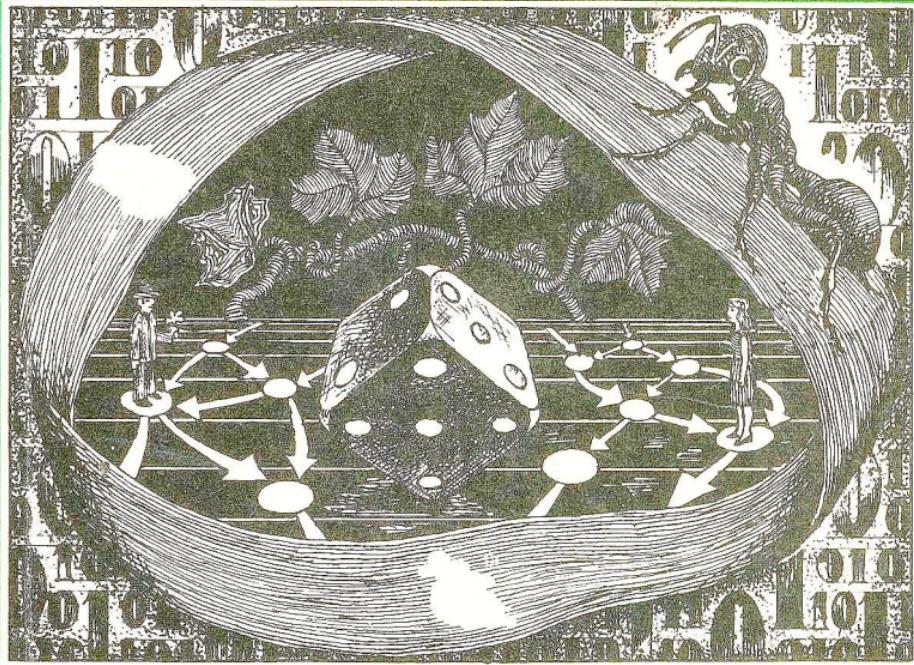


Did You Say Mathematics?

Mir Publishers

Moscow



Ya. Khurgin

Я. ХУРГИН
НУ И ЧТО?

Издательство
«Молодая гвардия»
Москва

Did You Say Mathematics?

Ya. Khurgin

Translated from the Russian
by George Yankovsky

Mir Publishers
Moscow

TO THE READER

Mir Publishers would be
grateful for your comments
on the content, translation
and design of this book.

We would also be pleased to
receive any other suggestions
you may wish to make.

Our address is; Mir Publishers,
2 Pervy Rizhsky Pereulok, I - 110,
GSP, Moscow, 129820, USSR

На английском языке

© English translation,
Mir Publishers, 1974

First published 1974
Second printing 1984

Contents

A FEW WORDS WITH THE READER	9
Mathematician and physiologist get together in September	12
Mathematician and physiologist get together in December	20
A radio engineer needs math	25
One last word to the reader	31
WHAT DO YOU THINK ABOUT MATHE- MATICS?	33
What is mathematics all about?	35
A little historical background	36
RUBBER-SHEET MATHEMATICS	41
Mathematics and art	47
Continuous transformations	47
A remarkable surface	55
Graphs	58
NUMBERS AND POINTS	71
THE MATHEMATICS OF A SADDLE	88
Extrema	96
Extremal curves	106

The epoch of Euler	108
Soap bubbles	110
MATHEMATICIANS ARE NOT ALL ALIKE	114
Where do axioms come from?	116
Two ways of reasoning	117
Induction and mathematical induction	125
The drama of equation solving—an historical sketch	129
ENGINEER CONSULTS MATHEMATICIAN	136
What is better?	141
Criteria	147
Optimization	150
How close?	152
Mary and Maude	155
Integrals—not so terrible after all	157
Space, distance, norm	162
Terms and where they come from	169
The problems of an oil engineer	172
Choosing a job	174
Model building	184
Mathematical models	187
Events and their models	191
Does one really need a mathematical model?	197
Modelling the oil-refining process	206
YOU PROBABLY LIKE THIS BOOK	209
Probability theory. Some background	212
Random events	213
Probability	214
An experiment and what came of it	216
Engineer consults mathematician	222
Experimenter and statistician	227
Decision making	231
Intuition and birthdays	234

Intuition and luck	239
Random walks	246
The drunkard's walk	258
The random-walk student	260
INFORMATION	268
Memory and codes	271
Information and what it's all about	278
Quantitative measures	282
The capacity of a communication channel	288
Coding	290
A language model and the transmission of information	295
Basic principle of the theory of the transmission of information	297
What about the content?	300
MATHEMATICAL MACHINES AND WHAT THEY CAN DO	303
The psychiatrist drops in for a talk	310
Pattern recognition	318
Technical diagnostics	327
Medical diagnostics	338
Replacing doc with a diagnostic machine	347
What is our life? A game...	349
One final word to the reader	360

A Few Words with the Reader

I like to argue and discuss things or just while away the time in friendly conversation. I don't like to write. Talking is better because it's a two-way affair. One gets a response, technically called "feedback".

During the past 25 years I have been involved in discussions with engineers, physiologists, doctors, geologists, and economists—people of different fields, views and talents. I've delivered numberless lectures and talks and I've conducted seminars. These talks deal with the problems and difficulties of various sciences about which I personally haven't the faintest notion most of the time.

I've never liked the idea of merely delivering a lecture—in some cases it amounts to simply reading a text. Nothing could be duller. I make every effort to carry on a conversation with my listeners.

It takes a long time to prepare a two-hour lecture, and even then I'm never sure of what it's actually going to be like because so much depends on the audience.

I imagine TV speakers have a hard time. After all, you can't laugh at your own jokes, and how does one

ask questions without getting so much as a silent nod for an answer?

Writing a book is like that too: there is no feedback. I find it very difficult to speak to an indefinite person, to an unknown reader. So here I'll speak to friends of mine from a variety of fields: physiologists, physicians, engineers, and geologists. We will talk about mathematics. The mathematician (the author, that is) will discuss matters with nonmathematicians. There have been many such conversations and there will be many more in the future. Why? For a very simple reason. A specialist is one who knows a great deal in a narrowly defined field of knowledge. Whereas it would take years for me, a mathematician, to collect the relevant facts of a problem concerning a specialist, the specialist can tell me all about his troubles and problems in a matter of minutes.

And he usually does it with the greatest of pleasure. My curiosity is satisfied and I do not even have to overcome my natural laziness.

In a word, I like conversations with specialists.

Why they come to me, a mathematician, is clear enough. We are in the midst of a mathematization of all sciences, even the descriptive sciences. At least that is what we read in the popular (and not so popular) scientific literature. That is what we hear over the radio and see on television. True, most people have a rather quaint idea of what mathematization is all about. Some think that the mathematician is capable of writing down equations for every imaginable practical situation. Others believe that electronic computers are about to take over and will do most of the thinking instead of human beings. Still others are sober enough to hope only for a certain amount of assistance from the mathematician.

Actually, of course, mathematical methods are no

cure-all for our many problems. But they are certainly applicable in every science if one takes the pains to apply them reasonably and properly.

Using mathematical methods is much like putting meat through a meat grinder—aside from having a good cutter and being able to turn the handle (and in the right direction!), you must put in quality products, otherwise you will grind out nothing but disappointment, in which case do not hurry to blame the theory because the blame lies elsewhere.

It is extremely important that the potential user of mathematical theory get acquainted with it and be capable of applying it appropriately, or at least be able to see when and where it is applicable. Users of mathematical theory will then be in a position to suggest new theoretical trends as they pose fresh problems, and the result will be of mutual benefit to all parties concerned.

The first encounters with specialists are in the nature of a competition, a clash. Each side is more interested in hearing himself than his adversary. Later, as the “battle” heats up, each side manoeuvres to establish its point of view. Then, finally, as a sort of mutual understanding takes place, both sides win.

Mathematicians, and I'm one, delight in such verbal fencing. We enjoy asking provocative questions like, “Now what is the question you are really interested in?” or “What sort of problem is it that's worrying you?” or even mere “So what?”

After a good deal of skirmishing, we finally arrive at a stage when the mathematician can begin to cooperate fruitfully with the specialist. Such joint undertakings are very satisfying and extremely fruitful to both parties.

If the reader finds these discussions exciting and useful, the author will consider his goal achieved.

MATHEMATICIAN AND PHYSIOLOGIST
GET TOGETHER IN SEPTEMBER

Autumn is always a fascinating topic for poets, writers and painters. For me, September means young people and the start of the school year. New students, new seminars, new problems.

My first encounter a few years ago was with a young and, so I heard, talented physiologist. He liked his subject and knew it well. He was enthusiastically seeking new pathways, new fields, and he earnestly wanted results. A person to my liking.

Mathematician (me, as usual). What topic are you working on?

Physiologist. I'm studying primary electrical responses of the visual zone of the cortex in the cat caused by flashes of light produced in front of the eye.

[I know what this is about. You insert a wire electrode into the cat's brain and bioelectrical potentials are recorded. The potentials are then fed to an electronic oscilloscope where they are displayed or photographed (see the upper curve in Fig. 1). The lower (periodic) curve is the time reference.]

Mathematician. Can you be a little more specific?

Physiologist. The stimulus is impressed as a pulse of light, the brightness of which can be varied. In the process the magnitude and shape of the positive and negative phases of the induced potential vary too.

Math. So what?

[How little one is able to put across on paper! The very intonation of the question contains a good deal of information. Right now it amounts to mere interest.]

Physiol. Just what do you mean? We have a definite relationship between the intensity of the light flash and all parameters of the electrical response.

[Note the words "definite relationship". What do they mean?]

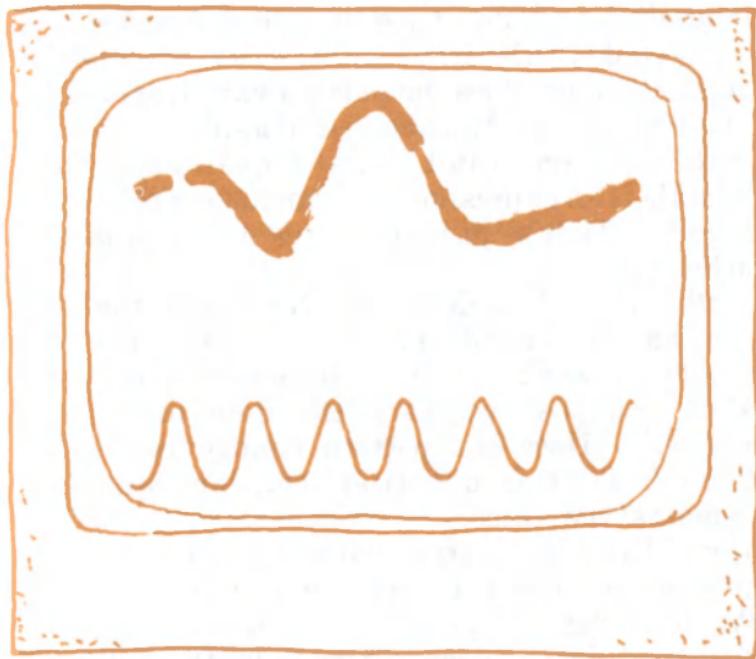


Fig. 1

Math. What kind of a relationship?

Physiol. For instance, with increasing intensity of the light flash the amplitude of the response at first builds up rapidly, then more slowly, and finally remains constant.

Math. That's fine, but where do I come in?

Physiol. I'd like to obtain a mathematical relation.

Math. Why do you need a mathematical relation?

Physiol. What do you mean "why"? Are you against applying mathematics to biology?

Math. Not in the least. I'm very much for it. By a mathematical relation you mean a formula, right?

Physiol. Yes, that's right.

Math. What will you do with the formula if I write it down for you?

Physiol. Please do. Then we'll run a series of experiments and verify it.

Math. Tell me, does the pattern vary from cat to cat?

Physiol. No, qualitatively it doesn't.

Math. But one can't write a qualitative formula. A formula is an expression of quantitative relationships.

Physiol. That's just what we need, quantitative relationships.

Math. That I understand. Now, are the animals under anaesthesia during the experiment?

Physiol. I work mostly with anaesthetized animals.

Math. Suppose we vary the dose or change the anaesthetic. Does the pattern change too?

Physiol. It does quantitatively, but qualitatively it remains the same.

Math. Does the pattern vary if you conduct experiments for a long time with one cat?

Physiol. Yes, it does. True, to varying degrees. But a good deal changes. It may be the cat gets used to it. Then, too, the depth of anaesthesia varies during an experiment.

Math. Why do you say that the relationship between the flash intensity and the duration of the phase is definite?

Physiol. Perhaps I did not put that quite exactly. But you get the idea, I'm sure. What I wanted to say is that a relationship exists.

Math. The point is this. Let's take the law of universal gravitation. It states a very definite relationship between the masses of two bodies, the distance between them, and the force of their attraction. Now in the process you are studying I don't see any definite, clearly defined relationship between the light intensity and the amplitude of the response (electrical reaction of the brain).

Physiol. But if we increase the intensity, then, as

a rule, the amplitude of each phase of the response increases as well.

Math. That is not a definite relationship by a long shot. What exactly is it you are studying?

Physiol. Academician A (or Professor B, or a famous scientist C) has developed a procedure for total registration of biopotentials from the auditory zone of the cerebral cortex. He worked with rabbits and investigated the auditory analyser. My chief posed the problem of investigating the visual analyser. We're used to working with cats, notwithstanding the extra trouble.

[Whenever someone refers to an authority instead of giving a direct answer, I get hot under the collar. I picture the experimental setup: a stuffy room, the strapped-down rabbit, dozens of instruments, and a multichannel loop oscillograph tracing out multitudes of curves: blood pressure, respiration, biopotentials from various parts of the brain, and more. A whole team of scientists carry out this involved many-hour experiment, and wind up, finally, by throwing out the poor little rabbit and also the metres upon metres of tracings—for the simple reason that it was never clear in the first place what they had intended to do with them.]

Math. What is the question that you want an answer to?

Physiol. (By this time also a bit exasperated.) Listen, I've already explained. We would like to know how the parameters of the primary response depend on the intensity of the flash.

Math. Suppose you have that relationship and a formula. How will they help?

Physiol. (Getting more excited.) Look, I told you that Academician A...

Math. O.k. What problem was he working on?

Physiol. Academician A was studying the effect of the intensity of an acoustic stimulus on the form of the primary response of the auditory zone of the cortex.

Math. As far as I can see, that isn't a problem but merely a descriptive topic. Did he get a formula?

Physiol. Of course not. He's one of the old type. He doesn't like any mathematics mixed into his biology and he can't stand formulas. His school holds that the task of the physiologist is to describe phenomena.

Math. Describe or explain?

Physiol. Classical physiology of course explains phenomena, only it does so descriptively.

Math. And how do you think they should be described?

Physiol. In an exact fashion. That's precisely why I want you to help me write down some formulas.

Math. I personally don't see much to admire in classical physiology. At any rate I find it hard to read physiology books: I find numerous facts and very arbitrary interpretations of them. That is to say, from the mathematician's point of view. And what you say isn't much better.

Physiol. (Resentfully.) Could you amplify that statement?

Math. Well, let's take an example. The leaves of various plants have distinct shapes. No one confuses a maple leaf with the leaf of a birch tree. Now take a flatiron and iron out the maple leaf, then trace around the edge. We get a curve. With some effort we can even write down the equation of that curve. Descartes, who invented the famous method of coordinates, made a study of a curve that goes by the poetical name of jasmine. The equation of this curve is

$$x^3 + y^3 = 3axy$$



Fig. 2

Now does that give us anything?

Look at Fig. 2. Here we have a graph. The portion of the curve in the first quadrant does indeed resemble a leaf. At least to some extent.

Physiol. Well, I don't know....

Math. Exactly. There's a whole literature dealing with the search for curves that describe the shapes of leaves. Many mathematicians, from Descartes to the present day, have studied this matter. At the end of last century the German mathematician Habenicht wrote a whole treatise entitled *The Analytical Shapes of Leaves*. A totally useless undertaking, in my opinion.

I think exercises of that kind compromise the

application of mathematics to biology, because no conclusions—at least conclusions worthwhile to the botanist—can be drawn from these formulas. Besides, the leaves of a single plant differ in shape, and so formulas only yield approximate shapes for the leaves. What is more, leaves are not, in reality, plane (flat) figures, but rather surfaces in space, and they are constantly undergoing change in the process of growth.

In short, it is not a matter of formulas. Mathematics does not reduce to formulas any more than music reduces to notes.

Physiol. My head's spinning. Where are we?

Math. Let's return to physiology. The study of the effects of the intensity of an auditory stimulus on the shape of the primary response is not a problem. It may be an intermediate stage.

What is the problem you want to solve?

Physiol. Electrophysiologists have a procedure for recording the biopotentials of the brain both in toto from large groups of cells and from separate neurons. We are investigating the responses of large masses of cells of the cortex to various stimuli.

Math. Tell me, if Academician A had never been born or were engaged in botany, and the new method of recording biopotentials of the brain were studied in standard courses of electrophysiology, would you be studying it now?

Physiol. If we knew as little as we do about the functioning of the brain, we would be engaged in the same work.

Math. What will you learn about the activity of the brain if you record primary responses? Methods and procedures aside, what, in the final analysis, is your problem?

Physiol. We're studying the relationship between the intensity of a light signal and the various charac-

teristics of the primary response. Here, take a look. [He extended a pile of photos—and the discussion went on, and on, and on.]

Math. Here's what I think. There is no direct unambiguous relationship between the two quantities that interest you (that is, the intensity of the flash and the duration of the primary response). The duration of the response depends on yet another dozen or so variables which cannot be recorded in an experiment. The relationship between the parameters of the stimulus and of the response is in your case of a statistical, probabilistic nature. For this reason, we are not able to write down a formula yielding a direct relationship between the quantities.

But the point is you do not need that. It would be silly, with no purpose in mind, to study relationships involving various quantities. *Your aim should be to experiment in order to obtain answers to specific questions, to build hypotheses, and then to verify them via experiments.* Which, of course, is what you do. But for some reason you don't seem to want to acknowledge it. Try to state explicitly the problem you want to solve.

Physiol. I'll try to do that since you've certainly got me cornered and I guess there's nothing much left to do anyway.

* * *

The reader may have the idea that our physiologist is rather weak even in his own field of research. Or, perhaps, he may have the view that neurophysiology is a second-rate science.

Neither is true. The neurophysiologist, like any biologist, has to do with living organisms, and a living entity (whether an animal or a single cell) is far more complicated than, say, any man-made machine. A living organism cannot be disassembled into its

constituent parts for each one to be studied separately. All the processes in such an organism are interrelated.

We can say that unlike man-made machines or systems, a living organism has a very large number of degrees of freedom. Practically speaking, an infinitude of degrees of freedom.

This puts the biologist in an extremely difficult position. It is precisely for this reason that biology has only just recently made the transition from the stage of passive observation of nature to broad active experimentation. At the present time, biology is working out a wide range of experimental procedures, it is in the search for new and more refined methods of investigation. And it is only just recently that biologists have become aware of the fact that in the study of living organisms there is a need for diverse mathematical methods, the development of new mathematical theories capable of describing adequately the complex laws of biology.

For this reason, the superiority of the mathematician in these (and other) discussions is in reality only superficial. It is easier to attack than it is to defend oneself.

For the mathematician to be of any real use and get beyond merely superficial criticism, he will have to study the branch of knowledge that the mathematics is being applied to. In this case it is neurophysiology. Only then can we expect fresh original ideas and proper conclusions in this new field of knowledge.

MATHEMATICIAN AND PHYSIOLOGIST GET TOGETHER IN DECEMBER

Three months passed after that first conversation in September. We met several times and went over the experiments and their results, and did a good deal

of arguing. I was often present in the laboratory during the experiments. I fiddled with the instruments and pitied the animals. And, as before, I fired the same questions at the physiologist and his coworkers. We organized a constantly functioning seminar. Gradually we worked out a common language and really got to understand each other. We felt we had grasped the true purpose of the experiments and were able to state the problem in clear-cut terms.

Mathematician. What's the story today? Have you obtained any new results?

Physiologist. We've got some results, but there's no news. On the other hand, I think it is now possible to state the problem more precisely.

Math. For the *n*th time?

Physiol. No, I hope this is the last time.

Math. Really! Let's hear what it's all about.

Physiol. The way I see it is this: the brain transforms—via the optic system—incoming light signals, and our problem is to figure out how it is done.

Math. We've already gone over that. The crux of the matter is not the signals but the information that they carry.

Physiol. That's precisely what I had in mind.

Math. What parameters of the light signal are carriers of information?

Physiol. That's just what I don't know.

Math. Hence, what we need to do is make a guess about these parameters and then verify our conjecture.

Physiol. Clearly, the most important parameter is brightness, the intensity of the light stimulus. Now, since there exists a certain statistical relationship between the intensity of the light signal and the amplitude of the primary response—you agree to that—we are consequently doing the proper thing by studying the primary response.

Math. Yes, there is apparently a close statistical relationship here. But what does it signify after all? It would appear that if isolated cells in the optic zone of the cerebral cortex respond to a stimulus, then the response is always of the same intensity. In the experiment we record the sum of the responses of many cells located in a specific zone. An increase in the overall response with an increase in the intensity of the stimulus apparently means that there is an increase in the number of responding cells as the intensity builds up.

Physiol. Yes, that's probably the way it is.

Math. But the cells do not respond simultaneously, do they?

Physiol. No, different types of cells have different delays in response to a stimulus. We say they have distinct latent periods. What is more, cells of different types respond in unlike fashion. Following a stimulus, a cell releases a series of pulses. Now the number of pulses and the intervals between pulses differ for different types of cells.

Math. Then this means the signals contain distinct information. Tell me please, is it true that each cell always responds to a stimulus in the same way? In other words, is it true that for a given cell the number of pulses and the intervals between them are constant quantities?

Physiol. It would seem so. At any rate, to a first approximation, as you would say. But if a stimulus is repeatedly delivered to a single cell via an electrode introduced into the cell, then the picture changes. However, it may be that this is not typical of a cell functioning jointly with other cells under normal conditions.

Math. What an awful number of reservations there are in physiology!

Physiol. Yes, a bit more complicated than a meat-grinder with a dozen or so parts where one can see at a glance what will happen if the handle is turned faster.

Math. That's clear enough, just as the principle of how a meat-grinder operates. But in physiology, it is the principles that lack clarity, and that's why it interests me. So it is precisely due to the spread in the latent periods and in the number and configuration of responses of different cells that the shape of the overall response (and not only the primary response) to a stimulus changes. Isn't that so?

Physiol. Yes, that's the way it is. But due to the intensive spontaneous activity of the cells of the brain there is a substantial background that is visible when stimuli are absent. The "tail" of the overall response to the stimulus is lost in the background and, for all practical purposes, cannot be isolated.

Math. Why is it impossible to isolate? I think it is not so impossible after all. My reasoning is rather simple. The background activity is the result of the activity proper of the brain. If we assume that the processes of the proper activity of different cells or groups of them are independent or but feebly related, we can then regard distinct sections of such a process as being independent too. Now what this means is that if we take, say, a hundred of such sections and superimpose them and then sum over them, we should get a resultant of just about zero.

Physiol. Background-activity experiments have been carried out. There appear to be some definite periodic processes. You've heard of alpha rhythms, beta rhythms, gamma rhythms.

Math. But aren't such processes slow in comparison with the responses we are studying?

Physiol. Yes, they are rather slow.

Math. Doesn't that allow us to hope for something?

Physiol. Have you a programme of some kind in mind?

Math. Suppose we attempt to extract information on the behaviour of the "tail"—the response to the light pulse—by means of a statistical treatment of a group of a hundred elicited responses. We will record them, compare the times of onset of the responses (or the delivery times of the stimuli), and then add them. The component of spontaneous activity will then, in the main, be eliminated, while the induced activity will be retained. In radio physics, that is a rather common procedure for isolating a weak signal from a background of noise.

Physiol. I have an idea of how to carry out the experiments, but how will we carry out the analysis? That's an extremely laborious undertaking.

Math. Yes, by hand it's too much, but we can make use of the technique of transferring continuous curves into discrete digital data and then we can process the material on a computer.

Physiol. Let's try it.

Which is what we did, and the results were very promising. I won't go into any more details now. The most important achievement was not the results but the fact that we had got to a point where we were understanding each other and could actually work together. We were able to formulate our immediate problem: to determine the parameters of a light signal to which the brain reacts. This was how a mathematician was able to help a physiologist by indicating a procedure for extracting information from observations. It was not a matter of formulas and equations but one of ideas and methods. It was only the first stage of a joint undertaking and I do not overestimate the importance of the results we obtained

It may be noted here that in reality the problem does not reduce to studying the parameters (of a signal) to which the optical system reacts. The problem is much more involved and profound. One thing is clear now: prior to processing the signals, the "system must know" why this has to be done. Only in that case can it reasonably select the parameters that are to be responded to; and only then will the received signal carry information that is useful to the system and not merely represent noise.

A living organism has to resolve an extremely diverse set of problems and it apparently has to reorganize itself depending on the problem at hand. We will return to this important range of problems later on.

A RADIO ENGINEER NEEDS MATH

This is a conversation I turned up by accident when rummaging through some old notes of mine.

A highly qualified engineer, a specialist in receivers, visited me recently. He is one of those who thinks up intricate designs. As engineers say, he has a feeling for circuits. Another thing. He is rather well-equipped mathematically. We got acquainted a long time ago when he was in post-graduate work in radio and I was teaching the graduate students mathematics and also learning their radio secrets myself.

Engineer. I've got an integral here I'd like you to help me calculate.

Mathematician. Whow! Where did you get such an enormous formula?

Eng. That's the way it turned out.

Math. Maybe there is a mistake in the computations.

Eng. No, I don't think so, I've checked it a number of times and I always get the same complicated integral.

I've gone through the reference books and haven't found anything remotely like it.

Math. What problem are you working on?

Eng. I'm working on noise stability in a "filter-linear detector-filter" system.

Math. Oh yes, that's an interesting field. What concrete problem were you tackling when you caught that monster of an integral?

Eng. If you don't believe I actually obtained that formula I can bring all the computations and you can check yourself.

Math. No, you don't really need to. I almost believe you as it is. But I don't believe that you needed it. A theory shouldn't have formulas that intricate.

Eng. Wait a minute. What do you mean "shouldn't"? That's what one gets if he considers an ideal linear detector. I took your advice and replaced the performance curve of a real tube with that of an ideal tube.

I recalled that he had indeed—about two months before—turned up with a request for a convenient analytical formula with a curve (graph) close to the performance curve of an ideal linear detector. Without thinking deeply about the problem I had suggested he try the function shown in Fig. 3. It was now clear that my advice had led to considerable complications.

Math. You know, I'm afraid it's all my fault. Let's take a better look at the problem. Just what precisely are you trying to solve?

Eng. A narrow-band pulse with noise is fed to the input of the "filter-detector-filter" system. We have to compute the ratio of the pulse signal to the noise at the output.

Math. Suppose you have already calculated the ratio of the signal to the noise, what have you got then?

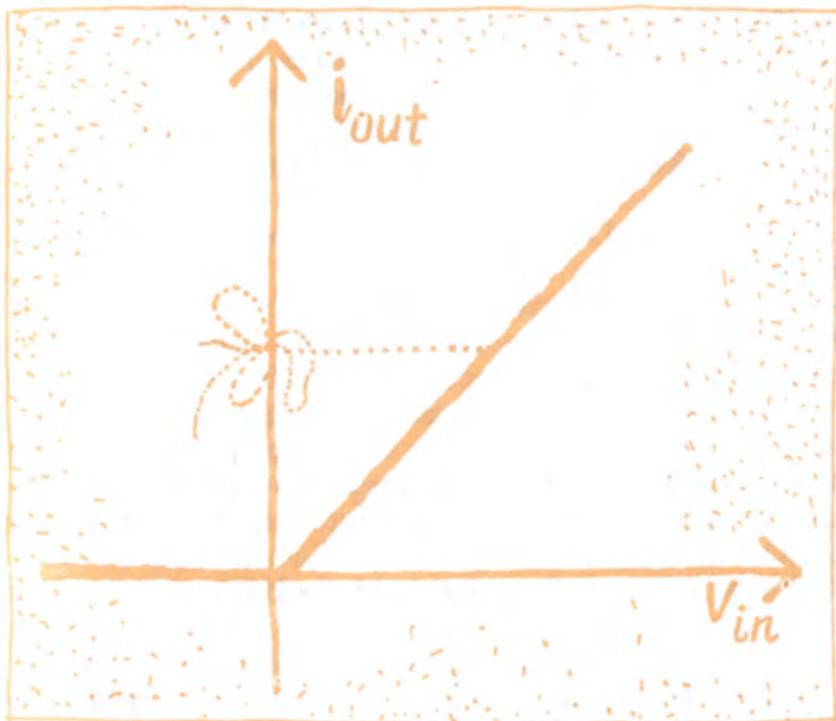


Fig. 3

Eng. What do you mean "what"?

Math. I mean, what do you want to do with the ratio?

Eng. I'll try to increase it.

Math. Now you're talking. If I see the matter correctly, the problem that interests you consists in selecting values of the parameters of the system for which (values) the signal-to-noise ratio will be as large as possible. Is that right?

Eng. Yes, that's it.

Math. Now what can be changed in the system, what parameters are in your hands, so to say?

Eng. If we assume the filters as given, then we only vary the performance curve of the detector.

Math. What accuracy can we actually obtain for this curve?

Eng. I would like to solve the problem in the general form.

Math. But in a practical sense what do you intend to do? Your system functions already, doesn't it? And rather successfully suppresses interference.

Eng. It works of course. The detector is a single electron tube. The circuit has two potentiometers. We can change the performance curve by varying the resistance. Then it's a simple matter of proper selection.

Math. What do you mean by "solving the problem in general form"?

Eng. Just writing down the general formulas.

Math. The equations have to depend on your initial parameters. If you cannot specify the initial data with absolute precision, then what sense is there in having an absolutely exact formula?

Eng. You see, this is material for my dissertation. There you need some theory, otherwise they'll say the material is not suitable.

Math. Is that the only reason you want to evaluate these integrals?

Eng. Listen, if I didn't have this dissertation paper to write, I wouldn't even be studying this business. I haven't got any extra time for such trivia. Actually, of course, if one has a set of neat equations, he can see what depends on what and then it is not hard to construct a system with better parameters. And that in turn can greatly boost the noise stability of the system.

Math. So the theory can be useful too, it turns out.

Eng. If the relations are simple.

Math. Then it's worth investigating. Tell me, what accuracy can be obtained for the performance curve of a detector?

Eng. Somewhere in the vicinity of one per cent, I'd say.

Math. And to be perfectly frank, how much?

Eng. I think we ensure an accuracy of only five per cent.

Math. That's more like it. What interval do you have for varying the input voltages?

Eng. Theoretically speaking, infinity, if one assumes that the noise has a normal distribution.

Math. Theoretically yes, but actually?

Eng. Practically speaking, there are no voltages outside the range between minus one and plus one volt.

Math. Now that's being specific. Let's try to state the problem. We have to choose a simple characteristic for the detector in the interval between -1 and $+1$ volt such that it will resemble the angle depicted in Fig. 3 and will ensure in that interval an approximate accuracy of at least five per cent. I think that we can make do here with a rather low-degree polynomial, say one of degree four or six.

Eng. That'd be marvellous! Then all the computations would be much simpler and the relationships between the parameters would be quite surveyable.

Math. That's what I think too.

Eng. I'll tell you what I'm afraid of though. My scientific adviser at the institute won't be particularly happy about such a formula. He'll say it's too simple.

Math. Listen, do you live a long way from the institute? How long does it take you to get there in the morning?

Eng. Oh, about 45 to 50 minutes, I'd say. But what has that to do with...?

Math. Merely that you're a specialist in microsecond techniques and my question is: could you measure the time it takes you to get to work to within a microsecond?

Eng. I suppose I could, but what sense would it make? One day hardly resembles the next, what with waiting for trolleybuses, being held up at home, and the like. Why would I want to measure the time with such an accuracy?

Math. That's what I think too. It could be done, in principle. The analogy with your problem here is complete.

Eng. Now I see your point. So what polynomial do we take?

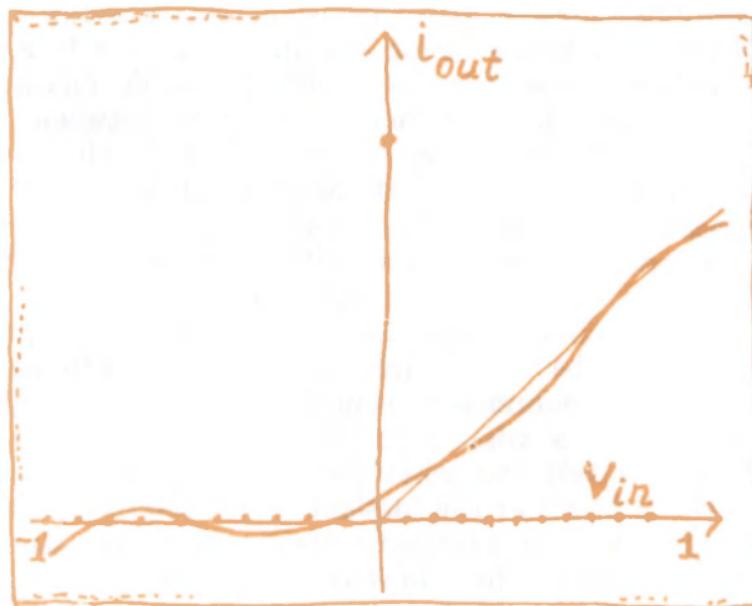


Fig. 4

Math. I'll try to work it out. Drop in in a couple of hours.

The polynomial didn't take much effort to figure out. The appropriate curve is shown in Fig. 4. Of course, it is not a question of choosing a polynomial, but of the general approach to such problems.

This conversation and a number of similar ones helped me to work out an important rule: when a mathematician is called in by a specialist of some other branch of science for consultation, he has to acquire a working knowledge of the subject matter and not merely answer questions.

ONE LAST WORD TO THE READER

You have just seen (and, I hope, participated in) two meetings of a mathematician with a biologist and an engineer. Later on there will be dialogues with specialists in other branches of science. These people needed the help of a mathematician, but each specialist viewed such aid in a different light. Our viewpoints concerning mathematics and its possibilities for utilization in the applied sciences did not coincide.

In the chapters that follow I will attempt to tell the story of mathematics and weave into an integral whole the various discussions I have had with my nonmathematical friends.

This will be a story of mathematics in popular language so that the nonmathematician will see what it is all about. This is not a course in mathematics but merely a series of sketches concerning ideas and methods. There will be no proofs to carry out and no need for paper and pencil. What I want to do is sketch a picture of the development of mathematics and show what mathematicians are presently engaged in—to some extent. It is hardly possible to examine

the whole field, but the fragments I hope to present will cover a rather wide range of outwardly unrelated mathematical theories and their applications.

It is best to read one chapter after another as they occur in the book, since they are all linked together. If some parts are uninteresting, don't throw the book down. Go over the earlier chapters once again. I'm sure you will find something that will hold your attention.

What Do You Think about Mathematics?

I find that at school the children like the teachers and not the subjects. A short time ago I delivered a lecture to would-be students of an engineering college. There were about five hundred in all. I asked them who liked mathematics and got two hundred raised hands. Then I asked them who were particularly fond of their math teacher, and again I got 150 to 200 positive responses. But when I asked them how many liked mathematics but did not like their mathematics teacher, I got four hands out of 200 lovers of mathematics!

Most of the students who finish grade school and high school, and even college with a short course in higher mathematics ordinarily forget almost all the details and even the mathematical methods they studied. They find it hard to recall even isolated fragments and they don't often know what use it has been to them.* What they usually remember are poor

* It ought to be very interesting to make a check of all the sciences and find out what a person remembers 5, 10 and 15 years after leaving school. What has been of use and what is totally useless or even harmful. Very interesting!

marks, funny or dramatic incidents and, finally, some of the theorems that caused them a headache or two. Generally they adhere to one of two contrary points of view.

The first—rather with haughty disdain—goes something like this: "Mathematics is a dull and truely boring science where all you do is count; oh, it's like book-keeping. Where does anyone find use for problems in pouring water from one pool into another? Time spent on things like that could be used more profitably. Why seek a complicated method for determining the third side of triangle via the other two sides and the angle between them? First, it's simpler to use a protractor for the angle and then measure the other side with a ruler, and second, that's another instance of something that is never needed in life." And so on in the same vein.

The other viewpoint is one of awe. "Mathematics? Oh, good heavens, that's very difficult, that's complicated, that's way out beyond the reach of the ordinary person. Only talents and men of genius can understand what mathematics is all about. Mathematicians pose fantastic problems and even find solutions to them."

But both camps firmly believe that mathematics consists of algebra, geometry and trigonometry, and also something called higher mathematics, which is pictured as a conglomeration of such intricate formulas as to be a complete mystery to both parties.

Arithmetic doesn't seem to be connected with mathematics; it has something to do with childhood and is as common as the alphabet, penmanship and babies' illnesses.

WHAT IS MATHEMATICS ALL ABOUT?

The sciences taught at school constantly undergo change. At school my parents never heard of Marx or

Lenin, Rutherford or Einstein, Gorky or Mayakovsky, Darwin or Popov. But the geometry of Euclid, the Pythagorean theorem, the formulas for solving quadratic equations, and the representation of the sine of a sum of two angles have all been taught to numerous generations of children, and will continue to be taught. This creates the impression of mathematics as of a fixed and finished edifice.

Can you picture for a moment the physics and astronomy of the 17th century prior to Newton's discovery of the law of universal gravitation and the famous three Newtonian laws of motion, prior to the discovery of electricity and electromagnetic induction, prior to Coulomb, Volta, Ampere, and Faraday?

It is still easier for the biologist or chemist to visualize the chemistry of the 17th century, just before Lomonosov and Lavoisier, or biology and medicine prior to the microscope of Anton van Leeuwenhoek.

But scholars of the 17th century actually knew everything that is covered by present school geometries and algebras and even much more, while a good deal of the information in these textbooks was common knowledge to Euclid himself—the third century B. C.

The oppressing antiquity of school mathematics, traditional as religion, is precisely the starting point for the conclusion that mathematics is a completely fossilized structure.

But this is not so. During the past 300 years, especially in the past century, mathematics has made tremendous strides. I will try to show the reader that mathematics is quite different from the boring aspects that are frequently handed out to school children.

Our first question is: "What is mathematics?"

We might start with the philosophical definition of mathematics given by Engels: "Mathematics is a science whose subject matter is spatial forms and quantitative relationships of the real world," or we could take advantage of the aphorism of the outstanding German mathematician of the end of the 19th century and the beginning of the 20th century, David Hilbert, who said that "mathematics is what competent people understand the word to mean". All well and good, but for a true understanding of any science one must at least roughly outline the spheres of its influence, describe the subject matter and the method it employs.

In a little book it is clearly impossible to consider separately and in sufficient detail the subject and the method of mathematics, although the mathematical method is the principal thing, as I see it. But it will be seen later on that the subject matter of mathematics is also of great interest to anyone concerned with science as such.

A LITTLE HISTORICAL BACKGROUND

At the dawn of humanity we see the origin of counting and then, with trade, the sharing of booty and products came the development of arithmetic.

Geometry, the measuring of land, also arose in remote antiquity. However, already two and a half thousand years ago, the works of the geometers of ancient Greece completely divorced geometry from the surveying of land and converted it into a science of spatial relations and the shapes of solids. Geometry was then constructed on the basis of a number of axioms acting as starting propositions and applied without proof, and theorems, which were derived from the axioms in a consistent deductive manner.

The construction was so faultless, so perfect, that for over two thousand years (right up to the start of the nineteenth century) no alterations were made in the foundations.

The more complicated problems of trade and industry called for the solution of equations and the introduction of literal symbols. Thus arose algebra, which at the time amounted to a science of equations. Even in antiquity, solutions had been found for equations of the first degree and for quadratic equations, those stumbling blocks of today's school children.

Enormous efforts were put into solving equations of degree higher than the second, and only in the 16th century were such solutions forthcoming for equations of the third and fourth degrees.

Another three centuries were spent in vain efforts to get the solution of equations of degree higher than the fourth. Later on we will return in more detail to this exciting problem and the dramatic events that accompanied it.

In the middle of the 17th century the demands of mathematics itself led the celebrated philosopher, natural scientist and mathematician René Descartes (in 1637) to a union of algebra and geometry, to utilization of algebraic methods in geometry. Thus arose analytic geometry in which straight lines, planes, circles and other curves and surfaces are specified by means of equations in a rectangular or, as it is sometimes called, Cartesian system of coordinates.

In Fig. 5 we see a straight line and a circle of radius r with centre at the coordinate origin and their equations in the Cartesian system of coordinates. Later on we will discuss coordinate systems in more detail for they are extremely useful.

The first step mathematics took after a lull of many centuries was the creation of analytic geometry. The

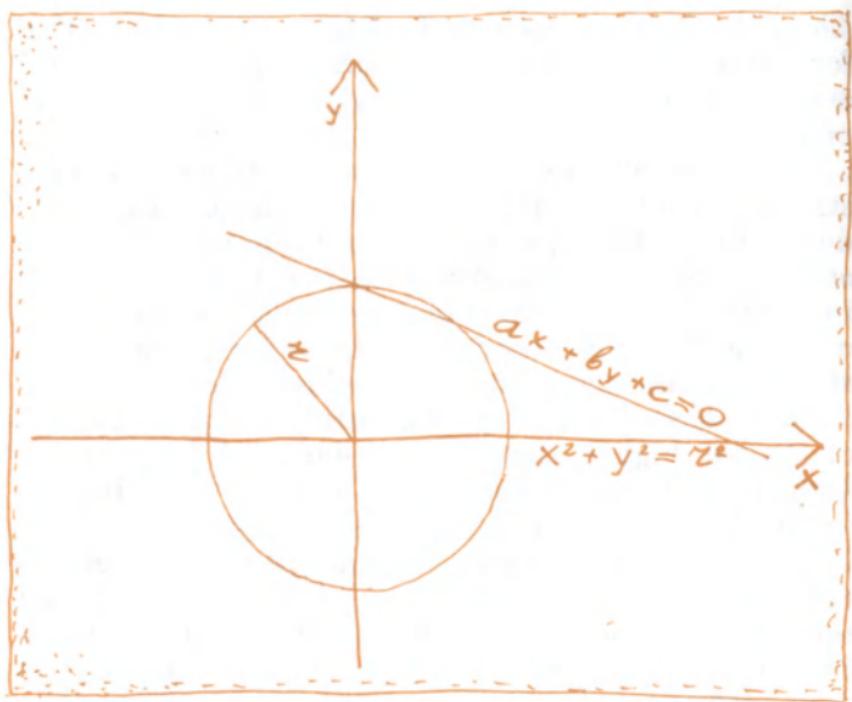


Fig. 5

end of the 17th century with its upsurge of astronomy, geodesy, mechanics and physics led the English genius of Isaac Newton and, independently, the great German scholar Gotfried Wilhelm Leibnitz to the setting up of the basic mathematical apparatus of classical physics—the differential and integral calculus, which in turn led to the development of the differential and integral equations of mathematical physics.

These new chapters of mathematics, united into a single section termed “Mathematical Analysis” or simply “Calculus”, carried physics and chemistry and their cognate fields to numberless victories, including the motion of machines, mechanisms, automobiles,

aircraft, and rockets, the development of electricity and radio, spectral analysis and weather forecasting. In a word, almost everything around us is indebted to calculus.

But what about geometry? At the beginning of the 19th century, after the triumph of analytical and differential geometry, it became possible to make a careful examination of the very foundations of geometry—its postulates.

This was undertaken by Nikolai Ivanovich Lobachevsky, the great Russian mathematician. He made a critical survey of the geometrical system of Euclid and excluded from Euclidean geometry the famous fifth postulate on parallel lines which reads as follows: "Only one line can be drawn parallel to a given line through a given point not on this line." In this postulate, Lobachevsky replaced the *assertion* of one line being parallel to a given line by the *supposition* that at least two parallel lines may be drawn through such a point.

Although such a supposition runs counter to our intuition, one must bear in mind that intuition rests on our observations, and what we observe is parallelism over extremely small portions of a plane.

What will occur if we assume the straight lines to be extended to infinity in both directions is therefore not at all obvious.

This theory, which became known as non-Euclidean geometry, or Lobachevskian geometry, was rejected by many of Lobachevsky's contemporaries. Later, however, it gave rise to other "non-Euclidean geometries" and—this is most important—it served as the mathematical basis for investigations, at the beginning of the 20th century, of actual physical space. These investigations culminated in Albert Einstein's celebrated theory of relativity.

At about the same time that Lobachevsky was working on his new theory, the Hungarian lieutenant Janos Bolyai was demonstrating the unprovability of Euclid's postulate of parallel lines and was constructing geometry on a new basis. Incidentally, the great Gauss, in a letter to Janos' father, wrote that he had already given thought to these problems and had laid the foundations of a non-Euclidean geometry but had not wanted to publish the results because of their extremely revolutionary and sensational nature.*

Many mathematical disciplines, which I will not discuss in this book, developed out of the requirements of mathematics itself but eventually proved to be extremely useful in physics, engineering and the natural sciences. An illustration is mathematical logic, which grew out of the necessity of constructing mathematics on a firm and consistent logical basis. Today mathematical logic serves as the foundation for constructing the theory of digital computers and, generally, is one of the most fundamental parts of the mathematical apparatus of cybernetics.

Further developments in algebraic theories and the establishment of profound relationships between algebra and mathematical analysis during the past three decades led to tremendous advances in what is known as functional analysis, which one of its founders, the Soviet mathematician I. Gelfand, described as the mathematical machinery of present-day physics.

Many more mathematical theories could be mentioned but there is no space; so let us take only a few and attempt to describe them in some detail.

* The exciting story of the birth of non-Euclidean geometry and the trials and tribulations of its creators is superbly told by V. Smilga in his *In the Search for Beauty* (Mir Publishers, Moscow, 1970).

Rubber-Sheet Mathematics

A good starting point is the familiar triangle. When studying any kind of objects, one tries to discover similarities and dissimilarities, distinguishing features.

What common features do the two triangles in Fig. 6 have? It would appear to be the sole fact that they are triangles, that is, they have three angles formed by straight-line segments. From this similarity there follow a good number of common properties: the sum of the interior angles in each is equal to two right angles; the area in each is expressed as half the product of any side by the corresponding altitude. The reader will probably be able to recall a good number of theorems dealing with triangles from his school math.

Now take Fig. 7. Do these figures have anything in common? Again, they are composed of straight-line segments, they have an odd number of vertices. And that's about all. How about the figures depicted in Fig. 8? There is some kind of resemblance of one to the other, but it is more difficult to state the properties they have in common.

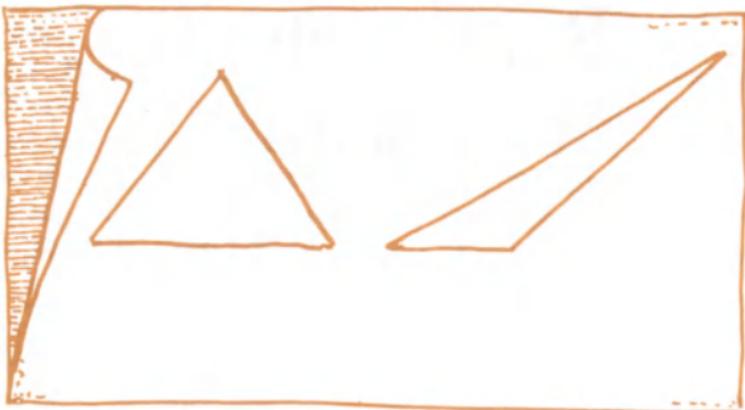


Fig. 6

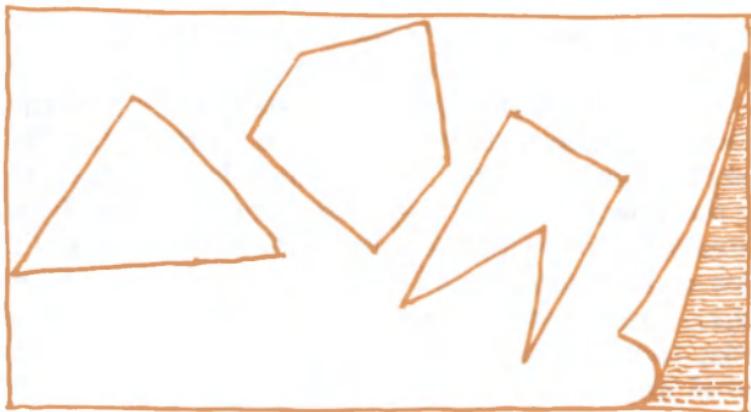


Fig. 7

Now let us return to the triangle. In Fig. 9 we clip off a similar triangle, that is, one having the same angles. Aside from the properties of all triangles, they have the additional property that they are similar. Now what does that mean?

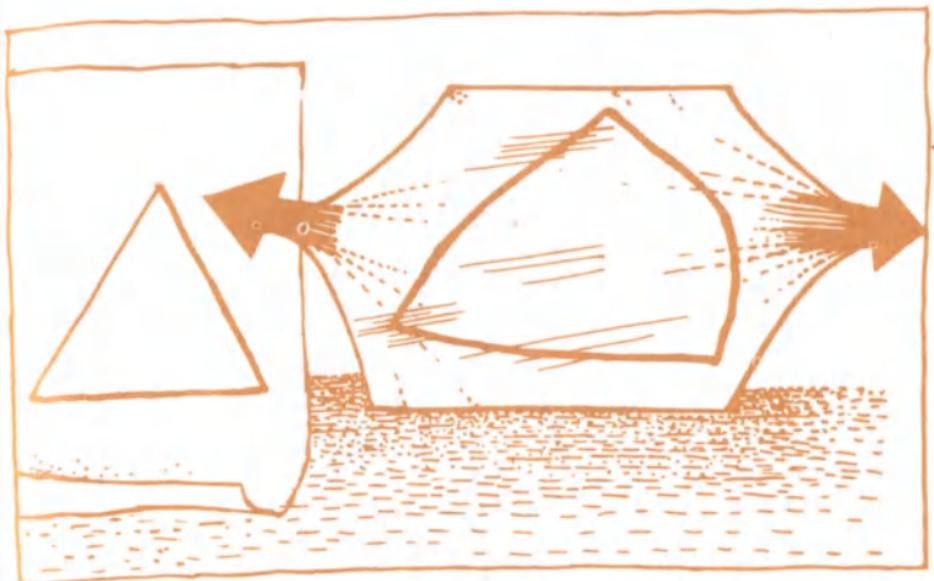


Fig. 8

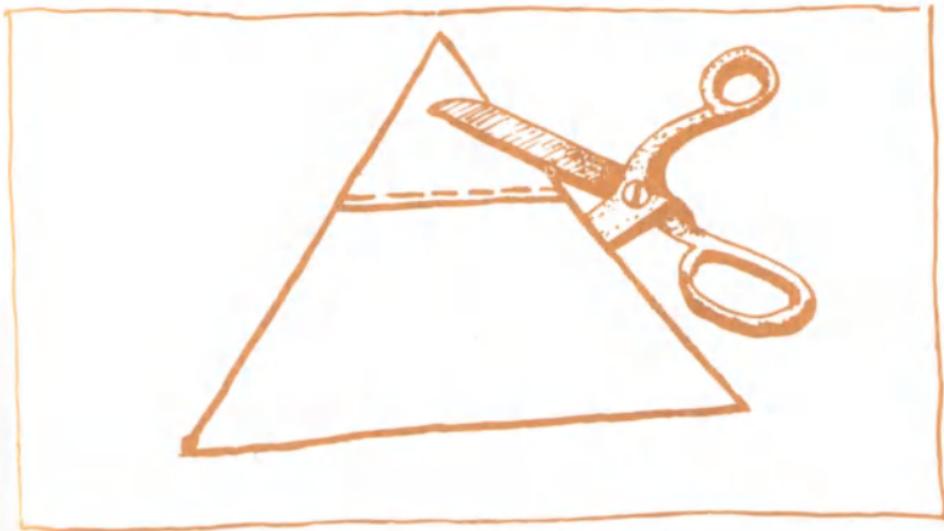


Fig. 9

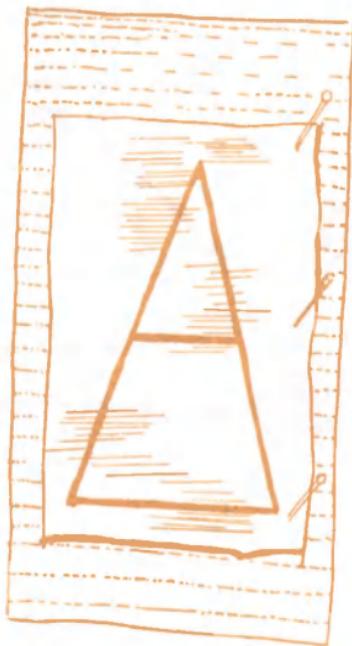


Fig. 10

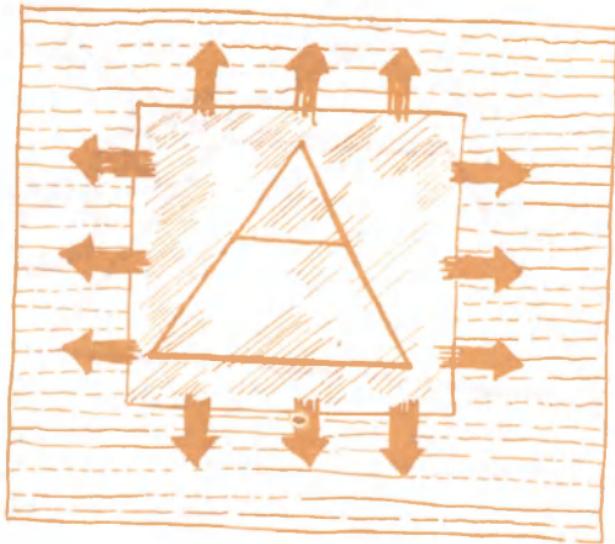


Fig. 11

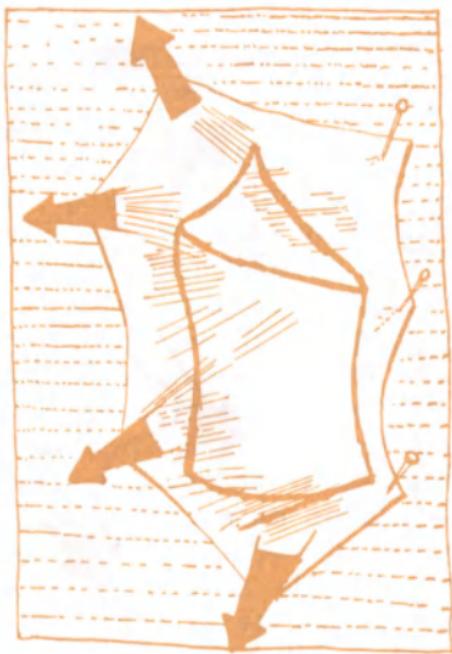


Fig. 12

Let us take a sheet of rubber and draw these similar triangles on it (Fig. 10). If the sheet of rubber is stretched lengthwise, the triangles will change but will remain similar (Fig. 11). (For the moment we can disregard the fact that when the rubber sheet is stretched lengthwise it becomes slightly narrower.) Thus, similarity is a property that is preserved under uniform stretching in some direction. However, if the sheet of rubber proves to be inhomogeneous or if the stretching process is not uniform, then the triangle may turn out to be something like we find in Fig. 12. The lines are no longer straight but there still remains something common between it and its predecessors. Let us try to figure out what this is.

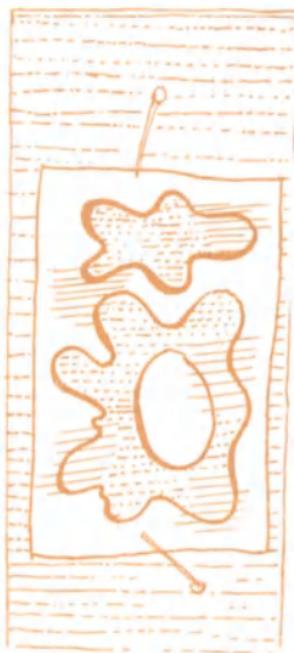


Fig. 13

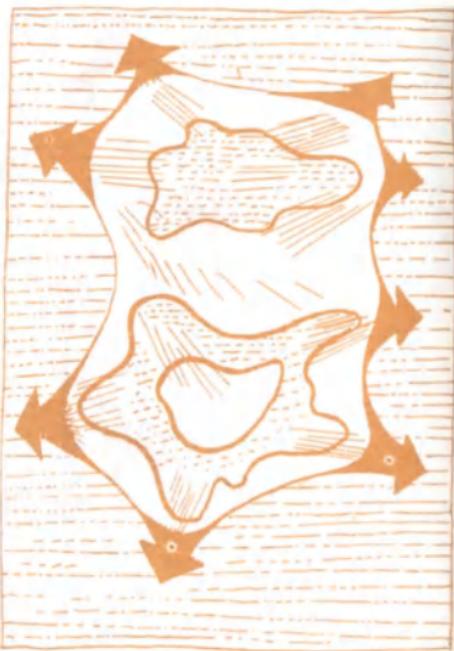


Fig. 14

The drawings in Fig. 12 are a sort of caricature of the neat triangles in Fig. 10, but they have vertices and the two triangles are separate and do not overlap. Now let us draw two amoeba-like figures on the sheet: one solid and the other with a hole in the middle (Fig. 13). Then stretch the sheet tight over a drum (Fig. 14). The amoebas will remain amoebas, and the hole will be preserved as well. No matter what kind of tension we apply without tearing the sheet, the hole remains.

We now have a good deal of observational material. What is there left in common to all these transformations of the rubber sheet?

MATHEMATICS AND ART

Like art, mathematics takes note of phenomena in real life, combines analogous events, processes and facts, and then generalizes.

People's Artist of the USSR, Obraztsov—the celebrated actor, painter and puppet-show man—shows us dogs and cats, and lions and rabbits and with their aid typifies (generalizes) certain humorous, pathetic or unpleasant qualities of human beings. Then Obraztsov replaces the puppets with ordinary balls attached to his fingers or just the fingers themselves. Using these very simple devices he is able to bring out a principal feature in the behaviour or character of a person and his relationships with other people. Here, art suggests an analogy and then says to the audience: fill in the rest by yourself.

With mathematics it is slightly different. After numerous and tiring observations, the mathematician discovers an important general property that describes a whole class of events. His work has just begun. He must formulate the properties he is interested in, then set up an appropriate theoretical scheme and make a thorough study of it; finally he must verify the correspondence of the newly constructed theory to reality.

CONTINUOUS TRANSFORMATIONS

The foregoing example showed us that under plane transformations, like those encountered in the arbitrary stretching of a rubber sheet, certain properties of the figures involved are preserved. The mathematician has a name for them. They are called *continuous* transformations. This means that very close lying points pass into close lying points and a line is translated into a line under these transformations. Quite



Fig. 15

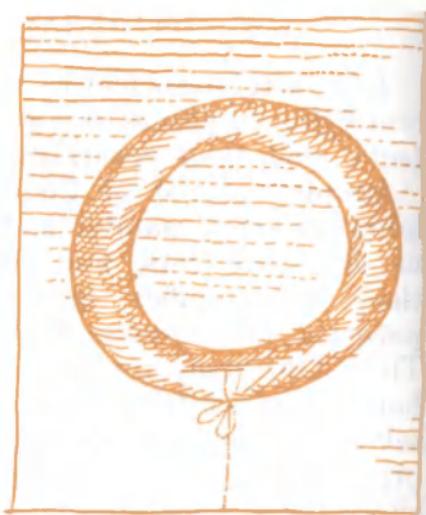


Fig. 16

obviously, then, two intersecting lines will continue to intersect under a continuous transformation, and nonintersecting lines will not intersect; also, a figure with a hole cannot translate into a figure without a hole or into one with two holes, for that would require some kind of tearing or gluing—a disruption of the continuity.

Such, in a word, are the starting principles of topology, a science that studies the properties of geometric figures that do not change under continuous transformations.

Is there any difference between a sphere and a doughnut (Figs. 15 and 16)? What is there in common between a cucumber and a ball (Figs. 17, 18)?

It is clear that if the cucumber is made of rubber, then it can be continuously deformed into a sphere but not into a doughnut. But a doughnut is nothing but a sphere with a handle (Fig. 19). It looks like a simple weight-lifting device.

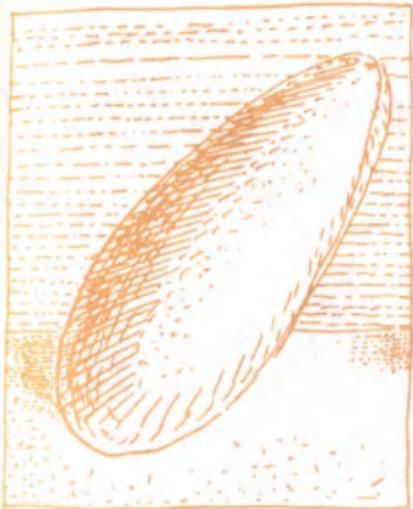


Fig. 17

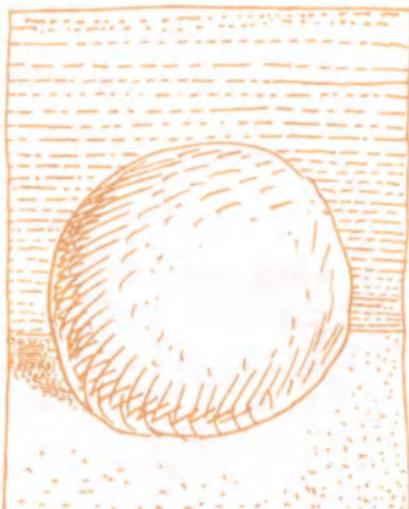


Fig. 18

Let us return to transformations in the plane. Draw a cat (Fig. 20) and then another one with a straight line through it (Fig. 21). Now if we compress the entire figure uniformly onto the line, we get a fat cat. Note that in the process all points of the figure, except those of the straight line, are displaced. The latter points stay fixed. Now (Fig. 22), take an arbitrary point O inside the figure and turn the cat about this point. Under this transformation only the point O will remain fixed while all other points will be translated to new positions. Let us now transform the cat by taking the point O as the centre of similarity. We will compress the figure (with different compression ratios) along various rays passing through O . Fig. 23 depicts such a transformation about another point (also designated by O) with a compression ratio of

$$K = \frac{1}{2 + \cos \varphi}$$



Fig. 19

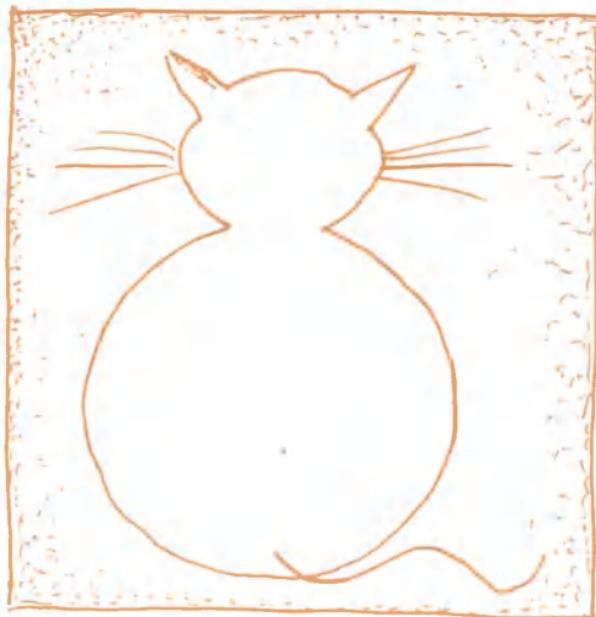


Fig. 20

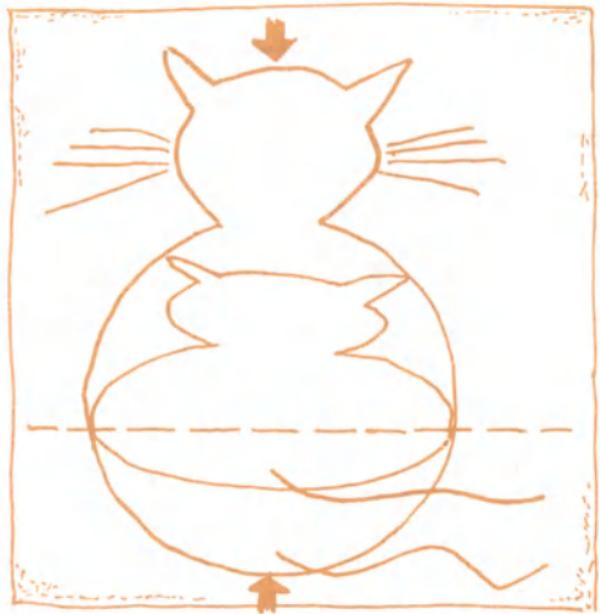


Fig. 21

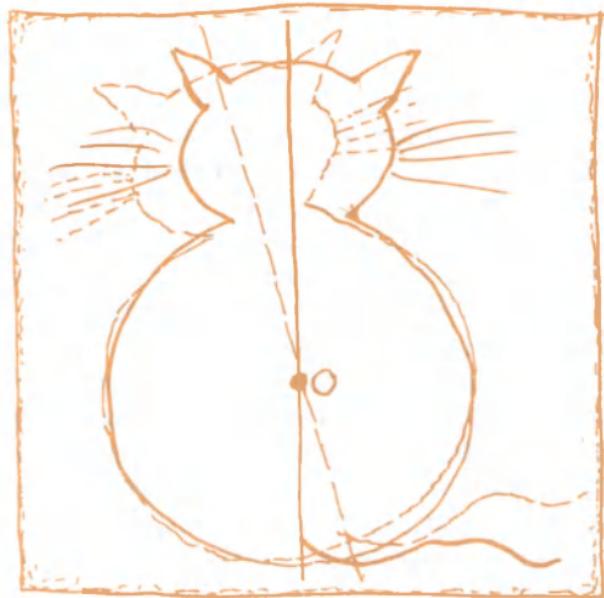


Fig. 22



Fig. 23

where φ is the angle between the direction of the appropriate ray and the horizontal straight line. The cat is distorted: all points have been displaced, with the sole exception of point O , which remains fixed.

Now (Fig. 24) let us displace the entire figure of the distorted cat in parallel fashion so that it remains inside the original figure.

These two consecutive transformations—nonuniform compression and parallel translation—may be regarded as a single transformation of the cat into itself.

The reader can now ask himself whether under such a transformation at least one point remains fixed or whether all points of the figure have occupied new positions.

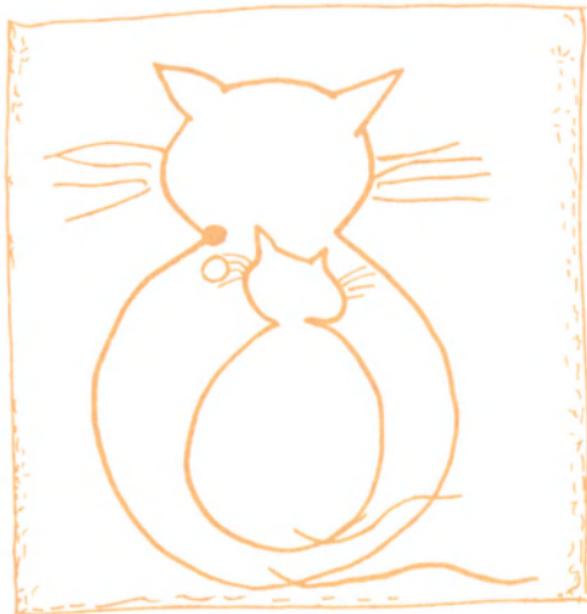


Fig. 24

Let us now take the rubber sheet and stretch it in different directions differently (at random, so to say), and then let us draw the same cat on the stretched sheet of rubber. Now release the sheet and let it take up its original normal position. The cat will contract and find itself inside the original drawing as the fantastic distortion shown in Fig. 25.

I think it will now be clear to the reader that in this complex transformation, all points of the original cat have taken up fresh positions and not a single one has remained fixed. At any rate, I have asked this question of many people and they have all stood firm in that opinion.

However, our intuition has played a trick on us: the assertion that all points of the figure have been

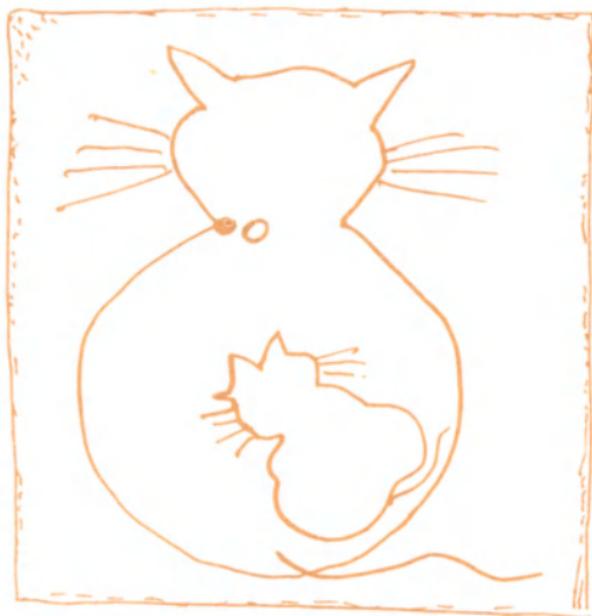


Fig. 25

translated to new positions is erroneous. In reality, just the contrary is true: *under any continuous point transformation of such a figure within itself, at least one point remains fixed.**

This famous Borel-Brouwer theorem, called the fixed-point theorem, was obtained at the start of the twentieth century and plays an important role in many problems of topology and mathematical analysis, particularly in the study of motions of dynamic systems.

* We are dealing here with figures that can be obtained via a continuous transformation of a circle. A similar theorem holds true for spatial figures (solids) that can be obtained by a continuous transformation of a sphere.

A REMARKABLE SURFACE

Fig. 26 depicts the emblem of the mechanics-mathematical department of the Moscow State University. It consists of a coordinate grid, an integral sign and a twisted strip which is called the Moebius strip.

To see what this strip is all about, take a strip of paper and paste the ends together to get a cylinder. We can draw horizontal lines on the outer side and vertical lines on the inner side (Fig. 27). Now do a mental experiment.

Put an ordinary ant on the outer surface and let it crawl round the cylinder but do not let it over the edge. Suppose it crawls along the middle line. After a time it will return to the starting point (much like Magellan's ships during their circumnavigation of the globe).

A roof, hat or automobile tyre have both an inner and an outer surface, as we know. From such observations it is easy to conclude that every surface must also have an outer and an inner side. Indeed, how could it be otherwise?

Now let us twist the same strip (call it $ABCD$) half over and then glue the ends together: point A



Fig. 26

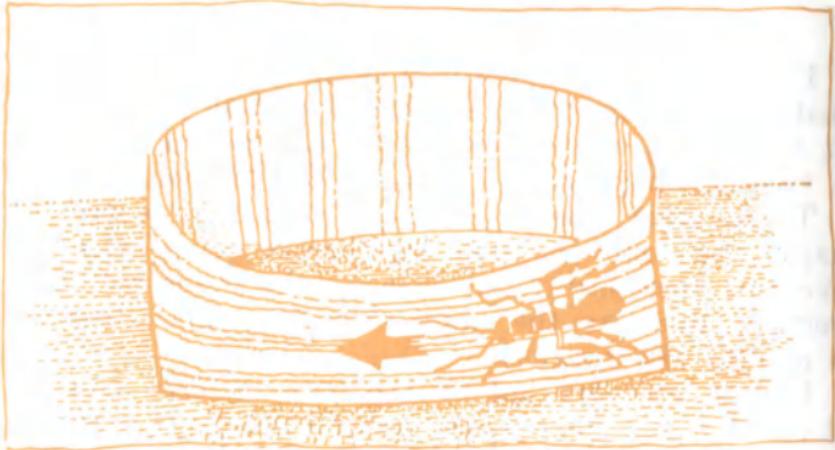


Fig. 27

to point D , B to C (as in Figs. 28-30). Now if the ant makes the same trip around the horizontal line you will be surprised to see it arrive at the starting point upside down!

If we tried to paint the sides of this surface in two different colours, we would get nowhere. It can't be done, and all because of the fact that this new surface has only one side! Our firm assertion has turned out to be erroneous.

This new figure is the famous Moebius strip, named after Moebius who discovered it in 1858. The Moebius strip has some other remarkable properties as well.

A cylinder, it will be noted, has two edges, upper and lower; now the Moebius strip has only one edge.

If one cuts a cylinder (Fig. 27) along the middle line that the ant travelled, he will get two cylinders. If the Moebius strip is so cut, we might get one of the following five results:

- (1) two Moebius strips;
- (2) two cylinders;

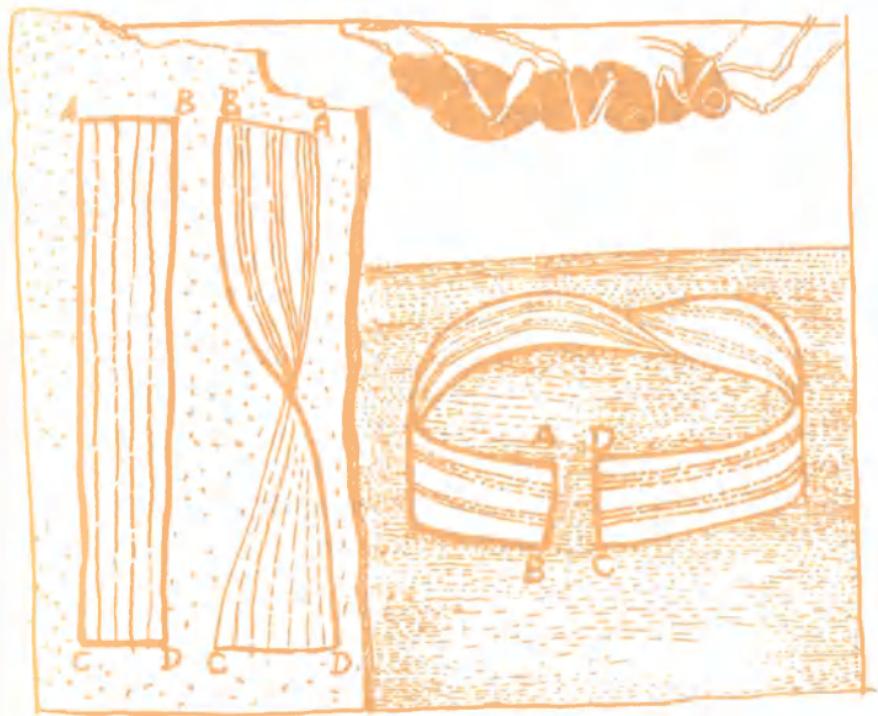


Fig. 28

Fig. 29

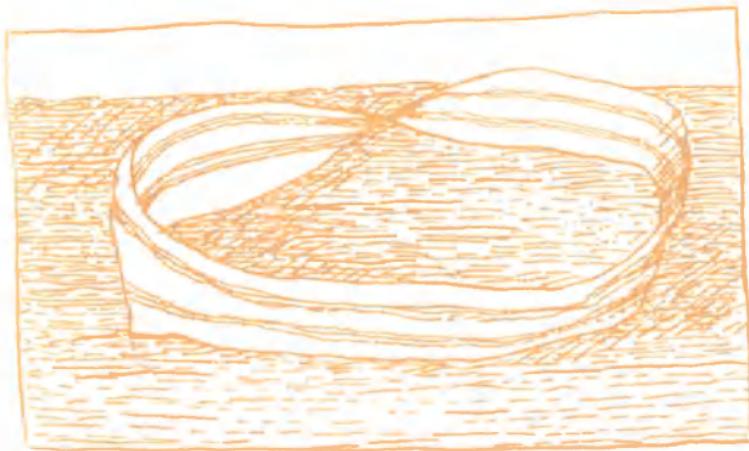


Fig. 30

- (3) one cylinder;
- (4) one new Moebius strip;
- (5) two linked loops.

The reader is invited to take his pick or suggest a possible fresh result. Now glue up the Moebius strip (that'll take a couple of minutes, but don't grudge the time). Now cut it down the middle line and see whether you get the figures you expected. Then cut the resulting strips along the middle lines once again. I don't think you'll get what you expected.

To summarize, we can say that the wealth of geometrical images was not exhausted by the ancient Greek geometers and it is not confined to polygons, cones and pyramids. The wealth of geometrical entities is unlimited— infinite—and continues to be studied with great intensity to this very day.

What is more, the apparently firm assertion that every surface has two sides proved to be wrong. This can mean but one thing; when a mathematician demands rigorously logical proof about any assertion, he does so not for his own pleasure but to verify the facts, which might easily appear to us to be obvious but which, when verified, prove to be erroneous.

GRAPHS

A railroad map of the country or a plan of the streets of a city represents a network of lines (see Fig. 31), each line segment connecting two points called vertices. Such a network of points and connecting lines is termed a *graph*. (Recall that the same term is used to designate the curve of a function.)

The network of water mains in a city is also a graph, but is essentially different from the street network because the water flows only in one direction. If the edges (lines) of a graph are arrows indicating the

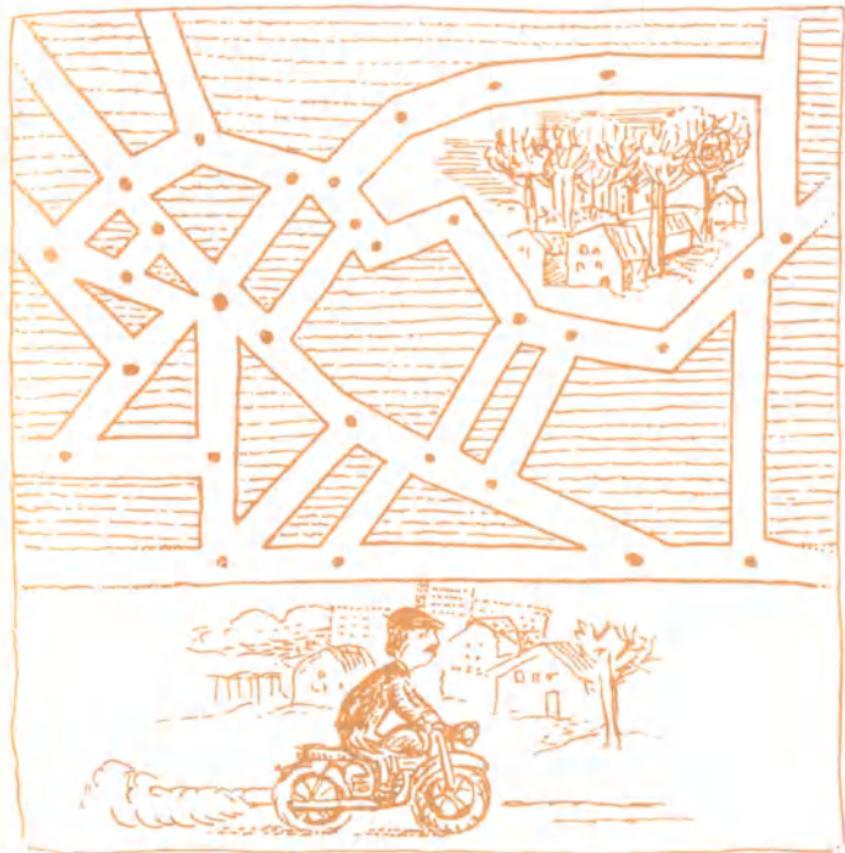


Fig. 31

direction of flow of water, then we have a *directed*, or *oriented*, *graph* (Fig. 32). Incidentally, some streets have one-way traffic, others two-way. If arrows are used to indicate one-way streets and an absence of arrows is taken to mean two-way streets, then we get a graph which is called *mixed* (Fig. 33).

A chess tournament can also be represented in the form of a graph. Draw circles on a sheet of paper to indicate the participants and equip each one with

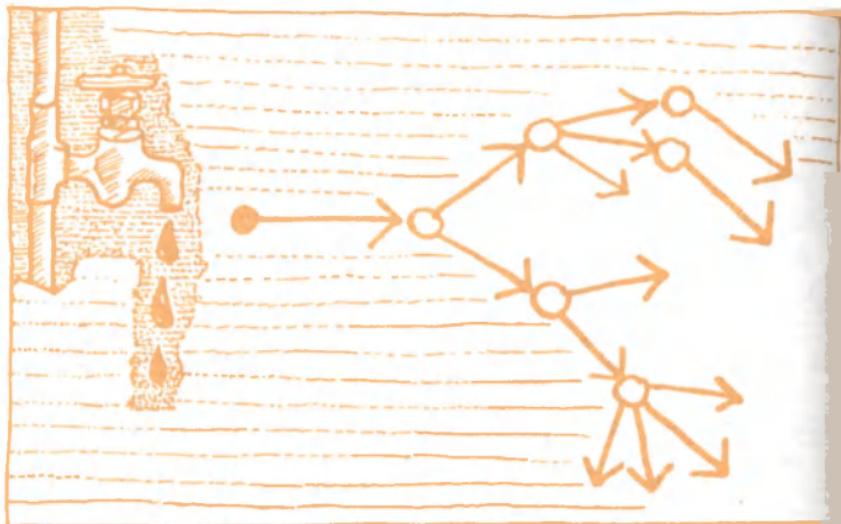


Fig. 32

a number in accordance with the lots drawn. Then the result of a game involving each pair will be represented as an edge joining the two appropriate points, and the direction of the arrow will be from the winning side to the losing side. No arrow on the edge will mean that the sides drew even (Fig. 34). The tournament will be over when each circle is connected with all the remaining circles. Such a graph is termed a complete graph. First place will have the largest number of arrows issuing from that circle. If each two players play twice (white and black), then two edges will be required. In Fig. 34 we have a situation in which all participants, except the fourth and sixth, have played two games each; the fourth and sixth have played one game each, and the leader is number two.

If there is any doubt, in a graph, about the intersection points of the edges not indicated by circles being meaningful, we can simply imagine the graph

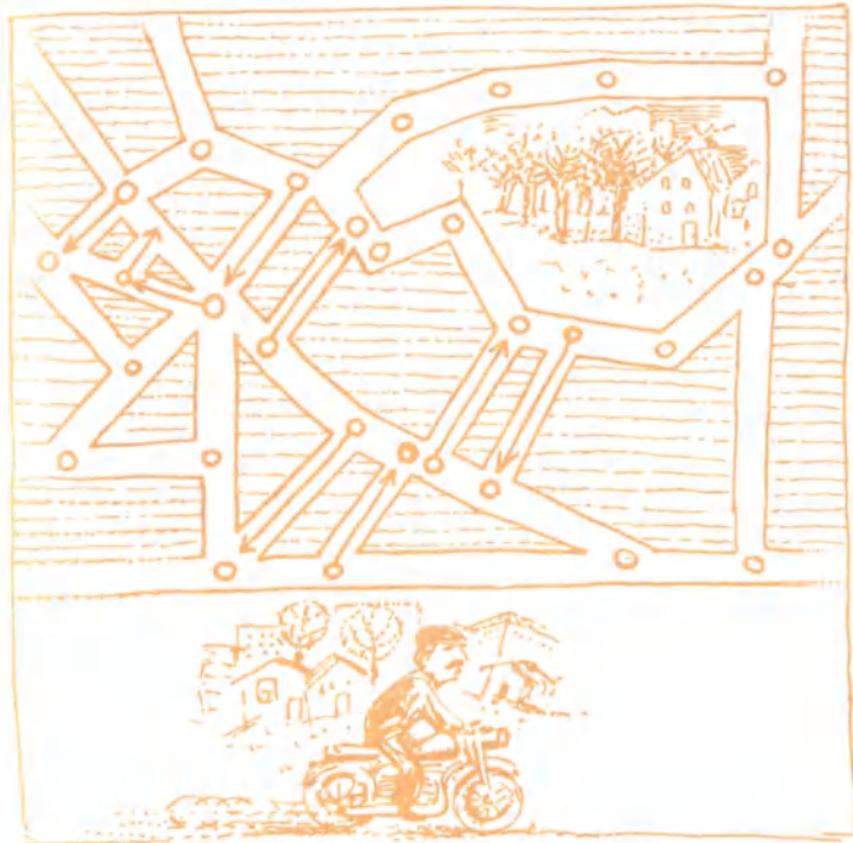


Fig. 33

as lying in space and then the edges will be like ropes with no intersections. Note please that the edges of a graph need not be straight line segments at all. The graphs in Figs. 34 and 35 are the same in the sense that one can be translated into the other by a continuous transformation. Mathematicians then say that such graphs are *isomorphic*.

Incidentally, it is not always immaterial whether the graph can be drawn so that the edges do not inter-

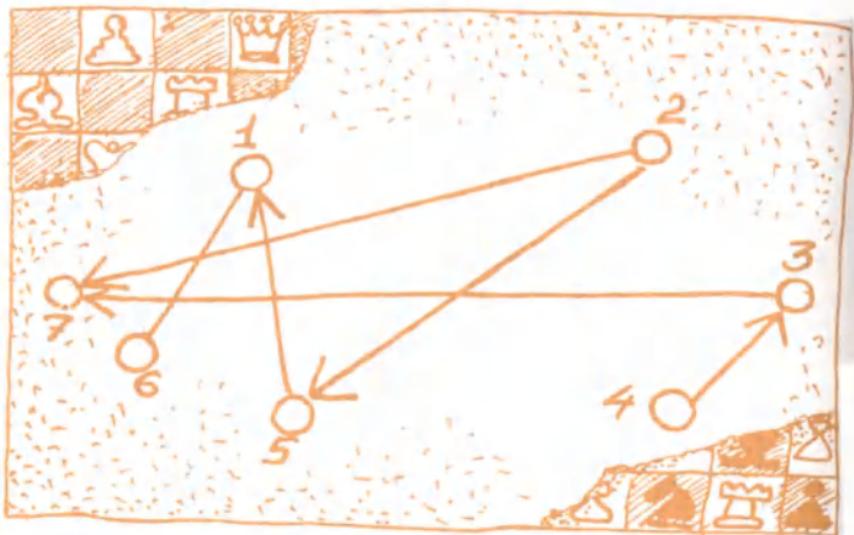


Fig. 34

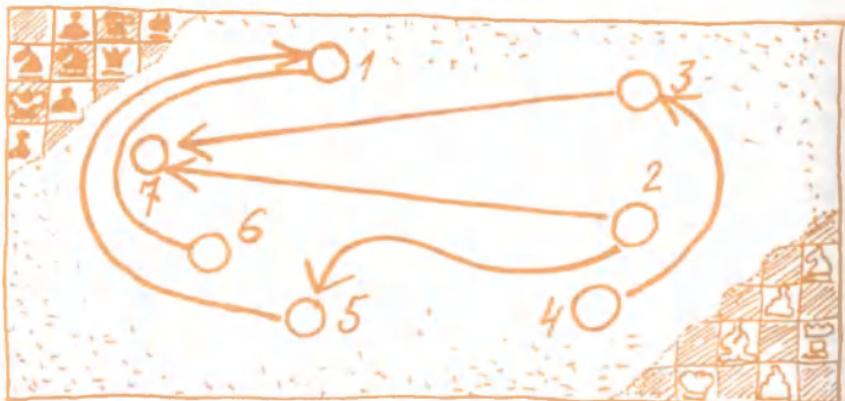


Fig. 35

sect. For instance, a radio-circuit drawing is a graph whose vertices are components such as resistances, capacitances, electron tubes, etc., while the edges are the connecting wires. It is not important that the

edges intersect or do not intersect in the drawing because in the actual circuit the wires will not intersect, and a short circuit is prevented by insulating the wires.

However, in recent years printed circuits have come into vogue. A printed circuit is a sheet of dielectric with a metallic film deposited that corresponds to the wiring diagram. In this case it is important that the vertices of the graph (circuit) be connected by nonintersecting lines, otherwise there will be a short circuit.

In other words, there are cases when a given graph must be represented in the plane so that the edges intersect only at the vertices. If this is possible, the graph is called a *plane graph*.

There is a method which enables one to check to see whether the graph is plane or not. (So you see this is a problem of practical significance.)

In setting up one-way streets, the city authorities have the problem of choosing directions of traffic on the various streets so that there will be no blind alleys or inaccessible points. For instance, in Fig. 36 we can drive from *A* to *B* but not from *B* to *A*. Then of course we must be able to drive (technically called choosing an orientation) from any point on a street to any other point without breaking any traffic rules. This may be stated as an exact problem concerning the structure of an oriented plane graph.

Incidentally, if the traffic rules were set up in accordance with this theorem, the militia would suffer from scathing criticism at the hands of drivers and taxi users.

Actually, the problem of traffic and its organization in a large city is extremely difficult and is becoming more so with the increasing number of motor cars on the streets. Yet in the final analysis, this is a mathematical problem which is closely related to the theory of graphs, although it involves more than graphs.

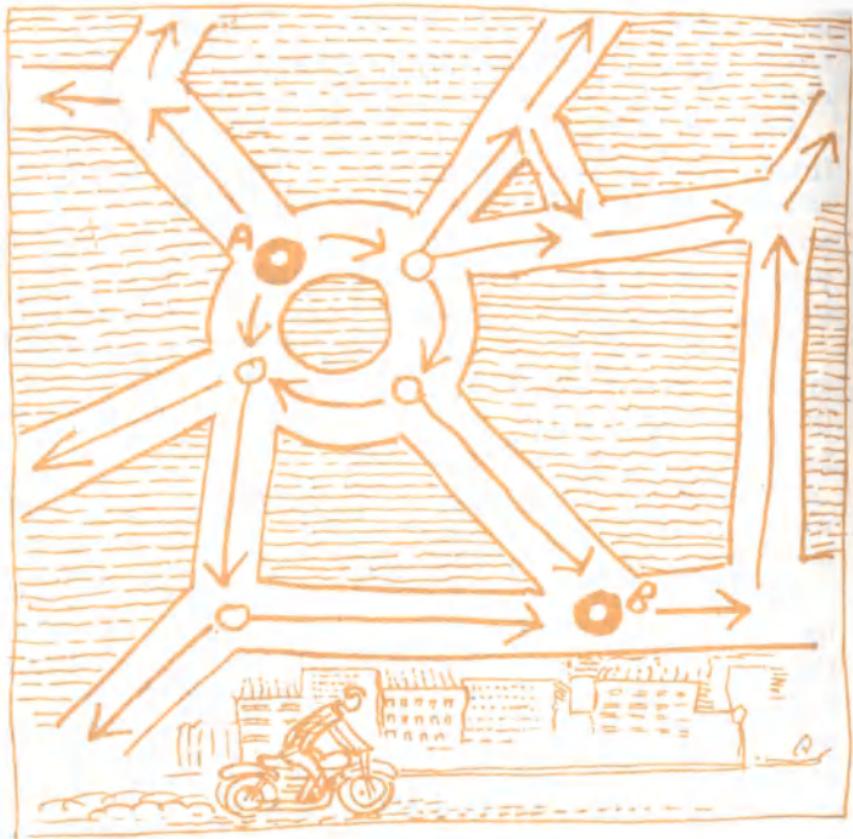


Fig. 36

The theory of graphs, together with certain other modern disciplines, now enables us to solve problems involving the planning of transportation, control of supplies of goods and reserves, etc. Let me illustrate by means of what is called a transportation problem.

Let us take a new town with a large number of construction sites. Under construction at one time are a school, a naval institute, a sixteen-story block of flats, and port facilities. Let us now imagine three brick factories supplying the sites with building

materials (further note that the sites are at different distances from the factories). We can now picture the chief of supply mulling over the problem of how to plan deliveries of bricks to the various sites. The aim is to satisfy the demands of the construction sites and at the same time minimize the very considerable costs of transportation between the factories and the sites. This problem can be solved but it requires some imagination and knowledge and a modern electronic computer as well, if of course the number of construction sites is large. I will give a brief outline of a solution to the problem.

Let us set up an oriented graph where the three brick factories are denoted by 1, 2 and 3, the construction sites by S (school), I (institute), H (house), P (port); the edges of the graph are drawn from the factories to all construction sites (Fig. 37). The numbers on the edges indicate the relative cost of transporting one thousand bricks over the indicated route.

The solution now appears to be obvious: the school will receive bricks from Factory No. 1, the port from No. 2, the institute can arrange for equal shares from No. 1 and No. 2, and the house construction site, from factories No. 2 and No. 3. It would appear that Factory No. 3 could be closed and this would minimize the transportation costs.

Actually, the matter is a bit more involved. Imagine that factories 1, 2 and 3 have different outputs, and the overall output barely meets the demands of construction. Suppose No. 3 has the highest output and No. 2 the lowest. These limitations then substantially complicate the problem. But if we run through all variants, the problem can be solved and the results enable us not only to ensure a regular supply of bricks to all sites but also to minimize the very high costs of transportation.

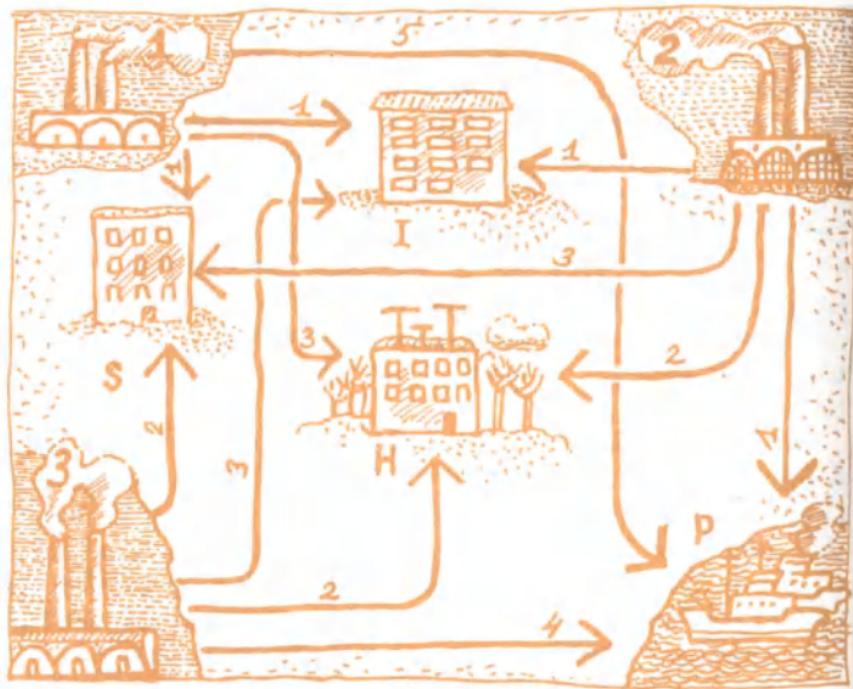


Fig. 37

It may be noted that if the overall output of brick does not markedly exceed the demands of construction, then ordinary rule-of-thumb planning of deliveries usually results in irregularities in deliveries. Planning via the optimum variant in deliveries of sand from Moscow's river ports to construction sites within the city has resulted in a tremendous annual saving of money.

Here are some other problems that involve the theory of graphs. We can begin with the exciting problem (for girls) of getting married. Suppose there are m boys in a village and n girls. The girls are very particular about their choices, and each regards only

a few of the boys as suitable for marriage, the others as unacceptable. How can we arrange the marriages so that each beauty gets an acceptable husband?

Unfortunately, the number of girls must not exceed the number of eligible young men ($n \leq m$), and this obvious condition greatly complicates life. But that is not all. If there were five girls and all five regarded as acceptable only the first two boys, the problem would have no solution. But suppose there are fewer girls than boys and their tastes and requirements are varied (or suppose they are quite reasonable and do not demand the impossible), and there are no basic objections to all girls getting married.

The situation is illustrated in Fig. 38 where the arrows coming from the girls indicate possible fiances. We ought to give all the young men nice names but it is easier simply to number them. In our situation we can make all the girls happy. If the first number stands for a girl and the second for a boy, then we can form five pairs (1,1), (2,2), (3,5), (4,3), and (5,6). And, as so often happens in real life, the fourth man with more girls interested in him than in any one else remained a bachelor. We could of course marry them off in other combinations. What is more, we failed to take into account the interests of the young men and a slew of other ordinary complications like jealousy, vanity and the like which so often spoil one's mood and even one's whole life.

When the number of girls and boys is very great and their interests are intricately interwoven, the problem, as you can well imagine, is far from simple. But it is possible to indicate the general conditions that ensure existence of a solution. I will not tire you with a statement of the appropriate theorem and will suggest a different, somewhat less dramatic, model of the same situation.

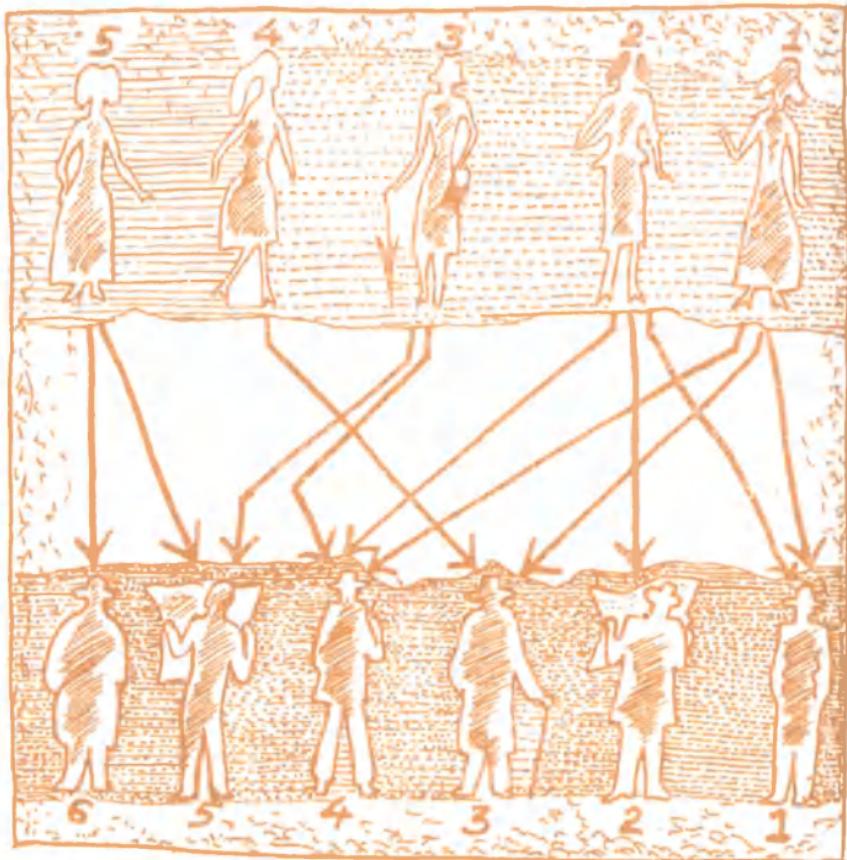


Fig. 38

Imagine a department in a factory with n distinct machine tools and m workers ($n < m$), and the qualification of the workers such that each machine can be tended by only a certain number of them. Under what conditions is it possible to ensure operation of all machines? The solution will be seen to be equivalent to the problem of pairing satisfactory marriage couples.

The so-called assignment problem is a modification of this situation. Suppose we have n officials and the same number of assignments; each one can perform

any one of the jobs (assignments), but the efficiency in each case differs. Denote by a_{ij} the efficiency (in certain units) of worker i performing job j so that, say, a_{24} signifies the efficiency with which worker No. 2 does job No. 4. The important thing is to distribute the machine operators so as to achieve the highest productivity. This situation is illustrated in Fig. 39. To gauge the efficiency of the entire set-up, we can take, say, the sum of the efficiencies. Then the situation shown in Fig. 39 yields an efficiency of

$$a_{12} + a_{24} + a_{31} + a_{43}$$

Now we can pose the problem of choosing the most effective distribution of work among the operators. This can be solved by considering all ways of distributing the work among the operators and then choosing the one that ensures maximum effectiveness. It is not the fastest solution but it yields the proper result.

We could also evaluate the overall efficiency by measuring the lowest efficiency. This would be the case when we are interested in utilizing even the weakest worker in the best possible manner. Then the problem of distributing jobs among the operators is formulated thus: distribute jobs among available workers so that the least efficient work is a maximum. Let us see what this gives us in terms of the graph depicted in Fig. 39. Suppose that in this graph a_{31} is the least of the efficiencies. We could change the job distribution and efficiencies from $a_{12}, a_{24}, a_{31}, a_{43}$ to, say, $a_{11}, a_{23}, a_{32}, a_{44}$. If a_{44} is the least of these four numbers and a_{44} exceeds a_{31} , then the second distribution of jobs is preferable to the first. We run through all possible job distributions and choose the arrow arrangement in which the smallest of the numbers (efficiencies) in the group of arrows is the largest possible.

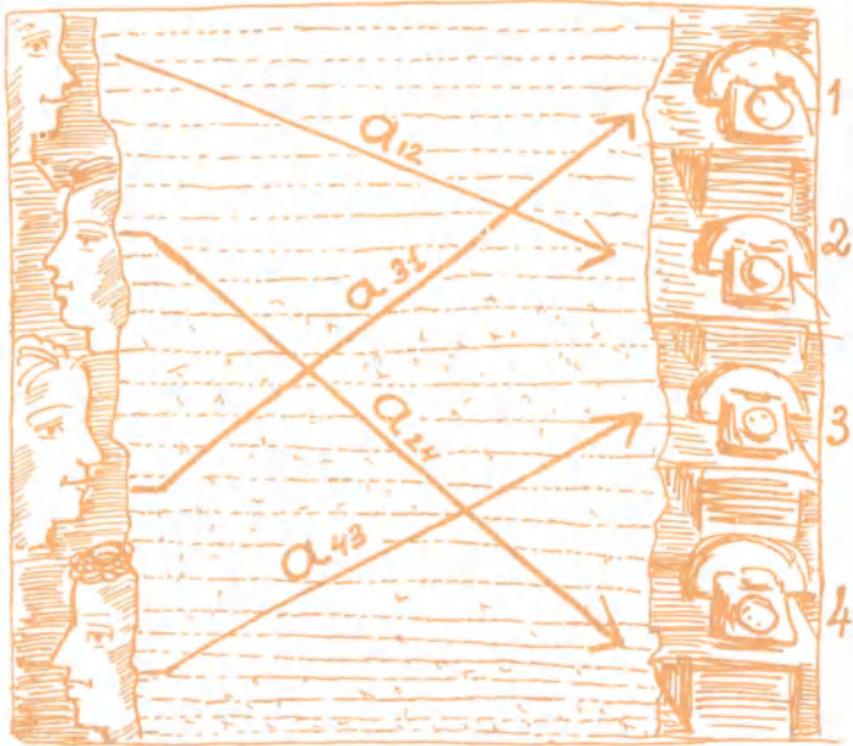


Fig. 39

Today the theory of graphs is extensively employed in different branches of science and technology, in particular, in a system called grid planning (in the United States it goes by the name PERT). This is a very exciting topic that I had wanted to spell out for the reader using such examples as making dinner, repairing a flat and presenting a graduation thesis, but so much has been written on the subject that I think the space and time will be better spent if I introduce some fresh topics.

Numbers and Points

For what follows we will need a few elementary facts from analytic geometry. If you have forgotten some of your school geometry, these will be welcome as a sort of refresher course, otherwise you can simply skip this chapter altogether. It is really very elementary.

You will most likely have noticed that on highways the distances between towns are often indicated by posts. This is a method of indicating the position of a point on a line (not necessarily a straight line) by means of numbers (Fig. 40).

We have already mentioned a method for indicating points in a plane by means of Cartesian rectangular coordinates.

In this very same way it is possible to specify points on any kind of surface by means of numbers. When the legendary Captain Nemo of the *Nautilus* had to fix his position on the surface he computed longitude and latitude.

Let us take a sheet of rubber and draw a grid of Cartesian coordinates, the mesh being of unit length. In order to get from one point of the grid to another by moving along the lines of the grid (Fig. 41), one



Fig. 40

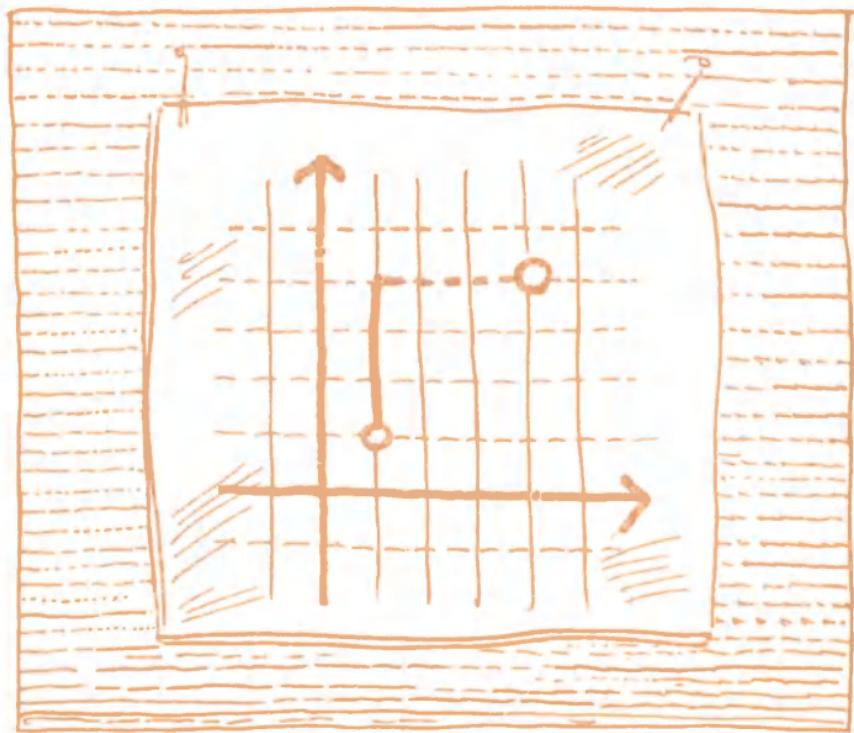


Fig. 41

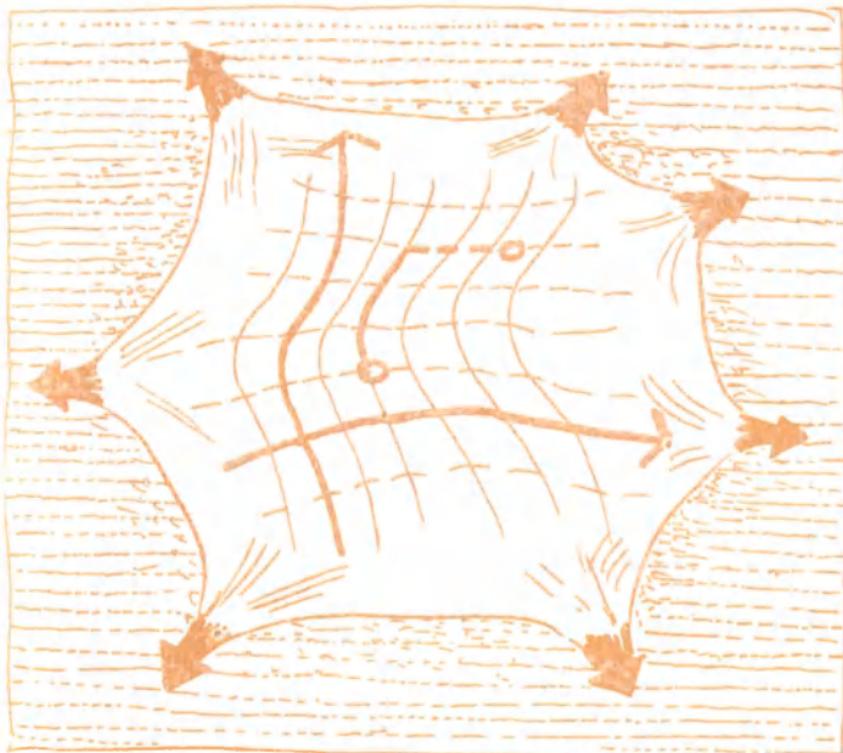


Fig. 42

has to move down a solid-line "street" and then along a dashed line. Of course, you get there the other way round too: first down a dashed street and then a solid-line street.

Now let us deform the rubber sheet by means of an arbitrary continuous transformation. The resulting curvilinear grid will also be a system of coordinates for travelling from one point of the grid to another: again travel down the solid street (which is no longer straight-line) and then along the dashed-line street (Fig. 42).

The situation is the same in space. To indicate the position of a hanging lamp, specify three numbers:

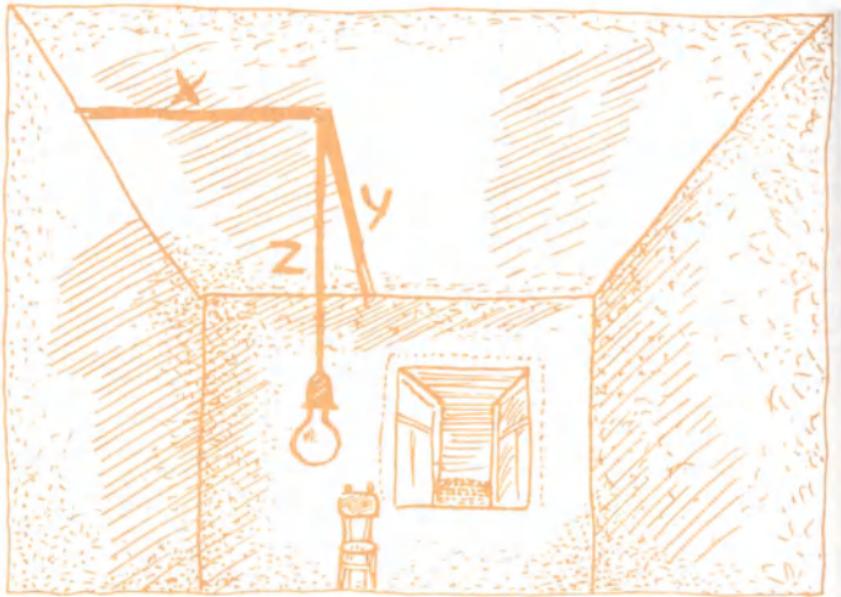


Fig. 43

say, the distance from two perpendicular walls to the point where the cord hangs from the ceiling, and the length of the cord (Fig. 43).

These are Cartesian rectangular coordinates in space.

If the captain of Jules Verne's *Nautilus* wanted his position in space, he took readings of the longitude, latitude and also measured the depth of submersion. These three numbers gave him the needed coordinates in space.

In astronomy, the positions of celestial bodies relative to the earth are defined by three coordinates: two angles (declination and right ascension) and the distance from the earth.

The method of coordinates enables us to describe any geometrical problem in terms of numbers. Geometrical images turn out to be equivalent to a definite set of numbers.

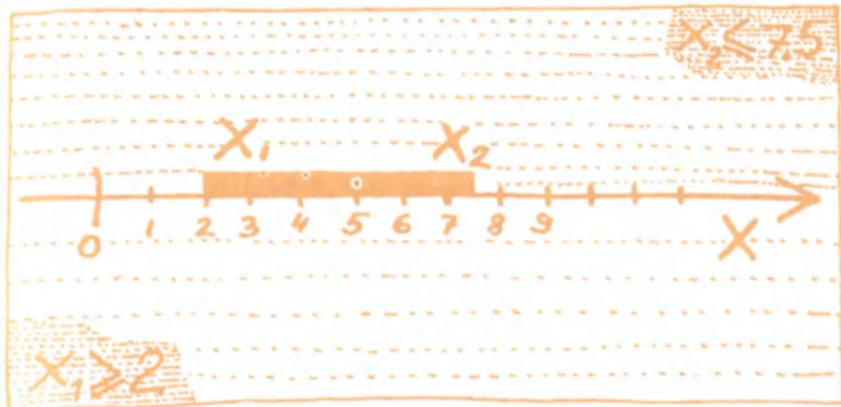


Fig. 44

For instance, the line segment between points with coordinates $x_1 = 2$ and $x_2 = 7.5$ (Fig. 44) consists of the set of all numbers x that satisfy the two inequalities

$$x \geq 2 \text{ and } x \leq 7.5$$

These two inequalities can be more compactly written thus:

$$2 \leq x \leq 7.5$$

A unit square in the plane, the vertices of the square being points with coordinates $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, is a set of number pairs (x, y) that satisfy the inequalities (Fig. 45)

$$0 \leq x \leq 1, \text{ and } 0 \leq y \leq 1$$

Thus, using the method of coordinates it is possible to present the whole of geometry analytically, beginning with the definition of a point on a straight line as the number x , a point in the plane as the number pair (x, y) , and a point in space as the number triple

(x, y, z) . A circle of radius 5 centred at the point $(2, 3)$ is merely the set of all number pairs (x, y) that satisfy the inequality

$$(x - 2)^2 + (y - 3)^2 \leqslant 5^2$$

A plane in space passing through the origin of coordinates is the set of all number triples (x, y, z) that satisfy the equation

$$ax + by + cz = 0$$

where a , b and c are given numbers.

The important thing is to note that the geometric and analytic approaches are equivalent: geometrical images may be expressed analytically in the form of

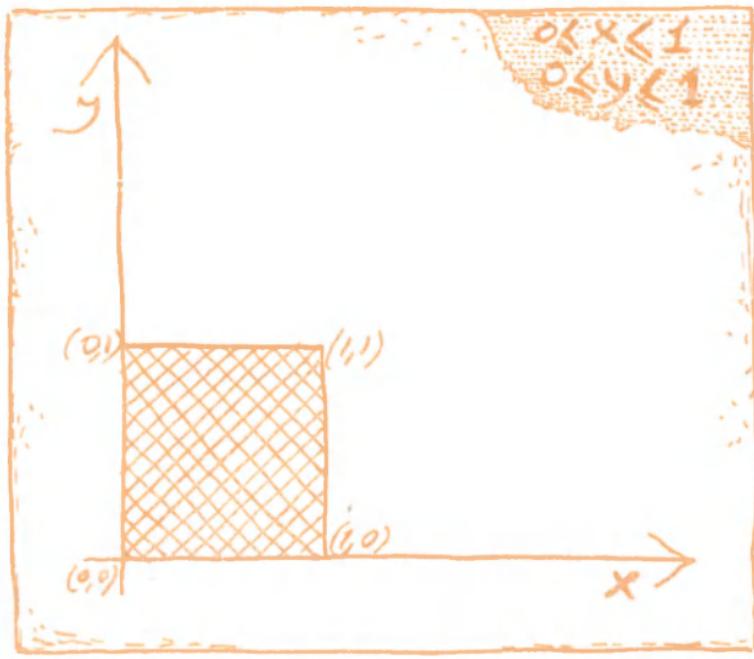


Fig. 45

equations or inequalities, and analytical relations may be represented in the form of curves, surfaces or figures.

The analytical approach to geometrical problems enables a doctor to visualize various characteristics of the human body. For example, we can lay off the height of a person on a straight line.

The height (h) and weight (p) of a person are indicated by a point (in a plane) with coordinates (h, p). Now if we also indicate age (t), then he will correspond to a point in space with coordinates (h, p, t).

Now what happens if a person is described by many parameters, say, height (h), weight (p), age (t), chest expansion (Q), strength of left hand (f_1) and right hand (f_2), vision (r)? Here we have seven parameters, and it would seem impossible to visualize them geometrically.

Actually, however, geometrical analogies are very widely used, and precisely for this reason the set of all possible number quadruples (x, y, z, t) may be regarded as a set of points in four-dimensional space; the set of all possible number septuples (x, y, z, t, u, v, w) as a set of points in seven-dimensional space. Finally, we can regard all sets of n numbers ($x_1, x_2, x_3, \dots, x_n$) as a set of points in n -dimensional space.

When a person first encounters the concept of four-dimensional space, he is usually at a loss as to how he is to visualize such a thing. Indeed, how do we picture four dimensions?

Let us take a narrow glass tube and put an ant in. If he wants to get out, he will have to move backwards. If two ants are let in at both ends, they will not be able to pass each other (Fig. 46). Such is the unhappy state of affairs in a space of one dimension (one line).

Now let the ants out onto the surface of a table or a pumpkin, and they will be able to move in any



Fig. 46

direction and skirt any obstacle (Fig. 47). Life on a surface (in a space of two dimensions) is much easier.

True, it too has its difficulties. If the ants are separated by a rivulet, they will never meet. They say that if you draw a white circle around a rooster he will be nonplussed to find himself inside the circle and will not have sense enough to step outside. Actually, what he needs is just a little common sense and enough courage to move out of two-dimensional space into three-dimensional space.

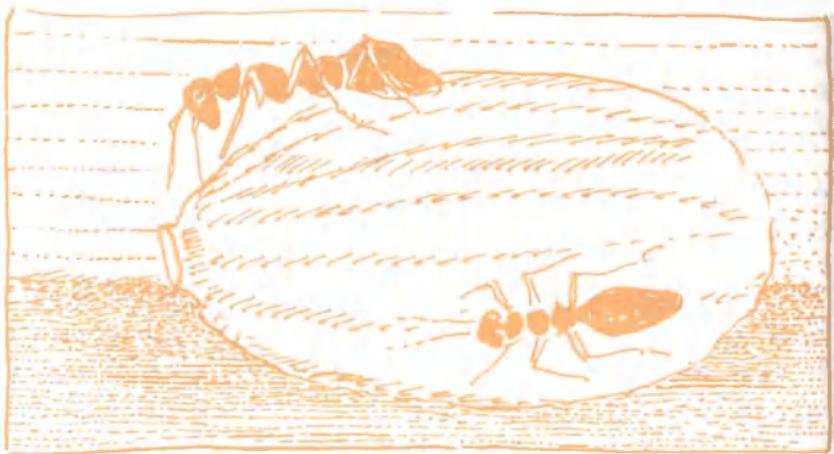


Fig. 47



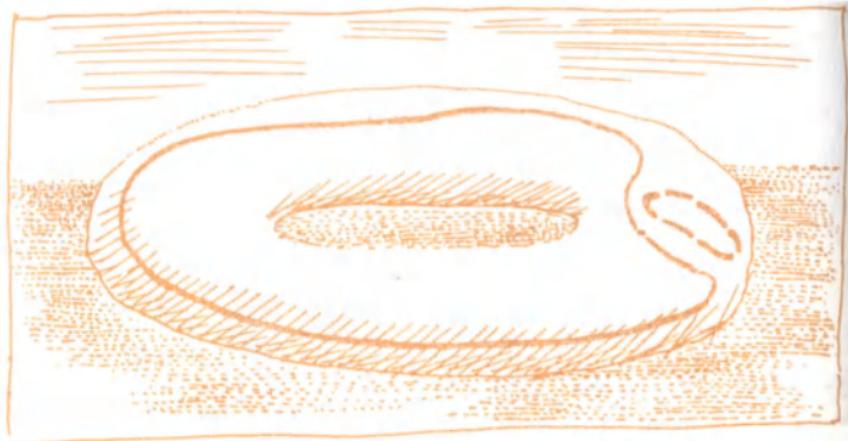


Fig. 48

A dragon-fly has it better than the ant because it can fly over a river. Dragon-flies live in three-dimensional space, and a closed line on a surface does not hamper their movements. But if we put the dragon-fly in a jar and put on the lid, it will be held prisoner. A closed surface divides its living space of three dimensions into two parts, an inner and an outer part, just as a closed curve divides into two parts (inner and outer) the living space (surface) of the ant.

Incidentally, not every closed curve drawn on a surface divides the surface into two parts so that it is impossible to get from one part onto the other while moving on the surface and not crossing the curve. A doughnut is an example: the dashed line in Fig. 48 divides its surface into two parts, while the solid line does not.

Here's something to think about: what is the situation for a sphere with three handles (Fig. 49) and for a Moebius strip? However, on every surface there are closed curves which divide it into two parts, an inner

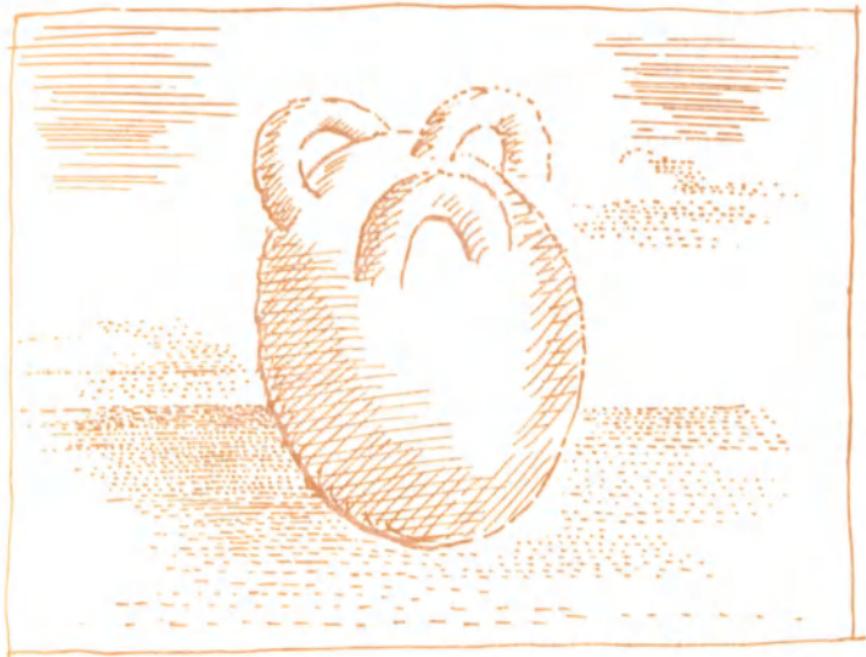


Fig. 49

part and an outer part. For the present this fact is more important to us.

Now imagine a living being existing in four-dimensional space. A closed jar will not be an obstacle, for it does not divide the living space into two parts. It will merely fly out of the jar via the fourth dimension.

The reader will also note that we ourselves live in a four-dimensional space where the coordinates are the three positional coordinates of ordinary three-dimensional space (x , y , z) and there is a fourth, time, coordinate t . These variables are not quite the same. Whereas x , y , z can assume arbitrary values as the signs and magnitudes vary, the time t can only increase. Yet even in this four-dimensional space it is possible

to get outside a closed room not through the doors or windows but by taking advantage of the fourth coordinate (time). By moving along this coordinate and retaining unchanged the other three, it is possible to find oneself in a different situation and ultimately to get outside the room (say, after the lapse of a certain time when the house falls to pieces and the walls are no longer an obstacle).

This of course may take quite some time but the fundamental possibility remains.

The situation becomes still more obvious if we allow for motion along the time axis in the reverse direction (backwards). The same point (x, y, z) of the space inside the closed room was once upon a time not surrounded by walls, floor and ceiling, for they had simply not yet been built. And so, by moving backwards a bit along the time axis we can extricate ourselves from any closed room.

Let us continue our sojourn into multi-dimensional worlds. In Fig. 50, we specify a circle in a plane (two-dimensional space), with centre at the origin of the coordinate system and with radius r , by the following equation:

$$x^2 + y^2 = r^2$$

A circle in a plane has an analogue in three-dimensional space. It is a sphere. A sphere of radius r with centre at the origin (Fig. 51) is given by the equation

$$x^2 + y^2 + z^2 = r^2$$

Now passing from three-dimensional to four-dimensional space, it is natural to give the name four-dimensional sphere of radius r with centre at the origin to a "three-dimensional surface" satisfying the equation

$$x^2 + y^2 + z^2 + t^2 = r^2$$

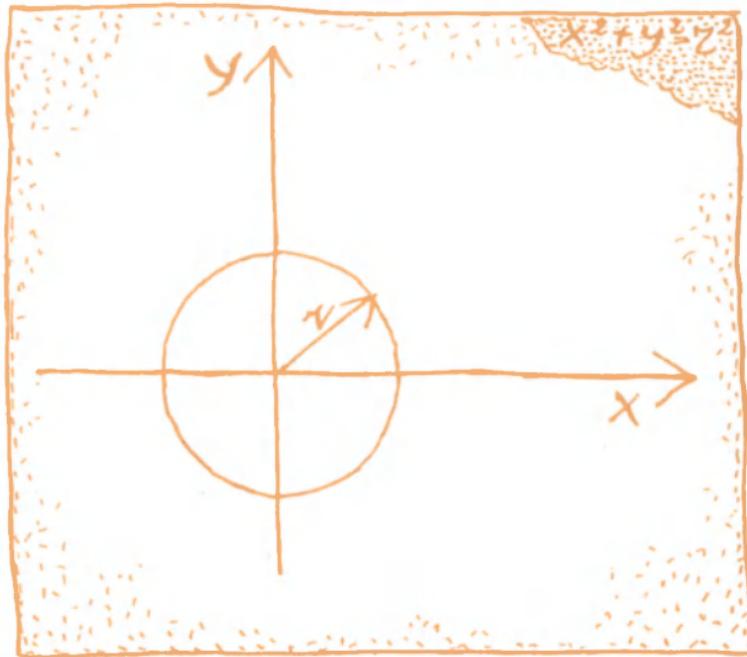


Fig. 50

Just as a chick living in the three-dimensional space of an egg cannot get out of the egg without breaking the shell, so a chick living in a four-dimensional space and placed inside a four-dimensional shell cannot get out of it.

To get out of a four-dimensional sphere the chick must break it in some way. But if the chick is living in a five-dimensional space, then the shell too must be similar to a five-dimensional sphere and not a four-dimensional one, since the latter cannot cover the embryo on all sides and it will be eaten by five-dimensional enemies before it grows up.

Now let us take a look at multi-dimensional space from another angle.

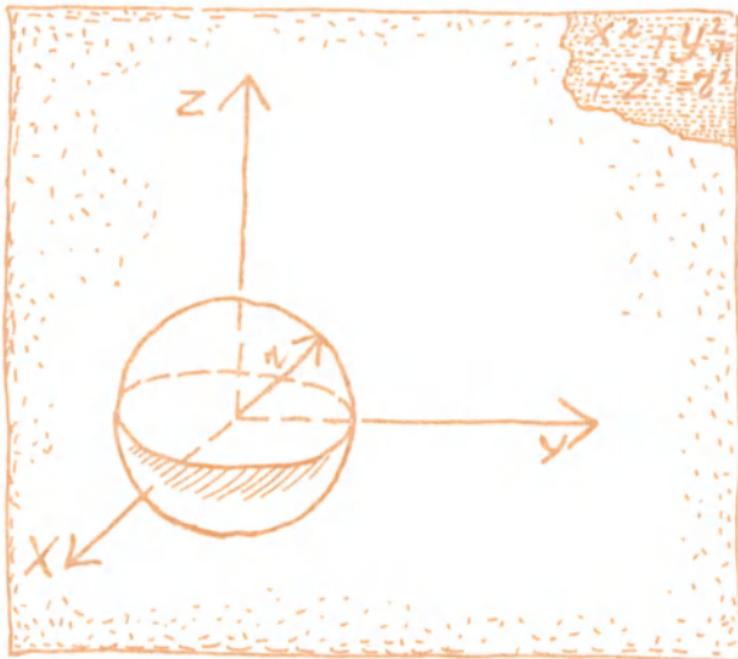


Fig. 51

Any point of a straight line divides the line into two half-lines without common points.

A plane cannot be divided by a point. But any straight line drawn on a plane divides the plane into two half-planes, and an ant moving from one half-plane to the other must cross the straight line dividing them.

In three-dimensional space the straight line will no longer be able to separate the space into two parts. But any plane will with ease separate the space into two nonintersecting half-spaces. A dragon-fly wishing to cross from one half-space to the other will have to cross the intersecting plane.

Similarly, a four-dimensional space cannot be divided by an ordinary two-dimensional plane. This

space is readily divided by any three-dimensional "hyperplane", which is to say, by any three-dimensional space lying in the four-dimensional space.

Thus, in any four-dimensional space we find certain subspaces of various numbers of dimensions: three-dimensional hyperplanes, two-dimensional planes, one-dimensional straight lines, and zero-dimensional points.

By analogy, n -dimensional space will have hyperplanes with different numbers of dimensions, from zero-dimensional (points) to $(n - 1)$ -dimensional hyperplanes. Only $(n - 1)$ -dimensional hyperplanes—those with the highest dimensionality—will be able to divide the n -dimensional space, whereas hyperplanes of dimensionality $n - 2$ and, all the more so, of smaller dimensionality will not be capable of dividing the n -dimensional space.

We have already established that the set of all points of a plane that satisfy the inequalities

$$0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1$$

represents a square. Its vertices are points with coordinates $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. In three-dimensional space, the figure analogous to this square is a cube. It may be defined as the set of all points (x, y, z) in space such that all three coordinates lie between zero and unity.

The vertices of the cube are points whose coordinates are equal either to zero or to unity (Fig. 52). As is readily seen, there are eight such points and each is defined by three coordinates: $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$, etc. A

It is natural, in four-dimensional space, to give the name four-dimensional cube to the set of points (x, y, z, t) , all four coordinates of which lie between zero and unity. The vertices will be points with coordinates

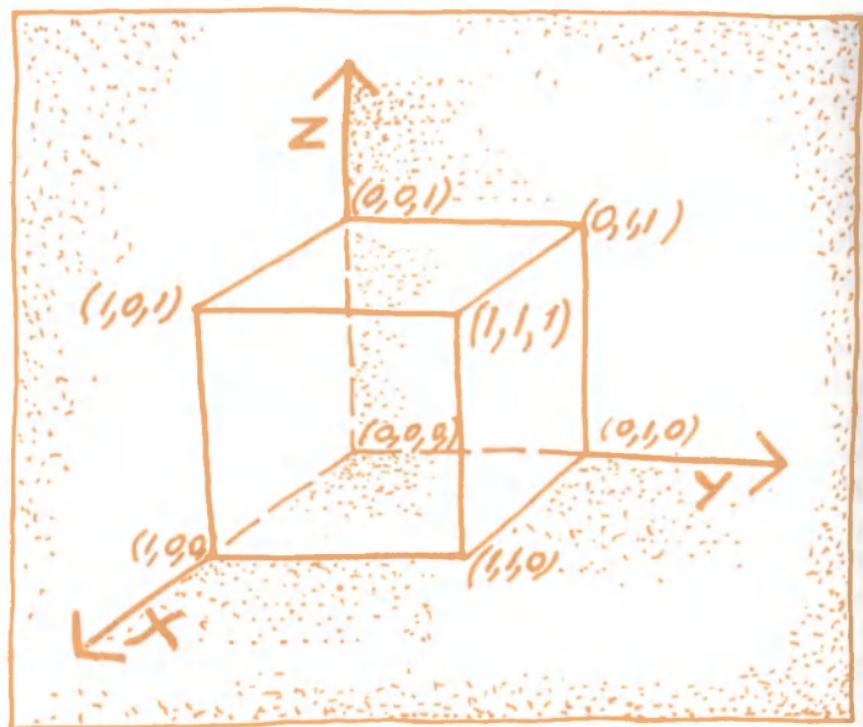


Fig. 52

equal either to zero or to unity: for example, $(0, 0, 0, 0)$, $(0, 1, 0, 1)$, $(1, 1, 1, 0)$, etc.

How many vertices does a four-dimensional cube have?

This is very easy to answer without writing out all possible points. Indeed, we already know that a three-dimensional cube has eight vertices. They constitute all possible combinations of number triples made up of 0 and 1. The vertices of a four-dimensional cube are obtained from these number triples by adding a fourth number (also either 0 or 1). Which means that a four-dimensional cube has twice as many vertices as a three-dimensional cube, or 16. Now note

that a two-dimensional cube, that is to say, a square, has $4 = 2^2$ vertices, a three-dimensional cube has $8 = 2^3$ vertices, and a four-dimensional cube has $16 = 2^4$ vertices.

It is easy to figure out that in n -dimensional space a unit cube is the set of all points whose coordinates lie between zero and unity. The vertices of such a cube are all points with coordinates equal either to zero or to unity. There will be 2^n sets of n numbers of zeros (0) or units (1), or, what is the same thing, vertices, in an n -dimensional cube.

All these facts and also the procedure for computing all possible groups of n zeroes and ones will be useful when we discuss other problems later on.

At this point in the manuscript, the editor said he couldn't see what contribution the ideas expressed here made to practical affairs, or to science or life in general. What he meant was that I as the author had not got that idea across in any way.

All I can say is: have a bit of patience. This is only the beginning. In what follows we will have to make use of multi-dimensional spaces and we will have need for the basic concepts of analytic geometry.

The Mathematics of a Saddle

Imagine a mountainous landscape with peaks and slopes and valleys and hills, and passes. It may not sound romantic, but such a surface can be represented analytically by writing

$$z = f(x, y)$$

where z is the vertical coordinate and x and y are coordinates in the horizontal plane (Fig. 53). The peaks correspond to maximum values of the function $z = f(x, y)$ and the valleys correspond to minimum values. If you are on a peak, there is only one way of going in any direction, and that is down; if you are in a valley, you can only go up. These points of maximum and minimum on surfaces will soon be of particular interest to us. If you are at some ordinary point of a surface, you can either go up or down. You can even choose a path that remains constantly at the same altitude. Such pathways are obtained by cutting the surface by a horizontal plane. Projections of such pathways onto one common horizontal plane are termed level lines (see Fig. 54). Those are the lines one sees on maps indicating height above sea level.

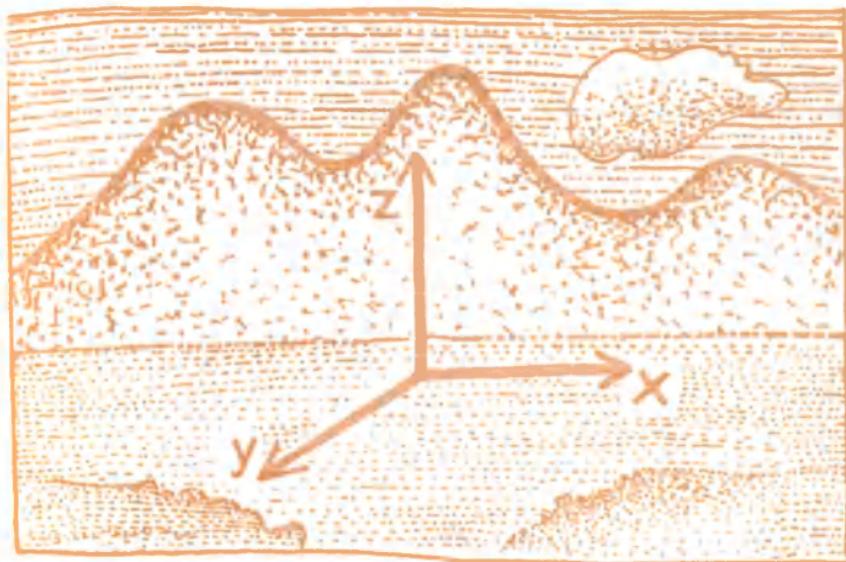


Fig. 53

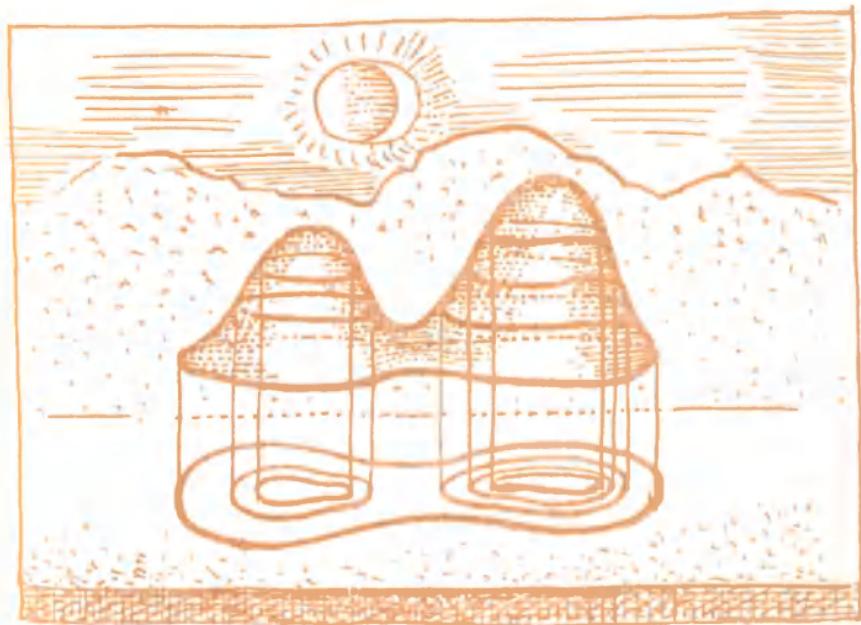


Fig. 54

An ellipsoid is a figure obtained by rotating an ellipse about its axis of symmetry. An ellipse has two such axes, one major and one minor. Rotation about the major axis yields an elongated, or prolate, ellipsoid, which looks like a cucumber, while rotation about the minor axis yields a compressed, or oblate, ellipsoid, which resembles a ball compressed from both sides.

We choose an arbitrary point P on the surface of an ellipsoid. It is always possible to intersect the ellipsoid with a plane so as to cut off a cap—the portion containing the point P . It is also possible then to choose the cutting plane so that the dimensions of the “cap” are very small (the mathematician would say: less than any preassigned number). Now let us take some point P on an arbitrary surface. If it is possible in any neighbourhood of this point to cut off a “cap” by means of a plane, then we will call this an *elliptic point*. By far not all points of a surface turn out to be elliptic. This will soon become apparent. We could also give a different definition of an elliptic point. We will draw various planes through P . If among these planes there are such that the entire piece of surface in the neighbourhood of the point P lies to one side of the plane, then P is an elliptic point.

Now let us come back to our mountainous terrain. Besides peaks and valleys we are particularly interested in mountain passes. A mountain pass resembles an ordinary horse saddle (see Fig. 55). Let us take two points A and B on different slopes of a pass (Fig. 56). One can travel from A and B along different routes (they are indicated by dashed lines in the figure) each one of which has a highest point denoted by an open circle. Quite obviously, among all these routes from A to B we can choose the highest point

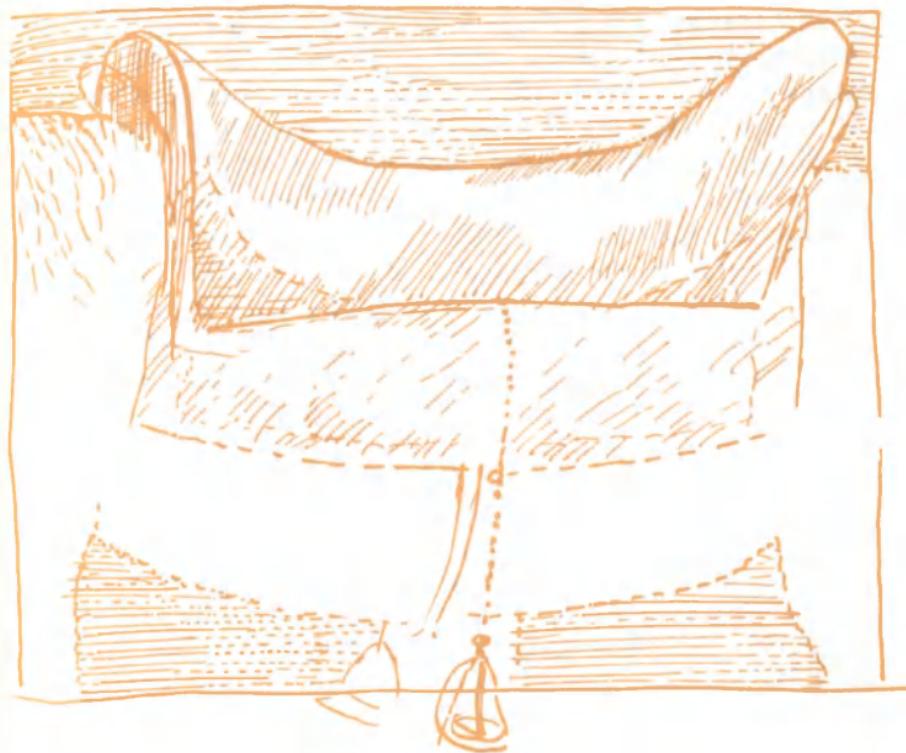


Fig. 55

that lies lowest. This route is depicted by the boldface dashed line.

Similarly, each solid-line route from point C to point D has a lowest point denoted by an open circle. From among all possible routes from C to D we choose the route whose lowest point is higher than all others. This route is depicted by the boldface solid line.

The highest point on the boldface dashed route and the lowest point on the boldface solid-line route coincide. We will call these points saddle points. If we incline the surface slightly, there will be a different saddle point.

We can give a different and perhaps more pictorial description of saddle points. Note that there is no plane that can cut off a "cap" in the neighbourhood of a saddle point. If we pass various planes through a saddle point, then in contrast to elliptic points, the plane in the neighbourhood of a saddle point will always intersect the surface so that there will be parts of the surface on both sides of the plane. In this description it will be seen that the point will be a saddle point irrespective of any inclination of the surface, or, in other words, irrespective of the choice of directions

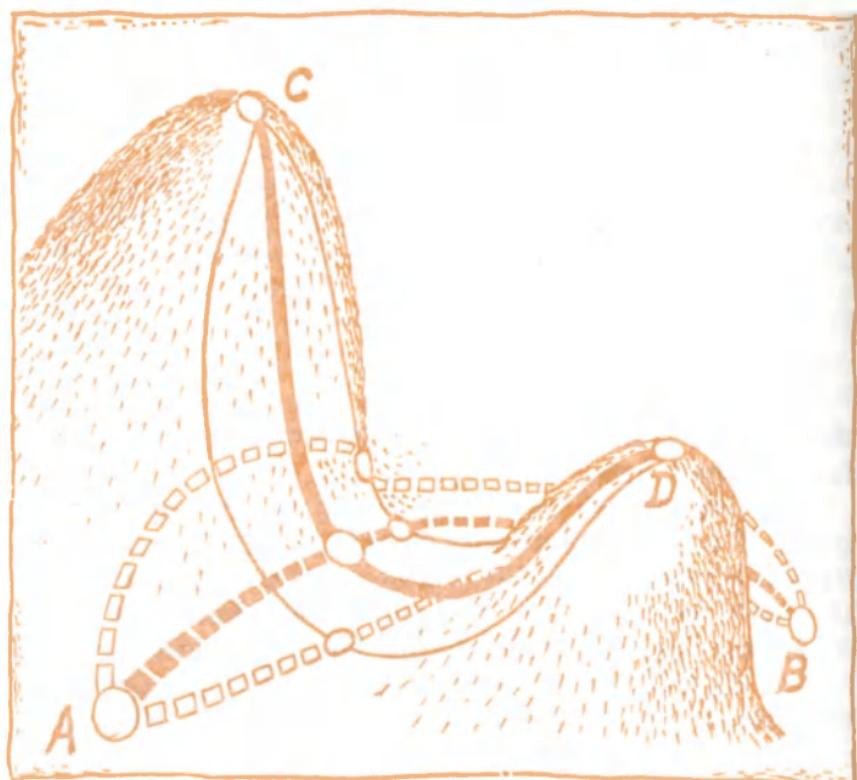


Fig. 56



Fig. 57

of axes of the Cartesian rectangular coordinates in space.

Just as a mountainous region can have several mountain passes, so a surface can have several saddle points.

Now I have a question for the reader. Can there be a very large number of saddle points on a surface? For instance, can a surface consist entirely of saddle points? If the answer is negative, then can a bounded piece of surface consist entirely of saddle points?

Before going on to what follows, try to imagine the situation we now have.

The answer is very simple. Take a look at the neck of an ordinary bottle (Fig. 57). All its points are saddle points. It is not difficult now to imagine an infinite surface all points of which are saddle points. To do so, take for instance a hyperbola whose equation is $x^2 - y^2 = 1$ (Fig. 58) and rotate it about the vertical axis. The resulting surface—a hyperboloid of revolution—will consist entirely of saddle points. The

hyperboloid is the most elementary surface possessing these properties. For this reason saddle points are also called *hyperbolic points*. Surfaces consisting entirely of saddle points play an important part in our lives.

Take for example the flat diaphragm of an ordinary telephone receiver. Clamp the edge of the diaphragm at several points and suspend small loads at certain other points (see Fig. 59). After the inevitable oscillations damp out, the diaphragm will assume a position in which all its points will be saddle points. It is not always possible of course to see this, but that is what the exact theory calls for, namely: under any deformation of the *edge* of a plane membrane (diaphragm), all its interior points will be saddle points.

If various parts of the edge of a membrane are heated in diverse ways, and the heat fluxes are held constant, the temperatures of its points will at first vary, but then will reach a steady state with the influx of heat equal to the efflux. If we lay off the temperatures on

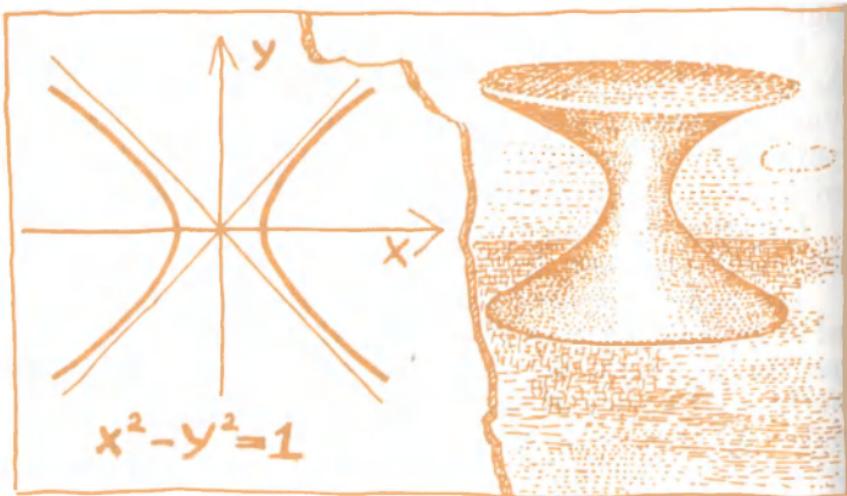


Fig. 58

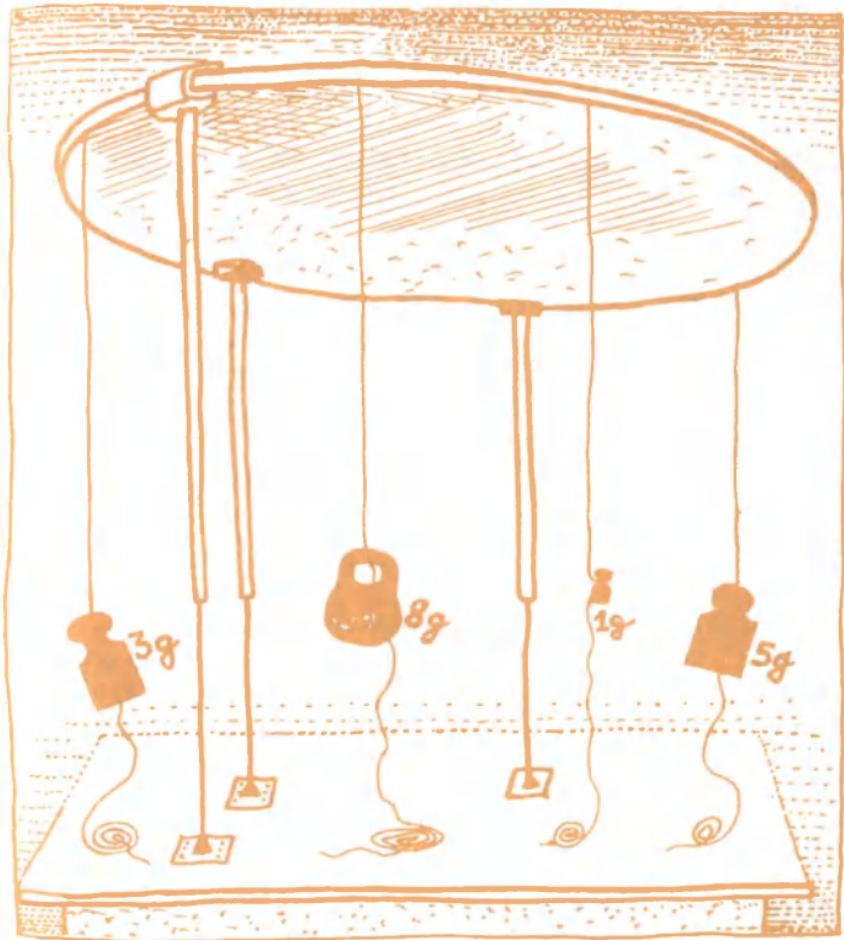


Fig. 59

a vertical axis, and position the membrane in the horizontal plane, then the appropriate "temperature surface" will consist entirely of saddle points.

The study of surfaces consisting solely of saddle points is closely connected with hydrodynamics, electrostatics, and other very important branches of science.

The shape of a fixed membrane is given by the solution of Laplace's differential equation (we will have more to say later on about this celebrated mathematician). The very same equation describes a steady-state irrotational flow of a noncompressible fluid and an established flow of heat, the distribution of forces in an electrostatic field, and a steady-state electric current, the diffusion of salt dissolved in water, and many other phenomena and processes. And all the functions—the solutions of these equations in a geometrical representation—prove to be surfaces that consist entirely of saddle points. That is why the investigation of such surfaces is very essential in a great diversity of fields of physics and technology.

EXTREMA

Extrema is the generic term for the concepts "maximum" and "minimum", like "parents" is the generic term for "father" and "mother". Extremal problems have to do with finding maxima and minima. We encounter them everywhere. It is hardly an exaggeration to say that all problems solved by living organisms are those involving a search for extrema.

Indeed, we are always seeking to extract the largest amount of something, produce the greatest effect, perform the maximum, and in the shortest time, or with least expenditure of energy. Also we want to get the maximum of pleasure or ensure the minimum of unpleasantness.

All problems of locomotion are extremal problems. When an animal wants to get from one place to another, it strives to do so via the shortest route, or as fast as possible, or by expending a minimum of energy.

Even when a person is standing still, he is constantly striving for some extremum. A standing person must

steady himself so as not to fall. And he cannot become rigid, like a pole, but must be ready to perform quickly all manner of movements. It turns out that what appears to be a stationary person is merely one that is in constant motion, always seeking a position of equilibrium. We will return to this interesting problem later on.

We begin the discussion of extremal problems with a problem that arises in tuning a television set. The picture on a television screen is always somewhat worse than the actual scene and so the problem of tuning consists in achieving as good a reproduction as possible. In other words, there is always an error of reproduction. Tuning consists in reducing this error to some attainable minimum.

Let us try tuning not during a transmission of programme material but when the usual resolution chart, or test pattern is shown (that's to keep family quarrelling at a minimum too).

According to the instructions, everything is extremely simple: "Using the knobs 'Brightness' and 'Contrast', set the brightness and contrast of the picture to that desired."

Let's begin. Turn the brightness knob as far as it will go: the brightness will fall off, the picture will deteriorate and we will have a high error of reproduction. Now gradually increase the brightness and watch the pattern. The reproduction error will gradually decrease to a minimal value, after which it will increase again until the pattern disintegrates entirely when the brightness becomes considerable.

If the electrical parameter that you control by turning the knob is denoted by V and the reproduction error by r , then the graph depicting their relationship (r as a function of V) will look just about like what we have in Fig. 60. The value of brightness cor-

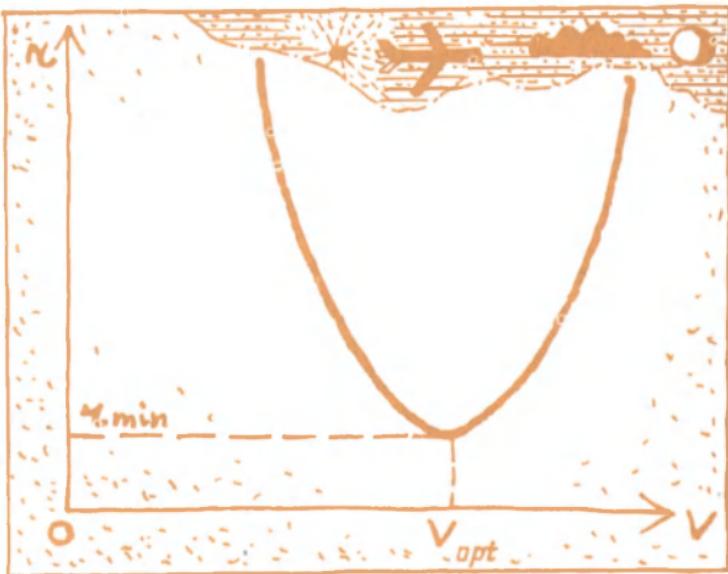


Fig. 60

responding to the minimal value of the error, r_{min} , is denoted by V_{opt} , which indicates the optimal value. We varied the brightness when the contrast knob occupied some given position. If you turn the contrast knob and then again vary the brightness, the r versus V curve will change, although the general trend will remain the same.

Use U to denote the electrical parameter controlled by turning the contrast knob. Other conditions being equal, in particular when determining the brightness V_1 , the dependence of the reproduction error r on U will be in the nature of a parabola. Varying the contrast gradually, we see that the reproduction error first decreases and then increases. But for various values of brightness, the r versus U curves differ too.

Fig. 61 shows several such curves corresponding to three values of brightness V_1 , V_2 , V_3 . The optimal

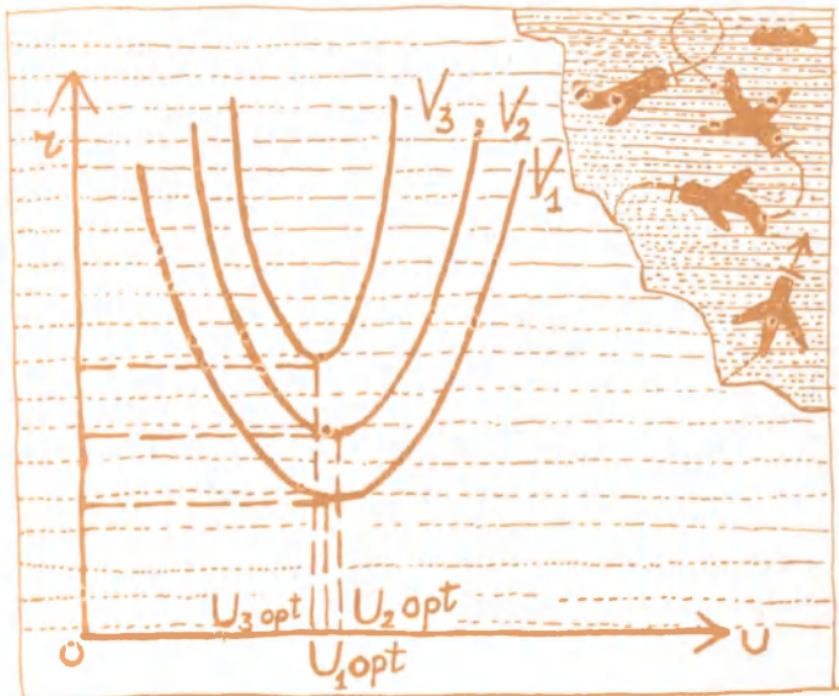


Fig. 61

values of U are indicated for each of the curves.

Thus, the reproduction error turns out to be a function of two variables, $r(U, V)$. Hence, to determine the least possible error of reproduction and, accordingly, the necessary values of the controlling parameters—brightness and contrast—we have to find the minimum of a function of two variables.

As we have already seen, a function of two variables is geometrically a surface. In the given instance, the surface resembles a cup and the extremal value for the reproduction error corresponds to its lowermost point (Fig. 62).

Maximum and minimum always exist together: if our cup-like surface is turned over, we get a cap, in which the highest point (maximum) corresponds to the lowest point of the cup (minimum). By climbing to the uppermost peak of a mountain we can find ourselves (via reflection in a nearby lake) in the lowest point of the valley below. Here, the mathematician calmly reasons to within an accuracy that amounts to the opposite, so to say, for if we find a maximum and then view the situation from another angle, we see a minimum. The answer thus depends solely on how we view the surface. That is why we always speak of seeking an extremum and not, separately, a maximum or a minimum. This kind of reasoning occurs

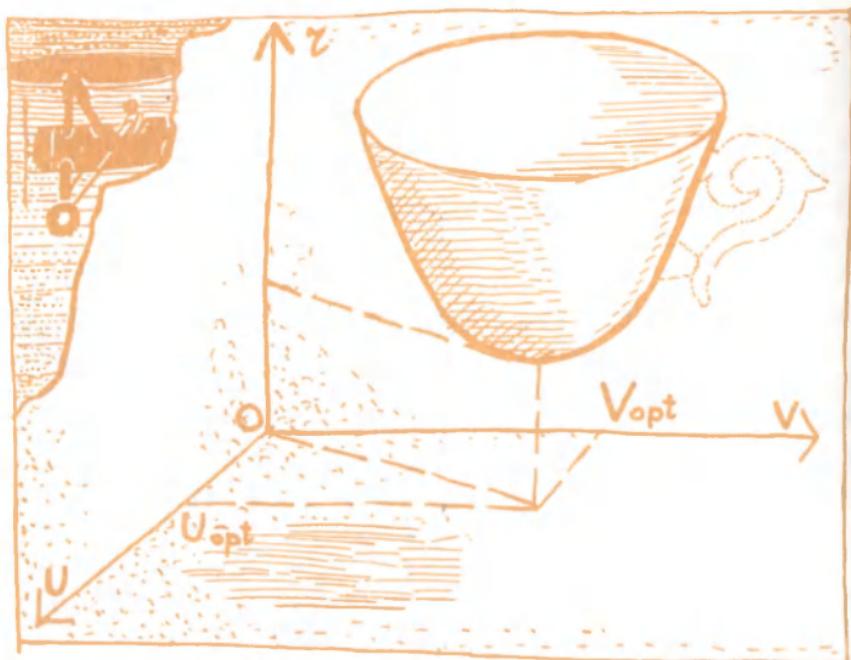


Fig. 62

in a wide range of situations and often greatly simplifies matters.

In the turbine drilling of wells we also seek extrema. It resembles the case we have just discussed. Normal operation of an oil well yields tens and hundreds of tons of oil per day. Drilling costs for a single well run into hundreds of thousands of rubles. For this reason, any cut in drilling time represents a considerable saving in money. In turn, increasing the rate of sinking a well results in a cut in the drilling time.

If the drilling is done with a turbo-drill, the drill solution is delivered under pressure via a column of steel pipes lowered into the well. The flow of the drill solution actuates the turbo-drill, which breaks up the rock. Also, the solution raises the drilled rock to the surface.

The boring tool—the boring bit—breaks up the rock if a certain pressure is exerted. For a constant rotational speed of the boring tool, an increase in pressure results in an increase in the rate of sinking the well. However, this increase in sinking rate continues only to a certain definite limit. If the pressure on the tool is too great, the tool will press on the rock and the rotation of the bit will slow down. The rate of sinking will fall to zero when the tool stops. The graph of the rate of sinking W versus pressure P at the heading is in the shape of an upturned parabola, that is, with the maximum facing up.

If we assume all other quantities constant, then we can determine the pressure at the heading for which the rate of sinking is a maximum.

True, the problem thus stated is rather simplified. Actually, the rate of advance depends on many other quantities. First of all, it depends on the consumption of drill solution, which is to say, on the quantity of liquid pumped through the turbo-drill per second.

An increase in consumption produces a higher rate of rotation of the turbine. Thus, the rate of advance depends now on two variables: on the pressure at the heading and on the consumption of drill solution.

The crust of the earth is not homogeneous but rather resembles a layer cake with a multitude of layers of diverse structure. It is clear that the rate of advance is substantially dependent on the hardness of the rock. Hence, maximum rate of advance is also a function of the properties of the rock and, thus, is now a function of three variables.

Here, ordinary geometrical representations are no longer feasible, for we are in four-dimensional space. (There is nothing to fear, since you are now used to multi-dimensional spaces.) However, here too we can invest with exact meaning certain concepts analogous to the lowest point of a valley or the topmost point of a peak.

A more careful study of the problem shows that actually the rate of advance depends not only on three but on a much larger number of variables. The rotational speed of the turbine is a function not only of the consumption of the drill solution, but also of its specific weight and viscosity. When in operation, the drill tool undergoes considerable wear, and the rate of advance depends to a very essential degree on the state of the tool at a given time. Other parameters could also be indicated as affecting the rate of advance of a well. Hence, the search for the highest possible rate of advance is a mathematical problem which involves finding the extremum of a function of a large number of variables.

If we know the type of functional relation between the variables, then it is possible, by standard mathematical methods, to find the extremum of the function and those values of the variables for which that extre-

mum is attained. Such methods are described in any textbook of mathematical analysis. After a few rather simple operations, the whole matter reduces to solving a system of equations. The system ordinarily contains just as many equations as there are variables, but the general aspect of the system may be formidable.

The reader will most likely recall the difficulties sometimes involved in solving an equation in only one variable. Not the quadratic equations of so many a headache, but something really complicated, like a trigonometric equation or one in exponential functions.

Further recall how it was usually done: thinking up a substitution, replacing the variables and—to one's ultimate joy—an equation that reduced to a linear or quadratic equation.

I must disappoint you, such neat problems only occur in school textbooks. In actual life situations, one rarely encounters an equation that can be reduced via substitutions to a quadratic equation; so rarely in fact that it is not worth the time to search for a suitable substitution, unless it can be guessed at a glance.

The point is that there are equations which cannot be solved for the unknown. For them it is quite impossible to write down the expression explicitly, that is to say, to derive a formula for finding the roots. Here is an instance:

$$a^x - ax = 0$$

One solution can be guessed at once: $x = 1$. But it is impossible to write down an explicit formula for finding all solutions, and the second root (there are two here) cannot be found in explicit form at all.

Incidentally, you of course know that algebraic equations of degree three and four cannot be reduced

to quadratic equations and their solution is anything but simple.

But let us return to the original problem of finding extrema. The mathematician solving an extremal problem is bound to meet a great many difficulties on his way. Imagine that it is required to determine a largest integer. I assert—and this runs counter to common sense—and will prove that it is unity.

Suppose the largest integer is greater than 1. We denote it by N (assume, say, that $N = 2$).

Thus, our assumption is that N exceeds 1. But then N^2 exceeds N (actually, $2^2 = 4$ is greater than 2) and N^2 is an integer. Which means that N is not the largest number (since 4 exceeds 2). Now the square of unity is equal to itself ($1^2 = 1$). Thus, unity is the largest integer.

Nonsense! And the reason is that I *assumed* that there is indeed a largest number; in other words, I assumed that there exists a solution to the extremal problem thus posed. Actually, of course, there is no solution since the number of integers is infinite.

As the German mathematician Hausdorff so neatly put it, if two times two is five, then we can believe in witches. Generally speaking, every incorrect assertion implies another incorrect assertion.

This instance and the conclusion drawn from it is apparently very important to every person involved in research, including experimental investigations, for if we proceed from an erroneous premise or use improper reasoning, the experimental research, even when carried out with meticulous care, can lead to erroneous, even paradoxical conclusions. The experimenter is often saved by the simple use of common sense, but I'm afraid we can't always rely on that (we will return to this question a bit later when we

discuss in more detail the methods of work of the mathematician).

Now suppose an investigation has been carried through and the existence of a solution to the equation has been proved, but the equation is so complicated that it cannot be solved in any simple and exact fashion. What do we do in such a situation? We resort to approximate methods of finding the roots. These include both analytic and graphic procedures. Say, for the solution of the equation

$$\left(\frac{3}{2}\right)^x - \frac{3}{2}x = 0$$

we can construct in one drawing the graphs of the functions $y = \left(\frac{3}{2}\right)^x$ and $y = \frac{3}{2}x$. The desired roots x_1 and x_2 are the abscissas of the intersection points of the curves (Fig. 63). Of course, the graphical solution gives the roots very approximately, but it can suggest a method for a more exact analytic procedure for determining the desired root in approximate fashion.

In short, the situation is not so bad for handling a function of one variable. But if we want to solve (even approximately) a system of equations, particularly a system involving a large number of variables, then the complications may become unbearable. Today we can resort to high-speed electronic computers in such cases.

To get some idea of the complications, note that a computer performing 20,000 operations per second requires about one hour of machine time in order to solve a linear algebraic system of one hundred equations in one hundred unknowns. But the point is that when seeking parameters that ensure maximum rate of advance in sinking a well, the system of equations will not be linear, and we need the answer at

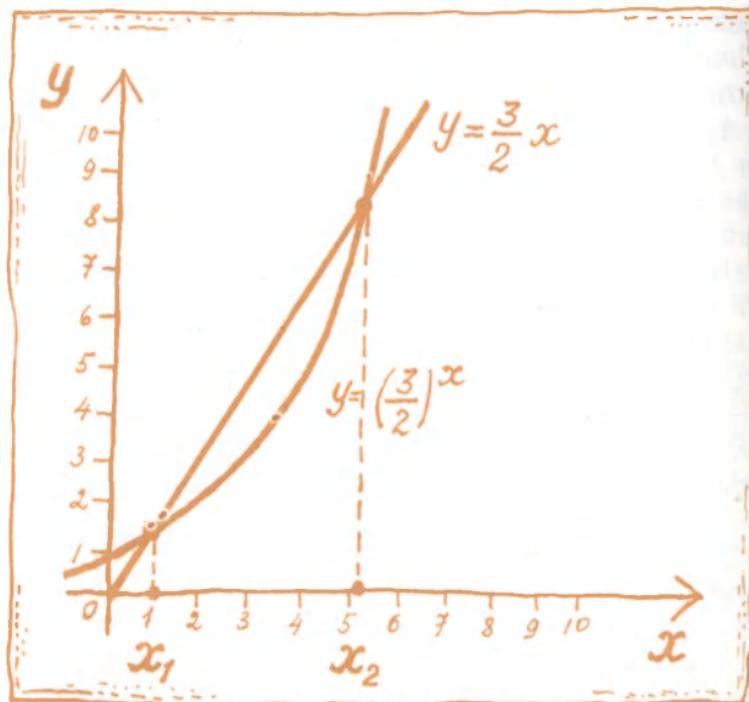


Fig. 63

once, within seconds or minutes, for within that time the situation will have changed, the properties of the layer of rock will be different, the tool will have worn to some extent, etc., and so our findings will be out of date.

What do we do in that case?

We shall soon see.

EXTREMAL CURVES

The shortest distance between a window and the door of a room is a straight line. But if the room is filled with furniture that cannot be moved, the shortest distance may prove to be harder to find.

Suppose we have to go from the traditional point *A* to the no less traditional point *B*, and there are two routes, one straight but difficult, the other tortuous and long but very easy. The straight-line route is shorter but knee-deep mud makes travelling well nigh impossible, with the result that a great deal of time will be needed. Now if the aim is to solve the problem for a minimum distance, we will have to choose the mud. But if the minimum sought is that of time expended or effort involved, then the longer route will do better.

In other words, when seeking the best of all possible routes, one must clearly state what is meant by best.

It is easy enough to indicate the shortest distance from a window to the door of a room, but how is this done when we want the shortest route from the peak of a mountain, say Elbrus, to the foot of the mountain? The answer is not obvious at all. True, a blind horse would all the time instinctively go down, in the direction that falls most sharply. Such also is the behaviour of running water.

The "physical" solution of this problem is very interesting. Imagine an extremely arbitrary smooth surface. Suppose, in addition, that it is convex. If we draw taut a fine rubber band between two points of the surface, it will take up the position of the shortest line connecting the two points.

Now if we undertake to construct a highway of shortest distance up a mountainside with the gradient not to exceed, say, five degrees, the problem becomes rather complicated.

Animals and human beings do not always properly resolve problems involving optimal routes or, as the mathematician would say, problems that call for determining an extremal.

I've heard that a dog running after a hare runs

straight for it at any given instant. It will catch up with the hare if it runs faster, but the dog will not do so in the shortest possible time. If the aim is to catch up with the hare in the shortest time, then the route of pursuit will have to be changed, so that the dog then aims not at where the hare is at the time, but where he will be within a short time, at the so-called point of prediction. Hunters and anti-aircraft men know what that means. True, the latter miss too, but this is due to inexact computations of the point of prediction for the missile to hit the target. True again, it is not always their fault for the simple reason that one cannot predict the behaviour of an animal or an enemy aircraft unambiguously. Computing the point of prediction is a very difficult problem that continues to be attacked by mathematicians and engineers.

A multitude of extremely important problems in the natural sciences and technology reduce to determining extremals, which are studied in a branch of mathematics called the calculus of variations.

Although certain extremal problems of this kind were solved even by the ancient geometers, it was the differential and integral calculus that formed the true basis for their investigation. The calculus of variations was created by Leonhard Euler in the middle of the eighteenth century. However, new problems in technology and physics, in particular automation and cybernetics, have required new methods in the calculus of variations which are now in full bloom.

THE EPOCH OF EULER

This is not going to be a history of mathematics. But since I have mentioned Leonhard Euler, I cannot help but give a brief picture of the life of this brilliant

luminary in a mathematical world which is so rich in remarkable talents.

Euler was born in 1707 in Basel (Switzerland). His father was a clergyman who hoped that his son would enter the ministry, but he also helped to instruct his son in the elements of mathematics, for he himself had been a pupil under the famous Jacques Bernoulli.

By the age of twenty, Euler had been broadly trained in theology, medicine, and oriental languages. In 1727 he was invited to St. Petersburg to the chair of physiology after he had failed to meet the requirements for a position with the chair of physics at Basel University. (How inexact the system of placement of scientists via ballot can be at times!) Incidentally, by this time he had already made considerable advances in mathematics and physics. For example, his essay on the masting of ships was published by the Parisian Académie des Sciences and was well received.

Euler lived for many years in St. Petersburg. In 1730 he became professor of physics at the St. Petersburg Academy of Sciences and in 1733 headed the chair of mathematics where he remained until 1741. Science in Russia was in a state of decay, the tsarist administration hampered scientific work and Euler left Russia and took up a position in Berlin. But in 1766 he returned to St. Petersburg for good. He died there in 1783.

Euler's rate of scientific output in the form of papers and books is phenomenal. A bibliographical list of his writings comes close to 900 items. His interests were extremely broad and the results he obtained were fundamental. For instance, in astronomy he extended the theory of lunar motion to practical application. He also made important contributions to hydrodynamics, optics, navigation, cartography, artillery and the theory of numbers. Euler made very

substantial advances in mathematical analysis, differential equations and the calculus of variations, which I have already mentioned. Incidentally Euler also wrote on medicine and physiology and even theology.

In 1736 Euler went blind in one eye due to overstrain, but this did not stop his constant outpour of scientific writings. Soon after his return to St. Petersburg in 1766 he went blind in the other eye as well, and yet he continued with a spate of papers and books which he dictated to his son and pupils right up to his death in 1783.

The publication of his collected works was started before the First World War under Swiss auspices for worldwide subscription and was originally planned at about 40 volumes. However, 50 volumes have already come out and there is still a great deal of work to do. It is now believed that the number of volumes may reach two hundred in all.

SOAP BUBBLES

The calculus of variations provides us with the machinery for solving a broad range of problems. It is not only used to find the shortest route between point *A* and point *B* but also to solve problems involving the search for a great diversity of extremal quantities.

It is common knowledge that, in a plane, of all figures having a boundary of a given length (or with a given perimeter, as we would say in elementary geometry), the circle has the largest area. In three-dimensional space, the solid of greatest volume for a given area of the bounding surface is a sphere. Conversely, of all solids of a given volume, the sphere has the least surface area. That is the precise reason

why soap bubbles appear in the form of spheres.

Let us take up some less obvious problems.

A circle can be the boundary of a surface, say of a pail. Now of all surfaces having such a boundary, the one with a minimal area is the plane disk stretched over that circle. Now distort the circle so that the curve can no longer be superimposed on the plane. There are any number of surfaces having such a boundary. But how does one find the minimum-area surface among them? That is already a difficult problem, and to solve it analytically requires applying methods of the calculus of variations. It turns out—Euler established this fact—that at every point such a minimal surface is a saddle-like surface.

It is interesting to examine a physical solution of this problem. Put a closed contour (circuit) made of thin tin wire into soapy water. The resulting soapsuds have a small surface tension. A soapy film will adhere to the contour and its area will be the smallest possible area. We have of course disregarded the force of gravity and other forces that prevent the film from attaining a state of stable equilibrium. Stable equilibrium is attained when the area of the film is minimal, since in that case the potential energy due to the surface tension is minimal.

You have probably forgotten the fun you once had making soap bubbles. Try it again. Take off a few minutes to return to childhood and we'll perform a number of experiments.*

Solder a soft wire into a circle with two handles (that'll make it easy to distort into a variety of sha-

* Try this solution: 10 grammes of pure dry sodium oleate dissolved in 500 grammes of distilled water. Then mix 15 parts of the solution and 11 parts of glycerine. The frames used should not be very large, not more than 10 to 15 centimetres in diameter.

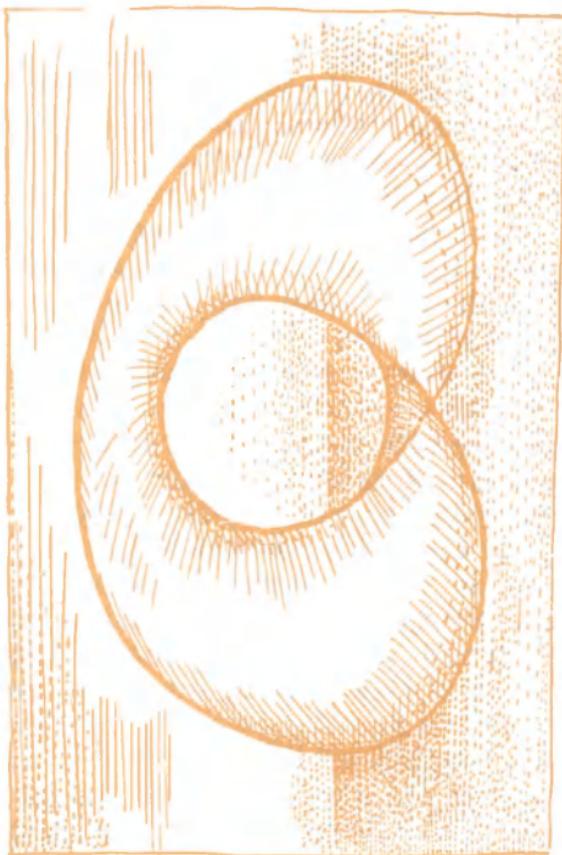


Fig. 64

pes) and dip it into soapsuds. A soapy film will stretch out around the contour. Now gradually twist it. You will see that by continuous twisting of the contour you can transform a two-sided membrane stretched onto a circle into a one-sided Moebius strip (Fig. 64). This is a remarkable fact, for the original surface and the resulting surface are not topological equivalents!

If you bend the circle into a space curve, as shown in Fig. 65, then you can stretch three different minimal

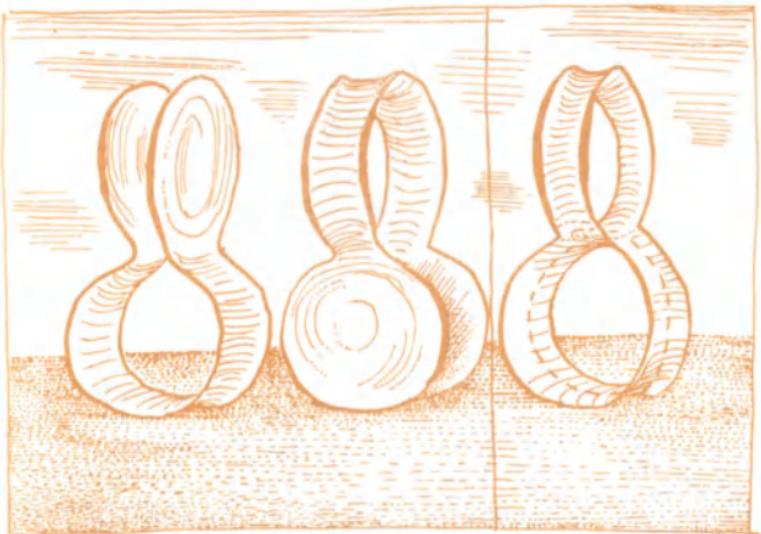


Fig. 65

Fig. 66

surfaces on your contour. On the latter (Fig. 66), it is possible to draw a closed curve, like the dashed line in the figure, which cannot be contracted continuously into a point without tearing. The other two surfaces do not possess this property. You will recall a similar situation when we compared a sphere and a doughnut.

All these beautiful geometrical figures are not only for fun. Surfaces of minimal area are the most rigid, and they find extensive application in the development of rigid structures in engineering practice.

Mathematicians Are Not All Alike

We have examined the subject matter of the science of mathematics, but we have not discussed the mathematicians themselves, though undoubtedly they make up quite an exciting group.

An elderly scholar is often pictured with goatee and sitting on a step-ladder, pawing through dusty volumes. Mathematicians are more usually depicted as young men, which statistically is true enough, for, like musicians, they develop early, by 25 to 30 years of age, while some make remarkable discoveries and become famous at 20. Now for some strange reason the mathematician is often depicted as a near-sighted, carelessly dressed unkempt person of indefinite age bumping into people or sitting in a crouch in a corner of the room and thinking his deep thoughts.

True, one does come across such mathematicians—once in a while. But I must disappoint you. In most cases, particularly in youth, such prodigies occasionally do try to play their part close to the stereotype of our literary hero. Actually, they are perfectly normal human beings, whereas oversolicitous relatives or friends give them the prodigy status instead of



washing their faces and laughing them off to the barbershop. A fitting definition for a child prodigy would appear to me to be "a normal child with abnormal parents".

Another type of mathematician is the dry pedant, always fully buttoned up, punctilious in everything and demanding a senseless learning by heart of all theorems and the solution of problems only according to a firmly established scheme. Such personages are quite convinced that science is dry and must be so, otherwise it wouldn't be science at all. Such mathema-

ticians are encountered once in a while, but actually they are merely a sad misunderstanding.

In reality, mathematicians more often combine their mathematics with mountain climbing, skiing, swimming and playing basketball. Then again we find the fashionable dressers, plain fellows, Don Juans and even pretty girls. What then is the difference between mathematicians and doctors, biologists or economists? Mainly, I would say, the difference lies in the way they argue.

WHERE DO AXIOMS COME FROM?

Most people imagine mathematics to be a *deductive* science in which all theorems, results and facts are obtained via logical reasoning by proceeding from certain starting axioms, primal assertions, assumed to be self-evident or not requiring any proof.

To a certain extent this is true, although I will soon have something to say about the imaginary self-evidence of axioms. But that is only half the matter.

We probably all recall from our school days how deductive mathematics is constructed. True, our impressions are as a rule quite distorted by the aridity of the ordinary school course. But what I want to dwell on now is the so-called *inductive* approach in the construction of every mathematical theory, on the processes of the birth and death of mathematical theories.

One might of course think that a mathematical theory is built up as follows: the mathematician thinks up some starting premises (axioms), proceeds to verify them as regards self-consistency and independence (otherwise he will not be able to extract anything worthwhile), and then derives from them a number of corollaries for his own pleasure or for some other pur-

poses, say, to increase the list of published papers.

It is paradoxical to say so, but even today there are some people (and not only those hardly acquainted with mathematics but even specialists, including mathematicians, who advocate the use of mathematics in the natural sciences) who take this primitive viewpoint.

Actually, the mathematician does not think up any system of axioms and does not build any theories devoid of purpose or meaning.

Every meaningful mathematical theory is a reflection of reality: the mathematician idealizes concrete phenomena into a rough scheme when he creates the starting propositions of a theory; later, when he already has drawn certain conclusions, he compares them with the phenomena of reality.

TWO WAYS OF REASONING

In life and in scientific work we apply reasoning. There are two types of reasoning: demonstrative and nondemonstrative (but still convincing). The latter is called plausible reasoning.*

* Outstanding mathematicians have always understood both the difference between demonstrative and plausible reasoning and the place occupied by plausible reasoning in all fields of science, including the role of such reasoning in the creative efforts of the mathematician. A great deal has been written on this subject both by classical mathematicians and by modern researchers. Perhaps the best discussion of these problems has been given in two excellent books by G. Polya, an outstanding Hungarian mathematician and teacher now working in the United States. The first book, entitled *How to Solve It*, 1946, uses the school course of mathematics to demonstrate ways of reasoning aimed at helping the student to solve problems, to learn to guess and reason. The second book, *Mathematics and Plausible Reasoning* (Vol. I, *Induction and Analogy in Mathematics*, Vol. II, *Patterns of Plausible Inference*), makes use of higher mathematics

Plausible reasoning is based on induction, analogy, observations, hypotheses and experiments; all these methods are used by natural scientists.

We are not speaking here of complete mathematical induction such as is used in school to prove, say, the famous binomial theorem of Newton, but about ordinary induction, about the observation of particular phenomena and the deriving, on their basis, of more general regularities (laws).

Mathematical knowledge is fixed securely by means of demonstrative reasoning, but the approaches to such knowledge are strewn with plausible modes of reasoning.

But plausible reasoning starts from conjectures. Conjectures of course are diverse, ranging from extremely reliable ones, like the Newtonian laws or Mendeleev's Table of Elements to not very reliable ones, like modern cosmogonic hypotheses or theories on the origin of life on the earth, where a single new fact can bring about a cardinal recasting of the whole theory. We also find conjectures of so low a quality that they may readily be classed as gossip.

The derivation of the Pythagorean theorem or the formula for solving quadratic equations belong to the class of demonstrative reasoning. Now the inductive reasoning used in deriving the law of universal gravitation, the Lomonosov-Lavoisier law, or Darwin's theory of natural selection are all examples of plausible reasoning. They are based on observations of a limited

as well as elementary mathematics. There are many examples given in this volume taken from mathematical analysis, the classical calculus of variations, and the theory of probability. However, the general ideas are equally applicable to all branches of science.

Apparently, it was Polya who introduced the term "plausible reasoning". It differs from what was used formerly—induction, which has a broader meaning.

number of experiments and are for that reason conjectures, albeit conjectures of genius.

In the nature of plausible reasoning and not proofs are the arguments of a doctor diagnosing a patient, the reasoning of a Sherlock Holmes following up a crime, the document-buttressed arguments of a scholar concerning the activities of the state of ancient Rome, or the statistical arguments of an economist on the usefulness or otherwise of payment by the piece.

Demonstrative reasoning differs from plausible reasoning just as the fact differs from the supposition, just as actual existence differs from the possibility of existence.

Demonstrative reasoning is reliable, incontrovertible and final.

Plausible reasoning is conditional, arguable and oft-times risky.

Every science is permeated with proofs, and to the same extent as mathematics, for demonstrative reasoning is an integral part of mathematics.

Note, however, that in carrying out the inexorable proof of the Pythagorean theorem we gain no new knowledge other than that our hypothesis (the square of the hypotenuse is equal to the sum of the squares of the other two sides) is true.

The element of newness was contained in the hypothesis itself, and the latter is what had to be conjectured prior to one's beginning the proof.

Thus, demonstrative reasoning in itself does not contribute any essentially new knowledge about the surrounding world. Everything that is new is connected with plausible reasoning.

The outstanding American mathematician Richard Bellman had this to say on the subject in his book *Introduction to Matrix Analysis*: "...Logic, after all, is a trick devised by the human mind to solve certain

types of problems. But mathematics is more than logic, it is logic plus the creative process. How the logical devices that constitute the tools of mathematics are to be combined to yield the desired results is not necessarily logical, no more than the writing of a symphony is a logical exercise, or the painting of a picture an exercise in syllogisms."

In mathematics the problem of the essence of proof has been thoroughly enough worked out and every mathematician must master the methods of demonstrative reasoning. Appropriate rules have been established for this purpose. These rules and the concepts of rigour and exactitude of reasoning vary from century to century, and at the present time every mathematician knows the level of rigour of modern mathematics.

Now there is no standard for plausible reasoning, there is no theory similar to the theory of demonstrative logic; yet every scientist needs them like the very air, for without them there can be no science.

Mathematics is the sole avenue for learning how to reason via proof. On the other hand, one must also learn how to conjecture.

It is hard to expect that a unified methodology will be worked out for learning how to conjecture and guess. The diversity of human individualities is too great for that.

Like many other kinds of human activity, plausible reasoning is mastered by imitation and practice. And thanks to the peculiarities of mathematics, this subject is better suited as material for learning how to reason plausibly than any other science. A complete mathematical theory appears as a pure theory of proof. But the assertion that "mathematics is a demonstrative science" describes only one aspect of the matter, because the process of creating a mathematical theory is the same as in other sciences. Before

proving a mathematical fact, one has to discover it, guess it, conjecture it.

In a rigorous case of demonstrative reasoning, the main thing is to be able to distinguish proof from conjecture, justified proof from an unjustified attempt. In plausible reasoning, one must distinguish a reasonable conjecture from a less reasonable conjecture and be able to substantiate the conjecture with the available facts, to find these facts, to search painstakingly for facts that contradict the conjecture, and to correlate the findings and again return to plausible arguments.

I stress the idea of searching for facts that run counter to the conjecture. In ordinary life, one does not always seek the truth. At times, ignorance is indeed bliss, whereas knowledge of the true facts makes it necessary to take undesirable decisions. But in science, self-complacency and faith in one's infallibility are invariably harmful.

Here are some examples of plausible reasoning which I am sure will convince you that our misgivings are fully justified.

If you put a cockroach on the table and then tap the table with your hand, the cockroach will scurry away. Now tear off the feet of the cockroach and tap the table: the cockroach does not move. Consequently, the cockroach hears by its feet.

Rather laughable reasoning. But does it differ so much from the once commonly accepted explanation of the cause of cholera, influenza and typhus—punishment by the Almighty and not by infection from one person to another via microbes or viruses? Hardly.

Now let us take an example of a "scientific" version of plausible reasoning. It is given in Arthur Clark's marvellously interesting *Profiles of the Future*.

"For a sample of the kind of criticism the pioneers

of astronautics had to face, I present this masterpiece from a paper published by one Professor A. W. Bickerton, in 1926. It should be read carefully, for as an example of the cocksure thinking of the time it would be very hard to beat.

"This foolish idea of shooting at the moon is an example of the absurd length to which vicious specialisation will carry scientists working in thought-tight compartments. Let us critically examine the proposal. For a projectile entirely to escape the gravitation of the earth, it needs a velocity of 7 miles a second. The thermal energy of a gramme at this speed is 15,180 calories The energy of our most violent explosive—nitroglycerine—is less than 1,500 calories per gramme. Consequently, even had the explosive nothing to carry, it has only one-tenth of the energy necessary to escape the earth.... Hence the proposition appears to be basically impossible....'

"Indignant readers in the Colombo public library pointed angrily to the SILENCE notices when I discovered this little gem. It is worth examining it in some detail to see just where "vicious specialisation," if one may coin a phrase, led the professor so badly astray.

"His first error lies in the sentence: 'The energy of our most violent explosive—nitroglycerine...' One would have thought it obvious that *energy*, not violence, is what we want from a rocket fuel; and as a matter of fact nitroglycerin and similar explosives contain much less energy, weight for weight, than such mixtures as kerosine and liquid oxygen. This had been carefully pointed out by Tsiolkovsky and Goddard years before.

"Bickerton's second error is much more culpable. What of it, if nitroglycerin has only a tenth of the energy necessary to escape from the Earth? That me-

rely means that you have to use at least ten pounds of nitroglycerin to launch a single pound of payload.*

"For the fuel itself has not got to escape from Earth; it can all be burned quite close to our planet, and as long as it imparts its energy to the payload, this is all that matters. When Lunik II lifted thirty-three years after Professor Bickerton said it was impossible, most of its several hundred tons of kerosene and liquid oxygen never got very far from Russia—but the half-ton payload reached the Mare Imbrium."

This quote hardly needs any commentary. We will not spend any more time on analysing plausible reasoning that pertains to the problem of the motion of bodies in cosmic space or to any other problems. All I wanted to demonstrate was the difficulty of reasoning in a plausible manner.

Plausible arguments convince to varying degrees in the different sciences. In physics one often encounters rather cogent arguments of this nature, whereas in the humanities and, oft-times, in the natural sciences too, the degree of plausibility is very low.

Here is a joke in the mathematical field.

Says the mathematician, "A physicist thinks that 60 is divisible by all numbers because he sees that 60 is divisible by 1, 2, 3, 4, 5, 6 and he then checks a few more numbers taken at random, say 10, 15, 20, and 30. And since 60 turns out to be divisible by all these numbers, he regards his experimental findings as sufficient."

"Yes," says the physicist, "but look at the engineer. He suspects that all odd numbers are prime (a prime number is a number divisible solely by itself and unity). At any rate, 1 may be regarded as prime.

* The dead weight of the rocket (propellant tanks, motors, etc.) would actually make the ratio very much higher, but that does not affect the argument.

Then there are 3, 5, and 7—all undoubtedly prime numbers. Then comes 9, which apparently is not a prime number. But 11 and 13 are of course prime. Let's return to 9, says the engineer, I must conclude that 9 is merely an experimental error."

"True," says the engineer, "but look at the physician. He allowed a hopelessly ill patient with uremia to eat a dish of cabbage soup and the patient got well. The doctor is now writing a thesis on the value of cabbage soup in curing uremia. Later he gave another uremia patient cabbage soup and the patient died. Then—in the proofs—the doctor made the following correction: 'Cabbage soup helps in 50% of the cases.'"

"All that may be true," says the doctor. "But the mathematician is some smart aleck too. When asked how to catch a lion in the desert, he answers: 'What does 'catch a lion' mean? It means isolating the lion from yourself by means of bars. I get into a cage and, by definition, the lion is caught!'"

I'm sorry if I have hurt any feelings but it seems to me that is roughly the way the various sciences regard plausibility of reasoning.

In many cases it is not the blame of the specialists but merely their sorry plight. The difficulties of many natural sciences and the humanities are at times so great that it is almost impossible to set up repeatable and specially devised experiments. And so one has to do with the available data. True, the situation is not so hopeless as one might think. It is often possible to enhance the degree of conviction and plausibility of arguments. To do this, we must learn to reason in a plausible fashion.

The foregoing examples have shown that induction may result in error, but this does not always occur, otherwise we would simply have to give up. All I want

to show is that in mathematics too we make just as extensive use of induction and analogy, experimentation and observation as is made in other sciences.

INDUCTION AND MATHEMATICAL INDUCTION*

Induction is the process of eliciting general laws via observation and the correlation of particular instances. All sciences, including mathematics, make use of the induction method. Now, mathematical induction is applied only by mathematicians in the proof of theorems of a particular kind. There is hardly any logical connection between these methods. However, there is a rather widespread terminological confusion in this case. We shall come back to this question again. But still there is a certain practical connection between induction and mathematical induction, and so let us illustrate both methods by one and the same example.

Noting that in the left-hand member of the equation

$$1 + 8 + 27 + 64 = 100$$

are cubes of successive natural numbers and the right-hand member is a square, we can rewrite the equation in the following interesting form:

$$1^3 + 2^3 + 3^3 + 4^3 = 10^2$$

Some readers may not find anything exciting about the fact that a sum of cubes is equal to the square of a number. But there certainly is!

A few years ago, at a lecture on the theory of analytic functions I was delivering to post-graduate engineers I remarked about one of the theorems: "Note this unexpected and remarkable fact." One of the engineers,

* In this section we have borrowed a good deal from G. Polya's book *How to Solve It*.

who looked bored, said "I don't see anything remarkable about it at all". It was now my turn to be surprised. A scientific worker should be excited and glad about any unexpected facts or turns of thought, otherwise he is not a scientist but a mere hack. Curiosity, a healthy curiosity, and a desire to learn is what leads the scientist from one problem to another. Once that feeling is lost and there is no excitement or pleasure in learning new facts, the scientist will no longer be able to discover anything new.

A well-known Soviet physicist once said jokingly that scientific work is a way of satisfying the curiosity of the scientist at the expense of the state. In the same way, we might say that acting is a way of satisfying the actor's vanity at the expense of the audience. For society, it is of course important that this work—whether science or art—be useful, ultimately, to other people as well.

I hope the reader doesn't think I have overdone this. I do not at all think that such curiosity is needed in equal measure in all sciences and is applicable to all problems, and so if you think the problem at hand is not so interesting—it was chosen merely to illustrate the method of induction—then please skip this section.

Now back to the problem. Does it often happen that the sum of the cubes of a succession of numbers is equal to the square of a number? What reason could there be?

When we formulate the question in that way, we are like the natural scientist who is still under the impression of a newly found plant or of a just recognized law in the alternation of rock strata, and who then poses a generalizing question. In our case, this generalizing question is connected with the sum of the cubes of the positive integers

$$1^3 + 2^3 + 3^3 + \dots + n^3$$

We discerned the general law on the basis of a particular instance, $n = 4$. What can we do to resolve the problem? Well, first of all we can do as the natural scientist does: he investigates other particular cases. The particular cases of $n = 2$ and $n = 3$ are simpler than the above. The case of $n = 5$ comes right after the one we considered. For the sake of consistency and completeness, let us also add $n = 1$. Writing down all these equations just as carefully as the geologist would put away his various rock samples, we get the following table:

$$\begin{array}{rcl} 1 & = & 1 = 1^2 \\ 1 + 8 & = & 9 = 3^2 \\ 1 + 8 + 27 & = & 36 = 6^2 \\ 1 + 8 + 27 + 64 & = & 100 = 10^2 \\ 1 + 8 + 27 + 64 + 125 & = & 225 = 15^2 \end{array}$$

It is indeed hard to believe that all these sums of numbers of successive cubes accidentally constitute squares. In such a situation, the natural scientist would hardly doubt that the observations so far suggest a general law. The general regularity is almost proved by induction. The mathematician of course thinks likewise, but he is more restrained. He would say that induction strongly suggests the following theorem:

The sum of the first n cubes is a square.

Thus, we arrive at the assumption of the existence of a remarkable and somewhat mysterious regularity. Why must the sums of the numbers of successive cubes be squares? Actually, they are, as you can see.

What would the natural scientist do at this point? He would continue to investigate his conjecture. He would carry forth his investigation in different directions and would accumulate experimental findings.

If we were to take that approach, we would have to verify the next cases: $n = 6$, $n = 7$, ...

The scientist might again investigate the facts that led him to this assumption. He might carefully compare them in an attempt to find some more profound regularity or some other supplementary analogies. We shall carry on with our investigation in the same vein.

Let us go back to our table and again examine the cases for $n = 1, 2, 3, 4, 5$. Why do the sums of these cubes turn out to be squares? What can we say about these squares? The bases of these squares are 1, 3, 6, 10, 15. What can we say about them? Is there any more profound regularity and any other analogies? It would appear, at any rate, that the way they increase is subject to some kind of law. How do they increase? It turns out that the difference between two successive bases also increases. Indeed,

$$3 - 1 = 2, \quad 6 - 3 = 3, \quad 10 - 6 = 4, \quad 15 - 10 = 5$$

The regularity in the increase of these differences is obvious at a glance, and we notice a similarity in the bases of the squares. After trying out a variety of cases, we dwell on the convincing regularity of the number sequence 1, 3, 6, 10, 15:

$$\begin{aligned}1 &= 1, \\3 &= 1 + 2, \\6 &= 1 + 2 + 3, \\10 &= 1 + 2 + 3 + 4, \\15 &= 1 + 2 + 3 + 4 + 5\end{aligned}$$

If this regularity is of a general nature (and it is hard to think otherwise), then the theorem which we assume to be valid takes on a more exact form, namely, for $n = 1, 2, 3, \dots$

$$1^3 + 2^3 + 3^3 + \dots + n^3 = (1 + 2 + 3 + \dots + n)^2$$

I will not go into any further details of the reasoning that must follow and will give the reader the final formula at once:

$$1^3 + 2^3 + 3^3 + \dots + n^3 = \left[\frac{n(n+1)}{2} \right]^2$$

If the reader is acquainted with the method of complete mathematical induction, he will have no trouble in proving the theorem stated above.

This law was discovered with the aid of induction. The whole course of our reasoning was somewhat one-sided and imperfect but at any rate plausible, and it gives the reader some idea of what this method is about. Induction is aimed at revealing regularities and relationships that are hidden behind the outer aspects of the phenomena under study. Its most common tools are generalization, specialization, and analogy. Generalization arises from an attempt to grasp the significance of observed facts and is then verified by further particular cases.

Such inductive reasoning, related however to much more substantive material and requiring quick wits, conjectures, analogies, is what goes to make up the working method of the mathematician.

THE DRAMA OF EQUATION SOLVING— AN HISTORICAL SKETCH

The fact that inductive arguments and analogies do not always by far lead to correct conclusions is well known. Recall the problem—already mentioned—of solving algebraic equations. During 300 years, right up to the start of the 19th century, mathematicians attempted to find formulas for solving algebraic equations of degree higher than the fourth: for example, the general quintic, or equation of the fifth degree,

$$x^5 + a_1x^4 + a_2x^3 + a_3x^2 + a_4x + a_5 = 0$$

where a_1, a_2, a_3, a_4, a_5 are arbitrary numerical coeffici-

ents. They sought a formula capable of expressing the root of this equation in terms of its coefficients by means of arithmetical operations: addition, subtraction, multiplication, division and the extraction of roots. It was precisely induction that compelled them to work in this direction. Formulas had already been found for equations up to degree four, although at times the going had been difficult. What is more, as Gauss had demonstrated, an algebraic equation always has roots, and the number of roots is always exactly equal to the degree of the equation. It required the genius of Abel and Galois to resolve this problem in its entirety.

At the beginning of the 19th century, a young Norwegian mathematician, Niels Henrik Abel, took up this problem. At first he thought he had found the solution of the quintic (an equation of the fifth degree). However, he was soon disappointed when he found a mistake in his calculations. He meditated long and painstakingly on this problem and finally came to the conviction that equations of degree higher than the fourth cannot, generally speaking, be solved by radicals. Abel demonstrated this assertion and his theorem became a turning point in the problem of equation solving. Abel became one of the most honoured names in mathematics. His papers on mathematical analysis are profound and diversified. Although during his lifetime Abel was recognized by the greatest European mathematicians, he died in poverty from tuberculosis at the age of twenty-seven.*

* O. Ore has written an interesting book of the tragic life of Abel entitled *Niels Henrik Abel*.

No less exciting are the following books: *Whom the Gods Love. The Story of Évariste Galois* by L. Infeld, and *Évariste Galois—Revolutionary and Geometer* by A. Dalmas and an epilogue by A. Yaglom, which has served as a basis for much of what follows.

At about this time, another young mathematician, Evariste Galois of France, "found" the solution to the fifth-degree equation. Like Abel, he was terribly upset when he detected an error in his reasoning. He too found the strength to continue his work.

There is no space here to give in detail the remarkable life story of this great French mathematician. But a few words simply must be said. Everything was unexpected in the brief but turbulent life of Evariste Galois. His mania for mathematics, his active participation in the political life of his day, failure at the entrance examinations in mathematics at the École Polytechnique and his later expulsion from the École Normale for political reasons, his later arrest and a jail sentence, and finally his death in a duel at the age of twenty. And yet Galois wrought a veritable evolution in science. The fate of his papers is likewise extraordinary. During his lifetime they failed to cause even a ripple of interest and were straightway forgotten after his death. Only half a century later were they rediscovered, and they exerted a tremendous effect on the development of mathematics. Galois' total output—what wasn't destroyed or lost—comes to only about sixty small sheets of paper. But their study requires great effort because Galois detested unwieldy computations and therefore gave extremely succinct statements.

In the problem of the solution of algebraic equations Evariste Galois took a new approach. To solve an equation means to find the roots. Galois made a study of the most general case of an equation of arbitrary degree. It will be noted that in practical situations nobody ever needs an exact solution of any concrete equation: mathematicians must only indicate methods for approximate evaluation of the roots. These approximate values are quite satisfactory for the needs of

physicists, chemists and engineers. I have already mentioned the fact that today we can obtain arbitrarily exact results by using computers. But general equations with literal coefficients cannot be studied by approximate methods.

You can write down a general algebraic equation and denote its roots by letters. The roots of course remain unknown. The first of Galois's discoveries was that he reduced the degree of indeterminacy of the values and established certain general relationships that the roots obey. Here is an instance: one root is a definite function of two other roots.

However, the name of Galois became famous not for the concrete results of solutions of higher-degree equations by radicals but for the general methods that he created for studying the properties of equations. The great contribution of Galois—the founder of modern higher algebra and one of the creators of modern mathematics as a whole—is the use, in the solution of a concrete problem, of the general concept of what is known as a *group*.

In mathematics, a group is a set of elements of any nature whatsoever for which a definite operation (called the group operation) is defined. This operation associates every two elements of a group, say, elements a and b —with a third element, their sum $a + b$. This process involves the execution of only a few operations similar to the rules of arithmetic. For example, the associative law holds: for any three elements a , b , c of a group, $(a + b) + c = a + (b + c)$, and sometimes also the commutative law: $a + b = b + a$.

To the average person, the customary rules of arithmetic always hold true and mathematicians must be wasting their valuable time on quite obvious things. Actually, the obvious here is not always true.

Indeed, our elements a , b , c , are of any nature whatsoever, and the group operation, say the operation of addition, need only be defined on the set of these elements and must satisfy the axioms. If you are given a pain-killing injection and then a tooth is extracted, you will say that that procedure differs substantially from what we would have if the order had been changed (the tooth extracted before the injection). Thus, as you readily see, the customary rule "the order in which numbers may be added is immaterial" does not hold true.

A group may consist of numbers, functions, rotations, or other motions. Actually, it is more convenient to study abstract groups whose elements are mathematical symbols the meaning of which is not specified at the moment. It is precisely this extraordinary generality of the concept of a group that constitutes its chief value. In mathematics proper and in its applications, and also in a great diversity of problems encountered in other sciences, it is convenient to utilize the fact that the entities under study form a group. This makes it possible to bring together and study areas of mathematical science that had earlier appeared to be totally unrelated.

An important example of a group is the so-called permutation group. The students in a class occupy definite seats. If some of the students (say Katya and Sergei and Ira and Alyosha) change places we have a reshuffling which mathematicians term a permutation. In the process, some of the students may remain where they are seated. The sum of two permutations (reseatings) is naturally called a permutation that arises from a successive reseating of the students in one way and then in some other way. With this definition of the concept of a "sum of permutations", the permutations themselves form a group.

This example may be developed further. The students of the given class may be distinguished in various ways: there are boys and girls, advanced students and failing students, undisciplined and disciplined ones, near-sighted and far-sighted, etc. When the students are reseated, these differences impose definite limitations on their arrangements. For example, near-sighted students require seats in the front rows; undisciplined students should be separated, etc. The set of permutations (reseatings that satisfy such requirements) form a certain "permutation group". It is closely tied in with the specific composition of students in the given class: in another class we would as a rule have a different permutation group. Somewhat simplifying the picture, we might call such a permutation group the "Galois group" of the class.

In his study of the properties of sets of equations, Galois operated in a similar manner. Instead of the students of a class, he considered the roots of a definite algebraic equation. The roots are connected by certain kinds of algebraic relations (for instance, one root may be equal to the sum of two others). Galois associated with each equation a permutation group of its roots—the group consisting of all permutations that obey the relations existing between the roots. One can then study the group and learn a great deal about the roots themselves. It turns out that when the Galois group of an algebraic equation has certain definite and readily verifiable properties (such groups are called solvable groups), the equation turns out to be solvable by radicals, which means its roots can be expressed in terms of the coefficients by means of explicit algebraic formulas involving solely the signs of addition, subtraction, multiplication, division, raising to a power and root extraction. Otherwise the equation is not solvable in that way. Consequently, to

decide the question of the solvability of a given equation in radicals, one must form its Galois group and check to see if it is solvable or not.

Thus it was that Galois gave the complete answer to a question that had agonized science for so long: when is an algebraic equation solvable in terms of radicals?

After Galois' works were rediscovered in the second half of the 19th century, the new methods began making inroads into all spheres of mathematics. At the present time, the concept of a group, like that of a number, a set, a function and a transformation, is one of the most fundamental in all modern mathematics.

Differential equations play a large role in mathematics. We have already mentioned them. Their solution and the study of their properties are more difficult still than the study of algebraic equations. Following the Galois pattern, we can associate with every differential equation a group similar to the Galois group of an algebraic equation. This method, suggested by the Norwegian mathematician Sophus Lie, makes it possible to study extremely important and profound properties of differential equations.

The introduction into geometry of the group concept substantially altered this branch of mathematics. In 1872 the famous German mathematician Felix Klein correlated a specific group with each division of geometry and proclaimed the basic task of geometry to be the study of the properties of corresponding groups.

The ideas of Klein and Lie later proved to be extremely fruitful for the most far-flung branches of mathematics and mathematical physics, and particularly for modern quantum physics.

To this day, the mathematical apparatus of group theory is one of the basic tools of theoretical physics.

Engineer Consults Mathematician

As I have already said before, meetings between mathematicians and specialists in other fields enrich both parties. Besides, their joint work often yields a very perceptible economic or production effect, or at least points the way for substantial savings. I recently had just such an encounter with an oil-refining engineer. (As usual, I am the mathematician in these conversations.)

Engineer. I would like your help in constructing a mathematical description of the process of primary oil refining.

Mathematician. That's a complicated problem, isn't it? From what I gather, the process is very intricate.

Eng. Yes, it is.

Math. Could you give me a rough outline of the process?

Eng. The raw oil is fed to an electric salting-out plant (ESP) where a considerable proportion of the salt is eliminated. From there, the oil goes to the first column where a heating system skims off the float fractions. The remainder is then heated again,

and the skimming process continues with more fractions removed. This process of skimming off definite fractions of the oil is repeated several times.

[The engineer then showed the mathematician a block diagram of the process or, rather, a scheme of partial automatization. The diagram was made to a small scale and still covered a broad sheet of paper one metre wide and 5 metres long. It was quite impossible to figure anything out. So a few days later the mathematician took a trip to an oil-refining plant, where he saw enormous 10-metre diameter spheres covered with concrete. These were the ESP—there were also 30- to 40-metre rectification columns, gas furnaces operating at hundreds of degrees Celsius, operator rooms with dozens of instrument panels recording pressures, temperatures, and other important parameters of the process. The distance from one plant to the next was in hundreds of metres. The mathematician had already seen such complexes in the movies, but reality made a far greater impression.]

Math. How many control parameters are there that determine the process?

Eng. I can't say exactly, I'll have to count them. At any rate, something in the vicinity of a hundred. But there are a few that do not require control, that is, they do not vary during the process. All that is needed is to maintain their values within certain limits.

Math. And what's the result?

Eng. A variety of fractions ranging from light gasolines to various oils.

Math. So what is it you want?

Eng. We would like to prepare a mathematical description of the process.

Math. What for?

Eng. We want to be able to control the process.

Math. But you already seem to be doing that.

Eng. Yes, that's true. But we don't do it very well. We barely keep it to normal, whereas if we were able to improve the process by as much as one per cent via optimization of control, the effect would be tremendous.

Math. You mean to say you're now operating more or less by rule of thumb.

Eng. Well, not exactly.

Math. A good many of the control problems of the process are handled by the operator on the basis of his own personal experience and intuition, is that right?

Eng. Yes, basically that's the way it is.

Math. What is the operator after? Or are there several operators?

Eng. There are several operators and each one attempts to vary the quantities under his control so that the process proceeds within a specific range.

Math. Another thing, is the initial raw material—petroleum—homogeneous in composition or does the operator have to keep tabs on the composition as well?

Eng. Petroleum consists of hundreds of hydrocarbons, and their percentage content in the petroleum varies perceptibly. But over a long period of time, the plant receives homogeneous petroleum or takes special measures to make it homogeneous if the composition differs.

Math. Does that mean that to a first approximation the composition of crude oil may be taken to be constant?

Eng. Yes, I suppose so.

Math. And yet there are certain varying characteristics of the starting product that have to be taken into account?

Eng. Yes, of course. For instance, the temperature of crude oil pays a very important role. That is why it is specially heated prior to delivery to the rectification column.

Math. Fine. Now how many parameters of the crude oil does the technologist deal with? What I mean is the parameters that require a change in the operating conditions.

Eng. Besides the temperature, he has to take into account the consumption, that is, the quantity of petroleum delivered per minute. It is also sometimes necessary to take into consideration the pressure.

Math. To summarize, then, we will assume that the original product, that is to say, the input of the system of primary oil refining, can be described by three varying parameters, or three numbers. Now let us describe the result of the refining process in similar terms.

Eng. To enumerate them, we can say that as a result of primary refining we obtain a variety of gasolines, jet fuel, diesel fuel, gas oil, various other oils, and oil tar. About ten quantities.

Math. Can we describe the properties of each one by a single number?

Eng. Hardly! Each of the components requires at least a few numbers for its description. Say, the quality of gasoline is defined by the octane number, the fraction composition, and the density.

Math. That makes things worse. How many variables (numerical quantities) must be specified so as to give a basic description of all output products? In other words, what characteristics of the quality of the output components may be termed essential?

Eng. It would take a long time to go through all these characteristics. There are a great many. But I think for a start we could take roughly 20 numbers.

Math. Now let us picture the situation. We have a process that can be described by three variables at the input, by twenty variables at the output, and by a hundred parameters subject to control. What is the problem now confronting us?

Eng. We need to describe the process mathematically.

Math. That's a tall order! You want to construct a mathematical model of an extremely complicated process. Just what do you think such a description will amount to? Writing down a system of equations relating all 130 variables?

Eng. Yes, that would be desirable.

Math. But do you know the relationships that obtain between all the variables?

Eng. Qualitatively I do.

Math. What do you mean by qualitatively? For example, how the specific weight of the clear fractions depends on the flash point?

Eng. No, we do not know such particulars. By qualitatively I mean to say that the flash point increases with the specific weight of raw oil.

Math. But how can you write a system of equations when the relations between the variables are not known?

Eng. If I knew the answer to that question I wouldn't be consulting you.

Math. You're confusing me with the Almighty. I am not God. But suppose that in some fantastic manner I deciphered these relationships and was then able to write down the equations. What would that give us?

Eng. Simply that if the equations were at hand, we would use them to figure out a system of optimal control.

Math. That's clear, but what do we have to optimize?

Eng. What I just said—control.

Math. No, control is a purposeful choice of values of the controlling parameters. The thing to optimize is the output. What characteristic of the output do you need to optimize?

Eng. It varies. Sometimes we want to optimize one characteristic, at other times, several.

Math. When you want a maximum of one characteristic that is more or less understandable, though not entirely. We could suggest say optimizing the output of diesel fuel and not impose any restrictions (demands) on the other components: just take what comes in that respect.

Eng. Yes, that's what is sometimes done.

Math: But then there may be different percentage ratios of the other components. What is more desirable? What is best?

Eng. The requirements vary from time to time. But generally speaking, hardly any technologist would be able to answer that question.

Math. But without answers to these questions, it is impossible even to formulate the problem of optimization. We will have to go into the situation in more detail.

WHAT IS BETTER?

What is better, to be rich and healthy or poor and ill? That's a joke. But what really is better: to be rich and ill or poor and healthy? It is impossible to give an answer at once. We have to agree on the actual content of the concepts of rich and poor, ill and healthy. But then comes the still more complicated question: what does "better" mean?

In Russian there is a phrase in frequent use which means "for the purposes of, say, improving, ...". The word "purpose" used here (in Russian) in the plural is incorrect, though many would argue that the grammar is all there, so to say. The point is that there cannot be several purposes all at once. You may not agree with me. You may say that it is possible to reach

new heights in science and sport or to overfulfil a plan involving a whole range of high indices. I will try to demonstrate a certain inconsistency in such a statement of the problem.

Let us begin with an evaluation of plan fulfilment. Suppose we have two identical factories, say, "Volga" and "Desna", producing men's and women's bicycles. The plans are the same: 900 men's and 600 women's bicycles per month. In the current month, the Volga factory turned out 1000 men's and 550 women's bicycles whereas the Desna factory produced 800 of each. The Volga factory overfulfilled the plan for men's bicycles and underfulfilled the plan for women's bicycles, the Desna factory did just the opposite, as may be seen from the accompanying table.

Let us calculate the output in terms of vehicles. The plan called for $900 + 600 = 1500$ bicycles per month. In reality, the Volga factory turned out $1000 + 550 = 1550$ items, the Desna factory, $800 + 800 = 1600$ bicycles. To summarize, both factories

"Volga"

Bicycles	Quantity		Gross output		Generalized index
	Plan	Output	Plan in rubles	Actual output	
Men's	900	1000		$1000 \times 100 =$ 100 000	
Women's	600	550	144,000	$550 \times 90 =$ 49,500	A = 94.5
Total	1500	1550		149,500	

“Desna”

Bicycles	Quantity		Gross output		Generalized index
	Plan	Output	Plan in rubles	Actual output	
Men's	900	800		$800 \times 100 =$ 80,000	
Women's	600	800	144,000	$800 \times 90 =$ 72,000	$\Lambda = 94.0$
Total	1500	1600		152,000	

overfulfilled the plan as to overall quantity, and yet both failed to fulfil it as to nomenclature. But the underfulfilment was according to different indices. Which factory worked better? It is clear that in a situation where we have a multitude of indices (not two) for assessing the work and plan fulfilment of a factory, say, quality, wage bill, economy of materials, etc., the question of choosing the best factory becomes still more involved.

This contradiction can be resolved in only one way: a generalized index has to be thought up that will describe the functioning of the plant and that can be used for comparison. For instance, factories can be compared on the basis of overall output. Then the monthly plan should be specified in rubles. Suppose a men's bicycle costs 100 rubles, and a women's bicycle costs 90, then we specify the monthly plan for the two factories as the following sum:

$$900 \times 100 + 600 \times 90 = 144,000 \text{ rubles}$$

Now there is no need to demand that the factories fulfil their plans as to nomenclature. In our example, the total output of the Volga factory came to $1000 \times 100 - 1 - 550 \times 90 = 149,500$ rubles, that of the Desna factory, to $800 \times 100 + 800 \times 90 = 152,000$ rubles.

To summarize, then, the Desna factory won out in overall output and, consequently, we can say it made a better showing, although both factories overfulfilled the plan in total output.

It might also be possible to assess the work according to some other general index that would take into account the nomenclature as well.

For example, in order to stimulate fulfilment of the plan as to nomenclature with account taken of overall output, we could take a composite index to describe the work of the factory, $A = D \times n$, where D is fulfilment of plan as to total output (in per cent). The quantity n is then defined as:

- (1) if the plan for nomenclature is fulfilled, then $n = 1$;
- (2) if the plan for nomenclature is fulfilled for all except one type of goods, then $n = \frac{n_1}{m}$, where m is the number of units of a given type according to the plan and n_1 is the actual output of that item;
- (3) if the plan for nomenclature is underfulfilled with respect to both types of goods, then $n = \frac{e_1}{e}$, where e is the number of units of both types according to the plan and e_1 is the number of actually produced items.

If the plan is fulfilled, $A = 100$; for overfulfilment, $A > 100$. In our case, a simple calculation yields

Volga, $A = 94.5$,

Desna, $A = 94.0$

According to this generalized index, the Volga factory did better than the Desna factory, although both failed to fulfil their plans.

Clearly, there can be any number of such indices and every time the results will differ.

How can we figure out a method for choosing the best (most advantageous) index?

Problems like this arise all the time. The Polish writer Anatoli Potemkowski has a delightful short story entitled "Lift" in which he gives a marvellous picture of the difficulties that can arise when choosing a criterion.

"Pan Zalzanewicz wrote an application to the house committee complaining that, strange as it may seem, he has to pay just as much for the lift as Pan Pataszonski, although Pan Pataszonski lives on the 13th floor, whereas he, Pan Zalzanewicz, lives on the second.

"We decided to examine this complaint right on the spot, all the more so since Pan Zalzanewicz's claims appeared to be quite justified.

"The higher one lives, the more he has to pay," remarked Pan Kuca. "Let's make up a table of the tenants of the house."

"But you have to take into account the size of the family," added Pan Zyzia. "The basic criterion will be man-floors."

"Kukuliak always goes up in the lift with his wife," said the baron's wife. "They are always together. That's two people, but the lift goes up only once."

"Let's introduce a coefficient of familial sentiment," suggested Pan Kuca.

"Then there is the weight that has to be considered," put in the baron's wife. "Pan Pataszonski and his wife weigh less than Pzeradska all by herself."

"True enough," agreed Kuca. "We'll have to take into account the total weight of each family."

"In summer or in winter?" That question was raised by Pan Zyzia, who added, "In winter a lot of the people wear heavy coats."

"All right then, let's introduce a monthly system of weighing the tenants of this house, because some may be gaining weight while others may be dieting."

"It would seem that we were close to a reasonable solution, but then somebody recalled the problem of guests.

"Let's introduce a coefficient for guests," said Kuca.

"But there are different kinds of guests," remarked Zyzia. Some go to the second floor but by mistake get out on the third floor. That means going down one floor to the second. So instead of one floor he travels three. The pay should be higher for stupid guests."

"True again, we need a coefficient for the level of intelligence of the guests," Kuca concluded.

"Let's not forget about their weight either," put in Zyzia. "An intelligent heavyweight can turn out to be more costly than a thin idiot."

"We'll have to give all this some serious thought," said the wife of the baron.

"After a detailed analysis, we dropped in to see Zalzanewicz, and each one of us (in the name of all the rest) gave him a good drubbing. What was the big idea of putting us to so much trouble, after all!"

"Then we all went up in the lift to the thirteenth floor and had dinner with the Pataszonskis."

The very statement of the problem is meaningless, since no reasonable answer can be given for all possible cases that may be encountered. It is something like choosing the best mode of transportation—train, airplane, boat, or donkey. It clearly depends on the circumstances. For a trip from Moscow to Novosibirsk, travel by air would be best, at least as far as time goes. A trip into the suburbs of Moscow would clearly indi-

cate train transport, whereas a boat might do better for a honeymoon, and a donkey might be the only mode of transportation in a mountainous country.

Thus, the answer to such a question depends on the problem and on the situation.

CRITERIA

In the foregoing example, the choice of the best type of transportation depended on the situation. For instance, a honeymooning couple would most likely choose a comfortable boat trip since time would not be a factor and the pleasant, romantic surrounds and changing scenery would probably be attractive. However, it is difficult in this case to assess quantitatively the advantages of a boat trip over those of a train trip.

In most engineering problems, particularly when solving optimization problems, one must have the opportunity of comparing different variants quantitatively. It is therefore important to be able to state a clear-cut quantitative criterion.

Recall the problem of getting from point *A* to point *B* when the straight-line route is through mud. If we want to avoid puddles, then we can formulate the problem thus: of all routes connecting *A* and *B* and not passing through puddles, determine the shortest.

Here the length of the route is the criterion used for comparing routes. The problem could have been posed differently: of all routes connecting *A* and *B*, find the route which can be covered in the shortest time.

The criterion for comparing the routes in this case will be different: it is the time required for getting from point *A* to point *B*.

It may turn out that in solving these two problems the optimum route will be the same, say, the dotted line in Fig. 67. But the problems are not the same.



Fig. 67

First of all, the original supply of routes from among which the best is chosen differs: when minimizing time, we have all the routes between A and B ; when minimizing distance, we have only those routes which do not pass through puddles.

What is more, the foregoing problems allow for a possible variety of solutions. For instance, if the second puddle is narrow enough, then when solving the problem as stated in the second case, we can choose the route indicated by the dashed line. When using

this route, all one has to do is step over the puddle and then it will be shorter than the dotted-line route.

Let us now return to the problem of the turbine drilling of wells. We posed the problem as follows: to determine the highest rate of sinking the well.

But why strive for the highest possible rate? The answer appears to be obvious: the higher the rate of advance, the sooner the well will go into operation, and that means extra hundreds of tons of oil per day.

However, the obvious is erroneous here.

Indeed, the attempt must be made to drill the well as soon as possible, that is, to spend as little time on drilling as possible. But this does not necessarily mean drilling with the highest rates.

All other conditions being equal, the higher the rotational speed of the turbo-drill, the faster the boring tool wears out and, consequently, the more often it has to be changed. Now changing the boring tool requires bringing up the entire column of steel pipes from a depth of several kilometres, which in turn requires time. "Slow but sure"—in strict keeping with the saying.

So we have to alter the statement of the problem: to determine the amount of pressure on the bottom of the well and the values of other essential quantities for which the entire well can be sunk in the shortest time interval.

The statement is changed and we have a different criterion. The criterion now for quality of drilling is the time spent on sinking the whole well and, hence, the problem consists in minimizing that time. In the sense of the new criterion, the optimal rate of advance turns out to be less than the maximal possible rate.

OPTIMIZATION

In the foregoing examples, the value of the criterion was determined by a single number.

The problem of optimization according to the given criterion thus reduces to seeking the objects (routes, values of parameters, etc.) over which the value of the criterion reaches an extremum. The question arises: is it not possible to think up a criterion whose value is specified by two quantities at once?

Is it not possible—in contrast to the comparison of the two factories that we gave earlier—to compare the functioning of factories both as to total output and as to nomenclature? All the more so, since it is possible to find the extremum not only of a function of one variable but also of a function of two variables.

This perplexing question is expressed quite often, either explicitly or implicitly. Actually, we are dealing here with different problems. Imagine two children on a seesaw (Fig. 68). When one goes down, the other goes up. Each one of the kids wants to be on top, and the one down below cries until he goes upwards. But there is no way of both occupying the top position at once.

The sum of their distances to the ground is a constant equal to twice the distance of the middle of the board from the ground. This means that their distances from the ground are not independent; they are related in such a way that the sum of the distances is a constant. Therefore, naturally, if one child goes up, the other has to go down.

It is hard to explain to children (and adults too, by the way) that moderation represents the optimum form of behaviour,* and the most that can be attained by both at once is to reach the same height (Fig. 69).

* As the poet Zhukovsky put it, "Moderation is the best feast."

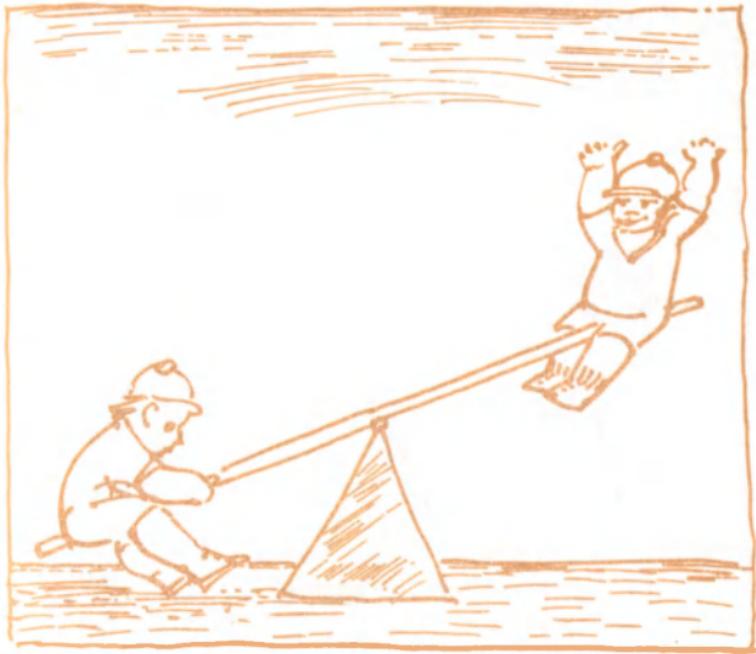


Fig. 68

Neither will be particularly pleased but the crying may stop.

We can summarize. When it is merely a matter of seeking the extremum for a function of several variables, it is assumed that these variables are independent, which means that any one of them can be changed and this will not affect the values of the others in any way. But when we seek the extremum of nonindependent variables, then the relationships that obtain between them must be taken into account. We then have to do with the concept of a conditional extremum.

Finally, when we discuss a quantitative criterion for the comparison of certain objects, then the criterion must always be expressed by one variable only. Its

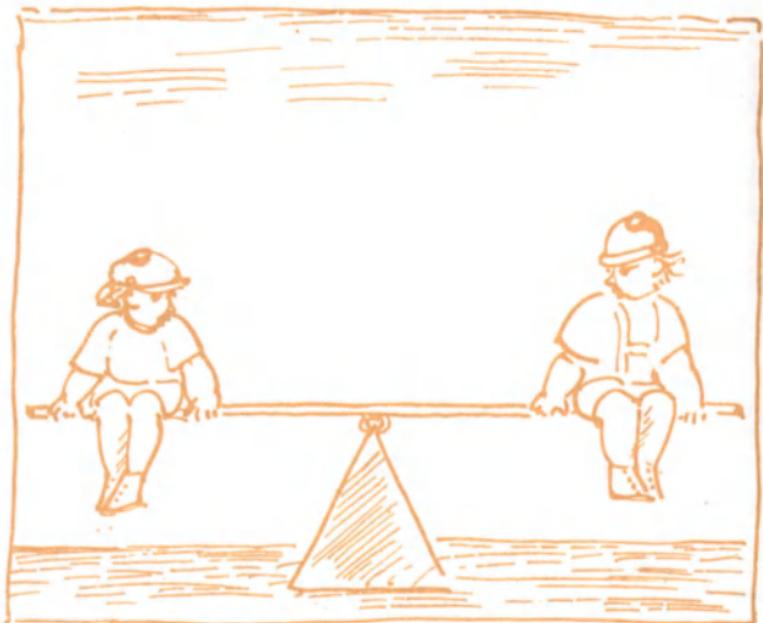


Fig. 69

values describe different entities and the comparison of the entities is via a correlation of the values of the criterion for these entities.

If the values of the criterion for two entities are the same, then from the viewpoint of their classification by means of this criterion the entities are indistinguishable. Only in this case is it meaningful to discuss the problem of choosing the optimal object (entity). Optimality is then understood thus: the criterion attains an extremum for the chosen object.

HOW CLOSE?

It is hard to introduce a quantitative criterion for measuring the degree of intimacy in human affairs, both cultural and the other kind, about which teenagers under sixteen are kept away from the movies.

What I want to discuss now is that aspect of proximity, or closeness, that permits introducing a quantitative measure. This is an extremely important concept.

We say geological eras are close, cities are close. What does "close" mean?

Suppose the distance between two towns is 200 kilometres. Are they close to each other or far away? What about geological eras spaced 200 million years apart? Are they close or distant? True, the careful reader might ask a counter question: close in comparison with what?

Geological eras are measured in millions of years, and if the time between the eras is less than their characteristic dimension, then we can speak of closely lying eras.

Distances between cities are measured in tens, hundreds and thousands of kilometres. The distance from Moscow to Leningrad is great when compared with that from Moscow to, say, Serpukhov, but small in comparison with the distance between Leningrad and Irkutsk. Hence, the notion of "close-lying cities" is a relative one depending on the situation.

It is clearly evident that in order to assess the closeness of two points on a straight line, on a surface or in space, one has to introduce a measure of the distances between points and indicate a unit of length. Even that is not enough: we must also indicate what we are comparing the closeness with. Sometimes we speak of closeness in comparison with unity, in other cases, closeness in comparison with the distance between some other points.

By this time, the reader will probably have concluded that he has learned nothing from these trivial ideas. I can almost agree. But tell me which of the two lines—dashed and dotted—in Fig. 70 is closer to the horizontal axis?

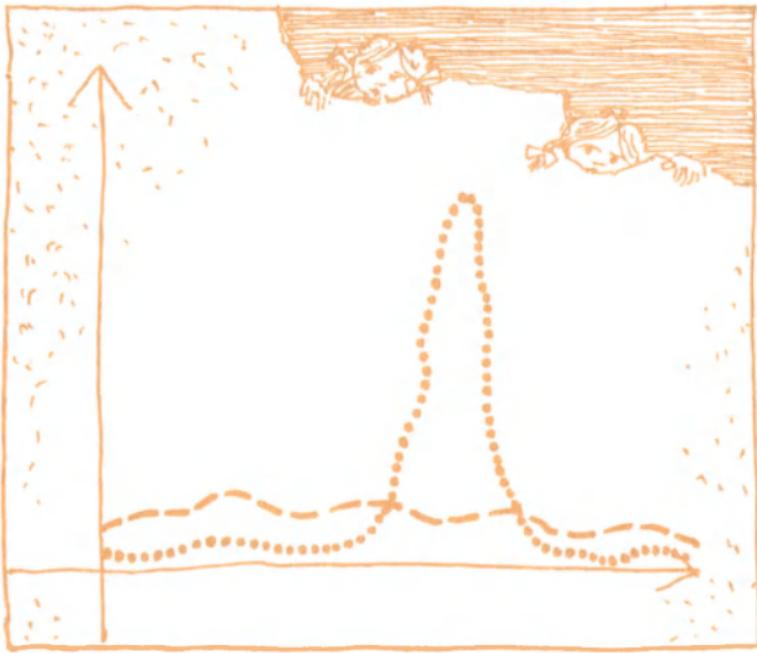


Fig. 70

You have probably found yourself in the difficult position of the parents of twins and only one extra ticket to a football game. One of the boys refused to eat his oatmeal all week, twice refused to brush his teeth, and was constantly biting his nails. The other twin was on his best behaviour all week long, except on Saturday when he hitched up a system of mirrors, pipes and levers and spent the whole evening watching his elder sister and her fiancé. How is one to choose the better twin?

The reader who has gone through the preceding sections of this book will probably smile and say, "These questions have already been answered: we have to introduce a criterion. In our case, this is the criterion of closeness of two curves."

I agree of course—we must think up a criterion. We already know that the criterion depends on the problem. But what criteria for the closeness of curves can we suggest?

MARY AND MAUDE

Two neighbours, Mary and Maude, are close friends living on the same floor. They are very much alike, even their names are practically the same, except for a couple of letters.

Mary and Maude come home in the evening and switch on the lights in their flats. The more energetic

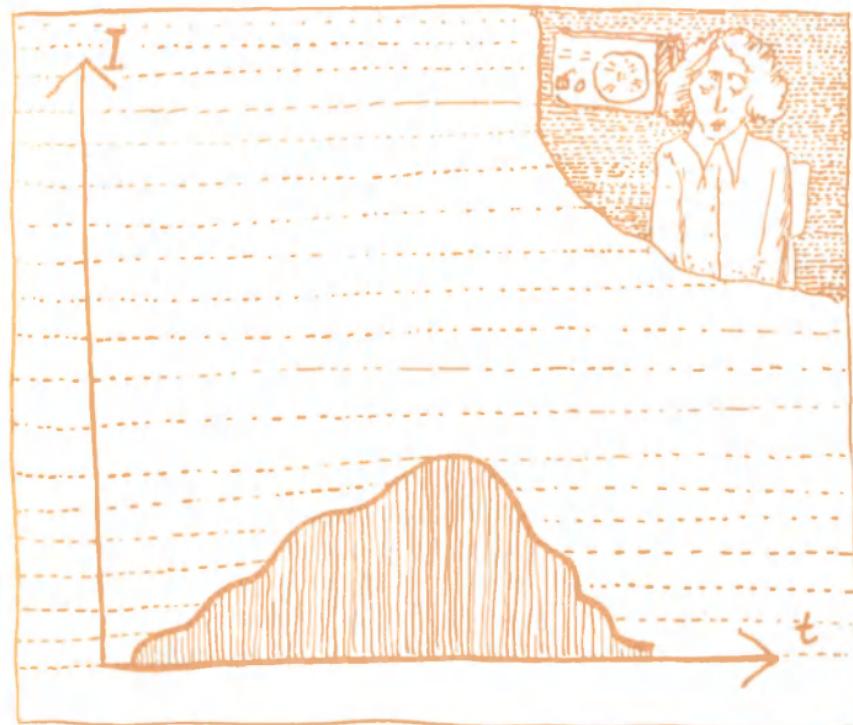


Fig. 71

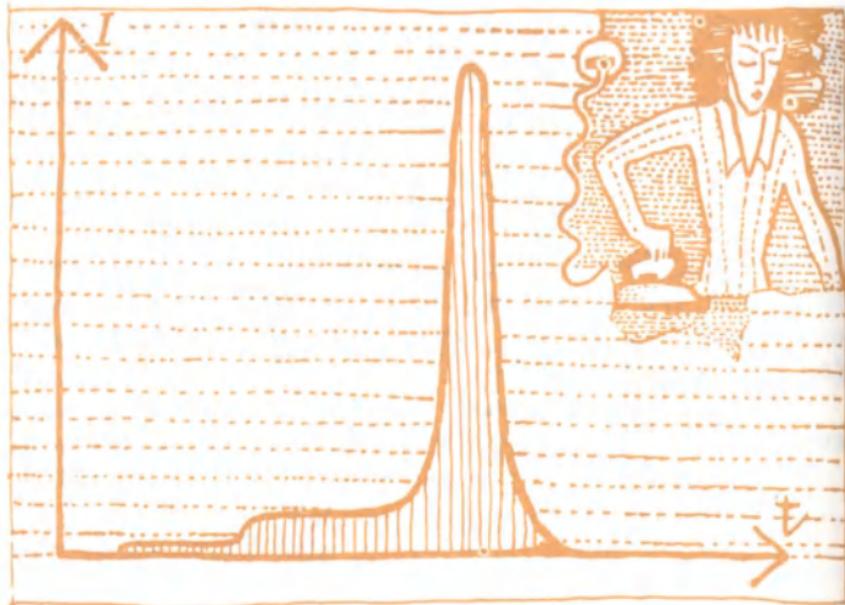


Fig. 72

Mary turns on all the lights, she likes a bright room, a bright kitchen, and keeps them that way all evening. Meanwhile Maude has turned on the flatiron, and blown a fuse.

Now let us take a glance at Figs. 71 and 72, which show how much electric power has been used in their respective flats. The horizontal axis is the time, and the vertical axis shows the power used.

When Maude switched on her defective flatiron, there was a short circuit, the power consumption went way up and a fuse blew, since it was calculated to withstand only a certain amount of electric power. The curve of power consumption then plunged to zero since the current supply came to a stop.

In any evaluation of power consumption, there can be at least two approaches. The readings of the meters

of Mary and Maude varied by quantities that are proportional to the areas under the curves of variation of power consumption. The area under Mary's curve is greater and hence Maude's curve is closer to the horizontal axis (zero line) than Mary's.

But if we assess the curves according to their maximum values, and this precisely is how the fuses react, then Maude's curve exceeds Mary's quite substantially. Using that criterion, we find that Mary's curve is closer to the horizontal axis.

INTEGRALS—NOT SO TERRIBLE AFTER ALL

I mentioned the area under the curve. The careful reader will note that in elementary geometry one determines the areas of figures bounded only by straight-line segments, whereas in our case we have an area bounded by an arbitrary curve. True, school geometry also gives the area of a circle defined directly by means of passing to the limit from the areas of inscribed and circumscribed regular polygons, but hazy reasoning based on a poorly substantiated approach to the limit concept only confuses the issue still more.

Area is a definite numerical characteristic of part of a plane bounded by a curve. To find this characteristic, it is necessary of course to indicate the rules for carrying out the computations, and to justify the rules requires a serious development of the theory of limits.

I shall now try to give the basic ideas and simple facts without resorting to the theory of limits, working only with intuition.

First of all, what we have to do is find the area, say, under the curve in Fig. 73. This area is bounded by the interval $a \leqslant x \leqslant b$ on the x -axis, the graph (curve)

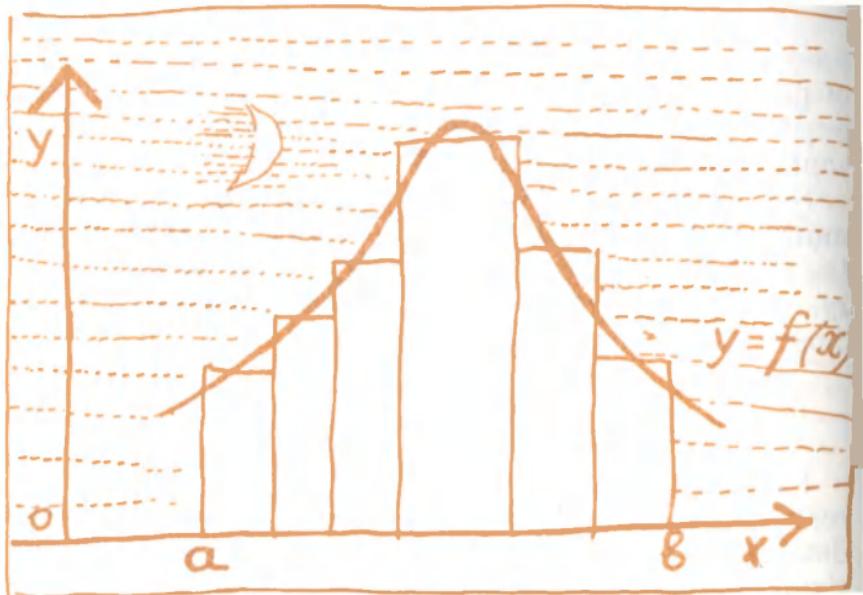


Fig. 73

of the function $y = f(x)$, and by two straight-line segments parallel to the y -axis and passing through the points a and b . The idea of computing the area S of such a curvilinear trapezoid consists in replacing the original curve by an almost identical step-like line, which is shown in the same figure. The area of each of the resulting rectangles is readily calculated, and their sum will be roughly equal to the desired area of the curvilinear trapezoid. The smaller the bases of the rectangles (the number of rectangles will then increase), the closer will the sum of their areas be to the area of the curvilinear trapezoid.

As the rectangles increase in number without bound and as, in the process, their widths decrease, the sum of their areas will approach the desired area. The resulting limiting area will then exactly equal the

area S under the curve $y = f(x)$. It is called the *definite integral of the function $y = f(x)$* on the interval (a, b) and is denoted as follows:

$$S = \int_a^b f(x) dx$$

The integral symbol \int was obtained by an elongation of the letter S , the first letter of the Latin word *summa* (sum). This symbol was introduced by the great Leibnitz, one of the creators of the integral calculus. The other was Sir Isaac Newton, the celebrated English scientist. Almost all the designations that we use today in integral and differential calculus are due to Leibnitz. The letters a and b at the bottom and top of the integral symbol \int indicate the initial and terminal points of the interval within which the area is sought. The symbol dx is not to be interpreted as the product of the letters d and x , but as a single symbol. It is called the *differential* and denotes an increment in the variable itself (the length of the bases of the rectangles).

The reader should not think that he has already mastered the essentials of integral calculus. Far from it! However, this is enough for the present, and we will not need any more integral calculus in what follows.

The integral calculus indicates methods for the approximate computation of areas bounded by intricately curved lines. But if you really have to determine the area of a complicated figure whose graph is known, then it is best to do so somewhat differently. The numerical value of a specific area is always required only to a certain degree of accuracy, say to two or three decimal points. Take a rectangular sheet of paper (the area of the sheet is readily computed by measuring

the sides of the rectangle and multiplying the numbers obtained) and weigh it.

Now draw (to a suitable scale) the figure that interests you, cut it out and also weigh it. Then a few simple manipulations with these numbers will give you the answer. This is a good way to obtain an approximate value of a definite integral. True, the approximation is very rough. If high accuracy is needed, then one resorts to the methods of mathematical analysis and the computation is performed on computing machines.

Before bringing to an end this short discussion of computing areas of plane figures, we may add that it is useful to introduce the concept of a negative area. If the curve—the graph of a function—lies beneath the horizontal axis (the x -axis in Fig. 74), then its area is taken to be negative, which is quite natural: the values of the function $y = f(x)$ are negative in

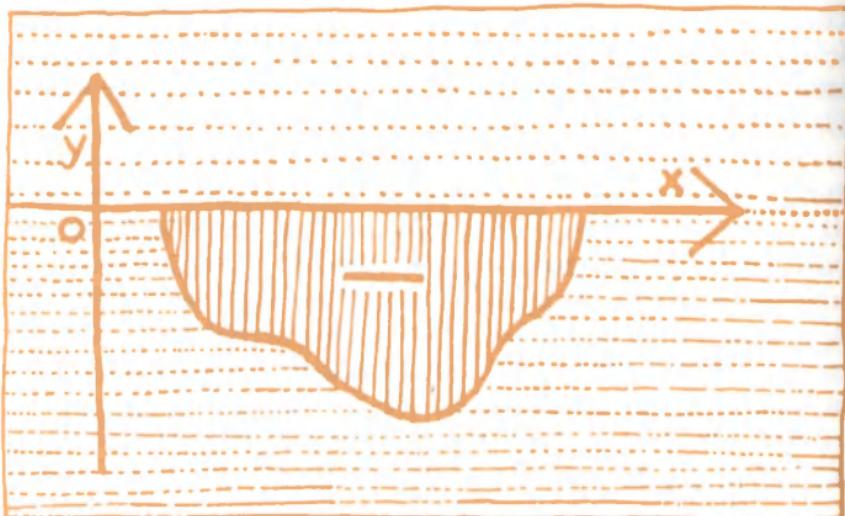


Fig. 74

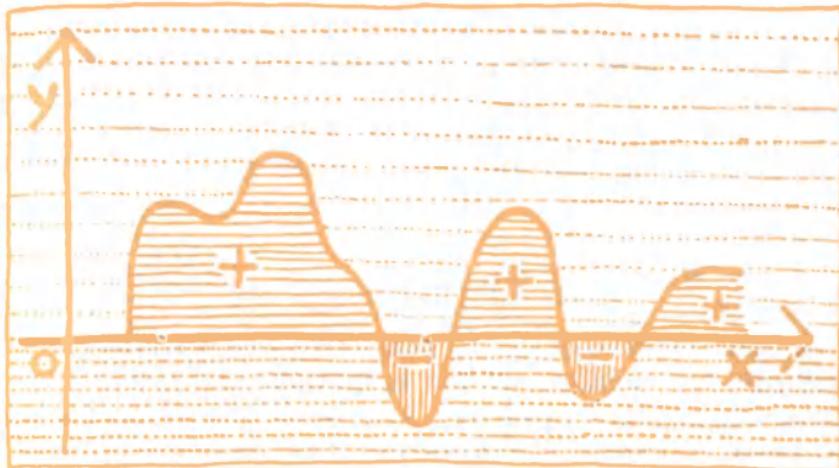


Fig. 75

this case and the base of the curvilinear trapezoid (the length of the line-segment on the horizontal axis) is a positive quantity.

If the curve $y = f(x)$ intersects the x -axis, then parts of the area above the horizontal axis are positive,

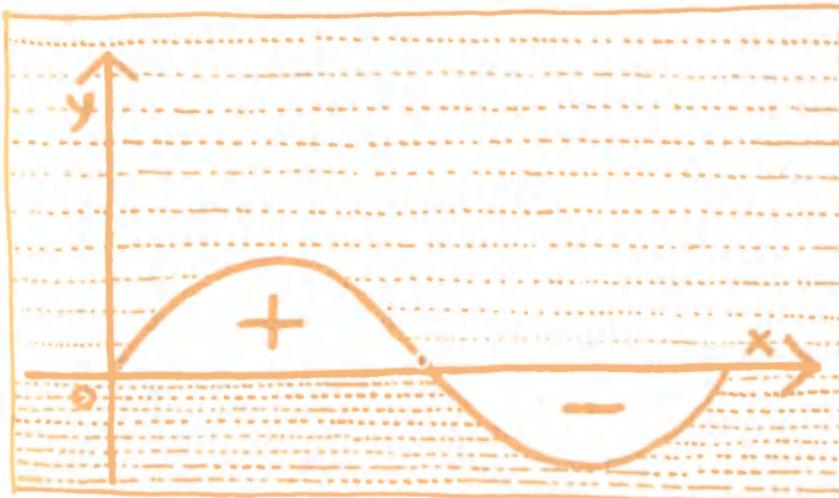


Fig. 76

while those lying below the axis are considered negative (Fig. 75). As a particular instance, the area bounded by a segment of the sine curve $y = \sin x$ over the interval $0 \leq x \leq 2\pi$ is equal to zero, since the area of the positive half-wave is equal to the area of the negative half-wave (Fig. 76).

SPACE, DISTANCE, NORM

In the course of time, many words of a language acquire new content, which is frequently more general than the original meaning. The word "mass" has come to mean a quantity of matter, a large body of persons, the body of people in contrast to the elite, and it is also one of the basic physical concepts.

Similarly, the word "space" signifying that which is capable of containing something has taken on a new and more general meaning. We have already discussed the notion of a multi-dimensional space as a generalization of ordinary space. What I want to do is now to show its further generalization that is closely connected with the concept of closeness.

We all know that in our ordinary space, the (shortest) distance between two points P and Q is the length of the line-segment connecting them. However, we do not live in empty space but on the earth; and if we consider the earth to be a sphere, then the distance between, say, Moscow and Alma Ata is not measured as a straight-line segment but as the length of an arc of the large circle between these points. An airplane could be used to fly the arc of a large circle. But if we travel by railway, then the distance must be taken to be the railway line, which of course goes around obstacles like deserts and is quite a bit longer than the arc of a large circle.

In a city, the distance between one's house and his

place of work is measured along the streets of the city and not by means of a straight line. We might add that travelling by foot or by motorcar gives different numbers of kilometres between these points. Incidentally as a rule we measure the distance between our homes and our places of work not in kilometres but as the time spent in getting there.

Now imagine a wire frame in the form of a parallelepiped. An ant moving from one vertex to another has to travel along the edges of the frame and consequently the distance for the ant is the sum of the distances covered along the edges. In the Mary and Maude section we did not figure out to the end the question of a measure for the closeness of curves. Now our problem is to think up how to measure the distance between curves.

All this compels the mathematician to meditate on the properties that various notions of distance have in common. We are led to the following principal properties.

The distance between two points P and Q must be a nonnegative number, and quite naturally we assume the distance to be zero only when P and Q coincide. Let us denote by $r(P, Q)$ the distance from point P to point Q . In ordinary space, the distances between P and Q and between Q and P are the same: $r(P, Q) = r(Q, P)$.

This property is known as *symmetry*. Do not think that such is always the case. In cities with one-way streets, as will readily be seen, the distance between two points by motorcar is quite different (from P to Q differs from the Q to P route). For the present, we will not deal with such nonsymmetric situations.

Finally, the most important property: the triangle property. According to this property, the sum of two sides of a triangle is not less than the third side. This

can be written as follows: if P , Q , S are three arbitrary points in space, then

$$r(P, Q) \leq r(P, S) + r(S, Q) \quad (1)$$

Suppose that we have at our disposal a certain set of entities of any nature whatsoever, say, points in the plane or in ten-dimensional space, vectors or polynomials, functions or transformations. Now let us proceed to construct the space of these objects. We call the objects of our new space points or vectors. There will be no confusion at all, for we will handle the elements of our new space (the elements may be, say, functions or transformations) the same way we handle points and vectors in ordinary three-dimensional space. Also, they will have the same designations: capital letters.

We are now in a position to give an exact definition of a metric space, which is a space that has a metric, or the concept of distance. This space can consist of a set of elements of arbitrary nature if for each pair P and Q of the elements of the set there is defined a real nonnegative number $r(P, Q)$ called the distance and having the following three properties: (1) the distance $r(P, Q) = 0$ if and only if the points P and Q coincide; (2) for any triple of points P , Q , S of the space the distance from P to Q does not exceed the sum of the distances from P to S and from S to Q (this is the triangle axiom, the formula of which is (1)), and, (3), the distance is symmetric.

The distance enables us to resolve the problem of choosing a criterion for closeness in the set of entities under study: if the distance between the entities is a small number, then the entities are close. Of course, there is still the question of what is a small number, but we have already discussed that.

I will now demonstrate how general the concept of

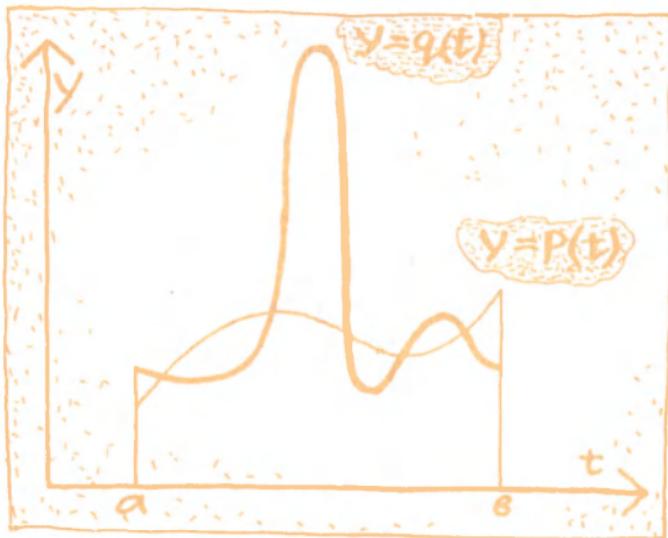


Fig. 77

a metric space has turned out to be, and how unexpected too. Suppose, to start with, that the points P , Q , S of our metric space are the functions $y = p(t)$, $y = q(t)$ and $y = s(t)$ specified on some time interval $a \leq t \leq b$. For this set of functions, it is possible to introduce the concept of distance in a variety of ways. Let us take, for instance, the functions describing the power consumption of Mary and Maude. For the distance between the functions we can take the largest difference (the distance must be nonnegative in absolute magnitude). Fig. 77 shows the functions $y = p(t)$ and $y = q(t)$. Fig. 78 shows their difference, and Fig. 79, the absolute value of their difference. The maximum value of this latter quantity is taken as the distance between the functions. It would be advisable to choose this notion of distance for assessing the power consumption from the viewpoint of protecting the fuses from blowing.

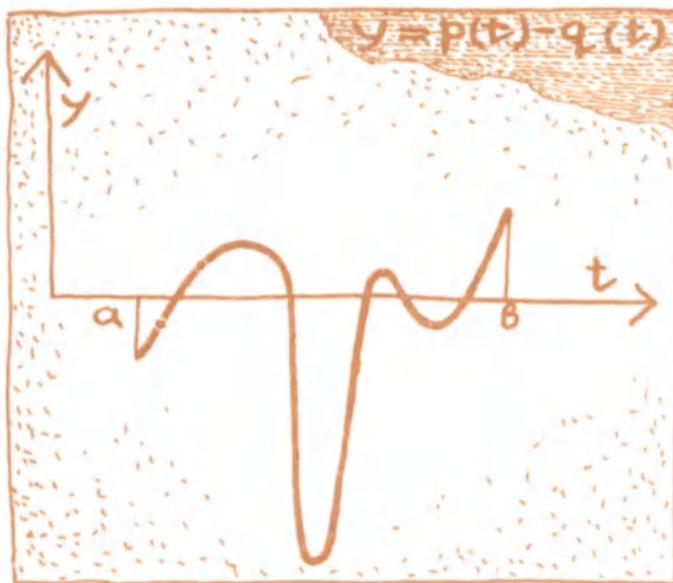


Fig. 78

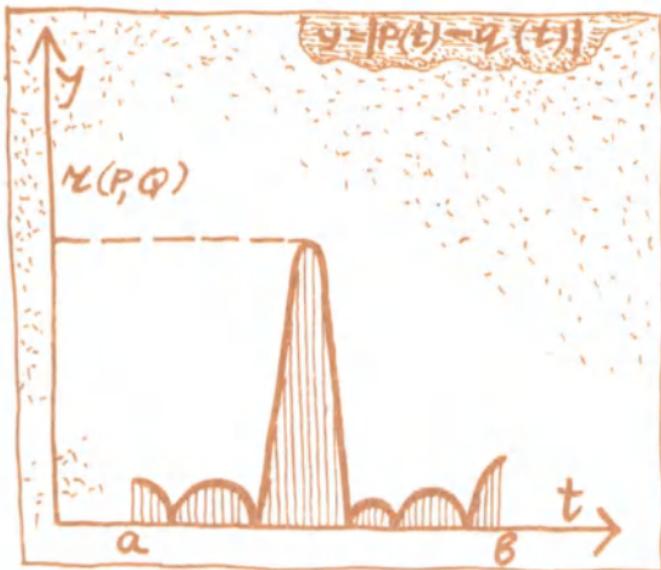


Fig. 79

But if we recall the slogans to economize on electric power, then for the distance between the functions we could take the area that is cross-hatched in Fig. 79. The formula will then be

$$r(P, Q) = \int_a^b |p(t) - q(t)| dt \quad (2)$$

I hope you will be able to stomach this formula, but if you can't, just forget it.

Both axioms of a metric space are fulfilled with respect to the two notions of distance: the maximum absolute value of the difference of the curves and the integral of the absolute value of the difference. If you are so used to the rigour of mathematical reasoning that you don't believe me, then make a check for yourself. It is not hard to do.

Let the distances on a sphere be the lengths of arcs of great circles. Of course, there are two arcs between two points on a circle, and their lengths are not the same if the points are not the ends of a diameter. We will take the length of the smaller arc for the distance between the two points. Here again the axioms of a metric space are fulfilled and the sphere (given that notion of the distance between points on its surface) is a metric space.

Space ordinarily is visualized as something enormous and all-embracing. However, our newly mastered metric space may now be regarded as consisting of, say, only three points, the vertices of a triangle. Indeed, if the points P, Q, S are the vertices of a triangle and the distance between the points is the ordinary length of a straight-line segment between them then both axioms of the metric space (the distance between coincident points being zero, and the triangle axiom) are fulfilled.

And since nothing else is required, that's it!

Later on I will give yet another exciting instance of a metric space.

Some may think that if a space can consist of only a few isolated points, then in the process of generalizing the space concept we must have chosen some of the less essential properties. For example, in ordinary space, vectors may be added and multiplied by real numbers to yield new vectors in the same space. In a metric space this may not occur, as the instance of a metric space consisting of three points indicates.

In constructing a new space, we can preserve the operations of addition of elements of the space and their multiplication by real numbers (scalar multiplication). The ordinary properties of these operations will then be preserved and, in particular, the elements of the space will form a group with respect to the operation of addition. We discussed that in connection with the great discoveries of Évariste Galois. This space is then called a *linear* space.

Vectors in a plane form a linear space under the ordinary operations of addition and multiplication by scalars. The set of all polynomials also forms a linear space. Indeed, a sum of polynomials forms a new polynomial; likewise, multiplication of a polynomial by a scalar yields a polynomial. But vectors have length. If, as is usually done, we refer all vectors to a coordinate origin, then the length of a vector is merely the distance between the terminal point of the vector and the origin.

If we introduce the distance notion into a linear space, that is, if we construct a space that is linear and metric, then we obtain a class of spaces called normed linear spaces, or Banach spaces (after the outstanding Polish mathematician Stefan Banach who

died in 1945. He was one of the founders of functional analysis).

A normed linear space has an analogue for the length of a vector. If an element of the space is denoted by P and the zero element is denoted by O , then the length of the element is the distance $r(P, O)$ between the elements P and O . This number is called the norm of the element and is denoted by $\| P \|$.

If we first introduce a norm into the space, then the distance between two elements P and Q will be the norm of their difference $\| P - Q \|$. It is clearly possible in many ways to introduce a norm into a set of functions of the form $y = f(t)$ given on an interval $a \leq t \leq b$. For example, utilizing for the norm of a function the concept of distance that we have already discussed, we can take its maximum absolute value $\| f \| = \max |f(t)|$ for $a \leq t \leq b$. Given the norm thus introduced and given the ordinary operations of addition of functions and scalar multiplication, the resulting normed linear functional space will have an infinitude of dimensions. We call such a space an infinite-dimensional space.

TERMS AND WHERE THEY COME FROM

We will now take some time out and rest up from the big formulas and heavy discussions concerning space. The reader may ask himself why the term "norm" is used as an analogue for vector length in space.

Which brings up the question of where terms come from in general. The topic is an interesting one.

A standard dictionary gives this series of definitions for the different meanings of "norm": (1) an authoritative standard, (2) a principle of right action, (3) a set standard of achievement, or a pattern or trait taken to be typical.

It will be seen immediately that the norm of a function does not fit any of these meanings.

Mathematicians widely use words with the root "norm". We have such notions as normal space, normal operator, normal divisor, normal distribution, normal equation, and simply a "normal". All these notions are quite distinct and they come from different branches of mathematics. And whereas we have normal people and also abnormal people (true, it's not quite clear what kind they are, ill or mad or what), there are no abnormal equations or distributions or operators.

Generally speaking, when a mathematician introduces a new term, he ordinarily pays little attention to whether it has a contrasting term to go with it. For instance, there is a class of ordinary differential equations but there are no "extraordinary" differential equations. Actually, ordinary differential equations are equations in one independent variable, whereas differential equations involving many independent variables are termed partial differential equations and not extraordinary differential equations.

In mathematics, a matrix is a rectangular array that looks something like this:

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \end{vmatrix}$$

Its elements may be numbers, letters, or functions. Note that the matrix used in typesetting is quite different. Now take what is called a square matrix with the number of rows equal to the number of columns:

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

The sum of the diagonal elements $a_1 + b_2 + c_3$ is termed the trace (or, in German, Spur) of the matrix. There is no connection here between the ordinary meaning of the word "trace" and the mathematical "trace of a matrix".

Incidentally, a relatively short time ago a well-known American mathematician, J. L. Doob, a specialist in the theory of probability, introduced a new term called the *martingale*. This term is used to denote a class of stochastic (or probabilistic) processes of a special type.

Here is the definition of a martingale taken from Doob's book *Stochastic Processes*:

A stochastic process $\{x_t, t \in T\}$ is called a *martingale* if $E\{|x_t|\} < \infty$ for all t , and, no matter what $n \geq 1$ and $t_1 < \dots < t_{n+1}$ with probability 1,

$$E\{x_{t(n+1)} | x_{t_1}, \dots, x_{t_n}\} = x_{t_n}$$

A few years ago Doob was here in Moscow and spoke at a seminar at Moscow University. He was asked where the term martingale came from. Although Professor Doob had delivered his talk in Russian, he did not find words enough to answer that question. Instead he drew a horse on the blackboard and then a circle around the neck of the horse. Then he pointed to the circle and said: "This is a martingale, and so is that which I defined earlier," or words to that effect.

I like that rash way of introducing new terms; no need to justify oneself before one's colleagues and describe the complicated chain of associations that brought the author to his new term. It is perfectly sufficient that the term rings well, is easily remembered and therefore has a right to exist.

Take the word "cybernetics". Nowadays few people know Greek, and arguments about the meaning of the word (pilot, governor) and its associative relationship

with problems of control were of slight use. However, the term was introduced by Norbert Wiener and is now commonplace and is used instead of a roundabout phrase like, say, "theory of automatic control".

Of course, when introducing new terms, one should be guided by something more than vanity. The objects and phenomena under consideration must be important enough and the class of phenomena or objects should merit having a new term to describe them.

THE PROBLEMS OF AN OIL ENGINEER

In our earlier talks between the mathematician and the oil engineer it became necessary to formulate three problems. Firstly, it was necessary to state the criterion of quality in the process of primary oil refining. Secondly, it was necessary to construct a mathematical model of the process, and thirdly, to indicate an algorithm (procedure) for controlling the process on the basis of the criterion and the model.

Unfortunately, we cannot boast about any solutions. These are extremely difficult problems and as yet no complete solutions exist anywhere in the world. However, the basic ideas that might lead to a solution are typical not only of the process of primary oil refining, but also of an extremely broad class of problems in the control of technological processes. We will therefore dwell only on the basic ideas without going into details about possible ways of solving our problems.

Let us start with the choice of a quality criterion of the process. The mathematician does not ordinarily know how to choose the criterion.

Here is the way the mathematician puts it (in the hope of emerging with honour from a complicated situation): "The choice of a quality criterion is the task of

the technologist or even, perhaps more so, the director of the plant."

One need not of course require that the mathematician have a working knowledge of the technology of the process and an understanding of intricate interrelationships, say, between factories and their suppliers, customers and superior bodies. Yet it is precisely this factor that makes for effective strategy of control, and, hence, the choice of an appropriate criterion (the word "strategy" may be understood here in its ordinary sense, but actually it has a more profound meaning that we will discuss somewhat later).

However, the executives and technologists have a hard time of it too because it is their duty to specify the criterion and at the same time satisfy the mathematician who will demand rigour and precision in all formulations. We can sympathize with the poor technologist. He does not exactly know what needs to be optimized and he is more used to plugging away at the job than discussing matters with mathematicians and juggling intricate mathematical formulations.

Therefore, either the technologist must become a mathematician or the mathematician must learn the technology, or—still better—both mathematician and technologist will have to learn to live together and cooperate on the job at hand. This third way out will surely lead to success with the smallest losses to both sides.

In short, the mathematician will have to "go to the people", as the old revolutionaries used to say, that is, go into production and learn what it's all about. Not permanently of course, but long enough (weeks at a time) to get the hang of things. He will have to talk with specialists and gradually dig up the information he needs in order to formulate a criterion. Contrariwise, the men on the job will have to find time and

meet the mathematician half way, so to speak, and give him sufficient explanations so that he can get on with his job. Here, the principle "It is better to see something once, than to hear about it a hundred times" is undoubtedly justified.

The reader will have to forgive me this little introductory section of slogans and pep talk. Now I promise to go straight to the heart of the matter. True, I will not take you to a factory, but will try to explain the method of constructing a criterion using a different example.

CHOOSING A JOB

Let us say an automation engineer—we'll call him Mike—is fed up with his factory job. We will say he is dissatisfied with the slight chance of getting a new flat, with the hurry-up-and-wait kind of work he has been saddled with, with the big turn-over of workers at the factory, the little time that he finds for private study in his field, and finally with the heavy-handed administrative manoeuvres of his superior.

Mike decides to probe out his friends for a new opening elsewhere and he gets five propositions.

At the North Factory of Canning Machinery (NFCM), the instrumentation laboratory working on computer-control systems needs an automation engineer. There's a good deal of work to do, no free time to speak of, no flats in view, and, what is more, the administration doesn't believe in the realistic possibility of computer control at the factory and is more interested in keeping the monthly plan fulfilled.

The heads of the Series Design Office (SDO) and the Experimental Design Office (EDO), and those at the Mathematical Machines and Automation Equipment plant (MMAE) answer the question of what the work is

like with a long story about seniority, bonuses and the sport-club facilities. True, some old college-day mates relate that the SDO is engaged mostly in introducing simple-type regulators in air-conditioning systems, and at the EDO they are for some reason being given refresher courses in pneumonics, while the MMAE is busy testing the vibration stability of apparatus designed in a different factory, and the instruments are sealed without permission to open them.

On the other hand, the MMAE has a system of post-graduate studies and last year there were vacancies because the department heads kept the clever ones, and those invited from outside couldn't make the grade.

The question of flats is always rather acute and in this respect the EDO is doing a good job by building a new block of flats. True, a new person hasn't much chance for the first house, but it'll probably be easier later on.

At the SDO and the MMAE, there are promises with respect to flats, but somehow they sound rather indefinite, more attention being paid to output.

At the Radiation Research Institute (RRI) the situation is quite different. First of all, it is a new institution and there's a doctor of biological sciences studying the effect of high-frequency oscillations on the growth of mushrooms, and they want to hire an automation engineer. Since they don't have any apparatus to speak of, Mike'd have to start from scratch. True, the head biologist promises help from a big specialist (with whom he goes hunting) from the Institute of Automation and Telemechanics. The pay here is 30 per cent higher than at the other jobs and there is also a possibility—in a year or two—of setting up an automation laboratory. But here's a drawback—it takes two hours to get there. That means changing

buses three times and two kilometres on foot in all kinds of weather. But then again, the RRI is planning to build a block of flats.

Now here's another angle. Lucy works in the SDO and if Mike gets in there, he'll be near enough to keep an eye on her. If he goes to the RRI, he'll be able to see her only twice a week. On the other hand, if Lucy hurries up with the answer and they get married, then they'll get a flat and she can come over to the RRI.

What would you say, dear reader? Mike couldn't find the answer either. So he went over to see a cybernetician friend of his, and neither did he go over all possible variants with all the pros and cons and pluses and minuses of a career. The cybernetician suggested making up a table.

The columns of the table listed the various institutions with vacancies. The rows indicated the various points of interest to engineer Mike. The two of them filled in the table row by row, assigning to each site a mark using a ten-point scale. This proved to be much simpler than trying to examine the whole situation at once. Now let us see what they got.

The most interesting job was in the RRI. It was regarded as not being such a big-scale idea but the job was one with a lot of independence. Accordingly it got a mark of 10, the highest.

The NFCM factory also had interesting work, but most of it was carried out by programmers and computer specialists, whereas an automation engineer would always be in the back seat doing subsidiary work. This case drew a mark of 8.

The SDO and MMAE would seem to be offering a pig in a poke, so to say. Perhaps the most exciting work is that offered at the EDO, so the boys over there say. Problems of pneumonics have a big future. True, one doesn't know if Mike will even be offered to take up

pneumonics, but of course he could try. To summarize, then, in the first row, SDO and MMAE get 2 points each, and the EDO gets 5.

Mike did not know any of the heads of these laboratories, but he gathered some information about them.

At the NFCM factory, the chief of the group engaged in introducing computers was a very energetic young man who had finished the same institute as Mike but two years earlier. The boys said he was all right, but he hadn't achieved much yet, and in that case of course one couldn't hope to learn much from him. He got a mark of 7 points.

The head of the SDO was a taciturn old man on the verge of retiring. The talk Mike had with him was highly unimpressive. Acquaintances said that he was no harm—just a stick in the mud. His mark came up to 5.

In the Experimental Design Office Mike was not able to see the chief because he was away on a commission. Mike was told, confidentially, that he was a rather tough character, envious and did not have good relations with a number of his superiors. He got 4 points on the 10-point scale.

At the MMAE, the head of the laboratory held the Candidate of Science degree but was somewhat of a bore. He didn't ask many questions and gave quite indefinite answers to the questions Mike asked. Mike learned that he was contemplating a different job in another organization. That yielded him 2 points.

The Doctor of Biological Sciences was very pleasant, wore a beard, spoke three languages and seemed to know everybody. He was expansive in his descriptions of the future in mushrooms, which he said was only the beginning. He thought it possible to appreciably speed up the growth rate of all plants in hothouses.

		Index	NFCM	SDO	EDO	MMAE	RRI	Weight coefficient	Index of weight coefficient	a _i
1.	Interesting content of work	x_1	8	2	5	2	10	15		
2.	Supervisor	x_2	7	5	4	2	9	12	a_2	
3.	Team	x_3	10	2	6	2	8	12	a_3	
4.	Prospects for writing a thesis	x_4	1	1	6	8	8	10	a_4	
5.	Material conditions (pay, bonuses, etc.)	x_5	7	8	3	9	10	10	a_5	
6.	Outlook for advancement	x_6	2	2	6	2	7	8	a_6	
7.	Possibility of obtaining flat	x_7	1	5	9	5	10	15	a_7	
8.	Travelling time to job	x_8	10	4	6	3	1	8	a_8	
9.	Sport facilities	x_9	2	2	10	2	5	5	a_9	
10.	Lucy	x_{10}	5	1	10	5	1	5	a_{10}	
11.	Sum		53	32	70	40	67			
12.	Criterion K — weighted sum		550	342	666	398	778			

Mike had heard very good opinions of him. But it turned out that he was rarely on the job, constantly travelling abroad, and a member of numerous societies. Busy, in a word. Mike would not be learning any automation from him, that was clear. But he could get as much help as he needed. With that the biologist got 9 points.

The "Supervisor" row was thus filled in.

Salaries at all places, except the RRI, were just about the same: 110 to 120 rubles a month. True, workers at the SDO and the EDO frequently got bonuses, which came out to roughly another 10 rubles per month. At the MMAE the bonuses were given at the end of each quarter and came out to a bit more: about 20 rubles per month. The RRI had a position as head of a group with a salary of 160 a month, which was given a mark of 10, and the other salaries were then marked accordingly.

The NFCM had the best location, only 10 minutes by foot. To get to the SDO and the MMAE plants required 40-minute bus rides, and the MMAE required an extra trip by the underground (Metro). To get to the EDO meant 25 minutes by trolleybus. The worst was the RRI: two hours and three changes en route. These figures were duly entered in the table.

Incidentally, one often wonders where the numbers come from. Of course, it is possible to suggest an algorithm (rule) for computing them. Say, we could assume that 10 minutes is 10 points, 2 hours and three transfers (roughly 150 minutes) is unity, and then join the appropriate points (10, 10) and (150, 1) with a straight line. This is done in Fig. 80. Then 25 minutes corresponds to 9 points, 40 minutes, to 8 points, and 40 minutes plus one transfer, or roughly 50 minutes, to approximately 7 points.

We can say that the inconveniences are inversely

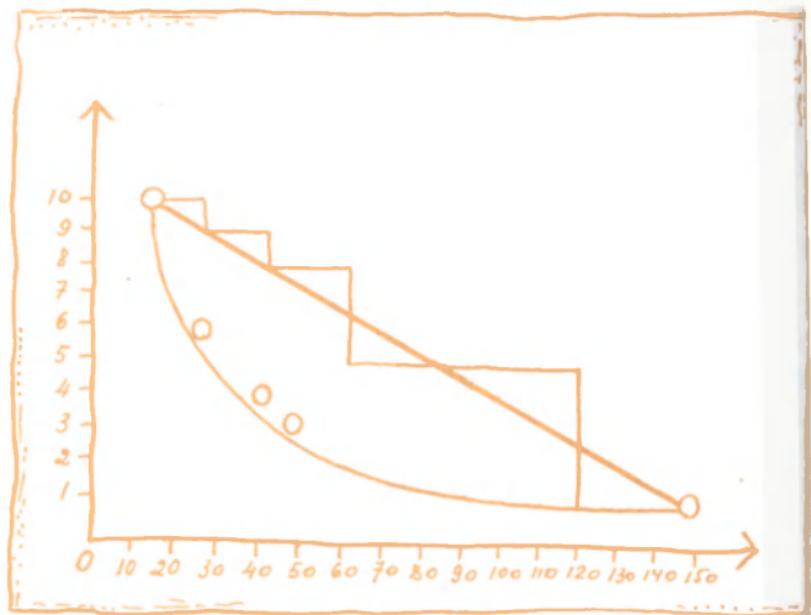


Fig. 80

proportional to the time en route, in which case the number of points will correspond proportionally to the numbers $1/10, 1/40, 1/25, 1/50, 1/150$.

To obtain a mark of 10 for the shortest time, multiply these numbers by 100 to get $10, 2\frac{1}{2}, 4, 2$, and $\frac{10}{15}$. Rounding off these numbers, we can enter 10, 3, 4, 2, and 1 in the table.

We can draw any decreasing curve passing through the same points $(10, 10)$ and $(150, 1)$ —say, the one shown in Fig. 80—and take the number of points accordingly.

Actually, 10 minutes of bus riding in peak hours can result in a loss of buttons, one's equanimity and even prestige—all of which might exceed the losses due to 30 minutes on the Metro, where the riding is much calmer and one can sometimes even read a newspaper.

or book. Therefore, the relative losses en route do not reduce merely to time and may be expressed in a more complicated manner than direct and inverse proportionality. So the numbers entered in the table must be those estimated by the interested person because the table as a whole is a strictly individual estimate of the situation. We will therefore not argue with Mike and will retain the numbers in the table.

Lucy positively did not want Mike to come to work at the SDO where she worked. She herself didn't like it there and was looking for a different place. However, one is not quite sure that that was the reason. She spoke of the Experimental Designing Office, where she claimed there was a better future. She lives close to the EDO and in case she gets married they can live with her parents. That'll be more convenient, she explained. Also, for some reason she had a grudge against the RRI (or is it that she doesn't want to be away from her parents?). Some more numbers were entered in the table.

The remaining numbers were filled in as the two young men continued their discussion of the pros and cons of the people they would be working with, the opportunities for writing scientific papers, the general outlook for advancement, the possibilities of sport activities, and the chances of finding a good flat.

Now the figures could be tallied in the columns to yield the total number of points for each institution. It turned out that the EDO and the RRI had substantial advantages over the others, as can be seen from a glance at the "Sum" row. Here, the EDO was one point ahead of the RRI. So the EDO was it.

However, one more factor had to be taken into account. Not all ten points of the table were of equal importance to Mike. For example, for the present time, career-making is definitely of less weight than the

content of actual work or the flat problem. This makes it necessary to introduce certain coefficients (we'll call them *weight coefficients*) for the various points of the table. These coefficients could also be marked on a 10-point scale, but the importance could be gauged in per cent as well. We have used per cent in our table. So as to avoid fractions, the percentages in the last column are taken from a hundred. They will serve as the weight coefficients.

It is quite obvious that their choice is subjective. The only doubtful coefficient was Lucy's opinion. But Mike did not attach much importance to it for the simple reason that, firstly, the question of marriage was not yet settled, and, secondly, even in case of it being settled there remained the not too bright prospect of living with the girl's parents.

Thus, the quality criterion K for various variants of a new job appears to be more than merely the sum $x_1 + x_2 + \dots + x_{10}$ but rather a weighted sum:

$$K = a_1x_1 + a_2x_2 + \dots + a_{10}x_{10}$$

In particular, for the coefficients denoted by the engineer—and it is these coefficients that reflect the degree of his interest in a particular index—this criterion is of the form

$$K = 15x_1 + 12x_2 + 12x_3 + 10x_4 + 10x_5 + 8x_6 + 15x_7 + \\ + 8x_8 + 5x_9 + 5x_{10}$$

The final results of all computations are given in the last row of the table. As before, here too the Experimental Design Office and the RRI have much higher marks than the other institutions. But in this last reckoning, the RRI has an appreciable advantage over the EDO, roughly by 15%.

A brief interview with his cybernetician friend, and Mike made his final decision to take the RRI job.

The foregoing reasoning devoted to the construction of a criterion of quality may be quite similar in many other problems. If any qualitative indices, say, salary, can be measured in an objective fashion, then they should be utilized. If a given index can only be assessed in a subjective manner, one should consult an expert or use the estimate of the interested party. For example, when choosing the quality criterion for primary oil refining, one can use the profit obtained by the refining plant. The different fractions obtained in the refining process (various grades of gasoline, gas oil, jet fuel, coke, etc.) serve as indices, while the weight coefficients are the selling prices of the fractions. Besides, it is necessary here to take into account the cost of raw materials, electric power, fuel, wages and the like. It should be noted, however, that even with such a clear criterion as profit, considerable difficulties arise.

It might be advantageous to have different departments of the plant do things differently. For example, one department might find it to its advantage to produce mainly light fractions—this is an easier process and involves fewer losses—whereas for another department, where the refining process is carried further, it might be better to obtain a smaller percentage of light fractions.

In short, what is profitable for one department might be to the disadvantage of another one. The plant director is then confronted with the problem of tying in the interests of the various departments.

If we set up an optimization criterion for the plant as a whole (let it be profit), then this might not reflect the interests of the ministry to which the plant is subordinate. There may be a variety of reasons for this. It might easily turn out that the production of petrol (gasoline) is good for the plant but not for the

ministry, which may be particularly interested in obtaining larger quantities of the heavier oil products for further treatment at neighbouring oil plants.

The optimization criterion specified by the ministry might also differ from that which would be to the advantage of the country as a whole. In country-wide planning, it is important to see that the output of the oil-refining plants ensure all regions of the country with all the required oil products, reducing at the same time transportation costs from plant to consumer to a minimum.

Other arguments might be introduced to complicate further the problem of choosing a criterion of optimization, but there is no need for pessimism, for optimization in the functioning of a plant according to any reasonable criterion is profitable.

MODEL BUILDING

A toy automobile is a model of a real motorcar, the game "cops and robbers" is a model of a real battle. The actual object and its model have something in common but they never coincide completely. The photographs of a film star in profile and full face are two different models. They can be small ones or they can take up half the side of a house. A child's balloon can equally serve as a model of the earth and of a tennis ball. In celestial mechanics it is common for the earth to be modelled as a point possessing terrestrial mass. This is called a mass point.

Quite naturally, a tennis ball can serve as the model of a balloon. But will a real automobile be a model of a toy car? And will a film star be a model of the photograph of an advertisement? I think it best to say 'yes' in this case and for the following reason.

The model of an object, process or phenomenon is some other object, process or phenomenon having

certain features in common with the original. It is ordinarily assumed that the model is a simplified version of the object of study. However, it is not always easy to give precise meaning to the concept "simpler than the original", for the simple reason that in reality all entities or phenomena are infinitely complicated and their study can be carried out with differing and constantly increasing degrees of accuracy.

The notion of a model is a reciprocal notion. Thus, a tennis ball can be considered the model of a balloon, and, equivalently, a balloon can serve as the model of a tennis ball. From this point of view, a film star may be regarded as the model of a photograph of the star. A real airplane can be viewed as the model of a toy plane, for there will always be properties of the toy not possessed by the real aircraft.

To summarize, then, when we build a model of some object, we must always specify the precise properties of the original object that are to be modelled.

It is possible to build models of processes and phenomena and not only of actual objects. Playing the accordion models the process of breathing (in and out), playing the organ models a choir. Preparing meals is a model of many technological processes.

Model building has for a long time served as a big help in studying a great diversity of phenomena. It is used today on a broad scale in technology and is making inroads into biology, psychology, and economics.

Models of ships are used to study their stability and manoeuvrability. Studies of the behaviour of model aircraft in wind tunnels make it possible to perfect the design of new aircraft.

Small-scale models are used in the designing of hydropower stations, bridges, and other large-scale

structures. Shipbuilding, aircraft construction, rocketry are permeated with all manner of models.

This kind of model building has to do with the actual construction of physical models and rests securely on the theory of similarity.

Pilots, navigators and astronauts study their arts on models of control systems. These models are no longer merely geometrically similar; the similarity lies in the functions of the appropriate systems.

Of fundamental importance are behaviour models.

These are physical models in the form of devices that interact with the environment and reproduce processes similar to the purposeful behaviour of living organisms. The sensitive elements of these models—they replace the sense organs of living organisms—are photocells, microphones, electromechanical relays and various measuring devices. The turtles, squirrels, mice, etc. that have been constructed or designed by scientists model motions that vary with reactions to light, touch, sound, and so forth. Of considerable interest are the learning models, say, a mouse that learns how to find the shortest route in a maze. Models of this kind have seen quite considerable development and use. Claude Shannon (we will come to him a bit later in our story) constructed an artificial mouse that was taught a variety of purposeful acts of behaviour. Models of this kind have become very widespread.

When modelling the functions of an animate or inanimate object, one frequently makes use of electronic or pneumatic models. Their design is based on the identity of the mathematical description of the processes which occur in the object being modelled and in the model. These models are finding more and more applications. Their use is based on a mathematical description of the object or process being studied.

MATHEMATICAL MODELS

"A rectangular playground is fenced in so that the length is 15 metres more than the width. The sum of the two long sides comes to 80 metres. Find the total length of the fence." This is taken from a standard school textbook in arithmetic.

It's too bad they put a fence up. What's more, I personally have never come across a situation like that in actual life. And this senseless problem is supposed to be solved in four steps (questions). What is still more, the child has to learn these steps and later completely forget how the thing is done. It is hard to believe that this method of teaching is optimal in any sense. It is always harder to undo what has been learned than to learn the right way from the beginning (a similar situation can be recalled from the trials and tribulations of little boys and girls first having to learn to use those old dip-in pens only to give them up in later years—why not fountain pens and ball pens from the very start? It would seem that we have just about coped with that problem now—at last). But still and all, a scheme of questions amounts to a mathematical description of any situation. I will describe a more convenient mathematical model. Denote the length of a rectangle by x and its width by y . Then the statement of the problem yields

$$x = y + 15, \quad 2x = 80$$

Find $2x + 2y$.

It is clear that this way is much simpler and more understandable: setting up equations is a very convenient way of obtaining a mathematical description or a mathematical model. A more general mathematical model of the same situation can be obtained by introducing literal coefficients.

Given: $a_1x + b_1y = c_1$ and also $a_2x - c_2$. It is required to find $Ax + By$.

Here, all the coefficients are regarded as given (but arbitrary) numbers, and for a numerical solution all we need to do is substitute numbers into the final formula.

For a model of the earth, one often takes a mass point with the terrestrial mass. In other situations, for an earth model we can take a sphere represented by the relation $x^2 + y^2 + z^2 = R^2$ (where R is roughly equal to 6400 kilometres and the coordinate origin is placed at the centre of the earth), or a geoid (a sphere compressed at the poles) whose surface is given by a more complicated equation than that of the surface of a sphere.

Depending on the type of problem, the earth is regarded as a homogeneous sphere, a rigid body with variable density, or a body covered with a liquid. Each situation requires its own mathematical model of the earth. For instance, in the study of tides it is of course impossible to compose a mathematical description without allowing for the fact that an enormous portion of the earth's surface is covered with water or disregarding the forces of lunar attraction.

Newton's second law of motion states that the product of the mass of a body by the acceleration is equal to the sum of the acting forces. For the sake of simplicity, we consider only the motion of a body in a straight line. If m is the mass of the body, a the acceleration, and F the sum of the forces, then the mathematical model for the relationship between the mass, the acceleration of the body and the acting forces is given by the equation

$$m \times a = F \quad (1)$$

This mathematical model gives a good description of physical phenomena so long as the velocities invol-

ved are not great. We know that if the velocities of the bodies are small compared with the velocity of light, the masses may be regarded as independent of the velocities. But when the velocities of the bodies become comparable with the velocity of light, then we get a substantial discrepancy with experiment, and this mathematical model then gives a poor description of the situation.

To refine the model, one has to resort to the concept of the derivative of a function.

Incidentally, if at this point you are fed up with formulas or if airplane speeds are the limit of your interest and you do not plan to participate, even vicariously, in interplanetary expeditions, or if you are totally indifferent to Einstein's theory of relativity, then you can calmly skip the next few paragraphs.

I do not intend to explain in detail the notion of a derivative and will only give a rough explanation of the symbolism that is used by mathematicians. Let v be the velocity of a body. We use the symbol dv to denote the differential of the velocity, which stands for the change in velocity v of a body during a very small interval of time dt (dt is the time differential; like dv , it is also a single symbol) so that the acceleration at time t may be expressed as

$$a(t) = \frac{dv}{dt} \quad (2)$$

The right-hand member of this equation is called the derivative of the velocity with respect to time. In the problem at hand, the main role is played by the quantity of motion, which is expressed as the product of the mass of the body by the velocity— mv . Then the earlier mentioned second law of Newton may be written thus:

$$\frac{d(mv)}{dt} = F. \quad (3)$$

which in words reads: the derivative of the quantity of motion with respect to time is equal to the sum of the acting forces.

If the mass m is independent of the velocity v , then

$$\frac{d(mv)}{dt} = m \frac{dv}{dt} = ma \quad (4)$$

and the earlier mathematical model, (1), is preserved.

However, if the velocity is close to that of light, then by Einstein's theory of relativity the mass depends on the velocity:

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (5)$$

where m_0 is the rest mass of the body and c is the velocity of light in vacuum. In this situation, we cannot take the mass m outside the sign of the derivative in formula (3) because the mass can change with time and the mathematical model of the relationship between mass, velocity and acting force in the mechanics of relativity theory is formula (3) together with formula (5).

If the force F and the mass m are given and the velocity v is unknown, then formula (3) is an elementary kind of differential equation. And that's as much as I will say on this involved topic of differential calculus and the still more intricate field of differential equations (which are equations that involve unknown functions and their derivatives). We need only add that differential equations are the basic mathematical model in physics, chemistry and other fields for an extremely diversified range of phenomena in which one has to take into account the dynamics (change) of the variables involved.

At the present time, differential calculus is taught only in higher educational establishments but not yet

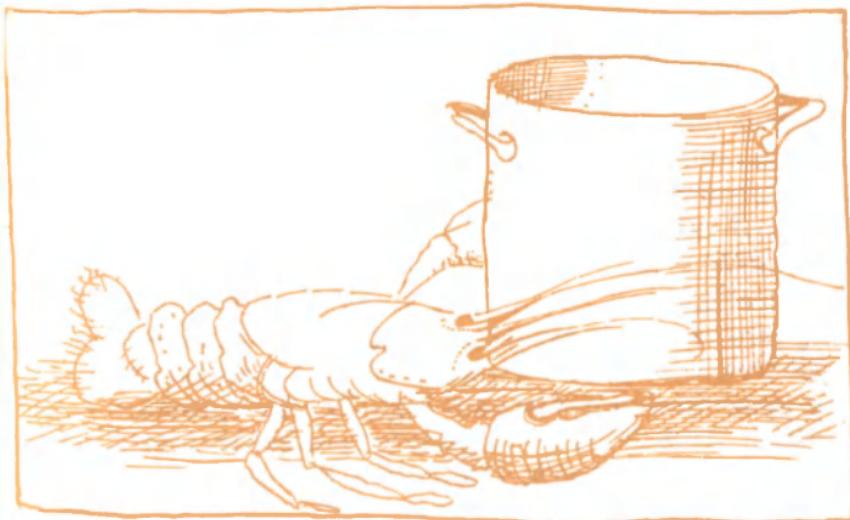
at all departments. Actually, however, this mathematical apparatus is much more needed, and it is much simpler and readily understandable than many of the sections of mathematics studied in secondary school.

EVENTS AND THEIR MODELS

"If all red boiled crabs are dead and all red dead crabs are boiled, does it follow therefrom that all dead boiled crabs are red?"

The reader will of course want to penetrate this dramatic situation on his own. By the use of common sense and elementary logic, he will do so in a matter of minutes, I'd say.

But the mathematician is not one to be fooled so easily. He is constantly on the outlook for confusion in words that look alike and does not trouble himself by running through all manner of versions in which he has the slightest possibility of being trapped. The mathematician would in such a case make use of the



algebra of events. Let us take a look at this algebra.

We will consider collections, or sets, of certain entities (objects or elements). For the questions under discussion, it is completely immaterial whether the sets are finite or infinite, whether they consist of crabs, beautiful girls, all possible routes between A and B , playing cards, or points in a plane. The important thing is that the constituent elements are homogeneous.

We will engage in experiments called thought experiments. They amount to dealing cards, choosing Miss Universe, checking the colour of crabs taken from a kettle, choosing routes less than three kilometres long, or indicating some set of points in a plane, or anything else you can think of.

The results of such experiments (or observations) will be called events. Say in checking a group of 10 crabs, it turns out that only three are red. This is an event. A no less important event is a flush in poker. In other words, practically any kind of result can be termed an event.

Let's give a definition: any set of initial elements (called a set of elementary events) is an event.

Now is the right time to introduce certain operations on events. If we have two events A and B , then it is always possible to relate two new events determined by the conditions " A and B occur" and " A or B or both A and B occur". In the former case we have a product of events $A \times B$, in the latter, a sum of events $A + B$.

To illustrate, we will assume that event A is the appearance of a point in a region hatched vertically, and event B is the appearance of a point in a region hatched horizontally. In Fig. 81, the product of the events, $A \times B$, is the region covered with a grid, while the sum, $A + B$, is the entire hatched region. It is bounded by a boldface outer contour.

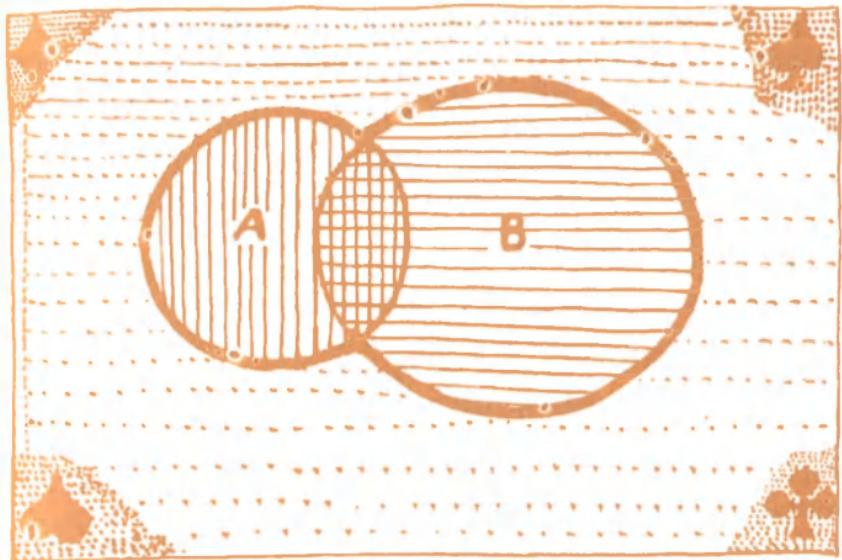


Fig. 81

One's first impression is the unjustified use of the familiar concepts of a sum and a product in such an unusual sense. But such is the accepted usage; furthermore it is quite justified. After dealing with sets for a while, you will get used to sums and products in sets just as you are used to arithmetical operations. Incidentally, other terms are sometimes used. For instance, in place of the expression " A and B occur", we can say "the intersection of events A and B ". The notation for this is $A \cap B$. In place of " A or B or both A and B occur" we can say "the union of the events A and B " denoted by $A \cup B$.

Every cow is an herbivorous animal, but not every herbivorous animal is a cow. Therefore, the event A —to find an herbivorous animal in a field—takes place every time that event B occurs—that is, every time a cow is found in the field. The converse does not hold true: we might find a donkey (also herbivorous,

and so event A takes place) but event B (the detection of a cow) would not have occurred. For such a situation, we say that event B is included in event A or that event B is part of event A .

When an event B is included in an event A , we use the notation $B \subset A$. This is the case in Fig. 82:

$$A + B = A \text{ and } A \times B = B$$

In particular,

$$A + A = A \text{ and } A \times A = A$$

Fig. 82 is an illustration of this situation. Here, A and B are events consisting in a point landing in the appropriate region, and the region B lies entirely inside the region A .

This of course runs counter to our customary rules of adding and multiplying, but if there weren't anything really new here I wouldn't be telling you about it.

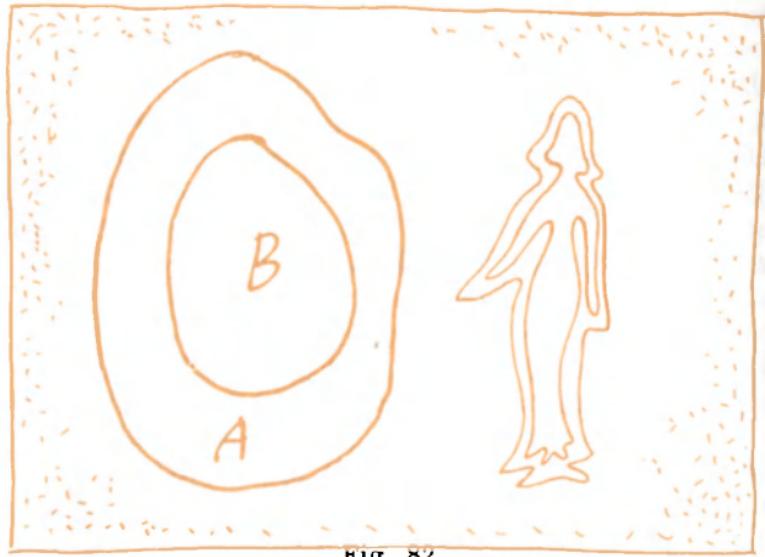


Fig. 82

In our exciting red-crab problem, let us denote the set of dead crabs by D , the set of boiled crabs by B , the set of red crabs by R .

Red crabs may be boiled or not boiled. In Fig. 83, the set B is hatched horizontally and the set R is hatched vertically. The set BR is cross-hatched into a grid and corresponds to red crabs which are boiled.

Since it is given that all red boiled crabs are dead, the set D of all dead crabs must contain within it the set BR . This may be symbolized as $RB \subset D$. The situation is then depicted as shown in Fig. 84, where the set D is hatched with oblique lines. The region marked with all three kinds of hatching is the product BRD , which means boiled, red, and dead crabs. The region with horizontal hatching and oblique lines corresponds to the red and dead but not boiled crabs.

Such is the general situation. Now let us take into account the second assertion: "all red dead crabs are boiled." Thus, there can be no red and dead but not boiled crabs, which is to say the region hatched with horizontal and oblique lines must be excluded. Then, in place of the situation depicted in Fig. 84, we get Fig. 85.

This picture solves our problem completely. Namely, the region with vertical and oblique hatching indicates the possibility, in the situation described, of dead boiled but not red crabs. For this reason, from the fact that all red boiled crabs are dead and all red dead crabs are boiled it does not follow that all dead boiled crabs are red.

This can be written formally as follows: from $RB \subset D$ and $RD \subset B$ it does not follow that $DB \subset R$. Quite succinct, is it not?

All this shows that the algebra of events, a portion of which we have just explained, enables one to construct a mathematical model not only by means of

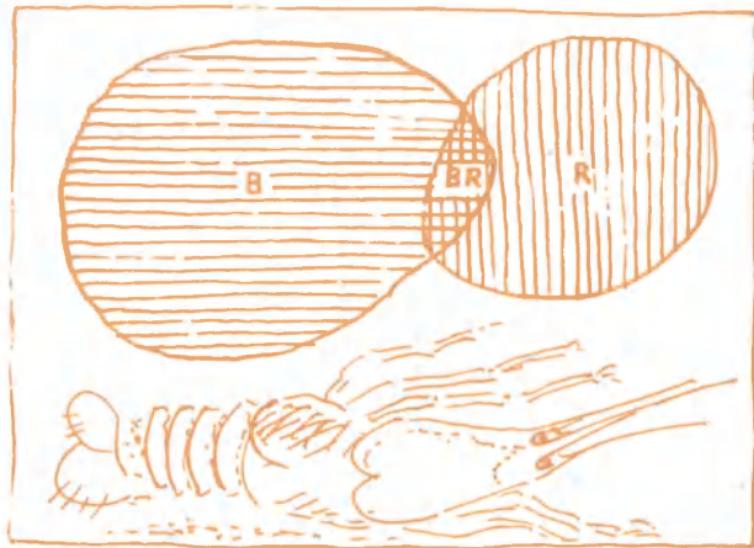


Fig. 83

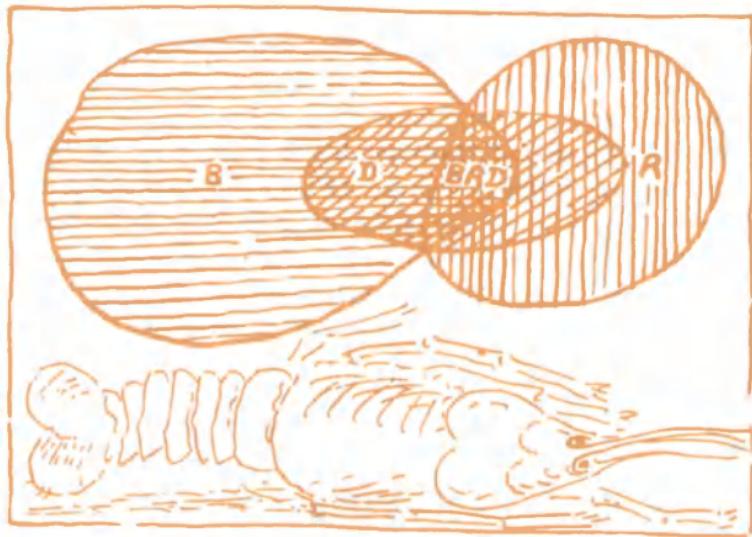


Fig. 84

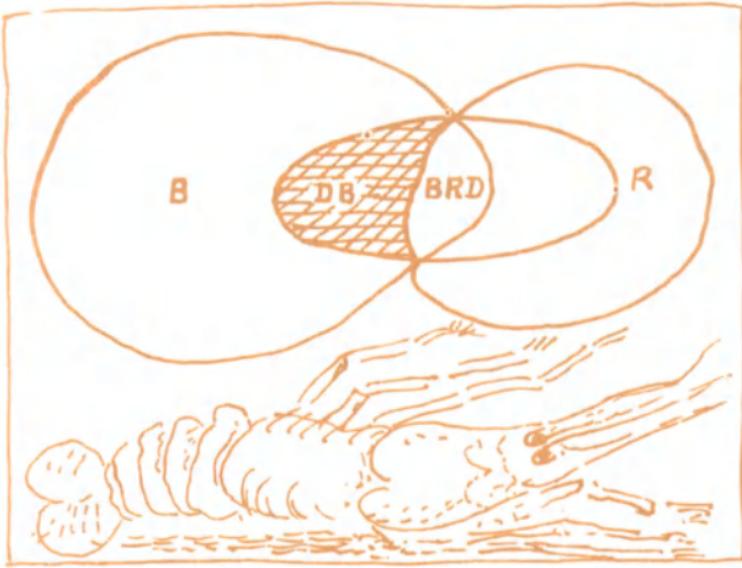


Fig. 85

ordinary elementary algebra and analysis. Incidentally, it is worth noting that Figs. 84-85 are also models of events. They are sometimes called Venn diagrams.

The algebra of events is also called *Boolean algebra* (after the 19th century English mathematician George Boole) or *symbolic logic*. The modern theory of probability rests on the Boolean algebra of events. Besides, this algebra is widely used in constructing mathematical models in many engineering problems, such as, for example, in the synthesis of relay circuits, in the theory of digital computers and the theory of finite automata.

DOES ONE REALLY NEED A MATHEMATICAL MODEL?

We can be quite sure that radar tracking of missiles, well drilling and oil refining are much simpler than ordinary walking. In the process of walking,

hundreds of muscles and millions of living cells participate, and every participating cell is a highly complex organism, the mathematical description of which is beyond the capabilities of present-day science.

Yet, cats, elephants and even you and I run about, eat meals and lavish love on our close ones, without resorting to mathematical model building. Living organisms are extremely economical and precise in their motions.

Having written that, I put down my pen and then took it up again with the question. 'How did I do it?' foremost in my mind. What is the mechanism of this apparently simple motion?

Rene Descartes, the founder of analytical geometry, was not only a mathematician but also a celebrated philosopher (a natural philosopher), a very interesting man of encyclopaedic knowledge. He too wished to grasp nature and find explanations to such remarkable phenomena as the purposeful motions of living beings. Descartes explained the reflex of the hand jerking away from a painful stimulus as follows: in the case of a stimulus (irritation), a steel cable in the nerve contracts, opening a valve in the brain; this releases a nerve gas which proceeds along a tube to the appropriate muscle, filling it and making it contract.

The above sounds naive, but remember that Descartes lived in the first half of the 18th century when nothing was known about electricity and, in particular, biological electricity. But it was a period of clock building and the construction of simple-type mechanical machines, and therefore Descartes could not think up anything as sophisticated as our children today read in their textbooks. However, the Descartes model was the first model of the reflex arc with all its basic elements.

In the third quarter of the 19th century, the celebrated Russian physiologist M. Sechenov published a book

entitled *Reflexes of the Brain*. He suggested that similar reflexes formed the basis of the nervous activity of human beings and animals. Later, the great Pavlov and other outstanding physiologists studied reflexes experimentally. Pavlov introduced and made a detailed study of conditional reflexes (a conditional reflex is a learned response to a stimulus).

Physiologists of Pavlov's school would explain my taking up my pen as follows: in the central nervous system (brain) there developed an order (to take the pen from the desk), which was transmitted via the peripheral nervous system to the muscles. These in turn contracted or relaxed in an appropriate manner, and I finally took up the pen.

However, this model fails to explain many phenomena having to do with motion, and a number of illnesses fall outside this scheme altogether.

1948 saw the publication of *Cybernetics or Control and Communication in the Animal and the Machine* by the outstanding American mathematician Norbert Wiener. This book represents one of the most important scientific events of the middle of the 20th century. Wiener presented here a different model of reflex action. Prior to the Second World War, Norbert Wiener had been interested in general methodological problems (including general problems of physiology) that unite various sciences. During World War II, Wiener became engaged in radar. He discerned a profound analogy between tracking a moving target by radar and the movements of living organisms: in both cases it is necessary to take into account feedback and to generate error signals. Wiener's model looks something like this.

In order to take up my pen, my brain must have worked out a definite order on the purpose of the motion and the initial actions. Then I act by moving in a defini-

te direction, all the time receiving signals about how much has been achieved. I compare these attainments with the task confronting me and work out a *discrepancy signal* (or *error signal*). The task—to pick up my pen—will be completed when the error signal is reduced to zero. My aim therefore is to continually work towards a reduction of the error signal.

Such is the control scheme based on the error signal. Using Wiener's model, we can get to understand many phenomena that could not be explained earlier.

It must be said that the outstanding Soviet physiologist Nikolai Bernstein had already in 1928 pointed to the very important role of feedback in explaining motion. Bernstein was not only one of the first to support cybernetics in the Soviet Union but was actually one of its founders.

Among the numerous—unfortunately they are numerous—illnesses of the nervous system there is one called *intention tremor* which is frequently connected with damage to the cerebellum. The patient with this disease is unable to do certain things: when he attempts to, say, pick up a pen, his hand misses the target, going wide of the mark in a totally uncontrolled manner. Such acts do not fit into the scheme of reflexes. But if we regard them from the viewpoint of the theory of feedback, then the involuntary swinging motions of the hand can be explained. In automatic control engineering, such phenomena occur in improperly adjusted systems of automatic control and go by the name of *over-correction*.

It would appear that the Wiener model is a universal one, all the more so that in this age of automation and high-speed computers, many visualize our brain as consisting of a great number of elements that go to build up an extremely compact, universal computing machine.

However, the Wiener model is in all probability still too primitive. The talented Soviet physiologist V. Gurfinkel has carried out a series of experiments relating to the posture of a standing man.

Try this experiment: put your arm on a desk so that your wrist hangs freely from the desk, and look at your fingers. You should see a slight tremor in the fingers that does not cease. Physiologists believe this tremor to be in the nature of random parasitic vibrations, something like the background noise of a radio receiver.

In studying the standing posture of a person, we can carry out some very interesting experiments. Here is one. A person stands on a special platform constructed in such a way that if the subject begins to sway, the movements of the centre of gravity are immediately recorded. The subject standing on the platform is asked to stand at ease, which he does, believing that he is indeed standing at ease. But we have known for a long time that in reality the centre of gravity is all the time in motion. The recordings of these motions exhibit a rather chaotic type of curve. However, such apparently irregular oscillations exhibit definite frequency components. If we analyse a curve of this kind statistically, we find quite definite oscillations with different frequencies, for instance, frequencies of eight to twelve oscillations per second and an amplitude of 0.1 millimetre, one per second and an amplitude of 2 to 3 millimetres, and also low-frequency oscillations of one per minute with an amplitude of up to 10 millimetres.

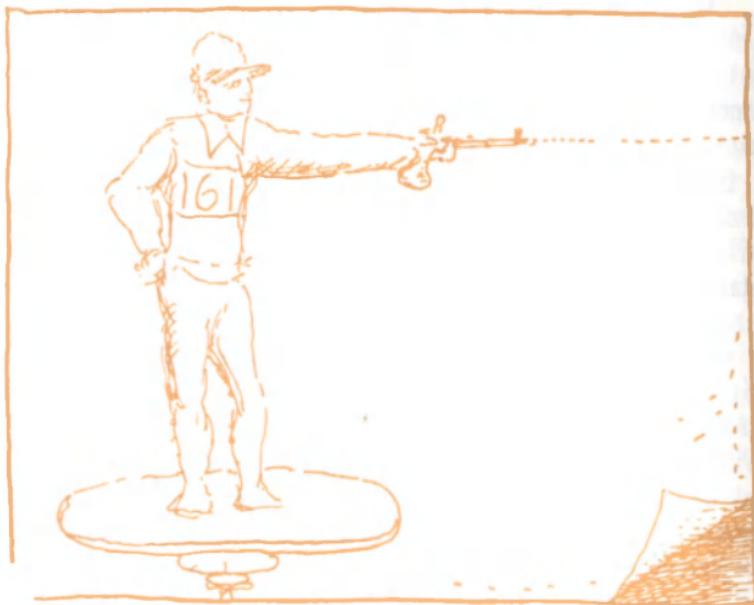
Such random oscillations cannot be accounted for on the basis of the feedback theory.

When purposeful movements are performed, feedback is needed to check their accuracy and to introduce corrections if need be. But a person merely standing in one place would not appear to have to oscillate, making feedback unnecessary.

Yet a person in a standing posture should be able not only to stand still but also to move directly from the standing position to a wide variety of other positions and to perform any kind of manoeuvre.

A person standing still has a great many degrees of freedom. The bones of the skeleton are covered with muscles and there are 28 vertebrae, each of which has three degrees of freedom, making a grand total of over a hundred degrees of freedom which ensure a human being a high degree of manoeuvrability. From the standing-still posture, a person can, if necessary, step in any direction, or waver in any direction, or jump, etc.

There must be a special mechanism for such rapid change-overs from one position to another. This mechanism must hold the whole body in readiness for a multitude of new postures.



During one of Gurfinkel's experiments, he noticed that the instruments did not give any readings at all. The platform and apparatus were checked and found to be in complete order. Then a check was made of the subject, who turned out to be a sharpshooter, a Master of Sports. When standing still on the platform, he hardly wavered at all, and the instruments were unable to record even the slightest tremor. This was unexpected. In subsequent studies of sharpshooters it was found that these sportsmen have beautiful control of their tremor and are capable of stopping it altogether when they aim. What is more, they are able to alter the frequency composition of the oscillations depending on the problem confronting them. When the sharpshooter stood on the platform and took aim, the instrument readings showed that the amplitude of oscillation of his centre of gravity diminished by a factor exceeding 10. We might therefore say that when sharpshooters shoot in a prone position, they do not aim with their eyes but rather with their legs or arms.

We can now hypothesize as to the special mechanism that enables the body of a person to retain a given posture and to move from one posture to another.

At the present time, it is believed that tremor is that mechanism — *a mechanism of continuous search*. For a person in the standing position, the tremor serves as a mechanism for seeking equilibrium. It also, in this position, makes for a rapid transition from one posture to another. In other words, it is a manoeuvring mechanism.

Most likely, search is one of the most universal and sophisticated mechanisms encountered in living nature. A bee in search of nectar performs what would appear to be random movements; a dog following a scent moves erratically; a person's eye examines an object performing outwardly irregular jumps of varying magni-

tude and direction. It is precisely this search that serves as a mechanism enabling a living organism to solve a multitude of problems in maintaining postures, and in moving—in particular, the problem of finding an extremum. (Note that in the standing posture the centre of gravity must be located in the highest possible—extremal—point, and tremor permits one to be near the equilibrium position at all times.)

A natural question arises: is it not possible to utilize search for control in engineering system? The answer is yes, and very successfully.

Imagine that you have to climb down a hill on a dark night. When you ascended the hill in daylight, it appeared to be even and smooth. When you start the descent at night, the hill seems to be covered with humps and ditches, ups and downs threatening you at every turn. With a good deal of swearing, you make trials at every step, putting out a foot to the right, the left, in front and thus choosing the line of steepest descent. You take small steps because in a large step it is easy to lose your equilibrium. That is what searching for direction amounts to, the direction of steepest descent.

In the very same way, we seek the minimum (or maximum) of a function, say, a function of two variables $z = f(x, y)$ or, speaking the language of geometry, of a surface like that shown in Fig. 86.

Let us place a square grid on a horizontal plane, the grid mesh (the side of a square) being equal to h . The points of intersection are termed nodes. We will now travel about the grid, bearing in mind the grid mesh length of h .

We choose an arbitrary node Q_0 and let P_0 be a corresponding point on the surface. We then take steps from node Q_0 to the adjacent nodes Q_1, Q_2, Q_3, Q_4 , and we choose the lowermost of the four corresponding points on the surface, $P_{01}, P_{02}, P_{03}, P_{04}$. Suppose that

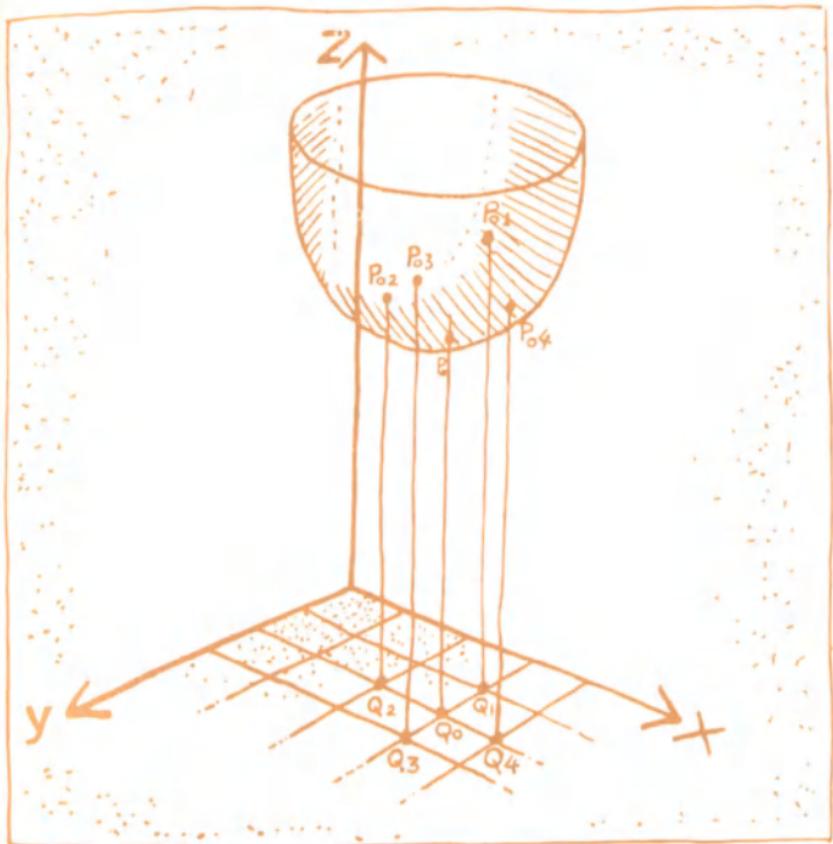


Fig. 86

point is, say, P_{02} . We compare it with the initial point, P_0 . If P_0 is located above P_{02} , then we will continue the search for lower points taking steps to nodes adjacent to node Q_2 . But if P_0 is below P_{02} and, hence, below all four adjacent values of the function at the nodes, our search is at an end. If the function has a minimum, the step-by-step process of search will lead us to the minimum or, to put it more precisely, to a point on the surface close to the minimum.

Of course, there are some subtleties here: the error in determining the actual minimum depends on the size of the mesh and on the form of the function itself. But let us not spoil things with pessimistic questions like "What if...?" All I wanted to do was to demonstrate how to search for a minimum via the "groping-in-the-dark" principle, without any preliminary construction of a mathematical model.

In the use of such a method of searching for a minimum, there is no need to know the function $z = f(x, y)$ for all values of x and y . All we need to do is to be able to find the values of the function at the nodes of the grid with mesh h .

Consequently, even if we do not have a mathematical model, it is still possible—via the search method—to control an object or process in a regime close to optimal.

MODELLING THE OIL-REFINING PROCESS

Now we have to discuss the problem of constructing a mathematical model of the technological process of primary oil refining.

Everything that happens in rectification columns, heat exchangers, furnaces, and elsewhere obeys certain physical and chemical laws, naturally. Hence, everything is very simple: all we need to do is write down the relations which are based on these laws and which relate the quantities that interest us—and there, on paper, will be our dream, a mathematical model.

Almost as simple as sculpturing. According to Michelangelo, here's the way it's done: take a chunk of rock, knock off the unnecessary part, and leave the rest.

Perhaps the reader thinks I want to compromise science and hint that certain very needed laws of nature have yet to be discovered. Nothing could be farther



from the truth. The basic laws of the theory of heat, thermodynamics, gas dynamics, chemical kinetics and other sciences involved in setting up the equations of the respective processes have already been discovered. The laws of sculpturing are also known, but it is no simple matter even to fashion the shoes of a great personage.

Let us recall the talk between the mathematician and the oil engineer. The process of rectification of oil consists in separating the oil into a variety of needed components (gasoline, gas oil, fuel oil, etc.). In the most simplified form, this process is described by tens of interrelated variable quantities.

Even in the static state (when the quantities are in equilibrium), it is no easy task to take all the variables into account. Now to obtain algorithms of optimal control requires dynamical equations of the process. The

details of the processes occurring in a column are not known. The important thing is to simplify these processes so that the describing equations are not too complicated and still represent the processes with sufficient accuracy (in the earlier discussed meaning).

It is a very delicate matter, however, to select important variables and reject variables that have little effect on the course of the process. It is important not to pour out the baby with the water, so to say. On the other hand, retaining a large number of variables can so complicate a mathematical model as to make it extremely difficult to handle. Unfortunately, I have no secret way of doing this. Occasionally it is apparent at once, in other cases a lot of time and energy is spent but the results are meager.

Yet there is still one more complication. One has to take into account all kinds of unforeseen circumstances, like a change in the composition of the raw oil, temperature variations of the air, or a change in the ambient pressure. All such random variations occur irrespective of how the process develops, yet they must be compensated for. Add in also the fact that a mathematical model only gives an approximate description of any process, and you will see that there will always be discrepancies, in which case we must constantly be in a position to rectify and modify the controlling action.

Thus, the construction of a mathematical model consisting of certain basic equations of a process is not yet sufficient for effecting optimal control. The mathematical model must also provide for the effects of random factors, the ability to react to unforeseen variations and ensure good control despite errors and inaccuracies.

All this requires a different mathematical approach, which we will discuss in the next chapter.

You Probably Like This Book

If you, dear reader, have gotten this far in the book and haven't thrown it out or stuck it in the back row of your bookshelf, you will probably read to the end.

When hubby and wife go visiting and—all dressed—she is still looking for her ear-rings, he is sure they will be late but she is confident they'll get there on time.

In ordinary life, probability is understood to mean something like a rough assessment of chance, conjecture or assumption—like the chances of getting caught, of coming down with a cold, or being late to a train. We assess them in a very subjective manner, depending on our nature, capabilities, bits of available information, and common sense.

A person will complain about failing memory, his state of health or lack of luck, but he will never complain about any breakdown in his common sense. But in identical situations, one's assessment of the chances of something happening may be quite different from another's.

The prominent French mathematician Émile Borel, in a small exciting book entitled *Probabilité et certitude*

(*Probability and Certainty*) pointed out that knowledge deserves the name of Science depending on the role played in it by number. In the most diverse problems of natural science, engineering, economics and sociology it is frequently necessary to obtain an objective assessment of the probability of certain events.

Here are some examples.

Ice cream is being sold on Lenin Prospekt, Moscow. If we put one kiosk per block, lines will begin to form and some people will not bother to stand in line preferring to go without ice cream. Hence, customers will be lost and profits will diminish. If we put 20 kiosks on every block, there will be many without any profit at all since there won't be enough customers. How many kiosks do we need? Changes in the weather will also have to be taken into account for they affect demand.

Say you buy a radio set, refrigerator or watch. You have a guarantee of one year, a year and a half, or two years, etc. What does this mean?



Of course, if something goes wrong in the refrigerator or if the watch stops in a month, the repairs will be done free of charge at a special shop. But you are not interested in repairs, you want the watch to keep running. Once you have a new watch and a guarantee of the manufacturer, you consider it highly unlikely that it will stop during the guarantee period. But why did the factory make the guarantee term one and a half years? Also, how much more reliable is a watch with a two-year guarantee period?

A local clinic writes out about 50 sick-leave certificates per day for influenza. On one occasion, there were 70 cases. Should we assume this to be the beginning of an epidemic and then take emergency measures, or is it merely accidental?

Such problems involving accidental factors arise constantly. It is often possible not only to state that the event is accidental, but also to assess quantitatively the indeterminacy of the event occurring or not occurring. Such an estimate is stated thus: "The probability of heads when a symmetric coin is tossed is equal to $\frac{1}{2}$ " or "The probability of receiving a trolleybus ticket with an even number is equal to one half, while the probability that this number will end in the digit 7 is equal to $\frac{1}{10}$ ".

We obtain these figures on the basis of an obvious symmetry, on the equal probability of various outcomes. A pack of playing cards and dice also have symmetry. It was precisely problems in gambling that started the theory of probability, the science of random events and the laws that govern them.

PROBABILITY THEORY. SOME BACKGROUND

The beginning of probability theory goes back to the 17th century, when such famous scientists as Pascal, Huygens, Fermat and particularly Jakob Bernoulli* laid the foundations of the calculus of probabilities. Although they were engaged in problems of games of chance, these outstanding scholars were clearly aware of the important natural philosophical significance of the theory of probability.

However, a number of problems went unsolved and it was far from clear when the scheme of classical probability theory could be applied. The problem of the conditions of applicability of a mathematical scheme or mathematical model is no idle question, and uncertainty in the basic notions of the theory led to rather dramatic events.

In 1812 the celebrated Laplace—astronomer, physicist and mathematician—in his book *Essai philosophique sur les probabilités* summarized the advances of probability theory of that period and included his own fundamental results too. However, in addition to the important mathematical results and applications obtained in the natural sciences, Laplace applied probability theory to the “moral sciences”, to the probability of witnesses’ depositions, to balloting, and to the assessment of equity in judicial sentences. The arbitrariness of estimates and the impossibility of determining in objective fashion the probability of human judgements resulted, in the middle of the 19th century, in probability theory being relegated to the rank of a mathematical recreation.

* Jakob Bernoulli was an associate of Leibnitz in creating the fundamentals of mathematical analysis. He was perhaps the most outstanding of the illustrious Bernoulli family of Swiss scholars that gave the world eleven (!) prominent mathematicians.

It took the genius of the Russian mathematician Chebyshev to separate the alien elements from the true theory of probability and convert the latter into a clear-cut mathematical theory with its own range of problems and its specific mathematical apparatus. P. Chebyshev devoted only four papers to these problems. They were written at large intervals between 1845 and 1887 and their significance to science is very great.

So much for the history of probability theory. Let us return to the matter at hand.

RANDOM EVENTS

Lying on the beach in summer can be dull. Take a toy pail and fill it with pebbles. I bet there are less than, say, a hundred pebbles in the pail.

Bets are dangerously pregnant with passion, whereas the material out of which bets are made is very sound: it is a random event consisting in the fact that the number of pebbles is less than one hundred.

Of course we can bet on a team winning the world football cup. That's also a random event. But there is a substantial difference between these two events.

The experiment with the pebbles can be repeated many times and under identical conditions: over tens of metres of the beach the number of pebbles we put into the pail will be about the same.

For such situations there is a characteristic statistic stability that reflects the regularities of mass phenomena.

Winning a football game is of a different nature. The games of the world cup finals cannot be repeated under the same conditions because the next time around will see new participants, veterans will acquire new experience and will be in a different state of health, the games will be played in a different country, etc.

Although such events are also random events (the point being that they may or may not occur and there is no way of predicting with certainty that they will), they are not characterized by any statistical stability.

Such random events are not studied in probability theory but are beginning to attract the attention of mathematicians in other branches. At the present time, situations similar to football games, wars, interrelationships between producers and consumers, etc., are beginning to be studied intensively. We shall return to that problem later on. For the time being our interest is probability theory, the science that studies random events of a mass nature and events possessing statistical stability.

PROBABILITY

You probably know that the probability of pulling a queen of spades from a thoroughly shuffled pack of 52 playing cards is equal to $1/52$. When Hermann, in the *Queen of Spades*, gets a three, a seven and an ace, that too is a random event, and its probability is easy to compute. Although Hermann played the little known game of "Stoss" (I was told this by mathematician Yu. Shreider), I will compute the chances of these three cards in the more familiar game of "21 points". Here the probability of extracting a three (there are four threes in a pack of cards) is $4/52$. When the three has been drawn, there remain 51 cards, so that there will be 52×51 possible pairs, and there will be 4×4 pairs in which a three appeared in the first drawing, and one of four sevens in a drawing from the remaining 51 cards. Consequently, the probability of obtaining a pair (three and seven) in the two subsequent drawings is $\frac{4 \times 4}{52 \times 51}$. Finally, from the remaining 50 cards we have to draw one of four aces. There will be $52 \times 51 \times 50$

of all possible triples of cards, and $4 \times 4 \times 4$ of all possible combinations we need (three, seven, and ace). Then the sought-for probability of the desired combination will be

$$4/52 \times 4/51 \times 4/50 = 0.00048$$

This is a rather small probability and we can imagine the joy of the player with such luck.

Divide a volleyball court into two equal parts. If you throw a ball at random, the probability of it landing in one of the two halves (assuming that the ball falls within the court) will be one half, while the probability of hitting a puddle on the court will be equal to the ratio of the area of the puddle to the area of the court.

But how does one determine the probability of heads or tails in coin tossing when the coin is nonsymmetric (say, crooked)? How can we compute the probability that pebbles picked up at random on a beach will weigh less than 20 grams? There is no sense of course in weighing every pebble and then finding the ratio of those weighing less than 20 grams to the total quantity on the beach. First of all, it is simply impossible to carry out such an experiment and, secondly, it is not obvious at all that every pebble will have the same probability of being chosen.

How can we determine the probability that a new electric light bulb will burn out only after 1000 hours of service? It is clearly impossible to carry out a series of experiments because when a bulb burns out it is simply thrown away. Still and all, we proceed from the existence of a definite probability in events like those mentioned above. Probability is an objective characteristic of events that does not depend on our attitude towards them. The existence of a probability concerning the events studied in probability theory is

similar to the existence of mass and velocity in the case of any body. The mass and velocity of a body are quantities that exist objectively. They characterize the object under study but we are not able to measure them with absolute precision. However, it is possible to indicate an approximate way of measuring the mass or the velocity. In the same way, we can indicate an approximate method for measuring the probability of an event.

To determine the probability of obtaining heads in tossing a crooked coin, we perform a number of tosses, say n , and count up the number of cases of heads. Let it be m . The ratio m/n is called the *frequency* of the event and serves as an approximate estimate of the desired probability.

Thus, when it is impossible to compute the probability by proceeding from some kind of general reasoning (say, like symmetry or equal probability of outcomes), then we take advantage of a frequency evaluation.

We are consoled by the assurance that as the number of experiments (n) is increased, the estimate thus obtained will be still more precise. Although this is true, the meaning of this statement merits further discussion.

AN EXPERIMENT AND WHAT CAME OF IT

A big experiment, a lot of work, numerous data recorded, and....

Yes, all well and good, but that is only half the job. We still have to draw conclusions on the basis of the data obtained. And that is no simple matter.

The result of a series of experiments may either be a qualitative conclusion like "The introduction of adrenalin causes a rise in blood pressure" or a quantitative description of a situation that looks like this: "In the case of one cubic centimetre of adrenalin injected into five rabbits, four exhibited a rise in blood pressure."

Actually, the result of a series of experiments is always some kind of quantitative characteristic—the total number of experiments at least. In most cases, the result itself can be described quantitatively, although it is not always clear at once how and in what units the measurements are to be made. Now every quantitative description of an experiment requires mathematical treatment.

Imagine a geological expedition in search of apatite deposits. In the course of their work they find diamonds, gold and uranium but pay no attention since these do not interest them. Can we justify their wastefulness? Yet does that differ from a physiologist performing a complicated series of experiments and utilizing only a tiny portion of the information obtained?

Recall the discussions between the mathematician and the physiologist. A 17-loop recorder takes down enormous quantities of information concerning the life activities of a rabbit, yet the conclusions drawn from this Mont Blanc of facts are mainly qualitative, something like: "After the introduction of adrenalin, the blood pressure rose." All the rest of the information is left untouched and hence is lost.

The treatment and analysis of any type of experimental material requires application of the methods of probability theory and mathematical statistics. The theory of probability serves as a theoretical basis for mathematical statistics. Books on probability theory or mathematical statistics state that the content of mathematical statistics consists in working out procedures of the statistical observation and analysis of statistical data and the results of experiments.

Experimental findings can yield a diversity of conclusions, particularly if the number of experiments is not very great.

The great physicist Niels Bohr had this to say about



experimentation: When we have a finite number of experiments and an infinite number of theories, then there exist an infinite number of theories that satisfy the finite number of experiments."

I would like to stress the fact that conclusions from statistical findings must be drawn with great caution, for it is not always clear whether we must attribute the data obtained to pure chance or whether we must consider them a confirmation of our hypothesis.

But the desire to obtain a definite result—which the doctor, geologist, chemist, or even physicist already believes to be true—is at times very great. Then the researcher uses his experimental findings merely to substantiate his own viewpoint, despite the questionable reliability of the results.

Here is an instance.

We treat a dangerous illness by two methods, which we will call the "old method" and the "new method".

An analysis of the data over a short period of time yields the following table:

	Number of patients	Died	Living	% of fatal cases
Old method	9	6	3	$\frac{6}{9} \cong 67\%$
New method	11	2	9	$\frac{2}{11} \cong 18\%$
Total	20	8	12	

From the table it follows that the number of fatal cases is perceptibly lower with the new method. But after thinking a bit, we may be in doubt as to whether the observational data are extensive enough to yield a reasonable degree of certainty regarding the computed percentages.

Now this ratio of fatal cases could also be a random result.

Imagine a sad situation in which both methods are equally effective or not effective at all, so that they do not affect the outcome of the illness in any way.

We will assume that the probability of staying alive is then the same for both methods and equal to 12/20. The probability of a fatal outcome is thus 8/20. What is the probability that the number of fatal cases with the old method is not higher than that given in the table; that is, of the 11 patients using the new method, not more than two will die (that is to say, either two will, or one, or none)? This probability will be approximately equal to 1/25. (That is the probability with which the number on your autobus ticket is divisible by 25, that is, the last two digits being 00, 25, 50, or 75. If you pay attention you will notice that rather often you have precisely that situation—4 times out of a hundred, on the average.)

Although one should not reject the findings in favour of the new method, I for one would not consider them sufficiently convincing, say, for putting out an order to go over to the new method. On the other hand, if I were ill and I had a choice between the two methods and there was no additional information available, I would probably prefer the new method. However, that is no longer an item of probability theory but of a different branch of science about which we will have something to say.

It is worth noting that an experimenter does wrong in rejecting some of the results of an experiment merely because they appear to fall outside the conditions of the experiment, because certain points are off the main curve. The experimenter is sometimes aware of the fact (though sometimes he is not) that he is merely trying to substantiate his hypothesis and certain undesirable results run counter to his desires. Of course, an experiment is occasionally a flop. But without apparent justification one should never reject any result in the analysis of an experiment.

Here is what the brilliant experimenting physicist P. L. Kapitza had to say in this respect in a speech dedicated to the memory of Rutherford:

"The study of nuclear collision processes contains a great weakness to this day. It is the necessity of a statistical method for analysing the results. It is a well-known fact that great caution is needed in order to derive a general law from a limited amount of statistical data. Speaking of the application of statistics, somebody said there exist three types of lies: lies, blatant lies and statistics. True, this was said of the statistics of social processes, but to some extent it can be applied to statistics in physics. In no other field of physics has there been more blunders and false discoveries than in the treatment of statistical data obtained

from nuclear collisions. Every year new particles, new elements and new resonance levels are discovered, which later turn out to be erroneous.

"Rutherford knew very well what danger lies in an unobjective interpretation of experimental findings of a statistical nature when the scientist wants to obtain a certain result. He carried out analyses of statistical findings with great caution. He had an interesting method. Scintillation counting was usually done by students who did not know what the given experiment was about. The curves were drawn by people who did not know what results to expect. As far as I can recall, Rutherford and his pupils did not make a single false discovery, whereas in other laboratories there were no small number of such discoveries."

For this reason, methods of mathematical statistics must be applied reasonably and with care, and then scientifically justified methods of analysing observational results will become a useful and everyday tool in the hands of the experimenter.

I have already mentioned the fact that in physiology and medicine the results of experiments are treated wastefully—a very small amount of information is extracted compared with what could be. To a large extent, the same goes for geology, geophysics, chemistry and engineering. Of course, the experimenter is not always by far to blame for not analysing his material in sufficient detail, for this not only requires a mastery of statistical methods but also the requisite apparatus.

The deciphering of an electroencephalogram or the complete interpretation of geophysical data measured in an oil well not only results in a great volume of computations but also requires complicated algorithms in the analytical stage. Desk calculators do not suffice, one needs high-speed electronic computers, automatic

input of experimental results and automatic print-out of results in the form of tables, punched cards, or curves.

That is not all. Interpretation of the results and choice of the analysis procedure require a deep and unbiased penetration into the problem at hand. By themselves, machines are useless. They must answer certain questions, and it turns out that the most difficult part is to pose a question correctly. Isn't that so?

ENGINEER CONSULTS MATHEMATICIAN

An engineer working on his graduate thesis came in for a consultation with a mathematician the other day. He was a man with extensive practical experience and had done a lot of work on his thesis, performing a large number of experiments. He wanted to bring them into a system and draw certain conclusions. (Let's listen in to their conversation.)

Engineer. I would like to consult you on a number of points. I have been studying the strength characteristics of pipes made of a variety of plastic materials and I have worked out the experimental procedures and carried out a number of experiments. I'd like to know whether the series of experiments already performed and the data obtained are enough.

Mathematician. Enough for what?

Eng. For conclusions concerning the strength of the materials under various conditions.

Math. Could you please give me a more detailed picture of the problem?

[He demonstrated tables summarizing the experiments he conducted. Each experiment lasted several weeks. The problem consisted in choosing a type of the strongest plastic for pipes capable of withstanding oil under pressure in production conditions. The expe-

riment consisted in pipes working for a certain time and then being broken by a special machine to determine the breaking force. After the experiment, the broken pipes were naturally thrown away.]

Math. Why did you vary the time of the experiments?

Eng. I worked out a procedure that would enable me to judge the strength of the pipes in a short time. It is necessary to establish a relationship between the time the pipes were in operation under load and the breaking force. I established this relationship on the basis of a series of experiments over the initial time period (10 weeks) and continued on the basis of theoretical reasoning concerning the mechanism of the phenomenon.

Math. But with that approach any conclusion about the breaking force in one, two or three years of operation of the pipe may be erroneous, isn't that so?

Eng. True, but in comparing different types of plastics, it will probably be possible to state that one type is preferable to another. A pipe can be in use for months or years, but I haven't got that much time in order to draw conclusions about the qualitative characteristics.

Math. Of course, you want your degree as soon as possible!

Eng. Joking aside, you mathematicians have it easy: prove a few theorems and pick up your degree. But we have to accumulate experimental material. That takes time.

Math. You're right, theoretical problems, if they are not too complicated, are very convenient for getting a degree in a short time. But even in theoretical questions you can bog down too. So you think it is obvious that with the methods at your disposal it is possible to determine the strongest type of plastic.... But suppose your hypothesis is correct. Over what intervals of time do you check the breaking force and how many pipes are under test at one time?

Eng. I take 20 samples and select 5 every two weeks. I can't test any more because that's the maximum the machine will take.

Math. What is your question?

Eng. These series of experiments have been carried out with three kinds of piping. Is that sufficient or should I continue?

Math. After the experiments, do you find definite advantages in any one type? For example, can you say that during two months of operation the first (best) type of pipe had a breaking strength that fell by only one per cent whereas that of the others fell by 20 per cent?

Eng. No, of course not. Take a look at this table. Here the values are roughly of the same order. But if we judge on the basis of averaged data, then I believe the first type of plastic is twice as good, and so pipes made of this plastic would last twice as long, which of course would mean a saving of millions of rubles.

Math. How much does one experiment cost?

Eng. What difference does it make? Let's go into the statistics of the matter. That's more to the point.

Math. Suppose you had to have an operation on the gall bladder. There are two surgeons that can do it. One has 10 operations to his credit, 9 of which were a success. The other has 100 operations, of which 90 were successful. Which one would you prefer?

Eng. The second one naturally, but....

Math. Now suppose the second had only 85 successes out of 100?

Eng. Then I don't know. Still I guess I'd prefer the second one because he has more experience.

Math. Let's not get worried about an operation. Suppose we have two marksmen: one scored 9 points out of 10, the other 85 out of 100, which one gets first place?

Eng. I came here with questions for you, and you

wind up asking me questions. Isn't that hitting below the belt?

Math. Nothing of the sort, all I'm trying to do is to help you pose your own questions. But if you don't like that procedure, let's try to answer some questions which you have not yet asked. There are several.

First question. How many experiments do we need to be quite confident that the average breaking force obtained is the true one?

To this question we can give an exact answer: an infinite number.

Eng. What do you mean, infinite?

Math. The point is that every experiment involves an error, the magnitude of which is not known beforehand and it varies from one experiment to another. For this reason, no matter what finite number of experiments have been carried out, the arithmetic mean of the values obtained will contain an error. Of course, if the experiments are conducted under identical conditions and the errors are random errors, then the error of the mean will diminish as the number of experiments is increased, but it cannot be reduced to zero for a finite number of experiments. And so we get the second question.

Second question. What can we guarantee for 5, 10, or 1000 experiments?

You feel more confidence in a surgeon with 90 successful operations out of 100 because of his greater experience (how proper the word is!). Consequently, the coefficient of confidence with respect to the assertion "This surgeon has achieved 90 per cent success in his operations" will differ. It will be higher if it is 90 out of 100 operations and lower if it is 9 out of 10. Intuitively, it is clear to you what *coefficient of confidence* means, isn't it?

Eng. Intuitively, yes of course, but how is it specified?

Math. This can be done in a variety of ways. Roughly speaking, it is the price that must be paid for more reliable data. This price may be expressed by the number of necessary experiments with allowance made for precision, or expressed in rubles needed to perform the experiments or in some other kind of units.

This is where statistics ties in with economics, though you wanted to keep them separate.

Eng. Yes, I understand, but tell me what am I to do?

Math. We'll discuss that later. Meanwhile I'd like to straighten out the questions and formulate them. For some reason you took 5 samples for identical time intervals, whereas the magnitude of the breaking force under study varies with time not in a linear fashion but rather hyperbolically or exponentially, as $y = 2^{-\alpha t}$, where t is the time and y is the breaking force, and α is a numerical coefficient. The most important part of the curve is the initial portion. That brings up the third question.

Third question. At what times is it best to perform the experiments and what quantity of pipes should be taken each time?

This can be put differently: we can assume that the total quantity of samples is specified beforehand and also the time devoted to the experiments. Then the problem will consist in choosing the times for the tests and in determining the quantity of pipes in each sample in order to ensure a maximum coefficient of confidence. This can be put in yet a third way: specify the coefficient of confidence and minimize the overall expenditures for all experiments. Other statements of the problem are also possible.

Eng. Now I'm totally at a loss. What am I going to do?

The two chose a statement of the problem, planned a new experiment, and in two months obtained certain results. So much for that job.

EXPERIMENTER AND STATISTICIAN

This talk with the engineer demonstrates a rather common situation in which the experimenter finds himself. Let us now take a closer look at the problems brought up in the preceding section.

Experimenters do not often consult statisticians. They ordinarily analyse the results of observations themselves, "as best they can", so to say. Sometimes the most fantastic conclusions are drawn, as we have already seen. But that is not all. The results of an experiment, their information content and significance, for a given amount of labour and money, depend to a large extent on how the experiment is conducted: at what times the measurements are made, how many measurements are made, and at what points, how to choose the values of the parameters or actions at the disposal of the experimenter, and more. (We could put a dozen exclamation marks at this point.) But the most important thing here is that the mathematician can suggest to the experimenter ways of getting out of the mire he is so often in. What has to be done?

Simply the statistician should come into the work not at the end of the experimental part but from the very beginning. As a rule, the experimenter doesn't pay much attention to extracting the maximum information about what interests him. Now the statistician will be concerned about this from the start. He has to plan the experiment, choose the number of necessary tests, think through the procedures, and see that the

data obtained are in a form convenient for direct analysis. The statistician will have his hands full. He will "get in the way" of the experimenter and demand all kinds of what would seem to be inessential conditions, but on the other hand when the experimental findings have been tallied, their analysis and subsequent interpretation will be both fast and effective.

For example, a biologist studying the effect of radiation on white mice takes a group of 30 mice, 10 of which will form the control group and 20 the experimental group. He divides the 20 into 5 groups, 4 in each. The groups of four are then subjected to different radiation doses. The setup would seem to be perfect: we have the experimental groups and a control group. But the statistician requires that all the mice be numbered. He then takes a table of random numbers and indicates which numbers are to go into what groups. He does this not only to exclude any purposeful choice by the experimenter (say, stronger mice for larger doses of radiation), but also to exclude any chance choice (for example, mice belonging to a definite set of offspring).

The choice of entities for an experiment must be perfectly random, so that even an apparently inessential cause could not lead to erroneous conclusions.

The same goes for inanimate objects as well. By way of an illustration, let us take quality control at an electron-tube factory. The qualitative index will be the service life of the electron tubes. If during some one month the factory produced 50,000 tubes of a certain kind with a service life of 500 hours continuous operation (over 20 days), we have a serious problem: how to check this index of quality for all tubes?

To check the useful life of the tubes, we select a certain number (a random test group) and test them under operation conditions for 20 days continuous service. During that time, the entire monthly output is kept

at the factory and does not go to the consumer. The reason is that the tubes may not conform to the specifications and the factory will not be able to guarantee the consumer 500 hours useful life. Now there are many different types of electron tubes, the warehouses are overloaded, the situation is critical.... If the method used is spot checking, then we have to know how many tubes should be tested and when the tests can be considered satisfactory.

About fifteen years ago I came up against this problem at one of our factories. The procedure was this. Ten tubes were selected from every month's output for the full 500-hour test. If during the test period, not one of the 10 tubes failed, then the situation was considered satisfactory and the batch was accepted for sale. But if even one tube failed, then there were meetings, discussions about the high percentage of defective goods, and all kinds of measures were taken to reduce it.

Yet it is easy to calculate that under that procedure the probability of shipping the consumer a very bad set of electron tubes is rather high. It turns out that if in the whole batch there are 5 per cent defective items, which is 5000 out of 100,000, then the probability of extracting 10 nondefective tubes in a random sampling is equal to 0.6. What this means is that on the average 60 batches out of 100 will be assumed non-defective. Now if the output contains 10 per cent defective goods, then on the average—with this rejection procedure—34 batches out of 100 will be regarded as nondefective. The consumer will hardly be satisfied with such low quality.

It would be a pity of course if the consumer used such unreliable tubes in his TV set or radio set. There would be no end of bad language used, but the factory heads would probably never hear it. But when the consumer is another factory using special devices and hundreds

of electron tubes, there would be trouble for sure.

Let us make a very approximate calculation. If the batch of tubes we are using contains 10 per cent defective tubes and the times of failure are uniformly distributed throughout the time interval of 500 hours, then the probability that a tube randomly chosen from the batch will fail during 24 hours comes out to roughly 0.005. Suppose that a control device uses 300 tubes. Then the probability of an event consisting in one of the 300 tubes of our control device not failing during 24 hours will be equal to $(1 - 0.005)^{300} = 0.995^{300} \approx \approx 0.2$. Thus, the probability that during 24 hours at least one of the 300 tubes will fail is equal to $1 - 0.2 = 0.8$.

If the failure of at least one tube in the control device results in errors of operation or in total failure of the whole device, then our calculation shows we have a catastrophic situation: only two days out of ten will the apparatus operate troublefree! Only suicide-prone passengers would risk travelling in an airplane controlled by apparatus with that kind of reliability.

Let us return to the procedure for choosing tubes for useful-life control checks. Suppose the task of a factory is merely to fulfil the plan and it is not responsible for the quality of the goods. The factory will then attempt, on purpose or otherwise, to choose the more reliable tubes for control tests. It might be done this way: night and day shifts turn out goods of different quality. Hence, choose the tubes manufactured in the day shift. Some operations are manual and so the output of the more qualified personnel must be chosen.

After control of this kind the consumer will, generally speaking, be getting some very low-quality goods, worse than in the case of the unbiased choice of 10 electron tubes.

In the foregoing procedure, it is of course necessary

for the sake of reliability to choose more than 10 tubes for control. But here again organizational problems arise. There are many kinds of tubes and if we take, say, 1000 tubes from each batch and keep them under test for 20 days, we will need space and we will consume a large amount of electricity. And what is more, for the most part the tested tubes are lost. All this means production losses.

A statistician in such a situation would require a test-control selecting procedure that would ensure an unbiased sampling. The procedure itself could also be improved substantially.

Today statisticians can already offer a more suitable *procedure for taking a decision* on the defectiveness or otherwise of a batch of goods.

Stop for a moment. Note the italicized words. Up to now they have been merely implied, but let us examine them in more detail.

DECISION MAKING

Each one of us has to make decisions at every step. Heads of factories, laboratories, departments, military leaders, members of governments, and team leaders have to make decisions of an organizational nature. A doctor makes a diagnosis, prescribes medicine, suggests a method of treatment or sends the patient back to work. A motorist or pilot takes decisions in selecting routes, increasing or decreasing speed, or putting on the brakes. A scientist takes a decision when he chooses a procedure for his experiment or proves a theorem, or concludes an experiment.

In going to the registrar's office to get married or to the court for a divorce, we first come to some decision, which unfortunately is not always well thought through. Incidentally, in many other cases too the decisions we



come to are insufficiently justified, weakly reasoned, and unreliable. The stumbling block here is not only light-mindedness and a lack of wisdom but also a lack of information. And when the information is unreliable or simply erroneous, then one has to resort to mere conjecture.

To summarize, we can say that rather frequently decisions are taken in conditions of uncertainty. In extremely rare instances (may be never) do the measurements or investigations themselves serve as the starting point of an experiment or the subsequent analysis of the data. Every research worker, whether physicist, engineer, physician, plant-breeder, or sociologist, conducts an experiment and draws conclusions with incomplete information at his disposal, with randomness present, and under conditions of uncertainty. Such decisions are highly diversified: to drill for oil in a given region, to consider a given region prospective for oil development; to consider streptomycin effective in treating pneumonia; to accept a batch of electron tubes if the number of defective tubes does not exceed 2 out

of 50; to consider the velocity of light in vacuo equal to 2.99793×10^{10} cm/s; to consider that a new elementary particle has been discovered; to go over to a new system of economic incentives for factories; to effect a five-fold reduction in enrollment in correspondence departments of engineering colleges.

For an experimenter to be able to choose the best (or at least a sufficiently good) decision from among all possible decisions, he must know the rules of selection and be guided by them.

Today, science in a number of cases is able to indicate the rules or, as the accepted term is, the strategy for making the best (or a sufficiently good) decision. Under other circumstances, there is no such strategy, but there are certain recommendations on how to pose questions in a more reasonable fashion, how to construct a suitable mathematical model of the situation and how to study the model.

A number of areas of mathematics are closely related to problems of decision making. These include the theory of games, the theory of optimal planning and control, the theory of operations and others. But the most important ones are *probability theory and mathematical statistics*.

Mathematical statistics does not only study procedures for analysing experimental findings but also elaborates methods for taking decisions under conditions of uncertainty, the uncertainty being such as is characterized by statistical stability.

It is well to bear in mind, of course, that decision making is based not only on statistical considerations. There is a great difference in whether a mistake is made in 10 cases out of 100 by a schoolboy solving an algebraic problem, a geophysicist interpreting the presence of oil in a stratum, or a surgeon to whom you have entrusted your life. Again, the difference is great between

putting an unreliable electron tube in your TV set or in apparatus of the aircraft you are flying in. And the probability of the tube failing in one hour is the same!

Thus, the statistician who gives the rules for taking a decision, and the experimenter who applies these rules must take into account the effects of his decision. An incorrect decision can mean loss of time and money involved in the experiment, harm done to society and the loss of one's reputation.

For this reason, one should never try to save on time and energy when working out the rules for decision making in conditions of uncertainty.

I spoke earlier about the unavoidability of taking decisions at every step, whether in ordinary workaday life or in scientific endeavours. The reader of course does this rather well, otherwise he would not have the time or desire to read this book. He relies on his common sense and his intuition that rarely lets him down. Let us put both to the test.

INTUITION AND BIRTHDAYS

If the birthday of one of your acquaintances coincides with yours (even if you are twenty and he is fifty) you are surprised. This is a rare event, you believe. I have known a couple who got acquainted on their common birthday. The fact that their birthdays fell on the same day already augured well for the future, but that they got acquainted on that day was surely a doubly good sign.

Imagine a hall with several hundred people, say at a lecture. Let us conduct a thought experiment: we ask every person present for his birthday and then note in pairs, triples, quadruples, etc. their common birthdays. But first let's make a bet. I pay you one ruble if there is not a single pair with a common birthday.

and you pay me a ruble if at least one such pair is found.

How many people must there be in the hall so that the chances of you and me winning are equal? Note that if there are 367 persons present then there will definitely be at least one pair with a common birthday. And you lose out. Indeed, there are 366 days in a year, and, generally speaking, there may be 366 persons with distinct birthdays (from 1 January to 31 December). Now there is no place at all for the 367th person, and so we get our first pair.

If there are only two people present—you and me—then there is little chance of our having a common birthday, so I will hardly win my ruble.

Now show your will power and do not look at the next few pages and answer (within five minutes) the question that has been posed.

Remember the number you gave. Of course, a few simple calculations yield an exact answer. To illustrate the method, let us solve a simple problem.

Write the word *школьник* (schoolchild) on a piece of stiff paper, and then cut out the separate letters, turn them over and shuffle them thoroughly like you would do in a game of dominoes (Fig. 87). We will now take one letter at a time and put them together.

What is the probability of three letters in succession yielding the Russian word *кoл* (which means “one”, or a failing mark at school)?

It is easy to calculate this probability if we assume that all letters have the same chances of being chosen. The probability that the first letter will be ‘K’ is clearly equal to 2/8 (‘K’ comes on two of the eight squares). Then, of the remaining seven letters the probability of choosing the letter ‘O’ is 1/7. Finally, the probability of choosing the letter ‘Jl’ out of the remaining six letters is equal to 1/6. Thus, the probability of building up



Fig. 87

the word *кол* is equal to

$$\frac{2}{8} \times \frac{1}{7} \times \frac{1}{6} = \frac{1}{168} \approx 0.006$$

This is a very small probability. Failing marks at school are definitely more frequent.

Now let us return to our bet. At lectures in probability theory I have repeatedly asked my listeners the same question. The answers are different: 100 persons, 150, 183 (which is 366 : 2). Nobody ever gave the number to be less than 50. Then each was asked (this doesn't take much time) and in an audience of 80, 50 and even 30 persons there were invariably several pairs with the same birthdays. This always made a big impression.

Now let us carry out the calculations. We will first compute the probability of the opposite event: all n listeners have different birthdays. For the sake of simplicity, we will assume the year consists of 365 days (in other words, we ignore the rare event of a birthday on Feb. 29). We will assume that each person can be born on one of the 365 days of the year and all possibilities are equally probable.

Number of persons present	Probability of coincident birthdays for at least two persons	Approximate condition for an honest bet
5	0.027	
10	0.117	
15	0.253	
20	0.411	70 : 100
21	0.444	80 : 100
22	0.476	91 : 100
23	0.507	103 : 100
<hr/>		
24	0.538	116 : 100
25	0.569	132 : 100
30	0.706	242 : 100
40	0.891	819 : 100
50	0.970	33 : 1
60	0.994	169 : 1
70		1200 : 1
80		12 000 : 1
90		160 000 : 1
100		$33 \times 10^5 : 1$
125		$31 \times 10^9 : 1$
150		$45 \times 10^{14} : 1$

The first person has the possibility of being born on any day: $\frac{365}{365} = 1$. The probability that the second person will be born on a day different from the first is equal

to $\frac{364}{365}$ (one day out of the 365 is already occupied). The probability that the third person will be born on some day not occupied by the first and second is $\frac{363}{365}$. The rest of the calculations are clear. The probability of a joint realization of all these n events, that is, the probability that no two persons out of n present have the same birthday is

$$Q_n = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \cdots \frac{365 - (n+1)}{365}$$

The probability of the event that at least one pair will have a common birthday is

$$P_n = 1 - Q_n$$

If we use the formulas for a series of values of n and carry out the calculations, we will get the numbers in the second column of the table.* The third column contains the values of n for conditions of an honest bet, that is, one in which the relationships between the stakes of the participants for which their average winnings would be the same. It is easy to compute this value:

$$\frac{P_n}{1-P_n}$$

From the table it is evident that the answer to the question posed at the beginning of this section is quite unexpected: our bet can be considered (approximately) an honest bet if there are 23 persons in the hall. Then the probability that there will be at least two persons with the same birthday is roughly equal to the probability that there will be no such pair.

* The table is taken from *Introduction to Finite Mathematics* by J. G. Kemeny, J. L. Snell, G. L. Thompson, 1957.

Now if there were 100 persons in the hall our bet would be honest if I staked 3,300,000 rubles to your one ruble. Now you see how hopeless it would be for you to win in the case of equal stakes (a ruble for a ruble).

INTUITION AND LUCK

Lateness, losses, unwanted encounters, unlucky marriages, bad weather and hapless fishing trips all plague humanity. But whereas the weather and fishing luck have little to do with one's personal efforts, being late or losing at games, or unhappy marriages are simply matters of one's being unlucky. When somebody is said to be lucky or his luck has run out, then we regard these words as merely a turn of phrase. Incidentally, if we dig a bit in our memory we are sure to dredge up periods of luck and then stretches of no luck at all.

Perhaps other less brilliant situations have simply been forgotten, or is this the way things actually happen? It is hard to say; one would have to observe events in a serious fashion for many years. However, for games of chance or for the more serious problems of the theory of diffusion, the question of luck or, to be more exact, the problem of the correspondence between our intuitive conceptions and the actual state of affairs can be investigated in sufficient detail.

In the accompanying table we find the number of "heads" that turn up in coin tossing (the coin is considered to be unbiased). Each two-digit number stands for the number of heads that turn up in a series of 100 tosses (the total number of tosses is 10,000). The first column indicates the number of trials, the last, the number of heads that turn up in the appropriate series of a thousand tosses. Heads turned up in 4979 cases out of 10,000 tosses. This number would appear to be satisfactory and the coin would seem to be unbiased.

Number of trials	Number of heads	Total number of heads
0—1 000	54 46 53 55 46 54 41 48 51 53	501
—2 000	48 46 40 53 49 49 48 54 53 45	485
—3 000	43 52 58 51 51 50 52 50 53 49	509
—4 000	58 60 54 55 50 48 47 57 52 55	536
—5 000	48 51 51 49 44 52 50 46 53 41	485
—6 000	49 50 45 52 52 48 47 47 47 51	488
—7 000	45 47 41 51 49 59 50 55 53 50	500
—8 000	53 52 46 52 44 51 48 51 46 54	497
—9 000	45 47 46 52 47 48 59 57 45 48	494
—10 000	47 41 51 48 59 51 52 55 39 41	484
0—10 000		4979

Take a good look at this table. Your reaction could well be summed up in the words: "So what?" I will now suggest a game of coin tossing. Invite a friend and take any coin at hand. Big-size coins are more fun. One of you start tossing the coin and continue for a long time. It will be convenient to assume that the tosses occur at equal intervals of time. If heads, you win one kopeck, if tails, your partner wins and you pay.

I hope you don't have the idea that I am pushing you into gambling. The stakes are small (1 kopeck) and so you can play for a good long time without any worry of being ruined. Also, wins or losses shouldn't distract your attention from the scientific aspect of the game.

It is quite clear that heads and tails will alternate in some kind of irregular manner. But we are not interested in what turns up on a definite toss, say the two hundredth, but rather what the sum of your winnings or losses is up to that point in the game. It is this over-

all gain and not the gain in some given toss that I will now discuss.

Suppose that your partner tosses the coin 200 times in a row and during that whole time you are never ahead (as to total sum of points). Will you regard this merely as bad luck or will you suspect your partner of cheating? If your partner is above suspicion, then perhaps you should explain this apparent unfairness as due to the coin being biased, in which case you can take a new coin.

On the other hand, maybe you figure 200 tosses is too small a number to discuss unfairness.

Let us say your mood gets blacker. You continue playing but suspicions build up. The coin is tossed for the thousandth time and still you have never been out ahead in total sum of points. How do you assess the situation?

You begin to suspect your partner. But what real cause is there for your suspicions? If the coin is symmetric and your partner is not cheating, then in every toss of the coin the chances are equal for heads and tails. Common sense tells you that with a long enough series of tosses, each player will be ahead just about half of the time.

Very convincing, but very wrong too!

Let us say that the leader is the player who at a given instant is ahead (as to total sum of points). It turns out that leadership in the game changes much more rarely than your intuition would suggest. No matter how large the series of tosses is, the most probable situation is that there will not be any change of leadership at all—one change is more probable than two, two is more probable than three, etc.

An ordinary interrogator (or psychologist) would be inclined to qualify most of the players as cheaters and most of the coins as biased. However, if we take

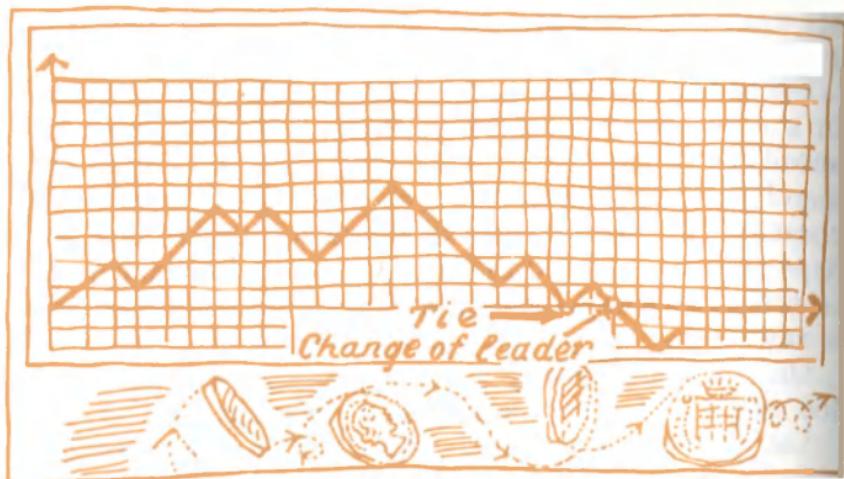


Fig. 88

a thousand coins and toss each coin 10,000 times, then most of the thousand coins will behave so that one of the players will be winning almost all the time. And only with respect to a few coins will the leadership pass from one to the other, as one would expect of an unbiased coin.

For the sake of pictorialness, let us depict the game in the form of a graph. The horizontal axis will indicate the tosses, and the vertical axis, at these points, the overall gain. If all this is plotted on graph paper, the game will take the form of a broken line, where the ordinates of the nodes of the square grid denote overall winnings for the appropriate toss. A typical graph of such a game is shown in Fig. 88.

Different graphs of this type represent possible outcomes of the game. In this game, a change in leadership is always preceded by a tie, that is, a situation in which the sum of the winnings of both players is equal to zero. Incidentally, a tie is not always followed by

a change of leader. This only occurs with a probability equal to one half.

You will of course agree with the following natural assertion: during two days of coin tossing at equal intervals there will be twice as many ties as during the first day.

But again this is not true!

It turns out that the number of ties increases as the square root of the time. This is hard to believe. But perhaps you will be convinced by quantitative data. I will have to make use of a characteristic of probability distributions that is called a *median*.

You remember the median in a triangle, I'm sure. It's the line that bisects the opposite side of the triangle. Also, in the theory of probability, the median (denoted M_e) is a number that divides the distribution of probabilities in half (the probability that a random quantity will assume a value less than M_e is equal to one half, just as the probability that it will assume a value greater than M_e is equal to one half).

In the problem of coinciding birthdays I asked you what the median number of persons was for which at least one pair has a common birthday. As you will recall, the median here turned out to be approximately 23.

Calculations show that the median in the number of ties in 10,000 tosses is equal to 67, but in one million tosses it becomes 674, which is a 10-fold increase and not a 100-fold increase, as one would expect from the common-sense viewpoint.

To corroborate these results which run so counter to our intuition, let us examine the experimental material. I have taken this material, like almost all of the factual material of this section, from a marvellous book by William Feller entitled *An Introduction to Probability Theory and its Applications*.

Let us return for a moment to the table at the beginning of this section, the one which caused us to exclaim "So what?" This table was compiled on the basis of an actual experiment.

To toss a coin 10,000 times requires about ten to fifteen hours. William Feller is a famous modern mathematician and of course did not spend all that time tossing a coin, as Buffon did in the 18th century. Instead of coin tossing, we can perform any other type of experiment with two equally probable outcomes. Such experiments are easy to run on high-speed computers, where in place of heads and tails we have the equally probable digits 0 and 1. All one needs is less than a minute of machine time for an experiment involving 10,000 "tosses". We give the results of one of these experiments below, but I will continue to speak of a coin, the overall gain and loss and change of leadership.

This experiment, the results of which are given in the table, involved the following changes in leadership.

First player was leader in	Second player was leader in
the first 7804 tosses	
the following 2 tosses	the following 8 tosses
the following 30 tosses	the following 54 tosses
the following 48 tosses	the following 2 tosses
the following 2046 tosses.	the following 6 tosses

In all, during 10,000 tosses the first player was the leader in 9930 tosses, and the second player in only 70 tosses.

The first player was extremely "lucky", as you can easily see. But this pattern is more the rule than the exception. On the average, one out of ten such experiments leads to results in which one of the players is in a still worse condition than the second player in our experiment.

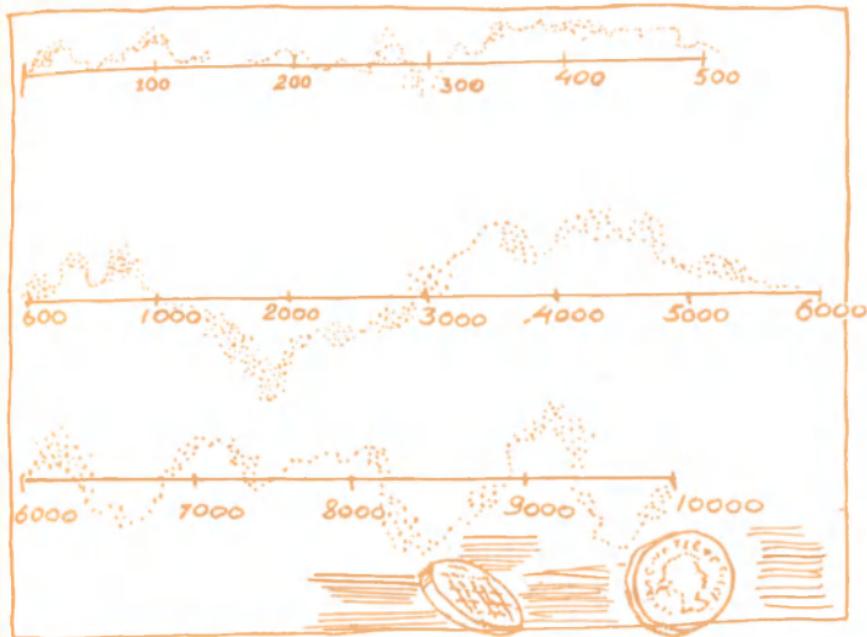


Fig. 89

Fig. 89 shows the graph of a similar experiment. On the horizontal axis we have the number of tosses, on the vertical axis, the overall gain of the first player. It is clear that a negative gain is equal to a loss by the first player, and hence, a gain for the second player. The graph has 142 ties, of which only 78 represent a true change in leadership. In the earlier described experiment there were 14 tied situations, of which 8 led to a change in leadership. It turns out that with 10,000 tosses, the probability of more than 140 ties is equal to 0.157, and the probability that there will be less than 15 ties is equal to 0.115.

Note how all these results contradict our intuition! I don't know whether this will console you any, but it would seem that "luckless" periods giving way to

"lucky" periods does not represent something out of the ordinary, but is rather in the way of a regularity.

Here is another instance where one's intuition fails. An eighteen-year-old girl is ordinarily eager to get married. At any rate, Elvira takes a very keen interest in the chances of her girl friends getting married. It is not so much that she herself is not yet the lucky one, but that she suffers so much when others are getting married and she is not. We can sympathize with the girl, but actually it is not in the least a case of luck. Everything is quite law-governed. A mathematical model of the situation presumes complete equity: firstly, the marriages of any number of girls have no effect on the marriages of any others, and the chances of getting married are the same for any age group; secondly, the waiting times (from age 18 to the happy event) are random variables with the same probability distribution.

Thus, there are many girls about to get married. We will say that there are infinitely many. Suppose that $(n - 1)$ of them are married, and then, finally Elvira gets married too and takes up the n th position.

Not too complicated reasoning and computations show that the probability of this event is $\frac{1}{n(n+1)}$, which means that the mean waiting time for Elvira is equal to infinity. Therefore, one should say that luck was with Elvira if she married fifth, one hundredth or ten thousandth. Do not speak of unlucky Elvira if next Saturday evening—or on following evenings—she does not pass by with her fiance.

RANDOM WALKS

About three hundred years ago, the Dutch trader Anton van Leuwenhoek, a self-satisfied ignoramus but a very inquisitive and persistent man, saw life through

the lenses of microscopes he himself had made. He found minute animals swimming about in rain water, animals a hundred times smaller than any being visible to the naked eye.

A hundred and fifty years after the extraordinary discovery of Leuwenhoek, the English botanist Robert Brown examined teeming microscopic life through the eyepiece of a rather powerful microscope. He noticed the random jumps and dancing of minute particles of pollen. Brown was educated and knowledgeable, he realized that what he was observing was not the movements of living beings but merely dust particles floating in the water. What is more, Brown did not confine himself to generalities. To prove the fact of these unaccountable motions he studied the behaviour, in liquids, of particles of an enormous number of objects, including a fragment of the Sphinx. He found a piece of quartz, inside which was a water-filled hollow. He placed it under his microscope and saw haphazard motions of particles suspended in the water. The water had probably gotten there a long time ago but the particles continued their dance. The year was 1827.

It was no easy matter to account for the random motions of minute particles in liquids.

The universal nature of the effect produced a great impression on Brown and he thought that he had discovered some kind of elementary form of life characteristic both of organic and inorganic matter.

By the end of the 19th century a variety of hypotheses—that the nature of the Brownian movement is connected with some kind of electric force, the evaporation of the liquid, or mechanical impacts—had been refuted. The Brownian movement was invariably detected after a sample had been in total darkness for a week or after it had been heated for many hours.

It finally became clear that the phenomenon itself—

Brownian motion—is of fundamental significance. The many experiments all pointed to one natural conclusion: the cause lies in the random bombardment of particles by the molecules of the containing liquid. But it required the genius of Einstein to analyse the problem in a clear-cut and unambiguous fashion.

Let us try to clarify a few of the questions that involve Brownian motion. We already know how to approach the problem. We have to construct a convenient model of the phenomenon and then also a mathematical model. A particle suspended in a liquid is subjected to a bombardment by the molecules of the liquid. The force of each impact differs (this is because the molecules move with different velocities and the impacts come from all sides), the directions are accidental, and the chances of being hit from the right, left, or from below or above are all the same. The number of collisions between a particle and the molecules is very great, of the order of 10^{14} per second. Incidentally, the absolute numbers of collisions and molecular velocities are inessential for the actual construction of a model of the phenomenon.

Let us try to determine how much the position of a jumping particle changes during an interval that is many times greater than the interval between two collisions.

We will construct a model of this phenomenon. We assume, firstly, that the molecular velocities are all the same in magnitude; secondly, that the impacts occur at equal intervals of time (if there are roughly 10^{14} impacts per second, we will assume that the impacts occur in a time interval $\tau = 10^{-14}$ second; for the scale unit we will take the time interval between two successive impacts); thirdly, we will assume that the particle moving in the liquid is sphere-like in shape.

The equal probability of various directions of motion

of the molecules means that if we isolate two equal areas on the surface of a sphere (the areas need not be of the same shape!), then the probability of any molecule hitting each of the areas will be the same. The probability itself of a molecule hitting any isolated region is equal to the ratio of the area of the isolated region to the entire surface area of the sphere. Such a distribution of collision directions is called a uniform distribution.

Besides we will assume that the various molecules hitting arbitrary nonoverlapping areas are independent events.

Under these assumptions, each successive step of a particle is independent of the preceding step and the steps are equal in magnitude, although their directions are random and uniformly distributed.

Let us now pass from the three-dimensional model to a two-dimensional model. The behaviour of a particle in a plane is similar to the behaviour of a drunkard meandering on a city square. He is tipsy to the extent that each step is made at random to some side, with equal chances for any direction. Here too, the direction of the next step is totally independent of the preceding steps. We will term this man to be an absolute drunkard.

Where will he find himself after a certain time interval? Of course, neither he himself nor you nor I know, nor can we make a prediction.

True, sooner or later he will probably find himself in the cooler, but let us not be distracted from a direct answer to our question. The absolute drunkard is a model of Brownian motion. We might change the image to a flea in a large empty hall with no distractions.

The absolute drunkard is capable, in n steps, of moving in some direction, and we can estimate the distance he will cover from the point at which we first

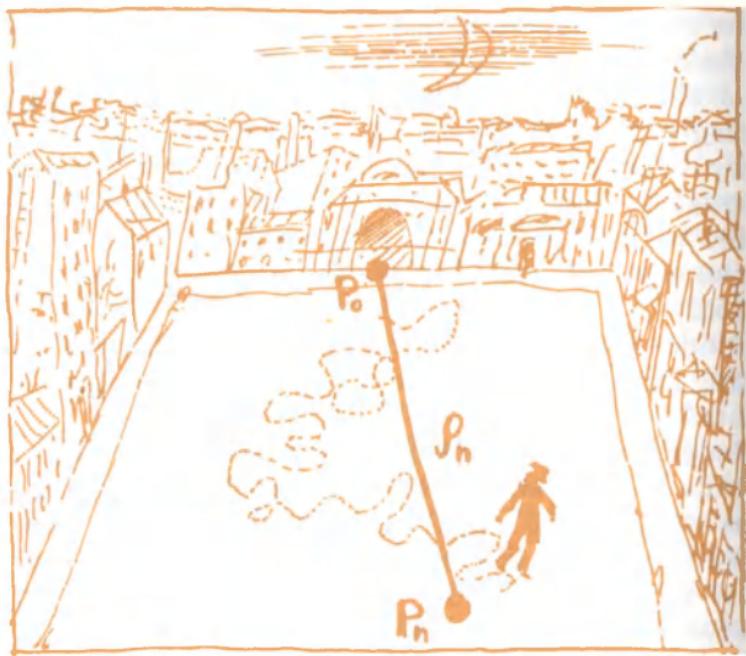


Fig. 90

found him. The distance ρ_n (see Fig. 90) between the initial point P_0 and the terminal point P_n of his walk (in n steps) will of course be a random variable. But what is the average magnitude of this distance, ρ_n ?

We can compute the quantity ρ_n on the basis of the initial assumptions. But before indicating this quantity I would like to simplify the model still more. By simplifying I mean reducing the number of coordinates or degrees of freedom.

Imagine the same drunkard in a narrow corridor, so narrow that he can only move forward or backward. His behaviour remains the same; that is, each step he takes is independent of the preceding steps and there is an equal probability that he will step backward or forward. His steps are all of the same size, and

the step size is l , then with each step he moves away from the starting point (or approaches it) by the amount $\pm l$ with probability $1/2$ (Fig. 91).

We want to know how far the drunkard will get in n steps. A drunkard in a corridor represents a one-dimensional random walk of a particle. A drunkard on a square is a model of a two-dimensional random walk, and the Brownian motion of a dust particle in a liquid is a three-dimensional random walk.

The one-dimensional model of a random walk can be rephrased immediately to yield the coin-tossing model that has already been discussed. Indeed, if you toss a symmetric coin at equal time intervals and your partner pays you l kopecks for heads and you pay him l



Fig. 91

kopecks for tails, then the sum of your winnings after n throws will be equal to the difference, multiplied by l , between the number of heads and tails that turn up. Numerically, it is exactly equal to the distance the drunkard covers in n steps, so that the distance is equal to the difference in the number of steps to the right and to the left multiplied by the size of one step.

You will recall the surprise you experienced while reading the preceding section when we discussed the problem of the length of leadership periods or the number of ties in coin tossing. We will get similar results here too.

Since the probability of steps forward and backward is the same and the steps are independent, on the average there will be the same number of steps forward and backward, and hence the mean distance the drunkard covers in the corridor is equal to zero. On the average, the absolute drunkard remains in one place.

Let me explain what this means. We will follow a large number of wandering particles. For each one we will record the position it occupies after n steps. We will obtain both positive and negative numbers. But the average (that is, the sum divided by the number of particles) will be close to zero. The mathematician would say "mean value" (or, in probability theory, the mathematical expectation) of the distance covered by a particle in n steps is equal to zero. But we are interested in the estimate of possible deviations of these distances from the mean.

In terms of coin tossing, the same statement reads thus: the expectation of a win by each of the players in a fair game is equal to zero. But we are interested in estimating the number of possible wins.

Denote by ρ_n the distance between the initial position of a particle and its position at the n th step. We could also study the absolute value of ρ_n (or, to put it

differently, of the winnings of either player). But it is more convenient to compute a different positive quantity called the square of the distance covered, ρ_n^2 (the square of the winnings).

The mathematical expectation of the square of the distance (that is, the mean value of this quantity when observing a large number of "wandering" particles) will be denoted by R_n^2 , this is no longer a random quantity but a deterministic quantity.

It turns out that R_n^2 —a quantity with the dimensions of the square of the distance—is proportional to the square of the step length l^2 . It is also clear that the quantity R_n^2 depends on the number of steps n , and at first glance it would appear to be proportional to n^2 . However, it may be demonstrated that, using the independence of the steps (or the independence of outcomes in a series of coin tosses), R_n^2 is actually proportional to the first power of n , and the expression looks like this:

$$R_n^2 = nl^2$$

If during unit time there are k steps of magnitude $\pm l$, then the mean deviation of the particle from its initial position during time t , R_t^2 , will be proportional to the time:

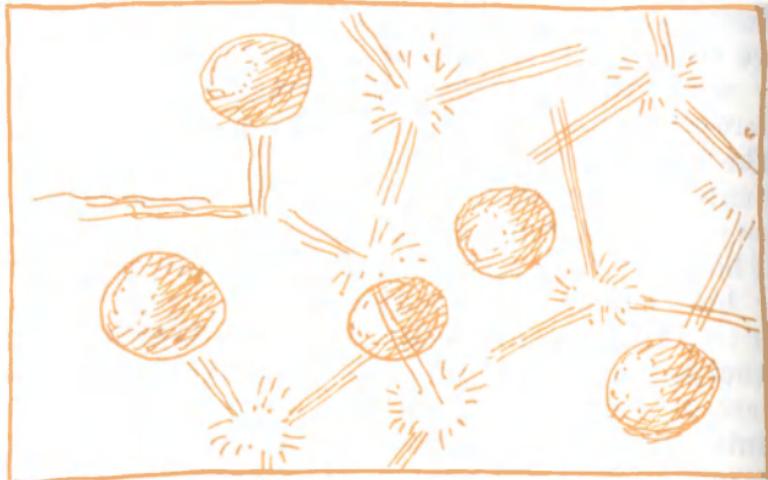
$$R_t^2 = ktl^2$$

This quantity has the dimensions of the square of the distance, but we of course find it more convenient to measure distance in linear units (centimetres and not centimetres squared). The appropriate typical deviation of particle in n steps is then

$$\sqrt{R_n^2} = R_n$$

Similarly, the typical deviation of a particle in time t is

$$R_t = \sqrt{R_t^2} = l \sqrt{kt}$$



The proportionality of deviation of a particle to the square root of the number of steps, \sqrt{n} (or \sqrt{t}) (and not to the number of steps n) is a fundamental factor in studies of such statistical phenomena. When estimating the possible winnings in a game of coin tossing, we can say that the typical gain (or loss) in n tosses of the coin will be proportional to \sqrt{n} .

As you will recall, the number of ties also increases in proportion to \sqrt{n} .

The model of a random walk has numerous interpretations. When motocars leave the centre of a large city at the end of a working day, the choice of route by each driver may be taken to be random.

The reader may be indignant to hear that the advance in his career is of a random nature and that this advance is close to a random walk. But we are of course talking about the careers of other people and we will not object to the proposed mathematical model, all the more so since we will never meet these people.

This paradoxical situation is a tribute to human vanity. It is illustrated brilliantly in the trivial argu-

ment about whether a machine can think or not. People far removed from mathematical thinking argue till they are hoarse about human beings having the sole right to think; they do not take the pains to define precisely what the argument is about and what the starting concepts actually mean (for example, such notions as "machine", "think", "ability to").

But let us return to the stream of motorcars leaving the city.

From the viewpoint of a transport engineer or a mathematician working on the problem of departure of motorcars, the simplest model of such traffic is a random walk. It is simpler and more convenient to regard the route of each separate car as being random than to attempt to predict it (although for the driver it is not random). A model like this is acceptable (at least to a first approximation, as the mathematician would say). Refinements may be needed later on, but this is a good beginning.

By considering traffic as a statistical (diffusion) problem, it is possible to work out the characteristics of streets that ensure a clearing of traffic during peak hours. And if as a result, returning home at 40 km per hour instead of 15, you do not get excited every time you have to wait for a red light, then perhaps this will soothe your hurt feelings about the accidental nature and indeterminacy of your route.

As you approach a T-intersection, you will have to turn either right or left. In constructing the model that interests us, it will be convenient to assume that the choice of direction of turn by each driver is accidental and the direction of turn of successive cars is independent of preceding ones. Let us also assume that left and right turns are equally probable for every motorcar. In this model we can, for example, estimate quantitatively the possible excess of cars turning right

to the number making left turns. This problem is quite similar to that of the behaviour of an absolute drunkard in a narrow corridor or to estimating the winnings in a coin-tossing game.

I have just mentioned the possibility of a statistical consideration of traffic and in parentheses wrote "diffusion". This was not accidental. The familiar diffusion of atoms or molecules can be studied with the aid of the same mathematical model.

Let us examine the motions of the molecules of a gas. This cannot actually be done, of course, but I hope you will make good use of your imagination. Take a freely moving molecule with nothing in its path. Suddenly it meets another molecule, a collision takes place and they fly off in different directions. The same thing happens when two billiard balls collide, only molecules move in three-dimensional space and not on a surface. Collisions of molecules take place very often (at normal pressure) and the mean distance l between collisions—called the *free-path length*—has a definite small value.

Let us now imagine that all distances between collisions are the same and are exactly equal to the free-path length.

Now the motion of our molecule will be very much like the behaviour of an absolute drunkard on a city square: it will move in steps of definite length, the direction of each succeeding step being random and uniformly distributed and also independent of preceding steps. The only difference is that the drunkard moves on a surface, while the molecule moves in space. But this will not prevent us from computing exactly how far the molecule will move in a specified time.

Electrons in solids participate in the thermal motion of the substance. Now let us examine, for example, ^{an} elementary oscillatory circuit consisting of a capacitor,

resistor, and inductor (coil). The thermal motion of the electrons gives rise, on the plates of the capacitor, to a time-varying (in magnitude and sign) electric charge, and, in the inductor, to a varying electric current. The mechanism of this phenomenon may be pictured as follows. The random thermal motion of the electrons in the substance of the circuit is equivalent to the action of very small and frequent randomly alternating (in magnitude and sign) electric pulses (short-term electromotive forces). These chaotic oscillations of charge or current are electrical fluctuations called thermal noise.

The magnitude of thermal noise is very small. At room temperature the equivalent fluctuation current that may be recorded in a resistor during one second is equal to approximately 10^{-10} ampere. Proceeding from general arguments, Einstein predicted such fluctuations but they were detected experimentally only 20 years later.

A similar phenomenon occurs in electron tubes. The stream of electrons moving from the cathode to the anode is also of a random nature: the number of electrons reaching the anode in unit time fluctuates. For ordinary currents of the order of a milliampere, roughly 10^{16} electrons pass from the cathode to the anode in one second with a transit time of 10^{-9} second. Deviations from the "mean current" that constitute the fluctuation component go by the name of shot effect, or fluctuation noise.

The nature of the fluctuation noise is such that the same model of Brownian motion helps in a quantitative study.

The thermal noise in conductors and the fluctuation noise in tubes are the Achilles heel of all radio engineering. These noises are heard during intermissions in radio broadcasts when the radio set is tuned to a station.

It is precisely the presence of fluctuation noises

(which cannot be eliminated because their cause lies in the discrete nature of electricity) that restricts the range of radio communication, the possibilities of radar and television and other spheres of radio engineering.

Whereas 20 to 25 years ago specialists in radio reception, radar and radio navigation had little interest in the methods of probability theory, during the past 15 years, as radio engineering approached the limits of precision and distance, the methods of probability theory have become one of the basic tools in the hands of communications, radio and automation specialists.

THE DRUNKARD'S WALK

Let us return to our drunkard in the narrow corridor. Suppose he sees another drunk in the distance. The contradictions that tear him are obvious. As before, he performs only steps forward and backward. But still and all he is more drawn to his fellow drunk, and we will describe his movements by saying that although his equal-magnitude steps forward and backward are random and independent, the probability p of taking a step forward is greater than the probability $q = 1 - p$ of taking a step backward. In this situation, the drunkard will, on the average, gradually and slowly move forward.

This mean shift is proportional to the product of the number of steps into the difference of the probabilities of taking a step forward and backward.*

* It is easy to compute that the mean (mathematical expectation) of his location in n steps will be equal to $S = (p - q) ln$.

If there is a great attraction to his fellow drunk so that, for example, $p = 0.9$, then in 100 steps the expectation of his location will be $S = (0.9 - 0.1) \times 100l = 80l$. If the attraction is slight, then the probability p will differ but slightly from $\frac{1}{2}$. For instance, for $p = 0.51$, we will have $S = (0.51 - 0.49) \times 100l = 2l$.

Our drunkard of course has nothing to do with all this. This is a model of a one-dimensional random walk in the case of a force acting on a particle in such a manner that the particle moves forward more often than backward.

It is easy to compute not only the mean deviation of the particle from its initial position but also the typical deviation. As in the case of a symmetric random walk, it too is proportional to the length of the step, to the square root of the number of steps, but besides it is proportional to the square root of the product of the probabilities of steps taken forward and backward, so that the final formula (in the notation of the preceding section) is of the form

$$R_n = 2l\sqrt{pq n}$$

Nonsymmetric random walks serve as a good mathematical model for many processes. The problem we considered earlier when estimating the traffic at a T-intersection is best regarded as a nonsymmetric random walk because the turns at such an intersection are ordinarily governed by the necessity to reach other routes whose attraction for the drivers differs. Hence the right-turn and left-turn probabilities will differ and we can estimate these probabilities by the frequency of appropriate turns, which can be obtained by observing the traffic.

Also important are problems dealing with the diffusion of particles (atoms, molecules, dust particles) when there is a flow in a given direction. The diffusion of ions of gas in an electric field is an example. The study of this phenomenon reduces to studying the same mathematical model of a nonsymmetric random walk. Here, of course, the random walk is two-dimensional and even three-dimensional, but this only slightly complicates the problem.

THE RANDOM-WALK STUDENT

In the preceding sections, a random-walk particle after each step could only move to neighbouring points.

It is of practical interest to study the behaviour of more active particle that can jump across one or two or more steps.

Let us take another model. A schoolboy takes test in mathematics and can get one of five marks (1, 2, 3, 4, and 5, which is the highest). The tests are conducted once a week. Since new material is studied each week we will assume that the results of each test are independent of the preceding tests. Besides, we will assume that the results are random and have a definite probability. To be specific, we will accept the following law of distribution of probabilities of the marks:

Mark	1	2	3	4	5
Probability	0.1	0.2	0.3	0.2	0.2

Of course the parents will be indignant if they fear that the outcome of the test is accidental when it comes to their child (they of course know exactly what was wrong!). All the more so about the outcome being independent of the past. Have patience. This example is meant to bridge the gap (later) to the case of subsequent work being affected by earlier work. Now about the accidental nature of the outcome of the test. To some extent it is always accidental (random). The main thing however is that it is more convenient, easier and more promising to consider it accidental (like the turns made by motorcars) than to take into account numerous factors that affect the outcome of a test. And if our problem is to estimate the degree of the calamity

to which such behaviour on the part of the student can lead if he gets a failing mark of 2 (or if he gets a five—in which case his father has promised a new bicycle), then a justified forecast is just what we want.

I will find it more convenient here to use a different terminology. I will speak of the state of a system and will assume that the system under study can be in one of several possible states and can pass from one state to another at each step. Here, the system is the student its states are the weekly marks, and passage (transition) from one state to another is the getting of a new mark.

Getting the same mark will also be considered a transition (merely a transition into the same state). From this standpoint, the life of the student is a dreary, emotionless stretch of life without awards or punishment.

Passing from one state to another occurs in accordance with the distribution of the probabilities of the states (the marks). This is a random walk through a multitude of possible states.

This random walk can be conveniently represented in the form of a graph (see Fig. 92). On the horizontal axis we lay off the time and instants of transition, that is, the times at which the student gets a new weekly mark. (Only the numbers of the weeks are indicated since it is immaterial on what scale the time is measured.) On the vertical axis we lay off the marks (or the numbers of the states). Any graph like this will represent the actual sequence of marks, but the probabilities of different graphs will not, generally speaking, be the same.

There are different ways of getting the final mark for a quarter or for the whole school year. This final mark is a criterion of how the student studied and how ~~he~~ can be classed: as failing, average or brilliant. Diffe-

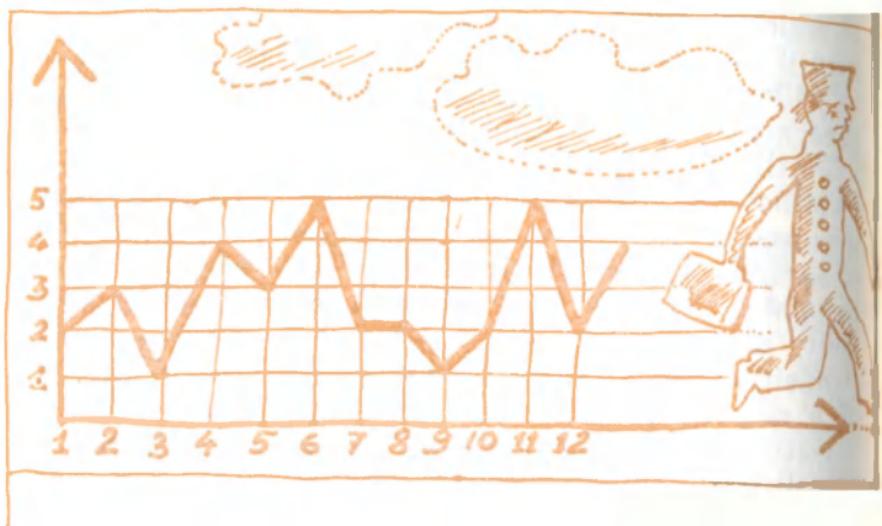


Fig. 92

rent teachers have different criteria. The simplest way is to take the average mark for the period. This merely amounts to the sum of all marks divided by the number of the marks.

The average mark of course gives us some idea of what the student has done during the period under consideration. However, different average marks for the quarter can result from one and the same set of probabilities recorded in the table. It may very well be that our student only got failing twos during the entire 10 weeks and so the average is 2. Let us recall our assumptions: every mark is accidental, independent of all others and obeys the given distribution of probabilities.

For this reason, the quarterly mark (the average for the quarter; we denote it by x) is a random quantity. And for our probability distribution the expectation (that is, the mean value of the quarterly mark; denote

it by \bar{x}) is equal to

$$\bar{x} = 1 \times 0.1 + 2 \times 0.2 + 3 \times 0.3 + 4 \times 0.2 + 5 \times 0.2 = 3.2$$

Which means that on the average such students will not be failing students.

What "on the average" means here is only that if we take the quarterly marks of a large number of students (a thousand, say) and compute the arithmetic mean of this thousand quarterly marks, then we get a number close to 3.2. But different students out of the selected thousand may receive a great variety of quarterly marks. It might even turn out that some student of the thousand never got a mark higher than 3 for all the 10 tests in a row. The probability of such an unpleasant situation is roughly equal to 0.006 so that on the average a situation like that will occur in 6 cases out of a thousand.*

It will be noted that the probability of getting an average mark below three will be considerably greater than the computed value since even after several fives and fours, a low average mark will result from a large enough number of twos and ones (which are failing marks). For example, from our graph it is evident that the average mark for the first 10 steps is 2.5, although there is one four and one five.

Thus, our student is barely scraping along and might easily wind up with a failing mark for the quarter. A slack pupil will be pleased of the opportunity to justify bad marks by referring to the appreciable probability of a failing mark, although he may be studying

* This can be computed very easily. The probability that a student will get a mark of 1, 2 or 3 is, according to the table, $0.1 + 0.2 + 0.3 = 0.6$. The probability that this event will occur 10 times in 10 independent tests is equal to 0.6^{10} , or roughly 0.006.

slightly above average. But we will not give him the satisfaction: his task, in reality, is to change the distribution of probabilities of various outcomes so that the probability of getting a failing two for the quarter is very small. To do so, he will have to study hard for tomorrow's math lesson.

Still and all, a teacher does not often get the quarterly mark merely by averaging. This would be too formal an approach, one that does not enable us to take into account the fact that the student may have begun to study hard and by the end may have mastered all parts of the course. So actually it is true that in mathematics the subsequent marks depend on preceding ones. There are several influencing factors here: the logical relatedness of various sections of the mathematics course, the faith of the student in his ability to catch up (or lack of faith), and the established attitude of the teacher towards the student. Although the results of each test cannot be predicted exactly (they are random variables), one has to allow for changes in the probability of certain marks for the current week being dependent on what marks were obtained in the preceding week. If last week's test was a failure (a 'one' or 'two'), then there is of course less chance to get a five than if the preceding test had been marked 'four' or 'five'.

In our terminology, a situation like this means that subsequent states depend on preceding states, that is the probabilities of subsequent states depend on the state that existed during the preceding step.

You of course rightly assume that every mark, though in itself accidental, depends on the whole preceding life of the student, on all his successes and failures in the field of mathematics and in other fields as well. But first let us make a study of a more primitive model and consider the probabilities of subsequent states (marks) as dependent solely on the states of the pre-

ceding step. In this kind of model, we can specify the probabilities of the marks at each stage by means of a table.

Preceding states.	Subsequent states				
	1	2	3	4	5
1	0.4	0.3	0.1	0.1	0.1
2	0.3	0.3	0.2	0.1	0.1
3	0.1	0.2	0.4	0.2	0.1
4	0.1	0.1	0.2	0.4	0.2
5	0.1	0.1	0.1	0.2	0.5

Here, the number 0.2 at the intersection of the second row and third column signifies that the probability of passing from state 2 to state 3 in one step is equal to 0.2. What this means for our student is that he will get a 'three' at the next test with a probability of only 0.2 if his mark at the preceding test was 'two'.

Ordinarily, mathematicians do not write the numbers of the states on the left and at the top. They merely write out a table of probabilities like this:

0.4	0.3	0.1	0.1	0.1
0.3	0.3	0.2	0.1	0.1
0.1	0.2	0.4	0.2	0.1
0.1	0.1	0.2	0.4	0.2
0.1	0.1	0.1	0.2	0.5

The above array is called a matrix. In this case, the numbers—the elements of the matrix—are called *transition probabilities* and the matrix itself is called a *matrix of transition probabilities*.

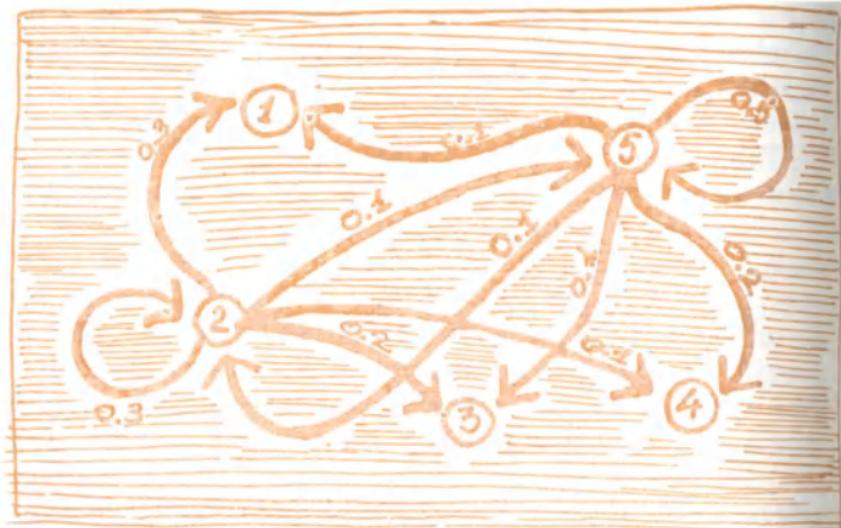


Fig. 93

Another and more pictorial way of representing transition probabilities consists in constructing a graph (see Fig. 93). The open circles indicate states, the arrows indicate transitions from one state to another, and the numbers attached to the arrows, the probabilities of these transitions. In Fig. 93 we see only the arrows of transitions from states 2 and 5 to all other states. The other arrows are not indicated because the web of lines would then be almost impossible to decipher.

For the sake of convenience (or due to tradition), mathematicians do not write the numbers of the states but the states themselves as letters with subscripts.

There are five states in our example: we can denote them thus: E_1, E_2, E_3, E_4, E_5 . This means that if a student gets a mark of 'four', then the state is E_4 . If he got marks as indicated in Fig. 92, then the sequence of transitions from one set of states to another set can be

written in the form of a chain thus:

$$E_2 \rightarrow E_3 \rightarrow E_1 \rightarrow E_4 \rightarrow E_3 \rightarrow E_5 \rightarrow E_2 \rightarrow E_2 \rightarrow E_1 \rightarrow \dots$$

Here, the probabilities of transition are specified by the matrix of transition probabilities.

Chains of random events or, in our terminology, chains of transitions of a system from one set of states to another set was first studied at the beginning of this century by Andrei Markov, a celebrated Russian mathematician who was a pupil of P. L. Chebyshev.

The example with student marks shows the model of independent outcomes to be insufficient for describing a sequence of states of a system. In most problems of physics and the natural sciences, the state of a system at a future time depends on which of a set of possible states the system is in at the present time.

This dependence may not be unambiguous: after a time the system may appear in different states, but the probabilities of future states ordinarily depend on preceding states.

If the probability of a transition of the system from one state to another depends solely on these states and does not depend on the prehistory of the system, then such transitions are termed a simple Markov chain. But if the probabilities depend on the preceding states of the system, then we have a complex Markov chain.

Information

The universal significance of feedback and the transfer of information in control processes led Norbert Wiener to a consideration of problems of control in engineering, living nature, and in society from a single point of view.

We have already spoken about feedback, now let us investigate the theory of information.

"The weather forecast for tomorrow is cloudy with intermittent rain and moderate winds...."

This weather report is transmitted by telegraph, radio, newspapers, telephone, by word of mouth and in many other ways. Quite irrespective of the method used, you will learn the weather forecast. The most important thing is the content, the physical carriers of which are quite distinct: electric current, electromagnetic waves, letters on paper, sound waves.

What do all these signals have in common? They contain the same information.

The telephone rings and you receive information about a phone call. You pick up the receiver, the bell stops ringing because the signal carrying information about the call has entered the telephone station and

has been transformed into a signal that switches off the calling device.

You touch a hot tea-kettle, cry out and pull your hand away. I have already explained the present-day view concerning the pain reflex. The important thing here to remind you is that information about the burning sensation of the skin is transmitted over the nerve network to the brain; here the information is transformed and, among other things, a new signal is generated that enters the nervous system and moves to the muscles of the hand. The final result is that you pull your hand away from the hot tea-kettle.

Fish transmit information (talk!) with the aid of ultrasonic vibrations in water. Bats utilize ultrasound for orientation in space.

Bees have a marvellous way of exchanging information. When a bee finds the "chosen land" where the whole swarm can drink the nectar of sweet flowers, it returns to the hive and puts on a dance: the figures of the dance in flight contain information about the direction of and distance to the newly found site. This has been verified by delicate and very ingenious experiments.

The life of every living organism is of necessity accompanied by an intensive exchange of information with the surrounding medium, and in the case of a more or less high organization it is accompanied by a mutual exchange of information between the organisms.

Take an automatic machine tool. Information is supplied beforehand in the form of a programme of operation. Besides that, the machine tool receives current information in the form of measurements of the workpiece; if the dimensions of the workpiece go beyond the tolerance levels, then information is transmitted on the necessity to readjust the tool.

When an aircraft is flown by radio (radio navigation), the information about the position of the plane and the weather conditions is transmitted to a control device that compares the information with the required course. As a result, information is generated on the necessity to change the positions of the rudders.

The control system of a factory or ministry uses information concerning the availability of materials and semifinished goods in warehouses, the state of operating machines and machines undergoing repairs, the work force, etc.; it starts with the indicated plan for output, which is also in the form of stored information.

Thus, everywhere we have information; in all systems of control there are communication channels for the transmission, reception and transformation of information.

Switching to a red light stops a train; pushing a button starts up a multi-ton press; a single phrase of the commanding officer is enough for a thousand-gun salvo.

The incoming message here can only have two values: red-green, on-off. This is elementary information. It is contained in the response to a question that requires only "yes" or "no" for an answer.

The recipient does not know the answer beforehand, otherwise he would not receive any information. From the point of view of the recipient, the answer to such a question is accidental (random), and one does not know beforehand which of two possible answers will be received.

The traffic lights at a city intersection have three colours: red, yellow and green. The message transmitted to the driver may be one of three possible ones: "Stop!", "Attention!", "Clear". In the transmission of letters via telegraph, the message assumes one of a series of possible values (letters of the alphabet). Incidentally, we can make use of different terminology.

We can say that as a result of an experiment (the reception of a letter) one of a series of outcomes of the experiment was realized (the letter 'z' was received).

Of course there are experiments that have so many possible outcomes that it is more convenient to regard them as experiments with infinite numbers of outcomes. For instance, in recording music the grooves on the phonograph record can have practically an infinitude of possible variants.

As in the case of two outcomes, so in the case of many outcomes, the recipient does not know beforehand which of the outcomes will be realized. The answer is therefore random—as far as the recipient is concerned. It is not known beforehand.

MEMORY AND CODES

Information may be accumulated and recorded. Actually, the entire process of education consists in the accumulation of information. Information is recorded or stored in books, articles, questionnaires, pictures, architectural structures, in musical script, phonograph records, magnetic tape. Electronic computers have special devices for storing information. Some of the information is kept in a long-term memory device (initial data, for example). But there is also a short-term memory, where the results of intermediate computations are stored until the given computation is completed.

The memory mechanism in the brains of animals and man is extremely complicated and, apparently, diversified. Its study has actually just begun. To some extent, certain of the devices of the human memory resemble the memory (storage) mechanisms of electronic computers, but there are also elements that differ substantially.

In all these so-called systems of memory the information is represented in quite distinct forms. Symbols on paper, grooves on a phonograph record, possible states (closed-open) of electronic relays, and states (excited-not excited) of nerve cells are all distinct forms in which information can be represented. What is more, one and the same bit of information may be represented in different ways: the number 5 may be represented as a digit, as a word (five), or five fingers may be bent down in turn.

The most important thing in such a representation is distinguishability. The information must be represented in such a manner that it is easy to distinguish one outcome from another. The messages must be uniquely distinguishable. For storing or transmitting digits from zero to nine, it is necessary to have ten distinct symbols, but it is immaterial whether they are Arabic or Roman numerals, written words, a sequence of electric pulses or some other kind of symbol.

The mode of representing possible outcomes of an experiment (or possible answers, or possible messages) in a distinct form is called a code, and the process of representing information in a distinct manner is called coding.

Thus, information may be recorded, transformed, represented in another form, which is to say, recoded. But for the present the most important thing for us is that

information may be transmitted

Rephrasing an aphorism of the famous physicist G. Thompson, we can say that information, like money, can be accumulated, but it is useful only when it is spent, that is, when it is transmitted.

The weather forecast for yesterday is useless, it is dangerous to be treated by an ignorant doctor or to

ride in a car driven by a drunken driver. Information must be timely and reliable.

When a speaker speaks very fast, the listener cannot make out what is said, he cannot reliably and unambiguously restore the information received. Given any mode of transmission and reception of information, reliability (confident distinguishability of the message being communicated) and speed of transmission of information are contradictory requirements.

Therefore, when designing a system for transmitting information the question arises of how to transmit the information so that both these basic requirements are satisfied in the best possible fashion.

The situation is complicated by the presence of interference in every communication channel.

When it is noisy in a room, it is difficult to hear the person you are speaking to, to say nothing of trying to listen to a lecture at the same time.

I have already mentioned the fact that in radio reception, telephone conversations, the transmission of telegrams and all other forms of electric or radio communication, there is always present a fluctuation noise that distorts reception. The level of fluctuation noise in a communication channel may be reduced but it is fundamentally impossible to eliminate it altogether.

Besides this there is a good deal of interference of a different character. For example, in radio reception we have the interference of neighbouring radio stations and atmospherics, interference due to passing trams, X-ray equipment, and many other sources of parasitic electromagnetic radiations.

When you talk over the telephone, you can hear a neighbouring telephone channel, a whistling or screeching (interference due to malfunctioning equipment).

In telegraph communications, interference can distort the letters and in place of a telegram reading "Bob tied

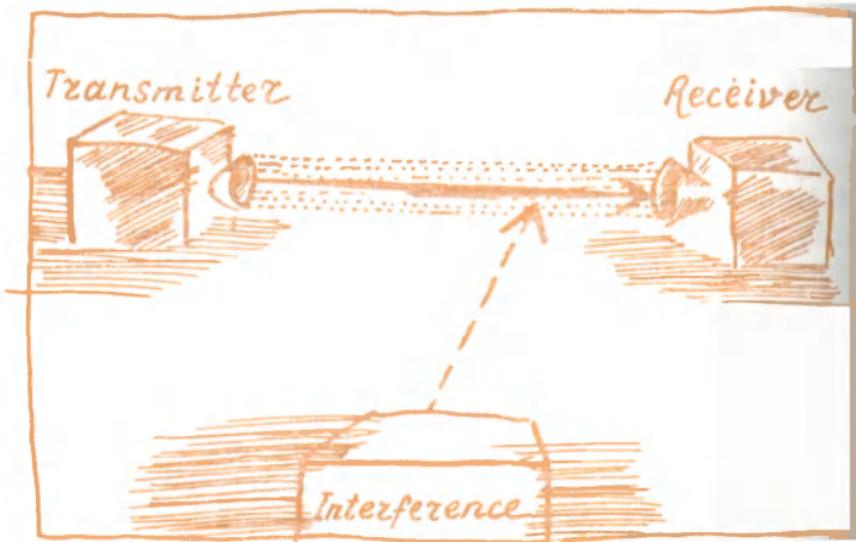


Fig. 94

400 relay event" the text might even read "Bob died 4:00 delay event". Quite disconcerting, to say the least. In printing, distortions give rise to misprints. To summarize, then, inherent in all systems of transmitting information is a diversity of interfering factors.

The various systems of transmitting information ultimately reduce to the simple diagram depicted in Fig. 94. Indeed, signals are transmitted via a communication channel from a transmitter to a receiver. They serve as carriers of the information.

How is it possible to ensure reliable error-free transmission of messages (information)?

Here is a likely argument that appears quite obvious and elementary. An erroneous deciphering of a message is due to a distortion of the signal by interference in the communication channel. Hence, we have to eliminate interference, suppress it at the point of origin. Or *find*

some mode of transmitting signals so that they are not distorted by interference.

And if this is not possible, then we have to increase the power of the signal. If one shouts loud enough in a noisy room filled with people, he will almost surely be heard.

From the invention of the electromagnetic telegraph in the middle of the 19th century up to the middle of the 20th century, this argument served as the guiding star of engineers and designers of telephone and telegraph systems, radio and television.

However, despite the tremendous amount of effort and appreciable advances in this direction, there are still many interference factors, as you have most certainly experienced yourself.

Now our aim is to increase the range of communication links. Not only between continents or even between the earth and the moon, but to such cosmic distances as to Venus and Mars. What we lack is the power output: there is a limit to the signal power that we are capable of delivering. What is more, the cost is prohibitive.

The problems of radar require rapid high-precision detection of high-speed aircraft or missiles. In the control of complex systems, such as missiles or technological processes, we have to determine parameters with ever increasing exactness, transmit the values of the parameters to control devices without loss of accuracy, and again transmit high-precision control commands. Noise and interference constantly plague the requirements of precision.

In a word, then, communications are expensive not only in outer space but on our planet as well. Our task is to learn to transmit information fast, reliably and inexpensively.

But is it really so important to transmit undistorted signals? Dispatchers at railway stations often make

announcements that are so hard to figure out that we are not always sure it is a man or a woman speaking. Yet this does not stop us from learning the track our train is arriving on. The signal—human speech in this case—is highly distorted yet we obtain the needed information.

Hence, the problem is not to reproduce the signal completely without distortion, but to see that the recipient is able to reproduce correctly the information contained in the message. This simple yet very important idea was clearly grasped and put into practice only a relatively short time ago, less than twenty years ago.

Note again that we have to do with a plausible argument, which is apparently very convincing but turns out to be erroneous.

In 1948 Claude Shannon, a noted American mathematician and engineer, published two articles which he called "A Mathematical Theory of Communications".

We will go into Shannon's theory a bit later, but for the present I want to point out that it was precisely Shannon who clearly understood and stated the problem that had to be solved in studying the transmission of information over communication channels.

In the diagram shown in Fig. 94, engineers carried away by the technical problems forgot about the two persons—the sender and the recipient of the messages. Now if we bring these two into the picture, then the scheme looks like that depicted in Fig. 95.

Suppose I am the source of the message (sender) and I want to send a telegram, or phone a friend, or give instructions to my colleagues to stop talking and get down to work. To transmit a message I have to put it in coded form, prepare it for transmission. I code a telegram by writing down the words on paper, I can speak over the phone, and I can quiet down my colleagues by lifting my hand and motioning to them.

Source of message

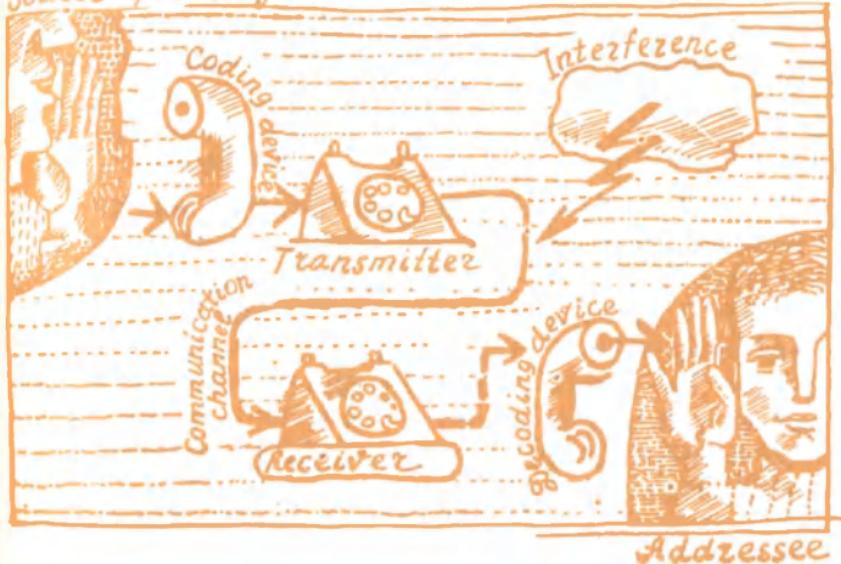


Fig. 95

The telegraph operator, in turn, will code my telegram by means of electric pulses, which will then be transmitted by cable through a communication channel. At the receiving end, the pulses will enter a decoding device and will be transformed back into letters. Then the printed telegram will be delivered to the addressee.

In the telephone conversation, the coding device is the microphone, in which sound waves are transformed into electrical oscillations, and the decoding device is the telephone of my friend, in which the electrical oscillations are transformed into sound vibrations of the membrane.

From these examples and the accompanying diagram it is evident that a designer is in a position to work on the transmitter, the receiver, and also to alter the coding and decoding devices, besides which he can choose at pleasure any method of coding and decoding.

Now comes the question of choosing a system of coding and decoding that will ensure error-proof transmission of information. And is that even possible?

INFORMATION AND WHAT IT'S ALL ABOUT

I have already used the word "information" a lot but I have not said anything about what it actually means. You of course know what it is, but I'm afraid you will find some difficulty in giving it a definition. This is because it is hard to define a broad concept. There is always the danger of floundering in generalities. But let's give it a try.

Ordinarily, when we speak about the information characteristics of a process or phenomenon, we have in view those properties which, in a certain sense, are contrasted with the energy or mass characteristics of the process. For instance, it is immaterial to us how we obtained a weather forecast—from the newspaper, by radio or in any other manner.

Information is neither mass nor energy. True, the transmission of information involves expenditure of energy, but this energy does not describe the information being transmitted either quantitatively or qualitatively: a single word of the commander-in-chief, and a war is started; a siren is sounded, and a factory or even a whole city comes to a standstill, yet the energy spent on the signal may be extremely minute.

However, information is just as objective a property of material processes as mass or energy. But whereas mass and energy are familiar notions, information is relatively new with a history of some 20 years or so. It takes time to get used to a new concept.

Mathematicians have not yet given us an exact definition of the term "information" that embraces the concept fully and can serve as a basis for constructing a

general theory of information, although fruitful attempts in this direction have been made.

Today when we speak about information theory, we mostly have in mind the ideas advanced about twenty years ago by Claude Shannon. However, the trend begun by Shannon does not refer to information in general but only to the problem of transmitting information over communication channels. It represents a unified scientific discipline. From the point of view of the mathematical methods employed, the theory of transmission of information belongs to the theory of probability. Today this is already a chapter in probability theory that is extremely fruitful and rich in content.

I shall first speak of problems relating to the transmission of information over communication channels and then touch briefly on certain other questions.

We can say "amazing information", or "valuable information", or "useless information". But all the modifiers here are emotional words which do not in the least characterize the information quantitatively. Now to construct a system of transmission of information it would be useful to have a quantitative characteristic of the information.

When we say that a certain message contains little or much information, we are comparing its quantitative characteristics, in the same way we compare the weight, length or cost of any item.

Let's try to think up a quantitative measure of information. We will take four different messages:

1. A tossed coin turns up heads.
2. The railway crossing is closed.
3. Today my wife gave birth to a baby girl.
4. The number of this tram ticket ends in the digit 7.

Can you tell which message contains the most information? Hardly. For me, the birth of a child is a big event, for you it probably makes no difference at all.

On the other hand, if you are a gambler and engrossed in a game of coin tossing, heads may call for rejoicing or the opposite—distinct information, and the quantity of information will depend on the betting. Now if you are in a hurry and the railway crossing is closed, such information will certainly mean a great disappointment.

Thus, the answer to my question depends on the point of view or on the circumstances. But how do matters stand if one takes the view of the designer of a system of communications, say a telegraph system? He is not interested in the subjective content of the information being transmitted. His aim is to set up a system capable of transmitting a message without errors, irrespective of what effect it will have on the recipient.

One of Claude Shannon's great achievements consists precisely in creating the concept of a quantity of information which is useful primarily in communications. The content of the information being transmitted over a communication channel is of no value per se and, hence, the concept of a quantity of information must rest on other characteristics of the message.

Now let us return to the question of which of the four messages had the most information. The first three messages can have only two versions: heads or tails, open or closed, boy or girl. To transmit such a message, we need two symbols, say, 1 and 0, plus and minus.

This is of course elementary information. It is contained in experiments involving only two outcomes.

(Note that an experiment involving a single outcome does not contain any information at all. Indeed, a grown-up person will not extract any new information from the statement 'if today is Wednesday, then tomorrow is Thursday'.)

A tram ticket can end in any one of the digits from 0 to 9, and so the last message has 10 possible versions.

To transmit any of the possible outcomes of an experiment involving a check of the last digit of a tram ticket we must have at least 10 distinct symbols. If we want to transmit a text in the Russian language, we need at least 32 distinct symbols.

To summarize, then, from the viewpoint of a designer of a system of communications, the information undergoing transmission is characterized primarily by the quantity of possible versions of the message or the quantity of possible outcomes (if one speaks in terms of transmitting messages on the outcome of an experiment).

Now recall the problem of coinciding birthdays. If you ask someone about the date of his birth and find that your birthday falls on the same day, then you will say the information is valuable, but if the birthdays do not coincide, you will most likely find the information of little value. (I'm almost sure you expected this answer.)

In the experiment in tossing a coin (a symmetrical coin), the quantitative estimate of information on the possibility of heads or tails turning up must be the same, for their chances are the same. We exclude any subjective assessment of outcomes (heads you win, tails you lose).

Now note the following: it is reasonable to express the information on the birth of a girl (not a boy) or heads (not tails) as quantitatively the same, because the indeterminacy of outcomes in these cases is the same. The probabilities of heads turning up and a girl being born coincide and are equal to $1/2$.*

* In different historical periods, the relationship between the number of girls and boys born varies somewhat, but it is always roughly the same. In an approximate estimate, we can assume the probability of birth of a girl and a boy to be the same, and, hence, equal to $1/2$.

In one toss of a coin, the indeterminacy of the outcome (that heads will turn up) is considerably greater than in a check of coinciding birthdays of two persons. The quantity of information obtained in the case of heads appearing is to be rated higher than when we get the news that two birthdays do not coincide, but is to be rated lower than the news that two birthdays do coincide.

Thus, the quantity of information describing a message to be transmitted is determined by the quantity of possible messages and their probabilities and does not depend on the semantic content (meaning) of the message.

That is the central idea in the theory of transmitting information advanced by Claude Shannon.

QUANTITATIVE MEASURES

The choice of a mode of transportation (automobile, train, airplane) for making a trip from one city to another does not depend on the exact distance between the cities (97 km or 5472 km) but on whether the distance is measured in tens, hundreds or thousands of kilometres. In a very large number of problems, the basic idea is not the exact value of the quantity involved but, as the mathematician would say, the order of magnitude of the quantity, that is, the number of digits in the numeral in the ordinary decimal system of numeration. Thus, the number 5472 lies between $1000 = 10^3$ and $10\ 000 = 10^4$ and it is convenient to say that 5472 is a number of the order of 10^4 . Incidentally, you will recall from your study of logarithms that in place of the inequality $10^3 < 5472 < 10^4$ we can write $3 < \log_{10} 5472 < 4$. It is thus very convenient to state orders of magnitude on the logarithmic scale.

In the theory of information one usually employs the binary system of numeration instead of the decimal system. In the binary system, numbers are written as sequences of zeros and ones. And, as we established at the end of the section "Multi-dimensional space", the number of all possible sets of n numbers made up of zeros and ones is equal to 2^n .

Hence, order of magnitude in the binary system is also determined by a logarithmic relationship, because n is equal to the logarithm of 2^n , however, not the common logarithm to the base 10 but the logarithm to the base 2.

It will be recalled that in our first conversation with the physiologist we spoke about the visual apparatus of man and animals as a highly sophisticated mechanism of nature. Here I want to remark on how this apparatus handles the extreme diversity of brightnesses that occur in life. The unit of luminance (brightness) in optics is called a stilb (sb). The following are some examples of the luminances of certain bodies:

the brightness of the sky on a moonless night is about 10^{-8} sb;

the brightness of the full moon seen through the atmosphere is 0.25 sb;

the brightness of a kerosene lamp is 1.5 sb;

the brightness of the metallic filament in an incandescent lamp is about 200 sb;

the brightness of the sun is about 1.5×10^5 sb.

It hurts the eyes if one looks directly at the sun, but objects not so bright can be examined and their brightness differentiated. Physiologists have established that the ratio of the least brightness to the greatest which the human eye is capable of distinguishing is of the order of 10^{12} . The visual apparatus is able to handle this enormous range of brightnesses because of its logarithmic scale. Numerous experiments have

demonstrated that the human eye does indeed react to the logarithm of brightness.

Thus, to estimate the quantity of information contained in the brightness of a luminescent body, it is natural to make use of a logarithmic measure.

Now note that if we want to transmit a three-digit number (in the decimal system of numeration) via a communication channel, it is by no means necessary to have a three-digit number of some kind of symbols. All we need to do is transmit three symbols, each one of which can assume one of ten possible values. This corresponds to the notation of this number by means of ordinary decimal digits.

The notation of that same number in the binary system involves no more than ten binary symbols because $2^{10} = 1024$. Therefore, if we use only two signals, say, current-no current, then no more than 10 such signals are required for the transmission of any three-digit number.

Hence, also in the transmission of information via a communication channel the logarithmic measure proves to be the most natural one.

It was Hartley, in 1928, who first proposed using the logarithmic measure for a quantitative estimate of the information conveyed through a communication channel. But Shannon went much farther. He utilized the probabilities of various measures.

If a message contains no indeterminacy, that is to say, if its content is known beforehand (a stone thrown up will fall to earth) and, hence, does not convey any information, then it is convenient in that case to regard the quantity of information as being equal to zero. The fewer the chances that the given message will be transmitted, that is, the smaller its probability, the higher one should express quantitatively the information obtained in the realization of that

outcome. Thus, the measure of quantity of information should be introduced in such a way that the quantity of information contained in a message increases as its probability decreases. It is natural to introduce the measure of the quantity of information so that in a message repeated twice (in an independent manner) the quantity of information is doubled, in a thrice repeated message, it is tripled, etc.

The income of a worker doing piecework is described by the mean daily wages and not the pay of any one day. In similar fashion, in the theory of information it is not the quantity of information acquired in the performance of an experiment that is the essential characteristic. And since the outcome of an experiment is accidental and subject to a definite distribution of probabilities, it follows that for the mean we must take our familiar mathematical expectation.

Suppose that the reserve of possible messages consists of only two messages with probabilities P_1 and P_2 (it will always be true that $P_1 + P_2 = 1$). Then, following Shannon, the mean quantity of information acquired in the transmission of such a message is equal to

$$I = -P_1 \log P_1 - P_2 \log P_2$$

(the minus signs are put there to indicate that the quantity of information is a positive number since the probabilities P are less than unity and, hence, their logarithms are negative). Given this definition of quantity of information, if an experiment is repeated independently twice, the acquired quantity of information is doubled, if it is repeated three times, the quantity of information is tripled, etc.

The foregoing arguments seem to be rather obvious, but the appearance of a logarithm (for no obvious reason) must seem unjustified.

It turns out that if we proceed from certain obvious properties with which we want to imbue the concept of quantity of information, then the sole measure having the needed properties is precisely the logarithmic measure. This is Shannon's theorem. I will not tire you with the proof: although it is rather elementary, it is quite long. Any reader interested in it can look up Shannon and read it through for himself.

Let us juggle our formula a bit. Since $P_2 = 1 - P_1$, we can rewrite the formula as

$$I = -P_1 \log P_1 - (1 - P_1) \log (1 - P_1)$$

Now I is a function of one variable—the probability P_1 —and it is easy to draw the graph of the function if, of course, we recall the concept of a logarithm or if we use tables. The graph of this function is shown in Fig. 96. It is seen immediately that the quantity of information I is equal to zero if and only if $P_1 = 0$ or $P_1 = 1$. But this means that either the first outcome is never attained and, hence, for any experiment the second occurs, or (when $P_1 = 1$) we always have the first outcome. In reality, this situation is equivalent to the case where there is only one message, which signifies the absence of indeterminacy in the experiment and, hence, the absence of information upon the receipt of such a message.

The quantity of information reaches a maximum when $P_1 = 1/2$, that is in the case of both messages being equally probable, as in the case of tossing a symmetric coin. Here the indeterminacy in the outcome of the experiment is greatest, and so we assess it as a maximum (in the mean). If we use binary logarithms, then $I = 1$ when $P_1 = 1/2$.

Let us examine experiments involving several outcomes. If the message consists of letters, an elementary

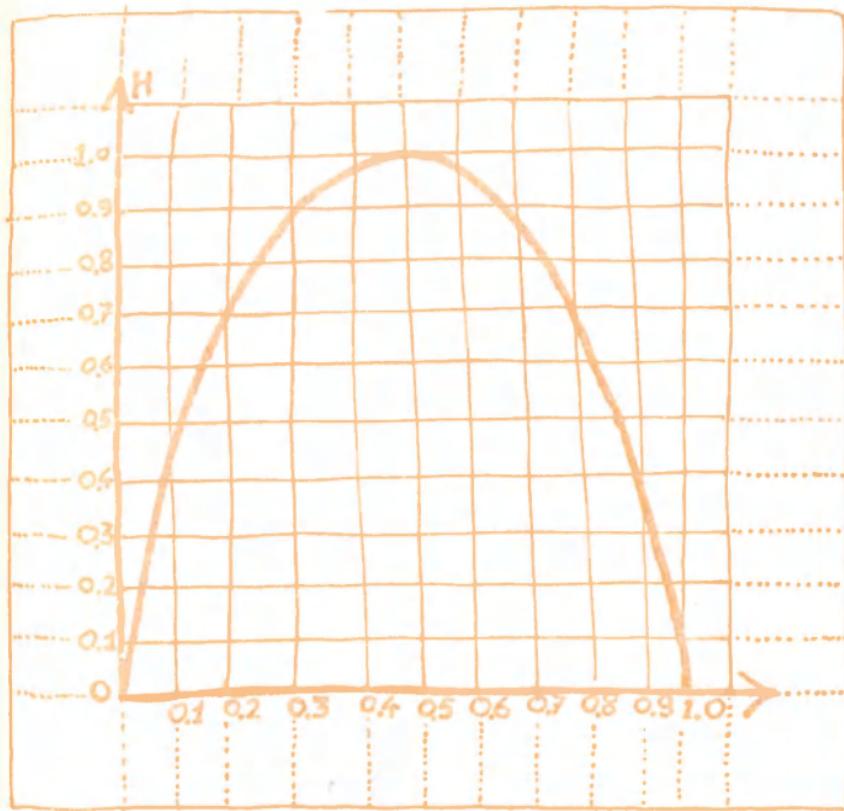


Fig. 96

outcome will consist in the appearance of a letter (there are 32 letters), then the average information acquired upon the receipt of one letter will be equal to

$$I = -P_1 \log P_1 - P_2 \log P_2 - \dots - P_{32} \log P_{32}$$

Here, P_1, P_2, \dots, P_{32} are the probabilities of occurrence of letters with the appropriate number labels.

On the average, we get the most information when all letters have the same probability of occurrence.

The picture is clear for the general case. We need only note that if the reserve of possible messages is equal to n , then the mean quantity of information reaches a maximum when the probabilities of transmitting all messages are the same, and, hence, equal to $1/n$. Then, as it is easy to compute,

$$I_{\max} = \log n$$

and, consequently, the average quantity of information increases very slowly with increase in the number of possible messages.

THE CAPACITY OF A COMMUNICATION CHANNEL

Let us take the railway line between Moscow and Leningrad. We can increase the freight in a train by increasing the number of cars and thus the carrying capacity of the train; but then the speed falls. It is natural to use the term 'carrying capacity' of a railway for the largest amount of freight (in tons) that can be carried in one hour on a given railway; that is to say, the amount of freight that can be delivered under the best distribution of freight among the locomotives and given the best freight-train timetable.

A similar situation is seen when we transmit information over a communication channel. For the sake of definiteness, let us consider a telegraph transmission with two elementary signals: current, no current.

These two elementary signals can be used to transmit letters, digits, and other necessary symbols. For instance, the Baudot code used in modern type-printing apparatus correlates each of the 32 letters of the

Russian alphabet with a definite combination of five current pulses or pauses of the same duration.*

Like loaded railway cars, these elementary signals are carriers of information. The quantity of information that can be loaded onto a single elementary signal becomes a maximum when the elementary signals are equally probable. Elementary signals have a definite duration and, hence, the quantity of information that can be transmitted over a channel in unit time is limited. It is natural to use the term 'carrying capacity' of a communication channel for the maximum amount of information that can be transmitted over the channel in unit time.

In the case of railway freight deliveries, some of the freight will be lost due to such random circumstances as accidents, natural disasters, negligence on the part of employees, etc. If these factors were common enough, it would be necessary to take them into account when distributing the freight among the trains and in making up timetables. Such is the situation during a war, when railway trains are bombed and lost.

However, even when taking account of accidental circumstances that lead to a loss of part of the freight, we can still speak of the carrying capacity of a railway, which is the maximum quantity of freight that can, on the average, be transported in one hour over the railway (with account taken of losses). It is of course less than the amount of freight that could be transported in the absence of any losses.

A similar situation occurs in a communication channel: the interference operating in the channel distorts the signals, and, as a result, part of the information being conveyed is lost.

* If each of the five positions can accommodate only one of two symbols (signal or no signal), then there will be a total of $2^5 = 32$ distinct signals.



However, here too the concept of capacity is retained. The capacity is the maximum quantity of information that can be conveyed, on the average, in unit time over the channel in the presence of interference.

The capacity of a channel in the presence of interference is determined solely by the number and duration of the elementary signals and the probabilities of their distortion by interference (that is to say, by the probability that one elementary signal was transmitted and in its place another was received) and does not depend on any other factors.

CODING

Recall some telephone calls when you couldn't hear well and had to yell "Hallo, I don't hear you! Please repeat that". Or maybe you used some stronger language. At any rate, even though the interference is considerable and hearing bad, if you repeat each word a sufficient number of times, your partner at the other end will finally make out what you want to say. But that will take up considerable time. The same

holds true for telegraph communication and other systems of transmitting information: multiple repetition of a message makes possible reliable reception of information, but it entails a sharp loss in the rate of transmission.

Now, is it possible to obtain reliable reception of messages without loss in the rate of transmission?

Let us take the familiar Morse code in which letters are represented by sequences of dots and dashes. Incidentally, it is not at all essential that dots and dashes be used; we could use zeros and ones, or current flow of different polarity. Here are some examples of the Morse code:

Letter (Russian) Code symbol	A	Б	В	Г	Д	Е	Ж
	01	1000	011	110	100	0	0001

Note that in the Morse code we have different numbers of symbols, and the more frequently occurring letters have shorter code symbols.

In the earlier mentioned Baudot code, all the letters are recorded with the same number of symbols, as witness

Letter (Russian) Code symbol	A	Б	В	Г	Д	Е	Ж
	10000	00110	01101	01010	11110	01000	00011

Such codes are called uniform in contrast to the nonuniform codes like the Morse code where elementary messages have unlike durations.

We see at once the advantage of a nonuniform code: one does not have to spend as much time in transmitting a frequently occurring letter (like 'E', say) as for the transmission of a rare letter like, say, 'Ж'. On the

other hand, uniform codes have a number of operational advantages. For example, when we use the Morse code we need an additional symbol to separate the letters, otherwise there would be complete chaos. When we use the Baudot code, it is clear that each succession of five symbols yields a letter, and so the messages are easily separated.

Incidentally, a special symbol for separation of letters does not always save the situation. A separating symbol, like any other symbol, may be distorted and the result again will be confusion. If, for example, in the word 'ДА' (100-01) the separating symbol is dropped, the resulting sequence, 10001, can be deciphered as 'НУ' (10-001) or 'НИТ' (10-00-1) or 'ТЖ' (1-0001), etc.

Still it is possible to construct a binary code without the aid of separating symbols. Here, we can take advantage of the theory of graphs. The idea consists in not using combinations, the initial parts of which have already been used as an independent combination.

For instance, we can use the combinations 10 and 001 but 10 and 100 cannot be used because if 10 has already been transmitted, we do not know whether the combination 10 is complete or whether the two elements are merely a part of the combination 100. A graph (tree) will help us to choose the needed combinations. At the two upper nodes (Fig. 97) we write 0 and 1 and then at each descending step we adjoin a 0 or 1 on the right. Thus, at the n th stage we will have written out all combinations of zeros and ones containing n symbols.

We will now construct a code without separating symbols by means of the following algorithm (rule). If we have already chosen some combination, say 010, then the portion of the tree topped by this vertex is no longer used. In the example shown in Fig.

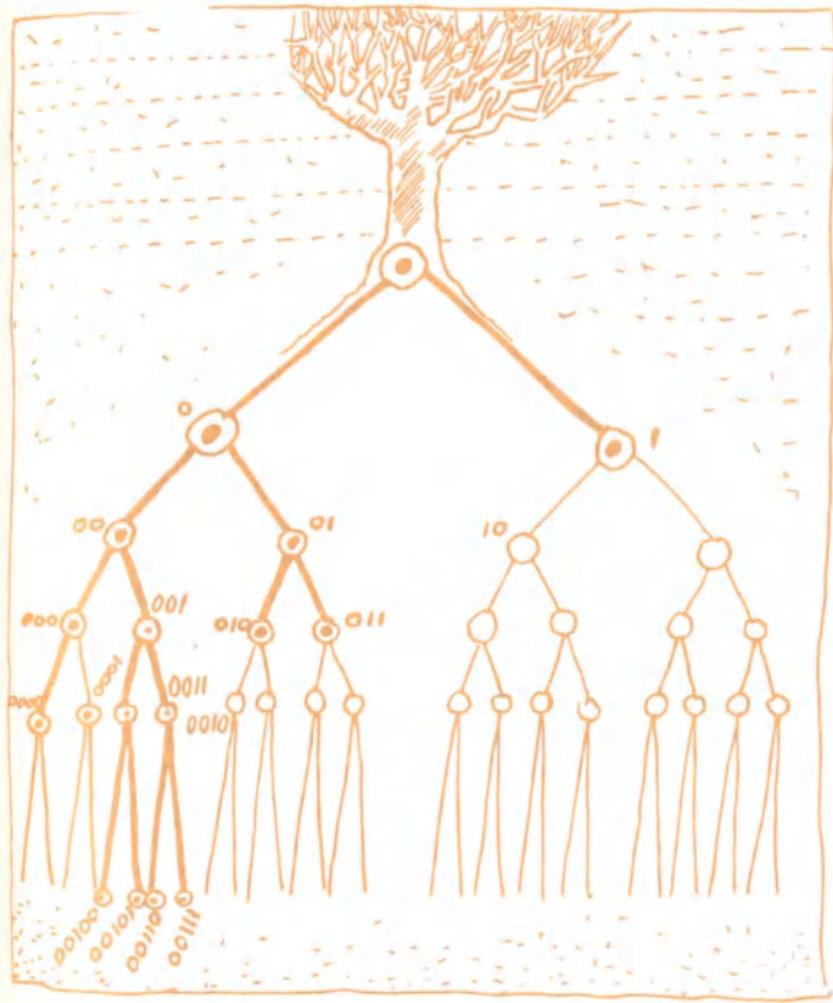


Fig. 97

the vertices that are used are denoted by circles and the edges to be used are heavy lines.

The choice of combination thus eliminates the possibility of using all subsequent combinations corresponding to ramifications of the graph.

In this way we can select any number of distinct code combinations.

When choosing the combinations 1, 010, 011, 0000, 0001, 00100, 00101, 00110, 00111 (see Fig. 97), any continuous transmitted sequence of these combinations is separated in a unique fashion. Say, the sequence 00100111001110100110001 is separated thus: 00100-1-1-00111-010-011-0001. And it cannot be separated in any other way, provided only those combinations that we have chosen are used.

The graph not only enables us to choose distinct combinations but also to decode the resulting sequence. Moving from the vertex to the right for a '1', and to the left, for a '0', we arrive at an open circle, which denotes one of the combinations. We then insert the separatrix, return to the vertex, and go down the graph again. It is of course easy to make this procedure automatic and hand it over to a computer.

To summarize, in codes like the Morse code or the Baudot code, each letter that is transmitted is coded separately. A proper choice of code can improve matters: we can reduce the number of improperly received letters or increase the rate of transmission for a given quality. It is not possible however to attain a radical improvement in this way. Let us therefore seek another approach to the problem of coding.

Happy New Year telegrams are sent at a reduced price rate. There are several standard forms to choose from. To transmit such a telegram it is not necessary to give the whole text—we can confine ourselves to the addressee and the number or code chosen by the sender. It is of course a big saving in time to send a code in place of the text of a telegram, and that is precisely what reduces the cost.

Now suppose we are sending routine telegrams ~~like~~, say, reports on bank operations. They too are standar~~d~~.

But the transmission of such special messages requires reliable reception. In coding the telegrams by means of appropriate numbers or codes, we will be able to utilize the extra time saved in order to enhance reliability of transmission. For instance, we can repeat the number of the telegram several times.

In studies of the structure of a language it has been noted that words are not made up of all possible combinations of letters but only of certain combinations.

The reader will recall the spelling games of his school days in which meaningful words are built up. I will change the conditions a bit. Let us take three letters (Russian letters) A, K, P and construct all possible words, including repetitions. It turns out that there will be $3^3 = 27$ such words. But out of these 27 "theoretical" words only 5 are meaningful ones. (The same can be done in English. Take, say, three letters A, C, P and we get at least one meaningful word, 'Cap'.)

A modern Russian dictionary contains 100,000 words. These are made up out of the 32 letters of the Russian alphabet. But if we took all possible seven-letter combinations of 32 letters, we would have 32^7 , which is over 30,000 million words. Yet the Russian language has quite a number of 10- 12- and 15-letter words too! In the formation of words, languages make use of only a small portion of all possible letter combinations. For every meaningful word there are millions of meaningless ones.

A LANGUAGE MODEL AND THE TRANSMISSION OF INFORMATION

Now let us examine a language model in the form of a Markov chain. In this model, the probabilities of succeeding letters depended on preceding letter com-

binations. Using transition probabilities, it is possible to compute the probabilities of various multi-letter messages. For instance, among all possible four-letter combinations (in English, say), the combination xyoz would have a zero probability of being transmitted via telegraph. Then there would be some like "ooze" with a very low probability, and others like "door" with a very high probability.

From now on I will divide all messages into high-probability ones and low-probability ones. Although such a division is clearly arbitrary it may be endowed with exact meaning. Here is the main point; not only is the probability of every low-probability message slight, but the sum of the probabilities of all low-probability messages taken together is extremely small.

When designing a system of transmitting information, it is reasonable to construct it so that the high-probability messages get reproduced reliably. As for low-probability messages, there is no need to take special measures for their errorless transmission, since there is every reason to believe they will hardly at all appear for transmission.

Claude Shannon considered the transmission, via a communication channel, of a sequence of messages that represent a Markov chain. This need not necessarily be human language. The sequence of commands for control of a machine, the development of a disease, and a sequence of chemical transformations can all be described with the aid of Markov chains.

You have already noted that it is possible to replace the coding of separate letters or symbols by the coding of whole words or, better, whole blocks of symbols. If the sequence at the input of the communication channel is a Markov chain, then it is possible, by taking sufficiently long blocks of symbols, to divide all the messages into a relatively small group of high-

probability messages, which are the ones to be transmitted, and not the very extensive group of low-probability blocks that does not require high-quality transmission. Their probability is so low that they need not be transmitted at all.

Now we have to choose a method of coding for the group of high-probability blocks. The longer the coded chain of symbols, the more inexpensive the coding system one can choose.

BASIC PRINCIPLE OF THE THEORY OF THE TRANSMISSION OF INFORMATION

If the quantity of freight (in tons) arriving on the average per hour at the Moscow Freight Station does not exceed the carrying capacity of the Moscow-Leningrad railway, then the freight can be delivered to Leningrad. The only requirement is that the train timetable be specially constructed for this purpose.

However, if the incoming freight exceeds the carrying capacity of the railway, not all of it can be delivered and some will have to be put into storage, and ultimately new measures will be required to transport all the freight to its destination.

A similar pattern occurs in the transmission of information over a communication channel. If the average quantity of information arriving at the input of a communication channel in unit time is less than the capacity of the channel, then it is possible to transmit all the information and properly decode it at the output of the channel. This can be done if a proper system of coding has been chosen.

But if the amount of information entering the channel in unit time exceeds the capacity of the channel, then it will be impossible to transmit all the information.



This at first appears to be obvious, but one must recall that there are interfering factors operating in communication channels that distort signals in a random fashion, whereas we are speaking about errorless decoding of the transmitted information. For this reason, if we take a closer look at our statement, it will appear to be faulty. Since there is interference in the communication channel that distorts the signals, and, hence, the information carried by the signals, it would seem to be impossible to receive information in the presence of interference.

Still and all, our first assertion is correct. Let us take a closer look at some aspects of this matter.

We have not forgotten that part of the signals may be distorted, but even in the presence of interference some portion of the information will be transmitted over the channel. The only thing to note is that the quantity of information being transmitted must not exceed the greatest possible quantity of information that can be transmitted over the given channel in the presence of interference; it is this quantity of information (conveyed in unit time) that was called the capacity of the communication channel.

To accomplish such transmission, we must of course choose a special method of coding. This method is chosen in such a manner that, despite the distortions of individual signals, the information carried by a

group of signals can still be unambiguously decoded. To achieve this, one has to code long blocks and not separate symbols or letters.

To be exact, we should also note the following: in order to code close to optimal, we would have to code all the long messages, which is, technically speaking, very difficult. Just imagine the complexity of, say, coding at once as one message (in a unified manner) all the telegrams that go in one day from Moscow to Leningrad. Therefore, our statement about the possibility of errorless transmission should be understood this way: the greater the reliability, that is, the smaller the number of errors with which one desires to transmit information, the more complex the method of coding.

To summarize, then, in place of nondistorted reproduction of signals that are distorted by interference in a communication channel, we have complex coding such that permits reproduction of the information with as small a number of errors as we please.

This marvellous idea and the theory justifying the possibility of such coding and also certain methods for constructing the necessary codes all belong to Claude Shannon.

But how, practically, does one go about building codes capable of realizing Shannon's ideas? There are many ways of constructing such codes. During the past decade considerable advances have been made in the theory of constructing interference-free codes.

The problem of constructing such codes represents an exciting mathematical problem. Many of the methods of constructing codes are quite elementary and of a recreative nature, but it would take up too much space to go into details here. As it is we have spent too much time on the theory of transmitting information.

WHAT ABOUT THE CONTENT?

If you are not a designer of communication systems but the father of a newly born baby girl or boy, then you will find information about the sex of your child far more interesting than the evenness or oddness of the number on a tram ticket. You are more interested in the content or meaning of a message than in its probability.

Quite right. But how is one to introduce a measure of the content-containing nature or value of importance of a message? How does one go about studying the semantic aspect of information? Is that possible at all?

A report on the discovery of a new antibiotic carries different semantic information to a child who has just learned to read, to a ninth-grade schoolboy, to an undergraduate microbiologist and to a specialist in antibiotics. What we have is this: as a rule the same information constitutes different values for different recipients.

In the preceding sections we discussed in detail the statistical theory of information. It was assumed that the recipient of the information is capable of extracting the entire information conveyed via a communication channel and it was precisely this maximum information that was estimated. In other words, it was a question of the potential possibility of extracting a certain quantity of information from a given message and not a question of what information a specific recipient of the message is capable of extracting.

At the same time, the ability to extract information from a message depends on the informational store (or store of facts) in the possession of the recipient. It is precisely for this reason that a report on the discovery of a new antibiotic conveys different infor-

mation to a child, a schoolboy, an undergraduate, and a specialist in antibiotics.

Let us picture the store of original information in the possession of a recipient in the form of a lexicon that not only enumerates all the words but also indicates relationships between them. For example, if the lexicon contains the words "student" and "book", then we must also include their relationship: 'the student is reading the book,' or 'the student has the book' (and of course not something like 'the book has the student').

The author of this approach to the study of semantic information is the Soviet mathematician Yu. Shreider. His term for the lexicon is "thesaurus", which comes from the Greek meaning "storehouse".

Due to the fact that the thesauruses of a child, a schoolboy, an undergraduate majoring in microbiology, and a specialist in the field of antibiotics are all different; a message (report) concerning the efficacy of streptomycin in treating pneumonia will convey distinct types of information. The child will be too young to comprehend anything (zero information), the schoolboy will obtain less information than the student majoring in pharmacology, and the specialist will, like the child, receive zero information since he already knows the facts. Hence, the amount of information received each time depends on the magnitude (or development) of the thesaurus of the recipient. If this information is depicted in the form of a graph, it will appear as the positive arch of a sinusoidal wave, where the maximum corresponds to the recipient with a thesaurus sufficiently developed to be able to comprehend the information but not developed to the point where the information does not involve anything new to him.

When a new message is received, the thesaurus of

the recipient is somehow changed, it is transformed. And the transformation is the greatest in the thesaurus of the most prepared (educated) recipient—however, the education must not be so great that the newly received information is so obvious that nothing new is apparent.

In this approach, for the measure of information acquired by a given recipient upon the receipt of a new message (or the measure of information contained in the given message relative to the given recipient) we can take the degree of change of the thesaurus under the influence of the message just received. Of course, what is needed is a quantitative measure of this change and this is what Shreider has done. True, it is only the beginning of a difficult field of investigation and of course I cannot explain the essence of the matter in a few words. At any rate, after nearly twenty-five triumphant years of the statistical theory of information and a critical revision of its initial premises there has now appeared an encouraging trend of studies that enables one to take into account the semantic content of a message. This augurs well for the fascinating field of information theory.

Mathematical Machines and What They Can Do

The cover of a magazine displays a pretty girl sitting in front of an enormous electronic computer console pushing buttons and keys. You learn that she can do the work of thousands of human calculators performing arithmetic operations at lightning speed and with unheard-of accuracy. For her there are no obstacles, no barriers. She can solve any problem.

The advertisement is a bad one, though. First of all, the girl has nothing to do with the whole business. It is not the operator nor the machine that does the work. Problems are solved by mathematicians. And not by the two or three mathematicians who bring the programme to the operator, although their work is often arduous. Behind the solution of every problem handled by a machine stands the genius of many generations of great mathematicians, including present-day scientists.

As for the electronic equipment itself, we can say that if the problem consisted merely in speeding up the work of human calculators, there would be no need to devise machines capable of performing tens of thou-

sands or millions of arithmetic operations per second, because such work could be done by much simpler and cheaper devices. What is really important is the possibility of solving qualitatively new problems. For even if all the three thousand million persons alive today were to try to compute the trajectory of a spaceship in flight, they would fail. Now an electronic computer can do that.

In a nutshell, then, we will be dealing with the potentialities of modern mathematical machines, computers. I left out the words "high-speed" and "electronic" on purpose. Right now, machines are being constructed that have no electronic components, and high speed is not their prime merit. For example, in certain systems for the control of oil processing and the chemical industry electronic devices cannot be used—an accidental short circuit or small spark might cause an explosion or a fire. What is more, the processes there develop relatively slowly and therefore in place of electronic devices we use *pneumatic* ones. A whole new field of science called pneumonics has come to life. In pneumatic computing devices, all the arithmetical and logical operations performed in electronic computers by means of transforming electric voltage and current are carried out by transforming currents of air under pressure that differs but slightly from normal ambient pressure.

Don't think for a minute that I want to convince you that high speed in mathematical machines is of no importance.

Problems of an economic character like, say, the compilation of a monthly or annual plan of work for a factory or the annual plan for the whole country require running through an enormous number of variants and choosing the best one. Of course, the best plan is one that optimizes a criterion, and we have

already dealt with that. However, there is a big gap between the general theory and practical applications of the theory. Today, huge teams of mathematicians and economists are engaged in introducing mathematical methods and computers into economics. The difficulties here are fantastic and they include purely computational ones. Recall the problem of distributing jobs described in the section entitled "Graphs". It would appear to be quite simple: just run through all possible variants. But if you have only 10 workers and 10 jobs, then the total number of possible distributions of jobs among the available workers will come to 3,628,800.

If the first person (out of 10) can take any one of the 10 jobs, then the second one has a choice of nine jobs, the third, one out of eight jobs, etc. The total number of all possible permutations will be equal to the product of these numbers:

$$10 \times 9 \times 8 \times \dots \times 2 \times 1 = 10! = 3,628,800$$

which is read: factorial 10.

This is a rather large number of variants if all you have at hand is a small desk calculator, but a modern electronic computer can handle a problem like that with ease.

However, if in the same problem we increase the number of workers to thirty, then similar reasoning yields the number $30!$. This is a colossal number despite the compact form of the notation. It exceeds 10^{33} , a one followed by thirty-three zeros, and is quite beyond our imagination. What is more, no computer can handle that number of variants. Indeed, even a computer capable of a million arithmetical operations per second would need over 10^{18} (a million million million) years to run through all the possible variants.

But then there are allocation problems involving 100 and more workers!

It is therefore impossible to resolve such problems by running through all variants. Before handing the problem over to a computer, an enormous amount of preliminary work often has to be done by highly qualified mathematicians.

Many economic problems are solved by methods of mathematical programming. For instance, the problems of linear programming reduce to the solution of algebraic equations and inequalities. You will recall from your school days that a system of two linear equations involving two unknowns—something like, say,

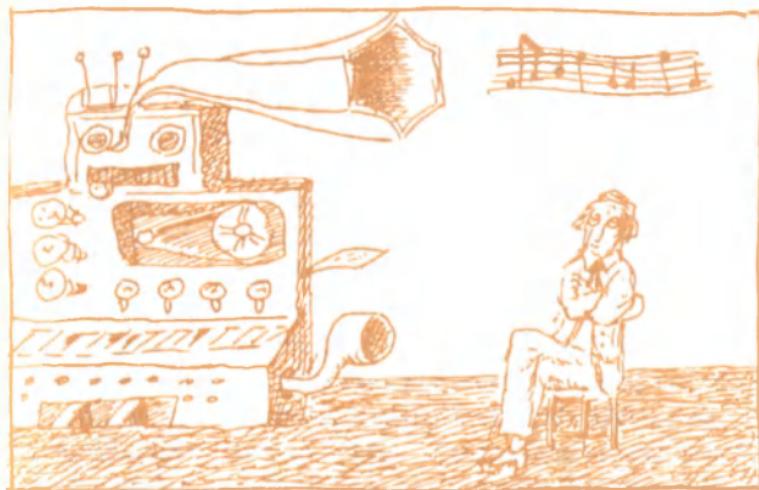
$$5x_1 + 4x_2 = 25$$

$$2x_1 - 6x_2 = -9$$

can be solved in two or three lines. All one has to do is multiply the first equation by 2, the second by 5 and then subtract the second from the first and, via addition and division, find that $x_2 = 2.5$, and then substitute this value into the second equation and by means of multiplication, subtraction and division find that $x_1 = 3$. All this requires only 15 arithmetic operations: 9 multiplications and divisions and 6 additions and subtractions.

But if we solve a similar system consisting of 800 equations in 800 unknowns, then we have to perform 250 million arithmetic operations. Yet there are many problems in economics and engineering that involve a still greater volume of computation. Here of course we need high-speed machines, devices for storing enormous quantities of information, and special techniques that speed up the process of finding a solution.

However, mathematical machines are capable of



doing quite different kinds of work. Apparently, it is already possible today to prove, by means of computers, many of the theorems of elementary geometry, even those not found in school textbooks, and perhaps even those which are still unknown. Appropriate programmes have already been compiled for that purpose.

It would seem that proving theorems is fundamentally different from performing the arithmetical and logical operations involved in solving equations. But, actually, computers perform only arithmetical and elementary logical operations, enumeration operations, comparison of numbers and the operations of choice (say, choosing the largest number in a group of numbers). Now all these are quite sufficient for the proving of theorems (for carrying out demonstrative reasoning).

A computer can even be instructed to compose music. R. Zaripov, mathematician and musician, got interested in modelling musical creativity on a universal mathematical machine. He analysed certain general laws of musical composition and then set up an appropriate programme for the computer which sol-

ved it ("wrote" the music). I personally heard some of that music. There are a number of quite acceptable pieces for the cello, some of which I really liked. True, I felt that I had heard that music before, but this often happens even at the concert of a living composer....

Zaripov's big achievement lies in the compilation of the programme for his musical opus to be executed on a universal mathematical machine. What the machine does, on the basis of Zaripov's programme, is to print out a sequence of digits that represent a definite code for ordinary musical notation. Now, performing a piece of music after recording (writing down the notes of music in the ordinary fashion) is a matter for the performer (musician).

Scientists are actively discussing the desirability of using mathematical machines in education. I have no doubts whatsoever about the efficacy of these machines in the learning process. Undoubtedly, the process of education will be greatly advanced by the introduction of learning machines. But education, unfortunately, is a field in which new methods penetrate with great difficulty.

Lectures are delivered just about like they were hundreds of years ago. I write formulas on a blackboard with chalk and then dictate definitions that the students could easily find in any textbook. I try to hold the attention of my audience by the same old tricks of a joke or two interspersed between the serious matter of equations, and so forth. I have to stop giggling girls with the same old stern schoolmaster glance, and I wake up dreaming boys who spent the previous night at a drinking party instead of reading up on the material of the previous lecture. And then at the exams I try to prove to the student that he deserves a failing mark. That takes up half an hour, sometimes

an hour, although a few minutes of questioning is enough to convince me. But the student should never think that anything is accidental at an examination, that he merely drew an unlucky set of questions. He must be made to feel and assess for himself the degree of the disaster that befell him.

Now a machine can handle the examination of a student and produce very decent results. There is no fear of the examiner, nor is there anyone to be accused of being unobjective, since the machine automatically puts down the mark depending on the number of correct and incorrect answers. Some advance has also been made in the use of mathematical machines in the actual learning process. It is to be hoped that this is only the beginning.

Mathematicians have compiled programmes for playing chess and there are already machines that play a fairly good game. Machines have also been programmed to play simpler games like dominoes and certain card games. The hard-core rationalist from among the readers of this book will of course ask who pays for all this recreation on the part of mathematicians. Let me remind the reader that the programming of a game like chess is not done for fun, it represents the modelling of intellectual activity of a human being.

At present, a good deal of attention is being paid to programmes for translating from one language to another. It is still cheaper of course to hire a translator, but the time will probably come when machine translation will be cheaper. But again it is not merely a matter of money: machine translation is also a model of the intellectual activity of man.

And what could be of greater interest than we human beings?

THE PSYCHIATRIST DROPS IN FOR A TALK

No other entity is more interesting than man himself and especially his intellect. Every psychiatrist comes into contact with a remarkable diversity of human characters with deviations from customary norms of thinking. And if he can tell his story, it makes the most exciting reading.

I've had my luck. A close friend of the family is a talented psychiatrist and a person after intellectual adventures. For many years she has kept us informed about the most exciting things in her field. Besides all this, she is a charming lady and a wonderful story teller. My luck again.

For her part, she has shown a keen interest in cybernetics, computers, biology, and medicine. I have had occasion to tell her about my conversations with physiologists and doctors, but up until recently she did not think these new approaches could be applied to her field. Now her views have changed....

Psychiatrist. I have some questions to ask you about my work.

Mathematician (me as usual). With pleasure. But how can I be of any help in your work?

Psych. I have to choose a narrow field for specialization. The time has come for a change. As you know, I have been studying a group of what are known as involution (presenile) psychoses. I have the catamneses* of some of my old patients and so I can follow their illnesses over a long period. All that material needs analysing and interpreting.

Math. What do you mean by analysing?

* The anamnesis is the personal history of the patient and of the onset of the disease, the catamnesis is the patient's history which follows the initial examination.

Psych. Well, drawing certain conclusions. It will be apparent today how correct or otherwise I diagnosed the cases 10 and 15 years ago.

Math. Suppose we work out the percentage of correctly diagnosed cases, what will we find? Only how qualified you were in those days. We might even pose the question of a cut in pay (15 years back) if you made a lot of mistakes, and a possible reconsideration of your salary if the percentage of errors turns out to be small. True, there is little hope of an increase in salary, I should say.

Psych. No, really, I'm serious.

Math. My question, then, is this: Can you be absolutely certain of that percentage? What I mean is how reliable is the diagnosis of the state of the patient at the present time.

Psych. In a certain English medical journal on psychiatry I came across an article that begins something like this: "A neurasthenic is a person building castles in the air. A schizophrenic is a person who lives in one of the castles. A psychiatrist is the landlord collecting the rent from these people."

So you see, the author distinguishes between a neurasthenic and a schizophrenic but he does not take into account the possibility of the individual first to build his castle in the air and then to set up residence there.

Seriously speaking, it is difficult to establish a diagnosis with any reliability.

Just recently I organized a consultation with my scientific adviser to examine a patient and we could not agree on whether his illness was schizophrenia or psychopathy.

Math. Are there many cases where the diagnosis is unambiguous?

Psych. Yes, quite a few. At any rate, qualified psy-

chiatrists of one and the same school can, as a rule, give the same diagnosis.

Math. What will happen if a different diagnosis is given? Will that improve the state of health of the patient?

Psych. Hardly. At first at least. But the type of treatment may be quite different.

Math. How many different diagnoses can be made in certain related situations that you have studied?

Psych. We are able to diagnose a rather broad range of illnesses. Within the range of senile psychoses that I have been studying, there are seven clear-cut clinical forms. Practically speaking, these are diagnosable illnesses. In a more profound analysis, with subdivisions added and allowance made for casuistry, if you will, there might be upwards of 20 distinct illnesses.

Math. Now about the types of treatment. Does each illness have its own particular treatment or are there fewer types of treatment than there are illnesses?

Psych. We do not yet have a specific treatment for each illness, so there are fewer types of treatment than there are illnesses.

Math. Why do you need more types of diagnosis than there are ways of treating the illnesses?

Psych. That's a tricky question. Maybe in the future we will have more types of treatment than at present. During the past 10 years or so a whole new branch of science has sprung up on the borderline between psychiatry and pharmacology; it's called psychopharmacology. The number of newly synthesized psycho-pharmacological drugs is constantly growing. We apply them separately and in combinations under clinical conditions and the effect is sometimes remarkable and sometimes insufficient. Clinicians try out new combinations. Gradually we will learn to make better and more precisely aimed drugs, and it is quite

possible that in the future there will be a strict correspondence between an exactly, and, what is more important, a timely formulated diagnosis and the treatment prescribed. And so today the problem of early diagnosis is definitely a topical problem. In the late stages of a disease, the diagnosis is quite evident, but treatment is then almost useless.

Math. All right. Now suppose you have been able to compare the diagnoses and the outcomes of diseases in patients you have treated. What will that give you?

Psych. That's just it. A large number of such descriptive papers have appeared in psychiatry. Perhaps the time has come to generalize in a more essential and objective way, to reliably elicit the characteristics of early stages of an illness that lead up to certain definite outcomes.

Math. But why is that needed if the patient covers the same ground irrespective of your intervention?

Psych. That's not exactly true. In some cases the outcome depends on the psychiatrist. First of all, in the case of a correct diagnosis the treatment will proceed differently. Secondly, a correct diagnosis is frequently of prime importance to a patient in forensic cases.

The point is that establishing whether a person is sane or not in a criminal case or establishing his competence or otherwise in a civil case depends a great deal on the diagnosis of the state of the patient. Hence, depending on the physician's conclusions, a person may be convicted of a crime if his sanity at the time of the commission of the crime has been established. Or, say, a person may be deprived of the right to rear his children or not be allowed to marry if his competence is called into question,

Math. Yes, those are important problems. How do you handle them at present?

Psych. In ordinary cases, an experienced psychiatrist has no trouble in making a diagnosis. But there are quite a large number of extraordinary cases too. For example, I was called in for consultation in the case of a patient in connection with repeated violations of the law and subsequent forensic examinations in different institutions. The diagnosis jumped from psychopathy to schizophrenia and back again, which signified first sanity and then insanity.

Math. Just a moment. Am I right in saying that schizophrenia is an illness whereas psychopathy is not?

Psych. In the case of schizophrenia, which is a rather common illness, the person's ability to think, feel and act is appreciably distorted. These three aspects lose the unity they ordinarily have when everything is normal (incidentally, this unity enables one to judge correctly the patient's motivations and thoughts via his gestures and acts). A certain dispareteness or splitting of the personality sets in (*schizo* means a splitting, and *phrenos* means personality).

Thus, a person cannot be responsible for his actions. Now in the case of upset psychic activity that goes by the name psychopathy the patient can control his actions. Here we have an instance of the differential distinguishing of complicated syndromes, and diagnostic divergences could easily be due to differences in distinguishing a number of the symptoms (signs) of the illness.

Math. What's a syndrome?

Psych. A syndrome is a group of clinical signs or symptoms which frequently occur together. One also speaks of a symptomatic complex.

Math. I see. How many symptoms go to make up a syndrome?

Psych. Syndromes differ. Some have three, five, others up to ten.

Math. And do you describe the state of a patient by a very definite syndrome? That is, does each symptom have a very definite significance?

Psych. Yes, it is more or less definite.

Math. I don't get that. Let's simplify the situation. We will assume each symptom to be binary, that is, capable of taking on only two values. Let us say the patient is either excited or not excited, jealous or not jealous, and so on.

Psych. All right, although that is definitely an oversimplification.

Math. You regard that as a simple situation? Let's count up what we have. If a syndrome consists of 10 symptoms, then there are a total of $2^{10} = 1024$ possible variants. And each of these variants has been described and signifies something.

Psych. Where did you get so many?

Math. And you thought I was oversimplifying. Here, take an example. Suppose you only have three binary characteristics: man-woman, excited-not excited, jealous-not jealous. We can then set up a diagram to illustrate all possible cases (Fig. 98). It will be seen that at each stage the number of variants is doubled. With three symptoms there will be $2^3 = 8$ variants, with 10 symptoms, $2^{10} = 1024$ variants.

Psych. Now I see. I didn't think there would be so many.

Math. How do you get out of a tangle like that?

Psych. It's hard to say since I rarely have to do any counting.

Math. Probably not all the variants are actually

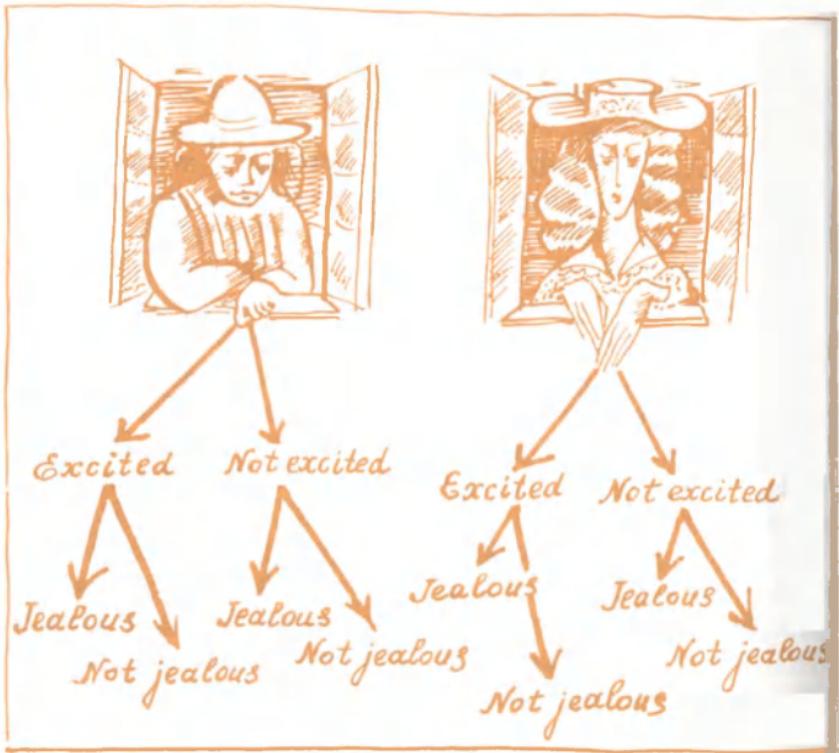


Fig. 98

encountered and you have to do with a much smaller number.

Psych. Yes, of course.

Math. What I would like to know is how you distinguish between essential symptoms and nonessential ones.

Psych. We do things somewhat differently. The case history of a patient in psychiatry is a whole booklet of 15 to 20 typewritten pages, and if it is well written up, the doctor can get a very good picture of the patient.

Math. What for?

Psych. People are different and psychiatric illnesses come in a great diversity of forms.

Math. Psychiatrists are different too. Suppose a series of outstanding writers—Dostoevsky, Tolstoi, Chekhov, Gorky, Tendryakov, and Nekrasov—went through a course in psychiatry and then set about writing up the case history of one and the same patient. I'm sure they'd all come out differently.

Psych. Yes, I guess so. That's precisely the difficulty.

Math. It seems to me that you yourselves create the difficulties. What actually happens is that you do not treat your patients individually but rather in standardized fashion. Then why the great diversity of descriptions? Psychiatry seems to be more of an art than a science.

Psych. Yes, psychiatry is indeed an art today. A mathematical analysis of the set of symptoms of a psychiatric disease might be compared to the range of colours used by a painter or the technique that a pianist acquires in performing scales. Therefore, we psychiatrists need your mathematical analysis, though we will not give up the art that we require in diagnosing a psychiatric disturbance. It is obvious to me that the question must be posed in a different way, but I don't know how.

Math. How many symptoms do you describe in an average case history?

Psych. A great many. It's hard to say just how many.

Math. Let's try and draw up a detailed list say a hundred items of all the main points. Some will be numbers, like age, blood pressure, and so forth. Then there are symptoms of a nondigital nature. But if we find any binary characteristics like jealous-not jealous, then we will write 1 and 0.

Now if a symptom is expressed in different ways, let's estimate the degree by means of a four-point

system. That should provide enough detail, since doctors differ mainly in estimates of degree. Then it'll be easier if we have a rough approximation.

Psych. Listen, that's a new angle—the rougher the approximation, the better. We always try to make th investigation as thorough as possible.

Math. And you get a wide spread of estimates.

Psych. Yes, that's true. All right, suppose I have this detailed questionnaire. Then what?

Math. Then you can take several hundred patients whom you are thoroughly acquainted with—case history and outcome. You fill out a detailed form on each case and then we'll try a cybernetic diagnostic procedure by means of a programme of pattern recognition.

Psych. What will that give us?

Math. First of all, we can automatize the process of making diagnoses, secondly, we can figure out the significance (information content) of the various symptoms and characteristics.

Psych. But that's a tremendous undertaking.

Math. Nothing here is very easy.

Psych. All right, then, let's try.

PATTERN RECOGNITION

Playing chess, composing music, solving equations and proving theorems can all be done by computers in accordance with definite rules that specify the sequence of logical or arithmetical operations. These rules, that is the programmes, are compiled by a human being.

Now can mathematical machines, like human beings and other living organisms, compile programmes of action themselves for achieving specific aims, or is it that without a detailed man-made programme they can do nothing?

Today this question is being vigorously debated. Biologists, physicians and specialists in the humanities are particularly active in defending the irreproducibility of the living entity when we speak of composing a programme of purposeful behaviour, in other words, the superiority of the living organism over the machine. In this context, the word "machine" is understood to mean something made by human beings using hammer, wrench and soldering iron.

Well, and what do mathematicians say? Here is the opinion of Polya, an outstanding mathematician and teacher whom we have already spoken about. In his book entitled *Patterns of Plausible Inference*, Polya writes: "From the outset it was clear that the two kinds of reasoning have different tasks. From the outset they appeared very different: demonstrative reasoning as definite, final, 'machinelike'; and plausible reasoning as vague, provisional, specifically 'human'. Now we may see the difference a little more distinctly. In opposition to demonstrative inference, plausible inference leaves indeterminate a highly relevant point: the 'strength' or the 'weight' of the conclusion. This weight may depend not only on clarified grounds such as those expressed in the premises, but also on unclarified unexpressed grounds somewhere on the background of the person who draws the conclusion. A person has a background, a machine has not. Indeed, you can build a machine to draw demonstrative conclusions for you, but I think you can never build a machine that will draw plausible inferences."

Thus, Polya does not believe that a machine can be entrusted with deriving plausible conclusions. That statement was made in 1954. Today—with a feeling of pride in man's prowess—we can say that Polya was mistaken. A machine can be taught to construct

plausible arguments and in this respect it has overtaken its human teacher, in a certain sense.

This is a complicated situation and I will begin from a distance. A baby learns to distinguish mother from father or grandmother. The words 'mama', 'papa', 'granny' are repeated and fingers are pointed and that's the way the baby learns. Now the mother all the time changes her appearance—different hair styles, different clothes, smiles or worry and so forth. But she always remains Mama. The same goes for the father. Gradually the baby learns to distinguish other men that are not his father. And so on and on and on. How is all this achieved? What is the process of learning and subsequent recognition of faces, cats, autos and so on? What is the mechanism like? We are not yet sure about anything in this sphere.

How does a person distinguish a portrait of a woman from that of a man, birch leaves from oak leaves? Particularly, the leaves—they are all different, actually only similar. Is it not possible to teach a mathematical machine to separate a variety of objects into classes of similar objects, just like we teach children to distinguish the letters of the alphabet written by different people—they are not exactly alike! Or take the case of making diagnoses when there are no two identical people or two identical diseases. In this process, no formalized criterion is given to the machine for classification of entities. We only supply it with several objects of the classes, say, a dozen oak leaves and a dozen birch leaves.

The same problem arises when designing a machine for reading handwritten or typed texts, when compiling programmes for a computer that classifies stages of schizophrenia or diagnoses cancer. Such automatic machines of course model the function of thinking.

The first automatic devices for recognition of visual

patterns were based on an analogy with the optic system of animals. The optic system—one of the most sophisticated and remarkable creations of nature—is an enormously intricate system. The human ocular fundus consists of roughly 130 million light-sensitive cells (rods and cones). Beyond the layer of these receptors (cells that receive stimuli) are several more layers of cells. They process incoming signals in a very complex manner and send them on to the brain. There the signals are processed a number of times. The way light signals are treated by the visual analyser is still not clear to scientists, and the models set up to help us understand how this highly intricate apparatus operates yield only a very rough picture.

One of the pioneers in the modelling of the functions of thought by means of automatic devices was the American engineer F. Rosenblatt. He gave the name perceptron (perception device) to automatic devices capable of modelling the functions of neurophysiological systems.

I will not dwell either on the theory of perceptrons or on the building of their models. The ideas behind various perceptrons are very interesting; however, so far practical steps to the solution of complicated problems encounter very considerable difficulties.

For this reason, many scientists engaged in the problem of modelling pattern recognition have set out in other directions, one of which I will now describe.

Fig. 99 depicts rectangles of two classes. Can a machine be taught to classify such figures?

Let us pose the problem more concretely. You are first shown only the eight rectangles depicted in the figure. Then a new rectangle is shown that does not coincide with any one of the earlier demonstrated 8 figures. Is it possible to construct an algorithm (rule)

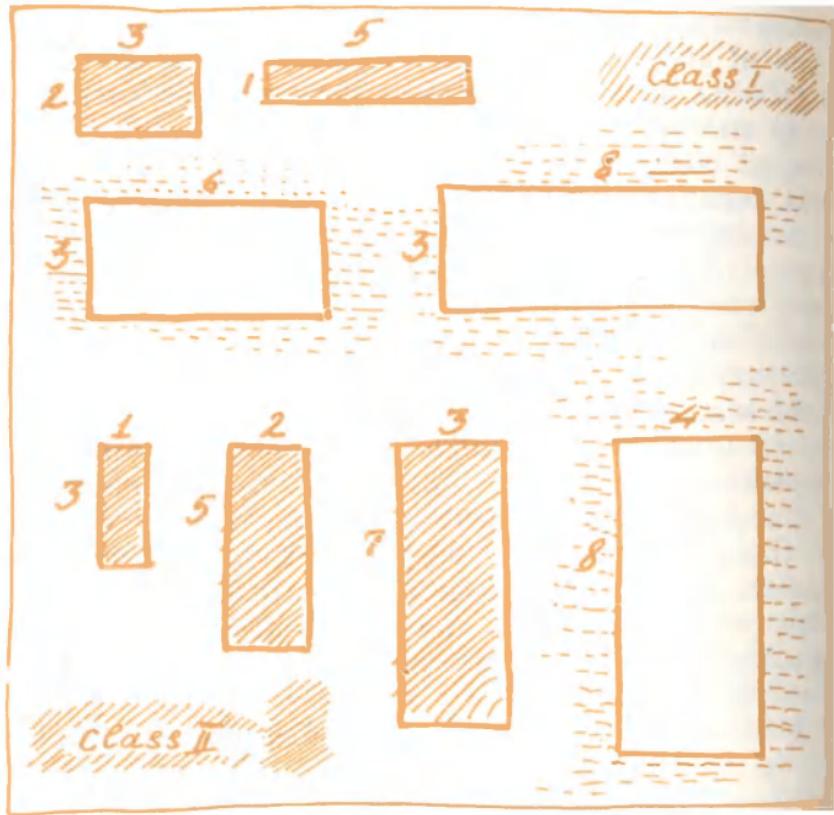


Fig. 99

for unambiguously placing the new rectangle in one of the two classes of figures?

You will say that it is very easy to formulate such a rule: horizontal and vertical rectangles. Perhaps you are right, but how does one explain to a machine what vertical and horizontal means?

That is not so difficult. We introduce the designations: x_1 for the width of a rectangle, and x_2 for the height. Then for rectangles of Class I and Class II we can form the following table:

x_1	x_2	x_1	x_2
3	2	1	3
5	1	2	5
(I) 6	3	(II) 3	7
8	3	4	8

Now, using a system of Cartesian coordinates (x_1 , x_2), we lay off the points corresponding to these number pairs (Fig. 100). The open circles correspond to Class I rectangles, the light crosses, to Class II rectangles. We can draw a line separating the two sets (circles and crosses). This may be an arbitrary line that



Fig. 100

separates the crosses from the circles. For example, in Fig. 100 we have drawn two possible lines — a light solid curve and a heavy straight line.

The rule for separating the classes is rather obvious: if a point (x_1, x_2) corresponding to a rectangle falls in Domain I we place it in Class I, if the point falls in Domain II, we put it in Class II.

For instance, if we make use of the algorithm specified by the light solid line, then the rectangles shown in Fig. 100 will be classified as follows. To Class I we refer rectangles $(1,1), (1,2), (2,2), (9,3)$ and to Class II, rectangles $(4,5), (5,5), (6,5)$.

Now if we take advantage of the rule specified by the heavy straight line, then all seven rectangles of Fig. 101 will refer to Class I. Perhaps you do not like the rules we have chosen. You are sure that the curve is wrong and that the line must be straight; what is more, a bisector of the right angle. For your pleasure I have drawn it as a dashed line.

Your assurance however would vanish if I had from the very start classified the rectangles of Fig. 99 as follows: hatched rectangles in Class I, unhatched rectangles in Class II. How would you draw the boundary line between the classes?

To summarize, then, let us note the important properties of such a classification algorithm. First of all, the rectangles were separated into two categories. Some of them were known at the start (they are depicted in Fig. 99). Then a rule was set up to separate them. After the rule (that is, the curve) was chosen, the instruction period was over. Then new rectangles were submitted. Using the given rule, we classified the new rectangles by referring them to different classes, depending on whether corresponding points in the plane appeared on one side or the other side of the curve.

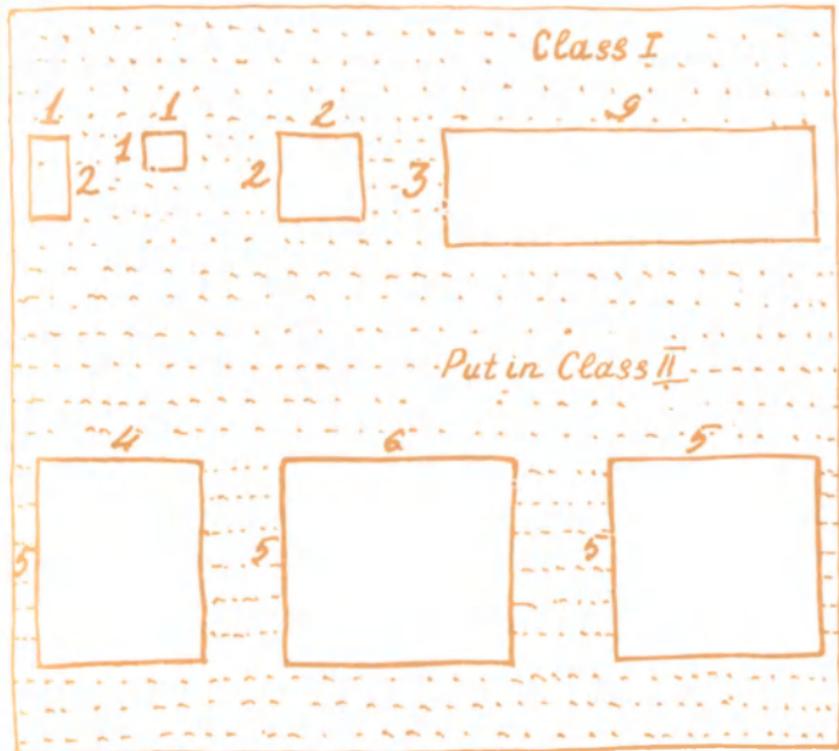


Fig. 101

It turned out to be something in the nature of an examination (of the curve of the chosen rule). Our assessment of the quality of the chosen rule depends on the result of the exam. Of course, we must know beforehand in which of the two classes each of the submitted rectangles belongs. The examiner always has to know the right answer. Then we can judge the quality of the chosen rule by the number of mistakes made at the examination.

Now let us discuss the results of the examination. To verify the quality of the two chosen rules of classi-

sification (the heavy straight line and the light solid curve), seven rectangles shown in Fig. 101 were submitted. From the start, it is given that rectangles (9,3) and (6,5) belong in Class I, (1,2) and (4,5) in Class II, and the squares (1,1), (2,2) and (5,5) may be put in either class.

When the classification was based on the "light solid curve", then rectangles (9,3), (1,1), (2,2), (1,2) were in Class I, while (4,5), (6,5), (5,5) were in Class II. This is illustrated in Fig. 101. Hence, classification via this rule led to two errors: (1,2) and (6,5) are not correctly classified.

When classifying with the aid of the "heavy straight line" rule, all seven rectangles submitted for the examination belong to Class I, which again led to two errors: (1,2) and (4,5) are incorrectly classified. Thus, if we judge by the results of this examination, then both rules are equally bad (or good), since they resulted in the same number of errors.

It is worth noting that to submit for examination squares when we are separating rectangles into vertical and horizontal types is just as unjustified as to submit one of the Beatles portraits when separating photos into men's and women's.

Thus, we see that firstly, the choice of rule for separation into classes depends on the material used for teaching, secondly, the separation is not always done unambiguously: there are different ways of establishing the separating boundary line.

Well, and isn't such a separation into classes a case of plausible reasoning? I will now examine a series of problems in classification—technical and medical diagnostics—the solutions of which by human beings represent typical examples of plausible inferences.

TECHNICAL DIAGNOSTICS

First a few words about a problem that confronts the geologist or geophysicist when drilling for oil. The deeper the oil lies, the fewer traces there are of it on the earth's surface (the day surface, as geologists say) and the harder it is to find. For this reason, geophysicists make extensive use of a variety of methods that enable them indirectly to detect properties of deep-lying rocks. They measure and study gravitational, electric and magnetic fields, nuclear and other radiations, elastic seismic oscillations obtained in special explosions. Geochemical methods permit detecting extremely small quantities of the mineral itself or of accompanying substances. These investigations are conducted in the air, on the ground and under the ground, in wells and mines.

Thus, the geologist and geophysicist have at their disposal a substantial amount of indirect evidence. But this information is hard to take advantage of, for not one of the prospecting methods yields an unambiguous answer about whether there is oil in the stratum or not.

This is similar to the situation of an examining magistrate who infers guilt from indirect evidence. Not one item of evidence taken separately constitutes full proof of the man's guilt, but all items taken together unambiguously incriminate him.

A geophysicist interpreting such material is frequently confronted by a very difficult problem—that of placing a given stratum in the category of oil-bearing or empty on the basis of measurements of a large number of distinct parameters and information concerning several qualitative characteristics.

Such a conclusion or, as we have phrased it, such decision-making leads to serious consequences. If the

decision is that the given stratum through which a well has been sunk is oil-bearing, then drilling is stopped, the well is cemented, shot, and the flow of liquid filling the porous rock of the stratum and flowing up the well is assessed. If the liquid is oil, that is good, but if it is water, then what? Then time, money and the labour of many people spent on the drilling are all lost. These losses are very considerable. It ordinarily takes about a year and close to a million rubles to drill to a depth of 4 to 5 kilometres. If the decision is that the stratum is filled with water, whereas actually it is oil-bearing and productive, then the losses are still greater, for millions of tons of valuable oil remain untouched deep inside the earth.

Some of the parameters that are measured in drilling are numbers, but most of them are curves describing the variations of a parameter along the well (for instance, the variation of electrical resistance of the rock).

Geophysicists have worked out a detailed series of methods for interpreting various geophysical parameters and also methods of a joint interpretation of two or even three parameters of a stratum. Although the methodology of joint estimates of two or three parameters enhances the reliability of the interpretation, it does not allow for reliable recommendations based on measurements or the avoidance of considerable errors. Even with regard to deposits that are easy to interpret we still have up to 5 and 6 per cent errors. But there are deposits of a difficult kind where the number of errors is very great. We will discuss those a bit later.

Yet a simultaneous account of the readings of all 10 to 15 available geophysical parameters is impossible. A problem of that magnitude considerably exceeds the potentialities of the human memory, the human possi-

bilities of analysis, synthesis, logical operations, arithmetical operations and enumeration of variants, which is to say, in short, the possibilities of processing the information.

So much for the difficulties and importance of oil prospecting. Let us now approach the problem from a different angle.

The human being is capable of doing a greater diversity of work than a hoisting crane, but the crane can lift tens of tons, while human weightlifters cannot even lift 300 kilograms. The same goes for technical diagnostics and interpreting the findings of geophysical measurements. A computer can do the job better, faster and more effectively than a man alone.

In the same manner that we described vertical and horizontal rectangles with the aid of number pairs (length and width), we will describe strata by means of sets of numbers (n -tuples).

Curves that describe the parameters of a stratum are likewise replaced by sets of numbers. These numbers are usually the mean values in specific intervals or certain characteristic values of the curves, say, the extremal values.

For instance, if in a stratum we have to measure 12 parameters, the 12 numbers x_1, x_2, \dots, x_{12} , then we will consider a 12-dimensional space where point P with coordinates $(x_1, x_2, \dots, x_{12})$ will correspond to a stratum with the given values of the parameters. I am sure the reader is no longer frightened by multi-dimensional spaces or by points with such a large number of coordinates. But don't try to imagine a space like that. All you have to do is imagine subsequent events in customary three-dimensional space and then calmly say that similar things occur in a 12-dimensional or a 100-dimensional space. We will call this space a parameter space. All possible sets

of measurements of the chosen 12 parameters that characterize oil-bearing strata of a given deposit will be represented by points in the parameter space. The set of all possible "oil-bearing" points in the space occupies a certain domain. We denote this domain by the word *Oil*. Similarly, points describing all possible porous strata not saturated with oil—empty strata—will occupy a certain domain in the parameter space. We denote it by *Emp* (empty).

Do you think the domains *Oil* and *Emp* can have some points in common? In other words, will these domains overlap?

The answer is ambiguous. Everything depends on how suitably the parameters have been chosen. Say, if we measured only three parameters—the thickness of the stratum, the magnitude of apparent electrical resistance measured by a 2.25-metre probe, and the relative amplitude of potentials of spontaneous polarization—then the domains *Oil* and *Emp* would have common points because for the same values of these three parameters a stratum may contain both oil and water.

If we measured only the depth of the stratum, its thickness and porosity (that is, the relative dimensions of the space between solid particles where a liquid might occur), then the domains *Oil* and *Emp* might be altogether indistinguishable.

Actually, however, oil-bearing strata differ substantially from empty ones: the former contain oil, the latter do not. And the basic hypothesis consists in the existence of a set of parameters that enables one unambiguously to distinguish empty strata from oil-bearing strata. Such a set of parameters may consist of a large number of elements (that is, elements in the set) and may be hard to measure, but it must definitely exist, since we know for sure that oil is not water.

If the parameters have been suitably chosen, then the domains *Oil* and *Emp* are situated in different parts of the space and can be separated by some kind of surface. The situation is illustrated in Fig. 102, and is quite similar in a 12-dimensional space (only in this case the separating surface is an 11-dimensional one).

Let us say we are lucky and the parameters have been chosen suitably and the domains *Oil* and *Emp*

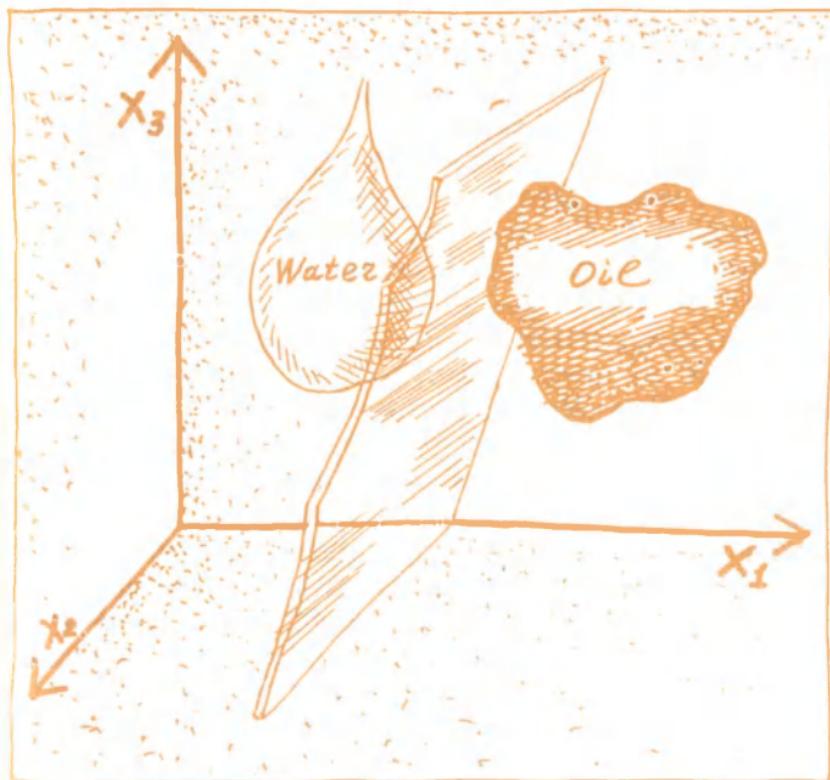


Fig. 102

are separable in the parameter space. If the two domains were known completely, it would be easy to find their separating surface. Actually, however, all we may know is certain sets of measurements that have been made in certain drilled wells, and also the results of tests made in the wells. In other words, speaking geometrically, all we know is certain groups of points in the domains *Oil* and *Emp*, and nothing else. Here is the task: using only these data, we must learn to classify any other strata that might appear later, that is to say, to place points corresponding to the strata either in the domain *Oil* or in the domain *Emp*.

It is now logical to proceed as we did with the rectangles. Let us take the available group of points in domain *Oil* and separate it into two parts. We do the same with points in domain *Emp*. Taking one subgroup from *Oil* and one from *Emp*, we use them to construct a separating surface—the decision rule. We will call these point sequences *learning sequences*. The remaining points will be used to verify the quality of the rule we have set up; that is to say, for the examination. We will refer to these points as the *examination material*.

The crux of the matter now is how to construct the decision rule (the separating surface). The method chosen must ensure not only the fundamental possibility of constructing the decision rule, but also that of constructing the rule in a sufficiently short time (machine time, that is). It must also ensure a subsequent classification of the material submitted to the examination, and this must be done with a small number of errors. These demands are contradictory: the simpler the type of separating surface, the easier it is to construct. At the same time, the simpler the separating surface, the more errors there may be. This is illustrated in Figs. 103 and 104 where any straight

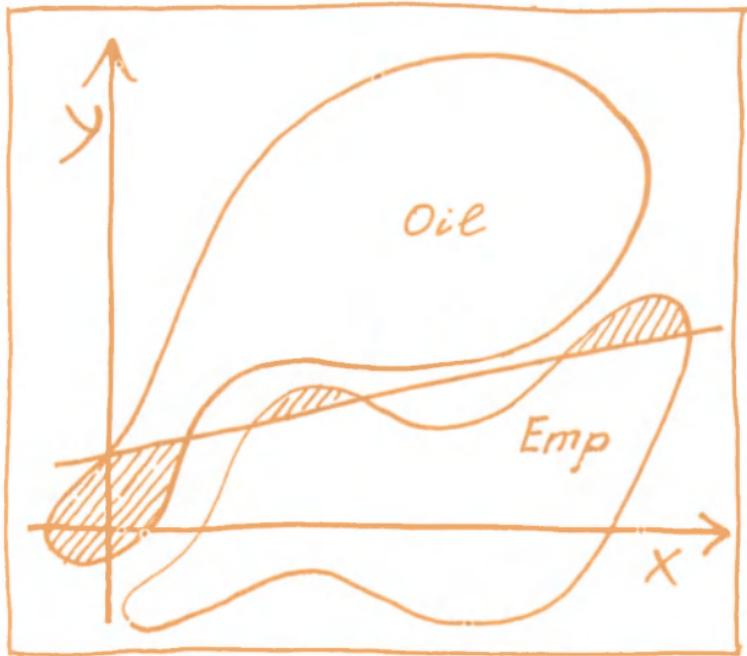


Fig. 103

line (the simpler rule) effects a worse separation than the curve (the graph of a third-degree polynomial).

I deceived the reader a bit when I said that in setting up the decision rule, we knew nothing other than the learning sequences. True, we do not have any points in the parameter space other than learning sequences, but there is one general fact without which all our conclusions would prove to be unpromising. That is statistical stability. The weather, the number of teachers needed in ten years, or the number of lung cancer cases can all be forecast only on the basis of previous experience on the assumption that the future will be "much like the past", that is, on the assumption that a definite probability distribution exists in the set of the phenomena under study.

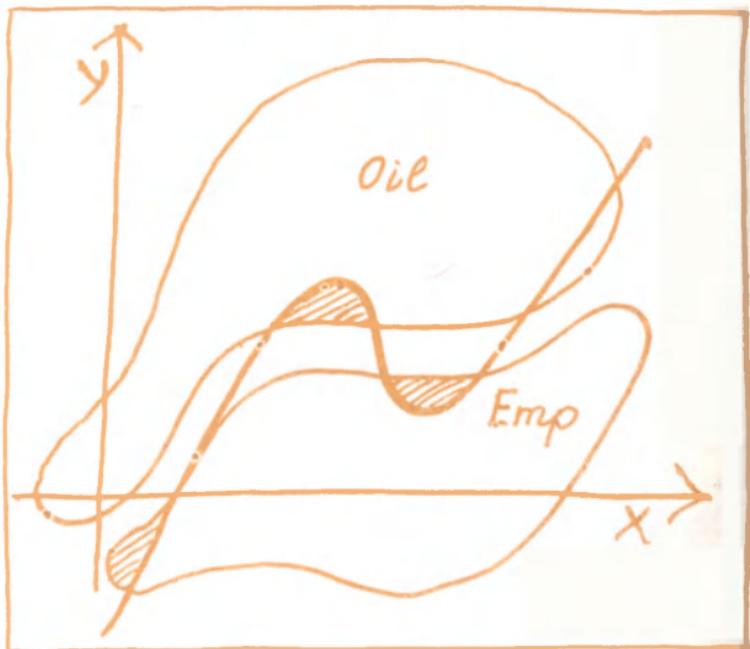


Fig. 104

For weather forecasting, this is a joint distribution of probabilities concerning temperature, pressure, humidity, and so on. For the number of teachers needed, it is the distribution of probabilities of children born, child mortality, and other factors that determine the number of children of a definite age in the coming ten years. To forecast the effectiveness of a chosen decision rule (forecasting the number of erroneous conclusions in a classification), it is not enough to know the number of errors at the examination. It is also necessary to be sure that the same trend will continue in the future. That is to say, we have to assume beforehand that the set of classified entities obeys a definite, though as yet unknown, distribution of probabilities. Only by proceeding on this assumption can we construct a statistical forecast.

We have already noted that not all random events possess statistical stability.

Whereas verifying a signature on a check against counterfeits is readily solved by a programme of pattern recognition (this is a clear-cut statistical problem), passing sentence on a criminal cannot be effected by such a programme.

I will not speak about the possible ways of constructing decision rules for pattern recognition, for it would take up too much space, although these methods can be discussed in a popular manner. All I want to note is that if the available parameters do not always make it possible to classify objects unambiguously and, hence, can lead to errors in any decision rule, it

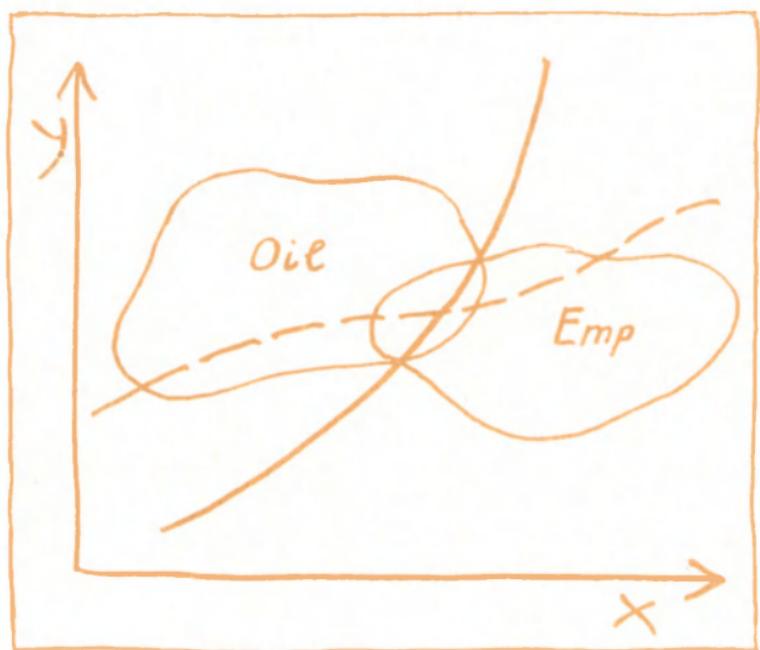


Fig. 105

is still possible to choose rules that will ensure the fewest errors.

This is illustrated in Fig. 105, where the domains *Oil* and *Emp* intersect (overlap). For the case where the appearance of points in the domains *Oil* and *Emp* obeys a uniform law of distribution, the classification rule specified by the dashed line yields, on the average, a perceptibly greater number of errors than the rule specified by the solid line.

It is hard to say whether the use of such recognition programmes is always justified. In some cases, it is like shooting sparrows with cannon, so to say, in others the difficulties of obtaining the necessary data are so great that they do not justify application of such methods.

However, in the case of integrated interpretations of geophysical measurements, the use of recognition programmes for isolating oil-bearing strata has been singularly successful. For example, regarding the Tartar deposits that are so amiable to "non-machine interpretation", interpreting geophysicists make up to 5 or 6 per cent errors, whereas interpretation by means of a pattern-recognition programme on the same material and with an M-20 computer has yielded only about one per cent errors. Using the materials of the Zhetybai oil deposit, ordinary geophysicists interpret with 35 per cent erroneous conclusions. Highly qualified geophysicists using the latest methods and only non-machine interpretation procedures yield 22 per cent erroneous conclusions. A machine interpretation of the same material by means of the pattern-recognition programme was in error by only 6 per cent. A spectacular result, as you can see.

It should be pointed out that in the interpretation of geophysical data via pattern-recognition programmes the computer is only used at the stage of choosing the

classification rule. After the rule has been chosen, interpretation reduces to the elementary arithmetical operations of multiplication and addition of numbers. The values of the parameters obtained in the stratum are arranged in a definite sequence, then each is multiplied by an appropriate coefficient, and the resulting numbers are added. If the sum turns out greater than a certain previously specified number (threshold, to use technical term), say, greater than unity, then the decision is that the stratum is an oil-bearing one. If the sum is less than that number, the decision is that the stratum is empty (non-oil-bearing). Such simple operations can easily be carried out with pencil and paper and by anyone with 7 or 8 grades of schooling.

Today these methods of machine interpretation are being widely used in oil production.

MEDICAL DIAGNOSTICS

New diagnostical methods for psychiatric and nervous diseases are only just being introduced and it is hard to predict their effectiveness in resolving such complicated diagnostic problems. What is more, just collecting reliable material about hundreds of patients concerning numerous aspects of each case is a tremendous undertaking. Several versions of questionnaires of the type discussed earlier by the mathematician and the psychiatrist have been worked out. One of them contains 130 symptoms. So far we do not know how many are essential symptoms in the diagnosis of schizophrenia and how many are not.

But before going on to other problems of medical diagnostics where the use of pattern-recognition programmes has led to some success, it is well to explain what kind of diagnostics we are talking about.

Diagnosing a disease is an extremely complicated thought process. The doctor starts with the complaints of the patient, asks questions and makes a study of the patient in order to get a picture of the possible range of illnesses. Then tests are made for further clarification and to be able to choose a method of treatment, which may undergo changes in the course of the illness.

When the patient complains of a pain in the arm, it may mean the nervous system is to blame, it may be a disturbance of the cardiovascular system, or it may merely be muscular, and so forth. We do not yet know how to apply mathematical methods to such general problems of diagnostics. At present, methods of pattern recognition only enable one to formalize and solve problems of differential diagnosis. Let's take a closer look at differential diagnosis.

In clinical practice, the great diversity of recognized illnesses is quite readily separated by the doctor into a number of distinct groups. Each of these groups consists of several illnesses having similar symptoms. Then comes the problem of defining one illness in a given group, this is a problem in differentiation. It is precisely differential diagnosis that is such a stumbling block even to highly qualified clinicians. That is why group consultation of several different specialists is so frequently needed. Incidentally, these extreme measures of collective discussion do not always save the patient. But the responsibility for an erroneous decision is less weighing on the members of such a meeting. This means something too, as you will recall from our discussion of criterion choosing and decision making.

A discussion of the problem of differential diagnosis in this book does not at all mean there is some profound relationship between differential calculus and differential diagnosis. Differentiation merely means

a splitting up of a whole into parts, a separation of a complex structure into simple elements. Now, whether the elements are the rather abstract intervals on a number axis or the quite concrete symptoms of a disease (tuberculosis, cancer, etc.) is a matter for the inventors of terminology who tack "differential" onto their specific terms. If you skipped the section on where terms come from, it is time to go back to it. But don't forget to return to differential diagnosis because the most interesting part is what follows.

Use of the methods of pattern recognition for the differential diagnosis of illnesses requires a preliminary isolation of characteristic symptoms or syndromes. That is the task of the doctor. Now, when the symptoms have been isolated and collected together as material for teaching and an examination (that is, patients of each of the illnesses being differentiated have been represented in sufficient numbers), then the machine works out the decision rule. Finally, when the rule has been chosen, we can use it in our practical work.

Numerous methods have been developed for pattern recognition in the diagnosis of illnesses. Their story goes far beyond the scope of a little book like this, but I would like to convince you of the effectiveness of these new methods. To illustrate, let me tell you about two pieces of research being done in Leningrad. One by a group of neurologists and a team of cyberneticians who have devised several programmes of pattern recognition. I cannot say that this work demonstrates the best methods or the most impressive results. Merely, I am acquainted with it better because I worked with the team.

The other was undertaken by a team of psychiatrists from the Institute of Psychiatry of the Academy of Medical Sciences together with my colleagues.

Disruptions in the blood circulation of the brain lead to cerebral hemorrhages or to encephalomalacia (softening of the brain). The causes of these two conditions differ. Softening of the material of the brain is often brought about by occlusion of vessels of the brain (thrombosis). To eliminate thrombosis, anticoagulants are injected into the blood stream of the patient. Anticoagulants are substances that prevent coagulation of the blood and clotting.

In the case of hemorrhages, we use substances with just the opposite properties. They are called coagulants and enhance coagulation of the blood and prevent bleeding from the blood vessel into the brain matter.

It is thus evident that errors in differential diagnosis of softening and hemorrhages can have fatal consequences for the patient. If in the case of softening an erroneous diagnosis of hemorrhage is made and the patient is treated with coagulants, then the processes of occlusion of the vessels and clotting will be strengthened. And if the blood ceases to flow to large sections of the brain, this will lead to severe damage and the possible death of the patient. On the other hand, if in a case of hemorrhage, the doctor mistakenly diagnoses softening, then the prescribed anticoagulants will further reduce the coagulation properties of the blood and hence increase the blood flow just when it should be halted.

Now the solution of this problem in differential diagnosis presents considerable difficulties even to experienced neurologists. The percentage of erroneous or indeterminate diagnoses is often rather high. The Leningrad cyberneticians and neurologists (A. Frantsuz, I. Tonkonogy and their colleagues) studied 278 cases of clinical anatomical observations involving softening of the brain and brain hemorrhages due to a paralytic stroke in cases of hypertension, atero-

sclerosis and rheumatic vasculitis. Here is a comparison of the diagnostic findings of a clinic with the results of a subsequent patho-anatomical study: the number of correct diagnoses, 75 per cent, indeterminate (when no decision could be reached) diagnoses, 13 per cent, and erroneous diagnoses, 12 per cent. For the patient an indeterminate diagnosis is just about the same as an erroneous diagnosis, because proper measures are not taken and the patient may die.

The low percentage of properly made diagnoses is remarkable in the case of such a common illness. What is more, the actual consequences of the illness are so distinct—occlusion of the blood flow in one case and just the opposite, hemorrhage, in the other. But outwardly the manifestations in patients are quite similar.

For example, loss of consciousness or nausea are considered to be signs of hemorrhage. But these same symptoms are sometimes observed when softening is the cause of disrupted blood circulation in the brain. The blood-red colour of the cerebrospinal fluid is considered characteristic of hemorrhages. In the case of softening of the brain, the cerebrospinal fluid is colourless. But cases in which there is no change in the colour of this fluid in hemorrhages are not rare.

Thus, each of these symptoms occurs separately in both illnesses. Apparently, the only reliable diagnostic procedure is a joint diagnosis covering all symptoms.

The Leningrad scientists applied the method of pattern recognition using 25 symptoms. Teaching was conducted on a sample of 100 cases out of the 278 available cases. At an examination covering the rest of the material, the machine yielded 88 per cent correct diagnoses. As you can see, the use of mathematical methods led to an appreciable increase in the reliability of

diagnosis: a boost from 75 per cent to 88 per cent. One would of course like to obtain 100 per cent correct responses. But these are only the first steps. Also it may turn out that the observable symptoms are not sufficient for an unambiguous diagnosis. This work will point up the necessity of seeking other determining symptoms and of resorting to supplementary methods of investigation. I would like to justify the machine-made errors by noting that in reality the doctor has access to a greater amount of information than those 25 symptoms that are fed to the computer. The doctor sees his patient and subconsciously takes note of many things. But it is hard to pass on these "many things" to the machine. Doctors, like all other people, find it hard to analyse the facts and motivations that govern their decisions.

I will now describe our work with a group of psychiatrists.

Further contacts led to a statement of the problem that differs somewhat from that discussed in the conversation with the psychiatrist.

Schizophrenia is a disease that often starts or manifests itself in early youth. It develops in a variety of ways. One of the forms of this disease, in which the symptoms are apparent throughout the patient's life, is called continual. But in the continual type that begins in early youth, the course of the disease may differ. Psychiatrists distinguish three forms: mild, medium, and severe.

We undertook the prognosis of the continual juvenile type of schizophrenia. The problem was to predict the development of the disease over 15 to 20 years on the basis of the data of the initial period of the disease. We will discuss that in more detail later on.

The study was conducted as follows. We had at our disposal an extensive volume of statistics: the case

histories of over 800 patients covering periods of 13 to 15 and more years and with just about the same age of onset of the illness. A careful clinical analysis was made of the initial stage of the disease (the first 3 to 5 years) of each patient and the most characteristic symptoms were isolated. Then a total of 130 binary symptoms were selected, and a card containing these 130 symptoms was filled out for each patient. If a given symptom was present, a "1" was entered in the card, if it was absent, a "0" was entered. These were the symptoms of the psychiatric disease in the most elementary form and as interpreted by a broad range of psychiatrists. This was done so that there would be no doubt on the part of the psychiatrist as to what must be entered in each column.

It is not easy to speak about psychiatric diseases and their symptoms. Sometimes it is even dangerous. The ordinary reader picks up the terms and without bothering to digest them begins to diagnose his own moodiness after a clash with a neighbour, or his dispair or elation as appropriate symptoms for some psychiatric disease.

When this diagnostic questionnaire appeared in our laboratory of young healthy scientific workers, they all took up self-diagnosis and filled in the questionnaire with zeros and ones (at their level of comprehension!). Quite naturally most of them found one or another form of schizophrenia.

But not every reader will regard his self-analysis in good humour. I will therefore forego any description of the symptoms of schizophrenia and will dwell only on the details needed for an understanding of the methods of work and their results.

In accordance with the classification that was developed, the patients were referred to one of the three forms of the disease based on their state of health at

the end of the observational period (that is, after 15 years had elapsed). Of course, this work took a long time. At the same time, a classification algorithm was being worked out to handle this great amount of starting material. The reader will recall that if all variants in the arrangement of zeros and ones in the questionnaire could occur, there would be 2^{130} in all, which is more than 10^{40} . This number of variants is not only far beyond the capabilities of any mathematical machine, but is quite beyond all comprehension. (To illustrate, picture every star in the portion of the universe accessible to our largest telescopes as having a planet and 3,000 million inhabitants—like the earth. Also imagine that they are all schizophrenics. Then their total number would fall short of 10^{40} .)

Building a classification programme and adjusting it on a big electronic computer was a tremendous undertaking too. But now we have the possibility of analysing experimental material containing several hundred binary symptoms which, apparently, is sufficient for the solution of practically any classification problem of this nature.

The classification rules that were worked out were based on the features in the diagnostic cards and on combinations of two and three features. Incidentally, the programme enables one to utilize even more complicated combinations of features. In constructing the rule, the teaching procedure was carried out on samples of 40 to 60 patients taken from each of the three classes. The remaining diagnostic cards were used to verify the rule (that is, for the examination).

In building up the diagnostic rule, we had the computer select the most informative symptoms (36 out of 130) to serve as the basis of the rule. All other symptoms were temporarily ignored. However, the classification based on only 36 symptoms yielded some good

results: from 92 to 94 per cent of the answers proved to be correct. It may be noted that most of the chosen symptoms and their combinations are in good agreement with the clinical picture.

It would be interesting to compare our results with prognoses done by psychiatrists if they were confronted with the problem of giving a prognosis of the state of the patients on the basis of the same questionnaire or solely of the symptoms chosen in the process of constructing the rule. Unfortunately, we did not do that because of the difficulties of organizing such an "examination". But our medical colleagues told us that predictions by specialists (doctors) would have yielded a substantially smaller per cent of correct prognoses.

Now a few words about the problem of predicting the course of a disease several years into the future. It is clearly very important to be able to estimate the severity of the future course of a disease on the basis of its initial manifestations. This is particularly true of such fields of medicine as psychiatry or oncology, where we do not as yet know the actual nature of the disease and we can judge the development of the process (especially schizophrenia) solely by the character and order of appearance of specific pathological symptoms. Appropriate treatment depends on whether the prognosis is correct or not. For certain categories of patients, rehabilitation in a social and occupational sense also depends on the prognosis. Besides, in the case of psychiatry, an important problem is forecasting the number of hospital beds needed in five, ten or fifteen years. This is essentially dependent on the prognosis of the diseases in the case of patients now under observation.

Let us summarize. The results of solving the classification problem—the problem of the prognosis of

the state of patients with continual juvenile schizophrenia-turned out to be good. However, I believe that these results do not represent the most important achievement. There is something more important. As we learned in our conversation with the psychiatrist, there is still a great deal of the subjective in the diagnosis of psychiatric diseases. Now the use of formalized rules of classification make the diagnostic process more objective, since such rules automatically rest on the assembled experience of many patients and can be refined as reliable material accumulates. All this enables us to raise the entire problem of differential diagnostics to a new and higher level. Of course, there is still a great deal of work confronting doctors and mathematicians, but it is already apparent that the results will be well worth the effort. Also, the use of such diagnostic algorithms may make examinations of patients faster and simpler due to better and more detailed formulations of questions in the questionnaire of the diagnostic card.

These same programmes will make it possible to correlate the traditional methods of examination with new methods (physiological, biochemical, electroencephalographic, and so on) and to determine the information content of the new methods. They will also permit verifying the effectiveness of methods of treatment, for that too involves classification problems.

REPLACING DOC WITH A DIAGNOSTIC MACHINE

If mathematical methods are capable of yielding a higher percentage of correct diagnoses, perhaps the time has come for mathematicians to handle the field of medical diagnostics.

Not by a long shot! One should not think that the

role of the physician reduces merely to diagnostics. Next come the most difficult problems—treating the patient and prophylaxis. Many other problems also confront the medical profession today.

But perhaps doctors should give up diagnostics and hand it over to mathematical machines. Machine diagnostics is of course quite impossible without the participation of the physician, for he is the only one who can select the necessary symptoms. And if he sometimes happens to err in assessing the amount of information carried by a certain symptom, this does not in the least mean that he can be dropped out of the game. Just the opposite, the physician must have at his disposal diagnostic machines for the purpose of simplifying his work and enhancing its effectiveness. But there is a danger here too which is precisely why I decided to write this rather unneeded section.

In actual fact, the doctor is a marvellous diagnostic machine. I hope doctors will not resent this comparison and will regard it as the highest degree of praise. One word of caution, though: the doctor must know how to observe his patient—not only look and listen, but also see and hear.

I had the privilege once of discussing some problems of medical cybernetics with Professor Votchal, a marvellous therapist, an erudite scholar and a brilliant man. Professor Votchal not only devises new instruments for his investigations but likes to fashion them with his own hands. He is very active in advancing the use of new devices in medical practice and is chairman of several authoritative commissions dealing with these questions. Therefore, his opinion concerning the role of electronics in medicine is of particular interest. Here is what Professor Votchal thinks about the matter. At the present time, he says, electronic instruments and computerized apparatus often

fail to help the physician and actually hamper his work because he places more confidence in the electronic device than in his own eyes and ears. And so instead of giving the patient a careful examination, the physician just takes a look at the electrocardiogram and views its sentence as final. These remarks refer not only to electrocardiography but in equal measure to any other method of examining a patient in which the role of the doctor is constantly being pushed into the background.

I am of course not calling on physicians to discontinue the latest methods of examining a patient and return to the days of "old doc and his black satchel". What I suggest is to use all weapons in diagnosing the patient. The eyes, hands and ears of the doctor are marvellous instruments created by nature, and so it is not a matter of machines taking the place of doctors but of doctors together with machines joining in a combined assault on disease.

"WHAT IS OUR LIFE? A GAME..."

During our lives we often encounter situations in which the participants have divergent interests and proceed by different pathways in the attainment of their aims. Such situations are often called conflicting situations, and the mathematical model of a conflicting situation is termed a *game*.

Let us recall another dramatic situation, this time from Pushkin's poem *Evgeny Onegin*—the duel between Lensky and Onegin.

The fighters cast their cloaks; the due
Paces, in number thirty-two,
Zaretsky, with due mensuration
Has taken. At the further ends

With pistols drawn he plants the friends.
"Approach"—and regularly, coldly,
Not aiming yet, the combatants,
Without a sound, but stepping boldly,
March on, four paces they advance,
Four fatal paces those! Not waiting,
And never his advance abaiting,
Evgeny is the first to lift
His pistol, quietly.—They shift
Two paces nearer; Lensky closes
An eye, the left—begins to aim
Also; Onegin at the same
Instant has fired. Thus fate disposes,
And strikes the hour. The poet lets
His pistol drop—his hand he sets
Hard to his bosom, never saying
One word, and falls—his clouded eye
No pang, but death itself portraying.

The combatants had different aims. It is natural to suppose that Lensky coolly regarded the possibility of dying but wanted to punish his wrong-doer. Onegin wanted to save his own life and was not at all interested in the death of his opponent. Each of them had one shot, and each could fire at the first step as they moved closer together, or then at the second step, or the third, and so on, right up to the barrier. Thus, each dueller could choose any one of the steps to fire from—one out of sixteen *strategies*. In the mathematical theory of games, strategy is the term used to designate the possible actions of each of the participants.

Now let us examine a rather dramatic problem but one that does not lead to such a tragic finale—the presentation and defense of a dissertation. Here we have an elementary mathematical model, a game between the applicant and the opponent. In an extre-

mely simplified version of this game, the applicant has two strategies: to write a good or a poor thesis (dissertation), and the opponent also has two strategies: to give a positive or a negative opinion.

The applicant will of course find it easier to write a poor dissertation but then the probability of a negative opinion is greater, and the aim of getting a degree will not be attained.

On the other hand, the opponent will find it easier to merely page through the dissertation and write a positive opinion. But if he writes a good opinion of a bad piece of work and that fact becomes apparent at the presentation of the thesis, then the opponent will sustain moral damage and his scientific prestige will be undermined. The applicant does not know what strategy the opponent will choose. He will have to examine his position in a most thorough fashion and choose a strategy.

Each of the players can choose one of a number of possible strategies. We will use the term *situation* to denote any set of possible strategies (one for each of the players). For instance, we have the following possible situation: the applicant chooses the strategy of writing a poor dissertation, the opponent chooses the strategy of writing a negative opinion.

It is natural to introduce a quantitative criterion, a measure of the preferability of each of the possible situations. If the choice of strategy is good, the player acquires something, if it is poor, the player loses out in some way. This criterion is termed *payoff*. A game does not of course always lead to winnings, and if a situation leads to a loss, then this corresponds to a negative payment. Incidentally, in some problems the losses are taken as the measure of preference in strategy, and then a positive situation leading to a gain corresponds to a negative loss.

Note that in our examples the gains of one player are not in the least equal to the losses of the other player, so that although their interests are different, they are not contrary, as is the case for instance, in games of chance.

The reader and the author of this book are also, by the way, playing a game. Here are my strategies: I write a good, mediocre, or bad book. The reader's strategies are: reading it from cover to cover, looking it through attentively, or merely paging it from time to time. This is what is known as a three by three (3×3) game because each of the contestants has three strategies.

If after reading the book from beginning to end you write a review full of praise, a cool review, or a killing review, and then either toss the book into the corner, or make a present of it to a friend, then I, the author, still have three strategies, but you already have six, and then we have a 3×6 game.

In our game, you pay money when you buy the book and spend time reading it, whereas I may be on the losing side even when it appears to me that I have written a good book. And so from now on we will speak both of winnings and losses. The participants in our game pay in quite different ways and so in one and the same game the losses of the different players can be measured in different units.

Every game between two persons with a finite number of strategies is conveniently represented in the form of an array called a *payoff matrix*, where the rows will correspond to the strategies of the first player, the columns, to the strategies of the second player. An example of such a matrix for a 2 by 2 game (degree applicant and opponent) is given below.

		Strategy of applicant	Good dissertation	Poor dissertation
Strategy of opponent				
Positive opinion		4	3	
	Negative opinion	-4	-2	

The numbers in the above matrix denote the payments of the applicant for each of the situations. The units are arbitrary because as yet we do not know how to express quantitatively the delight of the applicant if the presentation is a success or his dispair in case of failure.

The payoff matrix of the opponent may be quite different, as witness:

		Strategy of applicant	Good dissertation	Poor dissertation
Strategy of opponent				
Positive opinion		1	-5	
	Negative opinion	-4	-3	

It is now clear why the losses of the opponent are so great in the case of a positive opinion for a poor dissertation. But it is not very pleasant for the opponent to give a negative opinion even for a poor job,

all the more so for a good dissertation. And if the opponent is not able to drop out of the game in time and has to write a negative opinion, then he stands to suffer. In this case I assessed the payment as —3. The best situation for the opponent is a positive opinion given for a well-written dissertation and therefore the opponent here gets a positive payment; for him this is a gain, even though a small one. Take a close look at the payoff matrix and you will see why it is ordinarily so hard to find opponents willing to undertake the job.

In our example, the winnings (and losses) are evaluated in arbitrary units. However, in games of chance or in the analysis of many economic problems, the winnings are expressed in terms of money, in military affairs the losses are given as the losses of the sides, in engineering, the losses may for instance be time lost during repairs or idling of machines. We thus see that winnings may be expressed in a great diversity of units.

There can of course be any number of participants in a game. Say, in our game (author and reader), there may be a large number indeed—all the readers of the book. What is more, different readers have different interests, backgrounds and aims. Some read to extract new information, others read for recreation, and still others for.... There are any number of reasons for reading a book. I myself cannot often say why I pick up a book on genetics or architecture.

It is not always necessary to regard each individual person as a participant. In football, it is natural to consider two contestants—the two teams. In a war, the participants vary according to the problem at hand: several countries, groups of countries, but we might also consider separate army units as well.

The aim of the game for each of the participants is

to choose a strategy that maximizes his winnings. This would be simple if the player knew what strategies the other contestants have chosen. In that case he would review all situations involving the chosen strategies of the other participants and would then choose a strategy to maximize his winnings. But any one player does not know the strategies chosen by his adversaries. That is precisely the difficulty, the interest and sometimes the gamble involved in any game.

The simplest of all games is one involving two persons with conflicting interests (it is called a two-person game): in each play of a game the losses of one player are equal (with sign reversed) to the winnings of the other. Such games are sometimes called *antagonistic games*.

In this case the sum of the winnings of the players (winnings and losses, the latter being negative winnings) is always equal to zero. These games are therefore called *zero-sum games*.

Note that the zero sum of the game is a very essential limitation. Even in such sharp confrontations as military conflicts, we do not find the losses of one side equal to the winnings of the other, all the more so that the losses of the sides can, as we already know, be expressed in different units.

It is clear that when we specify a zero-sum game there is no need to indicate the winnings of both players. For this reason, a game of this kind is specified by enumerating the strategies of the two players and by a single payoff matrix.

Now suppose a game has been specified, that is, we know the strategies $A_1, A_2, A_3, \dots, A_m$ of the first player and the strategies $B_1, B_2, B_3, \dots, B_n$ of the second player. Let a_{ij} signify the winnings of player A if the players have chosen strategies A_i and B_j , respectively. The matrix of the game then

looks like this:

	B_1	B_2	B_3	...	B_n
A_1	a_{11}	a_{12}	a_{13}	...	a_{1n}
A_2	a_{21}	a_{22}	a_{23}	...	a_{2n}
A_3	a_{31}	a_{32}	a_{33}	...	a_{3n}
.
A_m	a_{m1}	a_{m2}	a_{m3}	...	a_{mn}

How do we find the solution of the game? For each of the players, solving the game means to indicate a course of action such that the average winnings are maximized in a large number of games. We assume here that both players are equally "brilliant" or "stupid", that is, each is equally capable of reviewing all possible situations and of assessing the degree of a calamity.

Quite naturally, the rules of the game provide that the players are ignorant of each other's choice of strategy in a given play of the game.

If each of the contestants acts in accord with games theory, he must each time pursue a strategy that maximizes his winnings when his adversary's actions are least favourable.

It is possible to interpret a zero-sum game as a choice of point on a terrain, the strategy of player A being to choose the geographic latitude of the point and the strategy of player B being to choose the longitude. The value of a winning is the altitude of the chosen point above sea level. If the relief of this terrain appears to be a mountain chain in the latitudinal direction and there is a relatively low mountain pass, then the situation of equilibrium that interests us corresponds precisely to this saddle point, *minimax*. For this reason such a strategy is termed *minimax*.

If the best strategy of one of the players is a minimax strategy, that is, one for which in the game matrix first the maximum element (number) in each row is taken and then the minimum of all chosen numbers, then the optimal strategy of the other player is *maxmin*, that is, a strategy in which the minimal elements (numbers) in each row of the matrix are taken, and then the maximal ones of those chosen. It may be proved that in all cases the maxmin does not exceed the minimax. But if they coincide, then there is a saddle point in the game: the saddle point is simultaneously a maxmin for one player and a minimax for the other. In this case their common value is called the *value of the game*.

If there is a saddle point in a game and one of the players chooses a strategy corresponding to it, then the best strategy of the other player will also be a strategy corresponding to the saddle point, any other strategy will only increase his losses.

Optimal strategies corresponding to the saddle point are called *pure strategies*. If there is a saddle point in a game, then there is no need to hide one's designs from the opponent because the best that both players can do—provided of course that the opponent is sufficiently wise—is to choose pure strategies. If the game does not have a saddle point, then there are no pure strategies for each of the contestants. Such games have a more complicated solution, and here, besides reasoning, the adversaries press into service the concept of randomness. It turns out that in this case the optimal behaviour is a change of strategy from play to play, the change being accidental (random) but with definite probabilities of occurrence of different strategies. These probabilities can be computed if we know the matrix of the game. This is what is known as a *mixed strategy*.

Let us now try to fix all these notions in mind by reverting to our coin game. We will alter the conditions somewhat, however. The game will now consist in the following: you place the coin on the table and cover it with your hand. Your adversary makes a guess as to whether it is heads or tails. If he guesses right, you pay him one kopeck, if he fails, he pays you. Here the matrix of the game is very simple:

-1	+1
+1	-1

But in this case the minimax (that is, maximum along the rows and then minimum along the columns) is equal to +1, whereas the maxmin is -1. Thus there is no saddle point.

What tactics will you pursue?

The simplest way is to keep the coin one side up all the time (say, heads). But then your opponent will soon perceive the situation and will win all the time. You can, say, change from heads to tails alternately. But then again your adversary will see your play and start winning. If you make the alternation sufficiently sophisticated but on a regular basis, then an observant adversary will sooner or later realize the situation and ultimately ruin you.

Consequently, your opponent must be deprived of any opportunity in the course of the game to extract useful information about your future intentions. For this purpose, your decisions at every step must be random (accidental) and independent. What is more, you will have to place the coin heads up or tails up with equal probability. It is easy to verify that for your opponent too the optimal strategy in this game is such that he will call out heads and tails independently and with equal probability. Thus, in this game the best alternative for the adversaries is to

use the simple tossing of another coin to give them their decisions.

This would seem to be paradoxical. Instead of purposeful action we recommend pure accident with no participation of "human reason". However, a more careful examination shows us that this is not so paradoxical as it is unexpected. But surprises of this nature are common. Throughout one's life one discovers new reasonable things which are hard even to suspect.

Although our conclusion does not appear to be very encouraging, in reality the theory of games has already made substantial advances in analysing the behaviour of animals, humans and social groups, in choosing an optimal course of action in a situation with conflicting interests and in the absence of complete information, and in the solution of problems that arise in military, economic, legal and production situations and elsewhere too.

However, the achievements of games theory do not consist so much in the resolution of specific problems as in the fact that for people dealing with highly complicated problems it offers a certain orientation when they encounter intricate situations involving conflicting interests.

At the beginning of this book the conversations between the mathematician and the physiologist led us to conclude that a living organism has to reorganize itself (change its state) in order to be able to solve the diversity of problems encountered in the course of its life. Modelling the adaptation of a living organism to external conditions in the solution of specific problems, that is, the modelling of purposeful behaviour, is now proceeding along the lines indicated by games theory. What we have in mind is the games played by automata among themselves and between them and "nature", that is, the adaptation of

an automation to changes in the environment, such changes being independent of the automation.

Here we have the marvellous investigations of the late M. L. Tsetlin, a talented Soviet scientist, into the behaviour of automata in "random media", chess games played by computers against humans and even between computers, the model-building of economic situations, and many other things. This is an ever expanding range of problems of extreme interest, but there is no time or space here to pursue the subject further.

Since one of my aims was to attract your attention to spheres of mathematics that might be of direct use, my advice is to keep in mind the mathematical theory of games. It might come in handy.

ONE FINAL WORD TO THE READER

I imagine you have spent a good deal of time reading up to this point. It is now my duty to explain why the book was written.

This book is not a text for self-instruction in modern fields of mathematics, neither is it intended for beginners, and of course it is not a textbook. It was written for those who are separated from mathematics by a wall of formulas, equations, proofs and graphs. It is indeed a hard job to break through such a wall. One never learns to play the violin by going to concerts and watching others play. In the very same way, to master mathematical methods of reasoning, to get acquainted with the numerous divisions of mathematics, and to learn to apply mathematics requires hard work and a lot of it.

My aim was to help the reader to see that beyond the wall of equations and symbols lays a land of exciting, understandable, and useful things. Of course, I was only able to make a few small holes in the wall.

The reader has merely glimpsed a few fragments of the overall picture. And perhaps these are not the most impressive fragments either. But can one enjoy a play seen from backstage?

I do not know whether I was able to demonstrate to the reader the greatness and significance of mathematics and to remove the oreole of mystery and inaccessibility of this science. If it is not too much trouble, perhaps the reader will find the time and energy to write to the publishers as to whether the author succeeded or not.

Let us conclude with some interesting definitions of mathematics given by outstanding scholars.

Friedrich Engels: "Mathematics is a science whose subject matter is spatial forms and quantitative relationships of the real world."

David Hilbert: "Mathematics is what competent people understand the word to mean."

In 1966, the 15th International Congress of mathematicians was held in Moscow. For the first time, a new section of the Congress was devoted to mathematical problems of control systems. Andrei Kolmogorov, the recognized Soviet, perhaps worldwide, leader of probability theory, opened the meeting of this section with the following words: "Mathematics is what people use to control nature and themselves."

The reader of course has to control himself, would like to control nature and perhaps even other people too. And so apparently your optimal strategy will be to master mathematics or establish contacts with mathematicians and work together with them.

Dr. Yakov Khurgin is professor of mathematics at the Chair of Applied Mathematics at the Gubkin Institute of the Petroleum and Gas Industry. He has written over a hundred scientific papers in pure and applied mathematics and has been particularly productive in the fields of radio-engineering, radiophysics, cybernetics, neurophysiology and psychiatry. At the present time, Professor Khurgin heads the laboratory of applied mathematics. He is also a member of the USSR National Committee of Automatic Control.

His extensive knowledge and wide range of activities have helped to make his popular-science book a great success.