 0). \tag{45.2}$$

Here the coefficients β_s should not be very small and the coefficients α_s too large. This is necessary for the changes of the payoff function (through which a player could estimate the efficiency of his moves) to be determined mainly by the function φ_s or, more exactly, to allow to understand how φ_s varies.

The objective 3 of the society also presupposes a maximization of the general win F via maximizing all individual wins φ_s, i.e. the maximum of the function F or the magnitude F close to it should be obtained through optimizing each of the functions φ_s over the variable x_s. It may be noticed here that if each of the functions φ_s depends upon x_s only, then from the optimization of all φ_s over x_s there will follow the optimization of the general win F as functions of all variables $x_1, x_2, \ldots, x_n$. We may hope that something similar will take place also in the case when the function φ_s has x_s as a leading and very essential variable. It is interesting to note here that the games with such a condition fulfilled will be most attractive.

Let us summarize what we have learnt about the functioning, organizational and management principles of the contemporary society. The functioning principle of the society is the game; the organizational principles are as follows:

1) Openness, an open access to the information regarding functioning, processes and events happening in the society;
2) The principle of minimal compulsion, minimality of constraints for the moves of the society members;
3) The principle of the human rights, i.e. the rules of life-like games being obligatory and equal for all members of the society.

The management of the society should assign the playing rules, i.e. the winning magnitudes $\bar{f}_s$ and the sets Ω_s of the admissible moves. This assignment should be exercised in such a way as

1) to make each player personally interested in the game and in the win, and to provide for him sufficient information and opportunities to optimize the game;
2) to make the total win F promoting all individual wins $\bar{f}_s$;
3) to make the optimization of the individual wins to simultaneously and adequately be the optimization of the general win F.

Besides, the management of the society should re provide an observance of playing rules and promote their improvement. The aspirations of people to perfect their needs may be implemented not only through more expedient moves in their games but also, in a wider aspect, through creating new games and new playing moves. New games and new admissible moves constitute new labouring tools, types and ways of production, extraction and processing, new technologies, industrial, agricultural and other processes, new living and working conditions facilitating a more complete satisfaction of human needs. Games are improved on the basis of the cognitive activity of people, due to their inquisitiveness and intellect, the desire to understand, investigate and improve. The game should properly

embrace these kinds of human activities as well and this leads us to a greater attractiveness of the joint playing life and promotes the general progress.

The core of these innovations is a cognitive, research and searching activity of people on the basis of a succession of knowledge and skills being provided through upbringing and education. These kinds of human activities are promoted by instincts and human nature, though not at all times they enjoyed a widespread support from the society and were included in the monetary win. The society in this question evidently demonstrated insufficient attention to its profit; and this lack of attention used to be only partially compensated by the tsar's favour, patron's and philanthropic activities. This underestimation has a powerful reason, viz. an absence of any monetary estimate for such actions, a distant nature of their effect and a distant material incentive for the society, as a whole but not for specific people. This has led to the situation that scientific and engineering research was highly stimulated mainly by the armed oppositions and conflicts. Unacceptability of this situation is evident but to correct it is not very simple. Aggressiveness and the aspiration to predomination lie in the human nature too and not yet recently were, as it seems, the main mechanisms of life evolution on the Earth. Armed conflicts are games as well, they are great from the historical point of view, but hardly acceptable at present and in the future. From the viewpoint of the theoretical-game model of the society these conflicts contradict the society's objective, but the introduction of games with cognitive actions (science, engineering, culture, education, etc.) is natural for the model, since these actions extend an appeal of the games, meet the inborn aspirations and dispositions of many people. Such games extend essentially, if not decisively, the effectiveness of the games of all people and the society as a whole.

45.3
An ideal public game

Proceeding from the theoretical game model for the public life of people, from their aspirations formulated in the section 45.2 (items 1 and 2), and from the objective 3 of the society, introduced in the item 3, we have come now to a necessity to observe a number of general principles. One cannot say that they are unexpected and new . Somehow or other, they have been expressed and even partially implemented in civilized countries. But alongside with them, absolutely other principles were declared and materialized as well. As for the win (45.2), one may say that it takes place in any society. The society, where $\alpha_s = 1/n$ and $\beta_s = 0$, may be called an ideal communism, and the society with $\alpha_s = 0$ and $\beta_s = 1$ a wild capitalism.

In a civilized society, the term $\beta_s \varphi_s$ in the formula (45.2) indicates an immediate earning and the term $\alpha_s F$ indicates, partly or completely, what a member

of the society possesses in the form of educational, medical, cultural, social, ju-
ridical supports and insurances and other duty-free services.

The new things we have now familiarized with are what principles must be obeyed if we accept the initial postulates and on what basis the coefficients α_s and β_s will be determined. Here, we clear out what functions are provided by each of these principles and what idea is held in each of these principles. In general, they are assigned to promote attractiveness and efficiency of games and, through this, the observation of playing rules.

In particular, there have been formulated the requirements for a universal game, including an economic game.. In their full size, these requirements can be hardly fulfilled, and this is not even demanded from them; instead, only some approximate observance of them will be sufficient. At the same time, a study of the games satisfying these ideal requirements is of a certain interest and allows to formulate the principles to be exploited as a basis for an approximate organization and self-organization of the games.

We remind here that a game implies that each of the players I_s, $s = \overline{1,n}$ chooses the value of his variable x_s and then obtains the win $f_s(x_1,...,x_n)$. A natural strategy of each player is maximizing his win over the variable x_s available for him. Let us call a game an ideal one, if this personal wins maximizing the strategy bring a maximization of the general win of all the players. This requirement can be written as the fact that

$$\max_{x_s} f_s(x_1,...,x_n), \quad s = \overline{1,n} \tag{45.3}$$

is achieved at the same point $O(x_1^*,...,x_n^*)$ that

$$\max_{x_1,...,x_n} F(x_1,...,x_n), \qquad F = \sum_s f_s(x_1,...,x_n). \tag{45.4}$$

We shall consider each of the players to be able to increase his payoff function over his variable x_s by varying it in accordance with the equations

$$\dot{x}_s = \varepsilon_s(t)\frac{\partial f_s}{\partial x_s}, \quad \varepsilon_s(t) > 0, \quad s = \overline{1,n} \tag{45.5}$$

or the analogous ones with small delays.

45.3.1
Conditions to obtain successful individual searches for the largest wins

The point being solved by us below touches the situation when   the strategy (45.5) brings a player to a successful fulfillment of the condition (45.3) and when the point O is unique.

Let us take the function V of the form

$$V = \frac{1}{2}\sum_s \varepsilon_s(t)(\frac{\partial f_s}{\partial x_s})^2 \tag{45.6}$$

and find its time derivative along a trajectory of the equation (45.5)

$$\dot{V} = \frac{1}{2}\sum_s \dot{\varepsilon}_s(t)(\frac{\partial f_s}{\partial x_s})^2 + \sum_{s,k} \xi_s \frac{\partial^2 f_s}{\partial x_s \partial x_k \xi_k}, \tag{45.7}$$

where $\xi_i = \varepsilon_i(t)\dfrac{\partial f_i}{\partial x_i}, \quad i = \overline{1,n}$.

The search strategy (45.5) can bring us to the fulfillment of (45.3) only if the function V converges to zero. The latter will occur when

$$\lim_{t \to \infty} \int_0^t \delta(t)dt = -\infty, \tag{45.8}$$

where $\delta = \dot{V}V^{-1}$. This is a necessary and sufficient condition.

From (45.6) and (45.7) it is evident that for the fulfillment of (45.8) it is natural to assume that the quadratic form

$$\Phi = \sum_{s,k} \xi_s \frac{\partial^2 f_s}{\partial x_s \partial x_k} \xi_k \tag{45.9}$$

is negative definite, so that

$$\Phi \leq -\chi(\xi_1^2 + ... + \xi_n^2), \quad \chi > 0. \tag{45.10}$$

Further, from (45.6), (45.7) and (45.10) we immediately determine that

$$\delta(t) \leq A - B,$$

where

$$A < \sum \left|\frac{\dot{\varepsilon}(t)}{\varepsilon(t)}\right|, \quad B \leq -2\chi \min_s \varepsilon_s(t),$$

and, therefore, for $0 < \underline{\varepsilon} \leq \varepsilon_s(t) \leq \bar{\varepsilon} \leq \infty$ the condition (45.8) is satisfied, since

$$\int_0^t \delta(\tau) d\tau < \sum \left| \ln \frac{\varepsilon_s(t)}{\varepsilon_j(0)} \right| - 2\chi \int_0^t \min \varepsilon_s(\tau) d\tau$$

$$\leq n \max\left\{ \left| \ln \frac{\underline{\varepsilon}}{\bar{\varepsilon}} \right|, \left| \ln \frac{\bar{\varepsilon}}{\underline{\varepsilon}} \right| \right\} - 2\chi \underline{\varepsilon} t.$$

From the last estimate, it follows that

$$V(t) < V(0)e^{a-bt}, \quad a, b > 0. \tag{45.11}$$

According to (45.11), a transition from any point of the domain for which $V(x_1,...,x_n) < c < \infty$ to the domain $V(x_1,...,x_n) \leq d$, where $d > 0$, is performed for the time not more than some $T < \infty$. From this it follows that all points of the domain $V(x_1,...,x_n) < c < \infty$ during some finite time will have passed to the points of the one and the same domain $V(x_1,...,x_n) \leq d$ and all points of this domain, for sufficiently small $d > 0$, will have passed to the unique point $O(x_1^*,...,x_n^*)$ for which

$$\frac{\partial f_1}{\partial x_1} = \frac{\partial f_2}{\partial x_2} = ... = \frac{\partial f_n}{\partial x_n} = 0, \tag{45.12}$$

and the domain $V(x_1,...,x_n) < c < \infty$ will hold no other similar points.

Since c is any finite positive number, then there will be no similar points anywhere at all. The point $O(x_1^*,...,x_n^*)$ is unique and for it, according to (45.9), one has

$$\frac{\partial^2 f_s}{\partial x_s^2} < 0, \quad s = \overline{1,n}, \tag{45.13}$$

and hence the maximum of each of the functions f_s over the variable x_s is attained in this point.

Thus, the equations (45.12) have a unique solution, the point $O(x_1^*,...,x_n^*)$, and this point is also a globally stable equilibrium state of the searching strategy

described by the equation system (45.5). This point is also the point of the maximum of all wins $f_s(x_1{}^*, \ldots, x_n{}^*)$ over the appropriate variable x_s.

45.3.2
Admissibility of small time delays

Let the strategy (45.5) be implemented with some time delays, so that

$$\dot{x}_s = u_s, \quad T_s \dot{u}_s + u_s = \varepsilon_s \frac{\partial f_s}{\partial x_s}, \tag{45.14}$$

where, for the sake of simplicity, ε_s is assumed to be restricted by the large constant $\varepsilon > 0$. For the proof of the statement stated in the heading, it is sufficient to take, as the earlier used function V, another function of the form

$$V = \frac{1}{2} \left\{ \sum_s \varepsilon_s \left(\frac{\partial f_s}{\partial x_s} \right)^2 + A \left(x_s - \varepsilon_s \frac{\partial f_s}{\partial x_s} \right)^2 \right\}, \tag{45.15}$$

where $A > 0$ will be then chosen properly.

45.3.3
When does the individual wins maximizing strategy bring to the global maximum of the general win?

An answer to the question in the heading is reduced to establishing the conditions under which the point $O(x_1{}^*, \ldots, x_n{}^*)$, being a unique globally stable equilibrium of the system (45.5) and, simultaneously, a unique point of the maximality of all wins (f_s over the variable x_s), could be also a unique global maximum of the payoff function for the general win $F = f_1 + \ldots + f_n$.

The uniqueness of the maximum, and, therefore, its global character for the function F, will take place if the quadratic form

$$\Phi_1 = \sum \xi_s \frac{\partial^2 F}{\partial x_s \partial x_k} \xi_k$$

is negative definite and an estimate of the form (45.10) is satisfied. It can be observed here that

$$\Phi_1 = \sum_{s,k} \xi_s \frac{\partial^2 f_s}{\partial x_s \partial x_k} \xi_k + \sum_{s,k,p\neq s} \xi_s \frac{\partial^2 f_p}{\partial x_s \partial x_k} \xi_k . \qquad (45.16)$$

The first quadratic form of this sum of two items is negative definite by virtue of (45.10); its main quadratic terms are

$$\frac{\partial^2 F}{\partial x_s^2} \xi_s^2 ,$$

and it is natural to presuppose that they are very essential, i.e. of larger magnitudes, since the variable ξ_s is, most likely, the most essential in the win f_s. The analogous term of the second quadratic form in (45.16)

$$\sum_{p\neq s} \frac{\partial^2 f_p}{\partial x_s^2} \xi_s^2$$

looks considerably more imposing, but it does not exceed the total result composed of rigidities of the playing effects of

I_s - player's moves upon the wins of all remaining players, which are not so many and may be averaged. The said is not more than an argument for accepting the condition of negative definiteness for the quadratic form Φ_1 and for the estimate of

$$\Phi_1 \leq -\chi(\xi_1^2 + ... + \xi_n^2),$$

which provides the fact that the maximum of the general win F will be unique and global; its coincidence with the point $O(x_1^*,...,x_n^*)$ will take place for

$$\frac{\partial F}{\partial x_s} = \sum_{p\neq s} \frac{\partial f_p}{\partial x_s} = 0, \quad s = \overline{1,n} . \qquad (45.17)$$

If $\dfrac{\partial f_p}{\partial x_s}$ is considered as a rigidity of the effect of the player I_s upon the

player I_p, then the sum (45.17) will be a sum of all rigidities of the pressure by the player I_s upon all remaining players I_p, $(p \neq s)$. Against what should this sum of rigidities be small in order to expect the closeness of the maximum for the general win to the general win in the point $O(x_1^*,...,x_n^*)$? This question will be studied in the following section.

45.3.4
The basic principle of the joint ideal game
in the society

An optimal general win is determined from the equations

$$\frac{\partial F}{\partial x_s} = \frac{\partial f_s}{\partial x_s} + \sum_{p \neq s} \frac{\partial f_p}{\partial x_s} = 0, s = \overline{1, n}, \tag{45.18}$$

i.e. the larger is $\dfrac{\partial^2 f_s}{\partial x_s^{\,2}}$, as compared against $\displaystyle\sum_{p \neq s} \frac{\partial f_p}{\partial x_s}$, the less essential will be the additional term in the equation (45.18). This follows from the first approximation for variations dx_s with respect to $x_s^{\,*}$, i.e. we obtain

$$\frac{\partial^2 f_s}{\partial x_s^{\,2}} dx_s + \sum_{p \neq s} \frac{\partial f_p}{\partial x_s} = 0, \quad s = \overline{1, n} . \tag{45.19}$$

A final conclusion on how close is the point of the global maximum of the general win F to the point $O(x_1^{\,*}, ..., x_n^{\,*})$ of the maximum of the individual wins can be formulated as a requirement of smallness for the sum of rigidities of effects of each player's moves upon the general win of all remaining players. This can be provided either through the smallness of each individual effect taken separately or through the smallness of the total effect produced; taking into account the possibility to compensate them. The said can be formulated as some behavioural principle in the general ideal economical public game: on the one hand, each player must strive to his personal win by optimizing it steadily, and on the other hand, he should not, as much as possible, bring harm in general and should not help others to win. It turns out that both the contributions and impediments to the wins of others will decrease the possible common win. Indeed, the first approximation of the general win over the sum of the individual maximized wins, according to (45.19), is equal to

$$dF = \sum_s \frac{\partial F}{\partial x_s} dx_s = \sum_s \frac{\partial F}{\partial x_s} \left[-\sum_{p \neq s} \frac{\partial f_p}{\partial x_s} \left(\frac{\partial^2 f_s}{\partial x_s^{\,2}} \right)^{-1} \right] =$$

$$= -\sum_s \left(\sum_{p \neq s} \frac{\partial f_p}{\partial x_s} \right)^2 \left(\frac{\partial^2 f_s}{\partial x_s^{\,2}} \right)^{-1} > 0$$

and is always positive; it will be the greater in size, the greater will be the squares of the sums of the effecting rigidities. It can be noted here that this excess, in any case for the first approximation, will be the smaller, the more "sharpening" will be the individual wins (i.e. the larger the second derivatives of f_s over x_s).

45.3.5
On expedient organization and self-organization of economic games in the society

Under an expedient organization there is understood a closeness to the ideal game. As one may judge from the preceding, the optimal economic functioning in the ideal game demands from its participants to be only skillful and successful in optimizing their personal wins. But the optimal game will be hardly possible; besides, who should arrange and define it. A more real thing will be some ap-proximation for the game obtained by using an expedient strategy of the eco-nomic games themselves and that of economic policy. The basic principle of this self-organization bringing the economic game close to the optimal game will imply the fact that each player should maximally strive to his individual win and simultaneously arrange the game in such a way as to make the total effects upon other players as less as possible. When following this strategy, he, while choos-ing his move, will maximally tend to escape from the effects of the remaining players. Such a self-organization naturally presupposes the players to be given sufficiently wide opportunities for choosing their moves; and this , in particular, implies them to be sufficiently informed and to have a full understanding of the consequences of their moves.

A trivial case of the ideal game is the case when each of the payoff functions f_s depends upon x_s only. This case may be reached through   uniting the play-ers with strong dependences into a "single" player, or through organizing the mutual relations between players, provided that these relations are not mutually strongly influential.

It is worth noticing here one more peculiarity of the ideal game: the success of optimizing the wins will entirely depend upon individual moves of each player, i.e. if his predictions are simpler and more reliable and he obtains a possibility to entirely concentrate himself only upon the efficiency of his moves and upon the facilities and technologies used by him. It may seem for somebody that the ideal game in its completely implemented form will exclude any competition entirely; but this is not so, because the competition will take place ,though not on the basic-level moves but on the level of comparing the playing results. And this is another important subject to deal with.

45.4
A problem of involving managers and authorities into a general playing interaction

Above we have defined a functional structure of the basic level of the society (as a playing society) and its main organizational and management principles following from the requirements of the game to be efficient and   acceptable. Further, there naturally arises the question: by whom and how is this basic level constructed and provided, and by whom and how are the games arranged and the observance of their rules provided? In   a living organism, this occurs , as it were, by itself, due to what may be called a self-organization. In the contemporary society, this should take place consciously and be provided by the activity of some personalities granted with special authoritative powers and by the authoritative structures, which can be divided into managers defining the games and the authoritative power itself, i.e. powerful   and   administrative authorities being responsible for the observance of playing rules and for the administrative serv-icing of the games. Alongside with this, there also arises the question of how these special powers can be made compatible with the demand that all the players should enjoy equal rights and play   one and the same common game.

The existence of the problem of managers was emphasized in our earlier mathematical model "producers – products – managers" , since in this model the managers arise   not because of the need in them and their usefulness, but be-cause they can be kept by the society.. In this sense, they look tragically similar to parasitical elements of the society – robbers, thieves, swindlers,… . From the viewpoint of the society's playing model, the problem consists in involving man-agers and both kinds of authorities into the equal - right universal playing inter-action, i.e. into the game general for everybody. The social status itself held by the   managers and authorities , so to say, hampers for them to enjoy equal rights. Though, it is the demand of the playing model that this equality of rights be observed. In my opinion, this is the very point where a central problem of the contemporary human society, i.e. the problem of power, is lying.

The description of   the authoritative functions and the functioning of the aµthorities lies beyond the framework of our playing model and requires its ex-pansion   and improvement; nevertheless, the problem of authorities , more exactly, the problem of inserting the authorities into a general playing interac-tion – this problem is already arising inside this model. From the positions of this model, there may be formulated the core of the problem and show the ways of its solution , i.e. , more exactly, to show what should be done   for its solution . Things should be done so that no one of managers or of authorities, in executing his functions, i.e. playing one of the assigned games, could have no possibility to use his power for changing or violating the rules of his game.

For this, there should be established   a purposeful and complete division of powers among   separate executors and management and authoritative structures. This   division should exclude not only an immediate use of their power for

their personal and corporative purposes but their collusion as well. Organizing the authorities , i.e. the games relevant to them, under the above listed requirements will automatically bring the so much needed high professionalism, competence and efficiency of managers and authorities, because only through these features they will be able to successfully play and win the competition laid in the playing rules. To some extent, it is in these very directions that changes in the contemporary civilized societies are occurring; namely, a separation of power, transparency of intentions   and actions, appointment through election and replacement.

45.5
Conclusion

No doubts, the theoretical game model considered above is primitively simple as compared against the real life of the human society on the Earth, with its various incomprehensible and complicated relationships between people, groups of people, countries and nations. All this cannot be embraced and adequately reflected by a mathematical model. We do not mean such a powerful model at all. Instead, we are discussing only the model for the basic level of the society, the foundation upon which the society is resting and the environment within which it is functioning. Our narration concerns only the environment   which will allow and promote an active and efficient activity of humans, the activity desired, convenient and acceptable for almost everybody. All the rest, i.e. above the basic level and its maintenance, is created, as if by itself, by active deeds of people and their capability for self-organization. All the rest is created by people during their general game, including also the mathematical simulation that helps them in describing all possible sides of their activities and life.

It is not necessary and impossible to demand from the basic-level model to be complete and adequate to a real human society. This demand is impossible, since any total model is so complicated that it will hardly make it possible to draw some general conclusions. We should and can demand from the constructed basic level the abilities to provide a deserving life for the people on the Earth, with its conditions, surrounding nature and human properties. This basic level must serve as some foundation for the efficient functioning of the society fully satisfying the needs of inhabitants.

The organizational , functional and management principles stated above are nothing but only the necessary conditions (they were obtained as necessary) and they will be sufficient, if this foundation, the basic level, is stable and self-sufficient.

The stability of this basic level demands that this basic level be acceptable and desirable by the overwhelming majority of the people. Also, this stability implies that this level can be implemented and maintained; and there should exist no actual possibilities to disorganize and destroy it .

The self-sufficiency of this basic level implies that the   basic level is sufficient to support and maintain itself, i.e. when the stability factors are incorporated into the basic level itself.

I am far from thinking that the basic level should be and can be implemented and organized forcefully . Conversely,   the above basic level, I think,   will be constructed on its own, thanks to natural activities of people; a beginning of the construction of this level is already vividly seen. Though, to accelerate and make this evolutionary process easier and to escape from erroneous actions and dead ends, it would not be superfluous to have a clear idea of its essential   purposes and perspectives. From the viewpoint of the above basic postulate concerning desirability and most efficiency of the game-like apprehension of life, stability and self-sufficiency of the basic level must take place. Here, the most difficult problem of humanity is to successfully involve all kinds of authorities into the general game and is also a presence of countries and nations falling largely behind in their development.

Of course, it may happen so that everything will be absolutely not so and what was said above is only a hypothesis unlucky in its implementation. That is why in the very beginning this playing model was called fantastic. Nevertheless, it is worthy here to say some more words about one specificity of the society self-organization.

For this, let us return to the above analogy between a living multi-cell organism and a human society. As an analogue for the basic level in cell organisms there serves the homeostasis of the biotope. However, except the homeostasis, as a necessary condition for life there appear humoral and nervous control systems that sustain the homeostasis and are not reduced to this function only. In a human society, as an analogue to humoral and nervous controls there exist deliberate authoritative structures being more or less centralized in the form of a State and its structures and institutions. This management, both in the organism and in the society, arises in the process of their self-organization. In the human society, authoritative structures are needed not only for sustaining the basic level; their power is considerably wider and this circumstance brings supplementary management structures and the necessity to involve them into a general playing interaction. Here, there arises the question about where and how runs the boundary between the authoritative management and the so-called spontaneous and natural self-organization demanding only a support of the basic level. Omitting our response to this not so easy question concerning this boundary, we note here that the problem of this supplementary management may be solved through properly dividing the power, through making the intentions and actions of the authorities more open and transparent for people, more predictable within possible limits, more liable to efficient public examination – all these requirements , as a whole, will make the authorities involved into the general game of the society.

References

1. Neimark, Ju. I., Ostrovsky, A.V. (1998): Flow-oriented model of economical dynamics. Bulletin of Nizhny Novgorod State Unuversity. Mathematical modelling and optimal control. 1(18), 105-115 (In Russian).
2. Neimark, Ju.I. (1999): Control theory as a science that is able and must help a human in solving social-economical and ecological problems. In: Proceedings of International conference on control problems, Institute of Control Sciences. Moscow, 131-138.
3. Neimark, Ju.I. (2000): A mathematical model of society that makes it possible to answer a question about principles of its functioning and management. Applied nonlinear dynamics. 1, 64-69 (In Russian).
4. Neimark, Ju.I. (2001): A theoretical game model of human society. Bulletin of Nizhny Novgorod State University. Mathematical modelling and optimal control. 1, 145-151 (In Russian).

Index

The numbers in the index refer to chapters, not pages; the number 0 refers to Introduction.

Foundations of Engineering Mechanics

Series Editors: Vladimir I. Babitsky, Loughborough University
Jens Wittenburg, Karlsruhe University

Palmov   Vibrations of Elasto-Plastic Bodies
(1998, ISBN 3-540-63724-9)

Babitsky   Theory of Vibro-Impact Systems and Applications
(1998, ISBN 3-540-63723-0)

Skrzypek/   Modeling of Material Damage and Failure
Ganczarski   of Structures Theory and Applications
(1999, ISBN 3-540-63725-7)

Kovaleva   Optimal Control of Mechanical Oscillations
(1999, ISBN 3-540-65442-9)

Kolovsky   Nonlinear Dynamics of Active and Passive
Systems of Vibration Protection
(1999, ISBN 3-540-65661-8)

Guz   Fundamentals of the Three-Dimensional Theory
of Stability of Deformable Bodies
(1999, ISBN 3-540-63721-4)

Alfutov   Stability of Elastic Structures
(2000, ISBN 3-540-65700-2)

Morozov/   Dynamics of Fracture
Petrov   (2000, ISBN 3-540-64274-9)

Astashev/   Dynamics and Control of Machines
Babitsky/   (2000, ISBN 3-540-63722-2)
Kolovsky

Svetlitsky   Statics of Rods
(2000, ISBN 3-540-67452-7)

Kolovsky/   Advanced Theory of Mechanisms
Evgrafov/   and Machines
Slousch/   (2000, ISBN 3-540-67168-4)
Semenov

Landa   Regular and Chaotic Oscillations
(2001, ISBN 3-540-41001-5)

Foundations of Engineering Mechanics

Series Editors: Vladimir I. Babitsky, Loughborough University
Jens Wittenburg, Karlsruhe University

Muravskii   Mechanics of Non-Homogeneous
and Anisotropic Foundations
(2001, ISBN 3-540-41631-5)

Gorshkov/   Transient Aerohydroelasticity
Tarlakovsky   of Spherical Bodies
(2001, ISBN 3-540-42151-3)

Babitsky/   Vibration of Strongly Nonlinear
Krupenin   Discontinuous Systems
(2001, ISBN 3-540-41447-9

Manevitch/   Mechanics of Periodically
Andrianov/   Heterogeneous Structures
Oshmyan   (2002, ISBN 3-540-41630-7)

Lurie   Analytical Mechanics
(2002, ISBN 3-540-42982-4)

Slepyan   Models and Phenomena in Fracture Mechanics
(2002, ISBN 3-540-43767-3)

Nagaev   Dynamics of Synchronising Systems
(2003, ISBN 3-540-44195-6)

Svetlitsky   Statistical Dynamics and Reliability Theory
for Mechanical Structures
(2003, ISBN 3-540-44297-9)

Neimark   Mathematical Models in Natural Science
and Engineering
(2003, ISBN 3-540-43680-4)