JPRS: 37,803

TT: 66-34231

23 September 1966

CYBERNETIC PREDICTING DEVICES

by A. G. Ivakhnenko, et al

- USSR -

U. S. DEPARTMENT OF COMMERCE

CLEARINGHOUSE FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION

JOINT PUBLICATIONS RESEARCH SERVICE
Building Tempo E
Adams Drive, 4th and 6th Streets, S.W.
Washington, D.C. 20443

Price: $6.00

I-N

FOREWORD

This publication was prepared under contract for the Joint Publications Research Service as a translation or foreign-language research service to the various federal government departments.

The contents of this material in no way represent the policies, views or attitudes of the U. S. Government or of the parties to any distribution arrangement.

PROCUREMENT OF JPRS REPORTS

All JPRS reports may be ordered from the Clearinghouse for Federal Scientific and Technical Information. Reports published prior to 1 February 1963 can be provided, for the most part, only in photocopy (xerox). Those published after 1 February 1963 will be provided in printed form.

Details on special subscription arrangements for any JPRS report will be provided upon request.

All current JPRS reports are listed in the Monthly Catalog of U. S. Government Publications which is available on subscription at $4.50 per year ($6.00 foreign) from the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. Both prices include an annual index.

All current JPRS scientific and technical reports are cataloged and subject-indexed in Technical Translations. This publication is issued semimonthly by the Clearinghouse for Federal Scientific and Technical Information and is available on subscription ($12.00 per year domestic, $16.00 foreign) from the Superintendent of Documents. Semi-annual indexes to Technical Translations are available at additional cost.

JPRS: 37,803

CYBERNETIC PREDICTING DEVICES
- USSR -

[Following is a cover-to-cover translation of the Russian-language book Kiberneticheskiye Predskazyvayushchiye Ustroystva (English version above) by A. G. Ivakhnenko and V. G. Lapa, Kiev, 1965, pp 1-214.]

Table of Contents

Page

[Author's English Abstract]

CYBERNETIC FORECASTING FILTERS

Forecasting programs designed for large general-purpose computers constitute an important new tool in the control of production and economics. An example of such «big» forecasting programming is the work of Professor Richard Stone of Cambridge, who computorized the economics of the United Kingdom for 1970.

Nevertheless, small forecasting filters have their own domain of application. They can be realized not only as programs for general-purpose computers, but also as simple analog devices with high quick response. The first of such devices was constructed on the basis of the operator of Academician Kolmogoroff's formula by Professor Dennis Gabor at Imperial College (London) in 1955. Since then many other forecasting filters have been designed for different purposes and in accordance with different formulae (algorithms) — for instance, at Kiev Polytechnic Institute, where the authors work.

These different forecasting algorithms are considered, and many new recommendations are given in this book.

The authors discuss three principal methods of forecasting in addition to some others.

1. Forecasting of determined processes, i. e. extrapolation and interpolation.

2. Forecasting of stochastic processes, based on statistical forecasting theory.

3. Forecasting based on adaptation or learning of the forecasting filters.

Professor Gabor's filter was a self-learning one. It is shown in the book that the perceptron — the best known cognitive system — can also be used as a simple forecasting filter. Thus, there is no dividing line between cognitive systems and forecasting filters, for forecasting is the cognition of the future. The theory of cognitive systems can be applied to the designing of forecasting filters and, vice versa, the well developed theory of statistical forecasting can be used in cognitive system design.

The main problem is realization of optimum forecasting precision, the comparison of the precision and simplicity of various algorithms of forecasting. Sometimes, as in the case of control, quick response of the forecasting filters is also important. Some recommendation are given on the basis of a study of the precision of forecasting in the general form; some, on the basis of calculation of examples. All calculations were performed on digital computers.

The examples are taken from chemical industry, biology, ocean turbulence processes, forecasting of the relief of the Dnieper river bottom, and so forth.

The most important is the original proposal to combine the forecasting method developed for non-stationary processes (presented by Professor Farmer at the second IFAC Congress) with Kolmogoroff's basic method, developed for stationary processes only. The combined method of forecasting yielded good results in forecasting intracranial pressure in neurosurgery.

A special part of the book is devoted to the use of forecasting filters or cognitive systems in production control. Extremum control of the plant should be effected by a combination of open loop control and a corrector, smoothly correcting the characteristics of the open loop part. Cognitive systems and forecasting filters can be used as correctors.

Forecasting filters furnish the only possibility of constructing a control system for periodical processes, since prediction of the result of the process is essential for its control. This problem is also discussed.

The book discusses some problems in the theory of predicting determinate and random processes. Special attention is devoted to the realization of various operator algorithms for prediction on digital computers. Space is devoted to problems of using recognition systems, and in particular the Alpha system, as predicting filters.

The methods described are illustrated by examples from power engineering,hydrology, petroleum chemistry, medicine, and the control of industrial processes.

The book can be of use to specialists working in the various branches of science and engineering who are interested in the methods of statistical prediction and the concrete applications of these methods.

Introduction

The use of automatic systems has made it possible to solve many complex problems of control without direct participation by man. As the structure of the objects being controlled becomes more complex and the amount of information about the processes occurring in them becomes larger, man is often not able to perform the control function in the best way. This can be explained by the lack of time in which to choose the optimal solution, the impossibility of mobilizing a large memory volume in a short time, the property of information forgetting, and a number of other factors.

Complex automatic control systems are very fast-acting and have a large number of memory devices.

Furthermore, they must perform many functions of an "intellectual" nature, such as comparing different variants of the solution of a problem, choosing the best variant in accordance with definite criteria, taking into account change in external actions and the consequent change in the nature of the solution and the criteria.

Since the nature of the thinking capabilities modeled in automatic systems continuously grows more complex, in creating analogous systems it is necessary to take into account one of the important problems characteristic of human thought-- the capability of learning to predict.

There is not one action performed by man in which he does not foresee the results of this action in a sufficiently definite form.

When we formulate theproblem of prediction in engineering, it is obvious that we must investigate how the corresponding functions are performed in living organisms. Soviet physiologists "... have indicated not only forms of prediction, but also some concrete physiological processes which aid in this." points out Academician P.K.Anokhin. "But this whole immense problem connected with the mechanisms of foreseeing in the brain's operation which give power over the future is still far from worked out." This problem is important both for neurophysiology and for engineering.

Cybernetics has already made it possible to explain many prediction mechanisms. Cybernetic self-teaching predicting filters in the form of actual electronic circuits can serve as models for the predicting mechanisms of the brain.

The Basis of Prediction is the Experience of the Past

One of the basic hypotheses on the nature of prediction of the future consists of the fact that conclusions as to the possiblity or probability of a future event or series of events are made on the basis of study, analysis and generalization of preceding experience, the history of the phenomenon being predicted. This idea, in particular, forms the basis of the statistical theory of predictions being developed at present.

However, we may encounter facts concerning prediction of the future which do not seem at first glance to be at all connected with the past. It is known that experience consists of a very much larger number of pieces of information than man can consciously elucidate. Hence statements to the effect that certain cases of prediction cannot be explained by preceding experience, since precisely such and such an event or situation was not observed in the past, cannot be considered to be well founded.

It has been proved that much of what is remembered by man is independent of his consciousness and is contained in his latent memory.

A number of works on neurophysiology bear witness to the fact that the information registered (consciously or unconsciously) in the memory does not disappear. The Canadian scientist W.Penfield has shown, in particular, that, when definite conditions are created, for example when a weak current is passed through electrodes attached to the temples, sensations relating to the past arise in the patient. Events experienced long ago and often forgotten are remembered. Well known is the phenomenon of hypertrophic sharpening of the memory, or hypermnesia, which arises as a result of some brain diseases. The person remembers completely forgotten facts which occurred in the past and can cite from memory whole pages of books read earlier.

The volume of information on the past, the size of experience on the past under different conditions cannot be the same. Proceeding from this fact, we may assume that the predictions the most unexpected at first glance, and especially the accuracy of their coincidence with reality, rest on firm "historical" ground. These predictions are based on the experience of the past, on the analysis of past events subconsciously registered in our memories and under the influence of a definite set of causes called into the spherd of consciousness.

It is possible that a great part of the experience of the past is made up of information genetically registered in the living organism and representing the "concentrated experience" of ancestors.

Before we try to explain the possible structure of the mechanism of accumulation of experience and prediction, let us acquaint ourselves with some basic concepts. Let us define the problems of predicting determinate and probabilistic, or stochastic, processes, and let us also explain the concept of unpredictable "pure" randomness.

Prediction of Determinate Processes

Determinate processes are those caused by the action of a number of known causes. If we know the result of the action of each of them, we can exactly compute thd final result. Ordinarily (in linear systems) the principle of superposition is operative; this principle can be formulated thus: *the total effect of the action of several causes is equal to the sum of effects of the action of each cause taken individually*.

The study of determinate processes is based on the inductive method, the method of studying cause and effect. The majority of the laws of classical physics are determinate, primarily those relating to the mechanics of solid bodies. The orbits of the planets and stars can be computed to any required degree of accuracy. Hence we can quite accurately predict a lunar or solar eclipse or compute the position of a satellite.

The time interval separating the moment of prediction of some phenomenon from the moment when it begins is usually called the *anticipation time*.

The scientific foreseeing of determinate processes is characterized by the fact that the anticipation time may be arbitrarily large. *Increasing the anticipation time does not lower the accuracy of prediction of determinate processes*.

This rule does not hold for probabilistic, or stochastic, processes. The fact that processes are non-stationary means that prediction is only possible for a comparatively short period. Increasing the anticipation time for a required quality of prediction is the basic problem in working out methods for statistical prediction.

Prediction of Random Processes

If we repeat some observation or experiment many times, each time trying to reproduce the same conditions exactly, then instead of obtaining identical results, in each separate measurement we will obtain a result different from the others. Influence is exerted each time not only by the conditions we have reproduced, but also by those which we are not able to reproduce. An event subject to this kind of variance is called random. Sequences of such random events, considered as a function of time, are known by the name of random processes. In a random process we can follow the result of the action of a number of causes, but we cannot calculate it.

The study of random processes is based on the deductive method-- the causal connection of phenomena cannot be followed, although such a connection has an objective existence.

In the real processes observed in life, three components should be distinguished:

1) a determinate part, subject to exact calculation by the inductive method;

2) a probabilistic part, which can be elucidated by the deductive method by prolonged observation of the process with the aim of determining the probabilistic laws of the process;

3) a "purely" random part, which in principle cannot be predicted in any way.

Let us first consider examples from the field of prediction of "random" quantities. Thus, in tossing a die, one of whose faces is colored red, and the other five blue, it is required to predict what color the upper face will be at the next toss. It is easy to establish the absence of a determinate part in the given example; a probabilistic prediction gives the number 5/6, i.e. this is the probability with which we can predict that blue will turn up.

In a coin-tossing game, it is required to predict whether the coin will turn up head or tail. When the number of tosses is large, heads will turn up approximately in approximately half the number of cases, and tails in the remaining half. This is an example of "pure" randomness, or an equally probable outcome which cannot in principle be predicted in any way.

Another good example is any sufficiently complicated game, for example soccer. In predicting the results of the game there is no determinate component (nothing can be calculated), but there is a sharply expressed probabilistic component which can be determined by observing a number of games of the teams in question. Furthermore, the game must have an in principle unpredictable element of "pure" randomness. Without this element the game would cease to be a game.

Let us consider an example of an actual random process.

For a long time there was uncertainty over the question of the causes and laws of the tides. Kepler and Newton connected this phenomenon with the moon. Later Laplace confirmed Kepler's and Newton's theory in a strictly mathematical fashion; this made it possible to predict each day's ebb and flow time with great accuracy.

Let us consider the tides problem from the point of view of dividing the process into a determinate, probabilistic and "purely" random part. All three parts are found in this process. The determinate part of the process is determined by the moon and, to a smaller degree, by the sun and can be exactly computed by Laplace's theory. Furthermore, there is a random part caused by the wind, change in the composition and density of the water, the temperature and many other causes, part of which qre known to us. By long-term observation of the result of the action of these factors, we can determine the probability of deviations from the exact calculation, somewhat in the manner of a "wind rose" for a given locality on the ocean shore.

The aggregate of the determinate and probabilistic parts is the best (optimal) prediction. Comparison of this optimal prediction with the actual tide makes it possible to determine the element of unpredictable, or "pure", randomness. Mistakes of measuring instruments ordinarily make up a large part of this "pure" randomness. As measurement technology develops, this "purely" random part decreases. Everything which has been said relates to predicting the time of the tide and, in even greater measure, to predicting the increase in the water level. In the last problem, waves are of essential importance. An example of predicting the amplitude of waves is considered in detail in the fourth chapter.

In tossing a die, the determinate part of the process is equal to zero, since ordinarily nothing can be calculated. In tossing a coin, the determinate and probabilistic parts are equal to zero, i.e. the process is purely random. In the processe reflecting the ocean's ebb and flow, all three parts are present: the determinate, probabilistic and"purely" random. As the exact sciences develop, the determinate part, which is subject to exact calculation, continuously increases. The development of the theory and techniques of statistical predictions increases the reliability of probabilistic predictions. However, in actual processes the"purely" random part cannot be reduced to zero. This part determines the maximum level which we asymptotically approach as we raise the quality of prediction of the determinate and probabilistic parts of the process.

The working out of methods for calculating determinate processes and the elucidation of the probabilistic part are the basic problems of the theoy of prediction.

If a process has not been thoroughly studied, a certain share of its determinate part should be assigned to the probabilistic part. And further, a certain part of the probabilistic part should be assigned to "pure" randomness. This sharply decreases the accuracy of prediction.

As material is accumulated, definite regularities become clear which make it possible to make more certain predictions on the basis of cause-and-effect relationships and later to theoretical constructs. Although in many processes the element of "pure" randomness, which cannot be predicted in any way, cannot in principle be reduced to zero, yet the basic problem of the theory of prediction is maximally to increase the causal, determinate part and continuously to increase the accuracy of the probabilistic prediction. The part of the process which we refer to "pure" randomness with the best, optimal prediction is minimal and cannot be further decreased.

In some processes, which are called stationary, the probability characteristics are constant. Here, as observation time passes, the probabilistic part is predicted with greater and greater accuracy.

In the ideal case, when the observation time is large enough, the anticipation time may be considered to be arbitrary. Thus, we can very accurately predict the July mean temperature for several years in advance. High and low tides taking into account the prevailing winds can also be an example of a stationary process.

It is much harder to reduce to the minimum possible the unpredictable "pure" randomness in quasi-stationary processes, and even harder in non-stationary ones, whose probability characteristics change with the passage of time. An example could be prediction of the mean July temperature for many decades or even hundreds of years taking into account change in the climate of the earth. In fact every real process is non-stationary, but we may consider it to be stationary if its probability characteristics change little during the anticipation time. Hence in real random processes, in view of the fact that they are non-stationary, the prediction accuracy falls as the time increases.

In connection with this, a basic problem of the theory of statistical predictions is the working out of methods (formulas or algorithms) of prediction for which the anticipation time is greater than with other methods.

Let us consider some examples.

Prediction of processes from their parameters at a given instant

The simplest method of predicting the future consists of the assumption that "tomorrow will be the same as today." Let us note that this primitive method of weather prediction turns out to be right in 70% of the cases. The probability of correct prediction by the "without change" rule decreases extremely rapidly as the anticipation time increases.

Prediction for a longer period requires taking into account not only the present state of the process, but also its speed of change. A somewhat better method of prediction is based on the assumption that the percentage

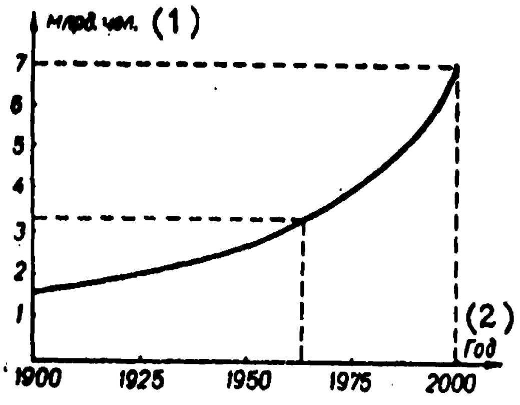


Figure 1. Growth in the population of the earth.
Key: 1) Billions of people; 2) year.

increase or decrease will remain constant. For example, this method is used in demography. Data on the population of individual countries and continents is processed by computers. The mean number of births and deaths per 1000 people is determined, and also the annual population increase in percent. Here the absolute growth increases from year to year. Figure 1 shows the curve for the growth of the world's population. From this curve we can predict that in 1975 the world's population will reach 4 billion. The assumption that the pecentage increase or decrease remains constant is only valid for a comparatively short period of time, when the conditions in which the predicted process takes place are almost identical. Hence there would be no sense in using this curve extrapolated to the 21st century.

Every real curve has its limitations. All physical quantities cannot exceed some "saturation".

Prediction for a prolonged period requires further complication of the formula by which the future values are determined. We may, for example, take into account not only the state of the process and the velocity of change, but also the acceleration, and possibly the third and higher time derivatives. In a number of cases this extended consideration gives good results, since the probability of correct prediction for longer periods increases. Even so, here too the period of correct prediction is determined by the properties of the process, the constancy of the coefficients of the prediction formula (state, velocity of change, acceleration, etc.). In a number of processes, which are called stationary, these coefficients are constant. For these processes, the indicated methods of prediction are very effective.

Prediction of processes from their parameters at a given instant and from their prehistory

In order to increase the anticipation time in predicting many processes, at a given instant it is necessary to take into account not only the parameters, but also their variation during the time preceding-- their <u>prehistory</u>. Weather prediction may serve as an example.

The first system of meteorological stations was organized in France in 1856, and in 1858 other countries, including Russia, joined this system.

The first meteorological observations in Russia belong to the time of the founding of St.Petersburg. Observations of the clearing and freezing over of the Neva dating from 1706, of the amount of precipitation from 1741, and of the temperature from 1753 are extant. A regular network of meteorological stations was organized ir 1830. However, only wide use of the telegraph made it possible to progress from predicting the weather from observations made at one point to more exact prediction of the weather by means of the preparation of synoptic maps.

Synoptic maps make it possible to follow the path of motion of cyclones and anticyclones. Thus, for the European continent there is a rule to the effect that if a cyclone moves to the east, there is a high pressure and high temperature region to the south of the center of the cyclone, and if a cyclone moves to the west, such regions lie north of the center, etc.

The use of special meteorological satellites has led to a great increase in prediction accuracy.

Long-term weather predictions are possible only when probabilistic methods are used. Determinate methods are evidently insufficient here.

The use of computers for weather prediction

In predicting the weather, we should also take in account the determinate part (influence of the sun, of the internal heat of the earth, etc.), the probabilistic part, and the element of "pure" randomness. For example, we can compute exactly that if the sun were to be extinguished, a uniform temperature of 141°C (i.e. a temperature much higher than absolute zero, 273°C) would be established on the earth's surface.

Increasing the accuracy of weather prediction means to reduce to a minimum the part which we assign to "pure" randomness, although this part will never be equal to zero.

At the present time approximately 20% of all weather predictions are wrong. There are reasons to assume that this figure can be reduced to 2-3%, while the predictions can simultaneously be made more concrete (can indicate the exact amount of precipitation, the exact limits of the region where it is precipitated, the exact temperature, etc.). The unpredictable "pure" randomness can be reduced to this small limit value.

For qualitative weather prediction it is necessary to solve a large number of equations describing the processes in the atmosphere with a large number of initial data which vary within a wide range. Thus, to predict the weather for a 24-hour period, approximately 3000 pieces of initial meteorological information must be taken into account.

For a 72-hour prediction, this figure has already risen to about 20,000. To solve the problem of long-term prediction, up to a season, about 100,000 pieces of information must be taken into account.

Processing such a huge volume of information is unthinkable without computers which are very fast acting and which have a large storage volume. Hence the Moscow World Meteorological Center has already completely switched over to weather prediction by means of computers.

The use of statistical methods requires the taking into account of various relationships and connections between active factors which have been brought to light by many years of investigation. At the present time a huge volume of information has been accumulated, and only the use of computers makes it possible to mobilize the "memory of the archives".

The computer makes possible continuous memorization of weather information arriving from numerous (counted in the tens of thousands) meteorological stations, processing of this information, and prediction of the weather on the basis of the direct solution of aerodynamical equations, and by computing probabilities (determinate and probabilistic methods). Hence weather prediction is a typical multivariate problem, since it requires indication of the variations of temperature, pressure and other quantities not only with time, but also over the surface of the planet.

Moscow State University is carrying on operative weather prediction. The use of fast-acting computers has made it possible for meteorologists to predict the pressure, wind velocity, temperature, etc. not by synoptic methods, as has been done up till now, but by the method of dynamic meteorology. The basic equations connecting pressure, wind velocity and temperature are the equations of motion, continuity, state and heat flow, in which all meteorologically unessential terms (so-called meteorological noise) are discarded. The problem of short-term prediction of meteorological elements consists of three stages: 1) analysis and processing of initial material; 2) prediction for time T of these intial data (T=12, 24, 36 or 48 h); and 3) prediction of the weather from the data obtained.

Solution of the problem of weather prediction for a 24-hour period takes 7 min of machine time.

At the NANWER laboratory (USA) a computer has been set up which prepares weather maps for the navy. The machine processes weather data arriving from 5000 meteorological stations and on the basis of these data prepares predictions for 24 hours in advance over the whole northern hemisphere. Weather information at the point of interest is found by interpolating the data obtained from meteorological stations located near this point. The computer processes individually data on pressure and temperature. Weather maps are drawn on the basis of the calculations. The weather prediction program was prepared on the basis of the statistical theory and laws of meteorology. Five minutes is required to predict one weather component (for example, pressure).

Important weather prediction data can be obtained from investigation of the upperlayers of the atmosphere, which is carried out by using meteorological satellites. Thus, satellites help to determine the places of origin

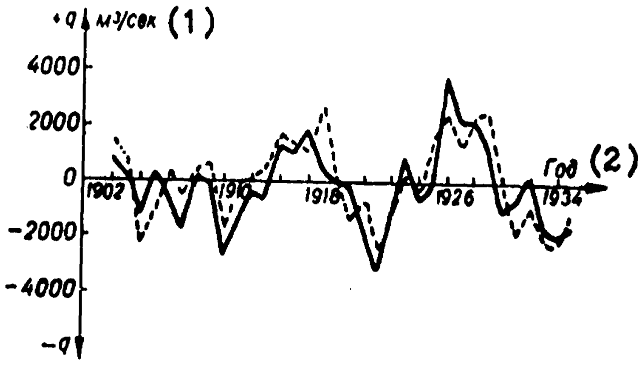


Figure 2. Prediction of the annual runoff of the Volga. Anticipation time 1 year.

Key: 1) m^3/sec; 2) year.

of typhoons and give a clear picture of overall planetary atmospheric processes. Color photos of the earth made by our cosmonauts also help in weather prediction.The huge quantity of rapidly arriving variegated information requires automation of observations and data transmission. Hence reliable weather prediction can be ensured only by detailed taking into account of meteorological data obtained on the earth and in space and by the development and use of suitable data transmission systems and computers.

Other geophysical predictions

It is hard to overestimate the value of prediction in determining the prospects and most expedient forms for using natural energy sources in the national economy /22/. Energy from river runoff is converted into electrical energy at numerous hydroelectric power stations. Solar

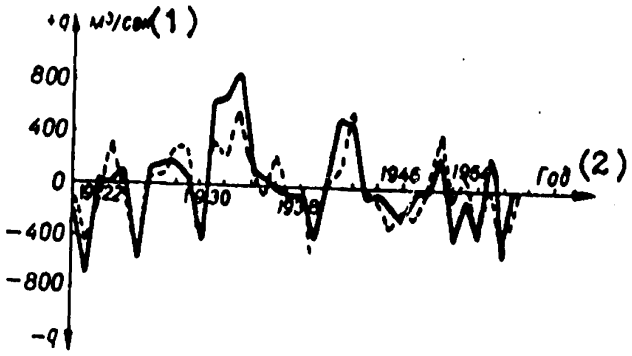


Figure 3. Prediction of the annual runoff of the Dnepr. Anticipation time 1 year.

Key: 1) m^3/sec; 2) year.

batteries feed the instruments and apparatus of satellites and space ships. The energy of the tides will be converted into electrical energy at our first maritime hydropower station, the Kislogub Tides Electric Power Station.

As for other forms of geophysical predictions, connected, for example, with the mean annual water discharge at rivers, the annual precipitation totals over large areas, the annual energy totals of earthquakes, etc., great success has been obtained due precisely to the use of probabilistic methods of prediction. Thus, Yu.M.Alekhin /1/ has successfully applied the method of linear extrapolation of random time sequences to predicting the annual runoff of rivers.

Figures 2 and 3 show graphs which reflect the results of predicting the annual runoff of the Volga and Dnepr. The results were obtained for an anticipation time of one year.

Prediction of earthquakes

Large earthquakes liberate a huge amount of energy, stored in the stresses of the rock layers of the earth, which is equivalent to the simultaneous explosion of several atom bombs. Earthquakes arise unexpectedly, and it has not been possible to find whether they depend on time in any regular way. At the same time, as far as space is concerned, the earthquake probability is clearly shown: 75% of all earthquakes occur in the seismic belt surrounding the Pacific Ocean, 20% are observed in a second seismic belt passing through Burma, the Himalayas, Iran, the Mediterranean Sea and the Azores. Only 5% of all earthquakes occur outside these two belts. Thus, prediction of the exact time of an earthquake is the most difficult problem.

A network of observation points is being organized to predict earthquakes in seismic regions with sufficient accuracy. Complex measuring devices are used to measure the contraction and inclination of the earth's surface. Indirect quantities are measured: speed of passage of seismic waves, changes in the electrical conductivity of the earth, and magnetic declination.

Japanese scientists, in particular, have shown that the change in the magnetic declination caused by contraction of the upper layers of the earth's surface is the most essential factor in making it possible to increase the accuracy of prediction of the time of an earthquake. Figure 4 shows a typical curve for the variation of the magnetic declination /5/.

The characteristic peak of the increase in the magnetic declination precedes a strong earthquake. It is evident from the curve that the time can be predicted for several months with an accuracy of up to two or three

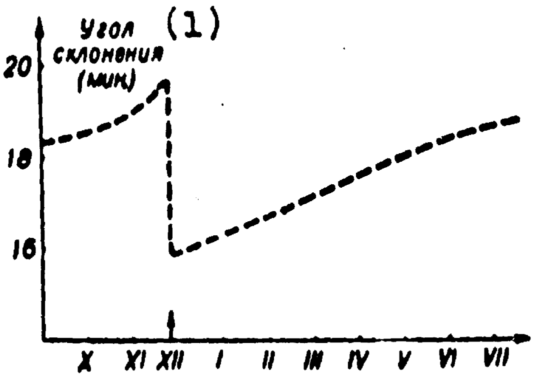


Figure 4. Change in the magnetic declination.
Key: 1) Angle of declination (min.).

weeks. Further increase in prediction accuracy is also possible.

Prediction of the level of ground water

In connection with the design and construction of hydrotechnical installations, for example reservoirs, the problem often arises of predicting the level of ground water in the surrounding mountains.

Variation of the ground water level is a protracted and aperiodic process. Prediction consists of calculating the displacement of the boundary of the free surface (depression surface), for which it is required to solve nonlinear equations of the parabolic type. Here there are concrete initial and boundary conditions which define the history of the process and the geological structure of the mountains and which take into account variation of the filtration factors in the volume being investigated.

The difficult part of the problem is the solution of these equations. By linearization, the function under investigation can be reduced to an equation of the type of the heat conduction equation.

Complex computing devices are used to solve such problems: grid integrators, models using electrohydrodynamic analogy, computers.

The determinateness of prediction in the given case is determined by a clearly expressed and exactly defined influence.

There are cases where the variation of the ground water levels is determined by a set of different causes. In such cases, variation of the ground water level can be interpreted as a stochastic process depending on irrigation, drainage, amount of precipitation, and fluctuations of the water level in rivers (the two latter influences are in themselves stochastic). In these cases, the whole theory of prediction of random processes is completely applicable.

Prediction of correlated processes

In the example of the problem of predicting earthquakes, it is important to note that, unlike the other problems, the prediction here is made by observation of processes connected (correlated) with the process in which we are interested (by magnetic declination). This method is a general one and has a wide range of application.

Often the prehistory of a process which we need to predict cannot be traced, but we have data on another process connected with the first by a functional or correlational connection. For example, in regulating an industrial process, we can predict the changes in some parameter or other without resorting to direct measurement of it, but using data on another parameter connected with the first. This is especially important in cases where some of the parameters are hard to measure, or where directly taking information from the object is undesirable.

In the most general prediction formula, besides terms which reflect the basic process at a given moment and its prehistory, there are terms which determine the parameters and prehistory of other quantities correlated with the given process.

Attempts at setting up formulas (rules or algorithms) for the most exact predictions, which take into account all the above factors, show that the use of such formulas is connected with a huge volume of computational work. Also, programming on large computers is only possible with comparatively simple algorithms. Hence a feedback system should be used. In the program for a computer or in a specialized element-by-element device, a search is carried out for the best prediction algorithm (with the given volume of the device). The machine automatically leaves in the program (prediction formula) only those terms whose effect on the prediction accuracy is shown to be essential. In chapter 4 we shall consider the examples of predicting the amplitude of ocean waves, operational indices of an enterprise, levels of a river bottom and prediction of atmospheric pressure, and these examples will give more concrete expression to this method of self-adjustment of the prediction formula.

One-dimensional and multidimensional prediction problems

In the simplest case it is required to predict the variation of a quantity or series of quantities with time. This prediction problem may be called one-dimensional. The prediction of the weather on the earth's surface is an example of a more complicated multidimensional problem, since it is required to consider the process not only with respect to time, but also with respect to space. Sometimes cases arise in which the processes are purely random with respect to time and at the same time are probabilistic with respect to space. As examples, we may cite the problems of predicting the load of power systems, the distribution of agricultural plant pests, the prediction of earthquakes, and many others.

The use of mathematical prediction in the planning and regulation of power systems

During the past years, computation centers having digital computers at their disposal have been set up in every large power system. These centers solve the problems of the optimal development of the system, expansion of the existing electric power stations and the construction of new high-power ones. Long-term predictions are made to determine the possible demands for electric power and heat. This prediction must determine the level of development of the generators and transmission network, as well as the yield of fuel and the development of other power sources.

This problem is obviously multidimensional, since it is necessary to determine the variation of the quantities involved not only with respect to time, but also with respect to space: it is required to point out the places of concentration of consumers and generating stations.

Prediction for protection of plants from pests and disease

Prediction for plant protection is also a multidimensional problem. It is required to predict where, when and in what quantity plant pests will appear in order to take the necessary protective measures. As of yet, mathematical methods have not been used to solve this problem. Prediction is made by purely empirical rules. For example, by counting the number of pupae in spring, we can predict the number of caterpillars in summer, etc. Reference [38] cites examples of the successful prediction of the appearance of the Coloradian beetle, potato blight, etc. There is every possibility of increasing the prediction period and accuracy with the use of digital computers.

Prediction in biology and medicine

The past decade has been marked by the intensive introduction of mathematical methods and the techniques of computation and technical cybernetics into medical and bilogical research and practical medicine. For a number of years "medical" mathematics has been in the main limited to the use of the methods of mathematical statistics for processing the results of observations and investigations and for quantitative evaluation and confirmation of the correctness of conclusions.

At present there has been a great rise in the interest shown by physicians and other biological scientists in various mathematical methods right up to the latest achievements in the field of technical cybernetics (information theory, game theory, queueing theory, theory of pattern recognition, etc.).

The use of the latest techniques of computation and technical cybernetics has made possible a qualitatively new approach to the solution of many problems connected with the investigation of living organisms.

The new sciences which have arisen as a result of the fruitful cooperation of mathematicians, biologists, and engineers-- biological cybernetics, bionics, neurocybernetics-- are developing rapidly and enriching biologists with new data on the living organism, at the same time helping those working in the exact sciences to take into account the experience accumulated by nature when they develop their high-efficiency technical devices.

Another big step in this direction is the use in biology and medicine of the theory of statistical prediction and corresponding techniques.

The development of neurosurgery and heart and lung surgery, with the ever increasing complexity of the methods of surgical intervention in vitally important organs, brings out the importance of the problem of developing automatic regulators for a number of physiological parameters of the human organism.

In the process of developing these regulators, it is necessary to take into account the special features of the reaction of the living organism which determine its compensatory possibilities. Sudden and sharp functional disorders do not arise immediately after the onset of the action of a noxious factor, but only after a definite time, during which the compensatory mechanisms are disturbed and then break down. Hence for timely connection of automatic regulators and in the process of their operation, a definite anticipation is necessary.

Predicting devices operating in sequence with instruments which register changes in various indices make it possible to forestall possible disturbances in the course of operation.

The use of predicting devices for processing the data of investigation of patients with progressive illnesses will make possible a more exact judgement as to whether a given method of cure is timely and indicated.

Prediction in control of industrial processes

Modern industrial enterprises are marked by a high degree of automation.

The continuous increase in the number of measuring and recording devices, modelling by means of computational techniques, the study of statistical and dynamic characteristics of units-- all these measures are intended to attain a basic goal: the optimal regulation of industrial processes.

The technological installations of chemical enterprises, the units of the metallurgical industry, large organizational and planning systems, and many other units are characterized by great inertia. For example in the oil industry, when automatic quality analyzers are used, the results of the analysis become known 20 to 25 min after selection of a product specimen.

Thus, regulation by quality index is performed with a large lag. It is obvious that if the instruments were capable of predicting the future changes in technological parameters on the basis of an analysis of their preceding changes, the quality of regulation could be greatly increased.

Large volume and laboriousness of operation characterize the enterprises of the mining industry. All-round automation of large coal and ore pits is an extremely pressing problem. The optimal regulation of mining machines and transport units, dispatcher's service, and many other preblems are solved by using the latest mathematical methods, the techniques of cemputation and technical cybernetics. In developing automatic centrel systems for mining-transport complexes and individual machines, great effect may be obtained by using the methods ef the theery ef statistical predictions and devices based on these methods.

Further success can be expected in connection with the use of the theory and techniques of statistical prediction in the sphere of organizing and planning the national economy taking long-term plans inte account.

At different stages of preduction autemation and for different centrol preblems, the methods and techniques fer solving these problems must be different. If, for example, a control problem can be solved with sufficient effectiveness by a single-circuit automatic control system, then in order to raise the quality of regulation with prediction, a specialized predicting device may be used which is based on a definite prediction algorithm. With multicircuit regulation we can alse develop a specialized device for predicting production indices depending on many factors.

However, taking into account the fact that centrel problems are constantly growing more complex and that controlling computers are consequently used, in a number of cases it is not necessary te develop specialized predicting devices. Their functiens can be successfully fulfilled by the controlling machines.

As concerns specialized devices fer predicting the future values ef various production indices, here great success can be expected from the predictive use of pattern and situation recognition systems. Intensive work is presently being done in the field both here and abroad.

In conclusion it should be said that the use of the achievements in the theory and techniques of prediction is a necessary condition for further improvement on the road to optimal control.

The examples considered do not at all exhaust the fields of application of the theory of statistical prediction. As the theory is further developed and perfected, its methods will undoubtedly be ever more widely introduced into the practice of scientific investigation and into various sectors of the national economy.

The succeeding chapters will consider methods for predicting determinate and random processes.

Much attention has been given to the modelling of predicting filters on universal digital computers and to the use of various methods for predicting actual processes.

Especial space is devoted to exposition of the problems connected with the use of recognition systems as predicting filters.

Chapter 1.

Forecasting determinate processes. Interpolation and extrapolation

Problems of interpolation and extrapolation

In determinate processes the random deviations are so small that these processes can be calculated in advance quite accurately. Examples of such processes are the motion of the heavenly bodies, and also the motion of simple mechanisms, for example a pendulum, every change or displacement occurring exactly according to a time table or a graph, etc. The laws governing such processes are sometimes known and can be expressed in the form of analytical functions, graphs or tables (for example, a train timetable).

Often these functions are unknown. But they exist, and in accordance with them this or that process or motion occurs. These functions are the solutions or, in other words, the integrals of the dynamic equations of the mechanisms or systems in which we are interested.

In studying determinate processes, there arise two types of problems connected with the determination of the values of some function at the points which interest us from the known values of these functions at other points. Let us consider these problems.

The problem of <u>interpolation</u> consists of finding values of a function within the segment of observation. Here, the function itself, as was pointed out, may not be known. But in the majority of cases it is necessary to know to which class of functions it belongs, i.e. whether it is expressed by a straight line, a parabola of the second degree, a cubical parabola, a harmonic function, etc.

Let there be known the values of the function $f(t_i)$; $i=1,2,\ldots,n$ at the points $t_0 < t_1 < \ldots < t_n$. It is required to determine the values of this function at the points t_j, which lie between the given points $t_i < t_j < t_{i+1}$.

For example, with linear interpolation the value of the function at some median point $t_0 < t < t_1$ is

$$f(t) = \frac{t - t_0}{t_1 - t_0}[f(t_1) - f(t_0)] + f(t_0). \qquad (1)$$

The problem of <u>extrapolation</u> consists of finding values of a function at a point lying outside the region of observation from its values within this segment. The most common types are linear and parabolic extrapolation, with which the function is expressed by a parabola of the second, third or higher order. Ordinarily, the less the time for which the process is extrapolated, the more exact is the determination of the future value of the function. This is due to the fact that the indicated functions only approximately represent (approximate) the actual laws governing the process.

Selection of an approximating polynomial

As has already been said, the form of an approximating function is determined by the physics of the process and, consequently, corresponds to the form of the solutions (integrals) of the dynamical equation of the system. For example, if it is known that some set of numbers expresses the angle of deviation of a pendulum, it is clear that they must satisfy the law of harmonic oscillations. The problem is much more complex if the physics of the problem is not known and we do not know the form of the solution function. Then we should choose the form of the approximating function so that it will pass through the given points in some optimal fashion.

In many cases the initial information is given in the form of a finite set of points (selection), and the problems of interpolation and extrapolation will be completely solved if we find an analytical expression which all these points satisfy.

Let us assume that we are given the following selection of data:

t =	1	2	3	4	5
f =	1,111	1,248	1,417	1,624	1,875

It is assumed that the selection is sufficiently representative, i.e. reflects sufficiently well all the basic properties of the function. Let us begin the selection of the approximating polynomial with the simplest expression. Let us assume that the process is described by the straight-line equation

$$f^{*1} = a + bt \qquad (2)$$

Arbitrarily selecting two points of the selection (for example, the first and the last), let us write the straight-line equation twice

$$1{,}111 = a + b \cdot 1$$

$$1{,}875 = a + b \cdot 5$$

We have obtained a system of two equations in two unknowns, i.e. the coefficients of the approximating polynomial a and b. Solving these equations simultaneously, we obtain

$$a = 0{,}920; \qquad b = 0{,}191$$

Now we can see whether we correctly guessed the form of the approximating polynomial. For this purpose let us find the values of the approximating function at the same values of the argument

*Here and following, an asterisk indicates a predicted value.

t =	1	2	3	4	5
f* =	1,111	1,302	1,493	1,684	1,875

The accuracy of the approximattion can be estimated from the variation

$$\delta = \frac{\overline{(f_i - f_i^*)^2}}{\overline{f_i^2} - \bar{f}_i^2}\, 100 \approx 2,2\%. \qquad (3)$$

The less the variation, the more exactly we have chosen the approximating polynomial and the nearer will be the predicted values to the actual ones.

Let us repeat the same investigation for a parabola of the second degree:

$$f^* = a + bt + ct^2 \qquad (4)$$

Arbitrarily choosing three points (for example, the beginning, middle, and end of the selection), we obtain a system of three equations in three unknowns·

$$\begin{aligned} 1{,}111 &= a + b\cdot 1 + c\cdot 1, \\ 1{,}417 &= a + b\cdot 3 + c\cdot 9, \\ 1{,}875 &= a + b\cdot 5 + c\cdot 25. \end{aligned}$$

Solving these equations, we find:

$$a = 1{,}015; \quad b = 0{,}077; \quad c = 0{,}019.$$

The quadratic approximating polynomial gives this sequence of values for the function:

$t =$ 1	2	3	4	5
$f^* =$ 1,111	1,245	1,417	1,627	1,875.

Let us find the variation:

$$\delta = \frac{\overline{(f_i - f_i^*)^2}}{\overline{f_i^2} - \bar{f}_i^2}\, 100 = 0{,}79\%.$$

We see that the variation has decreased. Hence the polynomial of the second degree is a much better approximation to the given function.

In order to raise the accuracy of approximation still further, let us pass to the polynomial of the third degree:

$$f^* = a + bt + ct^2 + dt^3. \qquad (5)$$

Preceeding analogeusly, let us write the system of four equations in four unknowns:

$$\begin{aligned} 1{,}111 &= a + b\cdot 1 + c\cdot 1 + d\cdot 1, \\ 1{,}417 &= a + b\cdot 3 + c\cdot 9 + d\cdot 27, \\ 1{,}624 &= a + b\cdot 4 + c\cdot 16 + d\cdot 64, \\ 1{,}875 &= a + b\cdot 5 + c\cdot 25 + d\cdot 125. \end{aligned}$$

Solving the equations simultaneously, we find

$$a = 1{,}0; \quad b = 0{,}1; \quad c = 0{,}01; \quad d = 0{,}001.$$

Determining the variation, we see that it is zero:

$$\sigma = 0.$$

Hence the third-degree polynomial exactly describes the initial function. If such a result cannot be obtained in other cases, we should stop at the approximating polynomial which gives a sufficiently small variation, of the order of a few percent. If this cannot be achieved and the variation remains large, this may be a sign that the initial process is not determinate, that, besides the regular component in it, there is a large random component. In this case the methods for selecting an approximating polynomial discussed here are no longer valid. We must have recourse to the methods for predicting random processes, which will be discussed in chapter 2, which is devoted to the prediction of random processos.

But if we obtain an expression which gives a small or (better) a zero variation, the problem of interpolation and extrapolation becomes trivial. Using the expression obtained, we easily find the values of the function in which we are interested at any moment of time both in the past and in the future.

In the numerical example considered, we can "predict" that when t=6, f= 2.187. We have deciphered the process; we have found and equation which describes it.

Let us consider another example. Let us cite data on the population of Europe for the period from 1850 to 1930 (in millions): 1850, 267; 1860, 284; 1870, 306; 1880, 332; 1890, 364; 1900,399; 1910, 441; 1920, 449; 1930, 491.

Let us assume that we know only the values of the population for 1860, 1870 and 1880. On the basis of this information let us determine the population in 1864; i.e. let us solve the problem of interpoltion. Using the formula for quadratic interpolation (4), we obtain:

$$a-b+c=284,$$
$$a=306,$$
$$a+b+c=332.$$

For values at different distances from one another it is convenient to denote the argument as follows: we take the first value, 1860, to be -1, the second to be 0, and the third to be 1. Then the value of the argument at the point to be predicted, 1864, is 0.6.

Solving these equations, we find the values of the coefficients

$$a=306;\quad b=24;\quad c=2.$$

Taking into account the computed values of the coefficients, let us write the interpolation formula in the form

$$f = 306 + 24t + 2t^2.$$

Substituting t=-0.6 into it, we obtain

$$f_{-0,6} = 306 + 24(-0,6) + 2(-0,6)^2 = 292,32.$$

Rounding off to integral values, we obtain a population of 292 million.

Now, proceeding from the assumption that the law which holds within the interval is valid outside of it, let us determine the population in 1900, 1910 and 1920. This is a problem of extrapolation.

To solve the problem, let us use the basic properties of interpolation formulas. For an interpolation formula of the n-th order, these properties consists of the following:

a) n-th order differences

$$\Delta_1^n = \Delta_2^{n-1} - \Delta_1^{n-1}, \Delta_2^n = \Delta_3^{n-1} - \Delta_2^{n-1} = \ldots = \text{const}.$$

b) differences of the n+1-th order

$$\Delta_1^{n+1} = \Delta_2^n - \Delta_1^n = \Delta_3^n - \Delta_2^n = 0.$$

For quadratic extrapolation we obtain

$$f_{1900} - 3f_{1890} + 3f_{1880} - f_{1870} = 0,$$

whence

$f_{1900} = 3 \cdot 364 - 3 \cdot 332 + 306 = 402$ (millions of people)

If we use cubic extrapolation, i.e. put the fourth difference equal to zero,

$$f_{1900} - 4f_{1890} + 6f_{1880} - 4f_{1870} + f_{1860} = 0,$$

we obtain a similar result:

$f_{1900} = 4 \cdot 364 - 6 \cdot 332 + 4 \cdot 306 - 284 = 404$ (millions of people)

In fact the population for 1900 was 399 million.

As we see, the deviations are not very large. By computing the variation, we can see which formula gives the best predicted value.

Let us use these formulas for predicting the 1920 population.

$$f_{1920} = 3 \cdot 441 - 3 \cdot 399 + 364 = 490$$
$$f_{1920} = 4 \cdot 441 - 6 \cdot 399 + 4 \cdot 364 - 332 = 494$$
(millions of people)

But in fact the 1920 census showed a European population of 449 million. The values 490 and 494 calculated by the formulas approximately coincide with the results of the 1930 census, 491. Thus the prediction was wrong. But it was not without use, since it enabled us to estimate the great damage done to the European population by the first imperialist war.

Automatic interpolation

The solution of problems similar to those considered in the preceding example is the task of statistics. The volume of information to be processed is constantly increasing, and the problems themselves are becoming more and more complex. Statisticians are being helped by universal computers.

The problem of automation of industrial processes and optimal control of various units has required the development of specialized devices which make it possible to solve interpolation and extrapolation problems. For example, for programmed control of metal-cutting lathes, it has been required to develop devices which could reproduce the whole path of the motion from the known coordinates of several points.

These devices have received the name of automatic interpolators. They are widely employed in laying out metal plates and sheets, in servomechanisms for controlling units which require high path accuracy, etc. Let us consider some examples of the simplest interpolators.

Linear interpolators

Figure 5 shows a block diagram of a linear digital interpolator whose initial data are the values sinα and cosα , where α is the angle of inclination of the path to be interpolated to the x-axis.

A similar interpolator is used for controlling the feed of the cutting instrument in automatic par.s working. When a segment of length l is being worked, the values of the sine and cosine of the angle of inclination of the segment being worked to the x-axis are introduced inte registers 4 and 6. The value of the segment with given angle of inclination at a fixed generater frequency is determined by the time it takes the pulses to arrive at the dividing circuit 2. This time in turn is given by the pulse sum x+y (the number of pulses corresponding to complete coordinate displacement). As soon as a pulse sum x+y equal to the number registered on counter 8 is established, the latter emits a cycle termination pulse, and pulses from the generator cease arriving at input 2. The insturment feed is controlled by signals at the busbars "x-axis" and "y-axis".

Let us cite some more examples of linear interpolators. In the interpolator whose block diagram is shown in Fig.6, the initial data are tanα and Δ x.

The linear interpolator in Fig.7 is constructed

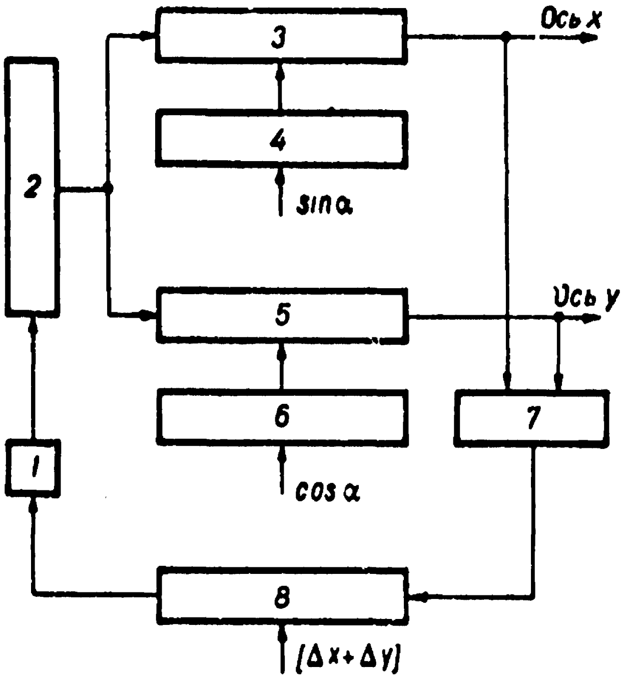


Figure 5. Linear interpolator: 1) pulse generator; 2) frequency divider; 3,5) rectifiers; 4,6) registers; 7) "or" circuit; 8) counter; 9) cycle termination pulse.

on the basis of a digital integrator [42]. Its principle of operation consists of the following.

If a constant number is introduced into the register of the interpolator, then in accordance with the expression

$$y = \frac{t}{T} x \qquad (6)$$

when $t=T$ we obtain

$$y = x$$

Here at the output during a time equal to the period

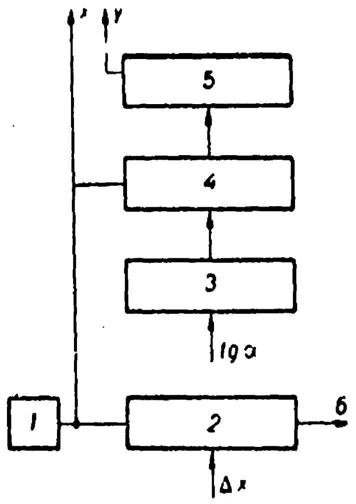


Figure 6. Interpolator with initial data Δx and tan α : 1) pulse generator; 2)counter; 3) register; 4) rectifiers; 5) summing device; 6) cycle terminatıon pulse.

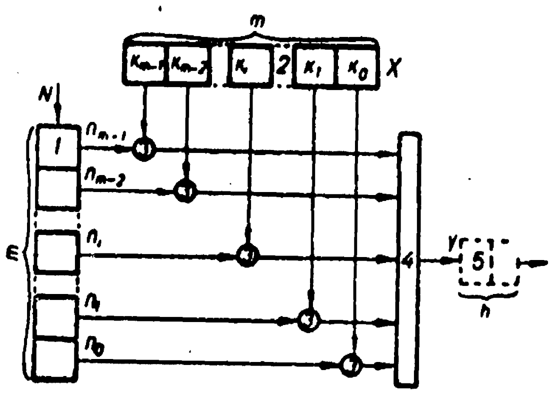


Figure 7. Interpolator on the basis of a digital interpolator: 1) frequency divider; 2) register; 3) coincidence circuit; 4) "or" circuit; 5) averaging cells.

T of operation of the frequency divider there appears a number of pulses equal to x.

The interpolator supplies discrete values of the function y, which we shall denote by y*.

Errors in automatic interpolation and methods for raising accuracy

Of especial interest is the problem of the magnitude of the error of digital interpolators and the methods for raising their accuracy.

In /42/ B.A.Sigov gives an expression for positive

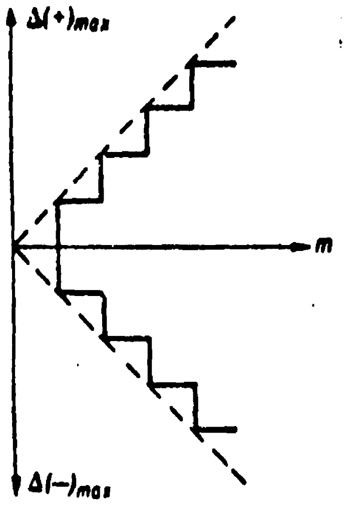


Figure 8. Interpolation error as a function of the number of discharges of the initial number.

and negative maximum error:

$$\Delta(\pm)_{max} = \pm \frac{3m+7}{18}. \qquad (7)$$

Here m is the number of discharges of the initial number x. It is obvious that as m increases the error increases, and when m > 3 we can assume that the error increases linearly. Figure 8 shows this function in the form of a graph.

To decrease the error Δ_{max} it was proposed to introduce a certain number of trigger cells, h, into the output circuits. In what follows we shall call these averaging cells. A dashed line surrounds these cells in Fig.7.

In the absence of averaging cells, the number of pulses at the output of the interpolator would be equal to the sum of the numbers of pulses along the open channels:

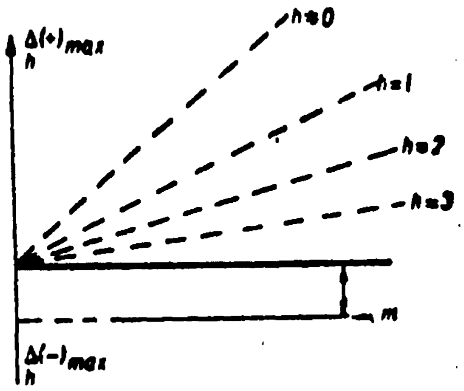


Figure 9. Interpolation error as a function of the number of averaging cells.

$$y^* = \sum_{i=0}^{m-1} k_i n_i. \qquad (8)$$

Now the value of the output quantity can be written in the form

$$y^* = \left[\frac{\sum_{i=0}^{m-1} k_i n_i}{2^h}\right], \qquad (9)$$

where h is the number of averaging cells.

For the maximum positive and negative errors, we will have the expressions

$$\underset{h}{\Delta}(+)_{max} = \frac{3m + 7 \pm 2^{-m+1}}{18 \cdot 2^h} \qquad (10)$$

and

$$\underset{h}{\Delta}(-)_{max} = -\left(\frac{3m - 11 \pm 2^{-m+1}}{18 \cdot 2^h} + 1\right). \qquad (11)$$

It follows from these formulas that, as the number of averaging cells is increased, the maximum positive and negative erros decrease (Fig.9).

The error can be further decreased by introducing some initial number s (bias) into the averaging cells.

The expression for y* can now be written in the form

$$y^{*}_{n,s} = \left[\frac{\sum_{i=0}^{m-1} k_i n_i + s}{2^n} \right]. \qquad (12)$$

Omitting the intermediate conversions for the above expressions for $\Delta(+)_{max}$ and $\Delta(-)_{max}$ and taking into account the bias s, we obtain

$$\lim_{\substack{h \to \infty \\ s = 2^{h-1}}} \Delta(+)_{max} = \frac{1}{2}, (13)$$

and

$$\lim_{\substack{h \to \infty \\ s = 2^{h-1}}} \Delta(-)_{max} = -\frac{1}{2}.$$

Indeed, the number h is unconditionally finite, it being in practice very small. Usually the properties of circuits with 4 to 5 averaging cells are very close to limiting. For example, for m=30 we obtain

$$\Delta(+)_{\substack{max \\ h = 0,\ s = 0}} = 5{,}4;$$

$$\Delta(-)_{\substack{max \\ h = 0,\ s = 0}} = -5{,}4,$$

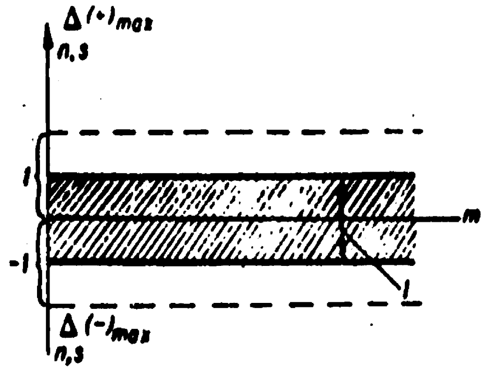


Figure 10. Decreasing the interpolation error by the initial bias method.

and

$$\underset{h=5,\ s=2^4}{\Delta(+)_{max} = 0{,}66;} \quad \underset{h=5,\ s=2^4}{\Delta(-)_{max} = -0{,}64.}$$

Decreasing the error by means of bias is shown in Fig.10. The cross-hatched error zone is equal to the discreteness step. By changing the bias, we can obtain only positive or only negative errors or locate the error zone symmetrically with respect to the horizontal axis.

The use of averaging cells and initial bias has also given positive results in the circuits of quadratic interpolaters. These circuits also showed a great increase in accuracy.

Linear-circular interpolator

When the path to be interpolated is a circle, linear-circular interpolaters are used [24].

Figure 11 shows a circuit for such an interpolator. The device consists of two integrators and an inverter and is intended to solve the differential equation of the form

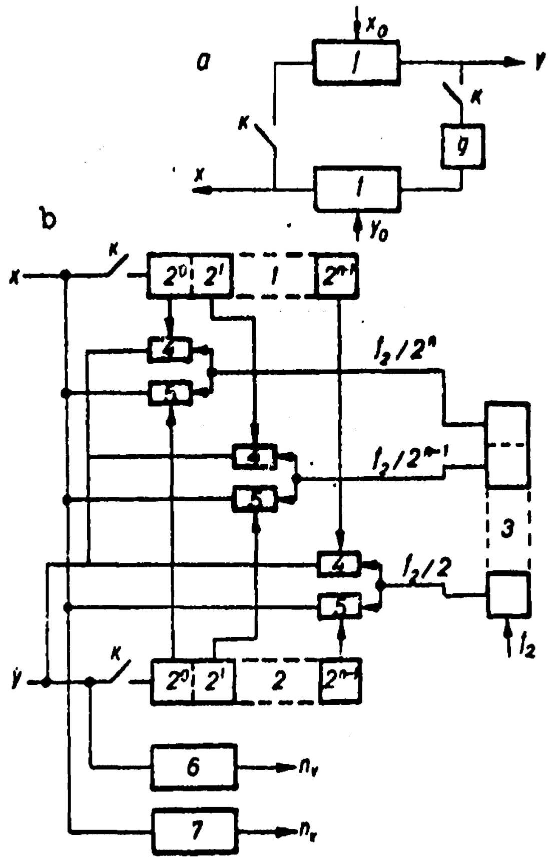


Figure 11. Linear-circular interpolator: a) functional diagram; b) block diagram; 1,2) reversible counters; 3) frequency divider; 4,5) rectifiers; 6,7) displacement measurers; K) keys; 9) inverter.

$$\frac{dy}{dx} = -\frac{x}{y} .$$

The solution of this equation is the circular equation

$$y^2 + x^2 = R^2$$

If the keys K are open, interpolation is done on a straight line with an angle of inclination to the x-axis of

$$\alpha = \text{arc tg} \frac{x_0}{y_0}$$

When one of the keys K is open, a parabola is reproduced; and, when the inverter 1 is switched out of the circuit, a hyperbola is reproduced.

At the beginning of operation of the interpolators, the total displacements are set in the displacement measurers 6 and 7 in an auxiliary code. The operation of the circuit continues until registers 6 and 7 are overfilled by the output control pulses arriving at them.

Quadratic interpolators

To interpolate second-order curves of the form

$$y = a + bx + cx^2$$

quadratic, or parabolic, interpolators are used. Figure 12 shows the block diagram of a parametric quadratic interpolator. Its operation, as with most parabolic interpolators described in the literature, is based on a difference method. The quantities x_i and y_i are accumulated in the summing devices 2 and 8, integral parts of these quantities being emitted in the form of pulses along the x and y-axes. The

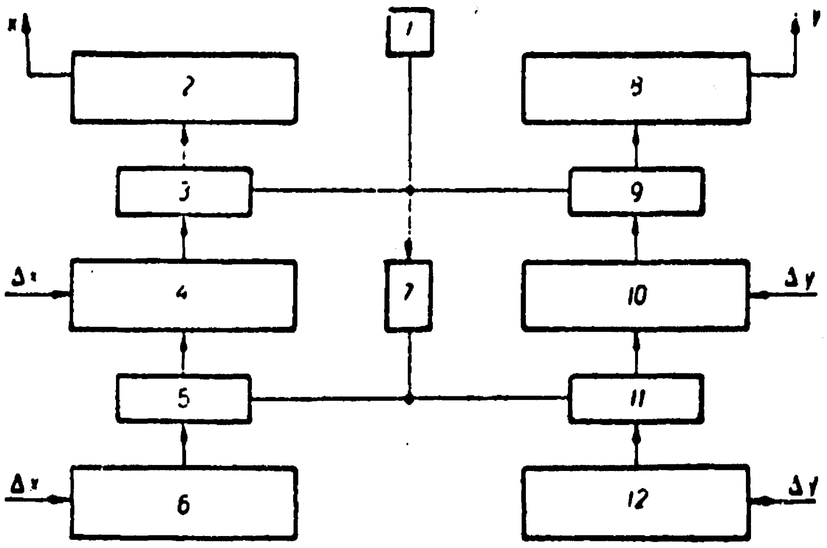


Figure 12. Parametric quadratic interpolater: 1) pulse generator; 2,4,8,10) summing devices; 3-9) rectifiers; 6,12) registers; 7) frequency divider.

current values of the first differences $\Delta x_i(t)$ and $\Delta y_i(t)$ are recorded in the summing devices 4 and 10. The values of the second differences $\Delta^2 x_i(t)$ and $\Delta^2 y_i(t)$ are kept in registers 6 and 12.

The operation of the circuit is described by the difference equations:

$$\Delta x_i (t) = \Delta x_{i-1} (t) + \Delta^2 x(t)$$

$$\Delta y_i (t) = \Delta y_{i-1} (t) \Delta^2 y(t)$$

Figure 13 shows the block diagram of an interpolator in which the x-coordinate is given by the equation

$$x = at^2 + bt + c,$$

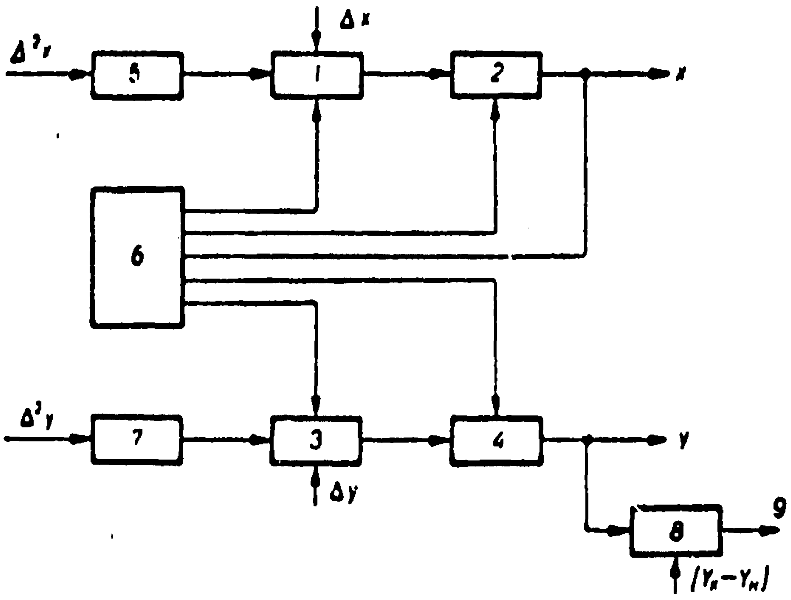


Figure 13. Quadratic interpolator with the argument given by the equation $x=at^2+bt+c$: 1,2,3,4) summing devices; 6,7) squares; 5) control device; 8) counter; 9) end-interpolation signal.

and the y-coordinate varies in accordance with the equation

$$y = a_1x^2 + b_1x + c_1.$$

A difference method is used to solve these equations. Differences with respect to the x-coordinate are summed in summing devices 1 and 2, and those with respect to the y-coordinate in devices 3 and 4. Before the beginning of operation, the difference between the finite y_K and initial values of the coordinate y_H is registered on counter 8 in an auxiliary code. After the counter is overfilled, a signal for transition to the following path section is emitted.

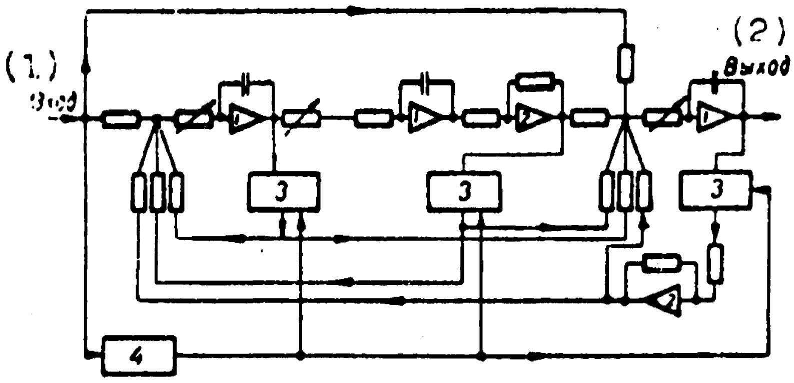


Figure 14. Cubic interpolator: 1) integrating amplifiers; 2) scaling amplifiers; 3) pulse elements; 4) pulse generator.
Key: 1) Input; 2) output.

Interpolators of higher orders

Figure 14 shows the circuit of an electronic device for cubic interpolation. A discrete sequence obtained by means of pulse-amplitude modulation is used as input signals. Interpolation is performed according to the formula

$$f(a+T+\tau)=f(a)+\Delta f(a)\,\frac{\tau}{T}\left\{\Delta f(a)+\frac{\Delta^{2} f(a)}{2}-\right.$$
$$\left.-\frac{\Delta^{3} f(a)}{6}\right\}+\frac{\tau^{2}}{2!T^{2}}\,[\Delta^{2} f(a)]+\frac{\tau}{3!T^{3}}\,[\Delta^{3} f(a)],$$

where f(a +T+τ) is the value of the function to be interpolated at instant a +T+τ ; a is the initial instant; T is the arrival period of the discrete values of the function f; Δ^1f, Δ^2f, Δ^3f are respectively the first, second and third differences of the function f; 0 $\leqslant \tau \leqslant$ T.

Δ^1f(a), Δ^2f(a) and Δ^3f(a) are computed from the formulas:

$$\Delta^1 f(a) = f(a+T) - f(a),$$
$$\Delta^2 f(a) = f(a+2T) - 2f(a+T) + f(a),$$
$$\Delta^3 f(a) = -f(a) + 3f(a+T) - 3f(a+2T) + f(a+3T).$$

The circuit of the interpolator contains three integrating amplifiers, 1, two inverting amplifiers, 2, three pulse elements, PE, which are controlled from the pulse generator, PG. PG is controlled by the input pulses. The pulse elements synchroneusly and in phase with the arrival of the input pulses takes and nemembers for one period the voltage from the outputs of the integrating amplifiers. The constant times of the integrating amplifiers and the transfer constant of the inverters, and also the weighting factors for the components in the two summing circuits at the inputs of the integrating amplifiers 9 and 12, are chosen so that the voltage at the output at instant a +3T+ τ will be equal to the value of the function to be interpolated at instant a +T+ τ . This is accomplished by suitable selection of resistances.

The interpolator operates with delay 2T. The PE circuit consists of a memory commutator, a key tube and two separation amplifiers. The key tube is a dual triode, half of which is connected in antiparallel. It is controlled by pulses from the PG.

The input and output amplifiers are cathode followers, with a tube instead of a cathode resistance, based on a dual triode.

Figure 15 shows the block diagram of an interpolator of the fourth degree. The external devices of the interpolator are: a counting device 1 and a magnetic tape re-

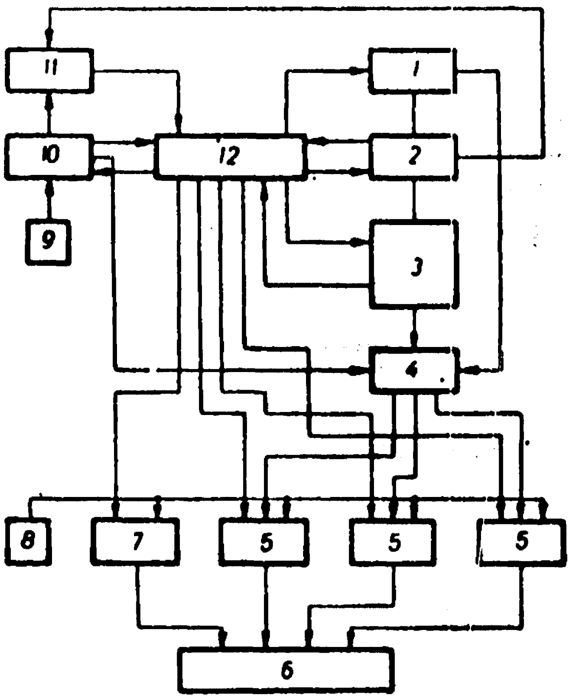


Figure 15. Interpolator ef the fourth degree: 1) counting device; 2) memory bleck; 3) summing device; [4) deceder;] 5) converters; 6) magnetic tape recording device; 7) pedestal frequency cenverter; 8) rectangular pulse generator; 9) generator; 10) circuit velecity bleck; 11) end-interpolatien block; 12) contrel bleck.

cording device. A pulse generator 9, which is contrelled by a circuit velocity bleck 10, serves as a timing element. A rectangular pulse generater 8 emits clearing signals fer the recording converters 5. The recording bleck is contrelled from the pedestal frequency cenverter 7. The initial data and intermediate results are recorded in memory device 2. With suitable initial setups in summing device 3, the circuit reproduces the function

$$x = f\ (y)$$

where x is a solution of the equation

$$ax^4 + bx^3 + cx^2 + dx - y = 0$$

The results of solution are fed through decoder 4 and converters 5 to the recording block onto magnetic tape.

Central controlling device 12 and end-interpolation block 11 control the operation of the circuit.

Automatic extrapolation

Automation of the solution of extrapolation problems is achieved by means of specialized computing devices, extrapolators.

If the input of these devices is fed some function, we obtain its anticipated values at the output. Here both the input and output signals may be both continuous functions and discrete sequences.

Discrete and continuous extrapolators

Let us consider some extrapolator circuits [18]. Let us assume that we need to find the value of the function $x(t)$ at the point t_4 from the known values of $x(t)$ at the points t_1, t_2, t_3 (see Fig.16).

Through the known points, we pass a second-order curve

$$x(t) = at^2 + bt + c.$$

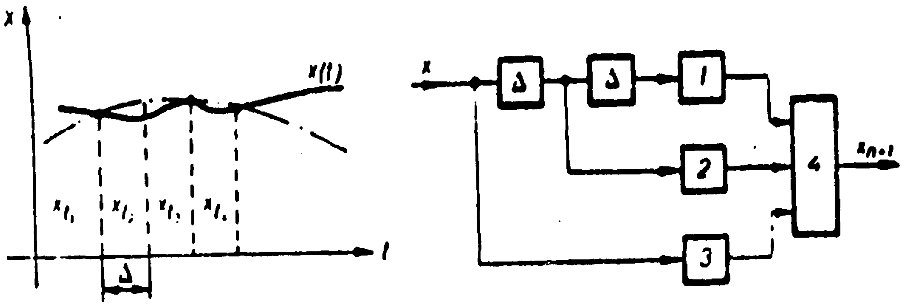


Figure 16. Extrapolation from three points: Δ) delay; 1,2,3) multiplication block; 4) summing device.

When t=0, t= $-\Delta$, t= -2Δ, we obtain x_{t_3} = c,

$$x_{t_2} = x_{t_3} - b\Delta + a\Delta^2,$$
$$x_{t_1} = x_{t_3} - 2b\Delta + 4a\Delta^2,$$

where Δ is the time quantization step.

From these equations we find an approximating polynomial in the form

$$x = x_{t_3} + \frac{1}{2\Delta}(x_{t_1} - 4x_{t_2} + 3x_{t_3})t + \\ + \frac{1}{2\Delta^2}(x_{t_1} - x_{t_2} + x_{t_3})t^2,$$

or

$$x_{t_4} = 3x_{t_1} - 3x_{t_2} + x_{t_3}.$$

The anticipated value of the signal is equal to the sum of the preceding values, which are separated from one another by the interval Δ, multiplied by the corresponding weighting factors. Figure 16 shows the block diagram of the extrapolator.

If we need to extrapolate a signal for an anticipation time τ, the anticipated values are computed in the form of a sum

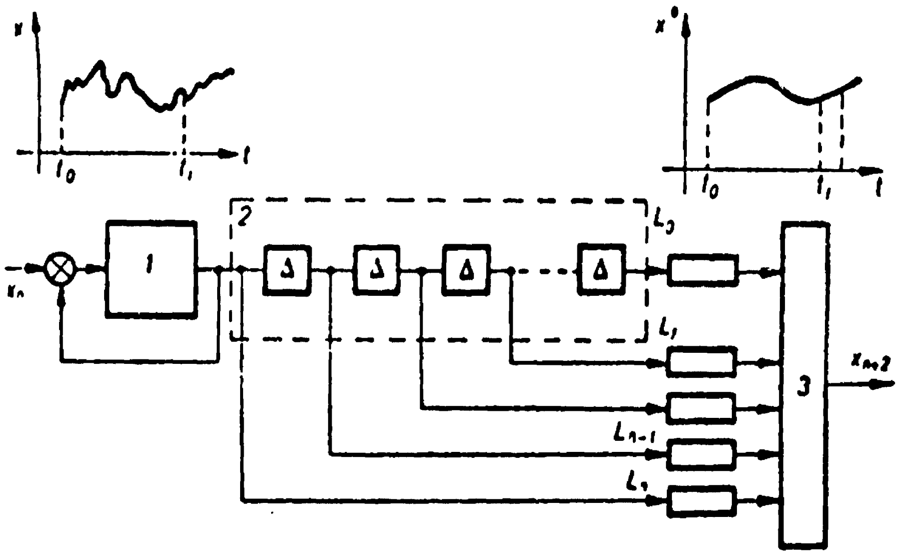


Figure 17. Continuous-signal extrapelator: Δ) delay; 1) filter; 2) bleck for emitting delayed values; 3) summing device.

$$x_{N+1}=\sum_{0}^{N} x_i r_i, \qquad (14)$$

where x_i is the value of the signal at the i-th point; r_i is the weighting factor of the i-th term. But let us point out at once that this formula is the first member of Kelmogorev's extended prediction operator, which will be considered in detail in the succeeding chapters. The block diagram of an extrapelator based en formula (14) is shown in Fig. 17.

A continuous signal is fed te a memory device, where it is divided into n=T/Δ equidistant values. The signal from each memory cell, multiplied by its weighting factor r, is fed to a summing device. Since the input signal

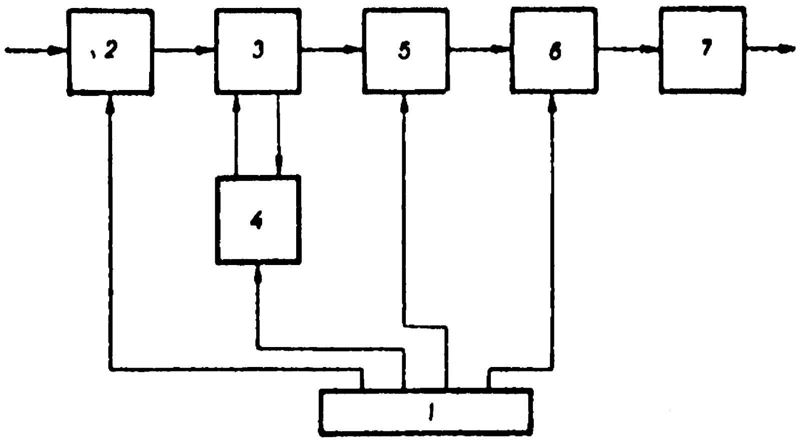


Figure 18. Discrete extrapolator: 1) control block; 2) converter; 3) shifting register; 4) weighting-factor block; 5) address formation circuit; 6) memory device; 7) summing device.

varies continuously, we obtain its continuous anticipated value at the output of the summing device. A similar extrapolator may either be in the form of an analogue computer or may use digital elements (Fig. 18).

Converter 2 of the discrete extrapolator converts the values of the continuous input signal into digital form. By means of shifting register 3 and weighting-factor block 4, we can, in accordance with (14), sequentially multiply the discrete values of the function to be extrapolated by the factors r_i.

Address forming circuit 5 provides for recording the products $x_i r_i$ on definite cells of memory device 6. The products $x_i r_i$ are then summed in summing device 7.

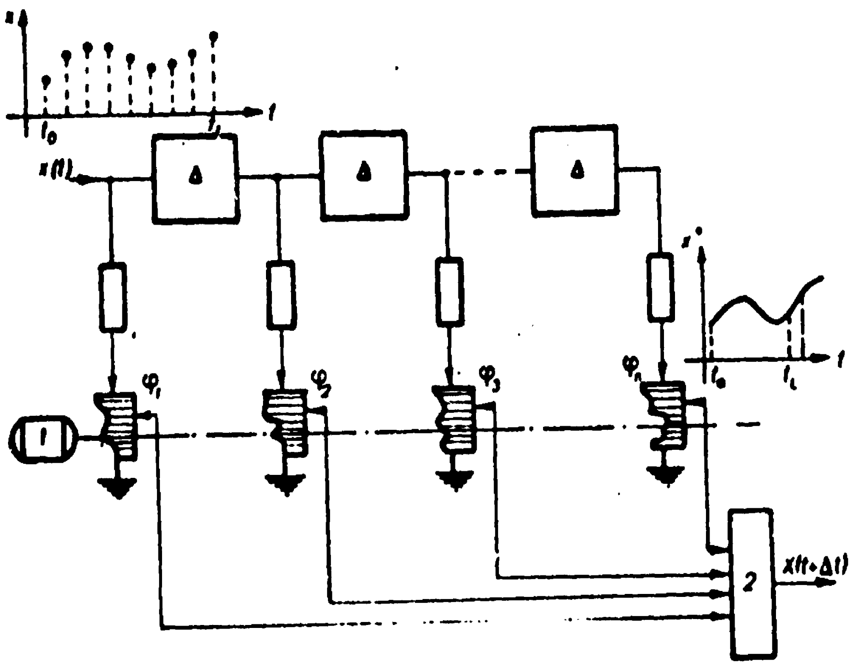


Figure 19. Extrapelator for obtaining a centimuous anticipated signal from discrete input data: Δ) delay; 1) moter; 2) summing device.

Extrapolator for centimueus anticipatien

Often the values of the quantity being measured can only be obtained at discrete times, and it is needed to know the probable value ef the signal not only at some future moment t+Δt, but alse te have the continueus value ef this signal in the interval [t,t+Δ t] . Using known mathematical methods, we can seek the law of variation ef the weighting factors during one discreteness interval.

The circuit of a device which performs the task of extrapolation in this form is shown in Fig.19 [18].

The discrete signal is fed to the memory cells with the following signal shifted with respect to the preceding by Δ . The remembered signals are continuously fed in the form of a voltage constant in the interval Δ to scaling potentiometers.

The resistance of these potentiometers varies according to the law $\phi(t+k\Delta)$, where k is an integer. The signal from the potentiometers, multiplied by r_i, is fed to the input of the summing device. At the output we obtain a continuous smoothed extrapolated signal.

Thus, with continuous extrapolation from known discrete values it is necessary that the weighting factors be functions of time, $r_i(t)$.

Invariance conditions and the synthesis of interpolators and extrapolators

The use of extrapolators and interpolators in control systems requires that their designers fulfill a whole series of special requirements .

Depending on the concrete problems, these devices must provide assigned accuracy, have a definite speed of action, and be reliable and as simple as possible. Engineers are aided by theoretical methods. Great success in solving these problems has been achieved thanks to the use of the theory of invariance.

Invariance conditions

Figure 20 shows the circuit of an open pulse servo-

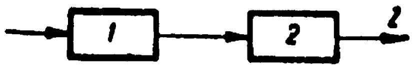


Figure 20. Open pulse servomechanism: 1) pulse moment; 2) continuous part.

mechanism. The absolute invariance condition for it is the condition that the input and output signals be equal at any time.

Let us write the mathematical expression for the transfer function of the open system

$$Z^*(q, \varepsilon) = K^*(q, \varepsilon)\, X^*(q). \qquad (15)$$

Let x(t) denote the input signal, K(p) the transfer function of the indicated continuous part of the system, and Z(t) the output signal.

Making the usual change of variables

$$\bar{t} = \frac{t}{T}, \quad p = \frac{q}{T}, \quad t = n + \varepsilon \; (n = 0, 1, 2, \ldots; \; 0 < \varepsilon \leqslant 1),$$

let us write the components of expression (15):

$$\begin{aligned} Z^*(q, \varepsilon) &= \mathrm{D}\,\{z(t)\} = \mathrm{D}\,\{z[n, \varepsilon]\} = \mathrm{D}\,\{Z(q)\}, \\ K^*(q, \varepsilon) &= \mathrm{D}\,\{k(\bar{t})\} = \mathrm{D}\,\{k[n, \varepsilon]\} = \mathrm{D}\,\{K(q)\}, \qquad (16) \\ X^*(q) &= \mathrm{D}\,\{x(\bar{t})\}_{\bar{t}=1} = \mathrm{D}\,\{x[n]\} = \mathrm{D}\,\{X(q)\}. \end{aligned}$$

To the condition that the input signal be equal to the output signal at any time

$$z[n, \varepsilon] = x[n, \varepsilon] \qquad (17)$$

let us apply the D-transformation.

We obtain

$$Z^*(q, \varepsilon) = X^*(q, \varepsilon). \qquad (18)$$

Taking (16) into account, we find the condition of abselute invariance for the open pulse system:

$$K^*(q, \varepsilon) = \frac{X^*(q, \varepsilon)}{X^*(q)}. \qquad (19)$$

In a number of problems with exact reproduction of the form of the input signal, a lag in the output signal is allowed.

In these cases, the invariance condition is written in the form

$$z(t) = x(t - \alpha), \qquad (20)$$

where α is the delay time, or the shift between the input signal x(t) and the eutput signal z(t).

Circuits based on invariance conditions

Let the input signal be known in advance. However, the information arriving at the input of the system is the values of the input signal at discrete instants. The problem of constructing interpolators and extrapolaters for such cases can be reduced te the construction of pulse systems invariant in the sense of (17)-(20).

Methods for synthesizing such systems can be found in more detail in Yu.V.Krementulo [27,28].

Let us consider here only one discrete-continuous system with an interpolator (Fig.21). The values of the signal at times t= nT arrive at the input of the interpolator. At the output we obtain a continuous function

$$x_1(\bar{t}) = x(t) + \Delta x(t),$$

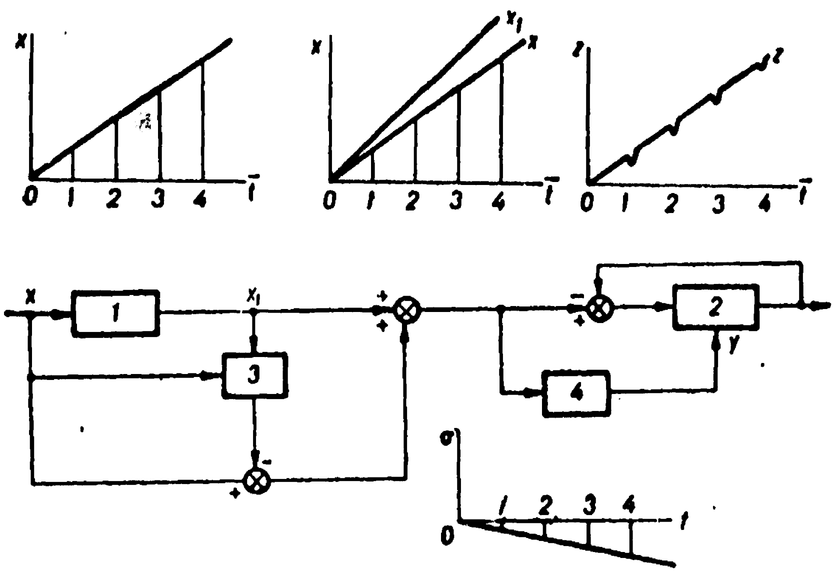


Figure 21. Discrete-continuous system with interpolator: 1) interpolator; 2) continous part; 3) comparison circuit; 4) input signal compounding connection; 5) compounding connection signal.

where $x(\bar{t})$ is a continous input signal, $\Delta x(t)$ the error due to inaccurate operation of the interpolator. By comparing signal $x_1(\bar{t})$ with $x(\bar{t})$ at time $\bar{t}=n$, we can decrease the interpolation error. The correction obtained by comparison is summed with the output signal of the interpolator.It is obvious that, when this method is used, correction occurrs only at discrete instants.

Figure 22a shows an improved circuit. This circuit uses an auxiliary memory device, also called an accumulating filter [20].

The correction obtained by comparing signals $x(\bar{t})$ and $x_1(t)$ at time $t=n$ is stored in the memory device in the interval $n < \bar{t} \leq n+1$ and is summed with the output signal

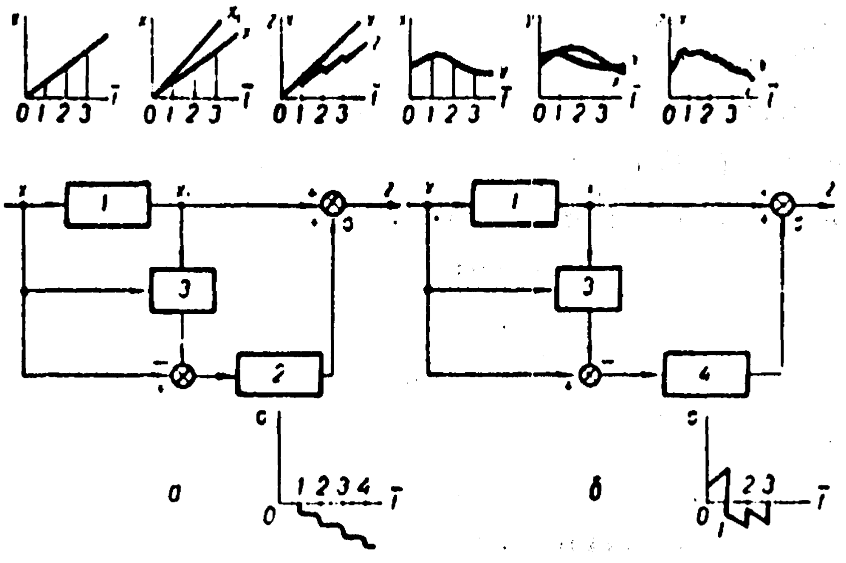


Figure 22. Systems with interpolator: a) with accumulating filter; b) rough-fine system; 1) rough interpolator; 2) accumulating filter; 3) comparison circuit; 4) fine interpolator

of the rough interpolattor. In the following interval $n+1 < \tau < n+2$, to the output signal there is added the correction obtahed in comparing $x[n+1]$ with $x_1[n+1]$, etc.

If, instead of a memory device, we use a second interpolator, we obtain a more general circuit (Fig.22b), which is marked by the presence of rough and fine interpolators.

The principle of combining rough and fine systems is widely used in technology. As examples we may point out discrete-continuous measuring and computing devices, and servomechanisms with rough and fine reading of angles of rotation. The use of rough-fine interpolators and extrapolators makes it possible in many cases to raise the accuracy of operation of a system and to simplify the circuit.

The "rough interpolator" (RI) is synthesized according to the invariance conditions. Small deviations in the output signal of the RI, $\Delta x(t)$, are decreased by the "fine interpolator" (FI). The FI converts the difference between $x[n]$ and $x_1[n]$ into the correction $\sigma[n,\varepsilon]$, whose law of change in the interval $n < t \leqslant n+1$ is determined by the interpolation law of the FI. In each concrete case this law is chosen from the conditions: 1) accuracy of operation of the RI; 2) accuracy of the whole system; 3) simplicity of design, etc.

In this chapter we have considered the problems of interpolation and extrapolation and have acquainted ourselves with the devices which make it possible to automate the solution of these problems. Here we have proceeded from the assumption that the processes are determinate and can be described by some analytical functions. The positive results obtained taking this assumption into account confirm the fact that the methods and devices used can be considered acceptable.

But if we observe large deviations of the predicted values from the actual ones,this indicates the presence in the processes of random components. In cases where the random factors exert great influence on the course of the processes, determinate prediction is invalid.

Then we can use the theory of probability, the theory of random processes, and mathematical statistics. On the basis of these disciplines, the past two decades have seen the creation of new methods united under the general name of the statistical theory of prediction.

Chapter 2

Prediction of stationary random processes

Topics in Brief from the Theory of Probability and the Thoory of Random Functions

Random events. Random variables. Random processes

The results of experiments, multiple measurements, the turning up of a number of pips on a die, and the falling of a projectile at some distance from the target are all characterized by an incenstant outcome.

In probability theory the various possible outcomes of trials are called random events. Every trial is determined by one or several variable quantities. If, as the result of a trial, these variable quantities can assume various values, these variables are called random variables. Let us assume that we choose at random some part frem a large batch of parts of one type. The dimensions of the chosen part are random variables. Since the results of investigations and measurements are ordinarily expressed in numbers, random variables can assume various numerical values. If random variables assume values which are separate and iselated from one another and which can be enumerated, these random variables are called discrete. Random variables which continuously vary as a function of some parameter ana whose values cannot be enumerated in advance are called continuous.

The classical theory of probability deals with "mass" random phenomena. A mass phenomenon is an aggregate of multiple repetitions of a phenomenon or actions "haphazardly" considered as a whole or without taking into account the chronological sequence.

As distinct from the classicat theory of probability, the theory of probabilistic, or random processes, developed in the main by A.N.Kolmogorov and A.Ya.Khinchin, operates with the processes and sequences (discrete processes) of random phenomena. Random processes and sequences are aggregates of random variables in the dynamics of their development. They are the same mass phenomena, but they are considered not in the form, for example, of a uniform ensemble of random numbers, but in the form of a sequence of numbers in the chronological order of appearance of the quantities to which they correspond.

Examples of random processes are changes in the coordinate of a brownian particle, fluctuations in electric circuits, vibrations of the units of a machine-tool during its operation, change in the temperature of a patient during the course of a disease, change in the bioelectric activity of the brain, etc.

Frequency and probability

Let us assume tnat in a group of 1000 people, there are people whose height is less than 165 cm. We carry out a series of trials. A trial consists of the measurement of the height of each person. As a result, it turns out that there are 250 people in the group whose height is less than 165 cm. We say that the frequency of appearance of a person less than 165 cm in height in the group of 1000 people is

$$w = \frac{250}{1000} = 0.25$$

In the overwhelming majority of cases, when a trial is repeated many time, the frequency of appearance of an event A in a series of N trials acquires a stability. It very seldom essentially deviates from some positive constant number.

This positive number, less than one, is the quantitative expression of the possibility of random event A and is called its probability.

The probability, usually expressed by the symbol P(A), is, as it were, a physical constant connected with the random event A. The frequencies of this event in various concrete series of trials are random manifestations of this constant characteristic, which expresses a completely definite objective connection between a complex of conditions and the random event. The value of the probability changes as soon as the basic complex of conditions does.

The integral distribution function of a random variable

Let Φ (t) denote the probability that the random variable x will assume a value less than t. Φ (t) is called the integral distribution functions of the variable x. Since any probability must lie in the interval between 0 and 1, for all values of t we have

$$0 \leqslant \Phi(t) \leqslant 1.$$

Let $t_2 > t_1$. Then the probability that $x < t_2$ will be greater than or equal to the probability that $x < t_1$, i.e. the function Φ (t) cannot decrease with an increase in t. Figure 23 shows a typical form for the integral distribution function.

If the random variable x is a result of the measurement of some characteristic of an object selected at random from among N objects, Φ (t) in practice determines the relative portion of the objects for which $x < t$. The group

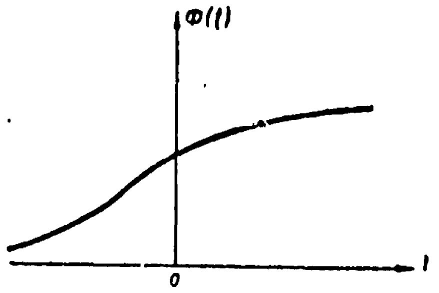


Figure 23. Integral distribution function.

of N objects is usually called the general aggregate.

Probability density function. Normal distribution law

Let Φ (t) be the integral distribution function of the random variable x. Then the probability that

$$t-\frac{\Delta}{2}\leqslant x<t+\frac{\Delta}{2}$$

(when $\Delta > 0$) is given by the difference

$$\Phi\left(t+\frac{\Delta}{2}\right)-\Phi\left(t-\frac{\Delta}{2}\right).$$

As $\Delta \to 0$, the limit of the ratio

$$\lim\frac{\Phi\left(t+\frac{\Delta}{2}\right)-\Phi\left(t-\frac{\Delta}{2}\right)}{\Delta}=f(t)$$

is called the probability density of the random variable x at the point x=t. The probability density f(t) is a function of t and is called the density function of the random variable.

If the random variable x is discrete, the integral distribution function is a step function,and the probability density function does not exist.

If we integrate the probability density function f(t) from t_1 to t_2 ($t_1 < t_2$), the integral

$$\int_{t_1}^{t_2} f(t)\,dt$$

will give the probability that x will assume a value between t_1 and t_2.

One of the most important probability density functions is the so-called normal probability density function (Fig.24). It is given by the expression

$$f(t) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \qquad (21)$$

where μ and σ are certain constants. We say that the random variable x is subject to the normal probability distribution if its probability density function is given by expression (21).

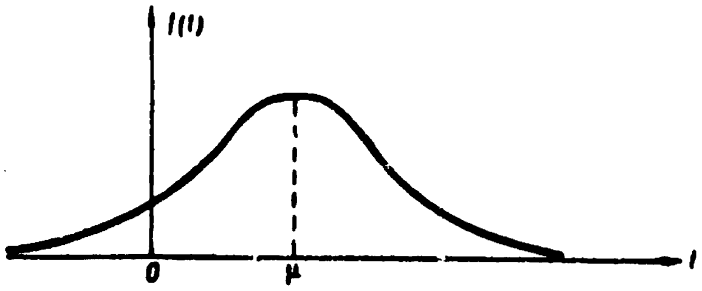


Figure 24. Normal probability density function.

Mathematical expectation and higher moments of a random variable

The mathematical expectation, or the mean value of a random variable, is the result of the probabilistic density function does not exist.

If we integrate the probability density function f(t) from t_1 to t_2 ($t_1 < t_2$), the integral

$$\int_{t_1}^{t_2} f(t)\,dt$$

will give the probability that x will assume a value between t_1 and t_2.

One of the most important probability density functions is the so-called normal probability density function (Fig.24). It is given by the expression

$$f(t) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \qquad (21)$$

where μ and σ are certain constants. We say that the random variable x is subject to the normal probability distribution if its probability density function is given by expression (21).

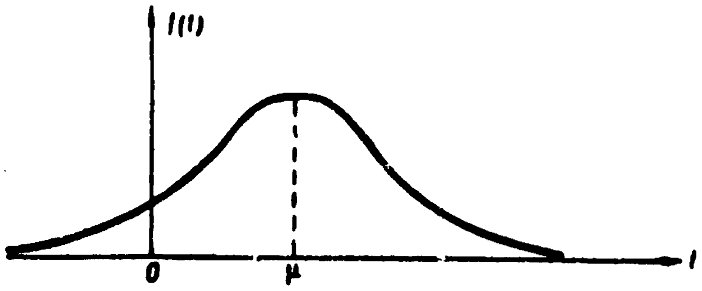


Figure 24. Normal probability density function.

Mathematical expectation and higher moments of a random variable

The mathematical expectation, or the mean value of a random variable, is the result of the probabilistic averaging of the possible values of this random variable. In this averaging, the probability of every possible value serves as a weight for this value.

In particular, the mathematical expectation (mean value) of a discrete random variable x whose possible values are N (finite) in number is equal to the sum of the products of each of these values by its probability

$$MX = \sum_{i=1}^{N} x_i P(x_i), \qquad (22)$$

where M is the sign of mathematical expectation.

A function ϕ (x) of a random variable is itself a random variable. The mathematical expectation of the function $(x-c)^k$, where k is any positive integer and c a constant, is called the kth-order moment of x with respect to c. Of especial interest is the case where c=MX. The mathematical expectation of the function $(x-MX)^k$ is called the kth-order moment of x with respect to the mean. The second-order moment with respect to the mean, i.e.

$$M(x-MX)^2 = DX, \qquad (23)$$

is called the dispersion.

The square root of the dispersion is called the standard deviation, or the mean square deviation.

For example, in the normal probability distribution function cited above

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

the mathematical expectation of the random varible x is equal to μ, and the dispersion to σ^2.

Conditional frequency. Conditional probability. Dependent and independent events

Often from a whole series of trial it is necessary to separate those as a result of which some event B has appeared and then afterwards to determine the frequency of the event in which we are interested.

If the frequency of an event A is computed not for all trials, but only for that sequence of trials as a result of which event B appeared, this frequency is called the conditional frequency of A with respect to B.

Conditional frequencies possess all the properties of frequencies, including the property of stabilization when the number of trials is increased without bound. As an objective quantitative characteristic of event A in its interrelations with event B, we can introduce the concept of conditional probability in a manner analogous to the way in which we introduced the concept of the probability of event A.

The conditional probability of event A with respect to B is the ratio of the probability that A and B will occur together to the probability that B will occur

$$P\ (A/B) = \frac{P\ (AB)}{P(B)} \qquad (24)$$

If the conditional probability of A with respect to B is not equal to the probability of A, event A is called dependent on B. But if the conditional prebability of A with respect to B is equal to the probability of A, event A is called independent of B.

Example [14]. A mechanical shop has produced 100 cylinders. Of them, 15 are elliptical, 50 are conical, 25 are simultaneously elliptical and conical, and 10 cylinders have no defects.

Event E consists of the fact that a cylinder taken at random will be elliptical, and event K of the fact that the cylinder will be conical:

$$P(E) = 15 + 25 / 100 = 0.4$$

$$P(K) = 50 + 25 / 100 = 0.75$$

$$P(EK) = 25 / 100 = 0.25$$

Let us assume that a randomly chosen cylinder is conical. But it may also be elliptical. Let us compute the probability that our randomly chosen cylinder with one defect also has the other.

From (24) we obtain:

$$P(E|K) = P(EK) / P(K) = 0.25 / 0.75 \approx 0.33$$

Analogously

$$P(K|E) = P(EK) / P(E) = 0.25 / 0.4 = 0.625$$

It is obvious that in this case $P(E|K) \neq P(E)$ and $P(K|E) \neq P(K)$. Hence E and K are dependent.

Basic concepts and definitions of the theory of random functions

Random functions. Distribution laws. Markov processes

A random function is a function whose value for every value of the argument (or several arguments) is a random variable. A function obtained as a result of one experiment is called a realization of a random function.

Random functions of time are usually called random, or stochastic, processes.

For every given value of the argument t, the value of a random function X(t) is an ordinary scalar random variable. The distribution law of a random function is a complete probabilistic characteristic of its value.

The one-dimensional distribution law of a random function X(t) depends on t as a parameter and can be given by the one-dimensional probability density f_1 (x,t).

The two-dimensional distribution law of a random function is the name given to the joint distribution law of its values $X(t_1)$ and $X(t_2)$ for two arbitrarily chosen values t_1 and t_2 of the argument t. In the general case, the name n-dimensional distribution law of the random function X(t) is given to the distribution law of the aggregate of its values $X(t_1)$, ..., $X(t_h)$ for n arbitrarily chosen values t_1, ..., t_h of the argument t.

An example of random functions which are exhaustively characterized by two-dimensional distribution laws are Markov random processes.

A markovian random process, or random process without aftereffect, is the name given to a random function with parameter t whose values when $t_1 < t_2 < \dots < t_h$ for any n form a simple Markov chain [15] In accordance with the definition of a simple Markov chain, the conditional distribution law of the value $X(t_{h+1})$ of the random function at a future instant depends only on the value of the ran-variable $X(t_h)$ at the present moment and does not depend on the values of the random variables $X(t_1)$, ..., $X(t_{h-1})$ at past instants.

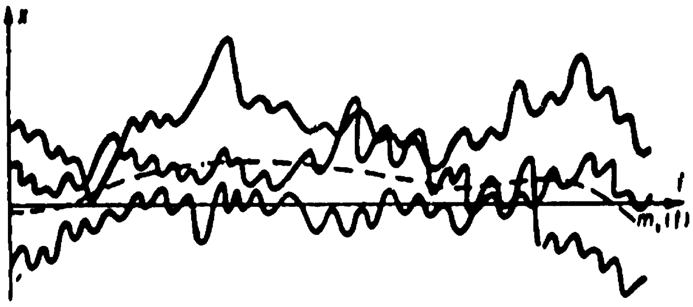


Figure 25. Random function.

Mathematical expectation and correlation function of a random function. Mutual correlation function

The mathematical expectation of the random function X(t) is the name given to the function $m_x(t)$ whose value for every given value of the argument t is equal te the mathematical expectation of of the value of the random function at the same t:

$$m_x(t) = M[X(t)].$$

This is a certain mean function, around which group and with respect to which oscillate all possible realizations of the random function (Fig.25).

The dispersion is taken as a measure ef the scatter ef a random function. This is a function whose value for every given value of the argument is equal to the dispersion of the value of the random variable fer this value of the argument.

In erder to take into account the influence of the values of the random function on each other for various values of the argument, besides the dispersion, the correlation moments of the values of the random function are given corresponding to all possible pairs of arguments.

The correlation moment of the values x(t) and x(t') of a random function X(t) is a function of the two independent variables t and t':

$$K_x(t, t') = M[X^{\circ}(t) X^{\circ}(t')]. \qquad (26)$$

This function is usually called the correlation (or autocorrelation) function of the random function X(t). $X^0(t)$ denotes the deviation of the random function X(t) from its mathematical expectation (centered random function).

The mutual correlation function, or correlation function of the connection of two random functions X(t) and Y(s) is the name given to the correlation moment of the values of these functions for arbitrarily chosen values of their arguments t and s:

$$K_{xy}(t, s) = M[X^{\circ}(t) Y^{\circ}(s)]. \qquad (27)$$

Random functions are called correlated if their mutual correlation function is not identically zero. But if the mutual correlation function of two random functions is identically zero, these random functions are called uncorrelated.

Stationary random functions. Ergodic property of a stationary random function

The random function X(t) is called stationary in the broad sense if its mathematical expectation is constant and its correlation function depends only on the difference between the arguments t and t':

$$m_x(t) = M[X(t)] = \text{const},$$
$$K_x(t, t') = k_x(\tau), \qquad (28)$$

where $\tau = t-t'$.

It follows from the definition that the correlation function of a stationary random function of one variable is a function of the one variable τ.

The dispersion of the stationary random function X(t) is

$$D[X(t)] = K_x(t, t) = k_x(0). \qquad (29)$$

The dispersion of a stationary random function is constant and is equal to the value of the correlation function at the origin.

An important class of stationary random functions is made up by the ergodic stationary random functions.

A stationary random function X(t) is ergodic if the absolute value of its correlation function $k_x(\tau)$ decreases without bound as $[\tau] \rightarrow \infty$, i.e. if for any $\varepsilon > 0$, we can find a quantity T_0 such that

$$|k_x(\tau)| < \varepsilon \text{ when } |\tau| > T_0. \qquad (30)$$

The random processes described by stationary ergodic random functions retain constant statistical parameters in realizations of any length, no matter how large.

We have become acquainted with the basic concepts and definitions of the theory of probability and the theory of random functions. In what follows, we shall operate with these concepts and definitions when we consider various prediction problems.

Prediction Quality Criterion. Optimality Criterion

Criterion of minimum mean-square error

The problem of interpolation and extrapolation can also be formulated for random processes and sequences.

However, instead of speaking about finding the values of a function within or outside of the segment of observation, we should rather speak about finding some function which can be determined from the initial values of a random process from the point of view of satisfying some general optimality criterion.

In most cases, such a criterion is the achievement of a minimum mean-square error of deviation of the unknown approximating function from the initial random function or from some point set (selection) whose points represent the latter function.

In the general case, there is a minimum mean-square error when

$$f^* = \int_{-\infty}^{\infty} f p(f/f_1, f_2, \ldots, f_N)\, df. \qquad (31)$$

In (31), f^* is the function being predicted; f_1, f_2, ..., f_N is a selection of the preceding values of the function; $p(f/f_1, f_2, \ldots f_N)$ is the conditional probability of obtaining f with $f_1, f_2, \ldots, f_N$.

A remarkable property of the criterion of the minimum mean-square error is the fact that this criterion gives a unique solution to the problem. Indeed, the equation

$$M(f - f^*)^2 = \overline{\varepsilon^2} \qquad (32)$$

describes a multidimensional paraboloid and, consequently, with any way of varying the parameters of the predicting model (of the mathematical operator or predicting device), achievement of a minimum is unavoidable. At the beginning of the fourth chapter we shall deal in more detail with devices based on this criterion.

Let us consider some more prediction optimality criteria.

Criterion of the minimum sum of the integral square error and the dispersion of the random error

If the function being predicted is the sum of a regular component (useful signal) and a random stationary component (noise)

$$f(t)=S(t)+N(t), \qquad (33)$$

we can formulate the optimality criterion as follews.

Letus write the integral square error in the form

$$\int_0^\infty \varepsilon^2 dt, \qquad (34)$$

where $$\varepsilon=S(t+\Delta t)-S^*(t+\Delta t). \qquad (35)$$

In expression (35), $S^*(t+\Delta t)$ is the regular component of the predicted function.

The predicted function can be written in the form

$$f^*=S^*(t+\Delta t)+N^*(t+\Delta t). \qquad (36)$$

Let us denote the dispersion of the random error by DN^*.

The prediction will be called optimal if a minimum is ensured for the sum with weight a of the integral square error (34) and with weight c of the dispersion of the random error while this condition is fulfilled:

$$\lim_{t\to\infty}\varepsilon(t)=0. \qquad (37)$$

Proceeding from this, we can formulate the problem of analytical design of an anticipator [16]. It is required to find functions $\varepsilon(t)$, $S^*(t)$ and some weighting function $\phi(t)$ which satisfy (35), (37) and which minimize the functional

$$I = a\int_0^\infty \varepsilon^2 dt + cDN^*. \qquad (38)$$

The criterion just considered can be expediently used for designing an optimal system for filtering and predicting processes described by expression (33).

Arbitrary optimality criteria

Let us consider a stationary random process which is gaussian in the broad sense of the word. This means that the signal is described by the expression

$$S(t) = m(t) + \sum_{i=1}^{N} r_i f_i(t), \qquad (39)$$

where $m(t)$ is a normal random signal; $\sum_{i=1}^{N} r_i f_i(t) \cdot$ is a linear combination of N known functions f_i with coefficients r_i; the coefficients r_i are random functions which also have a normal distribution.

A noise $N(t)$ having a normal disttribution is additively superimposed on the signal.

Thus, we are again dealing with prediction of a process of the form

$$f(t) = S(t) + N(t).$$

Optimal prediction is achieved with minimization of some error criterion (value function) $C(\varepsilon)$. In the case of the mean-square error criterion $C(\varepsilon) = \overline{\varepsilon}^2$, the optimal nonlinear prediction is equivalent to the optimal linear prediction in the Wiener sense [62] with the same error criterion. A similar result is obtained in the case of even, $C(-\varepsilon) = C(\varepsilon)$, nondecreasing criteria. The conclusion is also valid for asymmetrical nondecreasing error criteria.

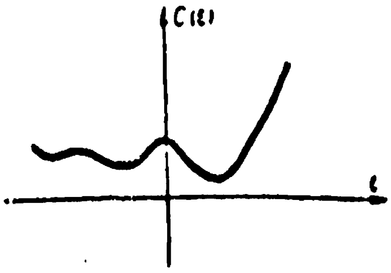


Figure 26. Graph of error-criterion variation.

Let an arbitrary error criterion be given, fer example in the form of a graph (Fig.26).

Pugachev's general theorem [39] states that in this case an optimal nonlinear system (in particular, a system which achieves optimal prediction) is an eptimal linear system in the Wiener sense with weighting function w(t, τ) with the mean-square error criterion, to which is sometimes added some censtant "bias" μ :

$$O[f(t)] = \int_{-\infty}^{t} w(t, \tau) f(\tau) d\tau + \mu. \qquad (40)$$

If the error criterion is an even function, the constant bias is always equal to zero.

Let us cite an algerithm for synthesizing a nonlinear eptimal system which minimises the mathematical expectation of an arbitrary error criterion.

1. We seek a linear optimal (in the Wiener sense) system with weighting function w(t, τ). As a criterion we here take the minimum mean-square error.

2. We will determine the mean-square errer by the usual methods.

3. We will determine the mathemtical expectatien and dispersion of this errer.

4. We will find the constant bias by minimizing the integral

$$E[C(\varepsilon)] = \int_{-\infty}^{\infty} C(\varepsilon) f(\varepsilon)\, d\varepsilon, \qquad (41)$$

where f(ε) is the probability density of the error of the optimal linear system with unknown mathematical expectation m'.

Let us put

$$\frac{\partial}{\partial m'} \int_{-\infty}^{\infty} C(\varepsilon)^{-\frac{(\varepsilon - m')^2}{2\sigma^2}}\, d\varepsilon = 0 \qquad (42)$$

and let us solve this equation for m'. Then

$$\mu = m - m'. \qquad (43)$$

The eptimal nonlinear system is described by the equation

$$O[f(t)] = \int_{-\infty}^{\infty} w(t, \tau) f(\tau)\, d\tau + \mu. \qquad (44)$$

The only limitation in Pugachev's theorem is the condition that the random component ef the signal and the random noise have normal distribution. However, in solving actual problems by the method of modelling en analog computers, good results have also been obtained when this condition is not fulfilled [60].

The optimal, most exact prediction is achieved when the initial data (keys) have normal or gaussian distribution.

It is clear that even with the best, optimal prediction, we cannot count on the exact coincidence of these functions: a random process always has some unpredictable element of "pure" randomness. The prediction accuracy can be estimated from the variation

$$\delta = \frac{\overline{(f_i - f_i^*)^2}}{\overline{f_i^2} - \bar{f}_i^2} 100\%.$$

But even with optimal prediction of random processes, the variation is not equal to zero. Only in predicting non-random determinate processes, the motion of the heavenly bodies for example, can the variation be zero, i.e. the function being predicted exactly corresponds to the actual process (with computational error)

$$\sum_1^N (f_i - f_i^*) = 0; \quad \delta = 0.$$

Besides the above criteria, attempts have been made of late to use various game criteria, for example minimax criteria, for synthesizing optimal systems. Especially effective is the use of such criteria in optimizing "large systems" /37/.

Below we consider a number of problems in the prediction of stationary random process and sequences. In all cases, our optimality criterion will be the minimum mean-square error one.

Prediction of Stationary Random Sequences

The method and formula of A.N.Kolmogorov

For stationary random processes whose values are known at discrete instants, the problem of extrapolation has been formulated as follows /26a/.

Let f(t) be a real random variable corresponding to every integral t in the interval $-\infty < t < \infty$.

If the mathematical expectation

$$m = M[f(t)] = \text{const}$$

and the correlation function

$$K_{\Delta t} = M[(f(t+\Delta t) - m)(f(t) - m)]$$

do not depend on t, f(t) is stationary. Without limiting the generality we can put

$$m = M[f(t)] = 0. \qquad (45)$$

Then

$$K_{\Delta t} = M[f(t+\Delta t) \cdot f(t)]. \qquad (46)$$

The problem of linearly extrapolating a stationary sequence which satisfies (45) consists of selecting for given $n > 0$ and $\Delta t \geqslant 0$ real coefficients r_i for which the linear combination

$$O[f(t)] = r_1 f(t-1) + r_2 f(t-2) + \dots + r_n f(t-n) \quad (47)$$

of random variables f(t-1), f(t-2),..., is an exact as possible approximation to tho random variable $f(t+\Delta t)$. As a measure of the accuracy of such an approximation we take the criterion

$$\overline{\varepsilon_e^2} = M(f(t+\Delta t) - O[f(t)])^2.$$

If we know several moments $K_{\Delta t}$, we can easily solve the problem of finding r_i for which $\overline{\varepsilon_E^2} = \overline{\varepsilon_E^2}_{min}$.

The problem of interpolation consists of estimating f(t) from the values of f(t+1), f(t+2), ..., f(t+n), f(t-1), ..., f(t-n).

Here, as another measure of accuracy we can take the criterion

$$\overline{\varepsilon_I^2} = M(f(t) - Q[f(t)])^2,$$

where

$$Q[f(t)] = r_1 f(t+1) + r_2 f(t+2) + \ldots + r_{-1} f(t-1) + \ldots \quad (48)$$

with constant real coefficients. The problem of interpolation is reudced te the determination of $\overline{\varepsilon_I^2} = \overline{\varepsilon_I^2}_{\min}$. A proof of the existence of limits for $\overline{\varepsilon_E^2}$ and $\overline{\varepsilon_I^2}$, as well as a solution ef the problem of finding their values is given in [26a].

Let us consider an example of the solution of a problem of linear extrapolation ef a stationary random process.

Prediction of the change im the quality index of a product of the petroleum-chemistry industry

The prediction problem

Figure 27 shows a diagram for automatic regulation of a thermocracking installation at a petroleum-processing plant [31].

The end product is thermecracking benzine, one of whose quality indices is the temperature at the end of boiling. An automatic analyzer determines this index every half hour and records the value of T_{KK} [temperature at end of boiling] in °C by means of a recording instrument on a cartogram.

Using the data on the values of T_{KK}°C for a certain interval of time preceding the instant t, we are required to predict the values of T_{KK}°C at some future time $t+\Delta t$. In practice the interval Δt is equal to the time interval between analyses.

Let us rewrite the prediction operator (47) in the form

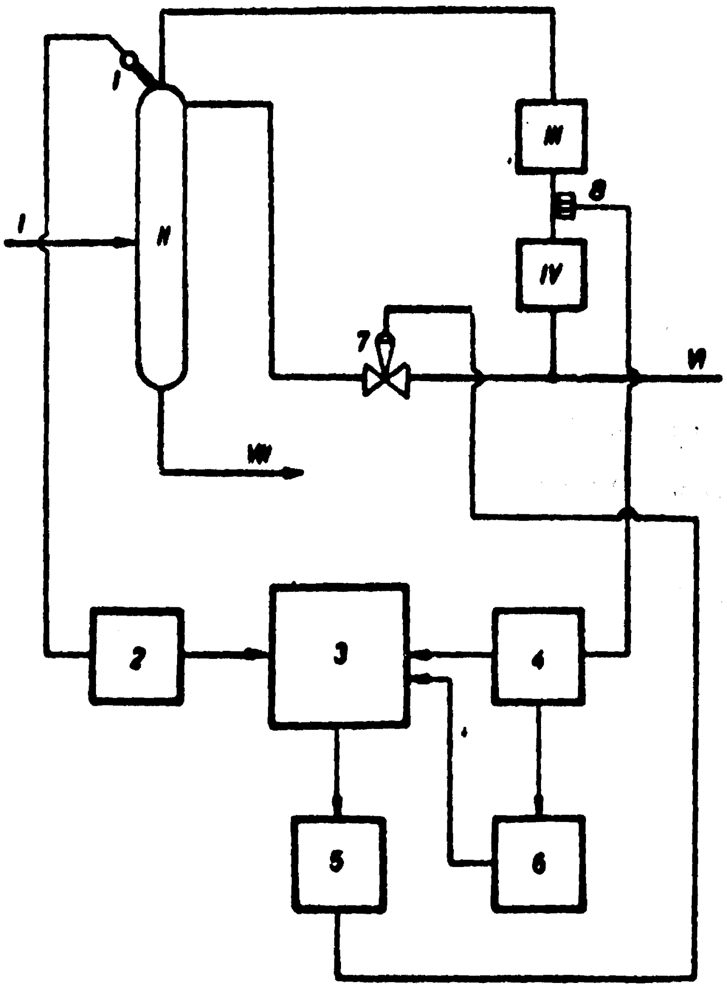


Figure 27. Diagram for regulation of thermocracking installation with quality analyzer; I) flow of raw material into column; II) fractionating column of thermocracking installation; III) cendenser; IV) feed tank; V) feed flow; VI) end product (thermocracking benzine); VII) debenzined product; 1) thermocouple; 2) potentiometer; 3) summing device; 4) quality analyzer; 5) regulator; 6) predicting filter; 7) regulating valve; 8) sampling point for analysis of petroleum product.

$$O[f(t)] = \sum_{i=1}^{N} f_i r_i = f^*(t + \Delta t). \qquad (49)$$

where f_i ($i=1,2,\dots N$) are the values of T_{KK}°C at the preceding instants, and $f^*(t+\Delta t)$ is the predicted value of T_{KK}°C.

The problem consists of finding coefficients r_i such that

$$\overline{\varepsilon^2} = M[f(t + \Delta t) - f^*(t + \Delta t)] = \overline{\varepsilon}^2_{\min}. \qquad (50)$$

The coefficients r_i for which condition (50) is satisfied are considered to be optimal, and we say of the operator (49) that it has learned to predict the future values of the given random sequence.

Learning algorithm for prediction operator

Let us write the logical scheme for the learning algorithm of the prediction operator

$$\overset{2}{\downarrow} ECc \overset{1}{\uparrow} S \overset{1}{\downarrow} R\omega \overset{2}{\uparrow}. \qquad (51)$$

In expression (51):

E is the operator for computing the mean-square error from the set of known values of the function (for various values of the coefficients);

C is the operator for comparing the computed error $\overline{\varepsilon^2}$ with the quantity chosen in advance $\overline{\varepsilon^2}_{\min}$ which satisfies the required accuracy;

c is a logical condition which is considered satisfied when

$$\overline{\varepsilon^2} < \overline{\varepsilon^2}_{\min};$$

is the operator for computing new coefficients;

ω is an identically false condition;

S is the operator for terminating the learning process.

The algorithm works as follows. The elements of the scheme operate one after another from left to right, beginning with the extreme left.If the following element of the scheme is a logical condition, two cases may arise. If the logical condition is satisfied, the next element of the scheme operates; but if it is not satisfied, there is a transition along the arrow. The symbol ↑ denotes the beginning of the arrow and ↓ the end [33].

First learning step. We determine the mean-square error of the representation of the function f(t) by the operator (49) on a known time interval with arbitrary values of the coefficients.

From the whole sequence we select the values $f_1, f_2, \ldots, f_k, f_{k+1}$ and we computo

$$(f^*_{\kappa+1} - f_{\kappa+1})^2;$$

$$f^*_{\kappa+1} = \sum_{i=1}^{\kappa} r_i f_i.$$

The sequenco f_1, $f_2, \ldots$ of length k we call the prehistory. Then we take the sequence $f_2, f_3, \ldots, f_{k+2}$ and compute $(f^*_{k+2} - f_{k+2})^2$, where

$$f^*_{\kappa+2} = \sum_{i=2}^{\kappa+1} r_i f_i$$

etc.

The mean-square error over the set of known values of the function f(t) at the first learning stop is given by the expression

$$\overline{\varepsilon_1^2} = \frac{\sum_{i=\kappa+1}^{N} (f_i^* - f_i)^2}{N - \kappa + 1}.$$

Comparing the obtained error $\overline{\varepsilon_1^2}$ with $\overline{\varepsilon^2_{min}}$, we evaluate the quality of representation of the function f(t) by operator (49) for selected values of the coefficients. If $\overline{\varepsilon_1^2} > \overline{\varepsilon^2_{min}}$, we carry out the second learning step.

We change the values of r_i in accordance with the minimization algorithm for the function $\overline{\varepsilon^2}(r_i)$.

We seek the value of the error at the second step $\overline{\varepsilon_2^2}$ and again compare the obtained value with $\overline{\varepsilon^2_{min}}$.

The learning process stops if, as a result of comparison of $\overline{\varepsilon_l^2}$, obtained at the l-th step, with $\overline{\varepsilon^2_{min}}$, we obtain

$$\overline{\varepsilon_l^2} \leqslant \overline{\varepsilon^2_{min}}.$$

Now the operator is considered to have learned te predictt the future values of the given function, and the coefficients used for computing operator (49) are optimal.

The following value of the function is predicted by means of realization of the algerithm fer computing operator (49) when $r_i = r_{ept}$.

Solution and results

The prediction problem was programmed and solved on a universal digital computer.

The function $\overline{\varepsilon^2}(r_i)$ was minimized by the method of steepest descent /4/. Table 1 shows the actual and predicted values of T_{KK} °C. Figure 28 shows graphs of the variation in the actual and predicted values of T_{KK} °C for a different number of peints of the prehistory (k=2,3,4,5). It is evident that the results of prediction depend on the length of the prehistory. This question will be considered in chapter 4, where we will solve the problems of prediction using an extended prediction operator.

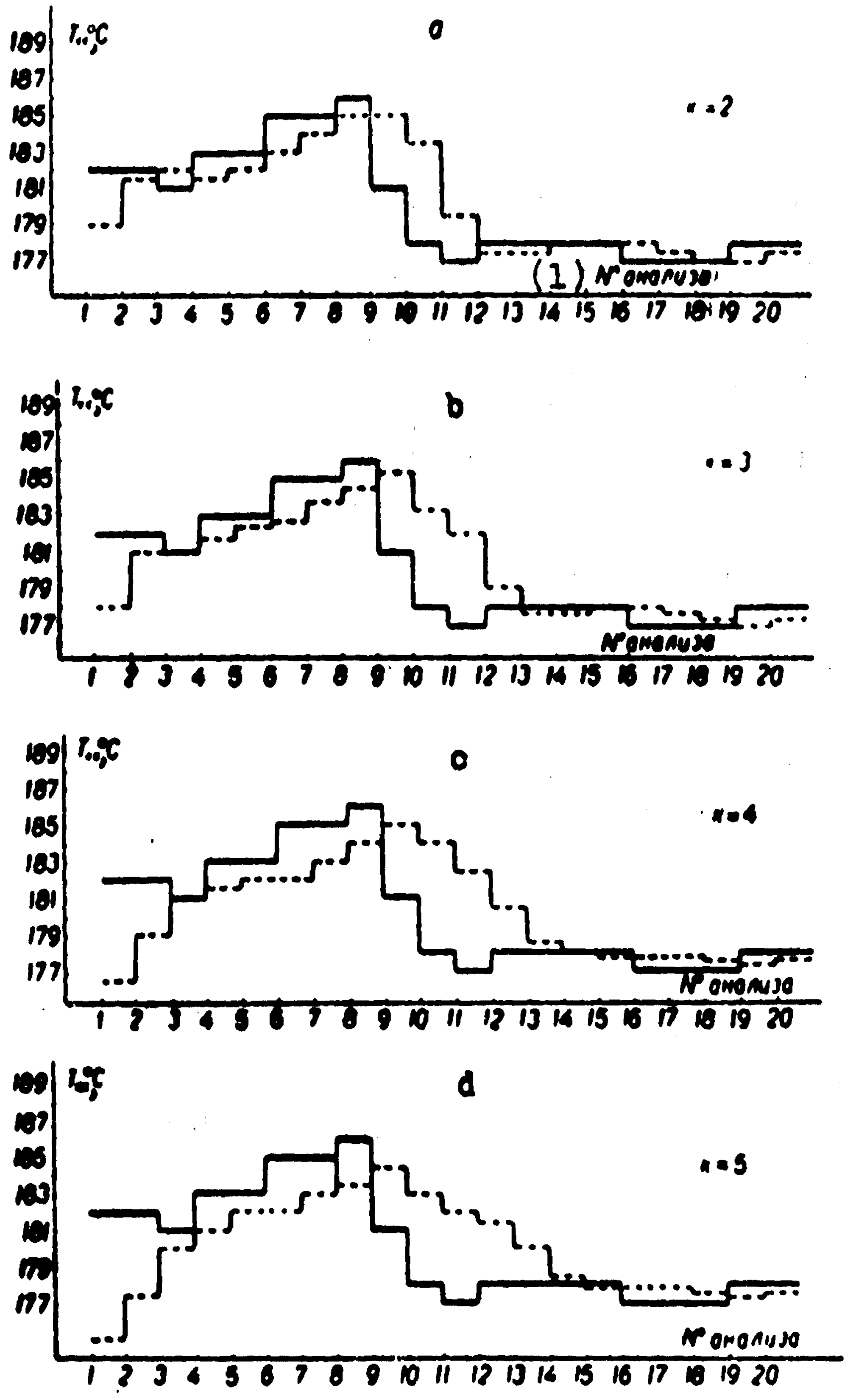


Figure 28. Prediction of the temperature at the end of boiling by the linear extrapelation method (k=2,3,4,5).

Key: 1) Number of analysis.

(1) Дейстгительные значения	(2) Предсказанные значения			
	к=2	к=3	к=4	к=5
182.	179	178	176,5	175
182	181	181	179	177,4
181	182 ·	181	181	180
183	181,5	181,6	181,5	181
183	182	182,3	182	182
185	183	182,6	182	182
185	184	183,6	183	183
186	185	184,6	184	183,4
181	185	185,3	185	184,4
178	183,5	183,3	184	183
177	179,5	182	182,5	182
178	177,5	179	180,5	181,4
178	177,5	177,6	178,5	180
178	178	177,6	178	178,4
178	178	178	177,75	177,8
177	178	178	177,75	177,8
177	177,5	177,6	177,75	177,8
177	177	177,3	177,5	177,6
178	177	177	177,25	177,4
178	177,5	177,3	177,5	177,6

Tablo 1.

Key: 1) Actual values; 2) predicted values.

Prediction by exponential smoothing formulas (Brown's method)

The theory of exponential smoothing [50,50a] has of late been greatly developed.

Exponential smoothing (Brown) is based on the assumption that the value to be predicted of some function f(t) can be expressed by a Taylor series:

$$f_{t+\Delta t} = f_t + \frac{df}{dt}\Delta t + \frac{1}{2!}\,\frac{d^2f}{dt^2}(\Delta t)^2 + \ldots + \\ + \frac{1}{n!}\cdot\frac{d^nf}{dt^n}(\Delta t)^n. \qquad (52)$$

The terms of the Taylor series are expressed by exponential smoothing formulas.Here we give a formula for an exponentially smoothed quantity of the first order:

$$S_t(f) = \alpha f_t + (1-\alpha) S_{t-1}. \qquad (53)$$

Thus, the new averaged value $S_t(f)$ is equal to the last known value of the function f(t) multiplied by the factor α (where $\alpha \leqslant 1$) plus the preceding averaged value S_{t-1} multiplied by $(1-\alpha)$. When $\alpha = 1$ we obtain a trustworthy transcription of the past values, i.e. a prediction according to the "no change" rule.

Less sensitive systems, which can be used with large noise, use values $1 > \alpha > 0.5$, and more conservative ones $0.5 > \alpha > 0.1$.

$$\begin{aligned} S_t^2(f) &= \alpha S_t(f) + (1-\alpha) S^2_{t-1}(f), \\ S_t^3(f) &= \alpha S_t^2(f) + (1-\alpha) S^3_{t-1}(f), \qquad (54) \\ &\dots\dots\dots\dots\dots\dots \\ S_t^n(f) &= \alpha S_t^{n-1}(f) + (1-\alpha) S^n_{t-1}(f). \end{aligned}$$

Now there remains to express the terms of the Taylor series in terms of the averaged quantities. Depending on how many terms of the series we use, the following formulas are used.

One term of the series:

$$f_{t+\Delta t} = f_t; \quad f_t = S_t(f). \qquad (55)$$

Two terms of the series:

$$f_{t+\Delta t} = f_t + \frac{df}{dt}\Delta t; \quad f_t = 2S_t(f) - S_t^2(f); \qquad (56)$$

$$\frac{df}{dt} = \frac{\alpha}{1-\alpha}[S_t(f) - S_t^2(f)].$$

Three terms of the series:

$$f_{t+\Delta t}=f_t+\frac{df}{dt}\Delta t+\frac{1}{2}\ \frac{d^2f}{dt^2}(\Delta t)^2;$$

$$f_t=3S_t(f)-3S_t^2(f)+S_t^3(f); \qquad (57)$$

$$\frac{d^2f}{dt^2}=\frac{\alpha^2}{(1-\alpha)^2}[S_t(f)-2S_t^2(f)+S_t^3(f)];$$

$$\frac{df}{dt}=\frac{\alpha^2}{2(1-\alpha)^2}[(6-5\alpha)S_t^2(f)-2(5-4\alpha)S_t^2(f)+$$
$$+(4-3\alpha)S_t^3(f)].$$

Numerical modelling

A predicting filter based on operator (52) should the following operations:

1. Compute the exponentially smoothed quantities and required orders in accordance with (54).

2. Determine the terms of the operator

$f(t), \frac{df(t)}{dt}, \frac{d^2f(t)}{dt^2},\dots$ in accordance with (55)-(57).

3. Sum the terms of series (52).

The algorithm for the operation of the predicting filter can be written in the form

$$\overset{3}{\downarrow} T \overset{1}{\uparrow} S\alpha \overset{1}{\uparrow} X \overset{2}{\downarrow} D\beta \overset{2}{\uparrow} \Sigma\gamma \overset{4}{\uparrow} O \overset{4}{\downarrow} \omega \overset{3}{\uparrow},$$

where T is the operator for emission of the values of the points of the prehistory (shifts along the sequence being predicted);

' is the operator for computing the exponentially smootl quantities;

α s a logical condition, fulfilled when S_t is obtained;

X is the operator for computing f(t) (first term of the series);

D is the operator for computing terms of the second, third, etc. orders.

β is a logical condition, fulfilled when an operator term of higher order is obtained;

Σ is the operator for summing the terms of series (52);

γ is a logical condition, fulfilled when the anticipation cycle is terminated;

0 is a stop;

ω is an identically false condition.

Figure 29 shows a flbw chart for the realization of the algorithm on a digital computer. Besides itself realizing the prediction algorithm, the program provides for operation under conditions with variable coefficient α . The quality of prediction is estimated from the mean-square error for various values of α . As will be proved later, the coefficient α depends on the statistical characteristics of the sequence being predicted and can take different values for different real processes.

The program allows during the process of operation with the initial data of the process being investigated the selection of an optimal value of α for which the best prediction quality is obtained (in the sense of $\overline{\varepsilon^2}_{\min}$). Thus, the digital model of the predicting filter operates in the learning mode and, after determination, is switched into the general program for solving the control problem.

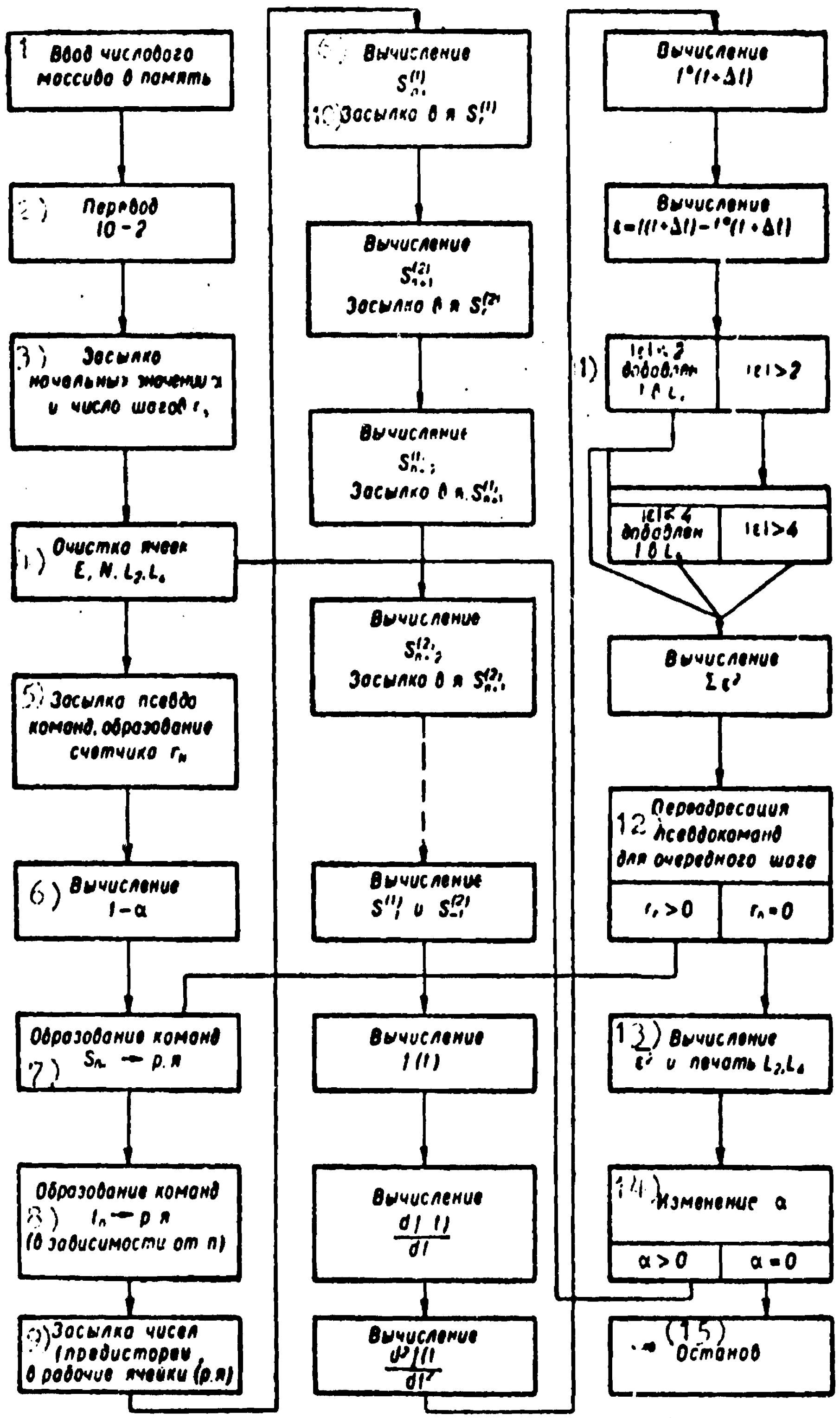

1 Ввод числового массива в память
2) Перевод 10 - 2
3) Засылка начальных значений и число шагов r,
4) Очистка ячеек E, N, L2, L4
5) Засылка псевдо команд, образование счетчика rн
6) Вычисление 1 - α
Образование команд Sn → р.я 7)
Образование команд 8) tn → р я (в зависимости от n)
9) Засылка чисел (предисторий) в рабочие ячейки (р.я)
Вычисление S(1)n+1
10) Засылка в я S(1)
Вычисление S(2)n+1 Засылка в я S(2)
Вычисление S(1)n+2 Засылка в я S(1)n+1
Вычисление S(2)n+2 Засылка в я S(2)n+1
Вычисление S(1) и S(2)
Вычисление f(t)
Вычисление df(t)/dt
Вычисление d²f(t)/dt²
Вычисление f*(t+Δt)
Вычисление ε = f(t+Δt) - f*(t+Δt)
11)
|ε| ≤ 2 добавлен 1 в L2
|ε| > 2
|ε| ≤ 4 добавлен 1 в L4
|ε| > 4
Вычисление Σε²
12 Переадресация псевдокоманд для очередного шага
rn > 0
rn = 0
13) Вычисление ε² и печать L2, L4
14) Изменение α
α > 0
α = 0
(15) Останов

Fig. 29. Block diagram of modeling a predicting filter based on the exponential smoothing algorithm

Key: 1. introduce numerical ensemble into memory; 2. convert; 3. send initial values of α and number of intervals r; 4. erase cells; 5. send pseudo commands, form counter; 6. compute; 7. form commands $S_n \rightarrow$ w.c.; 8. form commands $t_n \rightarrow$ w.c. (depending on n); 9. send numbers (prehistory) to working cells (w.c.); 10. send to cell; 11. add 1 to L_4; 12. readdress pseudo commands for next step; 13. compute $\overline{\varepsilon}^2$ and print L_2 , L_4; 14. change α ; 15. stop.

Prediction of the change in quality indices of petroleum products and investigation of a predicting filter

As an example let us consider the prediction of the temperature at the end of boiling of direct-distillation benzine. In solving this problem, we shall use the data with which we operated in solving the prediction problem by the method of linear extrapolation.

Let us set the value $\alpha = 0.1$. Let us take the prediction interval equal to 30 min as a unit. As in the linear extrapolation problem, prediction will be performed from 2,3,4 and 5 points of the prehistory.

For k=2:

when $t=-2$ $\quad S(f)=182;\quad S^2_{-2}(f)=182;$

when $t=-1$ $\quad S_{-1}(f)=0{,}1\cdot 182+0{,}9\cdot 182=182;$

$$S^2_{-1}(f)=0{,}1\cdot 182+0{,}9\cdot 182=182.$$

Let us use two terms of the Taylor series

$$f_{-1}=2S_{-1}(f)-S^2_{-1}(f)=2\cdot 182-182=182;$$

$$\frac{df}{dt}=\frac{\alpha}{1-\alpha}[S_{-1}(f)-S^2_{-1}(f)]=0;$$

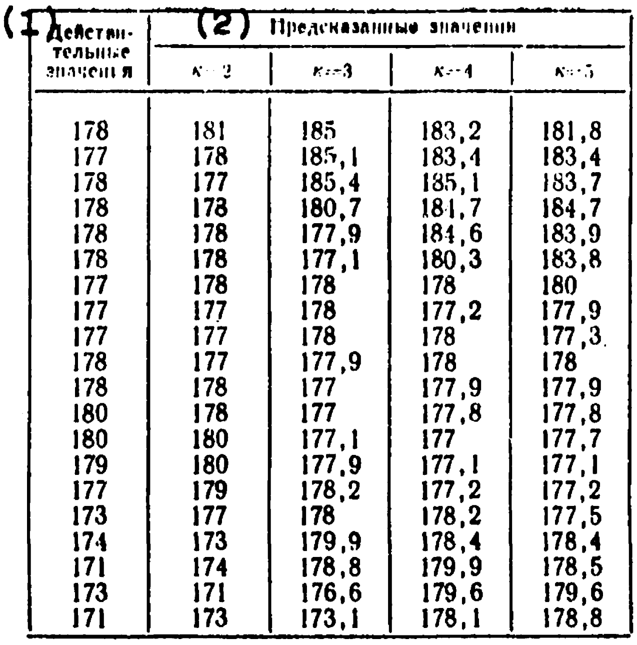
$$f_0=f_{-1}+\frac{df}{dt}\Delta t=182.$$

For k=3:

when $t=-3$ $\quad S_{-3}(f)=182;\quad S^2_{-3}(f)=182;$

$t=-2$ $\quad S_{-2}(f)=0{,}1\cdot 182+0{,}9\cdot 182=182;$

when $\quad S^2_{-2}(f)=0{,}1\cdot 182+0{,}9\cdot 182=182;$



(1) Действительные значения	(2) Предсказанные значения			
	к=2	к=3	к=4	к=5
178	181	185	183,2	181,8
177	178	185,1	183,4	183,4
178	177	185,4	185,1	183,7
178	178	180,7	184,7	184,7
178	178	177,9	184,6	183,9
178	178	177,1	180,3	183,8
177	178	178	178	180
177	177	178	177,2	177,9
177	177	178	178	177,3
178	177	177,9	178	178
178	178	177	177,9	177,9
180	178	177	177,8	177,8
180	180	177,1	177	177,7
179	180	177,9	177,1	177,1
177	179	178,2	177,2	177,2
173	177	178	178,2	177,5
174	173	179,9	178,4	178,4
171	174	178,8	179,9	178,5
173	171	176,6	179,6	179,6
171	173	173,1	178,1	178,8

Table 2.

Key: 1) Actual values; 2) predicted values.

when $t = -1$ $S_{-1}(f) = 0{,}1 \cdot 181 + 0{,}9 \cdot 182 = 181{,}9;$

$$S^2_{-1}(f) = 0{,}1 \cdot 181{,}9 + 0{,}9 \cdot 182 = 181{,}99.$$

Using two terms of the Taylor series, we obtain:

$$f_{-1} = 2S_{-1}(f) - S^2_{-1}(f) = 182,$$

$$\frac{df}{dt} = \frac{\alpha}{1-\alpha}\left[S_{-1}(f) - S^2_{-1}(f)\right] = 0{,}01,$$

$$f_0 = f_{-1} + \frac{df_i}{dt}\Delta t = 181{,}99,$$

etc.

Table 2 shows the results of solving the problem on a universal digital computer. Figure 30 shows graphs of the change in the actual and predicted values of $T_{KK}{}^{o}C$ for a different number of points of the prehistory (k=2,3,4,5).

Let us formulate the basic problems in investigating the predicting filter.

1. Investigation of the influence of the prehistory length k on the prediction quality.

2. Investigation of the prediction quality as a function of the parameter α .

3. Investigation of the prediction quality as a function of the anticipation time.

4. Investigation of the time parameters of the predicting filter.

Figure 30e shows a graph which reflects the change in the mean-square error of prediction due to the number of points k participating in the computation of the exponentially smoothed values and derivatives. The function $\overline{\varepsilon^2} = f(k)$ has a minimum when k=3 (for $\alpha \gg \alpha_{opt}$). For $k > k_{opt}$, the error rises as k increases.

Similar investigation conducted in the prediction of quality indices of other petroleum products have confirmed the conclusion that we should choose $k_{opt}=3$ for predicting such processes. This conclusion is valid only for the accepted speed of action of the automatic quality analyzer.

It can be seen from Fig.30f that for the processes under investigation there exists some definite value $\alpha = \alpha_{opt}$ for which the prediction error is minimal. In every concrete case, the value α_{opt} is characterized by the statistics of the process. For processes which reflect changes in the $T_{KK}{}^{o}C$ of petroleum products, the value of α_{opt} lies in the range 0.2-0.4.

If we compare Fig.30e and 30f, it becomes obvious

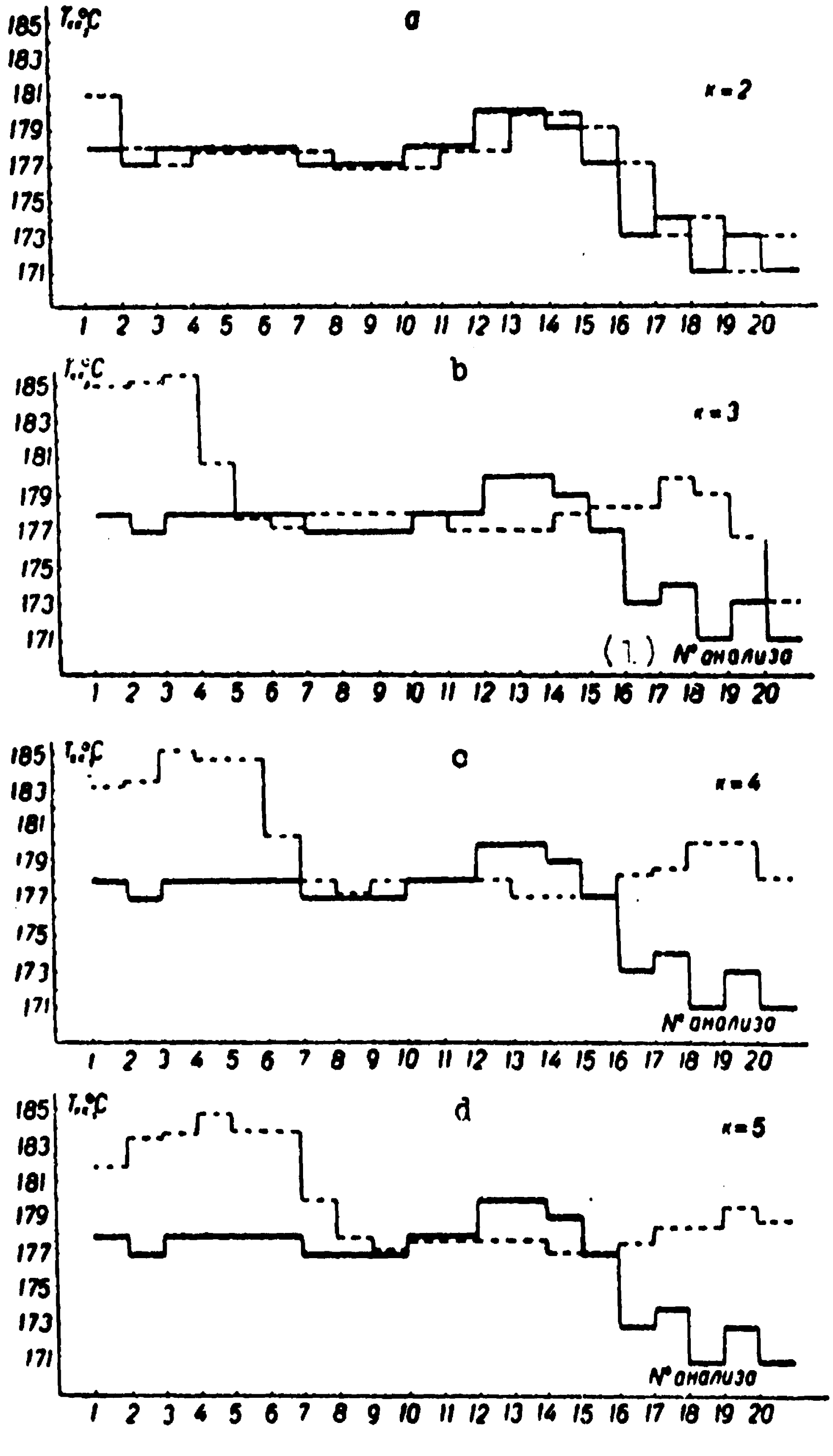
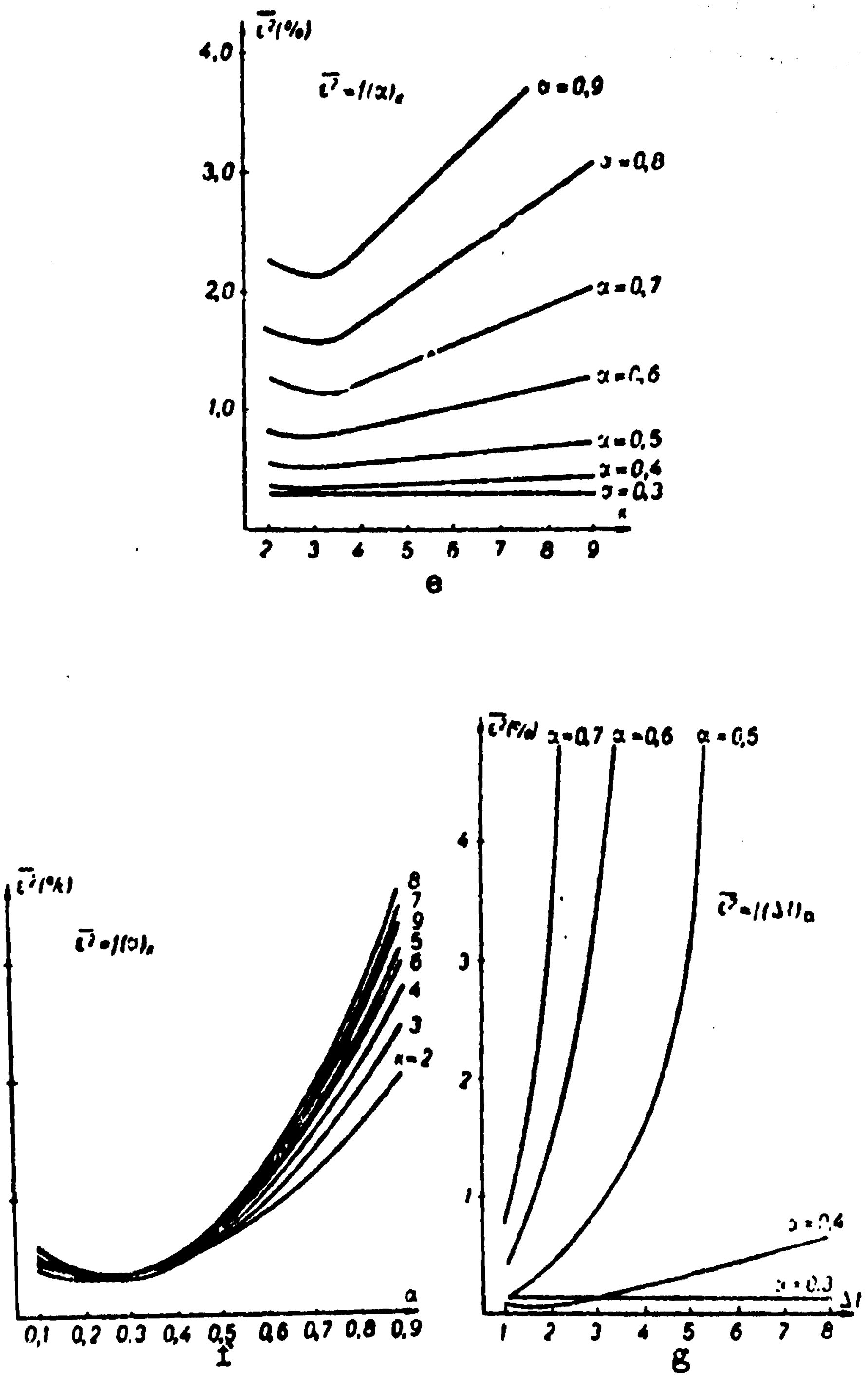


Figure 30 (a-d). Prediction of the temperature at the end of boiling by the exponential smoothing method (k=2,3,4,5).

Key: 1) Number of analysis.



i(%)
4,0
3,0
2,0
1,0
i = f(α)ₑ
α = 0,9
α = 0,8
α = 0,7
α = 0,6
α = 0,5
α = 0,4
α = 0,3
n
2 3 4 5 6 7 8 9
e
i(%)
i = f(α)ₙ
8
7
9
5
6
4
3
n = 2
α
0,1 0,2 0.3 0,4 0,5 0,6 0,7 0,8 0,9
f
i(%)
α = 0,7
α = 0,6
α = 0,5
4
3
2
1
i = f(Δt)α
α = 0,4
α = 0,3
Δt
1 2 3 4 5 6 7 8
g

Figure 30 (e-g).

that when $\alpha = \alpha_{opt}$ the prediction quality in practice depends little on the length of the prehistory.

In Fig.30g are graphs which show the prediction quality as a function of the anticipation time $\overline{\varepsilon^2} = f(\Delta t)$. Preceding investigation were conducted with Δ t=1, i.e. predicting the k+1-th value from the k preceding ones. The error was determined as the mean error over the set of values thus predicted.

In investigating $\overline{\varepsilon^2} = f(\Delta t)$ according to the k known values, the k+1-th value was determined, and the k+2-th value was computed taking into account the k+1-th predicted value, and not the actual one.

It can be seen from Fig.30g that the mean-square error of prediction rises sharply as Δ t increases. However, when $\alpha = \alpha_{opt}$, this error is minimal and depends little on the anticipation time.

The volume of computation, and hence the realization time of the algorithm described depend only on the number of points of the prehistory which participate in the computation of the exponentially smoothed values and derivatives. Since the cycle time for computing the exponentially smoothed quantities is constant, the function $t_{predict.} = f(k)$ is linear.

Thus, the results of these investigations make it possible to choose the parameters of the predicting filter in the best manner.

In connection with the change in the external conditions and the parameters of the processes being regulated, it is necessary periodically to switch over the predicting filter from the prediction mode to the learning mode. Here the greatest effect can be obtained if we use a computer operating on the multiprogram principle as a controlling machine. Such a machine allows the simultaneous realization of several independent programs and, in our case, makes it possible to achieve control with a predicting filter as a parallel corrector.

An extrapolating filter based on the exponential smoothing algorithm

Devices for predicting the future value of a function for linear and quadratic extrapolation can be assembled in accordance with the block diagram in Fig.31a /34/. As can be seen from Fig.31a, the circuit contains no sections with constant lag.

The circuit is made up of amplifiers, summing devices, and linear aperiodic section of the first order.

Figure 31b shows an experimental oscillogram of the operation of the extrapolating filter for the input function f(t). The extrapolator circuit was modelled on a type MPT-9 analog computer.

Prediction of Stationary Random Processes

Wiener's method

Let the total input signal of some system be given by the expression

$$f(t) = S(t) + N(t),$$

where is the signal carrying the useful information, and N(t) is the noise.

In the ideal case it is required to determine such a system so that the signal at its output will be equal to $S(t+\Delta t)$.

When $\Delta t=0$, the system is called a <u>filter</u>. When

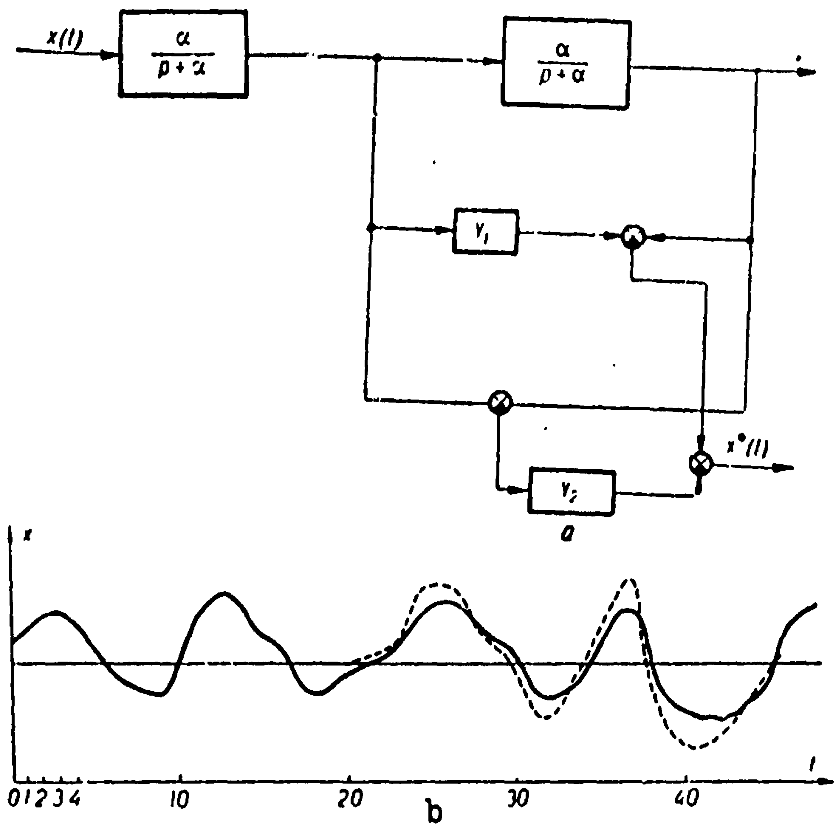


Figure 31. a) Extrapolating filter based on the exponential smoothing algorithm; b) automatic function extrapolation.

$\Delta t > 0$ and $N(t) \equiv 0$, the system is called an *anticipator*.

In the general case, the system must perform both operations, filtration and anticipation, simultaneously. In what follows, we shall call this kind of system a *predicting filter*.

N.Wiener [62] developed a theory of these systems based on the following assumptions:

1. The functions S(t) and N(t) are stationary and stationarily connected random processes.

2. The criterion for selecting the "best" possible system is the mean-square value of the difference between the actual signal and the desired signal at the output of the system

$$\varepsilon^2 = M\,[S(t+\Delta t) - S^*(t+\Delta t)]^2.$$

3. The operation performed for filtration and prediction is assumed to be a linear operation on the information at hand.

In other words, the system must be a linear physically realizable filter. Physical realizability should not be identified with the possibility of embodying the system in an actual design. The requirement of physical realizability consists of the fact that the reaction of the system to a unit pulse function becomes zero for t < 0.

Since the properties of a linear system are completely characterized by the pulse transfer function W(t), the output signal of the system can be written in the form of a convolution integral:

$$S^*(t+\Delta t) = \int_0^\infty [S(t-\tau) + N(t-\tau)\,W(\tau)\,d\tau. \qquad (58)$$

Then for the mean-square error we have the expression

$$\overline{\varepsilon}^2(t) = S^2(t+\Delta t) - 2\int_0^\infty [\overline{S(t+\Delta t)\,S(t-\tau)} +$$

$$\overline{+ S(t+\Delta t)\,N(t-\tau)]\,W(\tau)}\,d\tau + \int_0^\infty W(\tau_1)\,d\tau_1$$

$$\int_0^\infty W(\tau_2)\overline{[S(t-\tau_1)+N(t-\tau_1)][S(t-\tau_2)+N(t-\tau_2)]}\,d\tau_2. \qquad (59)$$

Noting that the correlation between x(t) and y(t) is

$$K_{xy}(\tau) = \overline{x(t)\, y(t+\tau)}, \qquad (60)$$

and making the notation

$$K_{SS}(\tau) + K_{NS}(\tau) = \psi(\tau), \qquad (61)$$
$$K_{SS}(\tau) + K_{NS}(\tau) + K_{SN}(\tau) + K_{NN}(\tau) = \varphi(\tau),$$

let us rewrite formula (59):

$$\bar{\varepsilon^2} = K_{SS}(0) - 2\int_0^\infty \psi(\Delta t + \tau)\, W(\tau)\, dt +$$
$$+ \int_0^\infty W(\tau_1)\, d\tau_1 \int_0^\infty W(\tau_2)\, \varphi(\tau_2 - \tau_1)\, d\tau_2. \qquad (62)$$

Now the problem can be formulated as follows. We know the correlation functions K_{SS}, K_{NS}, K_{SN}, K_{NN}. We must find a pulse transfer function W(t) such that

$$\bar{\varepsilon^2} = \min.$$

Rutting $W(t)_{t<0} = 0$ here, we automatically satisfy the condition of physical realizability.

From the general problem, there follow important special cases, such as the problem of filtering ($\Delta t = 0$) and the problem of "pure" prediction ($N(t) \equiv 0$).

The method of Zadeh and Ragazzini

A generalization of Wiener's theory for the case of a finite time interval was considered by Zadeh and Ragazzini [61]. Their method is based on the following assumptions:

1. The signals being considered consist of: a) a nonrandom time function which is representable by polynomials of a degree not exceeding some definite number n, and about

№ п.п	1) Отношение $S^*(t)$ и $S(t)$	(2) Оцениваемая величина	$Y(p)$	$w(t)$
1	$S^*(t)=S(t)$	(3) Настоящее значение $S(t)$	1	$\delta(t)$
2	$S^*(t)=S'(t)$	» » $S'(t)$	p	$\delta^{(1)}(t)$
3	$S^*(t)=S''(t)$	» » $S''(t)$	p^2	$\delta^{(2)}(t)$
4	$S^*(t)=S(t+\Delta t)$	(4) Будущее или прошлое значение $S(t)$ (Δt «+» или «—»)	$e^{\Delta tp}$	$\delta(t+\Delta t)$

Table 3.

Key: 1) Relationship of $S^*(t)$ and $S(t)$; 2) quantity being evaluated; 3) present value; 4) future or past value of $S(t)$ (Δ t "+" or "-").

which we know knothing except n; b) a stationary random time function whose correlation function is known.

2. The pulse transfer function $W(t) = 0$ when $\begin{cases} t \leqslant 0, \\ t < T. \end{cases}$

Let us consider the time function $f(t)$, consisting of $S(t)$ and $N(t)$.

The output of the predicting filter $S^*(t)$ is connected with $S(t)$ by the linear operator $Y(p)$:

$$S^*(t) = Y(p)\,S(t). \qquad (63)$$

Let us write S^* in the form of a convolution integral:

$$S^*(t) = \int_{-\infty}^{\infty} w(\tau)\,S(t-\tau)\,d\tau, \qquad (64)$$

where $w(\tau)$ is the pulse tranfer function of an ideal anticipator.

In the most general case, the quantity $S^*(t)$ being evaluated in prediction or filtration may be a functional

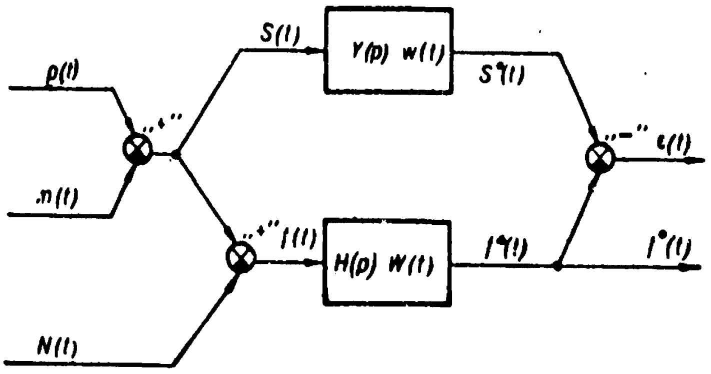


Figure 32. Zadeh-Ragazzini predicting filter.

of S(t).

Table 3 gives some possible values of Y(p) and w(t) for various S*(t).

It is assumed that

$$S(t) = m(t) + p(t), \qquad (65)$$

where p(t) is a nonrandom time function which can be represented by a polynomial in t of order not higher than n;

m(t) is a stationary random component;

m(t) and n(t) are described by the correlation functions $K_{mm}(\tau)$ and $K_{nn}(\tau)$.

It is further assumed that m(t) and n(t) are centered and uncorrelated.

Figure 32 shows a block diagram of a predicting filter for this problem.

In the absence of noise and with the physical realizability condition satisfied, there is no error

$$\varepsilon = f^*(t) - S^*(t) \qquad (66)$$

Here the operator of the actual anticipator H(p) is identical with Y(p). This case is trivial.

If H(p) and Y(p) do not coincide, it is necessary to determine W(t) such that

$$\bar{\varepsilon}^2 = \overline{[f^*(t) - S^*(t)]^2} = \min. \qquad (67)$$

The optimal predicting filter must satisfy the following conditions:

a) the mean error over the set is zero for all values of t;

b) the variation of ε over the set is minimal.

Let us write the output signal in the form

$$f^*(t) = \int_0^\infty W(\tau) f(t-\tau)\, d\tau. \qquad (68)$$

In practice it is necessary to bound the interval of the input function by some finite T. Then

$$f^*(t) = \int_0^T W(\tau) f(t-\tau)\, d\tau. \qquad (69)$$

Taking into account that

$$f(t) = p(t) + m(t) + n(t), \qquad (70)$$

and expressing

$$p(t-\tau) = p(t) - \tau p'(t) + \frac{\tau^2}{2!} p''(t) + \ldots + (-1)^n \frac{\tau^n}{n!} p^n(t),$$

we obtain

$$f^*(t) = \mu_0 p(t) - \mu_1 p'(t) + \frac{\mu_2}{2!} p''(t) + \ldots + (-1)^n \frac{\mu_n}{n!} p^n(t) + \int_0^T W(\tau) m(t-\tau)\, d\tau + \int_0^T W(\tau) n(t-\tau)\, d\tau, \qquad (72)$$

where $\mu_0, \mu_1, \ldots$ are the moments of W(t), equal to

$$\mu_\nu = \int_0^T \tau^\nu W(\tau)\,d\tau,\ \nu = 0,\ 1,\ 2,\ \dots,\ n. \qquad (73)$$

Since m(t) and n(t) are centered stationary functions, f*(t) and S*(t) depend only on the nonrandom components of the signal:

$$\overline{f^*(t)} = \int_0^T W(\tau)\,p(t-\tau)\,d\tau, \qquad (74)$$

or

$$\overline{f^*(t)} = \mu_0 p(t) - \mu_1 p^1(t) + \dots + (-1)^n \frac{\mu_n}{n!} p^{(n)}(t) \qquad (75)$$

and

$$\overline{S^*(t)} = \overline{Y(p)\,S(t),} \qquad (76)$$

or

$$\overline{S^*(t)} = Y(p)\,p(t). \qquad (77)$$

Comparing (75) and (77), we can write condition a) as follows:

$$Y(p)\,p(t) \equiv \mu_0 p(t) - \mu_1 p'(t) + \dots + (-1)^n \frac{\mu_n}{n!} p^n(t). \qquad (78)$$

Identity (78) determines the value of μ .

In other words, the ideal prediction operator Y(p) is determined from (78) by the first (n+1) moments of the pulse transfer function of the optimal anticipator. As an example, let us consider the case

$$Y(p)\,S(t) = S(t+\Delta t) \quad (\text{when } \pm \Delta t).$$

Formula (78) can be rewritten in the form

$$p(t+\Delta t) \equiv \mu_0 p(t) - \mu_1 p'(t) + \frac{\mu_2}{2!} p''(t) + \dots + (-1)^n \frac{\mu_n}{n!} p^{(n)}(t). \qquad (79)$$

Comparing

$$p(t+\Delta t)\equiv p(t)-\Delta t p'(t)+\frac{\Delta t^2}{2!}p''(t)+\dots+$$
$$+(-1)^n\frac{\Delta t^n}{n!}p^{(n)}(t) \qquad (80)$$

with (79), we obtain the system

$$\mu_0=\int_0^T W(\tau)\,d\tau=1,$$
$$\mu_1=\int_0^T \tau W(\tau)\,d\tau=\Delta t,$$
$$\dots\dots\dots\dots$$
$$\mu_n=\int_0^T \tau^n W(\tau)\,d\tau=\Delta t^n. \qquad (81)$$

When condition a) is satisfied, it follows from an examination of (66), (72) and (78) that

$$\varepsilon=\int_0^T W(\tau)\,[m(t-\tau)+n(t-\tau)]\,d\tau-Y(p)\,m(t), \qquad (82)$$

or

$$\varepsilon=\int_0^T W(\tau)\,[m(t-\tau)+n(t-\tau)]\,d\tau-$$
$$-\int_{-\infty}^{\infty} w(\tau)\,m(t-\tau)\,d\tau. \qquad (83)$$

Then

$$\overline{\varepsilon^2}=\lim_{L\to\infty}\frac{1}{L}\int_0^L \varepsilon^2 dt. \qquad (84)$$

After a number of intermediate transformations, the final expression for the mean-square error has the form

$$\overline{\varepsilon^2}=\int_0^T\int_0^T W(\tau_1)\,W(\tau_2)\,[K_{mm}(\tau_1-\tau_2)+K_{nn}(\tau_1-\tau_2)]\,d\tau_1 d\tau_2-$$
$$-2\int_{-\infty}^{\infty}\int_0^T W(\tau_1)\,w(\tau_2)\,K_{mm}(\tau_1-\tau_2)\,d\tau_1 d\tau_2+$$
$$+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(\tau_1)\,w(\tau_2)\,K_{mm}(\tau_1-\tau_2)\,d\tau_1 d\tau_2. \qquad (85)$$

The last term of (85) does not depend on w(t). Since w(t) is the kernel of the (n+1) equations (83), the problem of minimizing $\overline{\varepsilon^2}$ with respect to the class of w(t) which satisfy (73) is reduced to minimization of the expression

$$I = \int_0^T W(\tau_1)\,d\tau_1 \left\{ \int_0^T W(\tau_2)\,[K_{mm}(\tau_1-\tau_2) + K_{nn}(\tau_1-\tau_2)]\,d\tau_2 - \right.$$

$$-2\int_{-\infty}^{\infty} w(\tau_2)\,K_{mm}(\tau_1-\tau_2)\,d\tau_2 - 2\lambda_0 - 2\lambda_1\tau_1 - \ldots -$$

$$\left. -2\lambda_n\tau_1^n \right\}, \qquad (86)$$

where $\lambda_0, \lambda_1, \ldots, \lambda_n$ are lagrangian multipliers.

Letting I tend to zero, we obtain the minimum error $\overline{\varepsilon^2}$ for the value of W(t) satisfying the integral equation

$$\int_0^T W(\tau)\,[K_{mm}(t-\tau) + K_{nn}(t-\tau)]\,d\tau = \lambda_0 + \lambda_1 t + \ldots +$$

$$+\lambda_n t^n + \int_{-\infty}^{\infty} w(\tau)\,K_{mm}(t-\tau)\,d\tau, \quad 0 \leqslant t \leqslant T. \qquad (87)$$

The optimal predicting filter is found from equations (73) and (87). It should be noted that solution of the integral equations in synthesizing optimal predicting filters presents great difficulties. Reference [61] gives methods of solution for individual special cases.

The method of Bode and Shannon

This method of filtering and predicting random processes is based on expressing the mean-square error in terms of the spectral densities of the noise and signal powers.

The basic problem consists of determining Y(ω) (Fig.33). What will be the prediction error for this Y(ω)?

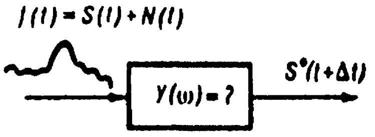


Figure 33. Bode and Shannon predicting filter.

The mean power of the error

$$\varepsilon = S(t+\Delta t) - S^*(t+\Delta t)$$

for incoherent frequencies can be computed by summing the components of various frequencies

$$\overline{\varepsilon^2} = \int_{-\infty}^{\infty} [|Y(\omega)|^2 N(\omega) + |Y(\omega) - e^{i\Delta t\omega}|^2 P(\omega)]\, d\omega, \qquad (88)$$

where $P(\omega)$ is the signal power, and $N(\omega)$ is the noise power.

It is required to minimize $\overline{\varepsilon^2}$ by suitable selection of $Y(\omega)$, taking into account the condition of physical realizability.

If $f(t)=S(t)+N(t)$ is passed through a filter with amplification $[P(\omega)+N(\omega)]^{-(1/2)}$, we obtain a flat spectrum. Let the minimal-phasefilter with characteristic $Y_1(\omega)$ have amplification $[P(\omega)+N(\omega)]^{-(1/2)}$. Then both $Y_1(\omega)$ and $Y_1^{-1}(\omega)$ are physically realizable filters.

If $f(t)$ were known in the interval $-\infty < t < \infty$, the best operation, applied to the input, would be the operation satisfying the equality

$$Y(\omega) = \frac{P(\omega)}{P(\omega) + n(\omega)} e^{i\Delta t\omega}. \qquad (89)$$

If the phase characteristic is $B(\omega)$), (89) is equivalent to the operation

$$Y_2(\omega) = \frac{P(\omega)}{[P(\omega) + n(\omega)]^{\frac{1}{2}}} e^{j[\Delta t\omega - B(\omega)]} \qquad (90)$$

on $Y_1(\omega)$, which has the character of white noise.

The corresponding weighting function is

$$W_2(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y_2(\omega) e^{j\omega t}. \qquad (91)$$

This is the function of a physically realizable filter. Let us put

$$W_3(t) = \begin{cases} W_2(t + \Delta t) & t > 0, \\ 0 & t < 0. \end{cases} \qquad (92)$$

$W_3(t)$ is the weighting function of the physically realizabbe filter with transfer function $Y_3(\omega)$. Then the transfer function of the optimal predicting filter for f(t)=S(t)+N(t) can be expressed in the form

$$Y_4(\omega) = Y_1^{-1}(\omega) Y_3(\omega). \qquad (93)$$

As in the case of the Wiener formulation, we can separate the special cases of pure prediction and pure filtration from the above general problem.

Nonlinear signal filtering

The problem of optimal nonlinear filtering

All the problems considered above were based on the general assumption that the operation performed on the initial information was linear. In comparison with linear systems, nonlinear ones under definite conditions can give a smaller mean-square error than the best (in the sense of mean-square error) linear systems.

Below we consider a class of systems described by the general relationship

$$S^*(t) = \sum_{n=0}^{N} \int_0^{\infty} w(\tau)\, \vartheta_n [f(t-\tau)]\, d\tau, \qquad (94)$$

where $S^*(t)$ is the output signal of the system;

$f(t)$ is the input signal of the system;

$w_n(\tau)$ is the weighting function of the linear part of the system;

$\vartheta_n[x]$ is a set of linearly independent functions. [29].

The problem consists of determining a set of optimal weighting functions $w_n(\tau)$, if there is sufficient statistical information on the input and desired output signals.

Such a system can be considered as a system of several parallel channels, each of which consists of a nonlinear element without memory and a linear element with memory connected in series (Fig.34). The weighting function of each circuit is equal to $w_n(\tau)$. If $w_n(\tau) = a_n \delta(\tau)$, the system turns into a multichannel system without memory.

The optimal system is defined as a system ensuring a minimum of the expression

$$\overline{\varepsilon^2} = \overline{[S(t) - S^*(t)]^2}.$$

Let us introduce the equality

$$S(t) = g[f(t)],$$

where g is an operator of a definite class. This operator acts on all values of f(t) in the interval $[-\infty, t]$ or on a part of these values and has a finite mean square.

The variation of the mean-square error $\overline{\varepsilon^2}$ caused by the variation $\sigma h[f]$ of the operator $g[f]$ is given by the equality

$$\Delta\overline{\varepsilon^2} = -2\sigma h[f]\,(g[f] - S^*) + \sigma^2 \overline{h}[f]^2. \qquad (95)$$

Here ε is a small constant, and $h[f]$ is an operator of the same class as $g[f]$.

In order that $g[f]$ should be an optimal operator, we must have the equality

$$\frac{\partial\Delta\overline{\varepsilon^2}}{\partial\sigma} = -\overline{2h[x]\,(g[x] - S^*)} = 0, \qquad (96)$$

since $\sigma = 0$; $h\overline{[f]\{g[f] - S^*\}} = 0$.

Expression (96) is a necessary and sufficient condition that $g[f]$ should be optimal.

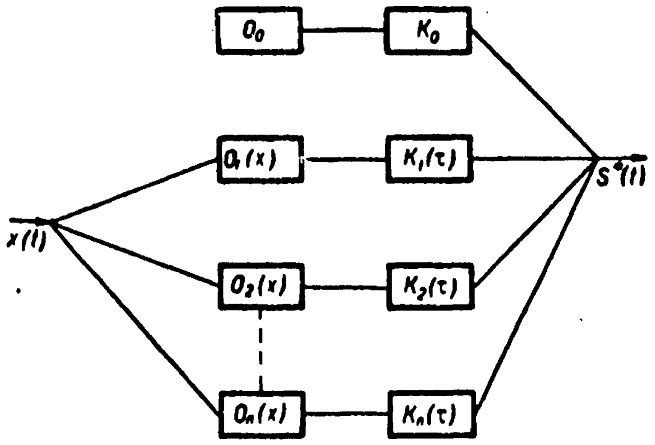


Figure 34. Nonlinear filter with memory.

Let us define the class of nonlinear systems (known by the name of multichannel systems without memory of degree N):

$$S(t) = g[f] = \sum_{m=0}^{N} k_{gm} \theta_m [f], \qquad (97)$$

where the k_{gm} are constant; the functions $\vartheta_m[f]$ are orthonormal polynomials of f of degree m.

Every polynomial $\vartheta_m(f)$ is characterized by some weight, p(f) being the probability density of the input signal.

The most general formula for $h[f]$ has the form

$$h[f] = \sum_{l=0}^{N} k_{hl} \theta_l [f]. \qquad (98)$$

Substituting (98) into (96), we have

$$\sum_{l=0}^{N} k_{hl} \left\{ \sum_{m=0}^{N} k_{gm} \overline{\theta_i [f] \theta_m [f]} - \overline{\theta [f] S^*} \right\} = 0.$$

Since

$$\int \theta_i [f] \theta_m [f] p [f] df = \theta_i [f] \theta_m [f] = \left.\right\} \begin{matrix} 0 \text{ when } m = i, \\ 1 \quad \text{"} \quad m \neq i, \end{matrix}$$

we obtain

$$\sum_{l=0}^{N} k_{hl} k_{gl} = \sum_{m=0}^{N} k_{hl} \theta_l [f] S^*. \qquad (99)$$

Since the k_{hi} take arbitrary values, then, in order that (99) should be satisfied, there must exist a unique solution for constant k_{hi}:

$$k_{gl} = \overline{\theta_l [f] S^*}. \qquad (100)$$

In virtue of $\vartheta_0[f] = 1$ $\quad k_{g0} = \overline{g [f]} = \overline{S(t)} = \overline{S^*(t)}. \qquad (101)$

Thus, the multichannel system without memory of order N is given by the equality

$$F[f(t)] = \sum_{l=1}^{N} \nu_{S^*} D_{l1}\, \theta_l = [f(t) + S^*(t)], \qquad (102)$$

where

$$D_{l1} = \frac{\theta_l [f] (S^* - \overline{S^*})}{\nu_{S^*}}\,.$$

From (102) we obtain

$$\nu_S^2 = \overline{[f[f(t)] - \overline{S^*(t)}]^2} = \sum_{l=1}^{N} \nu_{S^*} D_{l1}^2\,. \qquad (103)$$

Since $\theta_i[f] = 0$ when $n \neq 0$,

$$\theta_h [f]\, \theta_m [f] = \Big\} \begin{matrix} 1 \text{ when } l = m; \\ 0 \quad " \quad i \neq m. \end{matrix}$$

In (102) and (103), γ_S and γ_{S*} are respectively the mean-square deviations of the given output signal and the output signal of the optimal system.

Let us define the class of multichannel systems with memory of order N.

$$S(t) := g[f] = \sum_{m=0}^{N} \int_0^{\infty} k_{gm}(\tau)\, \theta_m [f(t-\tau)] d\tau. \qquad (104)$$

The most general expression for $h[f]$ is

$$h[f] = \sum_{l=0}^{N} \int_0^{\infty} k_{hl}(\tau)\, \theta_l [f(t-\tau)]\, d\tau.$$

Substitution in (97) gives

$$\sum_{l=0}^{N} \int_0^{\infty} k_{hn}(\tau_1) \left[\sum_{m=0}^{N} \int_0^{\infty} k_{gm}(\tau_2)\, C_{lm}(\tau_1 - \tau_2)\, d\tau_2 - \right.$$
$$- \theta_l [f(t-\tau_1)\, S^*(t)]\, d\tau_1 = 0, \qquad (105)$$

where $$C_{lm}(\tau) = \overline{\theta_l\,[f(t)]\,\theta_m\,[f(t-\tau)]}.$$

Using the fundamental theorem of the calculus of variations and noting that all the $k_{hi}(\tau)$ are arbitrary weighting functions of the physically realizable linear part of the system (i.e. $k_{hi}(\tau)=0$ for $\tau < 0$), we obtain

$$\sum_{m=0}^{N}\int_0^{\infty} k_{gm}(\tau_2)\,C_{lm}(\tau_1-\tau_2)\,d\tau_2 = \overline{\theta_n\,[f(t-\tau_1)\,S^*(t)]}, \quad (106)$$

when $\tau \geqslant 0$, $i=0,1,2,3,\ldots.$

Since

$$C_{0m}(\tau) \equiv \begin{cases} 0 \text{ when} m \neq 0, \\ 1 \;\; " \quad m = 0, \end{cases}$$

and

$$\theta_0\,[f(t)] \equiv 1,$$

then

$$\int_0^{\infty} k_{g0}(\tau)\,d\tau = k_0 = \overline{S^*(t)} \quad (107)$$

and

$$\sum_{m=1}^{N}\int_0^{\infty} k_{gm}(\tau_2)\,C_{lm}(\tau_1-\tau_2)\,d\tau_2 = \nu_{S^*}D_{ll}(\tau_1), \text{ when } \tau_1 \geqslant 0 \quad (108)$$

where

$$D_{ll} = \frac{\overline{0_i\,[f(t-\tau)]\,[S^*(t) - \overline{S^*(t)}]}}{\nu_{S^*}}.$$

It follows from (107) and (108) that

$$\nu_S^2 = \sum_{l=1}^{N}\sum_{m=1}^{N}\int_0^{\infty}\int_0^{\infty} k_l(\tau_1)\,k_m(\tau_2)\,C_{lm}(\tau_1-\tau_2)\,d\tau_1 d\tau_2 =$$
$$= \sum_{l=1}^{N}\int_0^{\infty} D_{l_1}(\tau)\,\nu_{S^*}\,k_l(\tau)\,d\tau. \quad (109)$$

System (109) can be solved by the method of undetermined coefficients [62].

The coefficients $C_{im}(\tau)$ and $\vartheta_{S*}D_{il}(\tau)$ can be determined by modelling. Here sufficiently long realizations of the input signal of the system and of the desired output signal should be known.

This optimization method can be used for designing nonlinear filters. As an example, let us consider the synthesis of a nonlinear filter for separating radio signals from their mixture with noise [29].

Calculation of optimal nonlinear filter

By using a nonlinear filter of the class under consideration, we can obtain a great decrease in the mean-square error. Let us assume that a useful signal which is a random pulse sequence S(t) arrives at the input of the receiver.

A noise N(t) is additively superimposed on this signal, i.e.

$$x(t) = aS(t) + bN(t).$$

When the signal is stationary, $f_1^n f_2^m$ is a function only of the time interval τ .

If S(t) and N(t) are statistically independent,

$$\bar{f}_n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} \overline{S^r N^{n-r}} \quad .$$

and

$$f_1^n, f_2^m = \sum_{r=0}^{n} \sum_{q=0}^{m} \binom{n}{r} \binom{m}{q} a^{r+q} b^{n+m-q-r} \overline{S_1^r S_2^q N_1^{n-2} N_2^{m-q}}.$$

Furthermore,

$$f_1^n S_s = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} \overline{S_1^r S_s N^{n-r}}.$$

In our example, the expression for $\overline{S_1^n S_2^m}$ has the form

$$\overline{S_1^n S_2^m} = \frac{1}{4}[(1+(-1)^m)(1+(-1)^n)+$$
$$+\rho(1-(-1)^m)(1-(-1)^n)].$$

The noise is characterized by a gaussian distribution; it has the same autocorrelation function as the useful signal.

$$N_1 N_2 = S_1 S_2 = \rho = e^{\rho(\tau)}.$$

Thus, the remaining functions $N_1^n N_1^m$ can easily be found.

Let us put

$$a^2 = 0{,}8;\ b^2 = 0{,}2;\ \frac{a}{b} = 2;\ a^2 + b^2 = 1.$$

In this case the orthnormal polynomials have the form:

$$\theta_0[x] = 1;$$
$$\theta_1[x] = x;$$
$$\theta_2[x] = 1{,}1785x^2 - 1{,}1785;$$
$$\theta_3[x] = 0{,}9836x^3 - 1{,}6918x;$$
$$\theta_4[x] = 8510x^4 - 2{,}6852x^2 + 1{,}2216;$$
$$\theta_5[x] = 6194x^5 - 2{,}7823x^3 + 2{,}3127x.$$

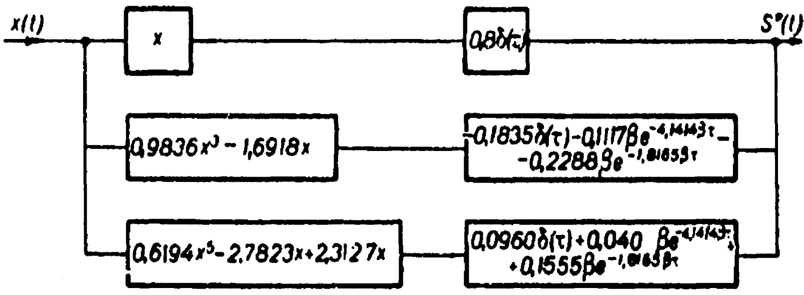
The coefficients aD_{hl} are:

$$aD_{11} = 0{,}8;$$
$$aD_{21} = 0;$$
$$aD_{31} = -0{,}2518;$$
$$aD_{41} = 0;$$
$$aD_{51} = 0{,}1414.$$

The equation of the fifth degree for the multichannel filter without memory has the form

$$S = 1{,}5530x - 0{,}6411x^3 + 0{,}0876x^5 \qquad (110)$$

for the required input signal aS*(t).



x(t)
x
0,8δ(τ)
S*(t)
0,9836x³ − 1,6918x
−0,1835δ(τ)−0,1117βe^{−4,14143τ} − −0,2288βe^{−1,81659τ}
0,6194x⁵−2,7823x+2,3127x
0,0960δ(τ)+0,040 βe^{−4,14143τ} + +0,1555βe^{−1,81659τ}

Figure 35. Nonlinear optimal filter.

The exact expression for the optimal nonlinear filter can be written in the form

$$S = \mathrm{ath}\,\frac{ax}{b^2}\,. \qquad (111)$$

Figure 36 shows graphs of the equations for filters with different a and b for the two forms of equations, (110) and (111).

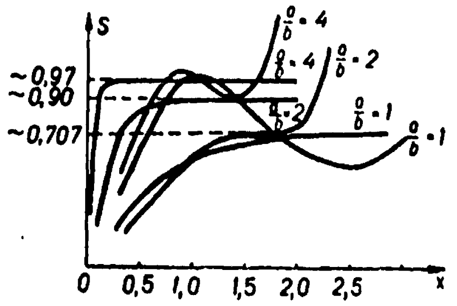


Figure 36. Graph of equations for filters with different a and b.

A multichannel filter for the case $a^2=0.8$ and $b^2=0.2$ is shown in Fig.35.

The relation of the mean-square errors of this nonlinear filter and the optimal linear filter has been computed and plotted as a graph (Fig.37).

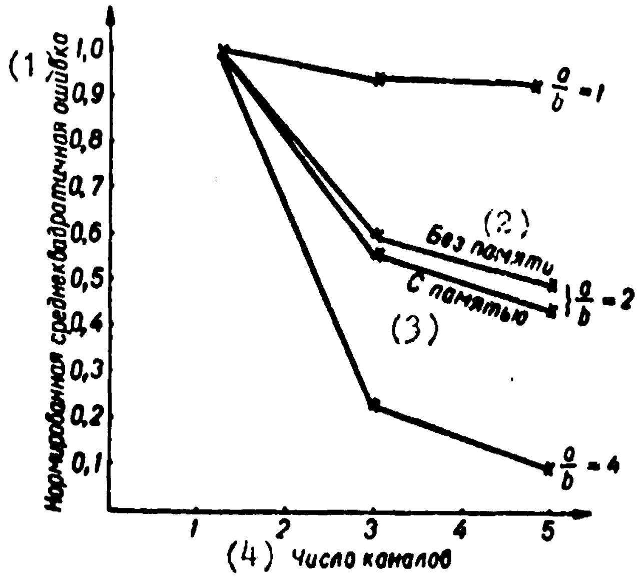


Figure 37. Mean-square error of nonlinear filter and optimal linear filter.

Key: 1) Normed mean-square error; 2) without memory; 3) with memory; 4) number of channels.

It should be noted that obtaining an exact filter equation requires the use of numerical methods of solution and the performance of a large volume of computations.

In the theory of prediction $S(t) = x(t+\Delta t)$.

Hence,

$$D_{i1}(\tau) = C_{i1}(t-\tau). \qquad (112)$$

Thus, if $C_{i1}(\tau)$ when $i\neq 1$, the optimal anticipator is linear, since system (111) takes the form

$$\int_0^\infty k_1(\tau_2) C_l(\tau_1 - \tau_2)\, d\tau_2 = \nu_x C_{11}(T + \tau_1). \qquad (113)$$

In the general case of filtering and predicting, just as with linear systems, we consider

$$x(t) = f[S(t);\ n(t)],$$
$$S^*(t) = S(t + \Delta t).$$

The methods described in this chapter are widely used for solving various proactical problems. These problems are: filtration and prediction of radio, telephone and telegraph signals; problems in transmitting and receiving television signals by the deflection method; problems of path determination and tracking of aircraft; and many others. The statistical theory of prediction can be widely used in biology and medicine. Thus, the hypothesis of the mechanism of visual pattern perception by deviations looks promising. According to this hypothesis, at every instant not all the information about the image being perceived arrives at the visual centers of the brain, but only information on deviations from the preceding image. This principle is used in work on the development of new systems for transmitting television pictures, this work at present being greatly expanded. The cooperation of engineers and biologists can be very fruitful from the point of view of elucidating many still unclear problems concerning the mechanism of pattern recognition. In turn, a correct answer to these questions will make it possible to develop more effective systems for recognizing not only visual, but also auditory and tactile patterns, for example in controlling various manipulators by means of muscle currents (miocontrol).

Chapter 3

Prediction of nonstationary random processes

Formulation of the problem

In the preceding chapter we became acquainted with the fundamentals of the Kolmogorov-Wiener theory and its use in predicting and filtering stationary random processes and sequences. Let us now consider the more general problem of predicting nonstationary random processes.

Let the values of x(t) be known in some interval $0 \leqslant t \leqslant T_0$. From these data we wish to determine the values of x(t) in the interval $T_0 < t \leqslant T$.

The values of x(t) can be written in the form of the following mean-convergent series [13]:

$$x(t) = \sum_{k=1}^{\infty} a_k \frac{\varphi_k(t)}{\sqrt{\lambda_k}}. \qquad (114)$$

In (114), the a_k are random variables such that the mathematical expectations

$$Ma_k = 0, \text{ and } Ma_k a_{k'} = \delta_{kk'},^1 \qquad (115)$$

$\phi_k(t)$ are eigenfunctions of the integral equation

$$\lambda\varphi(t) = \int_0^T r(t, \tau)\varphi(\tau)\,d\tau, \qquad (116)$$

and the λ_k are the corresponding eigenvalues.

[1]
$$\delta_{kk'} \begin{cases} 1 \text{ when } k=k', \\ 0 \text{ when } k\neq k'. \end{cases}$$

Let $L_2(X)$ denote the Hilbert space generated by the x(t) when $0 \leq t \leq T$, and let $P_{L_2(X)}$ denote the projection operator on this space. In this case, $L_2(X)$ coincides with the space A extended to the orthonormal system of vectors a_k (k=1,2,...), so that in order to find the predicted value at instant t+ Δ t, we must form the series

$$x^*(t+\Delta t) = P_{L_2(X)} = \sum_{k=1}^{\infty} a_k M a_k x(t+\Delta t). \qquad (117)$$

But

$$Ma_k x(t+\Delta t) = \sqrt{\lambda_k}\, Mx(t+\Delta t)\int_T x(t)\varphi_k(t)\,dt =$$

$$= \sqrt{\lambda_k}\int_T r(t+\Delta t,\, t)\varphi_k(t)\,dt, \qquad (118)$$

and since when

$$0 \leqslant \tau \leqslant T \quad \int_T r(t,\, \tau)\varphi_k(t)\,dt = \lambda_k\varphi_k(\tau), \qquad (119)$$

it is natual to put

$$Ma_k x(t+\Delta t) = \frac{\varphi_k^*(t+\Delta t)}{\sqrt{\lambda_k}}. \qquad (120)$$

Here the $\phi_k^*(t+\Delta T)$ are the eigenfunctions, extended to the point t+ Δ t, of the integral equation with kernel r(t,τ). Thus, the best predicted value can be computed from the formula

$$x^*(t+\Delta t) = \sum_{k=1}^{\infty} a_k \frac{\varphi_k^*(t+\Delta t)}{\sqrt{\lambda_k}}. \qquad (121)$$

This representation was proposed by Karhunen [56]. It is based on the definition of the best predicted value as the point of space $L_2(X)$ closest to x(t+ Δ t).

The method of characteristic components

The processes under investigation may often contain natural components whose values greatly facilitate the solution of the prediction problem.

For a wide class of processes, the realizations (or selection functions) can be represented by small numbers of "characteristic components". These components are determined by the physical nature of the object which generates the process.

Under such conditions, the use of a large number of selection functions is neither necessary nor desirable. Furthermore, it is the components, and not the correlation function, which characterize the process.

In these cases, Wiener prediction is no longer acceptable.

Let us consider a method for predicting nonstationary random processes proposed by E.D.Farmer [45].

Determination of the characteristic components

Let us assume that $x_m(t)$ $(m=1,2,3,...M)$ are M selection functions of a nonstationary random process. It is required to determine the characteristic (in some sense) components of the process.

One of the simple methods of determining the first component consists of finding a function $\varphi_1(t)$, a scalar multiplier $\sqrt{\lambda}_1$ and a sequence of coefficients a_{m1} such that the functions $\sqrt{\lambda}_1 a_{m1}\varphi_1(t)$, $(m=1,2,...,M)$ are approximately, in the sense of the minimum mean-squares, equal to the selection functions $x_{m(t)}$. The error for the m-th realization has the form

$$e_m(t) = x_m(t) - \sqrt{\lambda_1}\, a_{m1}\, \varphi_1(t). \qquad (122)$$

The mean-square error, averaged over time and the set, is

$$\bar{\varepsilon}^2 = \frac{1}{MT} \sum_{m=1}^{M} \int_0^T [x_m - \sqrt{\lambda_1}\, a_{m1}\, \varphi_1(t)]^2\, dt. \qquad (123)$$

The constants $\sqrt{\lambda_1}$ and a_{m1} and the function $\varphi_1(t)$ can be selected so that this error is minimal. A minimum is achieved when the variation ε^2 with respect to the variation of the first order of the quantities $\sqrt{\lambda_1}a_{m1}$ and the function $\varphi_1(t)$ is equal to zero, i.e. when

$$\begin{gathered}
\int_0^T \varphi_1(t)\, x_m(t)\, dt = \sqrt{\lambda_1}\, a_{m1} \int_0^T \varphi_1^2(t)\, dt, \quad m = 1, 2, \ldots, M, \\
\sum_{m=1}^{M} a_{m1}\, x_m(t) = \sum_{m=1}^{M} \sqrt{\lambda_1}\, a_{m1}\, \varphi_1(t), \quad 0 < t < T. \qquad (124)
\end{gathered}$$

Without harming generality, the function $\varphi_1(t)$ and the vector a_{m1} can be normed so that we have the relationships:

$$\begin{gathered}
\int_0^T \varphi_1^2(t)\, dt = 1, \\
\sum_{m=1}^{M} a_{m1}^2 = M.
\end{gathered}$$

The conditions for a minimum then take the form

$$\left.\begin{aligned}
a_{m1} &= \sqrt{\lambda_1} \int_0^T \varphi_1(t)\, x_m(t)\, dt \\
\sqrt{\lambda}\varphi_1(t) &= \frac{1}{M} \sum_{m=1}^{M} a_{m1}\, x_m(t)
\end{aligned}\right\}. \qquad (125)$$

Elimination of a_{m1} from these two relations gives

$$\frac{1}{M} \sum_{m=1}^{M} \int_0^T \varphi_1(\tau)\, x_m(\tau)\, d\tau\, x_m(t) = \lambda_1 \varphi_1(t)$$

or

$$\int_0^T R(t, \tau)\varphi_1(\tau)\,d\tau = \lambda_1\varphi_1(t), \qquad (126)$$

where $R(t, \tau)$ is the correlation function formed by averaging the M selection functions

$$R(t, \tau) = \frac{1}{M}\sum_{m=1}^{M} x_m(t)\,x_m(\tau). \qquad (127)$$

It can be seen from (126) that $\phi_1(t)$ is an eigenfunction of the modified Wiener-Hopf equation.

Combining equations (123) and (126), we obtain an expression for the minimum mean-square error

$$\overline{\varepsilon^2_{\min}} = \frac{1}{T}\left[\int_0^T R(t, \tau)\,d\tau - \lambda_1\right]. \qquad (128)$$

The error is minimal if λ_1 is the greatest or dominant eigenvalue and, consequently, $\phi_1(t)$ is the dominant eigenfunction.

The second component can be determined if we require that $\sqrt{\lambda_2} a_{m2}\ {}_2(t)$ be the greatest mean-square approximation. It follows from this that λ_2 must be the second eigenvalue in magnitude, and that $\phi_2(t)$ must be the corresponding eigenfunction. By continuing this reasoning, we obtain an expansion of the function in the form of (114).

The autocorrelation function $R(\tau, \tau')$ can also be expanded in terms of the characteristic components of the process

$$R(\tau, \tau') = \sum_{k=1}^{\infty} \lambda_k \varphi_k(\tau)\varphi_k(\tau'). \qquad (129)$$

We can also show that all eigenvalues are either positive or equal to zero.

An important property of the expansion of $x_m(t)$ consists of the fact that if the series is limited to K terms, the mean-square error integrated with respect to time is

$$\overline{\varepsilon^2}T=\int_0^T R(\tau,\ \tau')\,d\tau'-\sum_{k=1}^{K}\lambda_k. \qquad (130)$$

The expression

$$\int_0^T R(\tau,\ \tau')\,d\tau'$$

is equal to the mean energy of the process in the interval (0,T). Taking (129) into account, this energy can be expressed in the form

$$\int_0^T R(\tau,\ \tau')\,d\tau'=\sum_{k=1}^{\infty}\lambda_k. \qquad (131)$$

Combining (130) and (131), we obtain

$$\overline{\varepsilon^2 T}=\sum_{k=k+1}^{\infty}\lambda_k \qquad (132)$$

Thus, the integral of the mean-square error is equal to the sum of the omitted eigenvalues; the number of terms necessary for achieving the given accuracy can be easily found from (130) and (132). The eigenvalue λ_k, which is essentially nonnegative, can be interpreted as the energy connected with the k-th component of the process.

The weighting function of a Wiener anticipator can be described by the comp nents $\varphi_k(t)$ Let us assume that it is required that the anticipator evaluate the input selection function at the instant $T=t+\Delta t$ from the input selection function known in the interval $(0,T_0)$, where $T_0 < T$. The weighting function satisfies the relationship

$$\int_0^{T_0} R(\tau, \tau') g(T_0, \tau') d\tau' = R(T, \tau), \quad 0 < \tau < T_0. \quad (129)$$

If $R(\tau, \tau')$ is expanded in series (129), the integral equation will take the form T_0

$$\sum_k \lambda_k \varphi(\tau) \int_0^{T_0} g(T_0, \tau) \varphi_k(\tau) d\tau = \sum_k \lambda_k \varphi_k(T) \varphi_k(\tau). \quad (133)$$

Multiplication by $\phi_k'(t)$ and integration with respect to τ in the interval $(0, T_0)$ gives

$$\sum_{k'} A_{kk'} \lambda_{k'} g_{k'} = \sum_{k'} A_{kk'} \varphi_{k'}(T) \lambda_{k'}, \quad (134)$$

where

$$A_{kk'} = \int_0^{T_0} \varphi_k(\tau) \varphi_k'(\tau) d\tau. \quad (135)$$

Since the function defined by the series

$$\sum_k g_k \varphi_k(\tau),$$

must be equal to zero in the interval $T_0 < \tau < T$,

$$\sum_{k'} B_{kk'} g_{k'} = 0, \quad (136)$$

where

$$B_{kk'} = \int_0^T \varphi_k(\tau) \cdot \varphi_{k'}(\tau) d\tau. \quad (137)$$

The matrices A and B with elements $A_{kk'}$ and$B_{kk'}$, respectively, are idempotent, i.e. satisfy the equations

Since the functions of the componсnts are orthonormal in the whole interval (0,T),

$$\left.\begin{array}{l} A+B=I \\ AB=BA=0 \end{array}\right\}, \qquad (139)$$

where I is a unit matrix. The mean-square prediction error in equation (123) expressed by the characteristic components is:

$$\overline{\varepsilon^2(T_0)} = \sum_{k=1}^{\infty} \lambda_k [g_k - \varphi_k(T)]^2. \qquad (140)$$

This error is minimal if g_k satisfies (134) and (136), i.e. when

$$\left.\begin{array}{l} \sum\limits_{k'} A_{kk'}\lambda_{k'} g_{k'} = \sum\limits_{k'} A_{kk'}\lambda_{k'}\varphi_{k'}(T) \\ \sum\limits_{k'} B_{kk'} g_{k'} = 0 \end{array}\right\}. \qquad (141)$$

It follows from (138) and (139) that the sum of the ranks of A and B is equal to the rank of the unit matrix. Hence, in (141) there are as many independent equations as there unknown g_k, and the solution is unique. As a whole, these equation are completely equivalent to the Wiener-Hopf equation.

Prediction of a process from its characteristic components

The problem of prediction is the problem of estimating the values of a selection function x(t) of a nonstationary process in the interval $T_0 < t \leq T$ from the known values of x(t) in the interval $0 \leqslant t \leqslant T_0$.

It was established in (131) that if a process is represented in the form of a combination of its characteristic components, the integral of the mean-square error is

$$\overline{\varepsilon^2}T = \int_0^T R(\tau, \tau')\,d\tau' - \sum_{k=1}^{K} \lambda_k .$$

Corresponding to this accuracy, the selection function x(t) can be written in the form

$$x(t) = \sum_{1}^{K} c_k \varphi_k(t), \quad 0 < t < T, \qquad (142)$$

where the c_k are constant coefficients.

This expansion is valid in the whole interval (0,T), including that part of it in which x(t) needs to be predicted. Hence the prediction problem is reduced to the problem of determining the sequence of coefficients c_k. This method of prediction automatically describes the selection function as a combination of its characteristic components.

One of the methods of finding the coefficients c_k consists of satisfying the following condition: expansion (142) must be the best mean-square approximation to the function x(t) in the interval $0 \leqslant t \leqslant T_0$ in which this function is known. Here the coefficients are determined from the system of linear equations

$$\sum_{k'=1}^{K} c_{k'} \int_0^{T_0} \varphi_k(\tau)\varphi_{k'}(\tau)\,d\tau = \int_0^{T_0} x(\tau)\varphi_k(\tau)\,d\tau. \qquad (143)$$

Let us consider an example.

Use of the method of characteristic components for predicting changes in the load on electric power stations

Prediction of electric power consumption is required in order to organize the future operation of electric power stations and to ensure future breakdown-free operation of electric-power supply networks.

For prediction purposes, the curves of the daily change in the load must be divided into section corresponding to periods of several hours. These periods include the interval $T_0 < t \leqslant T$ and the immediately preceding segment $0 \leqslant t \leqslant T_0$.

The load depends on the weather conditions, the consumption of power by industrial enterprises, radio, television, etc. The most important influence here is the weather. The load x_{mn} in the m-th section for the n-th instant can be written in the form

$$x_{mn} = \alpha_n + f_1(T_m)\beta_n + f_2(L_m)\gamma_n + f_3(W_m)\delta_n + \dots \quad (144)$$

In (144), $f_1(T_m)$, $f_2(L_m)$ and $f_3(W_m)$ respectively denote functions of the temperature T_m, the light intensity L_m, and the wind velocity W_m. The quantity α_n is the basic load. The factors $\beta_n, \gamma_n, \delta_n, \dots$ take into account the effect of weather parameters changing with time. In accordance with (144), every load vector is linearly dependent on the vectors α, β, γ, δ, ... If the load is represented by K terms in the form of a combination of its characteristic components, x_{mn} can be written in the form

$$x_{mn} = \sum_{k=1}^{K} c_{mk}\varphi_{kn}. \quad (145)$$

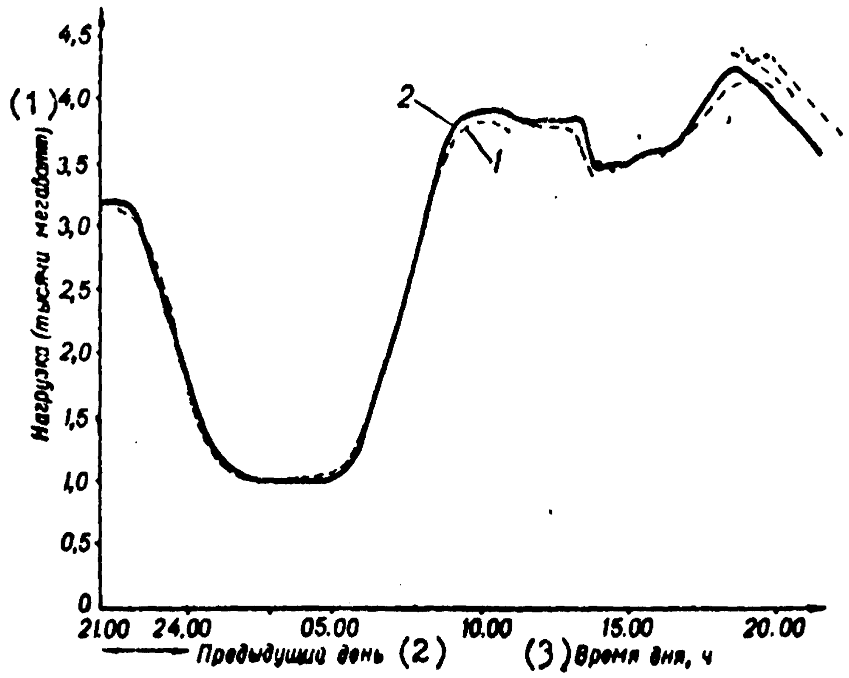


Figure 38. Curves of the actual and predicted load for 27 November.

Key: 1) Load (thousands of megawatts); 2) preceding day; 3) time of day, o'clock.

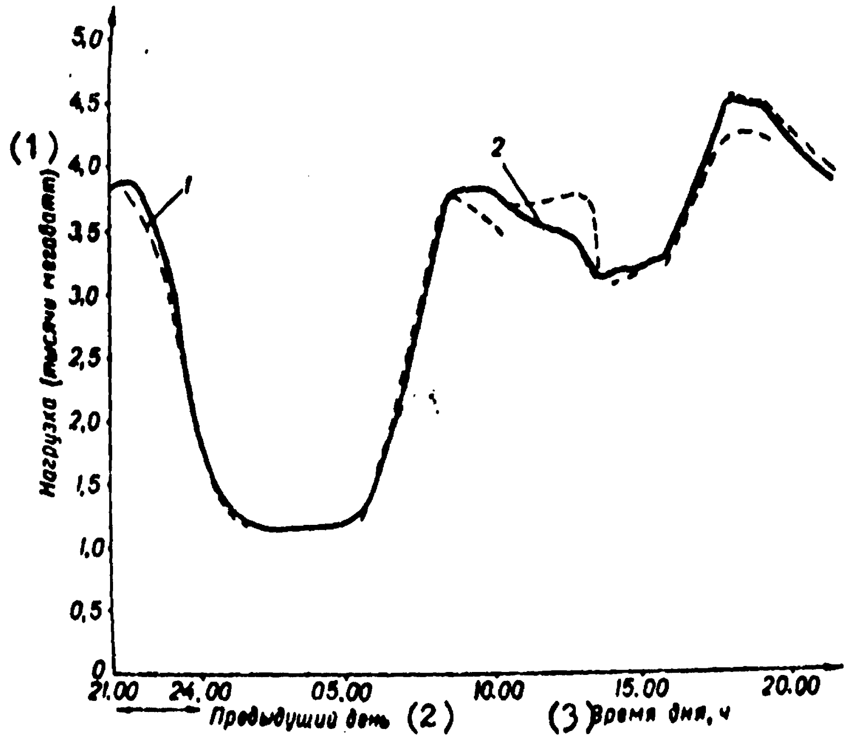


Figure 39. Curves of the actual and predicted load for 28 November.

Key: see Figure 38.

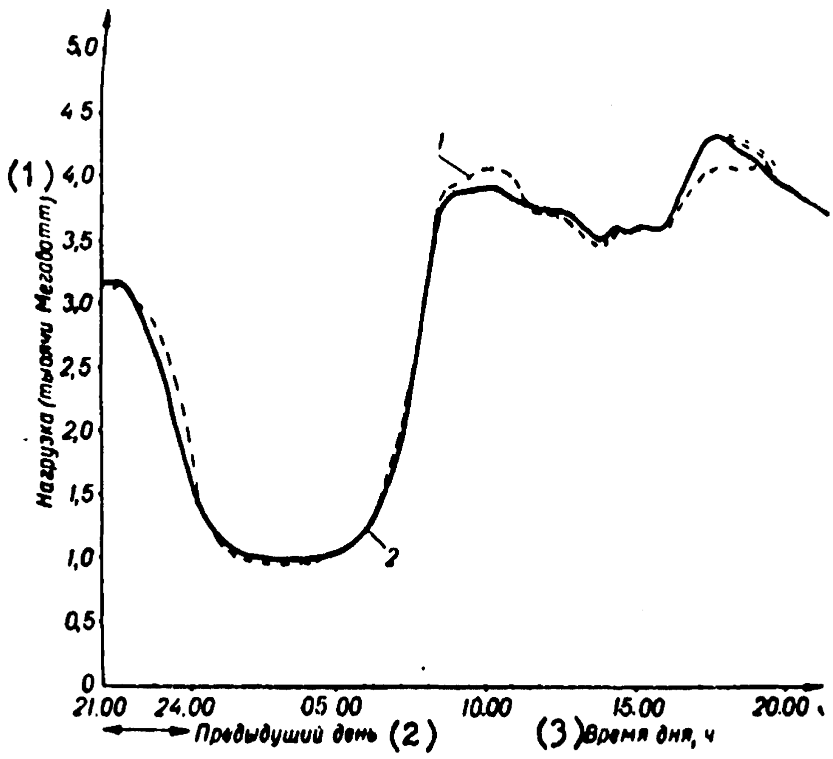


Figure 40. Curves of the actual and predicted load for 4 December.
Key: see Figure 38.

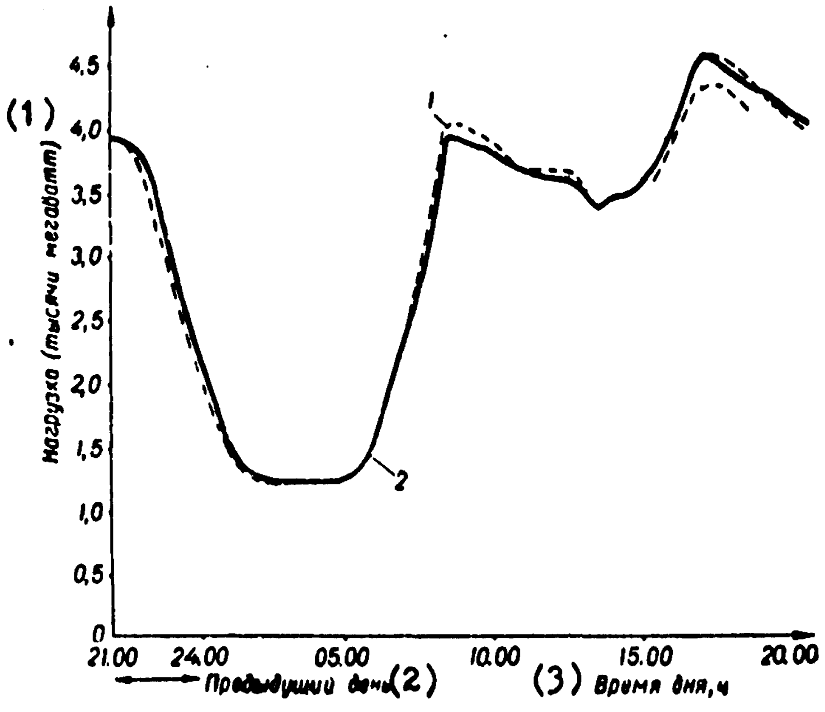


Figure 41. Curves of the actual and predicted load for 5 December.
Key: see Figure 38.

The expansion error will not change if the vector of the components is replaced by linearly independent combinations.

Since the vector minimizes the expansion error, the k-dimensional manifold formed by the vectors of the components and their linear combinations will, in comparison with other k-dimensional manifolds, give a minimal error. $\alpha, \beta, \gamma, \delta$, ... in equation (144) can be consdered as vectors belonging to the manifold of the vectors of the components. Thus, the components describe the basic tendencies of the load under mean weather conditions for the registration period. The weighting factors c_{mk} are functions of the weather parameters relating to the m-th part of the day.

Figures 38-41 show curves of the actual and predicted loads for two days in November and two days in December of 1961. The load was predicted for an interval of 8 hours over a large region with maximum consumption of 5000 Mwatt. The characteristic components were computed from data for the 20 preceding days. Each of the predicted curves was computed during the half-hour before it began.

The results obtained for the morning peak are in good agreement with the prediction made by the control center using weather data.

The prediction of the evening peak was less exact on the average [45]

The combined method of predicting nonstationary random processes

There is a wide class of real nonstationary random processes which can be represented in the form of some nonrandom function of time and a stationary random function z(t)

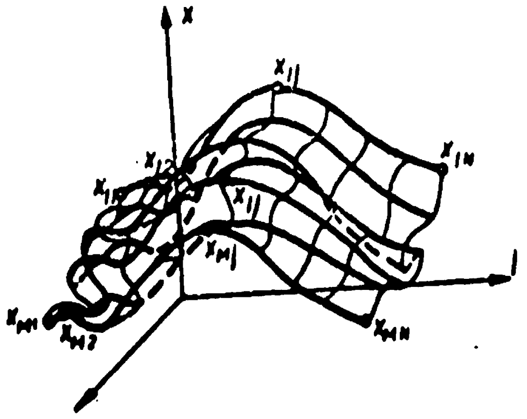


Figure 42. Nonstationary random process.

additively superimposed on it [39]:

$$x(t) = \psi(t) + z(t). \qquad (146)$$

Annual precipitation in a given region, consumption of materials and spare parts at an enterprise during an accounting period with a fixed plan, daily consumption of electric power, the variations in the temperature of patients identically suffering the same disease-- all these are examples of nonstationary processes which can, with greater or lesser accuracy, be described by expression (146).

Below we propose a method which makes it possible to solve the problems of predicting such processes.

Let M realizations of a nonstationary random process be known. For ease in solving the prediction problem on a digital computer, let us represent every realization in the form of a discrete sequence of values $\{x_j\}$, $j=1,2,\dots,N$ (Fig.42). From realization to realization, the values x_j undergo changes which are described by a stationary random function.

If we denote the j-th value of the i-th realization by x_{ij}, all the elements x_{ij} can be written in the form of a rectangular random matrix

$$X = (x_{ij}) \quad \begin{matrix} i = 1, 2, \ldots, M, \\ j = 1, 2, \ldots, N. \end{matrix}$$

Every column vector of this matrix is a stationary sequence of values $\{x_j\}$, and every row vector is one realization of the nonstationary random process.

In addition, from a set of N values of a realization with number M+1, let some number of values n be known. Let us denote these values by $x_{M+1,j}$.

The prediction problem consists of evaluating the values x^*_j $(j \in N-n)$ from the known values x_{ij} $(i=1,2,\ldots N; j=1,2,\ldots,M)$ and x_j $(j \in n)$.

Let us write the predicted realization (row vector) in the form

$$x^*_{M+1} = \sum_{k=1}^{K} c_k F_k. \qquad (147)$$

In expression (147), let us call the F_k component vectors.

It is required to find values for the F_k and coefficients c_k such that the vector x^*_{M+1} with components $x^*_{M+1,j}$ will approximate to x_{M+1} for $j \in n$ in the best way in the sense of some criterion.

Taking the stationary nature of the changes in x_j from realization to realization into account, we can determine the $x^*_{M+1,j}$ by using the extended prediction operator (158):

$$x^*_{M+1,j} = \sum_{i=1}^{M} x_{ij} r_i + \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} x_{i_1 j} x_{i_2 j} r_{i_1 i_2} + \ldots . \qquad (148)$$

The terms of the extended prediction operator are the corresponding components F_{kj} of the component vectors F_k:

$$\left.\begin{array}{l}\sum\limits_{l=1}^{M} x_{lj} r_l = F_{1j}, \\ \sum\limits_{l_1=1}^{M} \sum\limits_{l_2=1}^{M} x_{l_1 j} x_{l_2 j} r_{l_1 l_2} = F_{2j}, \\ \dots\dots\dots\dots\dots \\ \sum\limits_{k=1}^{M} \cdots \sum\limits_{l_k=1}^{M} r_{(l_k)K} \prod\limits_{1}^{K} x_{l_k} = F_{kj}. \end{array}\right\} \qquad (149)$$

As the criterion for the best approximation of the predicted valueś to the actual ones, let us take the criterien of the minimum mean-square error:

$$\bar{\varepsilon}^2 = \frac{1}{n} \sum_{j=1}^{n} \left(x_j - \sum_{k=1}^{k} c_k F_{kj} \right)^2, \quad (j \in n). \qquad (150)$$

Then the coefficients c_k can be determined from the equations

$$\sum_{k=1}^{K} c_k \left(\sum_{j=1}^{n} F_{kj} F_{k'j} \right) = \sum_{j=1}^{n} F_{k'j} x_j, \qquad (151)$$

which are analogous to those given by Farmer for computing the coefficients with characteristic components [45]. The value of K determines the number of components F_k necessary for achieving the given accuracy in accordance with the given criterion.

Finding the coefficients c_k and substituting them into (147), we obtain the predicted values of $x^*_{M+1,j}$.

Thus, the values of F_{kj} obtained earlier by means of the extended prediction operator are in essence the predicted values of the M+1-th realization of the process computed from the known realizations (prehistory). The values $x^*_{m+1,j}$ are more accurate values of the predicted M+1-th realization, this increase in accuracy being achieved due to a certain number of known values of the realization.

Second modification of the combined method

The values of the $x^*_{M+1,j}$ can be determined by using the method of exponential smoothing:

$$x^*_{M+1}(t) = \bar{x}_M(t) + \frac{d\bar{x}_M(t)}{dt}\Delta t + \frac{1}{2}\cdot\frac{d^2\bar{x}_M(t)}{dt^2}\Delta t^2 + \ldots,$$

where the $\bar{x}_M(t)$ are the exponentially smoothed values of the function.

Then the components F_{kj} of the component vectors F_k can be written in the form:

$$\bar{x}_{Mj} = F_{1j},$$

$$\frac{d\bar{x}_{Mj}}{dt}\cdot\Delta t = F_{2j},$$

$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot$$

$$\frac{1}{k!}\ \frac{d^k x_{Mj}}{dt^k}\ \Delta t^k = F_{kj}.$$

If as a criterion we take the criterion of the minimum mean-square error, the coefficients c_k in expression (147) can be determined from system of equations (151).

By means of the proposed method, we can solve the problem of predicting the M+$\underline{l}$-th realization of a non-stationary process (for $\underline{l} > 1$). However, the prediction accuracy in realizing both the extended prediction operator and the method of exponential smoothing drops as Δt is increased. Hence there is a decrease in the accuracy of determination of the component vectors F_k for the M+1-th realization as the number 1 is increased.

In predicting by means of the combined method, the accuracy can be increased by realizing an algorithm for continuously computing the coefficients c_k. The algorithm consists of the fact that the coefficients c_k computed from the n known values of the M+1-th realization are used for determining the future values only until the n+1-th actual value of the M+1-th realization becomes known. After this, the coefficients c_k are recomputed taking into account the n+1-th actual value.

As an example of the application of the combined method for predicting nonstationary random processes, let us consider the problem of predicting the variations in the load of a power system. These changes are shown in 24-hour graphs. For convenience in applying the method, the function of the change in the load with time is given in the form of a discrete sequence of values. The time discreteness step is taken to be the usual one for power systems, i.e. 1 hour.

Let it be required to predict the change in the load on some definite day of the week, for example Saturday. As prehistory let us use the 24-hour graphs of the change in the load for the corresponding days of the week in the past, i.e. for several preceding Saturdays. From the data of the graphs making up the prehistory and using (149), we find the values of c_kF_k. Further, using (150) and (151), we obtain the predicted values of the 24-hour graph in which we are interested.

It should be noted that, in predicting the daily variations of the load on complete working days (from Tuesday to Friday), the prehistory need not be made up of graphs from days with the same name. But as far as days

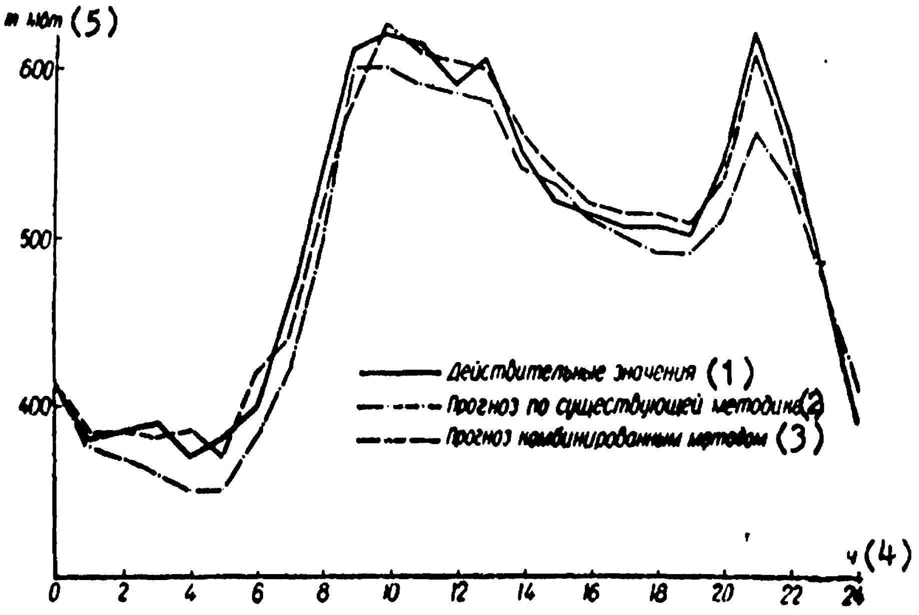


Figure 43. Variation in the load of a powar system.

Key: 1) Actual values; 2) prediction by existing methods; 3) prediction by the combined method; 4) o'clock; 5) Megawatts.

preceding days off, days off and holidays are concerned, the configurations of the 24-hour curves corresponding to them differ greatly from those of the graphs of complete working days.

For complete working days, the combined method makes it possible to obtain a prediction with a mean-square deviation at the peaks of $\sigma = \sqrt{\overline{\xi_i^2}} = 6-8$.

Figure 43 shows a graph for the load variation on one of the "nonstandard" days, Saturday. The predicted values obtained by means of the method described have been plotted on the graph together with the actual values. On the same graph are plotted the values obtained by predicting in accordance with the methods presently in use. The combined method gives a much higher prediction quality: $\sigma = 12$ as against $\sigma = 23$, obtained in predicting by the existing methods.

Let us consider one more example of the method's application.

Prediction of changes in intracranial pressure with a brain hemorrhage

Experimental reproduction of a cerebral hemorrhage (biological model)

The experimental neurosurgery laboratory of the Ukrainian Scientific-Research Institute for Neurosurgery conducted research on the variation of intracranial (fluid pressure in response to inhalation of carbon dioxide with an experimentally reproduced cerebral hemorrhage.

This reaction reflects the functional state of the brain vessels and makes it possible to make a judgement on the phase of development of the pathological process.

Into the intracranial cavity of an experimental animal there was introduced a needle, through which was passed blood from a femoral artery. Ruptures of vessels of different calibers were simulated by selection of needles with different diameters.

During the course of the experiments, inhalation tests of carbon dioxide were carried out at definite intervals. This test was chosen in view of the fact that carbon dioxide, which has a dilating effect on the small arteries and capillaries of the brain, has a different influence on the brain circulation, and hence on the nature of the change in the liquor pressure under normal conditions, in the presence of an intracranial seat of disturbance, when the initial pressure level is not yet changing (compensation stage) and on a background of incipient pressure change (subcompensation stage). The results of the experiment were recorded in the form of continuous curves on a polygraph and in numerical form by means of an electronic digital recording device in a complex with suitable pickups.

The change in the fluid pressure during one CO_2 test is a discrete sequence of the parameters of this test in which we are interested (see Fig. 45). The parameters were denoted as shown in Table 4.

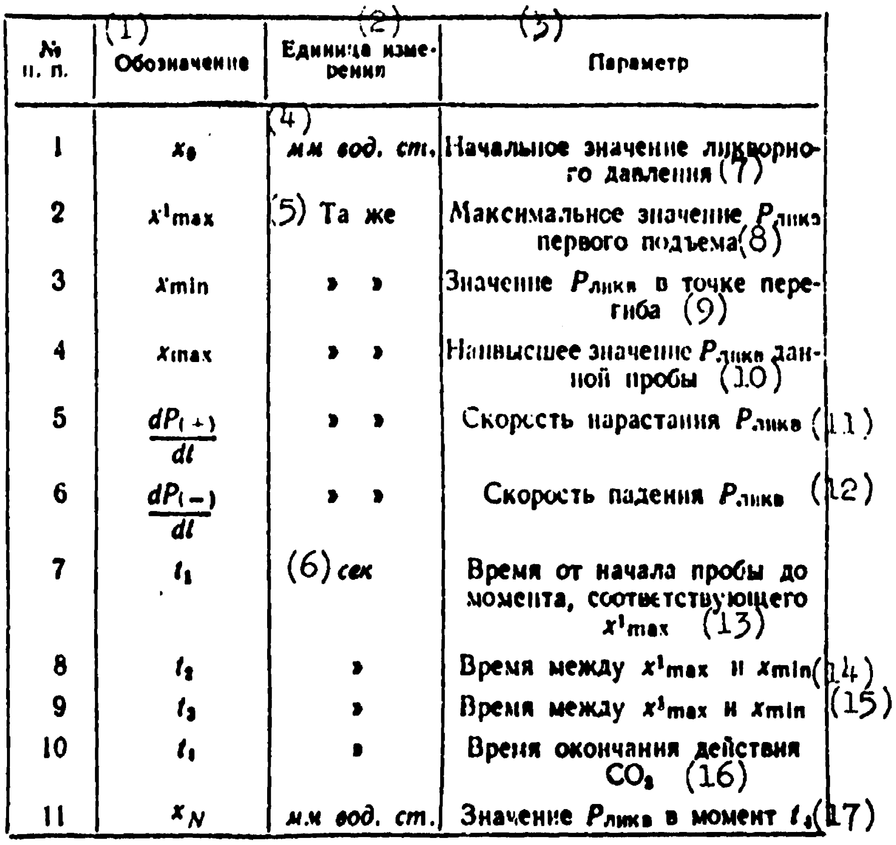
Prediction quality criterion

Ordinarily in filtration and prediction problems we use the criterion of the minimum mean-square error

$$M[x(t)-x^*(t)]^2=\overline{\varepsilon}^2_{\min}.$$

However, when this criterion is satisfied, there exists the probability that there will be individual large deviations of the predicted values from the actual ones. At the present time a number of rigorous criteria have been developed. These include the criterion of leas risk, the criterion of the minimum sum of the mean-square error and the dispersion, both taken with suitable weights, and several others. In solving the problem under consideration, we will require that the following additional condition be satisfied:

The absolute value of the deviation of the predicted value from the actual one must not exceed some previously chosen value Δ.



№ п. п.	(1) Обозначение	(2) Единица измерении	(3) Параметр
1	x_0	(4) *мм вод. ст.*	Начальное значение ликворного давления (7)
2	x^1_{max}	(5) Та же	Максимальное значение $P_{ликв}$ первого подъема (8)
3	x_{min}	» »	Значение $P_{ликв}$ в точке перегиба (9)
4	x_{max}	» »	Наивысшее значение $P_{ликв}$ данной пробы (10)
5	$\frac{dP_{(+)}}{dt}$	» »	Скорость нарастания $P_{ликв}$ (11)
6	$\frac{dP_{(-)}}{dt}$	» »	Скорость падения $P_{ликв}$ (12)
7	t_1	(6) *сек*	Время от начала пробы до момента, соответствующего x^1_{max} (13)
8	t_2	»	Время между x^1_{max} и x_{min} (14)
9	t_3	»	Время между x^1_{max} и x_{min} (15)
10	t_4	»	Время окончания действия CO_2 (16)
11	x_N	*мм вод. ст.*	Значение $P_{ликв}$ в момент t_4 (17)

Table 4.

Key: 1) Notation; 2) unit of measurement; 3) parameter; 4) mm water column; 5) the same; 6) sec; 7) initial value of liquor pressure; 8) maximum value of P_{flu} of first rise; 9) value of P_{flu} at point of inflection; 10) highest value of P_{flu} of given test; 11) rate of increase of P_{flu}; 12) rate of decrease of P_{flu}; 13) time from beginning of test to instant corresponding to x^1max; 14) time between x^1max and x_{min}; 15) time between x^1max and x_{min} /sic/; 16) time of completion of CO_2 operation; 17) value of P_{flu} at time t_4.

(1) Параметр обозначения	X_{M+1}	X^*_{M+1}	X_{M+2}	X^*_{M+2}	X_{M+3}	X^*_{M+3}	X_{M+4}	X^*_{M+4}
x_0	115	113	75	72	95	88	90	85
x_{max}	120	123	80	82	102	105	95	98
x_{min}	120	118	78	77	100	102	92	92
x_{max}	130	138	102	102	115	123	117	125
$\frac{dP(+)}{dt}$	0,277	0,444	0,778	0,889	0,556	1,007	0,556	0,666
$\frac{dP(-)}{dt}$	−0,044	−0,083	−0,222	−0,242	−0,111	−0,139	−0,139	−0,333
t_1	10,0	9,0	6,0	5,0	8,0	9,0	7,0	6,0
t_2	15,0	16,0	11,0	10,0	13,0	13,0	12,0	12,0
t_3	39,0	40,0	24,0	23,0	33,0	29,0	28,0	29,0
t_4	120	119	114	120	118	120	116	115
t_N	130	123	75	73	110	102	100	90

Table 5.

Key: 1) parameter notation.

Thus, the coefficients $r_{\{ik\}_K}$ in expression (148) are found from the condition

$$M(x_j - \overset{*}{x}_j)^2 = \overline{\varepsilon^2_{min}}$$

when

$$| x_j - \overset{*}{x}_j | \leqslant \Delta.$$

Solution and results

The problem of predicting CO_2-inhalation tests was programmed and solved on an electronic computer. Determining the parameters of the four predicted tests required about

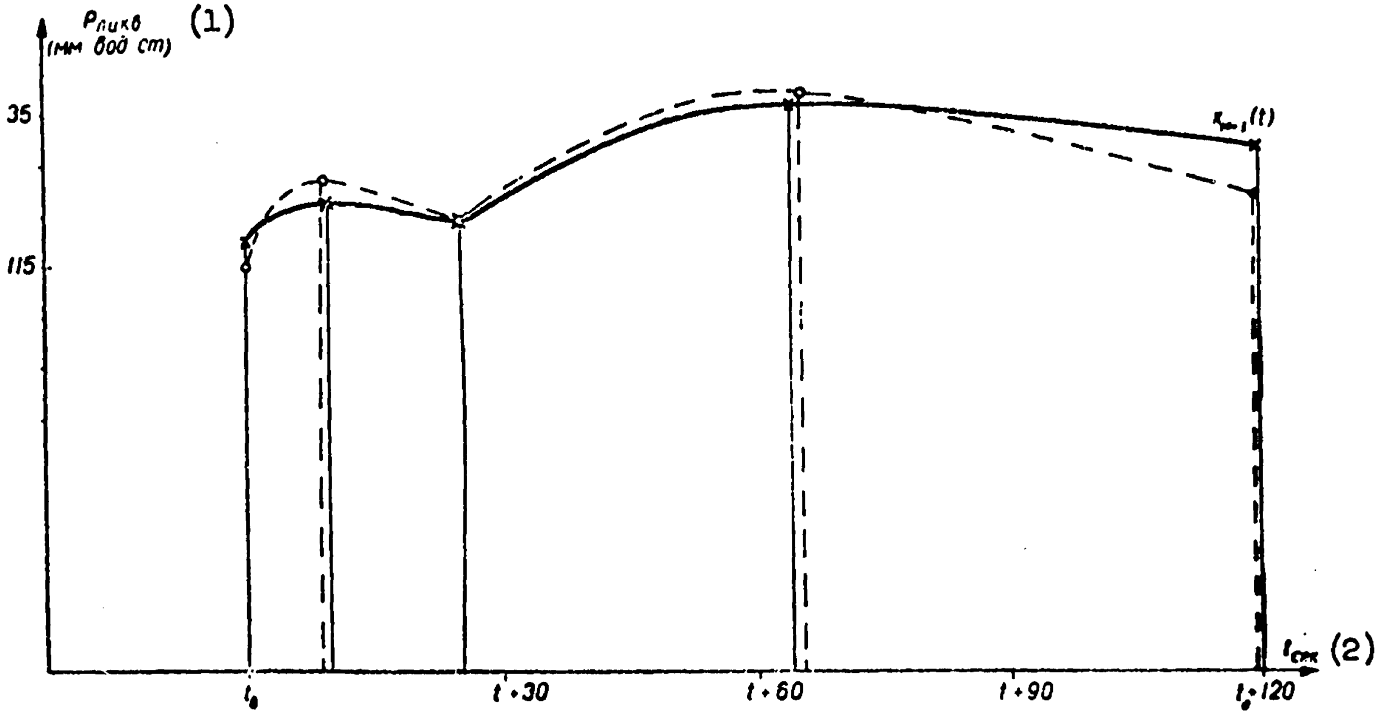


Fig. 44. Variation curve of fluid pressure. Test 1

Key: 1. P_{flu} (mm water column); 2. t_{sec}

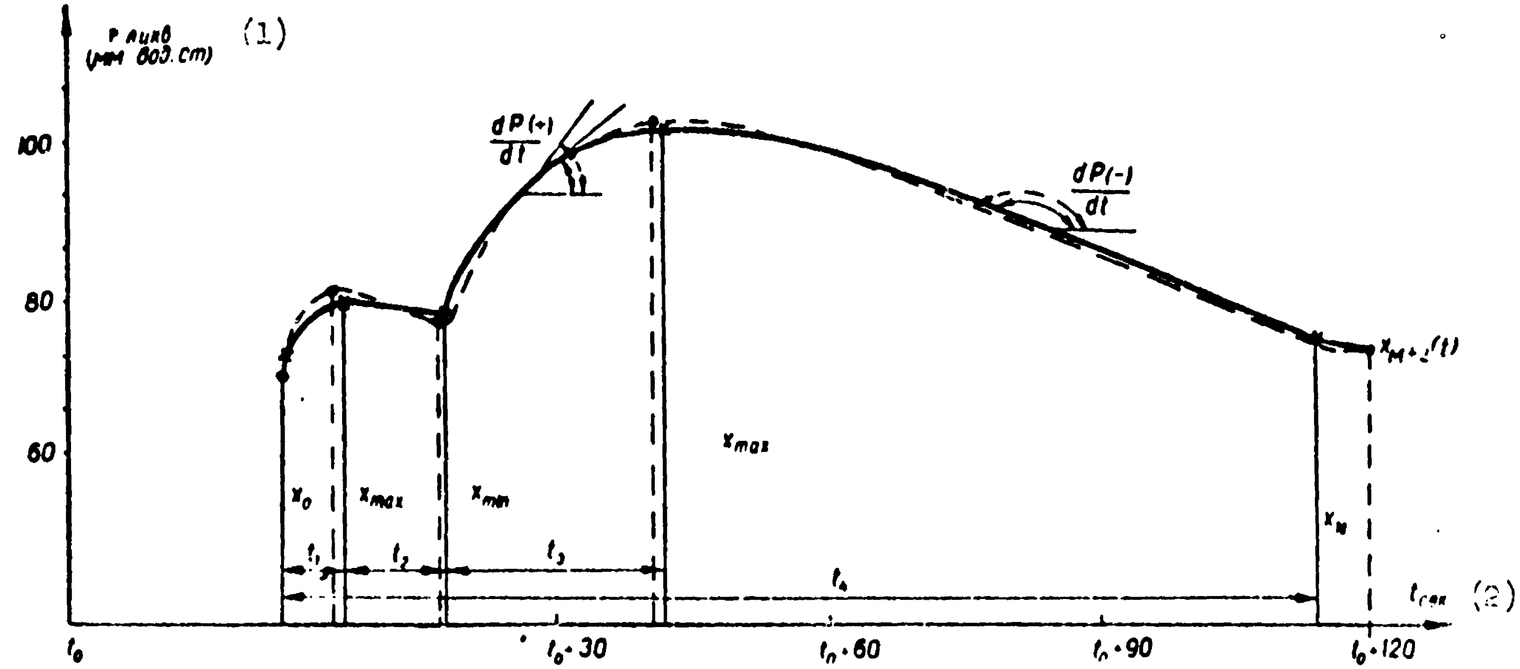


Fig. 45. Variation curve of fluid pressure. Test 2

Key: as in Fig. 44

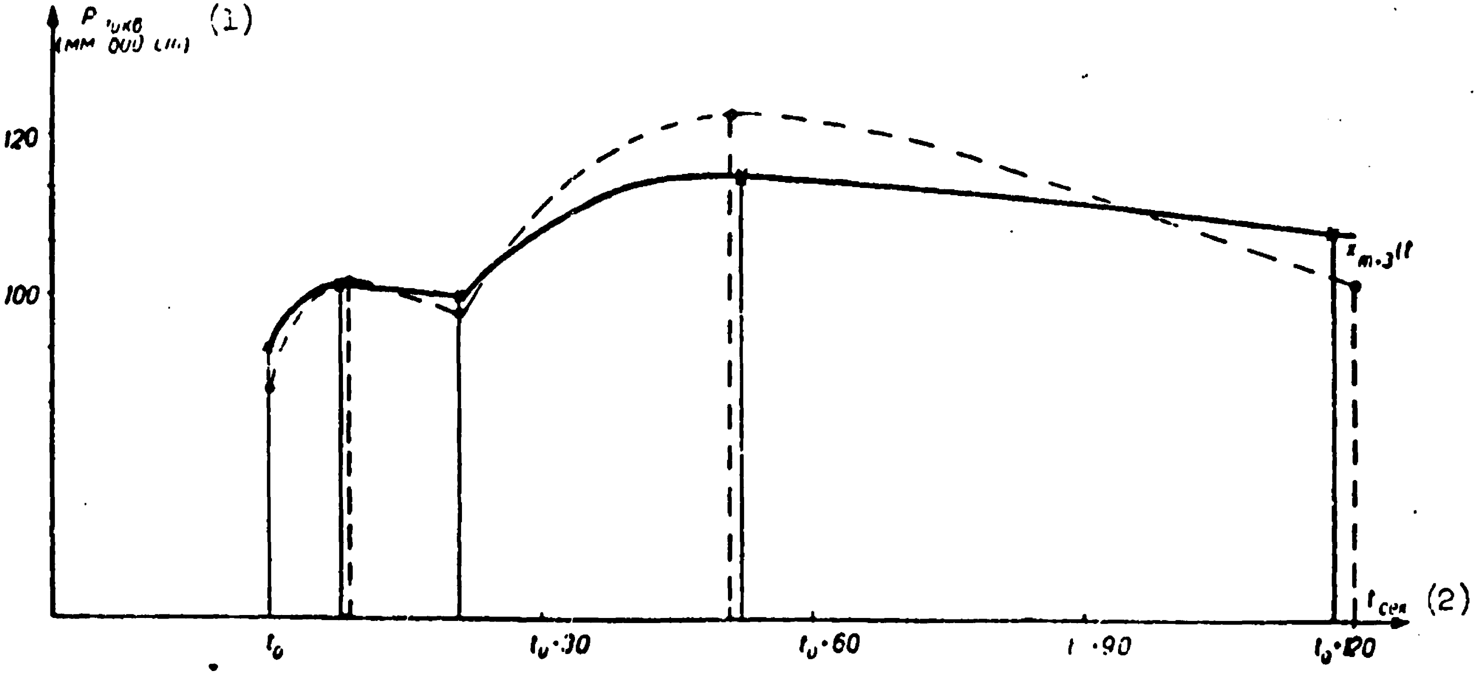


Fig. 46. Variation curve of fluid pressure. Test 3.
Key: as in Fig. 44

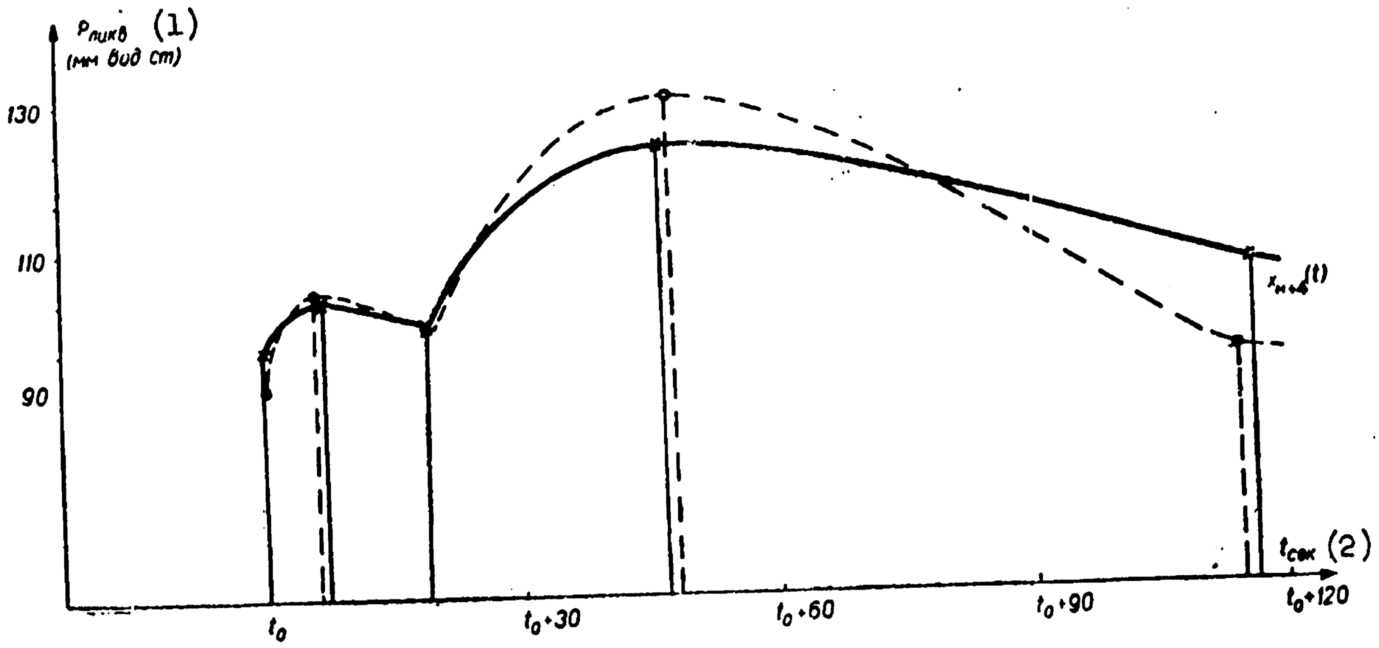


Fig. 47. Variation curve of fluid pressure. Test 4
Kay: as in Fig 44

six minutes of machine time (including print-out of the results) using a computer with a speed of the order of 3000 operations per second.

The results are shown in Table 5. Figures 44-47 show graphs of the actual and predicted curves for the variation of fluid pressure with inhalation of carbon dioxide.

The predicted curves give a good reflection of the qualitative aspect of the process. The accuracy fully corresponded to the requirements imposed on similar biological investigations.

Predictions of the development of processes with various acute and chronic illnesses can be used for diagnosis and prognosis in solving the problem of whether surgical intervention is timely and indicated.

Chapter 4

Recognition systems as predicting filters and regulators

During the past years, a number of scientists have proposed automatic devices which after a certain period of initial adjustment (learning) can quite accurately predict the course of various processes.

The material used to teach these devices is the past record of the process, the prehistory. The prediction quality of these devices can easily be evaluated if we compare their output signals with the actual val ues of the process being predicted. The circuit and parameters of the predicting device can be chosen so that the prediction accuracy increases in the course of time.

If this selection is made automatically by means of feedback, we obtain a self-adjusting predicting filter.

In realizing this or that prediction algorithm, the self-adjusting predicting filter automatically improves its structure and makes the values of its parameters more precise. This happens as a result of observation of the course of the process.

The system constantly takes into account newly arriving data on the course of the process and automatically makes the prediction more accurate.

A universal predicting filter with self-adjustment in the learning process was proposed by the English scientist Prof. D.Gabor [54]. This filter is based on a prediction algorithm which consits of finding the optimal weighting factors of the extended prediction operator.

The filter was designed in the formof an analog device using magnetic and piezoelectric multipliers. An example of the use of this device is the solution of the problem of predicting the amplitude of ocean waves. The prediction accuracy was of the order of several percent.

It was proposed to use the filter for predicting the indices of economic conditions, but the volume of filter input devices for this problem turned out to be insufficient. At the present time, experiments on a specialized analog device have ceased. The same algorithm is being realized on the fast-acting universal digital computer "Atlas" at London University.

In our experiments, we did not renounce the use of specialized self-learning filters, which is expedient only when the algorithm for self-adjustment of the parameters (coefficients) is greatly simplified. The basic proposal consists of transfer to the use of binary self-organizing recognition systems as self-learning predicting filters.

The most clear expression of the idea of self-organization is found in the works of F.Rosenblatt [59,a,b]. He proposed a statistical model of the brain which has the properties of learning and self-learning. This model was named the perceptron. This name then began to be applied not only to the model proposed by the author, but also to other analogous systems.

Perceptrons can independently, without the aid of man, recognize and classify input signals by attributes which were not previously given. The process of teaching a perceptron how to read letters was successfully demonstrated in June 1960.

Figure 48 shows a simplified block diagram of the "Perceptron". The letters or other images which the machine is to learn to recognize and classify are projected on a screen consisting of photocells. The photocells convert the images into a large number of electrical signals. Every photocell is randomly connected with a field of associating elements (cells). As a result of summation of the signals arriving from the associating elements, various slave elements are excited, and these indicate to which pattern the given image belongs.

In Rosenblatt's first perceptron, the field of photocells (about 400) was connected by random links to amplifiers, and then in the same way to servomotors.

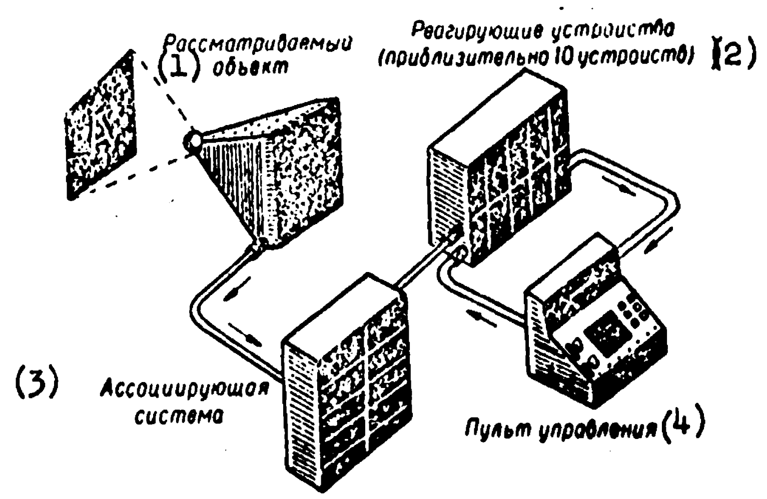


Figure 48. Simplified block diagram of the "Perceptron".

Key: 1) Object being considered; 2) reacting devices (approximately 10 devices); 3) associating system; 4) control desk.

The amplifiers are fed biases which can be changed either by the teacher, man, or a feedback pickup.

Let us consider the learning process of the perceptron. Let us assume that we want to teach it to distinguish the letters A and B; i.e. we want certain slave elements to operate when the letter A is projected on the field of the photocells, and others when B is projected. To do this we must "encourage", in the form of supply of suitable biases, those amplifiers which include the necessary outputs and hamper ("punish") the others which bring unnecessary outputs into operation. The encouragement and punishment laws may be different /59/.

The "self-learning" process of the perceptron occurs by a different method. With this method method, the biases are not changed by man, but arrive through feedback circuits from the outputs to the amplifiers.

The perceptron is called a statistical system, since it employs probabilistic inputs, and all its basic elements (pickups, associating elements and output elements) are connected by randomly selected links. If we use determinate pickups adjusted for reception of definite attributes and join the elements of the system by all possible connections, we obtain a system of the perceptron type. Such a system was designed by one of the authors. The circuit of the system is shown in Fig.49.

The essential distinguishing feature of this system is the presence in it of a separate positive-feedback circuit and a maximum-voltage indicator which indicates which of the groups of associating elements is giving the greatest voltage.

One group of associating elements corresponds to every pattern. The correct answer is given by the group at whose output (as a result of the "voting" of a large number of distributing elements) the greatest voltage is obtained. For the sake of simplicity, Fig.49 shows only three such groups. It shows the variant "with complete input information" with equal participation of the associat-

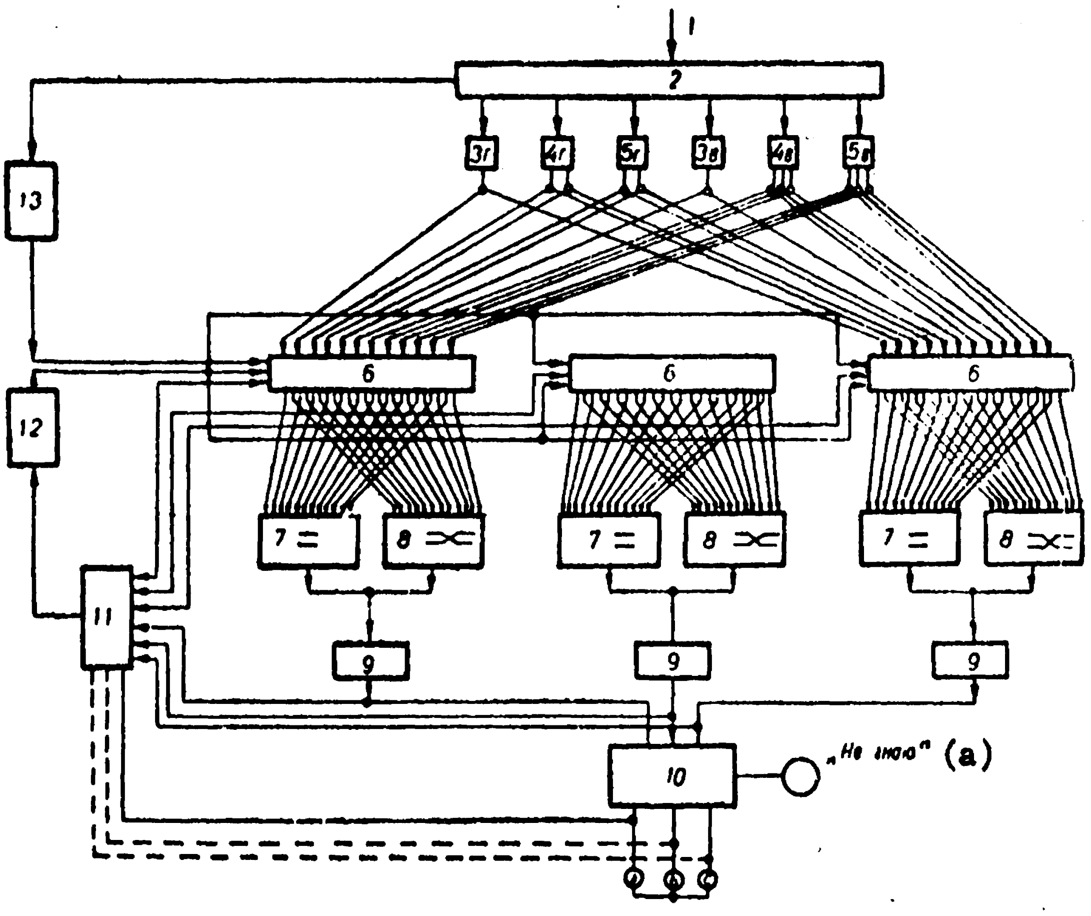


Fig. 49. Structural diagram of recognition system with positive feedback
1. pattern being distinguished; 2. input device; 3.- 5. pickups; 6. groups of associating elements; 7. direct amplifiers; 8. reversing amplifiers: 9. summators: 10. large-voltage indiicator [sic]; 11. control of order of self-learning; 12. positive self-learning feed; 13. open learning feed.

Key: a. "I don't know"

ing elements.

F.Rosenblatt has formulated two theorems which express the concept of self-organization. In accordance with these theorems, only an "infinite perceptron" having an infinite number of pickups can begin to act without initial organization. The greater the degree of initial organization, the smaller is the number of elements we can make do with, the cheaper is the perceptron.

Hence in designing practical circuits, it is desirable to proceed from some initial organization, although it is not necessary in principle.

In practice, the least initial information consists of the simple listing of the attributes of the input signals which can at any time be useful for distinguishing these signals without indicating to which signal they relate.

The above system (see Fig.49) was later improved and received the name of the "Alpha" system [21].

The basic advantages of binary systems in comparison with continuous ones, for example in comparison with the filter of Prof. A.[sic] Gabor, are the following:

1. Very little information about the process is required for the system to operate. It is only necessary to know whether the process index exceeds a definite level or not. It is not essential to assume normal distribution of attributes.

2. With a second positive feedback the system can automatically select the most useful initial data or attributes [21]. The most efficient use of the system capabilities leads to a great decrease in its volume. Furthermore, a recognition system can in principle be constructed so that signals will be emitted to the effect that the attributes being used are insufficient and that new data on the process are required. Below we will consider in detail

the circuit and operation algorithm of D.Gabor's analog self-adjusting filter and will then dwell on the predictive applications of the Alpha binary recognition system.

Universal predicting filter with self-adjustment in the learning process

The problem of synthesizing an optimal linear predicting filter was first formulated by Kolmogorov in 1942. As for the nonlinear filter, the opinion was expressed that formal solution of the problem would not have any practical importance. This is explained by the large number of computations and the huge volume of work collecting data. But from time to time scientists returned to the idea of developing such a filter, since the prediction of complex stochastic processes would have very great importance in solving important problems of control in production, of planning, economics and in other fields.

In 1954 the English scientist Prof. D.Gabor proposed the definition of a group of mathematical problems needing solution and investigation in developing an optimal filter. The filter must be the realization of an extremely flexible mathematical operator which takes into account the present and past values of any time function with which it operates. The parameters of the filter must constantly improve during the learning process. The device registers the values of the random sequences which must be filtered or predicted. The sequences, also called selections, must be sufficiently long to ensure complete representation of the function. This means that the selections must be of such a length that the statistical parameters computed from it can be assumed to be statistical parameters of the whole process.

The basic difficulty in the practical application of the Kolmogorov-Wiener theory of linear filtration and prediction lies in the fact that this theory considers signals in an unbounded frequency range.

In the solution of the problem of predicting by real devices with a bounded passband F, a continuously varying signal can be represented in the form of a discrete sequence of the values of this signal which follow at intervals of (1/2F) sec.

Extended prediction operator

The most general functional expression of the prehistory of a time function in a bounded frequency range is the following sequence of discrete values:

$$O[f(t)] = \sum_{i=0}^{N} f_i r_i + \sum_{i_1=0}^{N} \sum_{i_2=0}^{N} f_{i_1} f_{i_2} r_{i_1 i_2} + \dots \qquad (158)$$

The coefficients of this sequence are transfer functions which are only defined for integral values of the argument.

It is easy to see that the first sum is is fact the generalized linear predicting filter. The coefficient r_i determines the influence of the discrete value of the function f_i, this value corresponding to a moment i intervals earlier than the present moment.

The second sum is made up of pairwiseproducts of these discrete values, including squares. The coefficients $r_{i_1 i_2}$ determine the weights of the pairs of values at the instants i_1 and i_2 intervals earlier than the present instant, etc. These sums must include the whole prehistory.

If the prehistory is divided into N intervals, the operator contains N+1 terms of the first order (of the first sum), (1/2) (N+1) (N+2) terms of the second order, (1/6) (N+1) (N+2) (N+3) terms of the third order, etc. It is quite obvious that the number of terms of the operator increases sharply as the order increases.

Circuit of the predicting filter

A block diagram of the predicting filter is shown in Fig.50. The delay block 1 is a magnetic tape recording device with stepwise located heads.

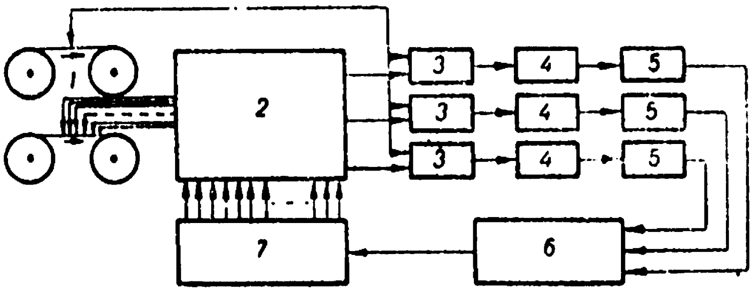


Figure 50. Block diagram of predicting filter with self-adjustment in the learning process: 1) magnetic tape recording device; 2) block of parameters being adjusted; 3) comparison block; 4) squaring devices; 5) integrators; 6) minimizing device; 7) block for adjusting variable parameters.

The sought function O*(t) is recorded on one of the tracks of the magnetic tape. In the case of prediction, this function is considered shifted forward by the head from the input tape.

The filter 2 itself consists of an arithmetical device and a set of parameters being adjusted. Potentiometers are used as the latter. The output signal of the filter and the sought function are regist red in the comparison block, which determines the prediction error. As in the case of linear filter theory, we take as our criterion that of the least mean-square error:

$$\overline{[0[f(t)]--0^*(t)]} = \bar{\varepsilon}^2_{\min}. \qquad (159)$$

The left-hand side of this expression is a positive definite quadratic form of the coefficients r. Hence a solution always exists. Furthermore, since the prediction is linear with respect to r, this solution is unique.

In the function of the arguments r, the mean-square error can be represented by a multidimensional elliptical paraboloid. Hence the use of any algorithm for decreasing $\bar{\varepsilon}^2$ must inevitably lead to achieving a minimum. This is shown in Fig. 51 for the case of one and two variables . Both parameters must be alternately adjusted.

Here the optimal coefficients can be determined from the expression

$$r_{opt} = \frac{1}{2} \cdot \frac{y_{-1} - y_{+1}}{y_{-1} + y_{+1} - 2y_0}.$$

But if we use differences, then

$$r_{opt} = \frac{1}{2} \cdot \frac{\Delta_{-1} - \Delta_{+1}}{\Delta_{-1} + \Delta_{+1}}$$

and

$$y_{\min} = y_0 - \frac{1}{2} r_{opt} \frac{1}{2} (\Delta_{-1} - \Delta_{+1}).$$

In order to increase the speed of action in the present predicting filter, the comparison blocks 3, the squaring devices r and the integrators 5 are thrice re-

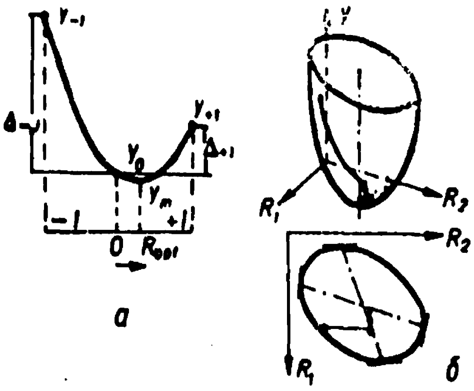


Figure 51. Optimization of adjustment of variable parameters: a) one variable parameter; b) two variable parameters.

peated; hence for every learning cycle the error is determined for three values of the selected parameter r_i: for r_i=0 and for the largest positive and negative values. On the basis of these values, the minimizing device 6 automatically determines the optimal value to which the parameter r_i and the quantity $\bar{\varepsilon}^2_{min}$ must be adjusted. The adjustmet block 7 establishes the necessary value of r_i. Then the learning cycle is repeated. As a rule, in the presence of M parameters to be adjusted, the number of learning cycles necessary to achieve $\bar{\varepsilon}^2_{min}$ is a quantity of the order of $M^2/2$. After this it is said that the filter has learned to predict the values of the given random sequence.

In other words, a predicting filter which has learned the best method (in the sense of $\bar{\varepsilon}^2$ = min) of predicting the values of the training sequences will operate analogously with other selection of the same stochastic sequence. This statement is based on the following assumptions:

1. The random processes under consideration are ergodic, i.e. retain the constancy of the statistical

parameters over any arbitrarily long sequences.

2. During the learning process, the machine determines all necessary statistical parameters and reproduces them.

The filter described above was used to carry out the following experiments:

1. Conversion of a sinusoidal signal into another sinusoidal signal shifted in phase and different in amplitude.

2. Filtering of a sinusoidal signal with a superimposed noise.

Prediction of the changes in the quality index of a product of a petrochemical enterprise

Let us consider an example of applying the Kolmogorov-Gabor predicting filter.

Figure 52 shows a diagram for automatic control of the fractionating column of one of the technical installations of a petroleum processing plant (AVT*installation).

The output product of this installation is direct-distillation benzine. A quality analyzer which automatically determines the temperature at end of boiling of the benzine is connected to the control circuit. In controlling the techinical device in accordance with the quality index of the output product, it is important to know not only the current values of the index and its deviations from the norm, but also the values of the index at some fu*ure times. This makes it possible to regulate the process with anticipation, taking into account possible future deviations. Such a method is known by the name of "anticipatory compensation". The predicting filter (see Fig.52) determines the future values of the quality index; these are used

*/atmospheric-vacuum pipe still/

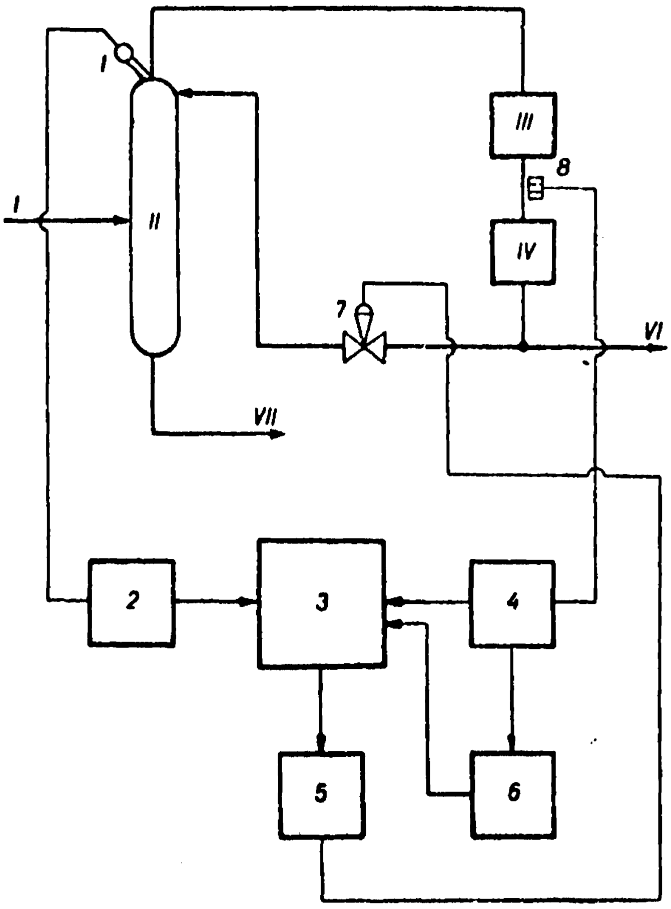


Figure 52. Circuit for automatic regulation of the fractionating column of an AVT installation:
I) raw material flow; II) fractionating column of AVT installation; III) condenser; IV) irrigation receptacle;
V) irrigation flow; VI) output product (direct distillation benzine; VII) debenzined product; 1) thermocouple; 2)potentiometer; 3) summing device; 4) quality analyzer; 5) regulator; 6) predicting filter; 7) regulating valve; 8) sampling point for product analysis.

as initial information in an automatic control system for the technical device. If there is a universal digital computer in the production process control system, the functions of the predicting filter can in many cases be conveniently assigned to this computer.

Modelling the predicting filter on a computer

Let us write the operation algorithm of the predicting filter. The learning process of the operator is:

$$\overset{1}{\downarrow} T\alpha \overset{1}{\uparrow} \overset{1}{\uparrow} \overset{2}{\downarrow} M'\beta \overset{2}{\uparrow} \overset{3}{\downarrow} \overset{5}{\downarrow} M'\gamma \overset{3}{\uparrow} \Sigma E p \overset{4}{\uparrow} S \overset{4}{\downarrow} R\omega \overset{5}{\uparrow} . \qquad (160)$$

In expression (160):

T is the operator of emission of the delayed values f_i, $(i=0,1,\dots,k)$, of the learning sequence;

α is a logical condition considered to be satisfied when $i=k$;

M^f is the operator for computing the products of the delayed values;

is a logical condition considered to be satisfied when all products of the delayed values have been obtained;

M^r is the operator for multiplying the values f_i and the products $f_{i_1} f_{i_2}$, $f_{i_1} f_{i_2} f_{i_3}$, etc by the corresponding weighting factors;

γ is a logical condition considered to be satisfied when all components of the prediction operator $O\ f(t)$ have been obtained;

Σ is the computation operator of $O\ f(t)$;

E is the operator for computing the mean-square error;

p is a logical condition considered to be satisfied if

$$\overline{[O[f(t)] - f^*(t)]^2} \approx \overline{\varepsilon^2_{\min}};$$

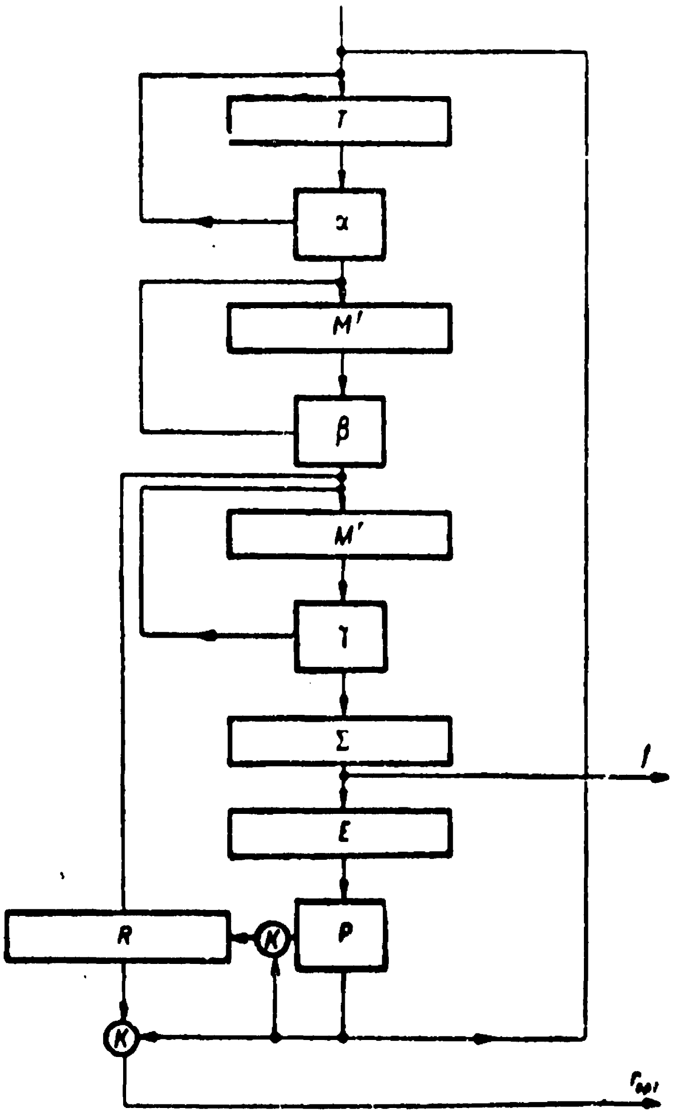


Figure 53. Flow chart of the program for modelling the predicting filter on a computer.

S is the end of the learning;

R is the operator for computing the new weighting factors r_{i_k};

ω is an identically false condition.

The prediction process is:

$$\overset{1}{\downarrow} T\alpha \overset{1}{\uparrow} \overset{2}{\downarrow} M'\beta \overset{2}{\uparrow} \overset{3}{\downarrow} M'_{opt}\gamma \overset{3}{\uparrow} \Sigma S. \qquad (161)$$

In expression (161), $M^{r_{opt}}$ is the operator for multiplying the products $f_{i_1} f_{i_2}$, $f_{i_1} f_{i_2} f_{i_3}$, etc by the optimal coefficients r_{opt} obtained as a result of the learning of the operator (160).

In the problem under consideration, the algorithm was realized on a universal digital computer. Figure 53 shows the flow chart of the modellling program. The actual and predicted values of the index T_{kk}°C are shown in Table 6. Figure 54 shows graphs of the actual and predicted values of T_{kk}. The operator learned with different numbers of values in the prehistory (k=2,3,4,5).

Influence of the prehistory length on the prediction quality

If we analyze the example just considered, and also the examples analyzed earlier in Chapter 2, it becomes obvious that an essential influence on the prediction quality is exerted by the number k of known values f_i (i=1,2,...,k) which participate in the computation of the operator $O[f(t)]$.

For clarity let us consider some examples.Figure 55 shows the prediction quality criterion as a function of the number of known values of the process being predicted taking part in the computation of the operator $O[f(t)]$. The cri-

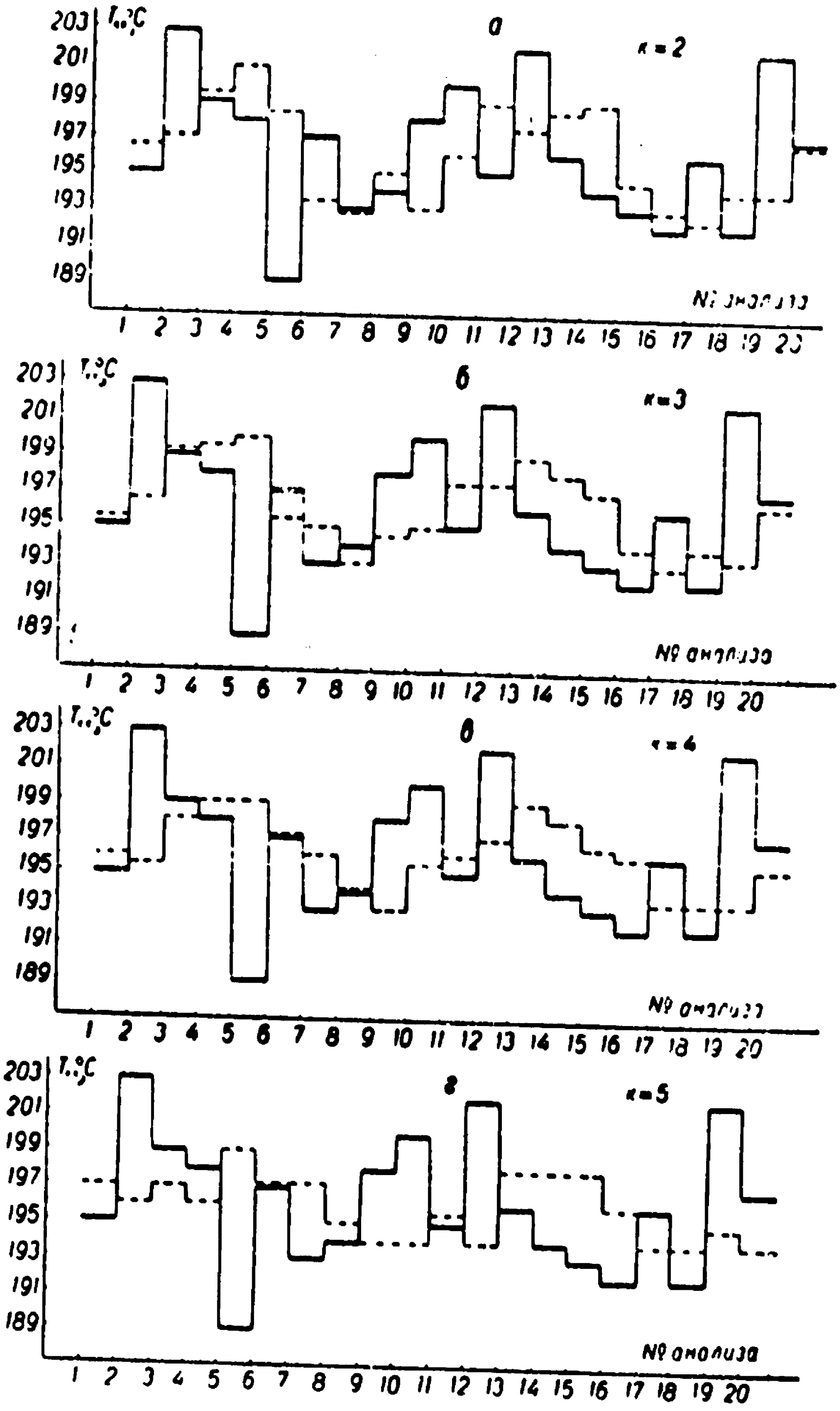


Figure 54. Prediction of the temperature at the end of boiling of direct distillation benzine.

Key: 1) Number of analysis.

Действительные значения (1	(2) Предсказанные значения			
	κ=2	κ=3	κ=4	κ=5
196	196,5	195,5	196	197
203	197	196,5	195,5	196
199	199,5	199	198	197
198	201	199,5	199	196
189	198,6	200	199	199
197	193,5	195,5	197	197
193	193	195	196	197
194	195	193	194	195
198	193	194,5	193	194
200	196	195	195,5	194
195	199	197,5	196	195,5
202	197,5	197,5	197	194
196	198,5	199	199	198
194	199	198	198	198
193	194,5	197	196,5	198
192	193	194	196	196
196	192,5	193	193,5	194
192	194	194	193,5	194
202	194	193,5	193,5	195
197	197	196,5	195,5	194

Table 6.

Key: 1) Actual values; 2) predicted values.

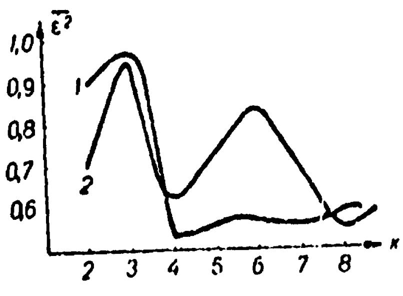


Figure 55. Mean-square error as a function of the length of the prehistory.

terion, as before, is the minimum of the mean-square error. It is evident from the graphs that, for the actual processes being considered by us, the prediction quality is notu a monotonic function of the length of the prehistory, as could have been expected in the case of stationary random processes. Thus, it is obvious from the graph in Fig. 55 that when k=2,4,8 the mean square error is less than when k=3,6. Similar processes relate to the class of periodically correlated (or almostt periodically correlated) random processes. Ye.G.Gladishev [10] will acquaint the reqder with the theory of these processes.

Thus it is ebvious that the quality of prediction of actual random processes essentially depends on the length of the prehistory. The selection of κ is the primary problem, on whose solution depends the accuracy and reliability of prediction.

Simple first-derivativd predicting device

The basic shortcoming characteristic of the Kolmogorov-Gabor predicting filter is the growth in complexity of the computations as the accuracy is increased. he complexity of circuit realization limits the sphere of application of test devices. This is especially true of the use of predicting filters in control systems.

In practice very often some lowering of the prediction accuracy is introduced for the sake of simplifying the construction of the system. In such cases, it is expedient to use simple anticipatory devices in which prediction is basdd on determination of the first derivative at a current point.

In order to raise the accuracy of prediction of such devices, the second derivative can also be used, but in practice it is most often enough to stick to the first. Raising the accuracy by using the second derivative leads to great complication of the system.

The accuracy of prediction by such devices drops in the region where the first derivative of the function being predicted with respect to time changes sign. It is obvious that, in order to raise the prediction accuracy, the prediction interval should be decreased as the frequency spectrum of the function being predicted is expanded. It is known that the width of the frequency spectrum of any process is sufficiently completely characterized by its autocorellation function. Let us compute the normalized autocorrelation function $\rho_x(\tau)$ at one point for some definite value of τ. If we now select the anticipation time Δt according to the formula

$$\Delta t = k\rho_x(\tau), \qquad (162)$$

we can select a relation between the magnitude of the prediction interval and the slope of the autocorrelation function for which the prediction error will not exceed an assigned value.

The predicted value is

$$x(t + \Delta t) = x(t) + \Delta x(t), \qquad (163)$$

where x(t) is the current value of the function.

The proportionality factor k in expression (162) is equal to the maximum anticipation time.

The slower the process changes with time, the greater is the anticipation time. The normalized autocorrelation function of such processes approaches unity.

Rapidly varying processes, which are characterized by a wide frequency spectrum, will have a very small anticipation time, since their normalized autocorrelation function approaches zero for the same .

By decreasing the anticipation time we can achieve any arbitrarily high accuracy. However, practical problems make it necessary not only to strive to raise the accuracy in all ways, but also to strive to increase the anticipation time where possible.

To evaluate the prediction quality, it was proposed to take the criterion of the maximum of the sum composed of a value inversely proportional to the prediction error and the ratio of the actual anticipation time to the maximum one /6/:

$$\varphi = \frac{1}{\varepsilon^2} + \frac{\Delta t}{k}. \qquad (164)$$

It is obvious that the function $\phi = F(\Delta t)$ has an extremum, since as the anticipation time increases, the mean-square error continuously increases. Using this dependency property (164), we can construct a self-adjusting system which by changing the anticipation time continuously would find the maximum of the prediction quality index. For this purpose we can use either a search extremal system or a non-search differential extremal system. The principle of operation of differential systems was described in V.I.Vasil'yev's book <u>Differential Regulating Systems</u> /7/.

Prediction of the contour of a river bottom

The method described above can be illustrated by the example of predicting the contour of a river bottom. The solution of this problem is very important, for example, for optimal control in seafaring.

The contour of a section of river bottom and its predicted values are shown in Fig.56. Computations have

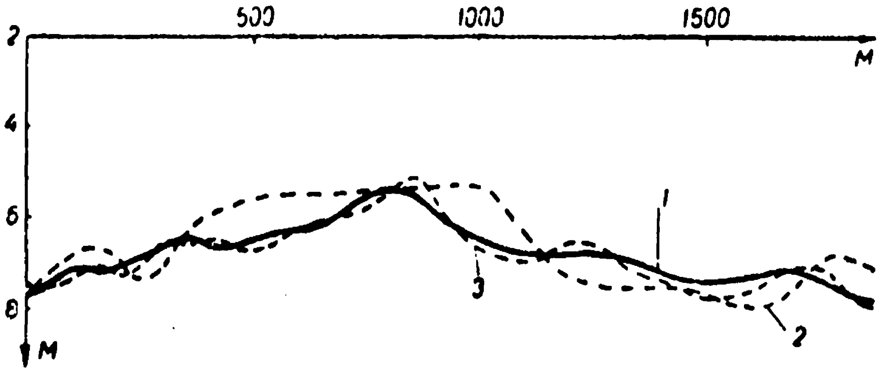


Figure 56. Prediction of the contour of a river bottom: 1) Contour; 2) prediction with constant interval; 3) prediction with variable interval.

shown that the introduction of a varible prediction interval sharply raises the prediction accuracy (Fig.56).

Alpha Recognition System as a Predicting Filter

Prediction of discrete results. If a number of similar cyclic processes differ only in the initial conditions, and they then proceed under approximately constant conditions, the results of these processes can be distributed according to the forms of the initial conditions with the aim of predicting the results of similar processes in the future. To carry out this generalization of experience we can use any self-learning recognition system with classification of patterns by output quantities [21, p 302]. The property of a recognition system of dividing a set of images into classes of patterns can be used to classify initial conditions (attributes) according to results.

Thus, V.L.Brailovskiy and A.L.Lunts in experiments on the prediction of the result of treatment of burns used twelve initial attributes (wound area, burn localization, degree of burn, age of patient, accompanying diseases, complications, data of medical analyses, etc.) Prediction was made of the outcome of the treatment, i.e. cure or death.

Prediction of continuous processes. As is known, continuous quantities can be approximately replaced by a number of discrete quantities. Hence, recognition systems can also be used to predict continous processes. An example is prediction of the duration of service of transistors from the form of the variation curves of currents observed during 10 min. The time, a continous quantity, is divided into a number of segments, and hence it is required to teach the system to predict the segment number.

According to Lyapunov's theorem, when the number of components of random quantities is increased, the distribution law of their sum tends to the normal probability distribution. Hence this explains the fact that very many stationary processes in nature have a normal distribution. The most general formula for predicting the future value of a random time function is Kolomogorov's formula

$$g[f(t)] = r_0 + \sum_0^n r_n f_n + \sum_0^n \sum_0^n f_{n_1} f_{n_2} r_{n_1 n_2} +$$

$$+ \sum_0^n \sum_0^n \sum_0^n f_{n_1} f_{n_2} f_{n_3} r_{n_1 n_2 n_3} + \dots,$$

where $g[f(t)]$ is the future, predicted value of the function; f_{n_1}, f_{n_2} are the values of this function in the past.

The first sum is a linear function with a constant

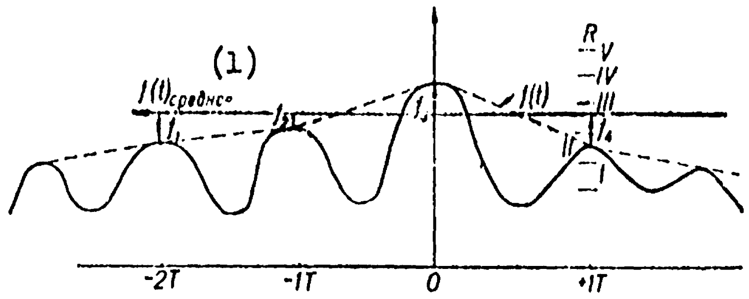


Figure 57. Problem of predicting the amplitude of the following wave from the values of the amplitudes of the three preceding waves. R is the number of levels of discretization of the output quantity.

Key: 1) $f(t)_{\text{mean}}$.

transfer function; the second is a quadratic filter; the third is a cubic one; etc. The coefficients r_n determine the influence (weight) of each term of the formula on the predicted value of the function. Let us clarify the formula by an example (Fig.57). For predicting the future value of the function from three data on the course of variation of this function in the past (prehistory) we obtain a formula in the form of a polynomial:

$$g[f(t)] = f_1r_1 + f_2r_2 + f_3r_3 + f_1^2r_4 + f_2^2r_5 + f_3^2r_6 + f_1f_2r_7 + f_1f_3r_8 + f_2f_3r_9 + f_1^3r_{10} + f_2^3r_{11} + f_3^3r_{12} + f_1^2f_2r_{13} + f_1^2f_3r_{14} + f_2^2f_1r_{15} + f_2^2f_3r_{16} + f_3^2f_1r_{17} + f_3^2f_2r_{18} + f_1f_2f_3r_{19},$$

where g is the value of the function f(t) at the future time +T; f_1 is the value of the function f(t) at time -2T; f_2 is the value of the function f(t) at time -T; f_3 is the is the value of the function f(t) at the given time 0; $r_1, r_2, \ldots, r_{19}$ are the influence (weighting) factors of each term.

The self-learning of the recognition system used as a predicting filter consists of determining the values of the influence factors on the basis of the data of some learning sequence. This process of self-establishmet of the coefficients for stationary random processes must be carried out once; the longer the learning sequence of data, the more accurate the prediction. For almost stationary processes it is better to limit ourselves to a local selection, a small learning sequence, and to have the self-learning of the coefficients carried out continuously only according to the latest data on the process. It is possible to determine the optimal duration of the memory of the system. The closer the process is to a stationary one, the greater is this duration. For extremely non-stationary processes, the optimal duration is small, and sufficient prediction accuracy is not ensured. In these cases we must have recourse to other prediction methods (for example, to the method of separating the periodic components of the components [45] or the combined method [31]).

Operation algorithm of the Alpha recognition system

Let us recall in brief the operation algorithm of the Alpha recognition system [21]. A sample circuit of the system is shown in Fig.58. In the given application the pickups of attributes by observation of the instantaneous changes in the function produce several functions of these quantities

$$x_1, x_2, \ldots, x_n,$$

which in the theory of pattern recognition are usually called attributes. The aggregate of attibutes forms an input "image" vector, or a "representation point",

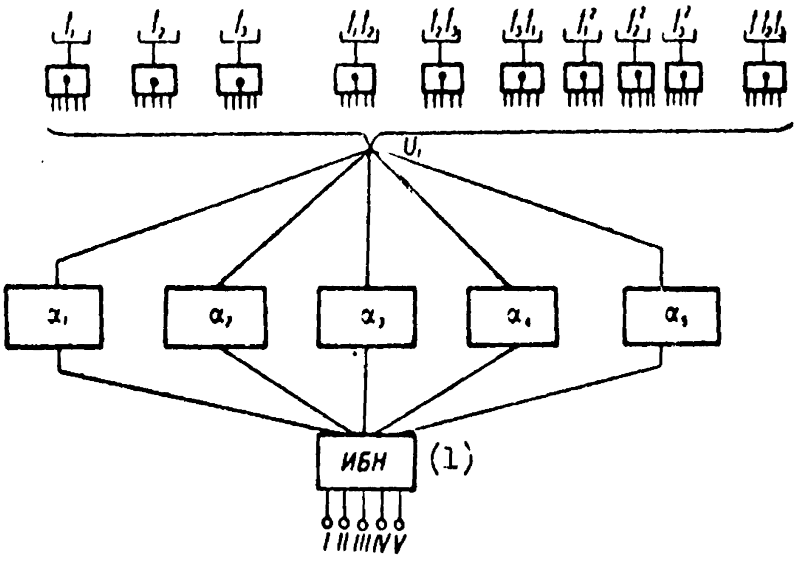


Figure 58. Alpha recognition system as a Kolmogorov-Gabor predicting filter.

Key: 1) Highest-voltage indicator.

$$v_i(x_1, x_2, \ldots, x_n).$$

The vector v_i is fed to the inputs of groups of associating cells (flip-flops or relays); these form the scalar products of the input vector v_i by several internal vectors $\alpha_k(r_1, r_2, r_3, \ldots, r_n)$ recorded in each of these groups; these vectors are called prototypes or poles,

$$\Sigma_1 = (\alpha_1 v_i),\ \Sigma_2 = (\alpha_2 v_i),\ \ldots,\ \Sigma_n = (\alpha_n v_i).$$

The number of groups is equal to the number of patterns being distinguished, i.e. divisions of the quantity being predicted.

The scalar products are fed to a comparator (a highest-voltage indicator, HVI), as a result of which the system selects the largest of them and thereby indicates the pattern (division of quantity being predicted). In self-learning the position of the poles is changed by feedbacks /21/. The learning of the system consists of an expedient selection of group poles.

The encouragement law of the Alpha system is expressed by the equation

$$\alpha_{m+1} = \frac{v_m + k_1 v_{m-1} + k_2 v_{m-2} + \dots + k_{m-1} v_1}{1 + k_1 + k_2 + \dots + k_{m-1}},$$

where m is the number of operations of the given group; k(m) is the "forgetting function" of the previous displays.

When k(m)=0 we obtain the "trustworthy" feedback

$$\alpha_{m+1} = v_m,$$

for which the corresponding prototype vector immediately takes the value of the vector of the latest image display.

When k(m)=1 we obtain an "averaging" feedback, where the prototype vector terminus is held at the "center of gravity" of the region of the given pattern:

$$\alpha_{m+1} = \frac{v_m + v_{m-1} + \dots + v_1}{m}.$$

It is possible to apply the exponential law of decrease of the coefficients as the given output operates:

$$\alpha_{m+1} = \alpha_m + (v_m - \alpha_m)\,\delta, \text{where } 0 \leqslant \delta \leqslant 1.$$

When δ=1, we obtain the encouragement law of "trustworthy" feedback.

An exponential encouragement law is most easily achieved in systems with continuous associating cells.

It is possible to have one more form ofthe "encouragement" law which is convenient for realization in relay systems. In this form the pole of the group being learned is displaced by one interval in the direction of the output image after every operation of the output corresponding to it. This interval may be constant or may decrease according to an exponential law.

The Alpha system generalizes individual images into regions-- patterns. Let us explain the generalization property by means of the example of an Alpha system having three groups of relays in all whose state at the n-th cycle of operation of the system is characterized by the three poles--

$$\alpha_{1n},\ \alpha_{2n},\ \alpha_{3n}.$$

Pickups feed the vector v_n to the input of the system. Then at the outputs of the group we obtain three voltages:

$$\Sigma_{1n} = (\alpha_{1n} v_n),\quad \Sigma_{2n} = (\alpha_{2n} v_n),\quad \Sigma_{3n} = (\alpha_{3n} v_n).$$

The LVI operates according to the following algorithm:

a) if $\Sigma_{1n} > \Sigma_{2n}$ and $\Sigma_{1n} > \Sigma_{3n}$, output 1 operates and the first group is relearned;

b) if $\Sigma_{2n} > \Sigma_{1n}$ and $\Sigma_{2n} > \Sigma_{3n}$, output 2 opera es and the second group is relearned;

c) if $\Sigma_{3n} > \Sigma_{1n}$ and $\Sigma_{3n} > \Sigma_{2n}$, output 3 operates and the third group is relearned.

All states satisfying the first condition will be related by the system to the first situation, all those satisfying the second condition to the second situation, and all those satisfying the third condition to the third situation. This constitutes the generalization.

In all cases the action of positive feedback only strengthens these inequalities and leads to still greater "anchoring" of the outputs:

$$\Sigma_{1(n+1)} = (\alpha_{1(n+1)} v_n) \geqslant \Sigma_{1n}; \quad \Sigma_{2(n+1)} = (\alpha_{2(n+1)} v_n) \geqslant \Sigma_{2n};$$
$$\Sigma_{3(n+1)} = (\alpha_{3(n+1)} v_n) \geqslant \Sigma_{3n}.$$

Let us recall that other recognition systems use other pickups, prototypes, comparison algorithms and encouragement laws. The system selecting the minimum square distance between the representing point and the poles uses the comparison algorithm

$$\Sigma_s = \sum_{m=1}^{K} (v_{lm} - \alpha_{sm})^2$$

with a following procedure for minimizing Σ_3 (as distinct from the procedure for seeking the maximum value of the scalar product in the Alpha system.

The Alpha system compares the distance between the poles and the image in Hemming space, and the system with selection of the minimum square of the distance compares them in ordinary euclidean space.

Prototypes need not be vectors; they can be replaced by the boundary equations of the regions of individual patterns in the multidimensional space of attributes, correlated pretetypes, etc.

Self-learning of a prediction filter by the method of regression analysis

The mean-square prediction error is defined by the expression

$$\Delta = \frac{1}{n} \sum^{n} (g_s - g)^2,$$

where g_0 is the actual value of the function (in the future);
g is the predicted value of the function.

It is a positive definite quadratic form of the influence factors. An error minimum always exists, and since the equation is linear with respect to r, it is an unique minimum. Manifold-correlation (regression) analysis makes it possible to select the influence factors in such a way that we obtain the minimum mean-square error of prediction.

For example, introducing for three values the notation

$$\begin{gathered} f_1 = x_1,\ f_2 = x_2,\ f_3 = x_3,\ f_1^2 = x_4,\ f_2^2 = x_5,\ f_3^2 = x_6, \\ f_1 f_2 = x_7,\ f_1 f_3 = x_8,\ f_2 f_3 = x_9,\ f_1^3 = x_{10},\ f_2^3 = x_{11},\ f_3^3 = x_{12}, \\ f_1^2 f_2 = x_{13},\ f_1^2 f_3 = x_{14},\ f_2^2 f_1 = x_{15},\ f_2^2 f_3 = x_{16},\ f_3^2 f_1 = x_{17}, \\ f_3^2 f_2 = x_{18},\ f_1 f_2 f_3 = x_{19}, \end{gathered}$$

we obtain the prediction equation in the linear form

$$\begin{gathered} g = r_0 + r_1 x_1 + r_2 x_2 + r_3 x_3 + r_4 x_4 + r_5 x_5 + r_6 x_6 + r_7 x_7 + \\ + r_8 x_8 + r_9 x_9 + r_{10} x_{10} + r_{11} x_{11} + r_{12} x_{12} + r_{13} x_{13} + r_{14} x_{14} + \\ + r_{15} x_{15} + r_{16} x_{16} + r_{17} x_{17} + r_{18} x_{18} + r_{19} x_{19}. \end{gathered}$$

The expression for the mean-square error takes the form

$$\Delta = \frac{1}{n} \sum_1^n (g_0 - r_0 - r_1 x_1 - r_2 x_2 - r_3 x_3 - \ldots - r_{19} x_{19})^2.$$

Since we wish to determine the minimum mean-square error, we find the values of the twenty partial derivatives and equate them to zero

$$\frac{\partial \Delta}{\partial r_0} = 0,\ \frac{\partial \Delta}{\partial r_1} = 0,\ \frac{\partial \Delta}{\partial r_2} = 0,\ \ldots,\ \frac{\partial \Delta}{\partial r_{19}} = 0.$$

These equations are the basic calculational equations ("normal regression equations"). In expanded form we obtain (lines above the correlation coefficients denote the operation of averaging over all data of the learning sequence):

$$
\begin{gathered}
r_0 + r_1\bar{x}_1 + r_2\bar{x}_2 + r_3\bar{x}_3 + \ldots + r_{10}\bar{x}_{10} = \bar{g}_0 ,\\
r_0\bar{x}_1 + r_1\overline{x_1^2} + r_2\overline{x_2x_1} + r_3\overline{x_1x_3} + \ldots + r_{10}\overline{x_1x_{10}} = \overline{g_0x_1} ,\\
r_0\bar{x}_2 + r_1\overline{x_1x_2} + r_2\overline{x_2^2} + r_3\overline{x_2x_3} + \ldots + r_{10}\overline{x_2x_{10}} = \overline{g_0x_2} ,\\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
r_0\bar{x}_{10} + r_1\overline{x_1x_{10}} + r_2\overline{x_2x_{10}} + \ldots + r_{10}\overline{x_{10}^2} = \overline{g_0x_{10}} .
\end{gathered}
$$

For a single-valued solution the number of equations can not be less than the number of unknowns. Hence the minimum duration of the learning sequence is equal to the number of terms of Kolmogorov's formula. If the prehistory covers N intervals, Kolmogorov's formula contains N+1 terms of the first order, (1/2)(N+1)(N+2) second-order terms, (1/6)(N+1)(N+2)(N+3) third-order terms, etc. For example, when N=3, there must be not less than 20 measurements. Every measurement must include three values of the function from the prehistory and a corresponding actual value of the function following these values (see Fig.57). In practice, the length of the learning sequence must be increased by a factor of five to ten in comparison with the minimum length. This makes it possible to eliminate the influence of function measurement inaccuracy. Here the number of equations does not increase, but there is an increase in the volume of computations of the correlation coefficients in these equations.

We have shown the algorithm for applying regression analysis towards determining the coefficients of Kolmogorov's formula chiefly in order to estimate the volume of the necessary computational work and to indicate the length of the learning sequence of the initial data.

Self-learning of a predicting filter by Gabor's iteration algorithm

Professor D.Gabor /54/ proposed an iteration algorithm based, like the method of regression analysis, on the search for the minimum mean-square error, i.e. giving the same result in a somewhat different way. The mean-square error as a function of the influence factors is a multi-dimensional elliptical paraboloid whose vertex is the aim of the search. The search for the minimum error can be carried out by various methods: the Gauss-Seidel method (i.e. successive variation of coefficients), the gradient method or method of steepest descent, etc. In his self-adjusting filter, Professor Gabor used an extrapolation method for the minimum search. The position of the minimum is computed from three points of a parabola. Deduction of the extrapolation formula is simple (see Fig.51). From the condition of the minimum mean-square error we find the optimal value of some coefficient

$$\Delta = a_0 + a_1 r_i + a_2 r'^2_i,$$

$$\frac{d\Delta}{dr_i} = 0 \quad \text{or} \quad r_{i\,opt} = -\frac{1}{2} \cdot \frac{a_1}{a_2}.$$

The parabolic equation for any three chosen points gives three calculational equations:

$$\Delta' = a_0 + a_1 r_i' + a_2 r_i'^2,$$

$$\Delta'' = a_0 + a_1 r_i'' + a_2 r_i''^2,$$

$$\Delta''' = a_0 + a_1 r_i''' + a_2 r_i'''^2.$$

By determining from these the values of the coefficients a_1 and a_2, we obtain the desired extrapolation formula

$$r_{i\,opt} = -\frac{1}{2} \cdot \frac{\Delta'(r_i''^2 - r_i'''^2) + \Delta''(r_i'''^2 - r_i'^2) + \Delta'''(r_i'^2 - r_i''^2)}{\Delta'(r_i'' - r_i''') + \Delta''(r_i''' - r_i') + \Delta'''(r_i' - r_i'')}.$$

In what follows, the determination of the coefficients of Kolmogorov's formula is reduced to a sequence of iterations: a) having assigned arbitrary values to the coefficients, we compute three values of the error Δ', Δ'' and Δ''' for three values of one of the coefficients r_i', r_i'', r_i'''. Applying the extrapolation formula, we use these quantities to find the optimal value of the coefficient r_i; b) taking this optimal value, we repeat the computation for the following coefficient r_{i+1}, etc, until the mean-square error stabilizes at some value. The error should decrease monotonically in the course of the iterations. The value of the error at the end of the iteration process is a measure of the prediction accuracy.

In volume of computational work, the iteration method is not much better than the regression analysis method, but it can be more easily programmed on computers. Requirements as to the length of the learning sequence are the same. The number of modes to be averaged in determining the mean-square error must in proactice exceed the number of Kolmogorov coefficients being determined by a factor of from five to ten. An advantage is the simplicity of constructing the self-learning filter with analog elements.

The Alpha recognition system as a predicting self-learning filter

The circuit of the Alpha recognition system (see Fig.58) completely reproduces all possibilities of Gabor's continuous self-learning filter, with the limitation that the former uses a discrete representation of quantities in the form of binary codes. For example, in a code "with change of sign" let there be given

$$f_1 = 0{,}3 = +1+1-1-1-1;$$
$$f_2 = 0{,}1 = +1-1-1-1-1;$$
$$f_3 = 0{,}9 = +1+1+1+1+1.$$

Then the input of the system is fed a general "image" or "representation point" code:

$$v_i = (0{,}3;\ 0{,}1;\ \dot{0}{,}9\ldots) = +1+1-1-1-1+1-1-1-1-1+1+1+1+1+1\ldots.$$

This code is fed to a certain number of groups of associating cells; this number is equal to the number of divisions R of the quantity being predicted. Figure 57 shows five divisions, corresponding to which Fig.58 shows five groups.

Every group is characterized by its pole (or prototype, standard). At the output of each of the groups a voltage is obtained which is proportional to the scalar product of the input code by the pole code. For example, if the pole of the first group is also

$$\alpha_1 = (0{,}3;\ 0{,}1;\ 0{,}9) = +1+1-1-1-1+1-1-1-1-1+1+1+1+1+1\ldots,$$

at its output we will then obtain the maximum possible voltage

$$\Sigma_1 = (\alpha_1 v_i) = U_{\max}.$$

It is clear that the largest voltage will be from that group for which the pole in the n-dimensional space of attributes is nearest to the representation point. When the pole α_k and the representation point v_i are equal, we will obtain the maximum possible voltage, which is shown in the given example. The large-voltage indicator LVI finds the pole nearest to the given representation point and thus predicts the future value of the function.

As a result of the learning process, the poles establish themselves in a position at which the prediction error does not exceed the error obtainable in the Gabor continuous filter, more than a half a discrete division. The system can be constructed with relays or flip-flops or can be programmed on a universal computer. The growth in volume of computations as prediction accuracy demands increase-- the basic difficulty inherent in the Gabor filter-- is also present here. The prediction accuracy increases both with an increase in the number of observed intervals N and with an increase in the number of digits of the discretizators n, and decreases with an increase in the number of levels of discretization of the output R. We will return to this question below.

Experiment on prediction of amplitude of ocean waves on an Alpha recognition system simplified as much as possible

With the aim of achieving a sharp cut in the volume of computations, it was proposed to decrease the volume of input information on the process. Let the discretizators of the Alpha system have one output; i.e. let them communicate to us only the sign of the deviation of the function from some mean value. The circuit of such a simplified recognition system is shown in Fig.59. It is clear that, in distinction from the complete system of Fig.58, the simplified system no longer has the potentialities of Gabor's continuous filter.

But how much does the prediction accuracy decrease here? Will such a system be capable of prediction at all?

To answer these questions, an experiment on predicting

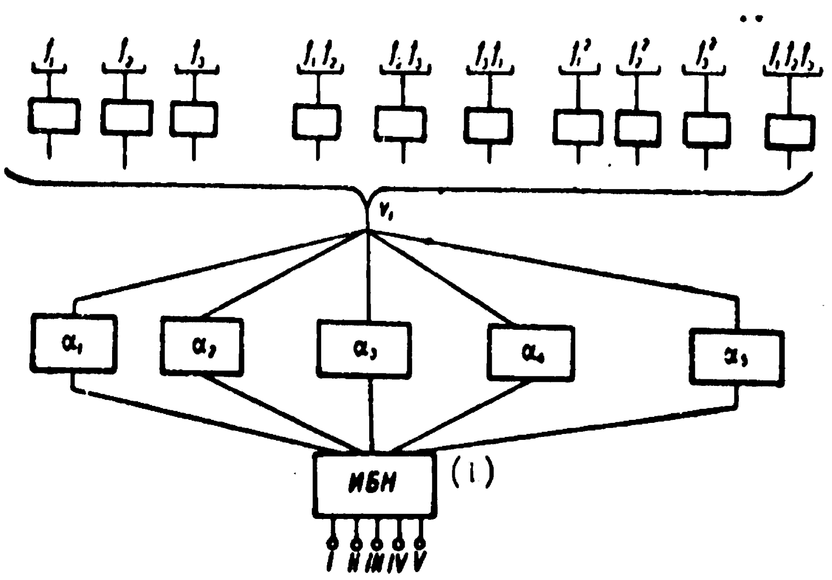


Figure 59. Simplified binary Alpha predicting system.
Key: 1) HVI.

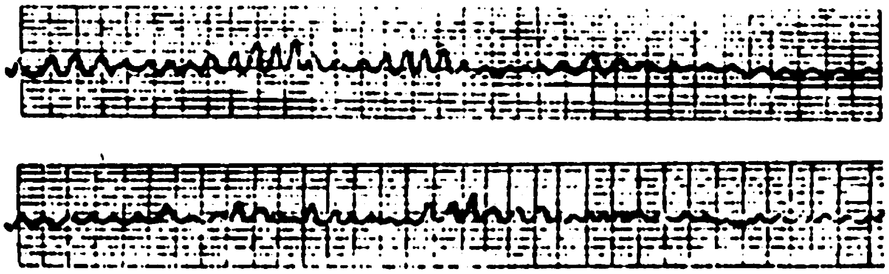


Figure 60. Recording of the variations in the amplitude of ocean waves.

the amplitude of ocean waves was conducted. The importance of the radical simplification of the prediction algorithm for this problem is determined by the fact that with complicated algorithms the determination time of the amplitude of the succeeding wave on a computer may be larger than

i	f_i	f_{i+1}	f_{i+2}	$f_{ср}$ (из трех)	$g_i \cdot 3$ (результат)
1	8 ±1	2 −1	3,5 −1	4,5	4 3
2	2 −1	3,5 +1	4 +1	3,2	4 3

Table 7.

Key: 1) f_{mean}(of three); 2) (result).

f_i	f_{i+1}	f_{i+2}	f_i, f_{i+1}	f_i, f_{i+2}	f_i^2	f^2_{i+1}	f^2_{i+1}	f_i f_{i+1}, f_{i+2}
+1 −1	−1 +1	−1 +1	−1 −1	−1 −1	+1 +1	+1 +1	+1 +1	+1 −1

Table 8.

the time interval between the waves (4-12 sec).

The recording of ocean waves (Fig.60) given in [51a] was used as initial material. A current mean value was determined for each three neighboring amplitudes. Only the deviations of the amplitudes from this mean value were recorded in the table of initial values: plus, if the wave was higher, and minus, if the wave was lower than the current mean value. The output scale of the system had five divisions. We took the value of each division equal to 3 (for trial purposes $(g_{max}-g_{min})/5=3$).

Tables 7 and 8 show a sample recording of wave deviations from the mean value and their products.

The feedback used was not exponential,but averaging.

The system used five groups, characterized by five poles α_1, α_2, α_3, α_4, and α_5. The system's self-learning of correct prediction consisted of self-establishment of the poles according to the averaging law

$$\alpha_{m+1} = \frac{v_1 + v_2 + \dots + v_m}{m},$$

where m is the number of arrival of the representation point giving the same result. For example, if it turned out that all measurements given in Table 8 gave the third division of the scale of the output of the wave amplitude, the pole of the third group could be determined thus:

$$\alpha_3 = +1 + 1 - 1 - 1 - 1 + 1 + 1 + 1 - 1 \quad \text{etc.}$$

In accordance with the evaluation given above, for self-learning of the poles a sequence was used made up of 140 measurements of not less than three neighboring values of the wave amplitudes and the following wave. After learning, the poles no longer moved, and the system was switched onto prediction. It turned out that it predicted the amplitude of the following wave with an accuracy of ±10% correctly in 80 cases out of 100; 10-20 cases represent an element of "pure" randomness in the given process; correct prediction with an accuracy of ±20% occurred with 96 waves out of 100. Thus, this highly simplified system, using a comparatively small volume of computations, is still capable of predicting processes with the indicated accuracy.

Investigation of prediction accuracy and attribute usefulness (terms of prediction formula)

An investigation was carried out on the influence of the number of levels on the prediction accuracy. Figure 61

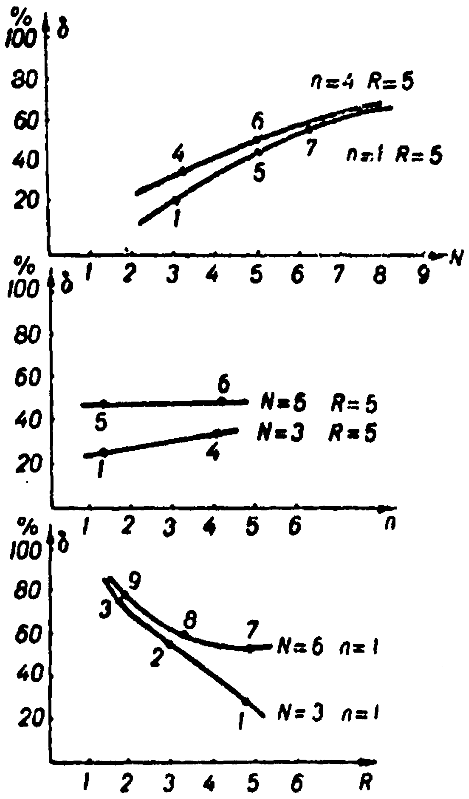


Figure 61. Prediction quality as a function of the number of points of the prehistory and the discretization levels: δ is the number of correct predictions (in %); N is the number of points of the prehistory; n is the number of levels of discretization of the input quantities; R is the number of levels of discretization of the quantity being predicted.

shows the functions which were obtained. The first graph in this figure shows that increasing the number of prehistory intervals N taken into account raises the prediction accuracy

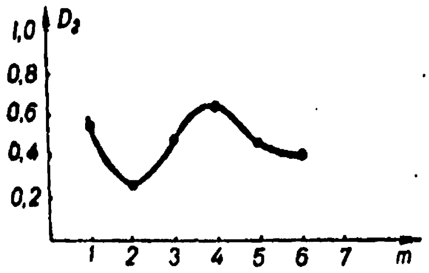


Figure 62. Usefulness of attributes D_2 selected on the basis of the Kolmogorov operator (prehistory 6 points) ; m is the order of terms of the Kolmogorov operator.

The second graph serves as a fundamental justification for the use of the simplified system of Fig.59 in place of the complete system of Fig.58. It has turned out that increasing the number of divisions n of the discretizators leads to an insignificant increase in the prediction accuracy. Finally, the dependence of δ on R (Fig.61) illustrates the obvious fact that increasing the number of divisions of the output scale lowers the number of correct answers of the system.

Self-arbitrary selection of most useful attributes

G.L.Otmezuri /36/ has pointed out that the selection of the most informative or useful attributes used for recognition can also be carried out in a determinate way by calculation of the value of some attribute usefulness criterion. It is possible to organize a self-arbitrary process of selection of the most useful attributes by using so-called secondary positive feedback. Here we can use any of the criteria proposed by various authors, for example the criterion of the number of distinguished disputes D_2 /21/, the criterion of the resolving power of the system R, the criterion of divergence or the criterion of change of entropy.

The example of predicting the amplitude of a wave was also used to check the possibility of discarding some terms of Kolmogorov's formula (Fig.62). Using the criterion of the "number of distinguished disputes" D_2, we determine the informative usefulness of the individual groups of terms of the formula. It turned out unexpectedly that the most informative attributes for prediction of wave amplitude were the terms of the fourth order, i.e. the terms $f_1f_2f_3f_4$, $f_2f_3f_4f_5$, $f_3f_4f_5f_1$, $f_4f_2f_1f_5$, etc. The investigation of the usefulness of combinations of attributes almost always gives unexpected results which are difficult to foresee from "common sense" considerations. Thus, in the experiments of V.L.Brailovskiy and A.L.Lunts, it turned out that the most useful attributes were not the initial attributes of the burns, but their pairwise combinations, i.e. the second-order terms in Kolmogorov's formula. The latter, together with the above method of determining the usefulness of attributes by the D_2 criterion is the most general algorithm which explains the success of given experiments.

However, we should not overemphasize the importance of such a system. There do not yet exist any systems which would "invent" or "think up" any unexpected attributes. Selection is carried out in a comparatively limited, preassigned set of attributes and their combinations. This is the chief weak spot of all complex work on the selection of attributes for pattern recognition.

It is enough to find one actually invariable attribute, and the whole picture of the division of the attribute space changes completely, and the whole investigation must be started again. A "noncompact" attribute set may become "compact", and the set itself sometimes decreases decisively.

Recognition Systems using Threshold Logical Elements

At the present time, much attention is being devoted to the creation and investigation of self-learning recognition systems on the basis of threshold logical elements.

Simple circuits based on threshold elements are being successfully used for speech recognition, in solving the problems of weather forecasting, in automatic control systems, and in solving the problems of diagnostics (for example, analysis of electrocardiograms).

The basic link in the self-learning machine is the threshold element, which is sometimes called an "adaptive neuron". Figure 63a shows the block diagram of such an element. Binary input signals $x_1, x_2, \ldots, x_n$ take the values +1 or -1. A linear combination of the input signals is formed within the neuron. The weighting factors are the amplification factors W_i, which may take both positive and negative values. The output signal of the element is equal to +1 if the weighted sum of the input signals is greater than a definite threshold, and -1 in all other cases. The value of the threshold is determined by selection of the factor W_{N+1}. The corresponding input x_{N+1}, which is called the threshold input, is constantly connected to a +1 source. If, for example, a threshold equal to 0 has been established, then the linear combination of the input signals

$$y = \sum_{i=1}^{N+1} W_i x_i$$

produces a signal

$$z = \begin{cases} +1, & y > 0 \\ -1, & y < 0. \end{cases}$$

at the output of the threshold element. When the weighting factor w_{N+1} is changed, the constant added to the linear combination of the input signals changes.

A_t given values of the weighting factors (amplification factors) each of the 2^n possible input combinations corresponds to one of the two values of the output +1 or -1. In the adaptive neuron, these factors are established during the learning process.

One of the structural variants of the learning machine is shown in Fig.63b.

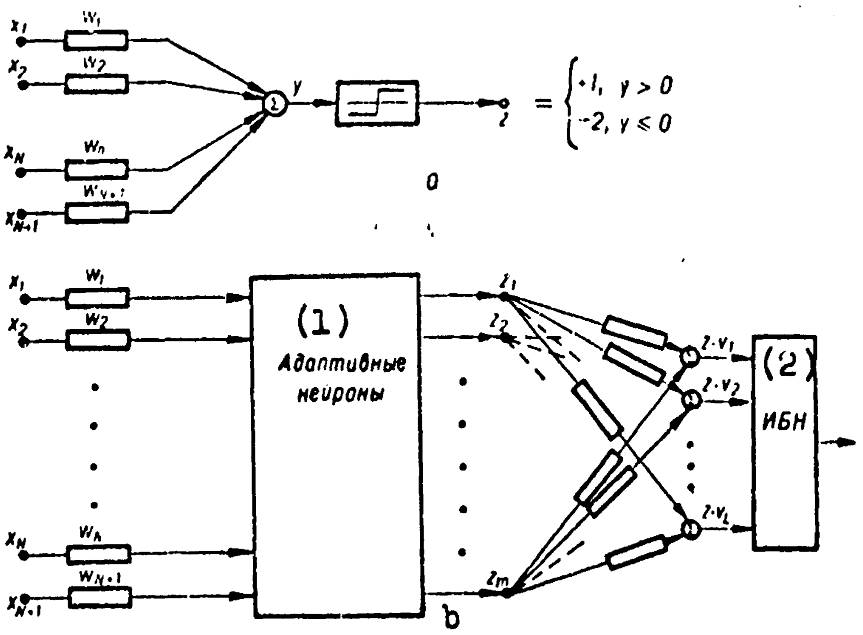


Figure 63. Recognition system using threshold elements.
Key: 1) Adaptive neurons; 2) IIVI.

The outputs of the threshold elements z_1, z_2, ...,z_M can be considered to be components of some M-dimensional vector z. Let us assume that we need to recognize L patterns. From the whole set of 2^n output vectors L vectors can be selected which best represent the patterns to be recognized. Let us denote these selected vectors by $v_1, \ldots,$ v_L. Then the vector z would be related to the pattern i whose representing vector v_i is nearer than all the rest to z. This means that the scalar product

$$z \cdot v_i = \max z \cdot v_j, \quad (i, j = 1, 2, \ldots, L).$$

The large voltage indicator shows which of the vectors v_i is the closest to z.

Let us consider the learning process.

The weighting factors are originally established equal to zero. The inputs are fed a combination of attributes $x_1, x_2, \ldots, x_N$ corresponding to the first pattern. If the answer given by the machine is correct, there are no changes. But if the answer is not correct, some of the components of the vector z differ from the corresponding components of the vector v_i. In this case, the output signals of some threshold elements (of those which did not coincide) change to the inverse ones in order to bring z near to v_i.

The number of elements to be corrected is taken to be some number $d' = \rho \cdot d$, where d is the number of all neurons which did not coincide, and ρ is some quantity varying within the limits $0 \leqslant \rho \leqslant 1$. The value of ρ is determined experimentally. If, for example, $\rho = (1/4)$, this means that a fourth of all neurons which did not coincide are corrected.

Prediction of changes in atmospheric pressure

If the machine just described is used for prediction purposes, in this case the prehistory serves as input signals, and the output represents the predicted values of the process in wkich we are interested. As an example of the use of the learning machine for prediction, let us consider the problem of the determining the future values of atmospheric pressure (Fig.64a). The part of the prehistory is played by four pressure values: the pressure at the given moment and the values of the pressure for the three preceding hours, which have changed on two different levels.

The prediction quality was determined from the amount of lowering of the variation

$$q = [1 - \frac{\overline{(f_t - f_t^*)^2}}{\overline{f^2} - \overline{f}_t^2}] \cdot 100\%.$$

In the example described, a variation lowering of q=63.3% was obtaned. Optimal prediction (according to the minimum mean-square error) for the same initial data gives a variation lowering of q=76.2% and, when linear regression analysis is used, of q=60.4%. Figure 64b shows the result of optimal prediction.

Let us note that, both in the case of prediction by means of a learning machine and in the case of optimal prediction, the errors are characterized by some general tendency. This bears witness to the fact that prediction errors are rather more to be explained by the probabilistic nature of the process and the presence of unpredictable "pure" randomness than by the use of this or that method (for example, the use of a learning machine).

The Use of Recognition Systems as Learning Correctors of Extremal Control

This section considers nonsearch extremal control systems which do not use test intervals of variation of the regulating influences on the object. The basic assumptions in this problem are based on the use of the methods of

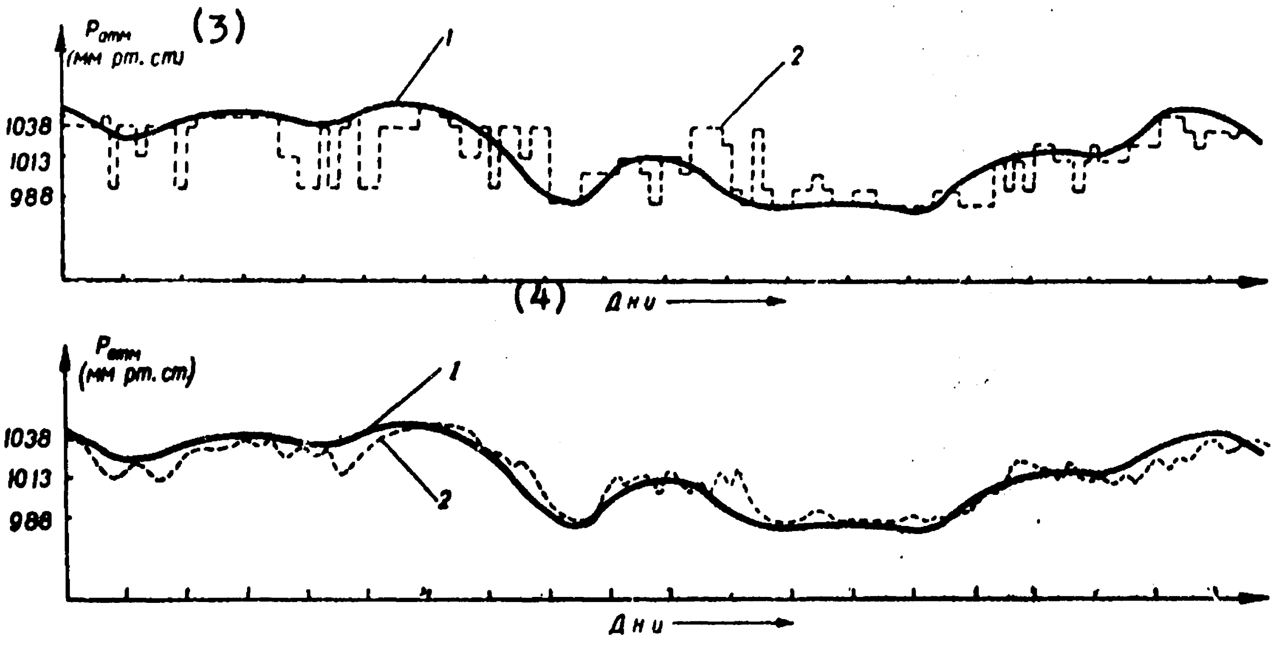


Figure 64. Prediction of atmospheric pressure: a) by means of a recognition system using threshold elements; b) optimal prediction according to the criterion of the mean-square error.

Key: 3) P_{atm} (mm Hg); 4) days.

"passive experiment" with further use of regression analysis [21a,44,57].

Among the number of difficulties which are encountered here, we may point out the very large volume of calculations in solving the equations on digital computers. Thus the problem of algorithmization of the process in a fractionating column leads to a system of 20 equations in 20 unknowns. The duration of the solution of the equations and the necessity of using large periods for averaging the input data [57] lead to low speed of action of the regulator.

The demands on a correlation (regression) regulator can be greatly decreased if the latter is used only as a corrector for a fast-acting open part of the control system which is in the form of a switching matrix of keys (Fig.65, left).

The thought arises of using the methods of "active" or "passive" experiment only for the learning of the recognition system, in which there is no need for solving equations. Using certain attributes, the system must, after learning, recognize "situations" and by this means give correct indications for correction of the characteristic of the open part. Below we give a definition of the concepts "state" (or "image") and "situation" (or "pattern") and synthesize a circuit and select the most useful (informative) attributes for the recognition system, the corrector.

The basic limitation which we assume is the assumption that the distribution of the perturbations remains almost constant, although it itself may be unknown to us. For large changes in the distribution, the system must be retaught anew. In what follows, this limitation will be removed by means of a special method of constructing the attribute pickups.

Furthermore, it is assumed that the inertia of the object is small or that we can connect to the output of the system a link with an inverse operator (anticipator), this link re-establishing the exact oscillogram of variation. of the quality index. Experiment shows that analog models carry out inverse transformation in measurement circuits with sufficient accuracy. The "Smith anticipator" [43] is used for objects with constant lag. Another possibility is the introduction of a delay into the circuit for measuring regulating and perturbing influences equal to the delay in the quality index circuit. However, this way of compensating for the inertia of an object, although it is simpler to achieve, is undesirable, since it slows down the action of the corrector.

The circuit for the open part in the form of a switching matrix of keys is not the only possible one. In another variant, the open part is made, in turn, in the form of a recognition system whose poles ("polar gas") are taught, for example, according to the algorithm in [25]. In this variant, the regression formula corrector plays the part of the teacher of the open part of the extremal control system.

It is desirable to supplement the algorithm used in this article with the laws of interaction of the poles among themselves as is done below.

Formulation of the correlation problem

Usually the extremal characteristic of the object can be approximated by a generalized power series of the second or third degree. For example, for an object which is a hydroturbine (see Fig.65), we can write:

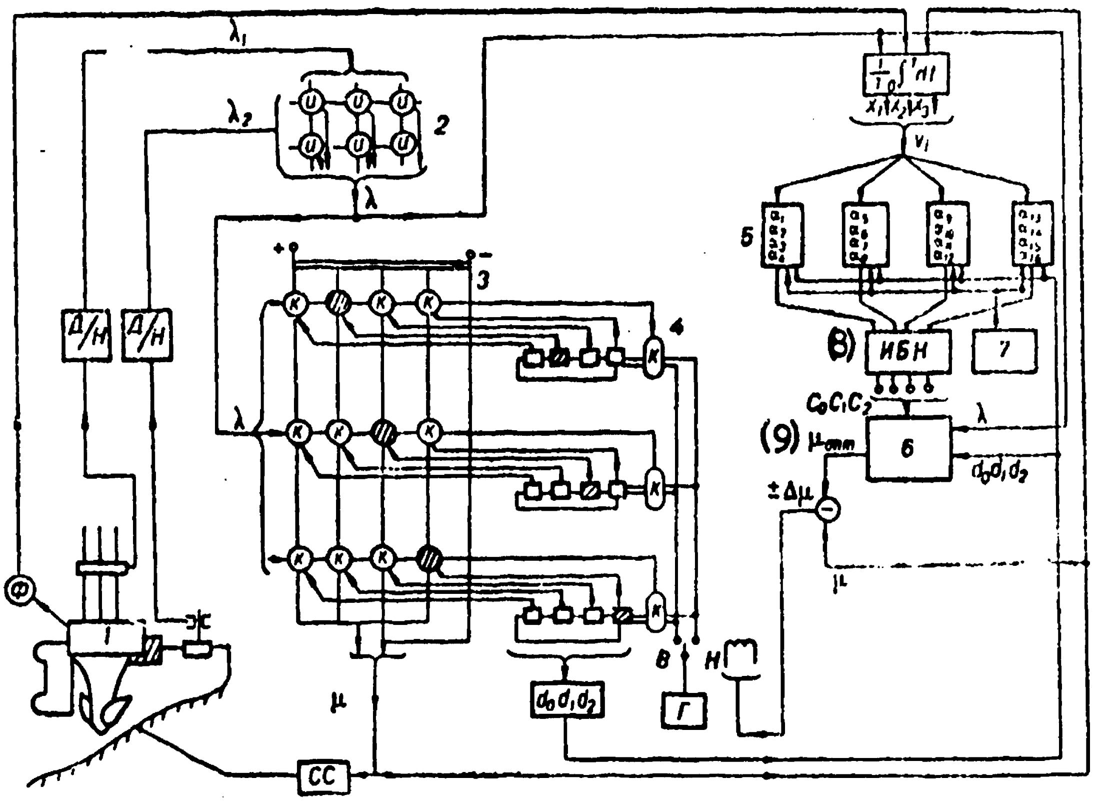


Figure 65. Example of combined system for extremal control with corrector, an Alpha recognition system: 1) object (turbine); 2) matrix convergence circuit; 3) matrix of keys of open part; 4) controlling reversible counters; 5) recognition system; 6) optimal characteristic model; 7) "teacher".

Key: 8) HVI; 9) μ_{opt}.

$$\varphi = a_0 + a_1\mu + a_2\lambda_1 + a_3\lambda_2 + a_4\mu^2 + a_5\lambda_1^2 + a_6\lambda_2^2 + a_7\mu\lambda_1 +$$
$$+ a_8\mu\lambda_2 + a_9\lambda_1\lambda_2 + a_{10}\mu^3 + a_{11}\lambda_1^3 + a_{12}\lambda_2^3 + a_{13}\mu^2\lambda_1 + a_{14}\mu^2\lambda_2 +$$
$$+ a_{15}\lambda_1^2\lambda_2 + a_{16}\lambda_1^2\mu + a_{17}\lambda_2^2\lambda_1 + a_{18}\lambda_2^2\mu + a_{19}\mu\lambda_1\lambda_2,$$

where φ is the extremum index (efficiency of turbine); μ is the regulating influence (angle of rotation of the vanes); λ_1, λ_2 are the basic perturbing influences (water pressure and turbine load).

If we use the convergence (rotation)of the matrices of discrete values of the perturbations into a row of generalized perturbation $\lambda(\lambda_1, \lambda_2)$, the same characteristic can be represdnted by a simpler polynomial with two arguments:

$$\varphi = a_0 + a_1\mu + a_2\lambda + a_3\mu^2 + a_4\lambda^2 + a_5\mu\lambda + a_6\lambda^3 + a_7\mu^3,$$

where (in the case of rotation in the usual order of row by row)

$$\lambda = [\lambda_1 + l_1(\lambda_2 - 1)]\,\Delta l;$$

l is the number of levels of discretization; Δ l is the discretization interval.

In the majority of cases we can select an order for rotation of the matrices for which the characteristic of the object is obtained as smooth so that it can be approximated by a straight line or a parabola of the second degree. Here the optimal characteristic of the object, on which it is desirable to operate all the time is defined by the equation

$$\frac{d\varphi}{d\mu} = 0 \text{ where } \lambda = \text{const} \text{ if } \lambda = c_0 + c_1\mu + c_2\mu^2,$$
$$c_0 = -\frac{a_1}{a_5}; \quad c_1 = -\frac{2a_3}{a_5}; \quad c_2 = -\frac{3a_7}{a_5}.$$

The best results are given by transposition in the order from the least mean value μ to the following larger one. If the characteristic becomes of too complicated form here too, we should not use convergence and raise the degree of the approximating polynomial. This only increases the volume of the recognition system and the duration of its learning.

In the case of a second-degree polynomial, the correction problem consists of keeping the mean line of the characteristic of the open part

$$\lambda = d_0 + d_1\mu + d_2\mu^2$$

as near as possible to the optimal characteristic of the object; i.e. with small displacements, rotations and deformations of this characteristic, it is possible to establish

$$x = c_0 - d_0 = 0; \quad y = c_1 - d_1 = 0; \quad z = c_2 - d_2 = 0.$$

more quickly. We are only concerned with the mean line, because in a nonsearch extremal system the characteristic of the open part should in essence not coincide in form with the optimal characteristic of the object. It is a straight line or a second-degree parabola with small "teeth" superimposed on it which take the place of the search oscillations at the object /21a/. The essential nature of this prohibition is connected with the well-known rule in interpolation theory, according to which the interpolation points (nodes) cannot be selected arbitrarily, in µrticular on one straight line.

Another definition of the concepts of "state" ("image") and "situation" ("pattern")

Earlier /21a/, we characterized the state of an extremal system by the coordinates of a point of the space

$$\Omega_{v_l}(\varphi_1, \varphi_2, \dots, \varphi_s, \mu_1, \dots, \mu_z, \lambda_1, \dots, \lambda_\gamma).$$

In accordance with this, we defined a situation as a definite region of this space.

We will now change our approach and will characterize the state of the system by the coordinates of a point of the space

$$\Omega_{v_l}(c_0, c_1, c_2, d_0, d_1, d_2).$$

In accordance with this, a situation should now be defined as some region of this new coordinate space. The correction problem consists of bringing the system into a region where $c_0=d_0$, $c_1=d_1$, $c_2=d_2$.

Now by a state (or image) we will mean all possible respective positions of the mean line of the characteristic of the open part and the optimal characteristic of the object. It is assumed that the coordinates $c_0, c_1, c_2, d_0, d_1, d_2$ can assume only a number of fixed discrete values. Hence the total number of possible states which it is necessary to distinguish is finite. Figure 66 shows the 16 states of the combined extremal system which are used below in the example.

The total number of possible states is equal to

$$S = l_0 + l_1(l_2 - 1) + l_1 l_2 (l_3 - 1) + l_1 l_2 l_3 (l_4 - 1) + \\ + l_1 l_2 l_3 l_4 (l_5 - 1),$$

where $l_0, l_1, l_2, l_3, l_4, l_5$ is the number of discrete levels of measurement of the coordinates $c_0, c_1, c_2, d_0, d_1, d_2$. It is easy to compute that with the actually used number of

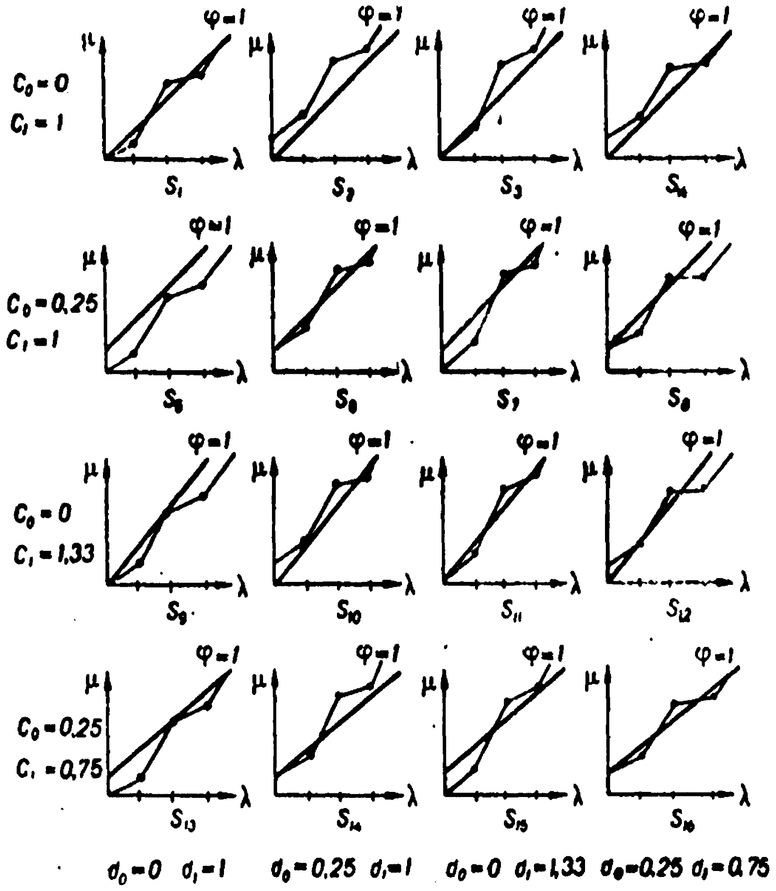


Figure 66. The central states of 16 situations which it is required to distinguish.

divisions, the total number of possible states is expressed by astronomically huge figures. No actually realizable system can have so many outputs.

A similar problem arises in the recognition of visual patterns. If, for example, 100 attributes are used in recognizing letters, the number of possible codes is 2^{100}.

Only to sort this number of variants on a fast-action machine with a counting speed of 10^6 comparisons per second would require more than a thousand years. A way out is found in the classifioation of images which are close in some sense as one pattern (property of generalization).

We should proceed in the same way in the case considered of classifying the states of an extremal system into situations. For example, using only two groups (two poles) we divide in this way the space of the coordinates $c_0, c_1, c_2, d_0, d_1, d_2$ into two regions, i.e. situations.

All states falling into the first situation will be indicated by the system at the first output, and the states falling into the second region will cause operation of the second output. Thus, the number of situations is determined by the number of poles, and their boundaries coincide with the boundaries of the "attraction regions" of these poles. Learning or self-learning of the system has the aim of a rational selection of the position of the poles and boundaries.

Usually in designing a system, it is possible to point out a certain comparatively small number of characteristic (central) states which are to be central situations after the learning of the system. In learning, the poles are located at points located as close as possible to these central states. Then the name situation can be given to the region of the space

$$\Omega_{vi}(c_0, c_1, c_2, d_0, d_1, d_2),$$

where the whole set of states is found which the system generalizes with a given central state (prototype).

The use of the Alpha recognition system for distinguishing situations

Figure 67 shows a sample system circuit. In this use of the recognition system, the attribute pickups, in accordance with observation of the instantaneous changes in the quantities φ, μ, λ produce certain integral functions of these quantities $x_1, x_2, \ldots, x_n$, which in the theory of pattern recognition are usually called attributes. The basic advantage of a recognition system consists in the fact that it is a very capable "student" and, after learning, acts by an order faster than its "teacher". The necessity for a "teacher" disappears after the recognition system learns correctly to distinguish a sufficient number of situations.

Let us assume that, as a result of an analysis of the usefulness of the attributes, we have selected three useful attributes x_1, x_2, x_3 (n=3).

The vector at the input of the system will be: $v_i(x_1, x_2, x_3)$. The poles of m groups of associating cells will also have three components each: $\alpha_1(r_{11}, r_{12} r_{13})$, $\alpha_2(r_{21}, r_{22}, r_{23})$, ..., $\alpha_m(r_{m1}, r_{m2}, r_{m3})$. The voltages at the outputs of the groups will be, respectively,

$$\Sigma_1 = (\alpha_1 v_i) = r_{11}x_1 + r_{12}x_2 + r_{13}x_3,$$
$$\Sigma_2 = (\alpha_2 v_i) = r_{21}x_1 + r_{22}x_2 + r_{23}x_3,$$
$$\Sigma_3 = (\alpha_3 v_i) = r_{31}x_1 + r_{32}x_2 + r_{33}x_3.$$

If, for example, wish to teach the first group to relate a given representation point to the first situation, we must select r_{11}, r_{12}, r_{13} so that the scalar product Σ_1 will be greater than the others: $\Sigma_1 > \Sigma_2;\ \Sigma_1 > \Sigma_3; \ldots; \Sigma_1 > \Sigma_m$.

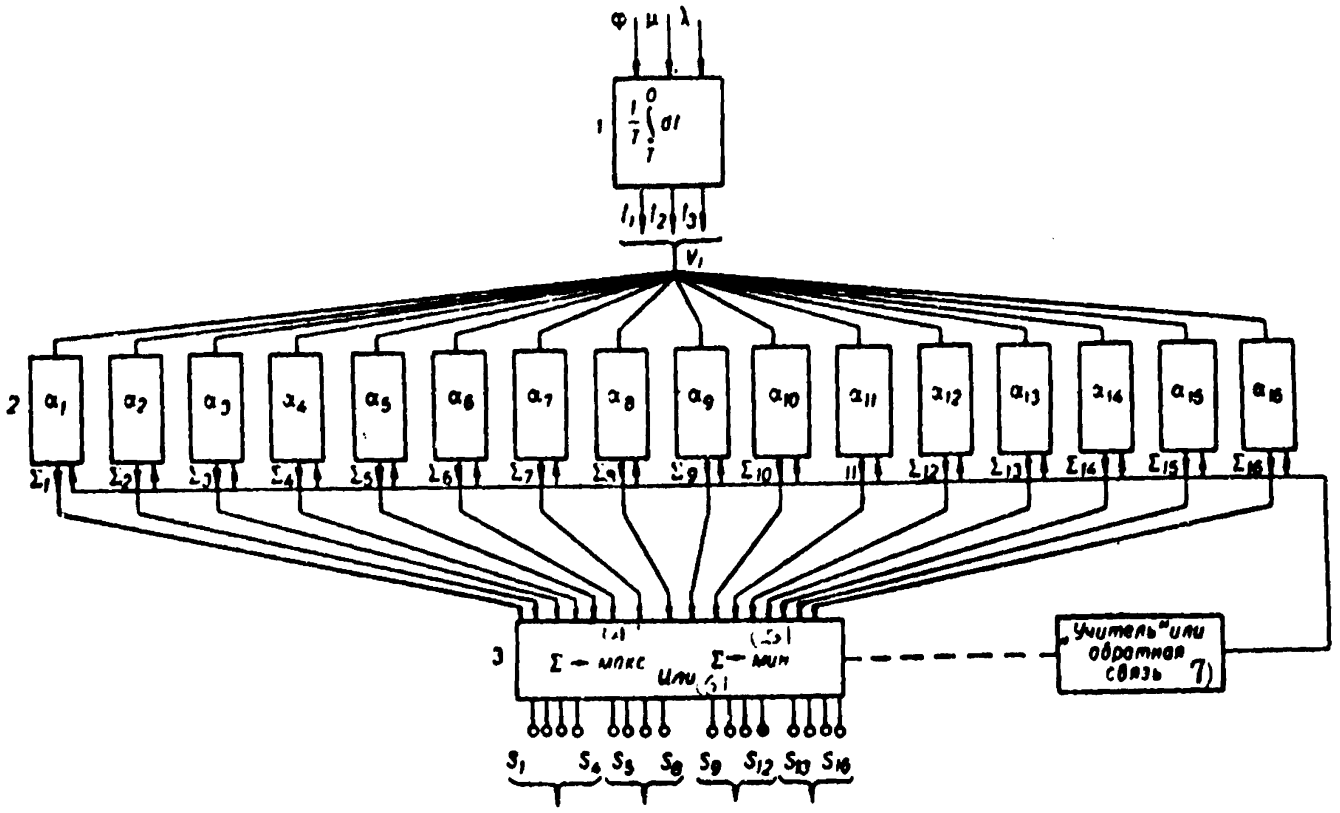


Figure 67. Circuit of Alpha recognition system: 1) attribute pickups; 2) groups of associating cells; 3) comparator, HVI or LVI.

Key: 4) max; 5) min; 6) or; 7) "teacher" or feedback.

It is well known that the scalar product of two vectors is equal to the sum of the products of the projections of these vectors. The maximum of the first scalar product is achieved when we have the equalities:

$$r_{11}=x_1;\quad r_{12}=x_2;\quad r_{13}=x_3.$$

If we want the first central state to produce operation of the output of the first group, the pole of this group must be established at a point corresponding to this state.

Above we spoke about splitting the space of the coorinates c_0,c_1,c_2,d_0,d_1,d_2into regions, i.e. situations. To every point of the coordinate space

$$\Omega_{v_i}(c_0,\ c_1,\ c_2,\ d_0,\ d_1,\ d_2)$$

there corresponds a definite geometric locus of points of the coordinate space of attributes $x_1,x_2,x_3,\ \ldots,x_n$. The latter space can also be split into regions, i.e. situations. In learning a pole should be established at the center of a situation corresponding to a given central state in both coordinate systems. But in practice we cannot protractedly and accurately maintain the system in a given central state, since the perturbation distribution and the form of the extremal hill are constant only to a first approximation. In reality they vary with time around some mean value. Hence we should establish the pole at the "center of gravity" of the points, i.e. the states relating to the given situation. With a large number of measurements, the "center of gravity" coincides simultaneously both in the space

$$\Omega_{v_i}(c_0, c_1, c_2, d_0, d_1, d_2),$$

and in the attribute space

$$\Omega_{v_i}(x_1,\ x_2,\ x_3, \ldots, x_n\ ).$$

This consideration indicates to us the rule for learning of the poles: the pole of the group being learned must be located at the center of the situation, this center being defined as the arithmetical mean of all points of the learning sequence relating to the given situation. By observing the operation of the object in all possible states and knowing (from the "teacher") the number of the situation to which they relate, we can prepare tables or graphs of the learning sequences (Fig.68). Having selected the data relating to one and the same situation, we find the position of the pole of the group corresponding to it in the learned state by means of averaging (where c_0=const; c_1= const; c_2=const):

$$r_{11} = \bar{x}_1 = \frac{1}{T}\int_0^T x_1 dt; \quad r_{12} = \bar{x}_2 = \int_0^T x_2 dt; \quad r_{13} = \bar{x}_3 = \frac{1}{T}\int_0^T x_3 dt.$$

This computationally second averaging is required only during the period of learning the poles. The first averaging is constantly required in working out the attributes $x_1, x_2, x_3, \ldots, x_n$.

Decreasing the duration of learning of the poles by means of interpolation

When the number of situations and groups is increased, the learning time increases accordingly. The mean values of the attributes $x_1, x_2, \ldots, x_n$ must be "displayed" for every situation in order to establish the group poles at these points. To shorten the learning, we can use self-arbitrary establishment of the poles according to the formulas of interpolation or regression analysis. Let us explain this.

It is easy to distinguish the "anchored" poles, which have already been indicated by the "teacher", from the "unanchored"ones, which have not yet been indicated by it.

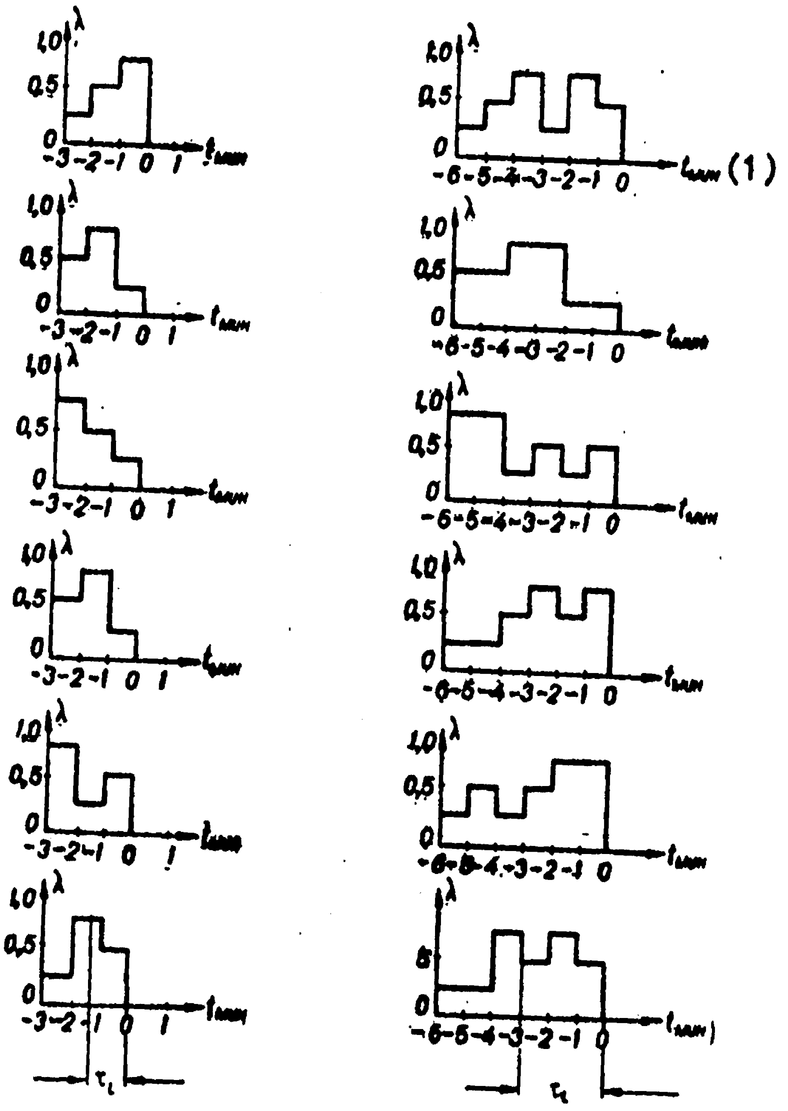


Figure 68. Examples of sequences with uniform distribution of perturbation probabilities.

Key: 1) t_{min}.

Unanchored poles do not remain constant, but displace as some function of already indicated poles. For example, with linear interpolation, the coordinates of unachored poles can be determined from the formulas

$$r_{1i} = \frac{r_{1(l+1)} + r_{1(l-1)}}{2}; \quad r_{2l} = \frac{r_{2(l+1)} + r_{2(l-1)}}{2}; \ldots;$$
$$r_{ml} = \frac{r_{m(l+1)} + r_{m(l-1)}}{2}.$$

With such an algorithm, unanchored poles are "repelled" from one another like particles of some "polar gas" [21a] and are located at equal distances from one another along straight lines connecting already indicated, "anchored" poles. As a result of this motion of the poles, the learning process is shortened.

Elucidation of the makeup of the "attribute" set

In order to elucidate the makeup of the set of attributes, a part of which we are going to feed to the input of a recognition system, let us return to regression analysis.

According to the regression method, the coefficients are determined from the conditions for obtaining the minimum mean-square error

$$\Delta = \frac{1}{n}\sum_{1}^{n} \varepsilon^2 = \overline{(\varphi_0 - \varphi)^2} = \overline{(\varphi - a_1\mu - a_3\mu^2 - a_5\mu\lambda - a_7\mu^3)^2}.$$

A minimum must exist and is unique, since the error is a linear function of the coefficients α_i. Taking derivatives when $\varphi_0 = \bar{\varphi}$, we find four normal regression equations:

$$\frac{\partial\Delta}{\partial a_1} = 0; \ \frac{\partial\Delta}{\partial a_3} = 0; \ \frac{\partial\Delta}{\partial a_5} = 0; \ \frac{\partial\Delta}{\partial a_7} = 0,$$

whence

$$a_1\overline{\mu^2} + a_3\overline{\mu^3} + a_5\overline{\mu^2\lambda} + a_7\overline{\mu^4} = \overline{\mu\varphi_0};$$

$$a_1\overline{\mu^3} + a_3\overline{\mu^4} + a_5\overline{\mu^3\lambda} + a_7\overline{\mu^5} = \overline{\mu^2\varphi_0},$$

$$a_1\overline{\mu^2\lambda} + a_3\overline{\mu^3\lambda} + a_5\overline{\mu^2\lambda^2} + a_7\overline{\mu^4\lambda} = \overline{\mu\lambda\varphi_0},$$

$$a_1\overline{\mu^4} + a_3\overline{\mu^5} + a_5\overline{\mu^4\lambda} + a_7\overline{\mu^6} = \overline{\mu^3\varphi_0}.$$

Four terms have been omitted since they have no effect on the coefficients c_0, c_1 and c_2. This halves the number of regression equqtions.

Solving the equations, we can find the coefficients

$$c_0 = -\frac{a_1}{a_5};\ c_1 = -\frac{2a_3}{a_5};\ c_2 = -\frac{3a_7}{a_5}$$

and, hence, determine the optimal characteristic of the object.

Let us make the notations:

$$x_1 = \bar{\varphi} = \frac{1}{T}\int_{-T}^{0} \varphi dt, \qquad x_6 = \overline{\lambda^2} = \frac{1}{T}\int_{-T}^{0} \lambda^2 dt,$$

$$x_2 = \bar{\mu} = \frac{1}{T}\int_{-T}^{0} \mu dt, \qquad x_7 = \overline{\varphi\mu} = \frac{1}{T}\int_{-T}^{0} \varphi\mu dt,$$

$$x_3 = \bar{\lambda} = \frac{1}{T}\int_{-T}^{0} \lambda dt, \qquad x_8 = \overline{\varphi\lambda} = \frac{1}{T}\int_{-T}^{0} \varphi\lambda dt,$$

$$x_4 = \overline{\varphi^2} = \frac{1}{T}\int_{-T}^{0} \varphi^2 dt, \qquad x_9 = \overline{\mu\lambda} = \frac{1}{T}\int_{-T}^{0} \mu\lambda c't$$

$$x_5 = \overline{\mu^2} = \frac{1}{T}\int_{-T}^{0} \mu^2 dt, \qquad x_{10} = \overline{\varphi^3} = \frac{1}{T}\int_{-T}^{0} \varphi^3 dt,$$

$$x_{11}=\overline{\mu^3}=\frac{1}{T}\int_{-T}^{0}\mu^3dt,$$

$$x_{12}=\overline{\lambda^3}=\frac{1}{T}\int_{-T}^{0}\lambda^3dt,$$

$$x_{13}=\overline{\varphi^2\mu}=\frac{1}{T}\int_{-T}^{0}\varphi^2\mu dt,$$

$$x_{14}=\overline{\varphi^2\lambda}=\frac{1}{T}\int_{-T}^{0}\varphi^2\lambda dt,$$

$$x_{15}=\overline{\mu^2\varphi}=\frac{1}{T}\int_{-T}^{0}\mu^2\varphi dt,$$

$$x_{16}=\overline{\lambda^2\varphi}=\frac{1}{T}\int_{-T}^{0}\lambda^2\varphi dt,$$

$$x_{17}=\overline{\mu^2\lambda}=\frac{1}{T}\int_{-T}^{0}\mu^2\lambda dt,$$

$$x_{18}=\overline{\lambda^2\mu}=\frac{1}{T}\int_{-T}^{0}\lambda^2\mu dt,$$

$$x_{19}=\overline{\varphi\mu\lambda}=\frac{1}{T}\int_{-T}^{0}\varphi\mu\lambda dt,$$

$$x_{20}=\overline{\varphi^4}=\frac{1}{T}\int_{-T}^{0}\varphi^4dt,$$

$$x_{21}=\overline{\mu^4}=\frac{1}{T}\int_{-T}^{0}\mu^4dt,$$

$$x_{22}=\overline{\lambda^4}=\frac{1}{T}\int_{-T}^{0}\lambda^4dt,$$

$$x_{23}=\overline{\varphi^2\mu^2}=\frac{1}{T}\int_{-T}^{0}\varphi^2\mu^2dt,$$

$$x_{24}=\overline{\varphi^2\lambda^2}=\frac{1}{T}\int_{-T}^{0}\varphi^2\lambda^2dt,$$

$$x_{25}=\overline{\mu^2\lambda^2}=\frac{1}{T}\int_{-T}^{0}\mu^2\lambda^2dt,$$

$$x_{26}=\overline{\varphi^3\mu}=\frac{1}{T}\int_{-T}^{0}\varphi^3\mu dt,$$

$$x_{27}=\overline{\varphi^3\lambda}=\frac{1}{T}\int_{-T}^{0}\varphi^3\lambda dt,$$

$$x_{28}=\overline{\mu^3\varphi}=\frac{1}{T}\int_{-T}^{0}\mu^3\varphi dt,$$

$$x_{29}=\overline{\mu^3\lambda}=\frac{1}{T}\int_{-T}^{0}\mu^3\lambda dt,$$

$$x_{30}=\overline{\lambda^3\varphi}=\frac{1}{T}\int_{-T}^{0}\lambda^3\varphi dt,$$

$$x_{31}=\overline{\lambda^3\mu}=\frac{1}{T}\int_{-T}^{0}\lambda^3\mu dt.$$

With these notations, the solutions of the normal equations can be written in the following form (we put $\alpha_7=c_2=0$):

$$C_0=\frac{a_1}{a_0}=\frac{x_7x_{21}x_{23}+x_{11}x_{19}x_{26}+x_{17}x_{15}x_{26}-x_{17}x_{19}x_{21}-x_5x_{26}x_{29}-x_{11}x_{16}x_{23}}{x_5x_{19}x_{21}+x_{11}x_{16}x_{17}+x_7x_{11}x_{26}-x_7x_{17}x_{21}-x_5x_{16}x_{29}-x_{11}x_{16}x_{19}};$$

$$C_1=\frac{a_2}{a_0}-2\,\frac{x_5x_{16}x_{23}+x_7x_{26}x_{17}+x_{11}x_{17}x_{19}-x_{15}x_{17}x_{17}-x_5x_{19}x_{29}-x_7x_{11}x_{26}}{x_5x_{19}x_{21}+x_{11}x_{16}x_{17}+x_7x_{11}x_{29}-x_7x_{17}x_{21}-x_5x_{16}x_{29}-x_{11}x_{11}x_{19}}.$$

These solutions show that, for recognition systems, the attribute set must be sought among the quantities $x_1, x_2, x_3, \ldots, x_j, \ldots$. It is clear that the set of attributes $x_5, x_7, x_{11}, x_{15}, x_{16}, x_{19}, x_{21}, x_{23}, x_{26}, x_{29}$ completely solves the problem of recognition of the situations, but it is too complicated and superfluous for solving the problem of distinguishing a given number of central states.

Selection of useful (informative) attributes

All the attributes found possess an important property: their value depends only on the situation and does not depend on the order of variation of the instantaneous values of the quantities φ, μ and λ if the distribution of the probability of discrete values of perturbation is constant. Figure 68 shows sample sequences of variation of perturbation λ having a uniform probability distribution $p(\lambda)=$ const. It is clear that both learning and test or working sequences of values of φ, μ, λ arriving at the recognition system must be of sufficient duration to be able to transmit the objectively existing perturbation probability distribution. It is precisely this which determines the speed of action of the recognition system as a corrector. The operation delay is equal to about half of the averaging time, i.e. half the duration of the represented sequence ($\tau_L=1.5$ min for sequences shown on the left-hand side of Fig. 3 [sic], and $\tau_L=3.0$ min for sequences on the right).

This delay is less than the delay in determining the characteristic of the object by the regression method [57].

Recognition theory knows many methods of evaluating the usefulness of attributes (by the number of disputes to be resolved D_2, entropy criterion, divergence criterion, etc [21]). But all of them have been worked out for binary "yes-no" attributes. A special feature of the problem under consideration is the fact that the attributes are not binary, but continuous quantities. Continuous attributes should be evaluated directly from the value which they assume in all situations to be recognized.

Let us consider an example of selection of the most useful attributes. Let us assume that the object can be described by the nonlinear equation

$$\ddot{\varphi} = 1 - (\mu - c_1\lambda - c_0)^2.$$

Then for the values of c_0, c_1, d_0, d_1 indicated in Figure 66, for any of the sequences shown in Fig.67, we obtain the values of the numerically first 19 attributes; these values are shown in Table 9.

The attributes can be divided into four groups:

group a: $x_7,\ x_{13},\ x_{15},\ x_{19}$;
group b: $x_1,\ x_4,\ x_8,\ x_{10},\ x_{14},\ x_{17}$;
group c: $x_2,\ x_5,\ x_9,\ x_{11},\ x_{16},\ x_{18}$;
group d: $x_3,\ x_6,\ x_{12}$.

The attributes of group a carry information on the variations both of φ and of μ and hence can be used for the construction of recognitionssystems operating according to one attribute in all. The attributes of group b contain information on the variations of φ, and the attributes of group c, on the variations of μ. These attributes can only be used pairwise (i.e. one attribute from group b and one from group c), since if this is not done we can find

Table 9

Признаки (1) / (2) Состояния	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
S_1	0,9975	0,4833	0,5	0,995	0,2775	0,29167	0,4823	0,4987	0,2833
S_2	0,9433	0,7333	0,5	0,8904	0,5816	0,2917	0,6907	0,4717	0,4083
S_3	0,9702	0,6510	0,5	0,9416	0,5003	0,2917	0,6286	0,4818	0,1811
S_4	0,9832	0,6093	0,5	0,9668	0,3966	0,2917	0,5997	0,4932	0,0657
S_5	0,9267	0,4833	0,5	0,8592	0,2775	0,2917	0,4498	0,4633	0,2833
S_6	0,9975	0,7333	0,5	0,995	0,5817	0,2917	0,7315	0,4987	0,4083
S_7	0,9832	0,6510	0,5	0,9617	0,5003	0,2917	0,6436	0,4939	0,3811
S_8	0,9754	0,6093	0,5	0,9520	0,3966	0,2917	0,5921	0,4841	0,3357
S_9	0,9595	0,4833	0,5	0,9218	0,2775	0,2917	0,4586	0,4737	0,2833
S_{10}	0,9887	0,7333	0,5	0,9775	0,5817	0,2917	0,7257	0,4952	0,4083
S_{11}	0,9974	0,6510	0,5	0,9947	0,5003	0,2917	0,6493	0,4986	0,3811
S_{12}	0,9802	0,6093	0,5	0,9615	0,3966	0,2917	0,5915	0,4857	0,3357
S_{13}	0,9752	0,4833	0,5	0,9515	0,2775	0,2917	0,4755	0,4912	0,2833
S_{14}	0,9834	0,7333	0,5	0,9673	0,5817	0,2917	0,7188	0,4902	0,4083
S_{15}	0,9877	0,6510	0,5	0,9658	0,5003	0,2917	0,6403	0,4918	0,3811
S_{16}	0,9970	0,6093	0,5	0,9952	0,3966	0,2917	0,6078	0,4987	0,3357

Sample calculation of the first number of the table: from Fig.68 we find: λ_1=0.25, λ_2=0.05, λ_3=0.75. Correspondingly we will have from Fig.66: μ_1=0.25-0.05, μ_2=0.5+0.05, μ_3=0.75-0.05. We then determine φ from the formula: φ_1=0.9975, φ_2=0.9975, φ_3=0.9975; whence

$$x_1 = \frac{\varphi_1+\varphi_2+\varphi_3}{2} = 0{,}9975.$$

Value of attributes in sixteen states (with p(λ)= const)

x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
0,9925	0,1725	0,1875	0,4807	0,4975	0,2769	0,1763	0,2910	0,1813	0,2877
0,841	0,4868	0,1875	0,651	0,4452	0,5477	0,3491	0,2758	0,2542	0,3855
0,9144	0,4178	0,1875	0,6033	0,4645	0,4798	0,3187	0,2801	0,2440	0,3657
0,9509	0,2703	0,1875	0,593	0,4866	0 3925	0,2341	0,2887	0,2074	0,3216
0,7973	0,1725	0,1875	0,4175	0,4296	0,2576	0,1763	0,3075	0,1813	0,2624
0,9925	0,4868	0,1875	0,7295	0,4975	0,5802	0,3491	0,3325	0,2542	0,4073
0,9437	0,4178	0,1875	0,6363	0,4881	0,4967	0,3187	0,2894	0,2440	0,3783
0,9295	0,2703	0,1875	0,5761	0,4689	0,3843	0,2341	0,2804	0,2074	0,3240
0,8869	0,1725	0,1875	0,4360	0,4495	0,2612	0,1763	0,2733	0,1813	0,266
0,9665	0,4868	0,1875	0,7195	0,4906	0,5761	0,3491	0,2894	0,2542	0,4048
0,992	0,4178	0,1875	0,6474	0,4973	0,4989	0,3187	0,2909	0,2440	0,3801
0,94380	0,2703	0,1875	0,5803	0,4722	0,3856	0,2341	0,2811	0,2074	0,3249
0,9289	0,1725	0,1875	0,4682	0,4827	0,2725	0,1763	0,2876	0,1813	0,2796
0,9516	0,4868	0,1875	0,7055	0,4806	0,5693	0,3491	0,2856	0,2542	0,3995
0,9491	0,4178	0,1875	0,6306	0,4837	0,4925	0,3187	0,8696	0,2440	0,3747
0,9927	0,2703	0,1875	0,6047	0,498	0,3957	0,2341	0,2909	0,2074	0,3349

Key: 1) Attributes; 2) states.

indistinguishable states which give equal values of x_1 or x_2 (see Table 9: states S_1 and S_6 for x_1 etc). Finally, the attributes of group d are useless for our problem, since they are connected only with $\hat{\Lambda}$.

The attributes of group $\underline{a}$ should be recognized as more useful also because, even when the slope of the extremal hill is increased, it is always possible to find among them an attribute which increases monotonically on both sides of the "crest" and hence uniquely determines its position. In the given problem it is sufficient to use only the attribute x_7 to distinguish all 16 given states.

The boundaries of situations with ideal and real attributes

If we provide the system with a device or program which, after every change in the location and form of the characteristic of the open part, perform a change in the locus of reference and a transformation of the coordinates μ and λ for which this characteristic is rectilinear, is always located at an angle of 45° and passes through the origin ($d_0 = 0$; $d_1 = 1$; $d_2 = 0$), the investigation takes a finer form. Instead of the six-dimensional space

$$\Omega_{\alpha i}(c_0, c_1, c_2, d_0, d_1, d_2)$$

we can consider the three-dimensional coordinate space $\Omega_{vi}(x,y,z)$, where

$$x = c_0 - d_0 = c_0, \quad y = c_1 - d_1 = c_1 - 1, \quad z = c_2 - d_2 = c_2.$$

The correction problem consists of bringing the system into the situation including the origin as center $x = 0$, $y = 0$, $z = 0$.

The larger the radius vector

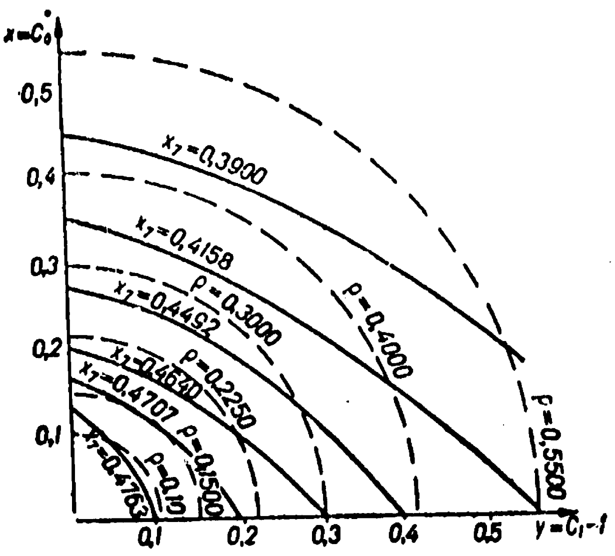


Figure 69. Situation boundaries for the attribute x_7 and for the"ideal" attribute ρ (where $d_0=d_2=c_2=0$ and $d_1=1$).

$$\rho=\sqrt{x^2+y^2+z^2},$$

the more is correction required. Hence a good division of the space x,y,z into situations would be division by concentric spheres wiith a common center at the origin (Fig.69, dashed line), and an ideal attribute could be the radius vector itself if it could be quickly and easily measured and computed. But this is not the case, and hence we should use the much simpler attributes indicated.

For each of the attributes, the situation boundaries can be constructed experimentally or computed pointwise. For the example considered above, the situation boundaries when only the one attribute x_7 is used are shown in Fig.69 by solid lines. We can only come to the conclusion that, although the form of the boundaries differs from the ideal, it is nevertheless by nature near the ideal in the square $x > 0$, $y > 0$.

Method of producing attributes in the presence of deviations of the perturbation distribution from the most probable curve (distribution transformation)

The method consists of selecting only those points which correspond to the mean perturbation distribution. Measurements which violate the given mean distribution are simply omitted.

Let us illustrate the method by an example. Let us assume that the mean or most probable distribution is the uniform distribution as in Fig. 68. Then by means of special selection filters, it is necessary to omit the points which disturb uniformity. For example, let

$$\lambda_1 = 0{,}25,\ \lambda_2 = 0{,}5,\ \lambda_3 = 0{,}25,\ \lambda_4 = 0{,}5,\ \lambda_5 = 0{,}75.$$

Then the filter must omit each value by one, i.e. the points $\lambda_1, \lambda_2, \lambda_5$. It is clear that this process of producing the attributes is delayed, since it is necessary to wait until all values of λ arrive. In the given example, the process will be delayed by two thirds of a period.

In practice, the attribute pickup can be constructed according to a principle which recalls the principle of construction of the eye of certain insects (for example, of the bee).

At all vertices of the teeth of the characteristic of the open part memory devices are placed which record

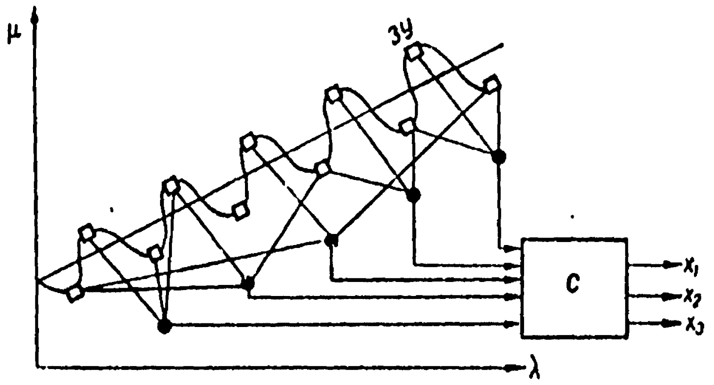


Figure 70. Structure of attribute pickup with separation of the effect of uniform perturbation distribution from any other (different from zero): 3Y) memory device for the latest value of tl ɜ quality index $\bar{\Phi}$; C) selector of group of three memory devices with minimal delay τ_L.

the latest value of the extremum indicator (Fig.70). The devices are connected in groups of three, but in such a way that these groups do not include points lying on a single straight line. Two extreme devices are placed at equal distances from the mean (in the case of separation of uniform distribution). All groups of three memory devices feed their signals to a selector. ^{T}he latter selects the group in which the delay time (equal to half the period during which all three devices have operated) is less than in the other groups. With a parabolic characteristic, the memory devices are connected in groups of four, in all possible combiaations, excluding those which give zero values of the determinants or violate the perturbation distribution to be separated.

Evaluation of attributes according to the criterion of resolving power

If the attributes are measured accurately, and the perturbation is a displacement of the extremal hill along the $\mu - \lambda$ plane without changing its form, and, furthermore, the perturbation probability distribution remains invariable, the recognition system can distinguish any number of states equal to the number of its groups. But in fact there always exists deviation from these ideal conditions. This brings it about that the system cannot distinguish states which are very close to one another. Hence there arises the problem of all-around raising of the resolving power of the syst , a definition of which is given in [21]. Let us recall that the resolving power is defined by the difference between the greatest scalar product and the one nearest it in magnitude.

The algorithm of a recognition system is such that the attributes can only be useful or useless in some respect (Wiener: "There is no evil, but there is an absence of good"). Hence the more attributes fed to the system, the higher its resolving power, although the volume of the system grows due to this. In striving to decrease the volume, we should select combinations of attributes which for an almost identical volume of the system give the greatest resolving power.

Let us illustrate the method of calculating the resolving power by an example. Let us denote by $x_1, x_2, \ldots, x_n$ the digits of the vector of the input image (attribute), and by $r_1, r_2, \ldots, r_n$ the corresponding digits of the poles. Then with $v_i(x_1, x_2, \ldots, x_n)$ and $\alpha_k(r_1, r_2, \ldots, r_n)$

$$\Sigma_k = (\alpha_k v_i) = r_1 x_1 + r_2 x_2 + r_3 x_3 + \ldots + r_n x_n \to \max.$$

It is obvious that the maximum of the scalar product coincides with the minimum difference of the individual digits. Hence in the algorithm of the Alpha system, we can, instead of the selection of the greatest scalar product (by means of the highest-voltage indicator HVI), use the selection of the least difference (by means of a lowest voltage indicator LVI)

$$\Sigma_k = (v_i - \alpha_k) = (r_1 - x_1) + (r_2 - x_2) + \ldots + (r_n - x_n) \to \min.$$

The scalar variant is convenient with binary attributes and for a unitary code. The difference variant of the algorithm is convenient for continuous (nonbinary) attributes, since it makes possible the simple use of a binary code. Let us dwell on the use of the difference variant of the algorithm of the Alpha system. Let us compare the resolving powers of three systems: 1) with one attribute x_7; 2) with two attributes x_1 and x_2; 3) with three attributes x_1, x_2 and x_7. In the learning process, after all 16 states have been displayed (see Fig.64), the poles of the groups will assume the following values:

In the system according to attribute x_7:

$r_1 = 0{,}4823$, $r_5 = 0{,}4490$, $r_9 = 0{,}4586$, $r_{13} = 0{,}4755$,
$r_2 = 0{,}6907$, $r_6 = 0{,}7315$, $r_{10} = 0{,}7257$, $r_{14} = 0{,}7188$,
$r_3 = 0{,}6266$, $r_7 = 0{,}6436$, $r_{11} = 0{,}6493$, $r_{15} = 0{,}6403$,
$r_4 = 0{,}5997$, $r_8 = 0{,}5921$, $r_{12} = 0{,}5945$, $r_{16} = 0{,}6078$,

In the system according to attributes x_1 and x_2:

$r_{11} = 0{,}9975$, $r_{21} = 0{,}4833$, $r_{15} = 0{,}9267$, $r_{25} = 0{,}4833$,
$r_{12} = 0{,}9433$, $r_{22} = 0{,}7333$, $r_{16} = 0{,}9975$, $r_{26} = 0{,}7333$,
$r_{13} = 0{,}9702$, $r_{23} = 0{,}6510$, $r_{17} = 0{,}9832$, $r_{27} = 0{,}6510$,
$r_{14} = 0{,}9832$, $r_{24} = 0{,}6093$, $r_{18} = 0{,}9754$, $r_{28} = 0{,}6092$,
$r_{19} = 0{,}9595$, $r_{29} = 0{,}4833$, $r_{1-13} = 0{,}9752$, $r_{2-13} = 0{,}4833$,
$r_{1-10} = 0{,}9887$, $r_{2-10} = 0{,}7333$, $r_{1-14} = 0{,}9834$, $r_{2-14} = 0{,}7333$,
$r_{1-11} = 0{,}9974$, $r_{2-11} = 0{,}6510$, $r_{1-15} = 0{,}9827$, $r_{2-15} = 0{,}6510$,
$r_{1-12} = 0{,}9802$, $r_{2-12} = 0{,}6093$, $r_{1-16} = 0{,}9976$ $r_{2-16} = 0{,}6093$,

In the system according to the attributes x_1, x_2 and x_7, the poles are the same as in the first two systems. Others will only be subscripts of r_k. Summing up, according to the above formula, the voltages across the outputs of all 16 groups appearing in the display of each of the 16 central states, we can select the least difference and by this means can determine the resolving power of each of the systems being compared.

For brevity, we have cited as an example only one table (Table 10), for the first system with one attribute x_7.

An inspection of Table 10 shows that, in the first system with one attribute x_7, R=0025; in the seoond with two attributes x_1 and x_2, R=0.003; and in the third system with three attributes x_1,x_2 and x_7, R=0.0073.

As the number of attributes being used increases, the resolving power does in fact increase. Thus, in this sense the best of the systems compared is the third system. It allows the greatest oscillation of the perturbation probability distribution curve and the form of the hill without making mistakes. When mistakes are present, the number of attributes should be increased.

A sample circuit for using the recognition system as a corrector

As shown in Fig.65, the recognition system is used for establishing on a model the parameters c_0,c_1,c_2 of the optimal characteristic of the object of control. The coefficients d_0,d_1,d_2 are always known.

Hence in order sharply to decrease the number of groups of the recognition system, we can use switching of the poles in dependence on the position of the characteristic of the open part $r_k(d_0,d_1,d_2)$.

Table 10

Выход (1) (2) Состояния	z_1	z_2	z_3	z_4	z_5	z_6	z_7
S_1	0	0,2084	0,1443	0,1174	0,0333	0,2492	0,1613
S_2	0,2084	0	0,0641	0,0910	0,1417	0,0408	0,0471
S_3	0,1443	0,0641	0	0,0269	0,1776	0,1049	0,0170
S_4	0,1174	0,091	0,0269	0	0,1507	0,1318	0,0439
S_5	0,0333	0,2417	0,776	0,1507	0	0,2825	0,1946
S_6	0,2492	0,0408	0,1049	0,1318	0,2825	0	0,0879
S_7	0,1613	0,471	0,017	0,0439	0,1946	0,0879	0
S_8	0,1097	0,0986	0,0346	0,0770	0,1430	0,1395	0,0516
S_9	0,0237	0,2321	0,168	0,1411	0,0096	0,2729	0,1850
S_{10}	0,2434	0,035	0,0991	0,1260	0,2767	0,0058	0,0821
S_{11}	0,167	0,0414	0,0227	0,0496	0,2003	0,0822	0,0057
S_{12}	0,1122	0,0962	0,0321	0,0052	0,1455	0,1370	0,0491
S_{13}	0,068	0,2152	0,1511	0,1242	0,0265	0,2560	0,1681
S_{14}	0,2365	0,0281	0,0922	0,1191	0,2698	0,0127	0,0752
S_{15}	0,158	0,0505	0,0137	0,0406	0,1913	0,0912	0,0033
S_{16}	0,1255	0,0829	0,0188	0,0081	0,1588	0,1237	0,0358

Key: 1) Output; 2) state.

Voltages across the outputs of groups in the system with one attribute with learning sequences (see Fig.68)

Σ_{8}	Σ_{9}	Σ_{10}	Σ_{11}	Σ_{12}	Σ_{13}	Σ_{14}	Σ_{15}	Σ_{16}
0,1097	0,0237	0,2434	0,1670	0,1122	0,0680	0,2365	0,1580	0,1255
0,0986	0,2321	0,0350	0,0414	0,0962	0,2152	0,2810	0,0504	0,0829
0,0346	0,168	0,0991	0,0227	0,0321	0,1511	0,0922	0,0137	0,0188
0,0077	0,1411	0,1260	0,0496	0,0052	0,1242	0,1191	0,0406	0,0081
0,143	0,0096	0,2767	0,2003	0,1455	0,0265	0,2698	0,1913	0,1588
0,1395	0,2729	0,0058	0,0822	0,1370	0,2560	0,0127	0,0912	0,1237
0,0516	0,185	0,0821	0,0057	0,0491	0,1681	0,0752	0,0033	0,0358
0	0,1334	0,117	0,0573	0,0025	0,1165	0,1268	0,0483	0,0158
0,1334	0	0,2671	0,1907	0,1359	0,0169	0,2602	0,1817	0,1492
0,1337	0,2671	0	0,0761	0,1312	0,2502	0,0069	0,0854	0,1187
0,0573	0,1907	0,0769	0	0,0548	0,1733	0,0696	0,0090	0,0415
0,0025	0,1359	0,1312	0,0548	0	0,1190	0,1243	0,0458	0,0133
0,1165	0,0169	0,2502	0,1738	0,119	0	0,2433	0,1648	0,1323
0,1268	0,2602	0,0069	0,0695	0,1243	0,2433	0	0,0785	0,1110
0,0483	0,1817	0,0854	0,0090	0,0458	0,1648	0,0785	0	0,0325
0,0158	0,1492	0,1187	0,0415	0,0133	0,1323	0,1110	0,0325	0

In the above example, this gives a r·duction in the grouos from 16 to 4 (Fig.65), in accordance with the number of combinations to be distinguished of the values of the coefficients c_0, c_1, c_2(Fig.66).

For a corrector of the "velocity" type, where it it only required. to indicate the direction of control, it is sufficient to compare the actual object with the model $\Delta\mu = \mu_{opt} - \mu$ in order to work out the corresponding signal:

when $\Delta\mu < -\delta$ -- "regulate, riase the characteristic of the open part into the region of the given λ".

when $-\delta < \Delta\mu < \delta$ -- "hold"

when $\Delta\mu > \delta$ -- "regulate, lower characteristic".

In order to construct a corrector of the positional type, the optimal value of the regulating influence μ_{opt} is directly established at the open part of the system by means of a servomechanism.

The possibility of corrector self-learning

By means of the above example of a corrector, we can again demonstrate the difference between two opposite approaches to the solution of the control problem.

The determinate approach consists of obtaining an algorithm for the control object and the perturbations acting on it and then solving the equations on computers.

In the problem considered, it is reduced to the computation of the coefficients c_0, c_1, c_2 from the formulas of regression analysis, which, in the first place, requires the presence of exact information on the object and the perturbations and, in the second, a large machine memory volume, while it it impermissible to spend large amounts of time data averaging and equation solving.

The cybernetic method consists of replacing exact calculations by learning of a recognition system from the results of experiments on a real object with minimal initial information. The algorithm of the object may be too complicated for control or unknown altogether. The perturbation distribution is also unknown. The complicated "teacher" algorithm is used only during the learning time. The "teacher" may be a man or an interpolator based on the methods of active or passive experiment. Learning is carried out according to records of the operation of the object in the past, as is done for recognition systems operating as prediction filters. In the problem considered, after completion of the learning, situation determination requires only computation of one simple attribute or several. Thus, the problem of determining the exact values of the coefficients c_0,c_1,c_2 is replaced by the problem of dividing the space c_0,c_1,c_2,d_0,d_1,d_2 (or the space of the attributes $x_1,x_2,x_3,\ldots,x_n$) into regions, i.e. situations.

If, instead of a"teacher", we use positive feedback, then, as in the case of self-arbitrary distinction of letters, the Alpha system teaches itself to distinguish situations.

However, just as the recognition system cannot correctly name the letters without indications from outside, in this application the "teacherless" system can by nature not assign values of the quantities c_0,c_1,c_2 for every situation distinguished by it.

The names or quantities can only be indicated by the "teacher", or in the other case they can be worked out in the process of concurrent "survival" from a larger number of systems in which those names are assigned by situations in a random manner.

The recognition system as a positional corrector for systems for controlling cyclical processes

First of all let us convince ourselves that the recognition system can distinguish input signals from any of their parts, and in particular from their initial parts. As an example we again use the algorithm of the Alpha system. Let there be two input signals:

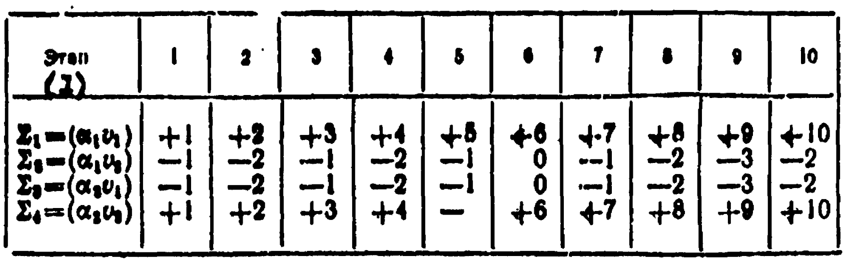
$$v_1 = +1 -1 -1 +1 -1 -1 -1 +1 +1 +1,$$
$$v_2 = -1 +1 -1 -1 -1 -1 -1 -1 -1 +1.$$

In the learned state, the poles of the system have the same codes:

$$\alpha_1 = +1 -1 -1 +1 -1 -1 -1 +1 +1 +1,$$
$$\alpha_2 = -1 +1 -1 -1 -1 -1 -1 -1 -1 +1.$$

We can obtain such a unitary code with several plusses if we do not use the convergence scheme at the input. When the latter are used, there will be a plus only at one place of the code, but this does not essentially change the rest of the outputs and is reflected only in the volume (number of elements) of the system.

Let us determine the scalar products at every stage. (It is assumed that the digits of the code are determined by degrees: first, only a few initial digits are known, then the following one is addod, etc.).



Этап (1)	1	2	3	4	5	6	7	8	9	10
$\Sigma_1=(\alpha_1 v_1)$	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
$\Sigma_2=(\alpha_1 v_2)$	−1	−2	−1	−2	−1	0	−1	−2	−3	−2
$\Sigma_3=(\alpha_2 v_1)$	−1	−2	−1	−2	−1	0	−1	−2	−3	−2
$\Sigma_4=(\alpha_2 v_2)$	+1	+2	+3	+4	−	+6	+7	+8	+9	+10

Table 11. Scalar products
Key: 1) Stage.

We have convinced ourselves from the example that the recognition system distinguishes the input signals v_1 and v_2 even from the initial parts of their codes, since we always have:

$$\text{for } v_i = v_1 \quad (\alpha_1 v_1) > (\alpha_2 v_1),$$
$$\text{for } v_i = v_2 \quad (\alpha_1 v_2) < (\alpha_2 v_2).$$

Precisely this serves as the basis for the roliability of recognition systems when a large number of pickups goes out of order. Just as the living organism continues to function when one of its parts goes out of order, the recognition system also continues to function correctly in similar circumstances. We can only show that this lowers its resolving power [21] When the length of the code is increased, the resolving power increases.

If two representation curves which lead to different results coincide at first, the recognition system gives the same voltages at two outputs, i.e, says "I don't know" right up until the moment that the representation curves diverge.

Let us use this property of recognition systems of predicting finite evaluations (at first unsurely, and then more and more accurately) to construct a combined determinate-self-learning system (Fig.71). The open part of the system is shown in Fig.72. The corrector is an Alpha recognition system with five groups of associating cells (neurons) according to evaluations 1,2,3,4,5. The input of the system is fed sequences of the variation of the coordinates L, T and K (in a code with a "signifying plus").

For example:

$$v_i = -1 - 1 + 1 - 1 - 1 - 1 + 1 - 1 + 1 - 1 - 1.$$

The coordinate K represents the number of the closed key of the open part and reflects the regulating influence μ.

Using the evaluations obtained at the end of the cycles, the system first of all learns to distinguish codes from the resulting estimates.

The pole learning algorithm of the recognition system, which is used as a postional corrector, was considered above, and hence we shall not go over it again.

Recognition predicting systems are used in a time region. If the same representation curve obtains the continuouslyvarying estimates $\phi_1, \phi_2, \phi_3, \dots \phi_n$, the process of variation of the evaluations being random and stationary, then, using Kolmogorov's formula, it is possible to predict the future code of variation of the evaluation during one cycle. This can be used for increasing the accuracy of control if definite conditions are satisfied (random process being stationary, where Kolmogorov's formula holds).

If during the variation of the evaluations there are observed periodic and some other repetitive variations, then for predictive purposes the method of characteristic

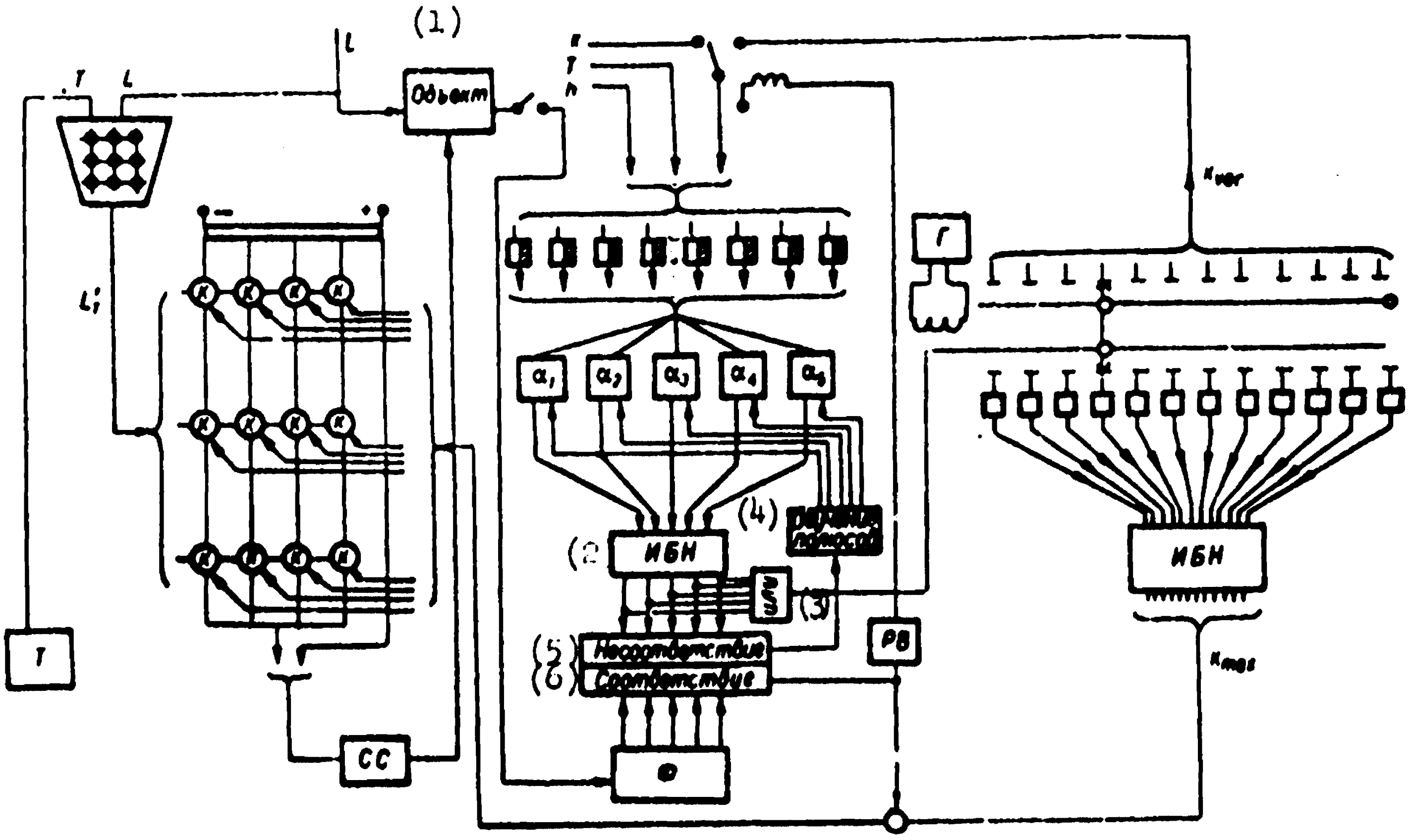


Figure 71. Corrector: recognition system for control of a cyclical process.

Key: 1) Object; 2) HVI; 3) or; 4) learning of poles; 5) non-corespondence; 6) correspondence.

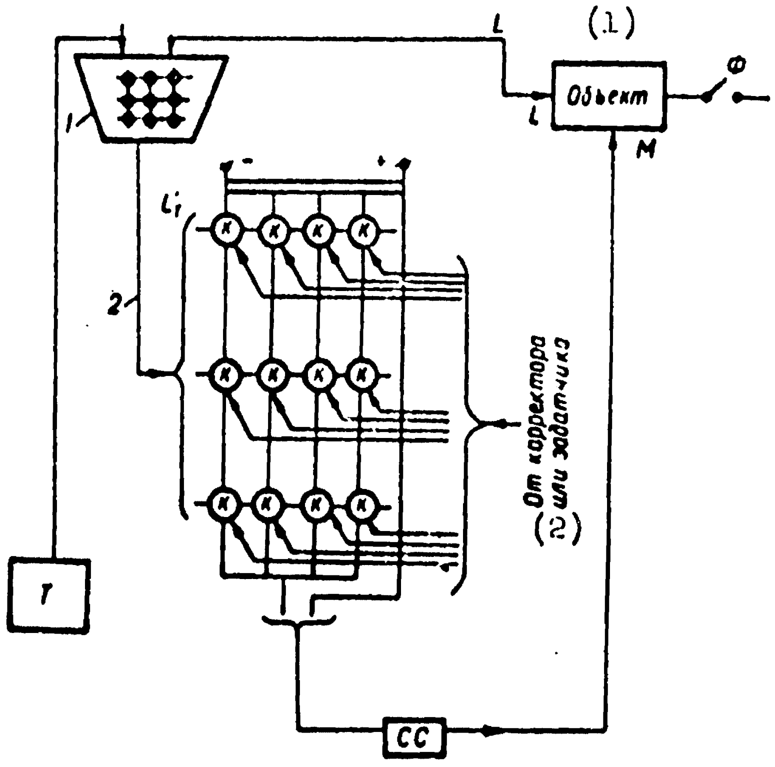


Figure 72. Open part of system for controlling a cyclical process.

Key: 1) Object; 2) from corrector or setter.

components or the combined method (see above) is used). In one way or another it is possible approximately to determine the future evaluation of the given cycle of influences.

We showed above that the recognition system can distinguish representation curves from their initial parts. This is also one of the forms of using recognition systems for prediction.

A quite different example of possible use of the Alpha predicting system is shown in Fig.73. Here the sys-

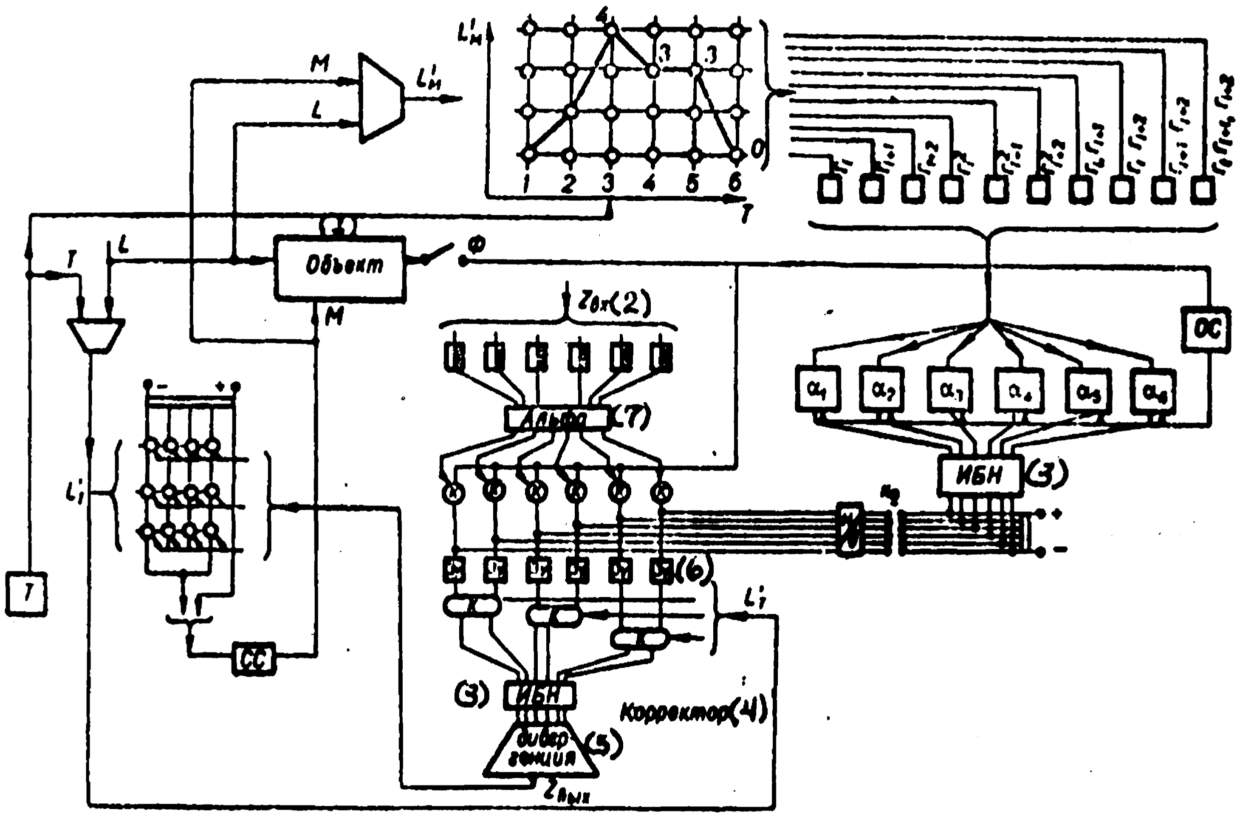


Figure 73. Nonreversible corrector with additional prediction of a given sequence of modes.

Key: 1) Object; 2) Z_{inp}; 3) HVI; 4) corrector; 5) divergence; 6) memory device; 7) Alpha.

tem is used for accelerating the obtaining of the evaluation of the cycle. The curve of the generalized perturbation L'_m as a function of the stage number of the program T is considered to be a random process. This means that we can use the Alpha system (or any other recognition system) to predict both future values of the perturbation L'_m and the directly correlated with it cycle quality index Φ, analogously to what we did in predicting the amplitude of sea waves. As soon as the system has learned and is essentially predicting correctly, a button K is pressed. When this is done, memory devices record the predicted value of the quality index. The system can either continue the motion (if the memory installation has remained the largest), or change the code and pass to the representation curve following in evaluation magnitude (if the memory installation has become lower than other memory installations). Thus, the system automatically takes into account the prediction of the result of its operation.

At the beginning of each given stage, we have at our disposal complete information on the value of all coordinates, except the coordinate k which is to be selected.

For this purpose, we can use a search on a taught recognition system (and not on the object). Feeding all possible values of k to the input of the recognition system, we select those for which the predicted evaluation is highest.

Perceptron for prediction of the result of cyclical processes

Prediction algorithm. The fundamental difference between the "complete" perceptron and the simpler recognition systems (for example, the Alpha recognition system [21]) consists of the fact that in the perceptron, images are not recognized from one averaged prototype or standard, but from many random prototypes. During the learning process there is established a "weighting factor" or degree of participation of each of the random prototypes in the formation of the given pattern, which is then used for classifying images into patterns or classes. Typical operation of this method is shown in the experiments of Bryan [51].

We shall use this principle of many random prototypes for predicting the result of a cyclical process. Figure 74 shows a circuit for the perceptron as a predicting filter.

The process whose result we are required to predict is known to us in some one of its initial parts during the course of n cycles (time intervals). The duration of the whole cycle is taken to be 100 units; hence $0 < n \leqslant 100$. During the course of the section of the process known to us, the latter can be represented by the vector

$$v_i(n) = x_1, x_2, x_3, \ldots, x_n,$$

which is also an "image" subject to recognition. With each new cycle, the number of meansurements of the vector increases by one (Fig.75). The coordinates $x_1, x_2, \ldots, x_n$ are called "attributes" of the given image. The problem is this: from observation of the changes of the vector v_i, to predict its coordinate at the end of the process h_{100} or indicate the maximum value h_{max}.

As random prototypes, we can of course use purely random point or curvilinear masks, as was done in Bryan's experiments. But then we would lose the information in the known realizations of the process, which would lead to an increase of the volume of the system and of the duration of its operation. To simplify the system, we can use as

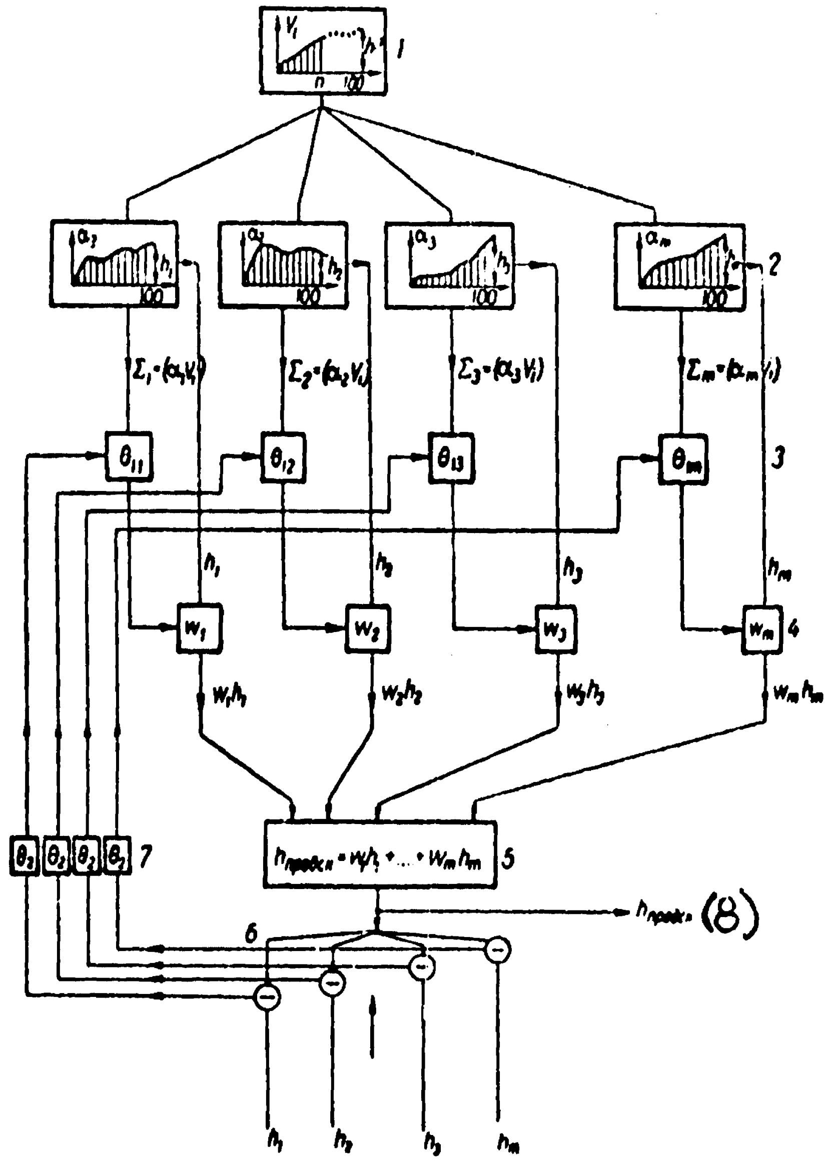


Figure 74. Perceptron circuit: 1) observed process; 2) standard processes; 3) threshold elements ϑ_1 (associating cells); 4) weight regulation; 5) summator; 6) feedbacks; 7) threshold elements ϑ_2. Key: 8) $h_{predict}$.

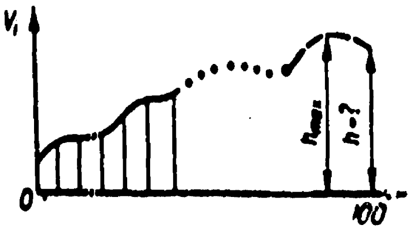


Figure 75. Observed process whose result requires to be predicted.

random prototypes previous realizations of the given process whose results are known.

Prototypes	Results
$\alpha_1(r'_1, r'_2, \ldots, r'_n)$,	h_1,
$\alpha_2(r^{\bullet}_1, r^{\bullet}_2, \ldots, r''_n)$,	h_2,
. . . .	. .
$\alpha_m(r_1^{(m)}, r_2^{(m)}, \ldots, r_n^{(m)})$.	h_m.

The dimension of the prototypes is equal to the dimension of the image and hence increases by one with each new cycle. Further, in accordance with the algorithms of operation of the perceptron with many random prototypes, these scalar products should determined:

$$\Sigma_1 = (\alpha_1 v_i),$$
$$\Sigma_2 = (\alpha_2 v_i), \ \ldots,$$
$$\Sigma_m = (\alpha_{,i} v_i).$$

These scalar products are a measure of the nearness of the image to the prototype in the attribute space. Since it is important to take into account only the presence of divergence of curves and not the sign of this divergence, the ordinates are divided by the square of the largest of them.

Example. Let us assume that we are given these initial data:

n	1	2	3	4	5
v_i	$x_1=30$	$x_2=40$	$x_3=40$	$x_4=50$	$x_5=70$
α_k	$r_1=20$	$r_2=40$	$r_3=50$	$r_4=50$	$r_5=60$

Then the scalar product is

$$\Sigma=(v_i\alpha_k)=\frac{1}{n}(x_1r_1+x_2r_2+x_3r_3+x_4r_4+x_5r_5)=$$
$$=\frac{1}{5}\left(\frac{30\cdot 20}{30^2}+\frac{40\cdot 40}{40^2}+\frac{40\cdot 50}{50^2}+\frac{50\cdot 50}{50^2}+\frac{70\cdot 60}{70^2}\right)=$$
$$=\frac{454}{525}\approx 0{,}862.$$

The terminus of the vector v_i is called the"representation point", and those of the vectors of the prototypes α_k the "poles". If representation point and pole coincide, i.e. $v_i=\alpha_k$, the scalar product of the vectors is equal to the greatest value, i.e. unity $(\Sigma_{max}=1)$.

Of course, we can also use other measures for the closeness of the representation point to this or that pole. For example, sometimes use is made of the square of the distance between them (square error). Let us limit ourselves to the use of the scalar products, which are also correlation coefficients.

Summing up the scalar products (for every n-th cycle of the observed process), in accordance with the perceptron algorithm, must select only the greatest of them, namely those values which exceed some threshold, when $0\leqslant\Theta_{lj}\leqslant 1$. This produces selection of the prototypes which are sufficiently near to the observed process. If we selected a very high threshold, so that there remained only one prototype, this would bring us from the complete perceptron to the simplified one, i.e. the Alpha recognition system.

The scalar products or, in other words, the voltages of the threshold elements (associating cells) which exceed the threshold Θ_{1j} determine the weighting factors (degree of participation) w_i with which the results of the corresponding standard processes are summed:

$$w_1 = \frac{\Sigma_1}{\Sigma_1 + \Sigma_2 + \ldots + \Sigma_m}, \quad w_2 = \frac{\Sigma_2}{\Sigma_1 + \Sigma_2 + \ldots + \Sigma_m}, \ldots,$$

$$w_m = \frac{\Sigma_m}{\Sigma_1 + \Sigma_2 + \ldots + \Sigma_m}$$

where the scalar products $\Sigma_j \leqslant \Theta_{1j}$ should be put equal to zero).

The predicted result of the observed process is determined by a summator

$$h_{\text{predict}} = w_1 h_1 + w_2 h_2 + \ldots + w_m h_m.$$

Perceptron learning. For correct prediction it is necessary to select the values of the thresholds Θ_{1j} of the association cells. This is achieved by means of learning from the known realizations of the process, which form a learning sequence. The processes which make up the learning sequence are not among the standard random processes. Learning is first carried out with sufficiently large and constant n=const, and then with n=var.

For learning, the prediction result h_{predict} is compared with the result of each of the standard processes which gave a voltage higher than the threshold value Θ_{ij}. Determination is made of the square error

$$\Delta_i = (h_{\text{предск}} - h_i)^2, \quad 0 < i < m.$$

предск = predict

If the square error is larger than some second threshold value Θ_2, the threshold Θ_1 of the corresponding associating cell increases by a small interval $\Delta\Theta_1$ or according to an exponential law $\Theta_{n+1} = \Theta_n + (1-\Theta_n)\delta$, where

$0 \leq \delta \leq 1$. This decreases "trust" in the given standard process and its role in the prediction. During the following cycle and in predicting other processes, the role of the given standard will be weakened. Conversely, if it turns out that the square error is sufficiently small, the threshold of the corresponding associating cell is lowered by a small constant interval or according to an exponential law.

The role of "correctly operating" standards is accordingly increased, which is required for raising the prediction accuracy.

Elements of Stability Theory and the Theory of Invariance of Combined Systems Containing Predicting Filters

In an automatic control system, the links containing predicting filters are called probabilistic links.

In dealing with the problems of stability theory and the theory of invariance of systems containing probabilistic links, it is first of all desirable to determine the transfer operator functions of the predicting filters. Let us consider the two simplest examples of linear filters.

Discrete predicting filters

A function $\varphi(t)$ is predicted from the first unary terms of the prediction formula and from the observation data from time $t=-T_1$ to $t=-T_2$ (Fig.76).

In the given example we use a formulation of the problem which is typical for self-learning pickups; the precdding values of a quantity are known; it is required to determine its value at a given time.

It is required to predict the value of the function at time t=0. We divide the interval T_1-T_2 into n equal

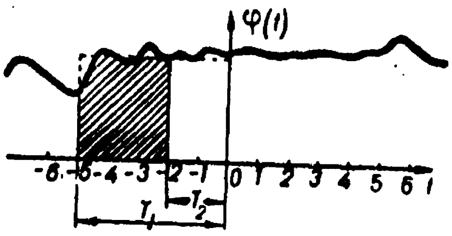


Figure 76. Prediction of future value of function $\varphi(t)$ as mean value during averaging time $T=T_1-T_2$.

segments with duration Δt and we find

вер = prob

$$\varphi_{\text{вер}}(0) = \frac{r_1\varphi_1 + r_2\varphi_2 + \ldots + r_n\varphi_n}{r_1 + r_2 + r_3 + \ldots + r_n},$$

where $r_1, r_2, \ldots, r_n$ are the coefficients of the "forgetting law" (weight), which are determined during the learning process of the predicting filter; φ_1 is the value of φ when $t=T_2$; φ_n is the value of φ when $t=T_1$.

For simplicity, let us first put

$$t = T_2,\ r_1 = r_2,\ r_n = 1.$$

Then

$$\varphi_{\text{вер}} = \frac{1}{n}\left[\varphi z^{-1} + \varphi z^{-2} + \ldots + \varphi z^{-n}\right],$$

where $z=1^{\Delta tp}$.

The desired transfer function is

вых = output
вх = input

$$P(p) = \frac{\varphi_{\text{вых}}}{\varphi_{\text{вх}}} = \frac{z^{-1} + z^{-2} + z^{-3} + \ldots + z^{-n}}{n}.$$

If the weighting factors are not equal to one, we obtain

$$P(p) = \frac{\varphi_{\text{вых}}}{\varphi_{\text{вх}}} = \frac{r_1 z^{-1} + r_2 z^{-2} + \ldots + r_n z^{-n}}{r_1 + r_2 + \ldots + r_n}.$$

Continuous predicting filter

Let us consider the simplest continuous proababilistic link, which predicts the future mos t probable value as the mean value during a certain observation time (see Fig.76):

$$\varphi_{\text{вср}} = \frac{1}{T_1 -- T_2} \int_{t-T_1}^{t-T_2} \varphi(t)\, dt.$$

Thus we are predicting many events: if in the course of a number of past days there has been good weather, it is highly probable that there will be good weather tomorrow, etc.

For more exact prediction, the time T_2 should be as small as possible (in some cases $T_2=0$), and the averaging interval $\Delta T=T_1-T_2$ is selected depending on the nature o the curve $\varphi(t)$. It must be several times larger than the period of the fundamental harmonic of the expansion of this curve in a harmonic series.

The operation of such a predicting device can be described by the equation

$$\varphi_{\text{вср}} = \frac{1}{T_1 - T_2} \int_{t-T_1}^{t-T_2} \varphi(t)\, dt = \frac{1}{T_1 - T_2} \left[-\int_{-\infty}^{t-T_1} \varphi(t)\, dt + \int_{-\infty}^{t-T_2} \varphi(t)\, dt \right] = \frac{1}{T_1 - T_2} \left[\int_{t-T_2}^{0} \varphi(t)\, dt - \int_{t-T_1}^{0} \varphi(t) dt \right].$$

In operator form we obtain the following transfer function:

$$P\varphi(p) = \frac{\varphi_{\text{вых}}}{\varphi_{\text{вх}}} = \frac{\varphi_{\text{вср}}}{\varphi} = \frac{1}{T_1 - T_2} \cdot \frac{1}{p} \left(e^{-T_2 p} - e^{-T_1 p}\right).$$

If the "forgetting function" is given, for example exp(-(t)/τ),in the above expressions we should replace

φ (t) by

$$\varphi(t)e^{-\frac{(t)}{\tau}}.$$

Invariance conditions for systems with probabilistic learning links

Below we introduce and partially investigate the conditions for absolute invariance and conditions for the stability of systems with probabilistic links, circuits for which are shown in Fig.77. This figure shows all the systems most important in practice:

a) one-circuit servo-system with links according to the fundamental perturbation (input signal),

b) one-circuit **stabilisation** system with links according to the fundamental perturbation (load of the object of regulation),

c) two-circuit (differential) servo-system without perturbation links,

d) two-circuit (differential) stabilization system without perturbation links.

The cross-hatched squares are devices which compute probabilistic values. It is not necessary that there be two probabilistic links in each of the systems considered. Some of the links may be determinate, i.e. the usual ones. In this case, the transfer function of the corresponding square should be taken equal to one. The square P_{Φ} (p) represents the learning feedback, P_{ψ} (p) an open learning link, P_{ψ} (p) [sic] a probabilistic link in a "system with learning prototype".

Let us proceed to the mathematical description and investigation of the systems of Fig.77. Table 12 shows

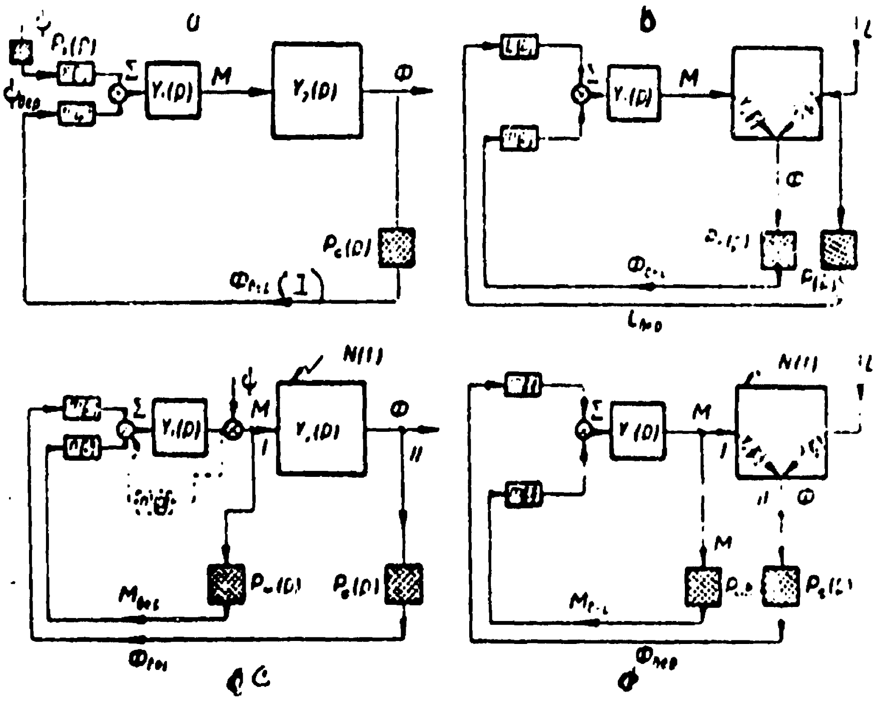


Figure 77. Basic circuits of systems with probabilistic links: a,b) with links according to basic perturbation; c,d) without links according to perturbation (differential systems).

Key: 1) probab .

the dynamical equations of the elements, and Table 13 those of the systems as a whole.

Stability. The characteristic equations, the signs of whose roots determine the stability of the systems, are given in Table 12. Even at this stage in our consideration we can make the following conclusions:

1. The stability of the stabilization systems and the servo-systems is determined by characteristic equations which are the same in structure.

2. Probabilistic (predicting) devices of the open links have no influence on the stability of the systems, since the operators $P_{\varphi}(p)$ or $P_L(p)$ do not enter into the stability conditions.

3. When probabilistic feedback is present in the system, when $P_{\Phi}(p) \neq 1$ or $P_M(p) \neq 1$, the stability of the system depends on the transfer functions of these links and in the general case greatly worsens in comparison with the determinate system (for which $P_{\Phi}(p) = 1$ and $P_M(p) = 1$), since the probabilistic links have a transport delay which is hard to compensate for.

4. Having the possibility of arbitrary selection of the coefficients of the external-feedback operator

$$m(p) = m_0 + m_1 p + m_2 p^2 + m_3 p^3 + \dots,$$

we can obtain a stable system for any sign and magnitude of the sign of the coefficients of the internal feedback:

$$n(p) = n_0 + n_1 p + n_2 p^2 + n_3 p^3 + \dots.$$

It is known that, when the absolute invariance conditions are fulfilled in systems without perturbation links, we should use positive internal feedback $n_0 > 0$. Quite recently many authors have affirmed that the system, as it were, must here arrive at the border of stability (loses

(1) Для рис. 77, *а*	(2) Для рис. 77, *б*
$\Sigma = k(p)\Psi_{\text{пер.}} - m(p)\Phi_{\text{вер}},$ $M = Y_1(p)\Sigma,$ $\Phi = Y_2(p)M,$ $\Phi_{\text{вер}} = P_\Phi(p)\Phi,$ $\Psi_{\text{пер}} = P_\psi(p)\Psi$	$\Sigma = -m(p)\Phi_{\text{пер}} + l(p)L_{\text{вер}},$ $M = Y_1(p)\Sigma,$ $\Phi = Y_2(p)M - \beta(p)L,$ $\Phi_{\text{вер}} = P_\Phi(p)\Phi,$ $L_{\text{вер}} = P_L(p)L$
(3) Для рис. 77, *в*	(4) Для рис. 77, *г*
$\Sigma = -m(p)\Phi_{\text{вер}} + n(p)M_{\text{вер}},$ $M = Y_1(p)\Sigma_1 + \Psi,$ $\Phi = Y_2(p)M,$ $M_{\text{вер}} = P_M(p)M,$ $\Phi_{\text{вер}} = P_\Phi(p)\Phi$	$\Sigma = -m(p)\Phi_{\text{вер}} + n(p)M_{\text{вер}},$ $M = Y_1(p)\Sigma,$ $\Phi = Y_2(p)M - \beta(p)L,$ $M_{\text{вер}} = P_M(p)M,$ $\Phi_{\text{вер}} = P_\Phi(p)\Phi$

Table 12. Dynamical equations of system elements (Fig.77)

(1) Для рис. 77, *а*	(2) Для рис. 77, *б*
$[1 + m(p)Y_1(p)Y_2(p)P_\Phi(p)]\Phi =$ $= k(p)Y_1(p)Y_2(p)P_\Psi(p)\Psi$	$[1 + m(p)Y_1(p)Y_2(p)P_\Phi(p)]\Phi =$ $= [l_1(p)Y_1(p)Y_2(p)P_L(p) - \beta(p)]L$
(3) Для рис. 77, *в*	(4) Для рис. 77, *г*
$[1 - n(p)Y_1(p)P_M(p) +$ $+ m(p)Y_1(p)Y_2(p)P_\Phi(p)]\Phi =$ $= Y_2(p)\Psi$	$[1 - n(p)Y_1(p)P_M(p) + m(p)Y_1$ $(p)Y_2(p)P_\Phi(p)]\Phi = -\beta(p)[1 - n$ $(p)Y_1(p)P_M(p)]L$

Table 13. Dynamical equations of systems as a whole (Fig.77)

Key (both tables): 1) For Fig.77a; 2) " 77b; 3) " 77c; 4) " 77d.

(1) Для рис. 77, а	(2) Для рис. 77, б
$1 + m(p)Y_1(p)Y_2(p)P_\Phi(p) = 0,$ $s = (1 + m_0\alpha_1\alpha_2P_{\Phi_0})$	$1 + m(p)Y_1(p)Y_2(p)P_\Phi(p) = 0,$ $s = (1 + m_0\alpha_1\alpha_2P_{\Phi_0})$
(3) Для рис. 77; в	(4) Для рис. 77, г
$1 - n(p)Y_1(p)P_M(p) +$ $+ m(p)Y_1(p)Y_2(p)P_\Phi(p) = 0,$ $s = (1 - n_0\alpha_1P_{M_0} + m_0\alpha_1\alpha_2P_{\Phi_0})$	$1 - n(p)Y_1(p)P_M(p) +$ $+ m(p)Y_1(p)Y_2(p)P_\Phi(p) = 0,$ $s = (1 - n_0\alpha_1P_{M_0} + m_0\alpha_1\alpha_2P_{\Phi_0})$

Table 14. Characteristic of equation of systems and expressions for rigidity S: (Fig.77)

(1) Для рис. 77, а	(2) Для рис. 77, б
$1 + m(p)Y_1(p)Y_2(p)P_\Psi(p) =$ $= k(p)Y_1(p)Y_2(p)P_\Psi(p)$	$L(p)Y_1(p)Y_2(p)P_L(p) - \beta(p) = 0$
(3) Для рис. 77, в	(4) Для рис. 77, г
$1 - n(p)Y_1(p)P_M(p) +$ $+ m(p)Y_1(p)Y_2(p)P_\Phi(p) = Y_2(p)$	$1 - n(p)Y_1(p)P_M(p) = 0$

Table 15. Conditions for absolute invariance of systems (Fig.77)

Key (both tables): 1) For Fig.77a; 2) " 77b; 3) " 77c; 4) " 77d.

its "coarseness"). It is obvious that this is not so. Systems without perturbation links can be adjusted both to positive and to zero or even to negative established and dynamic error with retention of their stability /20/. The systems (Fig.77c and d) can remain stable and "coarse" (i.e. such that small changes in the parameters do not essentially change their properties) in adjustment to absolute invariance.

Absolute invariance. The conditions for absolute invariance, for which for servo-systems $\Phi = \Psi$, and for stabilization systems $\Phi = 0$, are shown in Table 15. These conditions can be used for determining k(p), l(p) and n(p) which provide for ideal operation of the systems without established and dynamic error. The operators m(p) and $Y_1(p)$ will be used for selection of the necessary rigidity and stability according to the rules of compromise adjustment or by statistical methods.

When the invariance conditions in Table 15 are satisfied, the action is eliminated of all noise entering the system, in the "fork" of differential links, i.e. between points I and II (Fig.77c and d), including the action of statistically given perturbations, for example, noise of the "white noise" type (N(t) arrows in Fig.77c and d).

Example. Let us consider the synthesis of measurement links of the system in Fig.77b from the conditions of compromise adjustment and invariance for the case of the presence of only one probabilistic link-- according to the basic perturbation (load). Let us assume that we are given:

$$Y(p) = \frac{a_1}{\tau_1 p + 1},$$

$$Y_2(p) = \frac{a_2}{(\tau_2 p + 1)(\tau_3 p + 1)},$$

$$\beta(p) = \frac{\beta_0}{(\tau_2 p + 1)(\tau_3 p + 1)},$$

$$P_L(p) = \frac{1}{T_1 - T_2} \cdot \frac{1}{p} (e^{-T_2 p} - e^{-T_1 p}),$$

$$l(p) = \frac{l_0 + l_1 p + \dots}{l'_0 + l'_1 p + \dots},$$

$$m(p) = m_0 + m_1 p + m_2 p^2 + m_3 p^3 + \dots .$$

We select the operator of the closed link m(p) from the conditions for compromise adjustment, which ensures the optimal relationship between the rigidity and stability. If, for example, we require a rigidity s=100 for $\alpha_1 \alpha_2$=20, then obviously

$$m_0 = \frac{s-1}{a_1 a_2} = \frac{99}{20} \approx 5.$$

The remaining coefficients of the operator-- m_1, m_2, m_3-- are selected in order to ensure optimal damping of the free oscillations of the system (for example, so that in a system of the second order, the relative damping factor $C_{1,2}$=0.25; in a third-order system, the dimensionless parameters of Vyshnegradskiy x=1.2, y=3, etc). This procedure is well known, and we shall not dwell on it.

More complicated and interesting is the synthesis of the operator of the open link l(p) on the basis of the invariance conditions.

The conditions for absolute invariance (see Table 15) make it possible to determine (synthesize)

$$l(p) = \frac{\beta(p)}{Y_1(p)Y_2(p)P_L(p)} = \frac{\beta_0(\tau_1 p + 1)}{\alpha_1\alpha_2 P_L(p)}$$

when

$$P_L(p) = \frac{1}{T_1 - T_2}\cdot\frac{1}{p}(e^{-T_2 p} - e^{-T_1 p}).$$

We find

$$l(p) = \left[\frac{\beta_0(T_1 - T_2)}{\alpha_1\alpha_2}p + \frac{\beta_0(T_1 - T_2)\tau_1}{\alpha_1\alpha_2}p^2\right]\frac{1}{e^{-T_2 p} - e^{-T_1 p}}.$$

We draw the conclusion that in the given system the compounding perturbation link must have the form

$$l(p) = (l_1' p + l_2' p^2)\frac{1}{e^{-T_2 p} - e^{-T_1 p}},$$

i.e. contain two parallel-connected differentiators with the coefficients

$$l_1' = \frac{\beta_0(T_1 - T_2)}{\alpha_1\alpha_2} \text{ и } l_2' = \frac{\beta_0\tau_1(T_1 - T_2)}{\alpha_1\alpha_2}$$

and a series-connected anticipation section with the transfer function

$$\frac{1}{e^{-T_2 p} - e^{-T_1 p}}, \text{ где } T_1 > T_2.$$

This anticipation is easy to obtain in program control systems, where the future change in the perturbation is known and where it is possible to feed a signal to the input of the open probabilistic link with a definite anticipation.

It is just as easy to obtain any required anticipation in probabilistic links with cyclical repetition of the processes. For example, it is possible comparatively easily to predict the mean temperature for any future month of the year. The introduction of anticipation sharply improves the dynamics of transfer processes in systems with probabilistic links.

In a stabilization system, where the future value of the load is unknown, it is impossible to achieve such an anticipation section in practice. Hence we must accept approximate satisfaction of the invariance conditions. The problem reduces to the maximum possible approximation of the operator of the actually achievable quadrupole

$$l(p) = \frac{l_0 + l_1 p + \ldots}{l_0' + l_1' p + \ldots}.$$

to the ideal operator, which ensures absolute invariance in the presence of a probabilistic link,

$$l(p) = (l_1' p + l_2' p^2) \frac{1}{e^{-T_2 p} - e^{-T_1 p}},$$

It is necessary to select the coefficients of the actually achievable differentiator so that both functions differ as little as possible from one another. The problem can be solved by many methods (Chebyshev et al). Let us use one of the simplest methods: we use the expansion of the exponential functions in a series. We will limit ourselves to three terms of the series

$$\frac{1}{e^{-T_2 p} - e^{-T_1 p}} = \frac{1}{\left(1 - T_2 p + \frac{1}{2} T_2^2 p^2 - \ldots\right) - \left(1 - T_1 p + \frac{1}{2} T_1^2 p^2 - \ldots\right.} =$$

$$= \frac{1}{(T_1 - T_2) p + \frac{1}{2} (T_2^2 + T_1^2) p^2 + \ldots} =$$

$$= \frac{1}{p (T_1 - T_2) \left[1 - \frac{1}{2} (T_1 + T_2) p\right]}.$$

Setting both terms equal to one another, we find

$$l(p)=\frac{l_0+l_1p}{l_0'+l_1'p}=p(T_1-T_2)\left[\frac{\beta_0+\beta_0\tau_1p}{\alpha_1\alpha_2}\right]\frac{1}{e^{-T_2p}-e^{-T_1p}}=$$

$$=\frac{\beta_0+\beta_0\tau_1p}{\alpha_1\alpha_2-\frac{1}{2}\alpha_1\alpha_2(T_1-T_2)p},$$

whence we obtain

$$l_0=\beta_0,\quad l_1=\beta_0\tau_1,\quad l_0'=\alpha_1\alpha_2,\quad l_1'=-\frac{1}{2}\alpha_1\alpha_2(T_1-T_2).$$

The synthesis of the system is complete: the operators m(p) and l(p) which ensure optimal stability and invariance with respect to the load L(t) have been found.

Thus we have considered the invariance and stability conditions of systems with probabilistic links. The application of the general theory of combined systems to probabilistic learning systems is completely obvious.

Experimental method of finding the most effective prediction formula. If a process subject to prediction is so little studied that it is not certain that Kolmogorov's formula is the most general and best one, then it is possible almost mechanically to try various prediction formulas at random. Having obtained estimates of the usefulness of various terms of the formula, we can discard terms of little use and by this method can gradually work out the most suitable formula, i.e. one giving the highest percentage of correct predictions for a given volume of computational work.

L.I.Voronova, in particular, has shown that to predict the amplitude of waves we can use Taylor series. The prediction accuracy drops a little in this case, but on the other hand there is a great decrease in the volume of computations.

At the beginning of this section, we said that systems which achieve "pure" randomness do not exist. Even in tossing a coin, some kind of constantly acting factors (bend in the coin, manner of throwing, etc) make the coin fall more often on one side than the other; and hence, besides the "pure" randomness, a probabilistic law is active. However, pure randomness is completely real. However much we make the regression formulas more accurate, however much we raise the number of their terms and select the most useful of them, the success of prediction of a probababilistic process cannot be 100%. Such a result can only be obtained for processes which are completely determinate, hence subject to calculation. As the methods of predicting random processes are improved, the accuracy increases, but there always remains an unpredictable part which expresses the element of pure randomness. If we increased the number of terms in Kolmogorov's formula or passed the continuous quantities, it would be entirely possible that in the example with ocean waves we would obtain a prediction accuracy of more than 80%. But some some percentage of possible error would remain, for in this process, there is an element of pure randomness.

* * *

In conclusion let us note that Kolmogorov's formula (in its application to the Alpha discrete filter) can explain, and hence also direct, the success of many experiments on prediction.

In their experiments on prediction of the treatment burns, A.L.Lunts and V.L.Brailovskiy used 12 input attributes (area of wound, burn localization, age of patient, accompanying diseases, complications, data of blood analysis, etc), which were each used individually, and also in combinations of two, three, etc. It was established that the most useful information was contained in the logical products of several of the input attributes.

Kolmogorov's formula and the above method of determining the usofulness of its individual terms is a mathematical algorithm which explains the success of the indicated experiments.

In future the methodology of the organization of prediction experimonts based both on continuous and on binary input attributes must take into account the mathematical expectation and the structure of the extended prediction operator.

Bibliography

1. Yu.M.Alekhin, Statisticheskiye prognozy v geofizike (Statistical forecasts in geophysics). Leningrad Univ.Publ.House, 1963.

2. U.F.Barret, Zagadochnyye yavleniya chelovecheskoy psikhiki (Puzzling human mental phenomena). Moscow, 1914.

3. S.Bir, Kibernetika i upravleniye proizvodstvom (Cybernetics and production control). Fizmatgiz, Moscow, 1963.

4. K.But, Chislennyye metody (Numerical methods). Fizmatgiz, Moscow, 1960.

5. A.Val'd, Posledovatel'nyy analiz (Sequential analysis). Fizmatgiz, Moscow, 1960.

6. V.I.Vasil'yev and B.K.Svetal'skiy, "Accuracy of predicting devices", Avtomatika (Automation), 1965, 4.

7. V.I.Vasil'yev, Differentsial'nyye sistemy regulirovaniya (Differential regulating systems). Acad.Sci.UkSSR Publ.House, 1963.

8. L.L.Vasil'yev, Tainstvennyye yavleniya chelovecheskoy psikhiki (Secret human mental phenomena). Gospolitizdat, Moscow, 1964.

9. idem, Vnusheniye na rastoyanii (Suggestion at a distance). Gospolitizdat, Moscow, 1962.

10. Ye.G.Gladyshev, "Periodically correlated random sequences", DAN SSSR (Reports of Acad.Sci.USSR), 1961, 137, 1026.

11. idem, "Theory of periodically correlated random sequences and processes", Avtoreferat dis. (Author's synopsis of thesis), Moscow, 1963.

12. B.V.Gnedenko, Kurs teorii veroyatnostey (Course in probability theory). Gostekhizdat, Moscow, 1954.

13. U.Grenander, Sluchaynyye protsessy i statisticheskiye vyvody (Random processes and statistical deductions). For.Langs.Publ.House, Moscow, 1961.

14. I.V.Dunin-Barkovskiy, N.V.Smirnov, Teoriya veroyatnostey i matematicheskaya statistika (Probability theory and mathematical statistics). Gostekhizdat, Moscow, 1955.

15. Ye.B.Dynkin, Markovskiye protsessy (Markov processes). Fizmatgiz, Moscow, 1963.

16. D.K.Zhuk, "Applying the method of inverse operators to the synthesis of multicircuit systems", Avtomatika, 1963, 4; 1964, 6.

17. A.G.Zaytsev, "Analytical design of systems which reproduce the useful signal in the presence of noise", A i T (Automation and Telemechanics), 1963, 2.

18. E.G.Zel'kin, "Construction of extrapolators", A i T 1962, 23, 9.

18a. idem, "Construction of extrapolators", Nauch. dokl.vyssh.shkoly (Higher School Scientific Reports). Moscow Power Engineering Institute, 1958, 2.

19. A.G.Ivakhnenko, L.I.Voronova, "Alpha recognition system as a learning filter and extremal regulator without search oscillations", Avtomatika, 1964, 3.

20. A.G.Ivakhnenko, Elektroavtomatika (Electroautomation). Gostekhizdat UkSSR, Kiev, 1957.

21. idem, Samoobuchayushchiyesya sistemy s polozhitel'nymi obratnymi svyazyami (Self-learning systems with positive feedback). Acad.Sci.UkSSR Publ.House, 1963.

21a. idem, "Comparison of the properties of the basic circuits for combined extremal control", Avtomatika, 1964, 4 and 5.

22. Issledovaniya kharakteristik rezhima vozobnovlyayushchikhsya istochnikov energii vody, vetra i solntsa

(Investigations of the characteristics of the conditions of regenerative sources of hydro, wind and solar energy). UzSSR Publ.House, 1963.

23. B.B.Kazhinskiy, *Biologicheskaya radiosvyaz'* (Biological radio communications). Acad.Sci.UkSSR Publ. House, 1962.

24. V.V.Karibskiy, A.V.Chernyshev, *Tsifrovyye interpolyatory dlya sistem programmnogo upravleniya* (Digital interpolators for program control systems). Central Institute for Scientific and Technical Information, Moscow, 1962.

25. I.P.Kerekesner and Yu.N.Chekhovoy, "Learning algorithms of an open extremal control system", *Avtomatika*, 1965, 2.

26. A.N.Kolmogorov, "Stationary sequences in Hilbert space", *Byull.MGU* (Bulletin of Moscow State Univ.) , 1941, 2, 6.

26a. idem, "Interpolation and extrapolation of stationary random sequences", *Izv.ANSSSR* (News of Acad. Sci.USSR), Mathematics and Natural Science Series, 1941, 5.

27. Yu.V.Krementulo, "On the condition of absolute invariance for open pulse systems", *Avtomatika*, 1960, 2.

28. idem, "Synthesis of interpolators from the invariance conditions", *Avtomatika*, 1961, 5.

29. D.K.Labbok, Optimization of a class of nonlinear filters", *Tr. I kongr. IFAK* (Transactions of the First Congress of IFAC), Moscow, 1961, 3.

30. V.G.Lapa, "Combined method of predicting nonstationary processes", *Avtomatika*, 1965, 3.

31. idem,"Prediction of biological nonstationary processes by the method of characteristic components", *Avtomatika*, 1964, 4.

32. Dzh.Lening, R.G.Bettin, Sluchaynyye protsessy v zadachakh avtomaticheskogo upravleniya (Random processes in automatic control problems), For.Langs.Publ.House, Moscow, 1958.

33. A.A.Lyapunov, "On some general problems in cybernetics", in the book Problemy kibernetiki (Problems in cybernetics), Fizmatgiz, Moscow, 1958, 1.

34. B.Yu.Mandrovs'kiy-Sokolov, "Realization of extrapolating filters with exponential smoothing" (Ukr.), Avtomatika, 1964, 3.

35. P.V.Melent'yev, Priblizhennyye vychisleniya (Approximate computations). Fizmatgiz, Moscow, 1962.

36. G.L.Otkhmezuri, "On the properties of attributes and sixth feedback", Avtomatika, 1963, 2.

37. M.Pomortsev, Ocherk ucheniya o predskazanii pogody (sinopticheskaya meteorologiya) (Essay on the theory of weather forecasting(synoptic meteorology)). Saint Petersburg, 1889.

38. Prognoz v zashchite rasteniy ot vrediteley i bolezney (Prediction for protection of plants from pests and diseases). Acad.Sci.Latv.SSR Publ.House, Riga, 1964.

39. V.S.Pugachev, Teoriya sluchaynykh funktsiy i yeye primeneniye k zadacham avtomaticheskogo upravleniya (Random function theory and its applications to automatic control theory). Fizmatgiz, Moscow, 1960.

40. Ye.V.Saparina, Kibernetika vnutri nas (Cybernetics within us). "Molodaya gvardiya" Publs., 1962.

41. T.M.Sergiyenko, M.Ya.Voloshin, V.G.Lapa, "Use of methods of mathematical prediction in neurosurgical practice", Dokl. na Vsesoyuz. konfer. neyrokhirurgov (Reports at the All-Union Conference of Neurosurgeons), Leningrad, 1964.

42. B.A.Sigov, "Raising the accuracy of operation

of a digital integrator constructed on the basis of a frequency divider", Avtomatika, 1962, 1.

43. O.Dzh.M.Smit, Avtomaticheskoye regulirovaniye (Automatic regulation), Moscow, 1962.

44. V.V.Solodovnikov, Vvedeniye v statisticheskuyu dinamiku lineynykh sistem regulirovaniya (Introduction to the statistical dynamics of linear control systems), Moscow, 1960.

45. E.D.Farmer, "Method of predicting nonstationary processes and its application to the problem of load evaluation", Dokl. na II kongr. IFAK (Reports at the Second Congress of IFAC), Moscow, 1963.

46. Ya.Z.Tsypkin, Teoriya lineynykh impul'snykh sistem (Theory of linear pulsed systems). Fizmatgiz, Moscow, 1963.

47. B.S.Yastremskiy, Nekotoryye voprosy matematicheskoy statistiki (Some problems in mathematical statistics). Gosstatizdat, Moscow, 1961.

48. Bernard B. Measuring the worls population explosion. New Slentist, 1962, 313, Nov.

49. Bode H. W., Shennon C. E. A Simplifide Derivation of Linear Least — Square Smoothing and Prediction Theory. Proc. IRE, 1959, 38, 4,

50. Brown R. G. Statistical Forcasting for Inventory Control. McGrow Hill, N. Y., 1959.

50a. Brown R. G., Meyer R. F. The fundamental Theory of Exponential Smoothing. Opus. Res., 1961, 9.

51. Bryan J. S. Experiments in Adaptive Pattern Recognition, IEEE Trans. on military electronics, 1963, Apr.— July.

52. Duda R. O., Machanik J. M. An Adaptive Prediction Technique and its Application to Weather Forecasting. Wescon Techn. Papers, 1963, 7.

53. Electronics, 1960, 6.

54. Gabor D., Wilby W., Woodcock R. A Universal Nonlinear Filter, Predictor and Simulator wich Optimizes Itself by a Learning Process. Proc. Inst. Electr. Eng., 1961, 40.

55. Gabor D. Predicting Machines. Scientia, Rev. Int. Syntese Sci., Milano, 1962, 5, Mai.

56. Karhunen K. Über ein Extrapolationproblem in dem Hilbertshem Raum. 11. Skand. mat. Kongress, 1952.

57. Peshel M. Über die Anwendigkeit von Korrelations Metoden in der Regelungstechnick. Messen stevern regeln, 1965, 1.

58. Labbock G. K. The Optimisation of a Class of Non Linear Filters. Proc. Inst. Electrical Engineers, Pt C, Monogr. 1959, N 344E.

59. Rosenblatt F. Perceptual Generalization over Transformation Groups. In: Self — Organizing Systems, Pergamon Press, 1960.

59a. Rosenblatt F. Perceptron Simulation Experiments. Proc. IRE, 1960, March.
59б. Rosenblatt F. Principles of Neurodynamics. Spartans Books, Woshington, 1962.
60. Streets R. B. Arbitrary Non-Mean-Square Error Criteria. IEEE Trans. Autom. Control, 1963, 8, 4.
61. Zadeh L. A., Ragazzini J. R. An Extention of Wiener's Theory of Prediction J. Appl. Phys., 1950, 21, 7.
62. Wiener N. The Exstrapolation, Interpolation and Smoothing of Stationary Time-Series. J. Willey, N. Y., 1949.

E N D

2393
CSO: 13231-N