IFAC

INTERNATIONAL FEDERATION
OF AUTOMATIC CONTROL

WARSZAWA 1969

Translations of papers from Russian into English

Fourth Congress of the International Federation of Automatic Control
Warszawa 16–21 June 1969

Vol. II

TECHNICAL SESSIONS 28 – 46

NOT

Organized by
Naczelna Organizacja Techniczna w Polsce

INTERNATIONAL FEDERATION OF AUTOMATIC CONTROL

Translations of papers from Russian into English

Vol. II

TECHNICAL SESSIONS 28 – 46

FOURTH CONGRESS OF THE INTERNATIONAL FEDERATION OF AUTOMATIC CONTROL
WARSZAWA 16 – 21 JUNE 1969

Organized by
Naczelna Organizacja Techniczna w Polsce

Contents

28.3
DECENTRALIZATION PRINCIPLE IN OPTIMIZATION OF COMPLEX SYSTEMS

PERVOZVANSKY A.A.
Polytechnical Institute, Leningrad, USSR.

A paper consists of two parts. In the first part there is considered the possibility of reduction of optimal planning problem for systems of reasonably general structure to the series of extremum problems for separate elements. An influence of reciprocal supplies on the efficiency of the work of every element is studied and the simple local approximation for these functions is proposed. On this basis in the second part an influence of supplies irregularity during planning period on the average efficiency of the work of an element is analyzed with provision for the possibility of optimization of inventory policy. It leads to certain conclusions on the necessity of correction of the statement of a planning problem for a system as a whole.

I. A problem of a planning of a work of a system combined of interactive elements is considered. A structure of interaction may be defined by the incidence matrix of connections graph, i.e. for every i it is possible to define a set of indices j of the inputs $J_1(i)$ or a set of output indices $J_2(i)$. Define also an element $n+1$ (graph node), which has no arcs coming out of it. A perfomance of every element i characterized by its output P_i distributed to connected elements

$$P_i = \sum_{j \in J_2(i)} P''_{ij} \tag{I.I}$$

such that

$$\left\{ u_i ; P_i ; P'_{ji} \right\} \in R_i \tag{I.2}$$

where u_i - a vector characterizing an intensity of the processes in an element itself, P'_{ji}, $j \in J_1(i)$, - vectors characterizing a part of other elements output consumed by the element i, R_i - some closed sets.

Note that in some cases it is convenient to separate the constraints into two groups: system constraints characterized by the output consumption levels of system elements and own (local) constraints characterizing the possibilities of a given element only.

It is assumed that a perfomance of a system as a whole is governed by an extremum principle expressed in additive form

$$F = \sum_{i \in J_1(n+1)} F_i \left(P''_{i,n+1}\right) \tag{I.3}$$

and by the conditions of balance

$$P'_{ij} = P''_{ij}, \tag{I.4}$$

Extremum problem just formulated will be called optimal deterministic planning for a system. It is a mathematical programming problem and it represents a structural generalization of classical schemes given by L.V.Kantorovich and D.Gale which are used in econometrics. It may be interpretedd as a problem of a choice of output and distribution levels in the system combined of series of partially connected productive units (shops, enterprises, firms and so on). Although a time factor does not occur explicitly in a model, it is possible to suggest that given constraints determine a behaviour of each element of a system at a certain finite interval of time (planning subperiod). This interval may not be identical for different elements. It makes possible to consider so called determined perspective planning models although in this case one productive unit must be represented by several elements in a scheme corresponding to its state at successive planning subperiods.

After construction of a planning problem it may be solved at least in principle, by any known method of mathematical programming. However, special features of a model as a complex system impose essential limitations on the choice of a method. These limitations are of two types:

a) informational, connected with difficulties of collection and keeping of the data in one place (center).

b) computing, caused by practical inability of modern computers to operate with problems of very big volume.

For this reason, in this paper an attention is paid exclusively to the methods which use a decentralization principle. It means that the main problem must be separated into n extremal subproblems, independent in a sense that direct information about constraints on every element of a system is not required for their coordination.

In another treatment of the principle an additional requirement is put forward, namely, that a subproblem coordination may be done not in a single center but in several ones, each center possessing partial information only.

Mathematical treatment of an idea of decentralization has arised (although in a limited form) on a basis of works by Arrow, Hurwicz, Uzava on gradient schemes of search of extremum [I,2], then got a new impulse when the decomposition algorithm was created by Dantzig and Wolfe [3], and now is widely discussed, mainly in connection with different schemes of so called block programming ([4,8] and others).

It is possible to mark out (omitting some essential mathematical details) two different approaches on subproblem forming.

The first one has a general character and is based on utilization of a main theorem by Kuhn and Tucker on equivalence (with certain limitations) of a given problem and a problem of a search of a saddle point of Lagrange function

$$\varphi = \sum_{i\in J_1(n+1)} F_i\left(P''_{i,n+1}\right) + \sum_i \sum_j \lambda_{ij}\left(P''_{ij} - P'_{ij}\right). \qquad (I.5)$$

at the conditions (I.I-I.2). Then the problem is separated into n subproblems for the elements

$$f_i = \max\left[F_i\left(P''_{i,n+1}\right) + \sum_{j\in J_2(i)} \lambda_{ij} P''_{ij} - \sum_{j\in J_1(i)} \lambda_{ji} P'_{ji} \Big/ P_i = \sum P''_{ij};\ \{u_i; P'_{ji}; P_i\}\in R_i\right] \qquad (I.6)$$

where "prices" λ_{ij} are considered as parameters, and one problem for a "center", which is a problem of optimal coordination of "prices"

$$\varphi_0 = \min\left\{\varphi = \sum_{i=1}^{n} f_i\left[\lambda_{ij}, j\in J_2(i); \lambda_{ij}, j\in J_1(i)\right]\right\} \qquad (I.7)$$

This scheme may be called a scheme of "buy-sell" of intermediate products, the subproblems are treated as those of a profit maximization for every element, the prices on intermediate products are coordinated by the "center".

In the second approach it is assumed that every element of a system participates in forming of final products (or, final expenses), i.e. a set $J_1(n+1)$ includes all $i=1,2,\dots,n$. Then, considering the levels of distributed products as parameters, one comes to n subproblems of the type

$$F_i = max\left[F_i(P''_{i,n+1}) / \sum P''_{ij} = P_i ; \{u_i, P_i, P'_{ji}\} \in R_i\right] \qquad (I.8)$$

and one problem for a "center" which is a problem of optimal balance of intermediate products:

$$F = max\left\{F = \sum_i F_i\left[P'_{ji}, j\in J_1(i); P''_{ij}, j\in J_2(i)\right]\right\} \qquad (I.9)$$

This scheme may be called a scheme of resources distribution, or a scheme of "optimal balance". In this case an optimization of final products value is made inside of every element, the levels of reciprocal supplies are given by the "center".

The partitions described are purely formal. In fact, there is no way to specify the "prices" (objectively conditioned estimates in L.V.Kantorovich's terminology) or the levels of intermediate production, which give an optimal coordination of subproblems, exept the construction of a solution of a problem as a whole. However, a possibility of utilization of iterative procedures, i.e. the procedures of gradual coordination, makes this approach efficient enough.

Almost all of the known procedures are the versions of generalized gradient descent method or the method of feasible directions applied to the problem of a "center". In application to (I.7) the main idea is reduced to the following: an arbitrary set of "prices" is chosen by a "center" (in practical problems, the choice, naturally, is dictated by some practical considerations), and each element is informed about the choice. Then a problem of a type (I.6) is solved for every element, and this solution provides at given $\lambda_{ij} = \lambda^0_{ij}$ optimal values of activities u_i, supplies P''_{ij}, $j\in J_2(i)$, and external (for the element in question) consumption. In addition to it, a tendency of change of purpose function f_i due to small deviations of the "prices" from the level λ^0_{ij} is revealed, i.e. a local approximation $f_i(\lambda_{ij})$ in some small vicinity λ^0_{ij} is constructed.

Just the local approximation is sent to a "center", where on the basis of this information a tendency of a change of function φ as a whole is revealed, that makes it possible to find necessary direction of a change of "price" system. So far as the step in the chosen direction is accomplished, and a set is got of new values of $\lambda_{ij} = \lambda^1_{ij}$, a procedure, naturally, can be repeated. A choice of the way of local approximation and size of a step in the direction of decrease is a speciality of a particular method.

Note only some general features arising in a case when original problem, and hence - subproblems, are formalized as linear programming problems. Here a local approximation of the behaviour of functions f_i, which coincides exactly with a real behaviour in certain finite vicinity, is as follow

$$\hat{f}_i = f_i(\lambda^0_{ij}) + \min_{s\in S^0_i} \left\{ \sum_{j\in J_2(i)} \Delta\lambda_{ij} P_{ij,s} - \sum_{j\in J_1(i)} \Delta\lambda_{ji} P'_{ji,s} \right\} \quad \text{(I.10)}$$

where S^0_i - a set of optimal basises of a problem (I.7) at $\lambda_{ij} = \lambda^0_{ij}$, and $P''_{ij,s}$, $P'_{ji,s}$ -optimal values of the variables P''_{ij}, P'_{ji} corresponding to these basises. Hence, in particular, it is clear that the functions $f_i(\lambda_{ij})$, generally speaking, are not even diffetiable. If the solution (I.7) is unique at $\lambda_{ij} = \lambda^0_{ij}$, then a gradient $f_i(\lambda_{ij})$ exists at the point and may be readily found as soon as an optimal solution is known. If all the problems of a type (I.7) have a unique solution then a calculation of a gradient of a function φ as a whole is reduced to calculation of debalances on connections:

$$\left.\frac{\partial\varphi}{\partial\lambda_{ij}}\right|_{\lambda_{ij}=\lambda^0_{ij}} = P''_{ij,opt} - P'_{ij,opt} \quad \text{(I.11)}$$

Noted feature essentially facilitates an organization of a gradient descent. Moreover, it then follows a principally important result: a determination of the direction of a change of "prices" may be done without participation of a unitad center, but by coordination of the results of planning of directly connected elements. At the same time, it is obvious that a condition of uniqueness is not always fulfilled, it is certainly violated at the extremum point of a function φ, a consequence of this, being an absence of strict convergence of gradient descent at constant coefficients of proportionality [1]. However, there are some ways, a decrease of coefficients with a grow of its iteration number, for example, that permit to avoid the troubles noted.Moreover,when solving practical problems a requirement of strict convergence is not very essential,it would be more important to increase a spead of motion,that can be done,for example by utilization of finite-step schemes with a use of simplified ways of local approximation [IO] .Note also,that all that was said before may be completely applied to the second version of decentralization principle.Here the schemes of generalized gradient descent may also be used for the solution of the problem of a "center",but in a space of variables P'_{ij} , P''_{ji} , .The functions F_i are also piecewise linear if original problem is a problem of linear programming.Their local approximation may be constructed through the use of optimal basis solutions of the problems dual to (I.8).If these solutions are unique,i.e. the problems (I.8) at given values of parameters are not degenerated,then it is possible to find the gradients of functions F_i with the help of optimal dual variables (objectively conditioned estimates).In the vicinity of an optimum only piecewise linear locally-exact approximation of the following type is permitted

$$\hat{F}_i = F_i\left[P''_{ij,opt};P'_{ji,opt}\right]+\min_{\ell\in\ell_i^0}\left\{\sum_j \lambda'_{ji,\ell}\,\Delta P'_{ji}-\sum_j \lambda''_{ij,\ell}\,\Delta P''_{ij}\right\} \quad (I.I2)$$

whwre ℓ_i^0 , denote a set of optimal basises of a dual problem and $\lambda'_{ji,\ell}$, $\lambda''_{ij,\ell}$ - corresponding values of dual variables.Iterative procedures can be done in a manner described earlier for the first version of separation of a problem, this time,however,the conditions of a balance on connections ar are fulfilled exactly in every iteration,that makes its realization more complicated but increases a validity of constructed approximation,every of which represents a suboptimal plan of a problem as whole.

Note that a plan (program) is usually treated as a specification of final production and reciprocal supplies levels for every element of a system,these levels specified integrally at the planning period.However,due to inevitable fluctuations of conditions an exact realization of the plan turns out to be impossible,therefore in carrying out a control,i.e. in realization of a plan,a question of compensation of deviations from the program naturally arises.Since there are many plan indices, it is important to estimate the influence of their individual deviations on the efficiency of a system.In practice,this estimates is done subjectively.At the same time,decentralization principle permits not only to find optimal levels during construction of a plan,but also to esteem an efficiency of small deviations from these levels,because in the course of construction of a plan a local approximation of dependency of local purpose functions from external parameters is also constructed.It is essential that even in a small vicinity of optimal plan this approximation is nonlinear.This point predetermines impossibility of use of such "linear" criteria as summed value of deviations calculated in prices that do not depend on the level or direction of the deviations.

2. Let us now proceed to the analysis of influence of some fluctuative factors on the efficiency of realization of the program in the course of its accomplishment and of a back influence of these factors on a scheme of construction of a plan itself.

It is obvious that designation of a reciprocal supplies program integrally at the planning period does not specify a distribution of these supplies during this period.This uncertainty,peculiar to the method of planning,may lead (and practically leads) to a rise of irregularity in the level of supplies.

Consider certain element from a system and study a change of its final productions due to deviations in the level of supplies to this element from other elements. Divide a planning period in subperiods and enumerate them with the indices k ($k = 1, 2, \dots, K$).Assume that the resources of this element are not changed during a whole period and that its perfomance is governed by local extremal principle,formulated earlier,coordinated integrally with extremal principle for a system as a whole.Then at every subperiod a change of efficiency with respect to final production output is given by formula analogous to (I.I2),assuming that the requirements on supplies from a given element are fulfilled and the supplies to the element have small deviations from prescribed level.This peacewise linear dependence may be efficiently approximated by means of a system of coefficients (marginal values of a problem,following a terminology of [II]) characterizing a change of function at increase or decrease of every of the components of supplied production from optimal program.

Omitting indices, characterizing a number of an element and its suppliers, a local approximation may be given as follows [IO]:

$$\overline{F}_k = F_k(p_{k,opt}) + \sum_{\tau} \min\left(\lambda^+_\tau \Delta p_{\tau k}; \lambda^-_\tau \Delta p_{\tau k}\right) - F_k(p_{k,opt}) + \sum_{\tau} \Delta F_{\tau,k} \quad (2.I)$$

where $\lambda^+_\tau, \lambda^-_\tau$ - accordingly right and left partial derivatives of purpose function with respect to τ-th resource. It is possible to show that $\overline{F} \leq \widehat{F}$. An advantage of an approximation (2.I) consists in a possibility to consider an influence of changes of every component separately. A difference among the components demonstrates itself in a difference of marginal values only. In what follows, the indices of resource components $P_{\tau,k}$ and these of corresponding changes $\Delta F_{\tau,k}$ of purpose function will be omitted and following notations introduced:

$$\Delta p_{\tau,k} \equiv v_k \,;\ \Delta F_{\tau,k} \equiv \psi(v_k) = \begin{cases} \lambda^+ v_k, & v_k \geq 0, \\ \lambda^- v_k, & v_k \leq 0. \end{cases} \quad (2.2)$$

Note also that always $\lambda^- \geq \lambda^+ \geq 0$. The values $v_k, k=1,2,\dots,\mathcal{K}$, characterize a deviation of a quantity of a resource component by given element of a system from planned level, which is the $1/\mathcal{K}$-th part of planned supplies for a period as a whole. We may also introduce the values η_k, characterizing the deviations of real supplies during subperiod k from a given planned level. If an accumulation of resource surpluses is possible, then a choice of the value v_k is restricted by the quantity stored at a warehouse at the beginning of a subperiod and

$$v_k \leq y_k, \quad (2.3)$$

$$y_{k+1} = y_k + \eta_k - v_k, \quad k = 1,2,\dots,\mathcal{K}. \quad (2.4)$$

In this way a general problem of supplies irregularity is reduced to a classical onedimensional inventory problem (a particular case - water resource control was considered by Karlin and Gessford in [I2]). Both an analysis of a system as a whole and utilization of simplified local approximation (2.I) give a general character to this problem.

We shall now describe the concrete results of analysis with a use of optimal many-stage policy at the whole period and the simplest policy $v_{k+1} = \eta_k$ .Under these conditions, assume that η_k deviates from zero (planned level) with equal probability on the value $\pm a$,an average being equal to zero.

Then if $S_t(y)$ is a value of average loss due to irregularity at t subperiods to go to the end of a period, at the initial stock level y and optimal policy, then

$$S_t(y) = \min_{v \le y} [-\psi(v) + \int S_{t-1}(y+\eta-v)\, w(\eta) d\eta], \qquad S_1(y) = \min_{v \le y} [-\psi(v)]. \tag{2.5}$$

It is possible to show that an optimal policy is as follows:

$$v_t = \begin{cases} 0 \; ; & y \ge 0; \\ y \; ; & y \le 0, \end{cases} \quad t = 2,3,\dots ; \; v_t = y. \tag{2.6}$$

Then

$$S_t(y) = \begin{cases} \frac{1}{2} S_{t-1}(y+a) + \frac{1}{2} S_{t-1}(y-a); & y \ge 0 \\ -\lambda^- y + \frac{1}{2} S_{t-1}(a) + S_{t-1}(-a); & y \le 0 \end{cases}$$
$$S_1(y) = \begin{cases} -\lambda^+ y \; ; & y \ge 0 \\ -\lambda^- y \; ; & y \le 0 \end{cases} \tag{2.7}$$

A use of the simplest policy leads to

$$\bar{S}_t(y) = -\psi(y) + (t-1) \frac{\lambda^- - \lambda^+}{2} a . \tag{2.8}$$

Diagrams of the function $S_t(y)$ at $t = 1,2,\dots,6$ are given at fig.I.

It is now possible to make the following conclusions:

I. Supplies fluctuations with an average equal to zero lead to an increase of an average loss which generally speaking is not equal to zero.

2. In absence of initial stock the losses at the simplest policy are equal to

$$(n-1) \frac{\lambda^- - \lambda^+}{2} a$$

i.e. they grow proportionally to the number of subperiods.

3. At the presence of initial stock and under optimal use of accumulations it is possible to reduce average losses. Even a at the absence of initial stock, average losses at $n=6$ are reduced approximately twice compared with the results of the simplest policy.

It is possible to show also that the probability of an ab sence of losses essentially depends on the quantity of initial stock level y_0 and is equal to $(1/2)^{\mathcal{K}-N}$, if $N=\frac{y_0}{a}\leq\mathcal{K}$.

The described effect of an influence of supplies irregularity on the efficiency of a system is important itself, but it reflects one side of a problem only. In fact, in the course of analysis it was implicitly assumed that optimization problem for given element of a system may be solved not only at the planned level of supplies but also at the fluctuations near the level. In other words, it was assumed, that at the presence of fluctuations it is possible to satisfy the restriction on planned level of supplies from given element. Generally speaking, it is n not always true: at the deviations of supplies it is often impossible to fulfil the requirements on orders and hence in the case of decentralized current planning a necessity arises to commensurate optimally the deviations from the levels required.

A use of the local functionals in a form (I.I2) does not give a solution of this problem, because the functions $F_i(p)$ are defined only on the region of existence of the solutions. This problem becomes especially critical for a structure of a technological chain type [7,I3] where the elements are sequentally connected and the last element alone gives a final production with definite estimates of its components. At the same time a scheme of decentralization constructed on the basis of dynamical programming may be used for the analysis of such a structure. Here, generally speaking, it is necessary only to construct the dependence of the efficiencies of each element on given resources at small deviations of those from the optimally planned level. In a paper [I3] there was shown that locally-exact approximation of a purpose function for the p-th element of technological chain may be represented as a piecewise linear function

$$\hat{f}_p(v) = f_{p,opt} + \min_{l\in l_p^0} \lambda_p v \qquad (2.9)$$

where v - a column-matrix of the deviations from the leve optimal level of resources supplied to the p-th element from its predecessor in technological chain. The dependence of each type of products on v has a similar character. With the aid of simplified approximation of a type (2.I) it may be shown that the dependence of deviations from optimal plan for any product of the p-th element on small deviations of any component of a resource consumed from the $(p-1)$-th element may be represented by the function of a type $\psi(v)$ (look at 2.2). It leads to the effect of accumulation of an influence of fluctuations caused by the irregularity of reciprocal supplies along technological chain if a necessary level of initial reserve for each element is not provided in a chain. In fact, the fluctuations of supplies to the first element of a chain, lead to impossibility to maintain an average level of supplies from the first element to the second one, and so on. Due to this factor a real average levels of supplies to the following elements may considerably differ from the planned ones, and the estimates of the efficiency of deviations justified only in the vicinity of a plan will become incorrect. Therefore an optimal planning itself may be effective only if it takes into account necessity of expenditures on forming of the reserves that give on opportunity to localize an influence of irregularities.

It is natural that creation and maintenance of reserves seems to be not economical from the point of view of a purpose of a system. However, it is necessary and it has the same meaning as an introduction of redundancy into technical systems composed from elements that are not completely reliable. It is worthwhile to stress that it is not an effect of fluctuations of external conditions which is discussed (a necessity of reservation with respect to external conditions is generally accepted) but it is an effect of internal fluctuations stimulated by the fact that the system's planning is incompletely deterministic

I. Arrow K.J., Hurwicz L., Uzawa H., Studies in linear and non-linear programming, Stanf., Calif., I958.
2. Marschak T., Centralization and decentralization in economie organizations, Econometrica, v.27, 3 /july I959/.
3. Dantzig G.B., Linear programming and extensions, Princet., N.-j., I963.
4. Гольштейн Е.Г., Юдин Д.Б., Новые направления в линейном программировании, "Сов.радио", М., I966.
5. Волконский В.А., Оптимальное планирование в условиях большой размерности, Экономика и математические методы, I965, т.I, № 2.
6. Корнаи Н., Липтак Т., Планирование на двух уровнях, в сб. "Применение математики в экономических исследованиях", т.3, "Мысль", I965.

7. Первозванская Т.Н., Первозванский А.А., Распределение ресурсов между многими предприятиями, Экономика и математические методы, 1966, т.2, № 5.
8. Каценеленбойген А.И., Овсиенко Ю.В., Фаерман Е.Ю., Методические вопросы оптимального планирования в социалистической экономике, ЦЭМИ АН СССР, М., 1966.
9. Шор Н.З., Принцип обобщённого градиентного спуска в блочном программировании, Кибернетика, № 3, 1967.
10. Первозванская Т.Н., Первозванский А.А., Алгоритм поиска оптимального распределения централизованных ресурсов. Изв.АН СССР, Техническая кибернетика, 1966, № 3.
11. Mills H.D., Marginal values of matrix games and Linear programs, "Linear inequalities and related systems", Princet., N.-J., 1956.
12. Arrow K.J., Karlin S., Scarf H., Studies in the Theory of Inventory and Production, Stanf., Calif., 1958.
13. Первозванская Т.Н., Метод приближения в пространстве целевых функций при решении задач на "узкие места", ЖВМ и МФ, т.7, № 3, 1967.

17

28.5

ON A COMPLEX SYSTEMS CONTROL THEORY

A.I. Kukhtenko

Institute of Cybernetics of the Academy of Sciences
of the Ukrainian Soviet Socialist Republic
Kiev
USSR

It was already indicated in a number of publications [1-4] that the study of complex systems must be made, as to necessity by using different levels of an abstract description. Depending on a system designation, either a theoretical informational, or logical mathematical, or dynamical, or at last, a heuristical treatment of the problem may be used. But in reality one must use in most cases a few different levels of abstact description simultaneously.

Without dwelling here repeatedly upon a characteristic of the term "a complex control system", given in the paper [3], we shall mark the fact that the necessity of a description of a complex systems control at a few abstract levels simultaneously compels us to search for those mathematical means that enabling us to make it. However, attempts of applying for this purpose the well-known methods of the automatic control theory or, in general, the dynamic systems theory, the finite automatic theory, Shannon's information theory, etc. demonstrate an evident groundlessness of each of them for this aim. It may be affirmed at least, with respect to a state of things existing at present. Each branch of the whole scientific trend, connected with the control problem has been developing independently, and only recently the contacts between them are outlined. This report that is a summary statement of a part of a more complete work, prepared on this topic by the author for print is just dedicated to a description of the known things in this field and to a discussion on some possible ways of an investigation of the complex systems control.

§ 1. On an unique conception in the finite automata theory and the dynamic control theory

It is quite natural that, before " A General Control Theory" is created which will enable us to study in detail behaviour of a complex control system, different investigators try to solve a simpler problem, and try to creat a method covering simultaneously, at least, only two of a number of possible abstract treatments of problems. The authors of papers [5, 6], for example, make efforts of uniting the dynamic theory methods and those of the information theory. From this point of view, the opinions stated in work [7] are quite interesting, the author has shown that if one uses some ideas of the abstract algebra, profound analogies existing between the finite automata theory and the dynamic control theory can be revealed. If one represents the Milley's finite automaton, as one makes it usually [8], in terms of five quantities

$$M=\{X, Z, S_y, f_z, f_s\},$$

where X and Z are input and output alphabets of the automaton, respectively,

S_y is a quantities set determining the automaton state,

$Z_y = f_Z(S_y, X_y)$ is a characteristic function, by virtue of which the automaton output quantities are determined, if the input quantities and its state are known,

$S_{y+1} = f_S(S_y, X_y)$ is a characteristic function, by virtue of which the automaton state in the $y+1$ th tact is determined, if the input quantities and the automaton state in the y th tact are known,

then, as it is shown in [7], for a discrete time scale by setting that X, Z, S are representable through the Abelian (commutative) groups, the system (i.e., the automaton) is quite additive, if and only if such homomorphisms [9] exist

$$A: S \to S;\quad B: X \to S;\quad C: S \to Z;\quad D: X \to Z;$$

that means that for all $S_y \in S, x \in X$ the correlations

$$S_{y+1} = f_S(S_y, x_y) = AS_y + Bx_y$$

$$Z_y = f_Z(S_y, x_y) = CS_y + Dx_y$$

take place, and may be written in the more expanded form

$$S_{y+1} = f_S(S_y, x_1, \ldots, x_n) = A^n S_y + \sum_{m=1}^{n} A^{m-1} x_{n-m+1}$$

$$Z_y = f_Z(S_y, x_0, \ldots, x_n) = CA^{n-1} S_y + \sum_{m=1}^{n} CA^{m-1} Bx_{n-m} + Dx_n$$

If now one introduces the notations

$$\Phi(n) = CA^{n-1}$$

$$h(m) = \begin{cases} D & \text{при } m = 0 \\ CA^{m-1}B & \text{при } m > 0 \end{cases}$$

the expression for can be written in this form

$$Z_y = f_Z(S_y, x_y, \ldots, x_n) = \Phi(n) S_y + \sum_{m=0}^{n-1} h(m) x_{n-m}.$$

Substituting, at last the continuous time for the discrete one and passing from a groups homomorphism to a vectorial space homomorphism, the last expression may be given a form, well-known in a linear theory of the dynamic systems control [10]

$$Z(t) = \Phi(t-t_0)x(t_0) + \int_{t_0}^{t} h(t-\xi)U(\xi)d\xi$$

where $Z(t)$ is a P-dimensional vector characterizing the system output,

$x(t_0)$ is h-dimensional vector characterizing the system state at the moment to t_0,

$\Phi(t) - (P \times h)$ is a Z-dimensional vector characterizing the system output,

is a matrix, the i th column of which represents a system reaction at the moment t,

$h(t)$ is a pulse transient function of the system.

The possibility of this type of transitions from the finite automata theory correlations to the linear dynamic theory correlations reveals to us a close connection, which enables us to speak of an unique conception for these two branches of scientific knowledges earlier independently developed. However, it seems to be necessary to make one more step forward along the path of uniting logical and dynamical treatments of complex systems control theory problems. The section below is dedicated to a discussion of this possibility.

§ 2. On logical dynamical systems

It is not difficult to realize that, besides demonstrating of the fact itself of existence of a close connection, revealed between the finite automata theory and the linear dynamical systems theory, for a factual study of complex control systems, which consist simultaneously of logical and dynamical links united immanently as a whole, the construction of an absolutely new theory is required, consequently mathematical means are also required for its completion. The principal difficulty, arising in studying logical dynamical systems (we shall name them so, for the sake of brevity) consists in the fact that it is necessary to find such a supple language, which should enable us to operate in an equally convenient way both with mathematical analysis ordinary variables and logical ones. While creating this type of language one may apparently go along different paths. From this point of view, these things are wo*thy of attention: an -functions language [11] , a calculus of operators for Boolean functions [12] , a continuous logics language [13] and others. In particular, for the same purposes in the work [14] , the concept of an hybrid function $G(x_1,\dots,x_0)$ is introduced that represents a product of the ordinary function of real variable $\mathcal{F}(x_i,\dots,x_n)$ and the function of logical variables $f(x_j,\dots,x_m)$, i.e.

$$G(x_1,\dots,x_e)=\mathcal{F}(x_i,\dots,x_n)\,f(x_j,\dots,x_m).$$

The logical function $f(x_j,\dots,x_m)$ may be a predicate, a formula or a quantifier, but it may assume only two values: 1 (verity) and 0 (falsehood), however, the moments, at which this takes place, depend in a complicated way on a variables value of different nature.

They can be:

1) predicates depending on a real variable function,

2) partly real variables and partly logical variables,

3) only logical variables not depending on real variables.

All this creates various possibilities for describing logical dynamical systems. Unfortunately, the fact itself of hybrid function introduction does not mean yet that we have desirable language already.

We need further investigations due to a rules development, by means of which necessary operations with hybrid functions should be made (to differentiate them, to integrate them and so on), therefore the problem of creation of a mathematical tool, fitting for the study of logical-dynamical systems, remains still unsolved. There are already works, in which a tool based on the hybrid function concept [15] is developed, but there are also critical works [16] , in which the possibility of such a type are denied. Now it may be noted only that G. von Neimann's prediction about a necessity of a continuous and finite mathematics methods merging (i.e. of merging the mathematical analysis methods basing on the fact of variables continuity and the mathematical logics methods operating with discrete logical variables) comes true, and the needs of a logical dynamical systems theory construction will apparently impel many investigators to go at this aim more hurriedly.

It may be also noted that no less difficulties arise in studying systems, which require a simultaneous use of any two other abstract levels of complex systems description, for example, of the logical and the heuristic ones. Still greater difficulties arise, if one uses simultaneously not two, but three levels of the systems abstract description, for example, the informational, the logical and the heuristic ones. A creation of mathematical means permitting to realize investigations in such cases and in some more complicated ones is exactly the principal scope of the complex systems general control theory.

§ 3. On a multidimensionality problem in the complex systems theory

The above-mentioned difficulties are not the unique ones, which one meets while studying systems of similar type.

Essential difficulties arise while using one definite level of an abstract description too, if the system consists of many elements or subsystems, interconnected with each other.

By the way, " the multidimensionality curse" (by Richard Bellman's figurative expression [17]) is overcome equally with difficulty in using any of the levels of abstract description. Thus, for example, the problems are solved easily and elegantly in the finite automata theory with a low dimensionality of systems, and the difficulties increase essentially, as soon as a dimensionality of the system studied grows. Therefore a wish appears, quite naturally, to find paths to overcome difficulties conditioned by a system multidimensionality, which should be , for example, suitable simultaneously for the study of multidimensional dynamical and logical systems of a great dimensionality. In our opinion, one can struggle against a multidimensionality and a multirelationship of complex systems not only by a new mathematical methods research(to what many authors limited themselves on the whole) but by using physical and engineering informations concerning the systems studied. Thus, the fact may be used that the system (it would be of technical, economical, biological or social character) consists of big groups of elements of the same type. If this is the case, we may use a mathematical tool elaborated for the multiphase liquids description, as I.M. Gelfand and M.L. Tsetlin [18] proposed to make it for studying systems of this type. The second possible way of overcoming the difficulties, connected with the problem multidimensionality and considered in [3], is based on physical representations too. Namely: when the system is symmetrical in that or other sense, we may essentially simplify the investigation of a complex system of great dimensionality having both the dynamiwal and the logical treatment of problems. In these cases the groups theory methods [19] and, more exactly, groups representations theory methods [20] may be used, applied so widely in the quantum physics, the quantum chemistry and in the modern theory of elementary particles [21 - 23].

By the way, one should not think that the question is indispensably of a symmetry of purely geometrical character. By no means. The symmetry properties, hidden more deeply, are the most important ones for the investigation problem of great dimensionality systems. Thus, the symmetry property manifests, for many and quite diverse dynamical systems, that the Lagrangian (or the Hamiltonian) remains for them invariant with respect to linear transformations of coordinates. If the presence of symmetry is stated in that or other way (v.g. it may be revealed directly by applying a matricial form of equations), a formal apparatus of the groups representations theory enables us to replace an initial problem of great dimensionality by some problems of the same type, but of a considerably smaller dimensionality (for example, by 10 or 100 times smaller). As a matter of fact the corresponding initial matrix of a greater dimensionality can be reduced, by recipes quite determined and conditioned by the type of an existing symmetry, to a blocking diagonal form

$$\begin{Vmatrix} \Gamma_{ij}^{(1)}(p) & 0 & 0 & \cdots & 0 \\ 0 & \Gamma_{ij}^{(2)}(p) & 0 & \cdots & 0 \\ 0 & 0 & \Gamma_{ij}^{(3)}(p) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \Gamma_{ij}^{(n)}(p) \end{Vmatrix}$$

where each of the submatrices standing on the main diagonal is already irreducible, i.e. it can be represented once more through matrices of a smaller dimensionality. The groups representation theory enables us to answer, what a dimensionality of the submatrices $\Gamma_{ij}(p)$ may be, what their number is, in how many different ways such a type of representation of a matrix of a greater dimensionality by means of matrices of a low dimensionality may be realized and, in particular it gives criteria to judge, whether a further decomposition of the matrices $\Gamma_{ij}(p)$ into matrices of some a smaller dimensionalty and so on is or is not possible. In a concise summary it is impossible to describe all those procedures, which one must by the way, to fulfil practically, viz to divide groups elements into classes, to find the character for each class, i.e. a trace of the submatrix $S_p\|\Gamma_{ij}(p)\|$, to determine a canonical basis for an initial multidimensional vectorial space of the given problem and so forth. We can get familiar with all necessary concepts and theorems (equivalence, and homomorphism of groups, Shoor's lemmas, Lagrange's theorems and so on) and witha techniques of their application through the aforementioned books dedicated to an exposition of the groups theory and the groups representation theory 18 - 23 . So far as an investigator himself has often at his disposal a possibility to form a structure of the studied complex system (for example, while controlling economical systems and industrial objects), one has, in choosing the structure as a symmetrical one, a possibility to apply groups representation theory ideas due to which the latter theory acquires a highly great practical importance.

To explain the essence of the matter we shall consider elementarily a simple example, v.g. vibrations of some mechanical system (see more detailed 4).

Suppose that the equations written in a matricial form have the shape:

$$M\ddot{X} + KX = 0,$$

where M is a diagonal matrix of masses,

K is a rigidities matrix,

X is a variables column-vector.

$$K = \begin{Vmatrix} \frac{3}{2}K_2 & 0 & 0 & -\frac{\sqrt{3}}{2}K_2 & \frac{\sqrt{3}}{2}K_2 \\ 0 & \frac{3}{2}K_2 & -K_2 & \frac{K_2}{2} & \frac{K_2}{2} \\ 0 & -K_2 & \sqrt{3}K_1+K_2 & 0 & 0 \\ -\frac{\sqrt{3}}{2}K_2 & \frac{K_2}{2} & 0 & \sqrt{3}K_1+K_2 & 0 \\ \frac{\sqrt{3}}{2}K_2 & \frac{K_2}{2} & 0 & 0 & \sqrt{3}K_1+K_2 \end{Vmatrix}, \qquad X = \begin{Vmatrix} x \\ y \\ y_1 \\ y_2 \\ y_3 \end{Vmatrix}$$

$$M = \begin{Vmatrix} m_1 & 0 & 0 & 0 & 0 \\ 0 & m_1 & 0 & 0 & 0 \\ 0 & 0 & m & 0 & 0 \\ 0 & 0 & 0 & m & 0 \\ 0 & 0 & 0 & 0 & m \end{Vmatrix}$$

Thus, it is necessary to investigate vibrations of a system, having matrices of dimensionality 5. The problem consists in the fact that, using symmetry property, one should be able to solve, instead of an initial problem, a problem of smaller dimensionality. Applying the known methodics of the groups representations theory [18 - 23] and the fact that the vibrational system used is symmetrical (as it manifests in the matrices symmetry), the new variables $h_{11}, h_{12}, h_{21}, h_{22}, h_{23}$, which are related to the old variables x, y, y_1, y_2, y_3 through such correlations, so that the problem of dimensionality 5 x 5 splits into two identical problems of dimensionality 2 x 2 and one problem of dimensionality 1 x 1. Without expounding here all details of the transformations based on the groups representation theory ideas, we write out final correlations connecting the old and the new variables

$$y_1 = \frac{1}{\sqrt{3}} \sum_{\alpha=1}^{\alpha=3} h_{\alpha 1}, \qquad x = \frac{1}{\sqrt{2}} \left(h_{12} + h_{22} \right),$$

$$y_2 = \frac{1}{\sqrt{3}} \sum_{\alpha=1}^{\alpha=3} e^{\frac{2\pi i \alpha}{3}} h_{\alpha 1}, \qquad y = -\frac{i}{\sqrt{2}} \left(h_{12} - h_{22} \right).$$

$$y_3 = \frac{1}{\sqrt{3}} \sum_{\alpha=1}^{\alpha=3} e^{-\frac{2\pi i \alpha}{3}} h_{\alpha 1},$$

The same correlations may be written in a matricial form as follows:

$$X = RH,$$

where the matrix R and the column-vector H have the shape

$$\left\| \begin{matrix} 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{i}{\sqrt{2}} & 0 & \frac{i}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} e^{\frac{2\pi i}{3}} & 0 & \frac{1}{\sqrt{3}} e^{\frac{2\pi i}{3}} & 0 & \frac{1}{\sqrt{3}} e^{2\pi i} \\ \frac{1}{\sqrt{3}} e^{-\frac{2\pi i}{3}} & 0 & \frac{1}{3} e^{\frac{2\pi i}{3}} & 0 & \frac{1}{\sqrt{3}} e^{2\pi i} \end{matrix} \right\|, \qquad H = \left\| \begin{matrix} h_{11} \\ h_{12} \\ h_{21} \\ h_{22} \\ h_{23} \end{matrix} \right\|$$

The feedback of the new variables with the old ones in matricial shape is written in the form of equation

$$H = R^{-1} X,$$

where R^{-1} is a matrix, inverse to the matrix R. Substituting the linear transformation $X = RH$ for the initial equation and multiplying the left side by the matrix R^{-1}, we reduce the initial system of differential equations to a blocking diagonal form:

$$R^{-1} M R \ddot{H} + R^{-1} K R H = 0$$

where $R^{-1}MR$ is a masses diagonal matrix of the shape

$$R^{-1}MR = \begin{Vmatrix} m & 0 & 0 & 0 & 0 \\ 0 & m_1 & 0 & 0 & 0 \\ 0 & 0 & m & 0 & 0 \\ 0 & 0 & 0 & m_1 & 0 \\ 0 & 0 & 0 & 0 & m \end{Vmatrix}$$

and $R^{-1}KR$ is a blocking diagonal rigidities matrix of the chape

$$R^{-1}KR = \begin{Vmatrix} \sqrt{3}\,K_1 + K_2 & \frac{\sqrt{3}\,K_2 i}{\sqrt{2}} & 0 & 0 & 0 \\ -\frac{\sqrt{3}\,K_2 i}{\sqrt{2}} & \frac{3}{2} K_2 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{3} K_1 + K_2 & \frac{\sqrt{3}\,K_2 i}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{\sqrt{3}\,K_2 i}{\sqrt{2}} & \frac{3}{2} K_2 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{3} K_1 + K_2 \end{Vmatrix}$$

Thus, on the basis of a quite determined procedure, we have come to a blosking diagonal shape of matrices. One makes this not on the basis of a method of attempts, but according to quite definite algorithms in the method given, if the initial syste is symmetrical. The differential equations system in terms of the new variables has the form:

$$\left.\begin{aligned} m\ddot{h}_{11} + (\sqrt{3}K_1 + K_2)h_{11} + i\sqrt{\tfrac{3}{2}}K_2 h_{12} = 0 \\ m_1\ddot{h}_{12} + \tfrac{3}{2}K_2 h_{12} - i\sqrt{\tfrac{3}{2}}K_2 h_{11} = 0 \end{aligned}\right\}$$

$$\left.\begin{aligned} m\ddot{h}_{21} + (\sqrt{3}K_1 + K_2)h_{21} - i\sqrt{\tfrac{3}{2}}K_2 h_{22} = 0 \\ m_1\ddot{h}_{22} + \tfrac{3}{2}K_2 h_{22} + i\sqrt{\tfrac{3}{2}}K_2 h_{21} = 0 \end{aligned}\right\}$$

$$m\ddot{h}_{31} + (\sqrt{3}K_1 + K_2)h_{31} = 0$$

Thus, one can see clearly that the matter comes, instead of solving the problem of dimensionality 5 x 5, to a twofold solution of the same problem of dimensionality 2 x 2 and to a separate solution of an equation of dimensionality 1 x 1.

We take an interest in the complex systems control problem, and the method just exposed gives a possibility to investigate a controlled member taken separately, having a symmetrical structure, if the control system is asymmetrical. Therefore, in these cases, the stability, invariance, optimality problems and others are independent ones, each of which is solved on the basis of using the controlled member symmetry properties. Thus, the stability problem may be solved by means of an application of a decomposition method, according to which all the system is disintegrated into subsystems. The controlled member described by a symmetrical matrix enters the first of these subsystems, and the control system enters the second subsystem. At first one investigates a stability of each of the subsystems. Moreover, it is known that one uses, while studying the stability of a controlled member taken separately, their symmetry properties. Then, determining the subsystem stability and on the basis, for example, of the known Baley's theorem, the asymptotic stability conditions can be determined for the system as a whole. The optimal control problems are particulary interesting for controlled members of this type, when a common use of the decomposition method and of the groups representation theory ideas enables us to solve highly difficult multidimensional problems of optimal control. The factual demostration of a similar type of possibilities requires a separate report.

§4. On universality of electronic digital computers and on a complex systems control processes simulation

The principal particularity of the complex systems theory, as of the Cybernetics in general, becomes apparent in the fact that it must give possibilities to study objects of any nature (technical, economical, biological, social etc.) at a corresponding abstract level. Therefore it is necessary to take care not so much of a creation of " A general systems theory", from which particular theories as special cases should result (the linear dynamic systems theory, the information theory, the Markovian processes theory and so on), which [24] call for at times, as of a creation of many branches of a theory for different levels of a systems abstract description. A consideration of the problem at a singular abstraction level enables us to receive answers only for a certain group of questions, and in order to receive answers for other questions it is necessary to make an investigation at an abstract description level of other systems. To reach a maximally possible completeness of an information it is necessary to study the same system at all abstraction levels suitable for the case given. A path of methematical simulation is the most expedient for this purpose, and it is practically accessible. Although one usually looks at an electronic digital computer as a high-speed computer, but it is in reality a universal device, which besides a rapid computation, can make the processing of alphabetical or other symbols, transform an information into a form that we need, draw "conclusions", "deductions" and so on. All this enables us to study by means of an electronic digital computer not only informational processes in a complex systems control, but logical, dynamical and heuristic treatment. Nowadays special abstract languages are created (SIMSCRIPT,SIMPAC and others, [25]) permitting to economize the time and efforts due to a programming and a process itself of simulation. Therefore just this path of using the electronic digital computer for a simulation is now principal while developing really complex control systems. One often undertakes even the creation of special scientific centers intended exceptionally for the simulation purposes of a complex control system developed. As an example, one may mention a scientific center, created especially for developing an automatic system, controlling air traffic over the Western Europa territory [26] . To characterize a real complexity of control systems of this type one may give data about programs needed for a control by means of electronic digital computer and other technical means (radars, communication equipment etc.) of aircraft flows (with total number 300 - 600) over a territory of about 2000 km [27] . One needs efforts of 250 programmists during two years only for a development of such programs, and the total quantity of commands reaches 2.5 millions. In spite of the fact that the cost of scientific simulation centers of this type is sufficiently high, the economical expediency of their creation is doubtless while developing really complex control systems, and one follows this path in many cases both while solving technical, economical or defensive problems and fulfilling of great social investigations. Therefore the problem of developing a general theory of algorithms transformation (acad. V.M. Glushkov [28]) and of a complex systems theory in general becomes (some aspects of which were considered above) most significant, because so powerful means of mathematical modelling can be used completely only in a case of its presence. Of course, only some complex systems theory general questions are marked in this brief survey, and its numerous important aspects, which are due to a complex systems reliability problematics [29 - 30] and their efficiancy [31, 32] etc.

References

1. A General Systems Theory. Publishing House " The World", Moscow, 1966.
2. Acad. A.I.Berg, J.L. Cherniak. Information and Control. Publishing House " The Economics", Moskow, 1966.
3. A.I. Kukhtenko. Principal trends of a develpment of the complex systems control theory. In the collection "Complex Control Theory", Issue IV, Publishing House " The Scientific Thought", Kiev, 1968.
4. A.I. Kukhtenko. Principal Problems of the Complex Systems Control Theory. Transactions of the Seminar "Complex Control Systems". Issue I, Kiev, 1968. Publishing House of the Institite of Cybernetics of the Ukrainian Soviet Socialist Republic.

5. A.A. Krassovsky. Continuous Dynamic Systems Entropy Change. News of the Academy of Sciences of the USSR. "Technical Cybernetics", N 5, 1964.
6. R.L. Stratonovich. Theory Information Extremal Problems and a Dynamic Programming. News of the Academy of Sciences of the USSR. "Technical Cybernetics", N 5, 1967.
7. M. Arbib. A Common Framework for Automata Theory and Control Theory. I Soc. Industr. and Appl.Math., 1965, A.3. N 2.
8. A. Gill. Introduction into a Finite Automata Theory. Publishing House " The Science", 1966.
9. R.Fort, A.Kauffmann, M.Denis-Papin. A Modern Mathematics. Publishing House "The World", 1966.
10. V.M. Brown. An Analysis of Linear Systems Invariants in Time. Publishing House "Machine-Building", 1966.
11. V.L. Rvachev. Logics Algebra Geometrical Applications. Publishing House " Technics", Kiev, 1967.
12. I. Richalet. Calcul Opérationnel Booléen, "L'onde électrique", 1965, N 439, v. 43, p. 63.
13. S.A. Ginsburg. A Continuous Logics and its Applications. AIT, N 2, 1967.
14. R. Terno. Hybrid Functions Are a New Method of Complex Systems Description. News of the Academy of Sciences of the USSR. "The Technical Cybernetics", N 6, 1965.
15. M.M. Gerdov. One Numerical Analysis Method of Complex Systems. News of the Academy of Sciences of the USSR. " The Technical Cybernetics", N 6, 1967.
16. G.P. Sbrodov. Logical Relations and Complex Control Systems Description. Proceedings of the Conference on the Economical Cybernetics, Batumi, 1966.
17. R. Bellman. A Trend of Mathemetical Investigations in the Nonlinear Circuits Theory. The Collection "Mathematics", 8, 15. Publishing House of Foreign Literature, 1964.
18. I.M. Gelfand, M.L. Tsetlin. On Some Complex Systems Control Methods. "Successes of Mathematical Sciences", v. 16, Issue I, 1962.

19. A.G. Kurosh. A Groups Theory. Publishing House " The Science", 1967.
20. F.D. Murangan. A Groups Representation Theory. Publishing House of the Foreign Literature, 1950.
21. E. Vigner. A Groups Theory and its Application in the Quantum Mechanics. Publishing House of the Foreign Literature, 1947.
22. V. Heine. A Groups Theory in the Quantum Mechanics. Publishing House of the Foreign Literature, 1963.
23. G.I. Liubarsky. A Groups Theory and its Application in the Physics. Publishing House of the Physical and Mathematical Literature, 1958.
24. M.Mesarovic. A General Systems Theory. In the Collection " A General Systems Theory". Publishing House " The World", 1966.
25. G. Markovits et al. Simscript. An Algorithmical Language for a Modelling. Publishing House. "The Soviet Radio", 1966.
26. Control. II, N 105, 1967.
27. T.B. Steel. The Development of Very Large Programs. "Proc. IFI. P. Congr. N 1, vol. 1, 1965.
28. V.M. Glushkov. Perspectives of the Automation of a Computers Design. The Herald of the Academy of Sciences of the USSR, N 4, 1967. Publishing House " The Thought".
29. I.P. Popchev. Great Systems Structure. Principal Dependencies and Description Methodics. "The Technical Thought", N 1, 1967.
30. I. Popchev, M.Panova. A Determination of a Great Systems Efficiancy Index. News of the Institute of the Technical Cybernetics, v. VI, 1967.
31. B.S. Fleischmann. On a Complex Systems Vitality. News of the Acádemy of Sciences of the USSR."The Technical Cybernetics", N 5, 1966.
32. B.S. Fleischmann. Complex Systems Efficiency Statistical Limits. The Collection "Applied Problems of the Technical Cybernetics". Publishing House "Soviet Radio", 1966.

29.1

CONSTRUCTION OF CHECKING AND DIAGNOSTIC PROCEDURES FOR GENERAL-PURPOSE UNIFORM ARRAYS

I.V.Pranghishvily
V.V.Ignatushchenko

Institute of Automatics and Telemechanics
Moscow, U.S.S.R.

Introduction

Uniform arrays represent a new class of networks consisting of regularly interconnected identical cells capable to be preprogrammed to perform the desired functions. Investigations carried out in the Institute of Automatics and Telemechanics, and also in some other organizations in the U.S.S.R. and abroad showed that uniform programmable general-purpose arrays are among the most effective and promising vehicles for implementation of digital computers and control systems based on integrated circuits, especially LSI-based. Both uniform programmable microelectronic arrays and functions realized by them are advantageous in that they give general-purpose, flexible, economic, reliable and "hardy", easy-and-cheap-to-fabricate, high-capacity devices of a high standardization level.

We are going to show in this paper that regularity of uniform arrays leads to considerable simplification in the construction of checking and diagnostic procedures, and to drastic reduction in the time necessary for checking and diagnosis of arrays, irrespective of the function specified.

It is assumed here that any sequential function realized by a uniform array, is in order if all functional elements (cells) of the array used to perform it are in order. Then for any function realised by a uniform array, checking and failure diagnosis are reduced to testing of the array cells. Since all cells of a uniform array are identical and have identical interconnections, the tests that make the checking and diagnostic procedure for a uniform array as a whole, are sets of tests forming the procedure for one array cell.

Here we discuss those arrays described in Ref.[1)] whose functional element is a finite automaton with **k** inputs, **q** outputs, and memory that serves only to set up the cell. **p** of **k** inputs correspond to the interconnections, i.e. they are identified with the outputs of other elements of the array. We shall call them further "p-inputs" . The remaining **k-p** inputs are used for control, and setting up signals are fed along them from outside into the element.

Given a (k,q)-terminal network (functional element of the array), one can construct for each of its outputs a minimal (elementary) checking procedure Π min by the use of the technique of Refs.[2,3] . Let the minimal procedure Πmin for one output include **n** tests:

$$\Pi_{min} = t_1 \cdot t_2 \ldots t_i \ldots t_n \tag{1}$$

In general, a uniform array has external connections through peripheral cells, and through control inputs. For instance, Fig.1 shows that a uniform array $m_1 \times m_2$ whose cells have four inputs and four outputs (i.e. a signal may be transmitted through a cell in four directions), has $2m_1 + 2m_2$ external p-inputs and the same number of external outputs.

Since each p-input and output of an array functional element is identified with a certain direction of signal transmission, it is convenient to regard an uniform array functional element as a set of p=q identical elementary transmitting channels that interact inside it. Consequently the functional element of the array shown in Fig.1 is regarded as a set of four elementary channels.

From this point of view, elementary procedure Π min serves to test one of all channels with due regard to other channels. It is assumed here that the output of a cell elementary channel F_i is the input of a similar channel F_i of a neighbouring cell.

Let us give definitions for tests that are fed into adjacent cells a and b connected by the i-th channel.

Let cells a have at its inputs signals $(\sigma_1 \dots \sigma_i \dots \sigma_p, \sigma^H_{p+1} \dots \sigma^H_K)^a$ that correspond to the test t_i from Π min, where assumes the values 0 or 1, σ_i is the signal at the i-th p-input (i.e. the signal at the input of i-th elementary channel), σ^H_{p+i} is the signal at the (p+i)-th control (set-up) input. If cell a is in order, the signals $(\sigma_{out\,1} \dots \sigma_{out\,i} \dots \sigma_{out\,p})^a$ will appear at its outputs. If the i-th channel of cell a is out of order, or the cell a as a whole is out of order, the signal $\bar{\sigma}_{out\,i}$ will appear at the i-th output.

We shall call the test $(\sigma_1 \dots \sigma_i \dots \sigma^H_{p+1} \dots \sigma^H_K)^b$ fed into the cell b, <u>compatible</u> in the i-th <u>direction</u> (channel) with the test $(\sigma_1 \dots \sigma_i \dots \sigma^H_{p+1} \dots \sigma^H_K)^a$ fed into cell a, if value of $\sigma^a_{out\,i}$ at the i-th output of a good cell a coincides with signal value at the input of the i-th channel cell b in the test $(\sigma_1 \dots \sigma_i \dots \sigma^H_{p+1} \dots \sigma^H_K)^b$ i.e. $\sigma^a_{out\,i} = \sigma^b_i$. It is obvious that if $\sigma^a_{out\,i} \neq \sigma^b_i$, the test $(\sigma_1 \dots \sigma_i \dots \sigma^H_{p+1} \dots \sigma^H_K)^b$ may not be fed into the cell b at all.

Let for instance the test $(\sigma_1 \sigma_2 \sigma_3)^a = 001$ be fed into the cell a of the bidirectional uniform array shown in Fig.2, and let the signal $\sigma^a_{out} = 0$ appear at the horizontal output of the good cell a. Then the cell b may be fed only with such test which is compatible with $(\sigma_1 \sigma_2 \sigma_3)^a$, i.e. has at the horizontal input the signal value $\sigma^b_3 = 0$. If tests t_i and t are compatible with each other, we shall call these two tests <u>intercompatible</u>.

Fig.2 shows a special case where compatible tests for cells a and c that are neighbours along vertical line coincide $((\sigma_1 \sigma_2 \sigma_3)^a = (\sigma_1 \sigma_2 \sigma_3)^c = 001)$. We shall call such tests <u>self-compatible</u>.

We shall call the compatible test $(\sigma_1 \dots \sigma_i \dots \sigma^H_{p+1} \dots \dots \sigma^H_K)^b$, fed into the cell b, <u>conjugate</u> in the i-th direction (channel) with the test $(\sigma_1 \dots \sigma_i \dots \sigma^H_{p+1} \dots \sigma^H_K)^a$

fed into the cell a, if change of signal value from $\sigma^a_{out\,i}$ to $\bar{\sigma}^a_{out\,j}$ at the i-th output of the faulty cell a makes the cell b to change its i-th output from $\sigma^b_{out\,i}$ to $\bar{\sigma}^b_{out\,i}$.

If each cell, following the cell a in the i-th direction is fed with a test that is conjugate with the test fed into the preceding cell it becomes possible to obtain at array external output signals $\sigma^a_{out\,i}$ or $\bar{\sigma}^a_{out\,i}$ corresponding to the good and faulty states of the cell a .

The sets of tests $[(\sigma_1 \ldots \sigma''_k)^a, (\sigma_1 \ldots \sigma''_k)^b, \ldots$ $\ldots (\sigma_1 \ldots \sigma''_k)^r]$, where each test is conjugate in i-th direction with the preceding one, and the last with the first, will be called the test <u>period</u> (r).

In a special case conjugate tests may coincide. Such tests will be called <u>self-conjugate</u>, their period contains one test, (r=1).

If a period contains two tests (r=2), these tests will be called <u>interconjugate.</u>

The functional element of a uniform structure may be set up so that the same signal $\sigma_{out\,i}$ appears at several outputs corresponding to different channels (rather than at one output), i.e. the output signal of the cell is reproduced, and $\sigma_{out\,i} = \sigma_{out\,j}$.

Let a cell a, which is set up so that $\sigma^a_{out\,i} = \sigma^a_{out\,j}$ be fed with test ti from Πmin. If its neighbour in the j-th direction (a cell c) is fed with such test t, that change of the cell c j-th input signal from $\sigma^a_{out\,i} = \sigma^a_{out\,j} = \sigma^c_j$ to $\bar{\sigma}^a_{out\,i} = \bar{\sigma}^a_{out\,j} = \bar{\sigma}^c_j$,results in change of the cell c output signal value from $\sigma^c_{out\,j}$ to $\bar{\sigma}^c_{out\,j}$, we shall call the test t <u>conjugate</u> with the test ti in <u>intersecting directions</u> (i and j).

The problem of checking uniform array cells (detection of the fact of existence of faulty cells (x)) may

---

(x) It is implied here and henceforth that there is one or more non-compensating faults.

be formulated then in terms of checking array channels with tests from Πmin. In so doing one should feed conjugate tests into the cells situated in between checked ones and external outputs. The most convenient tests are those that make cells to perform connecting functions.

The problem of diagnosis consists in finding two co-ordinates of a faulty cell and is reduced then to the checking of the cell in two intersecting directions (channels).

The less is the number of tests in a period, the less is the time (number of clock cycles) necessary to check array cells along the given direction. Indeed, if a test is self-conjugate, it can simultaneously (during one clock cycle) check all cells along an i-th direction (from first to the last that has external i-output). For instance, 1-channels of all first row cells in the array shown in Fig.1, may be checked during one clock cycle. If, in addition, the given test is self-compatible in other channels all i-th channels of all array cells may be checked simultaneously during one clock cycle.

If the given test ti from Π min is interconjugate (in an i-th channel) with a test t from Π min, then the test ti can during one cycle check half of the cells (to the accuracy of the greater integer) and the test t checks the rest of them. At the next cycle the tests must be interchanged. In the case of a test t (interconjugate with the given one) not being included in the minimal procedure Π min, the given test ti needs two clock cycles to check the i-th channels.

Thus the time for checking array cells with the minimal procedure Π min depends on the number of tests in periods constructed for each test in Π min.

For each test ti from Π min the conjugated in an i-th channel tests are found as follows:

a) The values of ($\sigma_1 \ldots \sigma_i \ldots \sigma_k^{*}$)=ti are substituted into the truth table or algebraic expression of the function Fi (i.e. output function of the i-th channel of the cell), and $\sigma_{out\,i}$ = Fi is found. If $\sigma_{out\,i} = \sigma_i$

the given test is selfcompatible;

b) if the test t_i is selfcompatible, it is examined for being selfconjugate, to this end the values of $(\sigma_1 \dots \bar{\sigma}_i \dots \sigma_K^{"})$ are substituted into the truth table or algebraic expression of F_i. If $\bar{\sigma}_{out\,i}$, is the result, the given text is self-conjugate;

c) If the given test is not selfconjugate (in an i-th channel), one should pick out from the truth table those combinations of input signals $(\sigma_1 \dots \sigma_i \dots \sigma_K^{"})' = t'$, which have $\sigma_i' = \sigma_{out\,i}$; these combinations correspond to tests that are compatible with the given test t_i.

d) from the tests of t' type those tests t'' are chosen, which have $\sigma_{out\,i}'' = \sigma_i$. Then the given test t_i will have intercompatible tests of t'' type.

e) the tests interconjugate with the given one (t_i), are picked out from tests of t'' type. For this purpose the values of $(\sigma_1 \dots \bar{\sigma}_i \dots \sigma_K^{"})''$, and $(\sigma_1 \dots \bar{\sigma}_i \dots \sigma_K^{"})$ are substituted (similar to the point b)) into the truth table or algebraic expression of F_i. This corresponds to $\bar{\sigma}_i'' = \bar{\sigma}_{out\,i}$ and $\bar{\sigma}_i = \bar{\sigma}_{out\,i}''$. If in doing so the value of F_i changes to the inverse one (i.e. there are $\bar{\sigma}_{out\,i}''$ and $\bar{\sigma}_{out\,i}$, respectively), the corresponding test is interconjugate with t_i.

f) if no one test is found that is interconjugate with the given one, then from the tests of t' type compatible with t_i, the tests of t''' type having $\sigma_{out\,i}''' \neq \sigma_i$ are chosen. Now the tests of t'''' type are looked for, which are conjugate with the test of t''' type and have $\sigma_{out\,i}'''' = \sigma_i$. If they are not found, search goes on among tests of t'''' type until a test is found that is conjugate with the previous one and has $\sigma_{out\,i}^{(z)} = \sigma_i$, i.e. until a period of conjugated tests is found. For uniform array cells known to authors, the test period does not exceed three tests.

In order the given checking test t_i to be also diagnostic one, it is necessary to find out a test t, which is conjugate with the given one in a direction intersecting the i-th direction. Simultaneously with feeding the given test t_i into cells connected by the i-th channel, the neighbouring cells are fed with the conjugate test t . If it is necessary to diagnose e.g.cells situated in the 1st row of the array shown in Fig.1, the checking tests from Π min arefed into the cells of the 1 row providing for transmission of "in order/out of order" signal to the external output of the channel 1. At the same time the cells of all other rows are fed with tests that are conjugate with the given one in the 2nd channel. This provides for the transmission of "out of order" signals from faulty cells of the 1 row to vertical outputs of the lower row cells.

It is clear now that those checking and diagnostic tests are most effective that are selfconjugate in two intersecting channels (directions). One such test enables all appropriate channels of array cells to be examined simultaneously (during one clock cycle). It serves not only to detect the presence of a failure in the array, but to find out coordinates of faulty cells too.

The above discussion leads to the following algorythm for construction of uniform array checking and diagnostic procedures (Method 1):

1. By the use of the technique of Refs.2,3, a set of minimal checking procedures for the channels of an array functional element is found. For each test from the minimal procedure, the minimal conjugate test period (including the given test) is found.

2. Of all minimal checking procedures for the array cell that one is chosen which contains more selfconjugate (in two intersecting directions, or channels) tests. All array cells are checked and diagnosed simultaneously (during one clock cycle) with each of such tests.

3. For each test t_i which belongs to the elementary procedure (found in accordance with the point 2.) and is selfconjugate only in the i-th channel, a test t is found which is conjugate with the given one in a direction which intersects the direction i. A group of cells whose

i-th channels are set up so that $\sigma out_i = \sigma out_j$ is isolated. For example, the array shown in Fig.1 has cells of one row (connected with channels 1) isolated and set up to perform $\sigma out_1 = \sigma out_2$. The isolated group of cells is fed with the test ti (selfconjugate in the ith channel), and the rest of the array is fed with the tests t . If the isolated group has some faulty cells, the "out of order" signals will appear at the external output of the i-th channel and at the corresponding external outputs of the j-th channels.

4. Checking and diagnosing of the isolated group of the i-th channels with the rest of tests (included in the procedure found in accordance with point 2) is performed similar to point 3.

Points 3 and 4 are performed separately for each group of cells having connected i-channels. In Fig.1,for instance, each row of cells is examined.

The rest of channels is examined in a similar manner.

The total number of uniform array checking and diagnostic procedures ($M_{c.d.}$ by Method 1) and the time necessary for them (number of clock cycles (N)) is related to the number of array functional elements (elementary channels) as follows:

$$(M_{c.d.} = N) \leq \ell \cdot \left\{\tfrac{p}{2}\right\} + (\ell' + \ell'' + \ell''' \cdot z)(m_1 + m_2 + \dots + m_p) \qquad (2)$$

where: ℓ is the number of tests in the minimal procedure min that are selfconjugate in two crossing directions;

$\left\{\frac{p}{2}\right\}$ is the number of pairs of intersecting channels (directions) rounded off to the greater integer (e.g. for the Fig.1 array, where $p=4, \left\{\frac{p}{2}=2\right\}$);

ℓ' is the number of tests from Π min that are selfconjugate (r=1) only in i-th channel (direction);

ℓ'' is the number of tests from Π min that are interconjugate (r=2) in i-th channel;

$\ell^{\prime\prime\prime}$ is the number of tests from Π min that are conjugate in i-th channel and have period $r \neq 2$;

$m_1, m_2 \ldots m_i \ldots m_p$ is the number of groups of connected elementary channels of 1st, 2nd, i-th,...p-th directions (.e.g the array of Fig.1 has m_1 groups (rows) of connected channels of direction 1, m_2 groups (columns) of connected channels of direction 2).

Uniform array testing time may be made independent of the number of array elements. For this purpose, it is necessary to check all groups (rows, columns) of cells or a certain part of them with conjugate (in i-th channel) tests rather than one group. If some of the i-th channels have at their external outputs "out of order" signals, the cells are diagnosed by Method I to find second coordinates of faulty cells. In accordance with this, the following method for checking and diagnosis of uniform array (Method II) is proposed;

1. The point 1 of Method 1 is to be performed.

2. An elementary procedure is chosen which belongs to the set of array cell minimal procedures and has the greatest number of pairs of intercompatible (in j-th channel) tests selfconjugate in i-th channel. If there are several such procedures, that one is chosen which has the greatest number of selfconjugate (in two intersecting directions) tests.

3. The i-th channels of all uniform array cells are checked with each pair of the above mentioned (point 2) tests during two cycles. All other channels of array cells in this case, are fed with set-up signals providing for compatibility of their inputs and outputs. All channels are checked in a similar manner.

4. During $\left\{\frac{p}{2}\right\}$ cycles all channels of all cells are checked with selfconjugate (in two intersecting directions) tests.

5. All i-th channels of all array cells are checked then with the rest of tests during r cycles. Thereupon the rest of channels is checked.

The number of checking procedures (M'_c) and checking time (number of cycles (N') for uniform array (irrespective of number of cells) are defined in Method II as follows:

$$(M'_c = N') \leq \ell \cdot \left\{\frac{p}{2}\right\} + (\ell' + \ell'' + \ell''' \cdot \tau + \tau') \cdot p \qquad (3)$$

where r' is the total number of auxiliary tests t_{aux} that are fed into the elements situated in between m- and (m+r)-th groups of the i-th channels checked with the given test t_i. If, at the same time, t_i is selfcompatible in any other (j-th) channel, but the i-th one, or is intercompatible in the sence of point 2, tests t_{aux} are looked for that are not included in Π_{min}, but are compatible with t_i in j-th channel. Thus, tests t_{aux} serve to separate cells, if t_i may not be fed into neighbouring groups of elements whose i-th channels are not connected together. Only one coordinate of each faulty cell is found during N' clock cycles. In order to find out the second coordinates of m-th group faulty cells, a checking and diagnostic procedure is used that corresponds to checking (with the same test t_i) i-th channels of m-th group by Method I. Thus the total number of uniform array checking and diagnostic procedures ($M'_{c.d.}$) by Method II is as follows:

$$M'_{c.d.} = M_{c.d.} + M'_c - \ell \cdot \left\{\frac{p}{2}\right\} \qquad (4)$$

Only M'_c+S procedures are needed to diagnose S faults, since each fault needs only one procedure from $M_{c.d.}$. One additional clock cycle is needed to find out the second coordinate of each faulty cell. Thus S faults are diagnosed in N'+S cycles.

Examples of Procedures for Uniform Arrays.Majority Array.

The majority array of Ref[4] capable of performing an arbitrary switching function consists of three-input cells. (Fig.3). Only cells along edges have their horizontal and vertical inputs and outputs led outside. The central (control) inputs to all array cells are external. Each cell has one bifurcating output, i.e. $\sigma_{out\,i}=\sigma_{out\,j}$.

By the use of the technique of Refs.[2,3], a set of minimal procedures is found for checking the cell performing the function f=AB+AC+BC (Fig.4):

$$\Pi \min\,(\sigma_1=A,\ \sigma_2=B,\ \sigma_3=C) = (010,101,011,001) \vee (010,001,011,110) \vee$$
$$\vee\ (101,001,100,110) \vee (010,100,011,110) \vee$$
$$\vee\ (010,101,100,110) \vee (001,011,101,101) \quad (5)$$

According to the points 1 and 2 of Method I, the minimal procedure

$$\Pi \min\,(ABC) = 010,101,\ 011,\ 001 \quad (6)$$

is picked out, because it has two tests selfconjugate in two channels (010,101). Indeed, for A=0,B=1 and C=0 we get σ_{out} =f=0, i.e. $\sigma_1=\sigma_3=\sigma_{out}$, and the test 010 is selfcompatible in the channels A and C. Since alteration of signal value at the input of A, or C results in $f=1=\bar{\sigma}_{out}$, the test 010 is also selfconjugate in these channels.

The 4x4 majority array (Fig.5.a) is checked and diagnosed with the test 010 in one cycle T_1. To do this the external inputs of extreme array cells A and C are fed with zero signals, and the central inputs of all cells are fed with 1 signals. If any cell is out of order (e.g. in the 3rd row and 2nd column), itwill generate 1 instead of 0, and all cells to the right and lower the faulty one will change their outputs from 0 to 1. Thus both coordinates of the faulty element will be determined. In a similar manner all cells will be diagnosed with the test 101 during cycle T_2 (Fig.5,б).

The test 011, belonging to procedure (6), is selfconjugate in the horizontal channel (C), because $f = \sigma_{out} = 1 = C$ at A=0, B=C=1, and $f = \bar{\sigma}_{out} = 0$, if C changes to 0. According to point 3 of Method I, the test 011 is fed into all elements of the first row (Fig.5,б). All other cells are fed with the selfconjugate test 101 which is conjugate with 011 in the vertical channel. Thus, the test 011 checks and diagnoses all cells of the 1-st row during one cycle (T_3), all cells of the 2nd row during T_4, 3rd T5 and 4th T6 (Fig.5,г). If, for instance, cell (2,2) is faulty, it will generate 0 instead of 1 at the output. This will result in generation of 0 (instead of 1) at outputs of all cells to the right and below the faulty one, and both coordinates of the faulty cell will be determined.

The last test from (6) (001) is selfconjugate in the vertical channel (A). According to the point 3 of Method I, it checks and diagnoses all cells of the 1-st column during one cycle (T_7) (Fig.5,д). All other cells are fed with selfconjugate test 010 which is conjugate with 001 in the horizontal channel. The cells of 2nd, 3rd and 4th columns are checked and diagnosed during cycles T8, T9, and T10, respectively. Thus, all cells of the 4x4 majority array are checked and diagnosed with the tests of procedure (6) during 10 cycles.

Now let us construct a checking procedures M_C^I for the same array by the use of Method II. According to point 2 of Method II, the procedure

$$\Pi_{min}(ABC)=(010,001,011,110) \qquad (7)$$

from the set (5) of majority cell minimal procedures is chosen, since it contains two intercompatible (in the horizontal channel) tests (001 and 110). Each of them is selfconjugate in the vertical channel.

During the cycle T_1 the test 001 checks all cells of odd columns (see hatched cells in Fig.6,a), and the test 110 checks cells of even columns. In the cycle T_2 the tests are interchanged (Fig.6,б).

During cycle T_3 all cells are checked with test 010 (Fig.5,a). During cycle T_4 cells of odd rows are checked with selfconjugate (in horizontal channel) test 011 (Fig. 6,б). All cells of even rows are fed at the same time with the auxiliary test 100 which is intercompatible with 011, but is not included in (7). The cells of even rows are checked during cycle T_5 with 011 (Fig.6,г). If the cell (2,2), for instance, is out of order, the corresponding signal (shown in brackets in Fig.6,г) will appear at the output of the 2nd row when it is checked with the test 011. In order to find out the second coordinate of a faulty cell it is necessary to diagnose the array with a test, corresponding to checking cells ofthe 2-nd row with the same test 011 by Method I (Fig.5,г). Thus, the majority array is diagnosed during 5 clock cycles, and the time for diagnose by Method II is independent of array size.

The Uniform Array Capable of Performing an Arbitrary Sequential Function

Let us construct a checking procedure $M_C^!$ (Method II) for a four-directional array of arbitrary size (Fig.1). Fig.7 shows the functional diagram of one of four channels of the array cell, where A,B,C,D are external inputs serving to control (set up) the memory cells (the flip-flops TP5 and TP1), X is working input (or p-input), F is the channel output. The channel performs the logical function

$$F = X \oplus y_{TP5} + y_{TP1} \qquad (8)$$

where $\oplus$ is the "sum by moduls two" function.

It is assumed in the procedure construction for (8), that during all cycles T of channel F testing A(t)=0, i.e. the flip-flop reset signals are absent. It is also assumed that the inputs C and D get pulse signals i.e., their duration is less than T. The signal at a flip-flop output should be constant (either $y_{TP}=0$, or $y_{TP}=1$). Under these restrictions, and with signal duration at the B, C, and D inputs disregarded, one can determine y_{TP5} and y_{TP1}

as follows:

$$y_{TP5} = BC\,,\quad y_{TP1} = BD \qquad (9)$$

Substituting (9) into (8), one gets:

$$F=[X\oplus(BC)]+BD \qquad (10)$$

The procedure

$$\Pi\min\,(XBCD)=(1110,\ 0110,\ 1100,\ 0011,\ 0101) \qquad (11)$$

is one of the elementary procedures for circuits implementing the function (10). To see this, look through the Table 1, where the tests 1 and 2 are interconjugate, and the tests 3 and 4 are conjugate.

Table I

No.No.	X	B	C	D	F
1	1	1	1	0	0
2	0	1	1	0	1
3	1	1	0	0	1
4	0	0	1	1	0
5	0	1	0	1	1

As Fig.8 shows, the array functional element is an aggregate of four similar channels (F_1, F_2, F_3, and F_4). Thus, it is described by the system of four equations.

$$\left.\begin{aligned} F_1(i,j)&=[\underbrace{F_1(i,j-1)\cdot F_2(i-1,j)\cdot F_3(i,j+1)\cdot F_4(i+1,j)}_{X}]\oplus y_{TP5}+y_{TP1}\\ F_2(i,j)&=X\oplus y_{TP5}+y_{TP2}\\ F_3(i,j)&=X\oplus y_{TP5}+y_{TP3}\\ F_4(i,j)&=X\oplus y_{TP5}+y_{TP4}\end{aligned}\right\}\qquad (12)$$

Test 3 is selfconjugate in any pair of intersecting channels. The array testing consists in testing all elementary channels F of all cells with the tests of Table 1, hand in hand with testing conjunction in X of (12).

According to points 3 and 6 of Method II, the F_1 channels of all odd cells in each row are fed with the first test (1110), and even cells get the test 0110 which is interconjugate with the first one (See Fig.9a,).

All other channels (F_2,F_3,F_4) of each cell get test 0111 (compatible with 1110), or 111 (compatible with 0110) which correspond to channel blocking. If any of F_1 channels is faulty, the "out of order" signal will emerge at the horizontal output of corresponding row.

It should be noted that during the cycle T_1 in the even cells of each row conjunction in X is tested with the test $(F_1\ F_2\ F_3\ F_4)^x$ = $(0111)^x$, i.e. with one of five tests serving to check the four-input "AND" gate: $(0111, 1011, 1110, 1111, 1101)^x$. At the expiration of cycle T_1 (and each successive cycle) all cells are fed with the signal A=1 (at B=C=D=0) that results in resetting the cell flip-flops into initial state, i.e. y_{TP}=0.

During cycle T_2 the test 1 checks channels F_1 of even cells in each row and test 2 -- of odd ones. Channels F_2, F_3, and F_4 are tested with the same tests in a similar manner during cycles T_3 through T_8. During the cycle T_9 the channels F_1, and F_2 of all cells are checked with the selfconjugate (in a pair of intersecting channels) test 1100 (Fig.9,δ); the rest of the channels (F_3,F_4) is fed with the auxiliary test 1101. During cycles T_{10}, and T_{11} the channels F_1 and F_2 are checked with the selfconjugate (in one channel) test 4 from Table 1 (Fig.9,δ). During cycles T_{12}, T_{13}, T_{14} the other pair of channels (F_3 and F_4) is checked with same tests 3 and 4.

There is no one period of conjugated tests for the last test (0101) in Table 1, since input signals B=D=1 block the channel, i.e. the output of the cell becomes independent of its input. Such being the case, a test is used which is conjugate (without a period) and intercompatible with the given one in the i-th channel, and has conjugation period in the j-th channel. Thus, the "out of order" signal from the i-th channel will appear at the external output of the j-th channel. When, during the cycle T_{15}, test 0101 is fed into channels F_1, of odd cells in odd rows (Fig.9,2), the channels F_2 of even cells in all rows are fed with tests 1110 and 0110 (1-st and 2-nd in Table 1) which are interconjugate (in any channel, including F_2). Test 1110 is conjugate and intercompatible in the channel F_1 with the given test 0101. The odd cells in even rows are fed with any test which is intercompatable with the test 0110 in the channel. The "out of order" signal for the channel F_1 (0 instead of 1) will appear at the external output of the channel F_2. Channels $F_1 \div F_4$ of all other cells are checked in a similar manner with test 0101 during cycles T_{16} through T_{30}.

The checking time may be reduced if the array channels are tested with tests 1, 2, and 5, as shown in Fig.9,2. During a single cycle the test 0101 is fed into the channels F_1, of odd cells in odd rows, and the channels F_2 of even cells in odd and even rows are fed with interconjugate tests 1110 and 0110. Then the whole of the array, irrespective of its size, may be tested during 22 clock cycles. It is clear that at the same time all X-generating cell networks are tested.

It stands to reason that the checking time N' does not change if inputs B,C, and D_1 through D_4 of each cell are not external (as it was supposed before), but are connected to coordinate control buses as shown in Fig.10 for one row.

Summary

The uniform array checking and fault diagnosis (and consequently checking and diagnosis of any finite automaton (sequential function) realized by the array) are drastically simplified as compared with checking of non-uniform circuits, since they are reduced to checking and diagnosis of identical cells having regular interconnections. The number of checking and diagnostic procedures by the methods discussed above and time of array testing (number of working cycles) are nearly or completely independent of the number of cells (array size), i.e. of complexity of the function realised by the array.

References

1. Прангишвили И.В. и др. "Микроэлектроника и однородные структуры для построения логических и вычислительных устройств","Наука",Москва,1966.
(I.V.Pranghishvily et al. "Microelectronics and Uniform Arrays for Logic and Computation", "Nauka" Publishers, Moscow, 1966. (In Russian).
2. Чегис И.А.,Яблонский С.В. "Логические способы контроля электрических схем",Труды математического института им. Стеклова,№ 51,1958.
(I.A.Cheghis, S.V.Yablonsky, "Logical Checking of Electrical Networks". Trudy Matem.Inst.im. Steklova, No.51, 1958). (In Russian).
3. Карибский В.В. и др."Техническая диагностика комбинационных устройств" в сб."Абстрактная и структурная теория релейных устройств","Наука",1966.
(V.V.Karibsky et al. in "Abstract and Structural Theory of Relay Devices", "Nauka" Publishers, 1966). (In Russian).
4. R.H.Canaday. "Two-Dimensional Iterative Logic". Proc.of the 1965 FJCC, AFIPS, Vol.27, Part I.

29.2

ON DESIGN OF TESTS FOR DIGITAL DEVICES WITH DELAYS

Volkov A.F., Vasil'ev I.N.,
Vedeshenkov V.A., Petrov V.A.

Institute of Automatics and Telemechanics
(Technological Cybernetics), Moscow

U S S R

INTRODUCTION

A direct application of the conventional technique of tests construction[1] by tables of faults function leads to labor consuming cumbersome methods[2]. However, when the number and nature of faults that exist simultaneously and the action of contactless elements is known to be unidirectional more effective modular fechniques can be used to synthetize unitary tests.

We intend to discuss development and modification of tests for sequential units without memory elements such as flip-flops and feedbacks. These were chosen for analysis because:

a) a major portion of a control unit circuits satisfy the requirements of such a model;

b) the transitional nature of the model analysed (from combinational units to units with memory elements) makes it possible to resort to tests construction techniques used for combinational elements [3,4] and allow for specific features of sequential units;

c) if memory elements are included, the algorithm will have to be altered rather than invalidated.

The functioning of an element without fixation of the internal state can be described by the equation

$$W_t = f(u_1, u_2, \ldots, u_m)_{t-\tau}, \qquad (1)$$

where W is a variable at the output of the element;

$u_1,\dots,u_m$ are the corresponding variables at the inputs of the element;

τ a delay between changes in $u_1,\dots,u_m$ and W.

Any possible fault will change the logic or time response of the element and can only be detected if there is such a set of variables $u_1^o,\dots,u_m^o$ and such a time instant t^o for which the test function turns to unity

$$P_{t-\tau} = W(u_1^o,\dots,u_m^o)_{t-\tau} \oplus W^*(u_1^o,\dots,u_m^o)_{t-\tau} = 1, \quad (2)$$

where $\oplus$ is a symbol of addition for modulo 2,

W^* is an element operation equation for a given fault. The tests of separate elements are best developed experimentally by introducing faults into the elements and obtaining test tables.

The element checked is assumed to be connected to an output terminal of the unit through a group of elements whose operation is described as

$$y_t = F(z_{1,t-\tau_1},\dots,W_{t-\tau},\dots,z_{n,t-\tau_n})_{t-\tau_0}. \quad (3)$$

Then the function of the test for detection of faults in this element for output y can be represented as an intersection of two functions:

$$R_{t-\tau_0} = P_{t-\tau-\tau_0} \cdot \varphi^y_{w,t-\tau_0}. \quad (4)$$

The function of the test $P_{t-\tau-\tau_0}$ from (2) defines the conditions for detection of the fault when the output of the faulty element is being checked.

The function $\varphi^y_{w,t-\tau_0}$ determines the conditions for transfer of the control point from the output W of the element to the output y which is linked to the output W directly or through a number of intermediate elements that are connected arbitrarily. This is the function of the output y sensitivity to changes in the variable W^4. Formally the sensitivity function is detemined as

$$\varphi^{y}_{w,t-\tau_0} = F(z_{1,t-\tau_1},\ldots,W^{0}_{t-\tau},\ldots,z_{n,t-\tau_n})_{t-\tau_0} \oplus F(z_{1,t-\tau_1},\ldots,\overline{W^{0}_{t-\tau}},\ldots,z_{n,t-\tau_n})_{t-\tau_0}, \tag{5}$$

where

$$W^{0}_{t-\tau} = W(u^{0}_{1},\ldots,u^{0}_{m})_{t_0-\tau}.$$

By equation (5) one can find the sensitivity functions for elements of any type. E.g. for elements OR (NOR), AND (NAND), the sensitivity function for the i input will be (for n input elements)

$$\varphi^{OR}_{i} = \varphi^{NOR}_{i} = \overline{x_1}\cdot\overline{x_2}\cdots\overline{x_{i-1}}\cdot\overline{x_{i+1}}\cdots\overline{x_n}, \tag{6}$$

$$\varphi^{AND}_{i} = \varphi^{NAND}_{i} = x_1\cdot x_2\cdots x_{i-1}\cdot x_{i+1}\cdots x_n.$$

For a follower, inverter, the delay line $\varphi \equiv 1$.

As follows from the formula a solution to the equation enables to find such values of variables at other inputs of the element (circuit) at which the values of y depend unambiguously on the values of W:

$$y_t = W_{t-\tau-\tau_0}$$

or $$y_t = \overline{W_{t-\tau-\tau_0}}.$$

We will say that there is a sensitive path between the terminals W and y if changes in the value of W leads necessarily to changes in the value of y. Eq. (5) determines the conditions for the existance of a sensitive path between W and y. A solution to Eq. (5) would yield these conditions, but this may appear too cumbersome when the size of the circuit increases.

Therefore we undertook to analyze the relation between the structure of the device which implements equation (3) and properties of the sensitivity function. That analysis has shown that the sensitivity function of a sequence of ele - ments connected in series is the product of the sensitivity functions of all the elements in the sequence. Therefore, the conditions for existance of a sensitive path through a sequence of elements connected in series can be achieved by a consistent unification of conditions for existance of separate parts of the path which are associated with each element of the succession.

If there are several paths for a signal from the terminal W to propagate to the terminal y along, the sensitivity path chosen may prove impossible to realize. Therefore an essential part of construction of a unitary test suggested here is an analysis of conditions for realizability of the sensitive path chosen.

11. A modular technique for development of a unitary test in case of sequential device with delay elements

The algorithm suggested employs local characteris - tics of separate elements making up a sequential unit: tests tables, operational equations, magnitudes of delay, sensitivity functions, tables of links to neighbouring elements through inputs and outputs. Its steps are:

1. Ranking of the circuit.
2. Making up a primary tests table.
3. Finding the sensitive path.
4. Finding and analysis of parallel paths.
5. "Widening" of the sensitive path.
6. Making input variables consistent.
7. Making the times when values of variables change consistent.
8. Minimization of the test total duration.

Each step will be discussed for the case of a unit whose block diagramme is shown in the Figure. The unit is made of conjunction ($\wedge_i$) disjunction ($\vee_j$), inverters (Jn_k) and delay, whose magnitude (τ) is taken arbitrarily and shown in relative units.

1. The first part of the algorithm is to denote the

output components by numbers in the increasing order of the elements response when input variables are fed si - multaneously to terminals 1, 2, ..., n .

The next numbers $n+1$, . . . , $n+m$, where m is the number of elements in the unit, will be assigned to the output of that element whose all inputs have already been numbered and for which the maximal delay of response is the lowest (e.g. for terminal 9 $\Sigma \tau_i = 6$, for terminal 10 $\Sigma \tau_i = 9 > 6.$).

2. The second part of the algorithm is to memorize the initial information on the device analysed in a form convenient for construction of a test by a digital computer. This information makes the primary tests table which con - tains test tables of all elements placed in accordance with their position in the circuit. The last column contains data on the type of the logival element and the magnitude of delay in the relative units (T_0). The table is filled out in the increasing order of numbers of the element output terminals. A portion of the table for the unit shown in Figure is represented in Table 1. The tests reveal all kinds of logical faults, i.e. physical faults due to which the outputs of the element look as though one or several its input channels are constantly receiving either 0 or 1.

TABLE 1

No. of row \ No. of terminal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Type of element (τ_i)
1				0				1								$IN_1(4)$
2				I				0								
3			0					0	0							
4			0					1	1							$OR_2(2)$
5			1					0	1							
6						0		0		0						
7						0		1		1						$OR_3(5)$
8						1		0		1						

No. of row \ No. of terminal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Type of element (τ_i)
9					1				1		1					
10					1				0		0					AND_1(1)
11					0				1		0					
12							1			1	1		1			
13							1			1	0		0			
14							1			0	1		0			AND_2(2)
15							0			1	1		0			

An analysis of test tables for elements AND (NAND), OR (NOR) of Table 1 shows that each test set can be written as

$$T_j = q_i \cdot \varphi_i^{\,j},$$

where q_i is the value of the input y_i necessary to find logical faults of that input; $\varphi_i^{\,j}$ is the sensitivity function of the output of the j element to changes of the variable at the i input.

3. The third part of the algorithm enables contruction of a possible sensitive paths for the next row of a primary table. This construction as well as results of all the following steps will be shown in the secondary test table, which apart from the $(n+m)$ columns denoted by the numbers of the circuit components include a column of time units numbers which acts as a time scale of the test as well as a column of notes containing the numbers of rows from the primary table.

Let us take from the primary table the next row of the i element and rewrite it in the free row numbered a_i in the secondary table; the value of the output from column y_i will be written τ_i rows below. To construct a part of the sensitive path through an element whose input is connected to the terminal y_i we will find in the primary table a row which in the column y_i contains a value of q_i equal to the value of y_i in the row chosen while the content of other columns coincides with the sensitivity function for the input y_i. The content of row chosen is rewritten in the row numbered $(a_i + \tau_i)$ the secondary table while shifting the value of the output y_j by τ_j rows downwards. If y_j is not an output terminal of the unit, then we proceed to construct the sensitive path by finding in the primary table the sensitivity function of the element which follows y_j.

Table 2a shows how a version of the sensitive path is constructed for the second row of Table 1.

TABLE 2

No. of time unit \ No. of terminal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Note
1				1												2
2																
3																(a)
4																
5						0	0									6
6																
7																
8																
9																
10						1				0	1					14
11																
12												$\underline{0}$				
1			$\underline{1}$													2
2			1^x													(b)
3																
4		0^x						1^x								
5		$\underline{\underline{0}}$				$\underline{0}$		$\underline{0}$								6 3
6		0^x		1^x				0^x	1^x							
7				$\underline{\underline{1}}$			1^x			$\underline{0}$	1^x	1^x				10
8				1^x			$\underline{\underline{1}}$			0^x	$\underline{\underline{1}}$	$\underline{\underline{0}}$				
9							1^x				1^x	0^x	1^x			
10							$\underline{1}$				$\underline{0}$	$\underline{0}$	$\underline{\underline{0}}$			14
11													0^x			
12													$\underline{0}$			

4. Let us say that between terminals i and j of the circuit there are loops (groups of elements connected in parallel) if one can construct two or more paths along which the signal propagates from i to j and differing by at least one elements. Such terminals in the circuit of Figure are 8 and 13, 9 and 15. A linking element will be the one whose output terminal coincides with the final point of the loop. For each branch of a loop one can determine:

a) whether the number of signal b_i inversions is even (the phase of signal propagation);

b) total delay time of signal T_i, the index 1 relating to that branch for which a sensitive path was constructed during the third step of the algorithm.

If $b_1 = b_2$, the two branches of the loop will be termed synphase, in the opposite case, counterphase. After a loop has been found with a sensitive path passing through a branch, the difference $(T_1 - T_2)$ of the summed signal delays has to be computed for the branches. If the difference is zero, the phase relations of the signals transmitted $(b_1 - b_2)$ has to be found. If the branches of the loop are in phase we have to find the function $\varphi_{1,2}$ of the link element sensitivity to a simultaneous synphase change of its two input variables. If $\varphi_{1,2} \neq 0$ the sensitive path constructed has to be adjusted. For linking elements of the types AND (NAND), OR (NOR) this reduces to the value the variable obrained at the output of the first branch being rewritten to the column which corresponds to the variable at the output of the second branch.

If the branches are counterphase, or $\varphi_{1,2} \equiv 0$, it is necessary to prevent the signal from transmission through the second branch. For this purpose each branch should contain fixation elements or elements that make it possible to fix at the output of the giwen branch those values which are determined by the sensitivity function of the linking element.

At $|T_1 - T_2| = \Delta T > 0$ the branches can be assumed non-shunting and the third step of the algorithm has to be repeated for the sensitivity path, crossing the second branch of the loop.

Table 2b differes from Table 2a in that records have been added that describe the sensitivity path through terminals 8, 9, 11, 13 with time $T_2 = 9 < T_1 = 11$. The records which describe the first and the second sensitive path are underlinded once or twice, respectively. Besides, 1 in the cell of the 11th column and 10th row is replaced by 0 taken from the 8th row to simplify the stage of consistency. (The numbers with asterisks will be explained below).

5. Changes in the quantities τ_i of delay elements caused by faults will make the correct value of the output variable appear at the terminal checked at time t which does not coincide with the desired time t_0. When the output variable $f(t)$ values at time $(t_0 - 1)$ are fixed and that value is compared with the quantity $f(t_0)$, we may check whether the magnitudes τ_i are correct for the delay elements on the given sensitive path.

A test which characterises the sensitive path constructed can be modified so that not only a qualitative reply (yes or no) to the question, whether the delay elements are in order or not, can be obtained, but also we will be able to quantitatively estimate the total change of delays. To do this we will "expand" the sensitive path.

Assume that the number of time units d is the maximal permissible deviation of the total delay. Then starting with the row numbered $(a_i + \tau_i)$ we will repeat the records which correspond to sensitivity functions of separate elements d times upwards and d times downwards; in the lower rows the columns will have the values of y_i^0 and in the upper lines the values of $\overline{y_i^0}$. The columns of input variables which are related to d rows above the a_i-th row will be left blank, while in the same columns of the lower rows we will write the values of variables from the a_i-th row.

In Table 2b this construction was complted for a sensitivity path with a shorter delay time at $d = 1$. (The values added at this stage are marked with asterisks (*)).

The values of vatiables which expand the sensitive path fix the time t_0 when the values of variables at the output of each delay elements change (the boundaries between y_i^0 and $\bar{y}_i^0$) and maintain the constant values of variables that describe the sensitive path during the interval (t_0-d, t_0+d) . As a result when τ_i of one of delay elements on the sensitive path changes by the magnitude $|\pm \delta\tau_i| < d$ the time t of the output value change will be shifted with respect to time t_0 to by the same value $\pm \delta\tau_i$. A comparison of the value $f(t_0+d)$ with the value which corresponds to the operative unit gives an answer to the question: "Do the changes in the delay magnitude exeed the permissible limits ?" If the values coincide, the answer is no, other wise, yes, they do.

6. The sensitive path which results from the previous parts of the algotithm depends on the values of input variable elements which form the path. Some of these variables are outputs of other elements and thus cannot be fed from outside. The sixth part of the algorithm is to make consistent the values of input variable unit which are essential to feed the appropriate values of variables to the input of elements which form the sensitive path. Since for most logical elements there are no unambiguous relations between input and output variables the process of specification can lead to inconsistent result and a trial of a certain number of versions will be required until a satisfactory result is obtained.

The values of all variables that are required for a sensitive path are made consistent starting from the element which corresponds to a filled colimn with the greatest number in the row with the greatest number of time units. After input variables have been made consistent for remaining elements of the same row, in the decreasing order of their numbers, one can proceed to the elements of the nearest row which describes the value of at least one variable with the number $(n+1, \dots, n+m)$. Thus at the stage of making variables consistent the secondary table is looked through from the right to the left and upwards until there are no values of variables with the numbers $h+1$, ..., $h+m$, which are not determined by the values of variables with the numbers 1, 2, ..., h.

To choose one possible version of making the variables consistent one can resort to a relatively simple technique at which the chosen input variables coincide with the values of the same variables in the nearest upper or lower rows of the secondary table. Table 3 represents the results of making the variables consistent for the data of Table 2b whose numbers of rows are given in the first column in parenthesis.

TABLE 3

No. of time unit \ No. of terminal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1				0											
2				0											
3				0											
4(1)				1											
5(2)				1			1								
6(3)	1			1			1								
7(4)	1		0				1								
8(5)	1		0		0		0								
9(6)	1		0		1		0		1						
10(7)			0		1		0		0	1	1				
11(8)					1				0	1	0				
12(9)									0	1	0		1		
13(10)										0	0		0		
14(11)													0		
15(12)													0		

7. Among the columns of the secondary table let us find those that contain 0 and 1 with a group of blanks between

them [x/] and fill out the half of cells closer to 0 with zeroes and the half closer to 1 with ones. To avoid contradictions, the blanks will be filled in this order:

a) fill out one of the intervals by the above rule; the next interval will be the one which is in the column with the greatest number;

b) make the input variables consistent.

8. The secondary table resulting from the making the values of input variables consistent and times when these are fed contains a great number of blank cells. Their content is of no importance for correct realization of the part of the test which is related to the sensitive path chosen . Therefore blank cells of a secondary table can fe filled for various purposes in particular to minimize the total length of the test when separate parts of the test are linked. Efficiency of various minimization techniques depends on the structure of a unit which feeds the sets of a test. Therefore we will not deal with minimization in this paper.

The above algorithm for construction of unitary tests for sequential digital devices yields tests which guarantee detection of faults that cause changes in logical or temporal characteristics of one element of the device at a time. An advantage of the algorithm is that construction of the test does not require a faults function table for the entire device. This results in a comparative simplicity and ease with which the algorithm is realized on computers.

REFERENCES

1. Чегис И.А., Яблонский С.В. Логические способы контроля работы электрических схем. Труды Математ.ин-та им.Стеклова В.А., т.51, 1958, стр. 270-359.

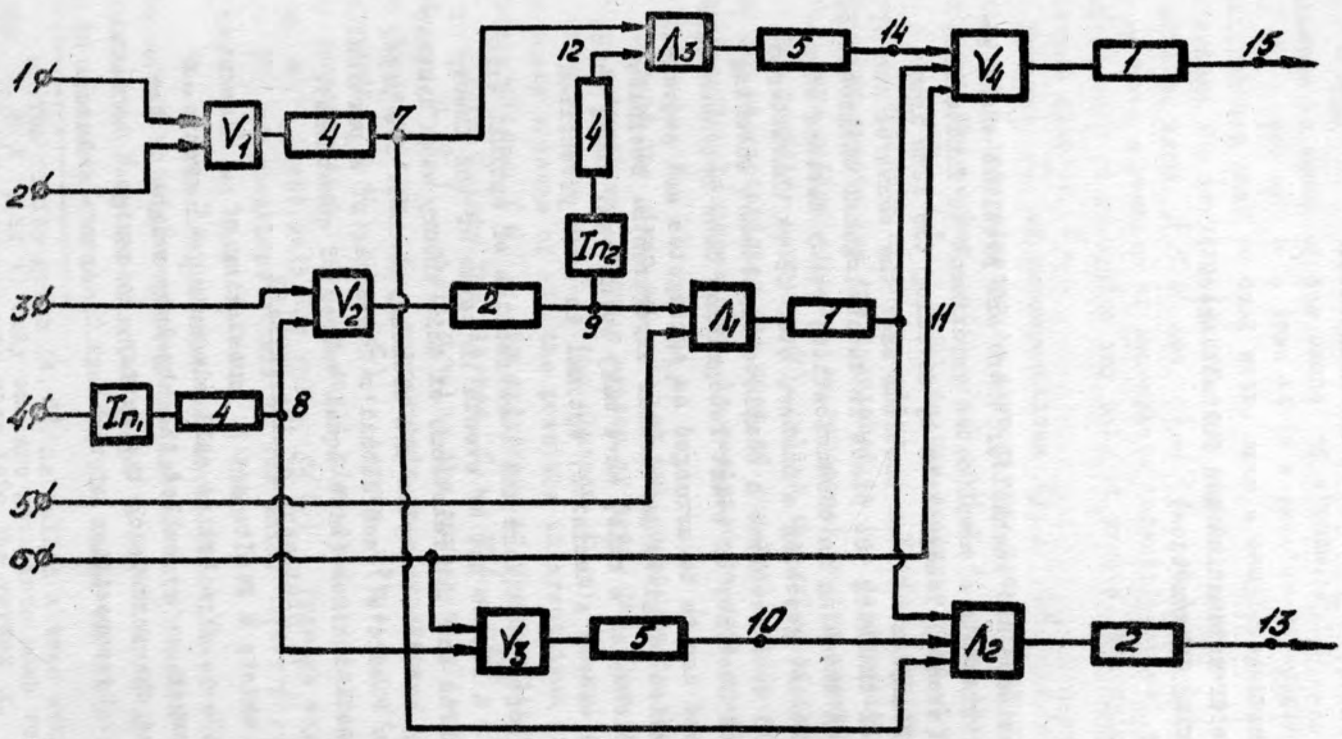
2. Gill A., Introduction to the theory of finite-state Machines, McGraw-Hill Book Company, 1963.

3. Roth J.P., Diagnosis of automata failures: a calculus and a method. IBM J. of Research and Development, July 1966, vol. 10, No. 4, pp. 278-291.

---

x/ This may be the case for a sensitive path containing a loop with counterphase branches.

4. Ведешенков В.А. "О построении таблицы тестов для обнаружения логических неисправностей бесконтактных комбинационных устройств". Автоматика и телемеханика, том XXIX, № 3, 1968.
5. Ведешенков В.А., Волков А.Ф. "Блочный метод построения единичного проверяющего теста для бесконтактных комбинационных устройств без обратных связей". Тезисы докладов научно-технического семинара МДНТП им.Ф.Э.Дзержинского "Проектирование радиоэлектронной аппаратуры с использованием вычислительной техники", июнь 1967, Москва.



Black-diagram of the sequential device analysed.

29.3

THE THEORY OF QUESTIONAIRES AND PROBLEMS OF TECHNOLOGICAL DIAGNOSTICS

P.P.Parkhomenko
Institute of Automation and Telemechanics
(Engineering Cybernetics)
Moscow
USSR

In science and technology there are problems whose solution represents a complicated experiment in finite sets elements (events) recognition by dividing the sets into classes. Examples of such problems are: the construction of programs of checking the plant in technological diagnostics, information encoding in communication, relay devices structure synthesis, guessing a number, etc. It is characteristic that in such problems a complete experiment ensuring the desired accuracy of event recognition does not necessarily exist and has to be arranged as aggregates and sequences of realization of given particular experiments. Besides, these problems, as a rule, have many solutions, among which one should choose a decision, optimal in some prescribed sense.

This paper considers the construction of optimal experiments for a finite set of events in terms of the theory of questionnaires and the relations of this theory with dynamic programming and the branches and boundary method. The paper is based on Picard's [1] and Dubail's [2] studies of a particular kind of questionnaires (with equal values of questions).

1. Statement of the Problem

There exists a finite set E consisting of elements y_i, $i = 1,2,\ldots,N$. Let us call elements $y \in E$ events. A positive number $w(y)$ called the absolute weight of the event y and characterizing the latter, is assigned to each event $y \in E$. Decomposition of the set E into λ classes of

E_μ, $1 \leqslant |E_\mu| < N$, $\mu = 1,2,..,\lambda$ is recorded. If the number of classes is equal to the number of events, $\lambda = N$, and $|E_\mu| = 1$ for all μ, then it is a problem of complete identification that we deal with, when a complete experiment must ensure the recognition of each single event among all the others. When $\lambda < N$ and $|E_\mu| > 1$ at least for one μ, we have a problem of incomplete indentification, which requires only to recognize any pair of events, belonging to different classes E_μ.

The set T_1 of decompositions t_j, $j = 1,2,\dots,|T_1|$ of the set E into classes is also given. Let us call $t \in T_1$ questions. The number $a(t)$, $2 \leqslant a(t) \leqslant N$ of the classes $E_{\gamma(t)}$, $\gamma(t) = 1,2,\dots,a(t)$ in the decomposition $t \in T_1$ is called the base of the question . Features by which classes of events are distinguished in the decomposition $t \in T_1$ are called answers (or outcomes) to the question t.

The positive number $c(t)$, called the value of the question and characterizing the cost of the realization of the corresponding decomposition, is assigned to each question $t \in T_1$.

Expand the given set T_1 of the questions in the following natural way. Form a system of sets, the elements $\mathcal{E}$ of which are the set E, the classes $E_{\gamma(t)}$ as well as non-empty classes of all the possible intersections of the decompositions $t \in T_1$. An aggregate of decomposition τ_j of the sets $\mathcal{E}$ into the classes $\mathcal{E}_{\gamma(\tau_j)} = E_{\gamma(t_j)} \cap \mathcal{E}$ $\gamma(\tau_j) = 1,2,\dots, a(\tau_j)$, where $a(\tau_j)$ is the number of classes in the decomposition τ_j corresponds to a question $t_j \in T$ Apparently, $a(\tau_j) \leqslant a(t_j)$. Assume $c(\tau_j) = c(t_j)$. The decomposition τ_j we shall also call a question on the base $a(\tau_j)$ and with the value $c(\tau_j)$. Form an expanded set T of questions of all τ for which $a(\tau) \geqslant 1$. Since with $\mathcal{E} = E$ we have $\tau_j = t_j$, then $T_1 \subseteq T$.

The complex experiment in identification of events of the set E by decomposing the latter into classes can be represented [1] by an oriented graph $G = (Q \cup E, \Gamma)$ of the type of a tree with the root x_0 [x]). The inner vertices and the

---

x) The finite graph (X, Γ) is called a tree with the root $x_0 \in X$, if 1) only one arc goes into each vertex $\neq x_0$; 2) not a single arc goes into the vertex x_0; 3) the graph (X, Γ) does not contain any loops.

root of the graph form a set Q of questions; the arcs outgoing from the vertex $x \in Q$ are called the outcomes of the question x , while their number $a(x)$, $1 \leqslant a(x) \leqslant N$ is called the base of the question x ; the value $c(x) > 0$ is assigned to each question x . In the problems of the complete identification the events $y \in E$ are compared with and the weights $w(y)$ are assigned to the finite vertices of the graph G . In cases of incomplete identification of the events of the set E the finite vertices of the graph G are compared with v, $1 \leqslant v < N$, of the subsets $E_k \subseteq E_\mu$, $k = 1,2,\dots, v$, when $\bigcup_{k=1}^{v} E_k = E$, and the weights $w(E_k) = \sum_{y \in E_k} w(y)$ are assigned. The graph G is called a questionnaire for E [x)].

Concerning the vertex $z \in Z = Q \cup E$ of the graph G we shall distinguish the set Γz of its followers, the set $\hat{\Gamma} z \setminus z = \Gamma z \cup \Gamma(\Gamma z) \cup \dots$ of its descendants the set $\Gamma^{-1} z$ of its forerunners and the set of its ancestors. Instead of the absolute weights $w(y)$ of the events it is more convenient to consider their relative weights $p(y) = \frac{w(y)}{w}$ where $w = \sum_{y \in E} w(y)$, further termed weights. The quantity $p(x) = \sum_{y \in \hat{\Gamma} x \cap E} p(y)$ is called the weight of the question $x \in Q$. The sum $C(x_0, z) = \sum_{x \in \hat{\Gamma}^{-1} z \setminus z} c(x)$ is called the value of the path from x_0 to $z \in Z$. The value of the path from x_0 to $y \in E$ characterizes the costs of identification of the event y .

All the practically interesting characteristics of the experiments, described by questionnaires, can be tackled by these questionnaires. One of the general enough and at the same time practically useful features of a questionnaire is [3] the value of scanning determined by the expressions

$$C(x_0, E) = \sum_{y_i \in E} c(x_0, y_i) \cdot p(y_i) = \sum_{x_j \in Q} c(x_j) \cdot p(x_j) \qquad (1)$$

corresponding to the mean weighed costs of identification of events in the entire questionnaire. For the questionnaires with equal values of questions one may obtain [1] from (1) the length of the scanning, corresponding, for example, to the mean lenght of a coding combination in Shannon-Fano codes or to

---

x) Further, if the reverse is not specified, the questionnaires for complete identification problem will be considered.

the average number of operations in sorting problems. If the value of a question is the cost (time) of realization of a separate checking, and the weight of an event is the probability of the operable or inoperable state of the plant then eq. (1) gives us the mean cost (mean time) of determining the states in the conventional sequential checking program.

With the same sets E and T, one can, generally speaking, construct various questionnaires for E, differing in aggregate and in succession of question realization and having different values of scanning. The questionnaire for E with the minimum value of scanning will be called optimal. We shall discuss the features and the methods of constructing optimal questionnaires in this report.

2. Optimal Questionnaires

By transforming the given questionnaire G for E, invariantly relative to the number q_m of its questions with the base a_m, $m \in M$ and the number q_ℓ of its questions with the value c_ℓ, $\ell \in L$ (M, L being some numerical sets), it has been revealed[3] that an optimal questionnaire, allowing to identify N events by means of $\sum_{m\in M} q_m = \sum_{\ell \in L} q_\ell$ questions, having bases a_m and costs c_ℓ is a tree with a root x_0, where $N = \sum_{m \in M} q_m (a_m - 1) + 1$ and such that its vertices, arranged in a non-decreasing order relative to their ranks, have weights assigned to them in non-increasing order and the values of paths leading to them in a non-decreasing order, the questions being arranged in a non-increasing order relative to their costs; among all the vertices having the same value of the paths leading to them as the value of the followers of the question on the base a, there is not a single vertex, whose weight exceeds the sum of the weights of all the other a vertices with the same value of the paths leading to them.

Assume that to construct a questionnaire G for E there must be and there are q_m questions with the base a_m, $m \in M$, among which q_ℓ questions are of the value c_ℓ, $\ell \in L$. Then the following algorithm (let us call it A1) of constructing an optimal questionnaire results from the above definition.

Make list 1 of weights $p(y_i)$, $i = 1,2,\dots,N$, arranged in a non-decreasing order, and list 2 of pairs (a_j; c_j) of question bases a_j, $j = 1,2,\dots,\sum_{m\in M} q_m = |Q|$, arranged in a non-decreasing order and of question values c_j, $j = 1,2,\dots,\sum_{\ell\in L} q_\ell = |Q|$ arranged in a non-increasing order. Eliminate the first pair (a_j, c_j) with the smallest base from list 2 and refer it to the question x_j. Eliminate a_j first (smallest) weights from list 1 and assign them to a_j followers of the question x_j. Determine the weight $p(x_j)$ of the question x_j and include it into list 1, and preserve the non-decreasing order of weights there. Return to list 2 and take its first not-eliminated pair (a_j, c_j), etc. Repeat the described operations until the last pair ($a_{|Q|}$, $c_{|Q|}$) with the largest base and least value of a question is eliminated; the last $a_{|Q|}$ weights will be eliminated out of list 1.

The described procedure is simple, almost does not require any search through all versions, ensures that the value of scanning of the questionnaire obtained will not exceed the value of scanning of any questionnaire for E, containing q_m questions with the base a_m and q_ℓ questions of the value c_ℓ.

A questionnaire, in which each question from Q is a question from the given set T of questions, i.e. for each question $x \in Q$ a question $t \in T$ can be found, such that $a(x)=a(t)$, $c(x) = c(t)$ and $\hat{\Gamma} z \cap E = \mathcal{E}_{\gamma(t)}$ for each $z \in \Gamma x$, or briefly, a questionnaire in which $Q \subseteq T$ will be called a realizable questionnaire. It is clear that the algorithm A1 does not always lead to an optimal realizable questionnaire. The condition $Q \subseteq T$ is always met for the problems, in which the set T contains all the possible decompositions of the set E into $a(t) = a =$ const classes, the values of all decompositions being the same, $c(t) = c =$ const. The problems of encoding and sorting mentioned above are examples of such problems. It is not difficult to show that the algorithm of constructing optimal questionnaires with equal values of questions [1], representing a generalization of the well-known [4] algorithm of constructing redundants codes, follows from the algorithm of constructing optimal questionnaires with unequal bases and values of questions.

The construction of optimal realizable questionnaires with restictions existing for the given set T of questions is discussed below. We shall note here, that if from the given set of questions we choose $\sum_{m\in M} q_m$ questions with as large bases as possible and $\sum_{\ell \in L} q_\ell$ questions with as low values a_ℓ as possible, then the value of scanning of the optimal (not necessarily realizable) questionnaire, constructed according to the algorithm described, represents the lower boundary of the value of scanning the questionnaires for E. This lower boundary is comparatively easy to calculate in the process of implementing the algorithm A1 operation and is achievable, when the condition $Q \subseteq T$ is met for the constructed optimal questionnaire. When the lower boundary is known it is possible to estimate the quality of questionnaires obtained by "approximated" methods and to construct optimal realizable questionnaires by the branches and boundaries method.

3. Recurrent Calculation of the Value of Scanning

Let us denote questionnaire vertices by the two indices - $[z, s]$ where z is the rank of a vertex and s determines its location among other vertices of the rank z. Let $G = (Z, \Gamma)$ be a questionnaire for E and $z_{z,s}$ - its vertex. The subquestionnaire with the root $z_{z,s}$ of the questionnaire G is a graph, $G_{z,s} = (Z_{z,s}, \Gamma_{z,s})$ where $Z_{z,s} = \hat{\Gamma} z_{z,s} \subseteq Z$ and the reflection $\Gamma_{z,s}$ is determined in the following way: $\Gamma_{z,s} z = \Gamma z \cap Z_{z,s}$. The subquestionnaire $G_{z,s}$ is a questionnaire for $E_{z,s} = \hat{\Gamma} z_{z,s} \cap E$.

If the vertex $z_{w,t} \in Z_{z,s}$, then its weight in the questionnaire G is $p(w,t)$ while its relative weight in the questionnaire $G_{z,s}$ is equal to $p_{z,s}(w,t) = \frac{p(w,t)}{p(z,s)}$. Assume that besides the subquestionnaire $G_{z,s}$, there is a questionnaire $G'_{z,s}$ for $E_{z,s}$. The operation of substituting subquestionnaires in the questionnaire G for E determines a new questionnaire, G'', for E, obtained from by replacing the subquestionnaire $G_{z,s}$ by the questionnaire $G'_{z,s}$ with further re-calculation of the vertices weights of the latter by multiplying them by $p(z,s)$.

By eq. (1) for the value of the scanning $G_{z,s}$ we have:

$$C_{z,s} = C(z_{z,s}, E_{z,s}) = \sum_{y_i \in E_{z,s}} c(z_{z,s}, y_i) \cdot p_{z,s}(y_i) \qquad (2)$$

Instead of (2) we may write the following:

$$C_{z,s} = \sum_{y_i \in E_{z,s}} [c(z_{z,s}, y_i) - c(z_{z,s})] p_{z,s}(y_i) + \sum_{y_i \in E_{z,s}} c(z_{z,s}) \cdot p_{z,s}(y_i) \qquad (3)$$

Noting that $E_{z,s} = \bigcup_{n=1}^{a(z_{z,s})} E_{z+1,n}$;

$$p_{z,s}(y) = p_{z+1,n}(y) \cdot p_{z,s}(z+1, n); \qquad c(z_{z+1,n}, y) = c(z_{z,s}, y) - c(z_{z,s});$$

$$\sum_{y \in E_{z+1,n}} p(y) = p(z+1, n); \qquad \sum_{z_{z+1,n} \in \Gamma z_{z,s}} p(z+1, n) = p(z, s),$$

and taking into account (2), we obtain the following from eq. (3):

$$C_{z,s} = c(z_{z,s}) + \sum_{z_{z+1,n} \in \Gamma z_{z,s}} p_{z,s}(z+1, n) \cdot C_{z+1,n} \qquad (4)$$

Therefore, the value of scanning $C_{z,s}$ of the subquestionnaire $G_{z,s}$ of the rank z is equal to the sum of its root's $z_{z,s}$ value $c(z_{z,s})$ and the sum of the values of the scanning $C_{z+1,n}$ of the subquestionnaires $G_{z+1,n}$ of the rank $z + 1$, whose roots $z_{z+1,n}$ are the followers of the root $z_{z,s}$.

If $z_{z,s} = y \in E$, then $\Gamma z_{z,s} = \phi$ and it follows from eq. (4) that

$$C_{z,s} = c(z_{z,s}) = c(y) = 0 \qquad (5)$$

Formulae (5) and (4) allow to calculate recurrently the value of scanning every subquestionnaire $G_{z,s}$ of the questionnaire G, including the value of the scanning G itself:

$$C = C(x_0, E) = c(x_0) + \sum_{z_{1,n} \in \Gamma x_0} p(1, n) \cdot C_{1,n} \qquad (6)$$

Let us prove, following Ref. 2, the following statement: an optimal questionnaire consists of optimal subquestionnaires.

Let G be an optimal questionnaire. Suppose that in G there is a non-optimal subquestionnaire $G_{z,s}$ with the value of scanning $C_{z,s}$. Then there exists an optimal questionnaire $G'_{z,s}$ for $E_{z,s}$ with the value of scanning $C'_{z,s}$ and

$$C'_{z,s} < C_{z,s} \tag{7}$$

Let be the minimum rank of non-optimal subquestionnaires in G . If $z = 0$ then it is the questionnaire G that is non-optimal, which is impossible according to our condition. Assume that for all $w < z$, $0 < z \leq z_M$ (z_M being the maximum rank of a questionnaire) all subquestionnaires $G_{w,s}$ in G are optimal. Take a subquestionnaire, $G_{z-1,t}$, with the root $z_{z-1,t} \in \Gamma^{-1} z_{z,s}$. This subquestionnaire is optimal and, according to eq. (4), its value of scanning is

$$C_{z-1,t} = c(z_{z-1,t}) + p_{z-1,t}(z,s) \cdot C_{z,s} + \sum_{z_{z,n} \in \Gamma z_{z-1,t} \setminus z_{z,s}} p_{z-1,t}(z,n) \cdot C_{z,n} \tag{8}$$

In the questionnaire $G_{z-1,t}$ substitute the subquestionnaire $G'_{z,s}$ for $G_{z,s}$. For the new questionnaire $G'_{z-1,t}$ obtained we have:

$$C'_{z-1,t} = c(z_{z-1,t}) + p_{z-1,t}(z,s) \cdot C'_{z,s} + \sum_{z_{z,n} \in \Gamma z_{z-1,t} \setminus z_{z,s}} p_{z-1,t}(z,n) \cdot C_{z,n} \tag{9}$$

Comparing eq. (8) and eq. (9) we obtain by virtue of eq. (7):

$$C'_{z-1,t} < C_{z-1,t} \tag{10}$$

which is incorrect since $G_{z-1,t}$ is optimal by definition. The statement has been thus proved and it is also true for optimal realizable questionnaires.

4. Optimal Realizable Questionnaires

Let $\bar{\mathcal{E}}$ be a set of all possible different non-empty subsets $\mathcal{E}_j$ of the set E , containing $m_j \leq N$ events. When $m_j \geq 2$ we shall call the given decomposition of the subset $\mathcal{E}_j$ into $a(t)$, $2 \leq a(t) \leq m_j$ classes $\mathcal{E}_{j,n}$, $n = 1,2,\dots, a(t)$ a question $t \in T$ feasible for $\mathcal{E}_j$. Denote the set of questions feasible for $\mathcal{E}_j$ by T_j . If $m_j = 1$, we assume $T_j = \emptyset$. Then $T = \bigcup_{\mathcal{E}_j \in \bar{\mathcal{E}}} T_j$ is a set of feasible questions, which is called a compatible set if for it there is at least one realizable questionnaire for E. Let us call the pair ($\mathcal{E}_j$, T_j) a situation and m_j elements in $\mathcal{E}_j$ - the order of a situation.

It has been mentioned above that if there are constraints for the set T of the given questions the algorithm A1 may give an optimal realizable questionnaire only by chance, when all the questions of at least one of all versions possible according to the algorithm turn out to be feasible. Let us state the general algorithm A2 for construction of optimal realizable questionnaires, based on the procedure of recurrent calculation of the value of scanning a questionnaire.

Let $\mathcal{G}_2$ be a set of all realizable questionnaires for E and for the compatible set T of feasible questions and $S\mathcal{G}_2$ - a set of all (realizable) subquestionnaires of the above questionnaires from $\mathcal{G}_2$.

Let us call the situation, to which the realizable subquestionnaire from $S\mathcal{G}_2$ corresponds, a possible situation. Let us call the questions of the realizable questionnaires from $\mathcal{G}_2$ possible feasible questions. Denote the set of feasible questions by T'.

Assume $(\mathcal{E}_j, T_j')$ to be a possible situation of the order m_j. If $(\mathcal{E}_j, T_j')$ is a possible situation of the first order, then the realizable subquestionnaire corresponding to it is a degenerated *) one and therefore optimal with the value of scanning

$$C_0(\mathcal{E}_j, T_j') = C_0(y_i, \phi) = 0, \quad j = 1, 2, \ldots, N \qquad (11)$$

Now assume that for every possible situation $(\mathcal{E}_k, T_k')$ of the order m_k, $1 \le m_k < m_j \le N$ there is a corresponding optimal realizable subquestionnaire already constructed. Take a possible situation, $(\mathcal{E}_j, T_j')$. The question t feasible for $\mathcal{E}_j$ with the base $a(t)$ is a decomposition, $\mathcal{E}_j$, into $a(t)$ non-intersecting and non-empty subsets $\mathcal{E}_{j,n}$, each of which has the number of a elements $m_{j,n} < m_j$. By our condition, an optimal realizable subquestionnaire has already been constructed for each of the possible situations $(\mathcal{E}_{j,n}, T_{j,n}')$. Then the subquestionnaire with the minimum value of scanning will be an optimal realizable subquestionnaire. This value is given by

*) We call questionnaires for one event degenerated [1].

$$C_o(\mathcal{E}_j, T_j') = \min_{t \in T_j'} \left[c(t) + \sum_{n=1}^{a(t)} p_n \, C_o(\mathcal{E}_{j,n}, T'_{j,n}) \right] \qquad (12)$$

where

$$p_n = \frac{\sum_{y_i \in \mathcal{E}_{j,n}} p(y_i)}{\sum_{y_k \in \mathcal{E}_j} p(y_k)}$$

It is easy to observe that recurrent relation (12) is a functional equation of dynamic programming with the operator of the minimum [5]. Thus, the procedure of constructing optimal realizable questionnaires by the algorithm A2 consists in examination of all possible situation of the order m_j, from $m_j = 1$ to $m_j = N$, and in constructing the corresponding optimal realizable subquestionnaires. The questionnaire corresponds to the possible situation (E, T').

The algorithm A2 allows to construct realizable questionnaires, optimal not only in terms of the minimum value of optimization; the algorithm makes it possible to obtain either one or all the decisions, it can be easily modified for for the solution of incomplete indentification problems. The drawback is the great number of calculating operations, determined in the worst case by $2^N \cdot |T_1'|$ * references to eq. (12). Let us turn to the algorithm A3, representing a procedure of the branches and boundaries method [6] and using the algorithm A1 for calculating the lower boundary of the value of scanning a questionnaire. In the version described the algorithm A2 constructs the desired questionnaire from its finite vertices towards the root x_0. We shall describe algorithm A3 in a version, where a questionnaire is being constructed, on the contrary, from its root x_0 towards its finite vertices.

Assume that there is a tree with the root x_0, realizing the questionnaire G for E. In G let us single out a subtree $G_2 \subseteq G$ with the root x_0, belonging to the set Q_2 of the inner vertices x_2, and with a set E_2, of finite vertices, $y_2 \in E_2$. Apart from the subtree G_2 the

---

*) In a number of cases it is possible to reduce the number of operations by determination of all possible situations in advance.

tree G contains subtree G_{y_2} with the roots $y_2 \in E_2$, realizing subquestionnaires G_{y_2} for E_{y_2}, corresponding to them (and including, perhaps, degenerated ones), while $\bigcup_{y_2 \in E_2} E_{y_2} = E$. The value of scanning the questionnaire G can be represented in the following way:

$$C(x_0, E) = \sum_{x_2 \in Q_2} p(x_2) \cdot c(x_2) + \sum_{y_2 \in E_2} p(y_2)\, C_{y_2} \qquad (13)$$

where $p(y_2) = \sum_{y \in E_{y_2}} p(y)$ and C_{y_2} is the value of scanning the subquestionnaire G_{y_2}.

Let us determine the initial possible situation ($\mathcal{E}_j = E$, $T_j' = T_1'$) of the order N. Any question $t \in T_1'$ may turn ou out to be the first question x_0 of the optimal realizable questionnaire sought for, i.e. there are $|T_1'|$ versions (branches of the tree of solutions). Let us determine possible situations ($\mathcal{E}_{j,n}$, $T_{j,n}'$) for each version, by the algorithm A1 construct optimal (non-realizable in the general case) subquestionnaires, corresponding to each of these situations and find the value of scanning the latter $c_0(\mathcal{E}_{j,n}, T_{j,n}')$ Then by eq. 13, determine the value of scanning the entire questionnaire for each version and for the continuation of the solution choose the version that gives:

$$C(x_0, E) = \min_{t \in T_1'} \left[c(t) + \sum_{n=1}^{a(t)} p_n \cdot C_0(\mathcal{E}_{j,n}, T_{j,n}') \right] \qquad (14)$$

At the next steps the operations described above are repeated conformably to the possible situations, chosen at the previous steps. If the versions of the given step give a higher value than any of the versions of the preceding steps, then it is chosen for the continuation of the solution. The first realizable questionnaire for E obtained in the course of solution will be the optimal one.

In the majority of cases the branches and boundaries method allows to reduce the volume of calculating operations as compared with the dynamic programming procedure and is as general as the latter.

5. On "Approximated" Method of Constructing Questionnaires

The set E of events can practically always be represented as a complete system of events, whose probabilities are equal to their weights. Hence, the conventional entropy of the rank τ is rather often used in the questionnaire construction procedures as a function of preference in choosing the solution at every step. This corresponds to maximization at every step of the function

$$H = \sum_{n \in \Gamma z_{\tau,s}} P_{\tau,s}(\tau+1, n) \cdot \log_{a_{\tau,s}} \frac{1}{P_{\tau,s}(\tau+1, n)} \qquad (15)$$

i.e. to the choice of such a decomposition of the subset $E_{\tau,s}$, which makes it possible to obtain the closest to each other weights of the questions $[\tau+1, n]$, which are the followers of the question $[\tau, s]$.

Another "approximated" procedure of questionnaires construction, also often used in practice, is a procedure, known as Shannon-Fano method for codes with variable length of encoding combinations. By this procedure the best decomposition at every step is such that allows to obtain the weights of the questions $[\tau+1, n]$, closest to each other, provided that the weights of the events are arranged in a non-decreasing (or non-increasing) order.

The investigation of the features of the questionnaires, optimal in terms of the minimum value of their scanning, as well as of the questionnaires, optimal from the point of view of information reveals that neither of the preference functions mentioned guarantees that optimal questionnaires will be obtained even in case [1,2] of questionnaires with equal values of questions. In the general case the principle of the conventional entropy maximization gives decisions, which are farther from being optimal, than Shannon-Fano method. This is explained by the fact that in the first case there is a more pronounced violation of the property, characteristic of optimal questionnaires, which is that among all the vertices with the value of paths leading to them equal to that of the followers with the base a must be none whose weight exceeds the sum of weights of other a vertices with the same value of the paths leading to their forerunners.

In conclusion let us point out that the theory of questionnaires enables us to study the characteristics of questionnaires optimal in terms of the minimum or maximum of other objective functions of optimization differing from those mentioned in the paper.

References

1. Picard C. Theorie des questionnaires, Gauthiers-Villars, Paris, 1965.

2. Dubail F. Algorithmes de questionnaires realisables optimaux au sens de differents criteres. These presente a la faculte des sciences de l'Universitō de Lyon, 1967.

3. Пархоменко П.П. Оптимальные вопросники с неравными ценами вопросов.ДАН СССР, I84, № I, I969 г.

4. Huffman K.E. A method for the construction of minimum-redundancy codes. Prec. IRE, vol. 9, 1952.

5. Беллман Р., Динамическое программирование, ИЛ, Москва, 1960 г.

6. Lawler E.L., Wood D.E., Branch-and-bound methods: a survey, Operat. Res., 14, No. 4, 1966.

29.4

EFFECT OF MONITORING PERIODICITY ON RELIABILITY OF RESTORABLE DEVICES

Garcavi A.L., Gogolevsky V.B.,
Grabovezky V.P.
Reliability Problem Scientific Council of the Academy of Sciences of the USSR.
Moscow, USSR

Considered in this report are some methods of determining reliability characteristics of restorable devices under check. The proposed methods are used to solve two groups of problems.

I

The first group of problems is related to the reliability characteristics of devices, operating episodically (by special command), though they are under current for a long period of time. To be determined in this case is an effect of the periodicity and the character of the monitoring tests on the device efficiency. To be found further is the probability of the event when the device will be sound at a random moment t and will show trouble-free operation during time period τ - $\mathcal{P}(t,\tau)$.

It is supposed, that the efficiency testing of these devices is of intermittent type, since a continuous monitoring is either impossible or unsuitable.

The following assumptions are taken: the time intervals between the failures of the device and the time required for restoration are distributed by the exponential law with parameters λ and μ corresponding to the above values respectively. The device is monitored instantly; the monitoring equipment is fully reliable.

Two cases are considered for solving the problem.

a) The monitoring is accomplished at random time intervals $t_{\kappa i}$, distributed by the exponential law with parameter α (Fig.1a). The device under check, switched on at time instant $t=0$, is in the failure-free state during a random time interval ξ_1 till the failure appears. The failure is detected only in a random time interval $t_{\alpha 1}$ at the instant, when the device is monitored. After detection of the failure the latter is eliminated for a random time t_{B1} ; the device is switched on, and the process is repeated all over again.

Thus the device under check can be in three states:

State 1 - the device is sound;

State 2 - the device is faulty and fails to be restored;

State 3 - the device is faulty and restorable.

The random operation process of the device under check and the monitoring device, shown in the fig.1a, can be described by a following system of equations:

$$P_1'(t) = -\lambda P_1(t) + \mu P_3(t)$$
$$P_2'(t) = -\alpha P_2(t) + \lambda P_1(t)$$
$$P_3'(t) = -\mu P_3(t) + \alpha P_2(t) \qquad (1)$$

where $\alpha = \frac{1}{T_K}$ - is monitoring intensity;

T_K - is mean time between the monitoring testings.

$P_1(t), P_2(t), P_3(t)$ are probabilities of the fact that the device under check will be in state 1,2 or 3 at time instant t .

The equation system (1) is solved under conditions that:

$$P_1(0) = 1 ,\ P_2(0) = P_3(0) = 0,$$

i.e. it is accepted that in the switching-on instant the device is always sound.

As result of the solution of the equation system (1) the expression for the probability $P_1(t)$, that is for the probability to find the device in the working order at any random time instant t will be represented as:

$$P_1(t) = \frac{\alpha\mu}{\alpha\mu + \alpha\lambda + \mu\lambda} +$$

$$+\frac{\lambda\left[\lambda-\alpha-\mu-\sqrt{(\alpha-\mu-\lambda)^2-4\mu\lambda}\,\right]}{(\alpha-\mu-\lambda)^2-4\mu\lambda-(\alpha+\mu+\lambda)\sqrt{(\alpha-\mu-\lambda)^2-4\mu\lambda}}\cdot e^{-Ct}+$$

$$+\frac{\lambda\left[\lambda-\alpha-\mu+\sqrt{(\alpha-\mu-\lambda)^2-4\mu\lambda}\,\right]}{(\alpha-\mu-\lambda)^2-4\mu\lambda+(\alpha+\mu+\lambda)\sqrt{(\alpha-\mu-\lambda)^2-4\mu\lambda}}\cdot e^{-Dt}, \quad (2)$$

where $$C=\frac{\alpha+\mu+\lambda-\sqrt{(\alpha-\mu-\lambda)^2-4\mu\lambda}}{2}$$

$$D=\frac{\alpha+\mu+\lambda+\sqrt{(\alpha-\mu-\lambda)^2-4\mu\lambda}}{2}$$

The probability to be found will be determined by the expression:

$$\mathcal{P}(t,\tau)=P_1(t)e^{-\lambda\tau} \quad (3)$$

The limiting cases for the probability $P_1(t)$ are:

1a. The permanent value $(t\to\infty)$

$$P_1(t)\to K'_{rk}=\frac{1}{1+\varkappa+\sigma},$$

where $\varkappa=\frac{\lambda}{\mu}$, $\sigma=\frac{\lambda}{\alpha}$

2a. No monitoring $(t_k=\infty,\ \alpha=0)$

$$P_1(t)=e^{-\lambda t}$$

3a. Continuous monitoring $(t_k=0,\ \alpha=\infty,\ t=\infty)$

$$K_{rk}=\frac{1}{1+\varkappa}$$

b) The device is monitored periodically in equal time intervals with duration h.

The random operation process of the device under check and the monitoring device for this case is shown in Fig.1b and can be described by the following system of the recurrent algebraic equations, based on the complete probability formula:

$$\begin{aligned}
P(t)&=P_n e^{-\lambda\tilde{t}}+\bar{P}_n z(\tilde{t})\\
P_n&=P_{n-1}e^{-\lambda h}+\bar{P}_{n-1}z(h)\\
\bar{P}_n&=P_{n-1}(1-e^{-\lambda h})+\bar{P}_{n-1}\bar{z}(h)\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
P_1&=e^{-\lambda h}\\
\bar{P}_1&=1-e^{-\lambda h}
\end{aligned} \quad (4)$$

The following symbols are used in these equations:

$P(t)$ - the probability of the fact that the device is in the working order at random time instant t ; it is accepted, that before the time instant t n monitoring tests have taken place;

$\tilde{t}$ - the time interval elapced from the instant of the last (n-th) monitoring test up to the present time instant t ;

P_i - the probability of the sound state of the device under check at the time instant of the i-th monitoring test;

$\bar{P}_i = 1 - P_i$ - the probability of the faulty state of the device under check at the time instant ih ;

$z(\xi)$ - the probability of the fact that the faulty device under check will be restored and will not fail again during the time interval $(0, \xi)$;

$\bar{z}(\xi) = 1 - z(\xi)$ - the probability of the fact that the faulty equipment will not be restored, or, when restored, will fail again during the time interval $(0, \xi)$.

The equation system is solved as follows:

$$P_n(t) = \frac{\mu[e^{-\lambda\tilde{t}}(1-e^{-\mu h}) - e^{-\mu\tilde{t}}(1-e^{-\lambda h})]}{} \pm$$

$$\pm \frac{(1-e^{-\lambda h})(\mu e^{-\mu\tilde{t}} - \lambda e^{-\lambda\tilde{t}})}{\mu(1-e^{-\mu h}) - \lambda(1-e^{-\lambda h})} \cdot \left(\frac{\lambda e^{-\lambda h} - \mu e^{-\mu h}}{\mu - \lambda}\right)^n \qquad (5)$$

The sign "+" is used at even n , the sign "-" is used at odd n . The probability to be found will be expressed by the formula:

$$\mathcal{P}(t,\tau) = P_n(t) e^{-\lambda\tau} \qquad (6)$$

The limiting case. At $n \to \infty$ and $\tilde{t} = 0$

$$P_n \to K_r' = \frac{\mu(e^{-\lambda h} - e^{-\mu h})}{\mu(1-e^{-\mu h}) - \lambda(1-e^{-\lambda h})}$$

The influence of the monitoring nature on the failure-free operation probability of the devices for the cases considerated above, is illustrated by the dependences $P_n(t)$ and $P_1(t)$ shown in Fig.2.

II

In the following group of problems are considerated the reliability characteristics of the devices, for which, under certain conditions, the operation with intervals is allowed. The intervals can be caused by necessity of performing monitoring tests and the restoration repairs. The principal requirement for such devices is the requirement of the necessity to process the preset volume of information for the preset time. The preset volume of information is divided into separate groups, which are processed in several stages. At the end of each stage the results of work at the given stage are checked by applying one of the known methods.

Then to be found are the following reliability characteristics of the device under check:

1. Probability $\mathcal{P}_z(v, t)$ of the processing information volume v for the time $t \geqslant v$ (the time excess) at z stages;

2. The expectation $M[T]$ of time spent for processing information volume v at z stages;

3. The optimal number z_0 of stages, into which information processing cycle v is to be divided to get the minimum value of $M[T]$.

The determination of these characteristics makes it possible to connect the device reliability with their efficiency.

The problems are solved if we assume that:

- the failure flow is of Poisson's form with parameter λ;

- the checking time τ_n and the checking and restoration repair time τ_b is taken constant and equal to d and h correspondingly, with the exception of the cases specially mentioned, when these values are taken as random ones with expectations d and h.

Further on are considered three cases of the solution of the problem .

a) The monitoring is accomplished at the end of each stage by using special fully reliable monitoring devices. In case of failure the results of all stages, preceding to the failure, get lost, because the intermediate storages are not present(Fig.3a).

Thus, the preset volume of information will be processed, provided the device operates successfully during z stages continuously for the time interval t .

Let us take $\tau_n = \tau_6 = d$. The stage number possible for time interval t is determined through the expression:

$$n = \left[\frac{t}{\frac{v}{z} + d} \right] ,$$

where $[x]$ is a whole part of x .

When we consider the particular stage as a test, and the trouble-free operation of the device during the stage as a event A with probability $p = e^{-\frac{\lambda v}{z}}$, a following problem arises: to find probability $P_z(n)$ of the event, that among n tests will be at least one series of z successive tests, where event A will occure. Let us consider the series of z of successive successful tests as series $\mathcal{S}_z$.

Series $\mathcal{S}_z$ is among n tests only in two incompatible cases:

1) series $\mathcal{S}_z$ is among the first ($n-1$) tests;

2) series $\mathcal{S}_z$ appeared for the first time at the n-th test.

Thus,

$$\mathcal{P}_z(n) = \mathcal{P}_z(n-1) + \Pi_z(n) , \tag{7}$$

where

$$\left.\begin{aligned} &\Pi_z(n) = 0 & &(n < z) \\ &\Pi_z(n) = p^z & &(n = z) \\ &\Pi_z(n) = [1 - \mathcal{P}_z(n-z-1)]\,q\,p^z & &q = 1-p \end{aligned}\right\} \tag{8}$$

Taking into account (8), with recurrent formula (7) successively get the following:

$$\left.\begin{aligned} &\mathcal{P}_z(v,t) = \mathcal{P}_z(z+k) = p^z(1+kq), \quad (k=0,1,\dots,z) \\ &\mathcal{P}_z(v,t) = \mathcal{P}_z(2z+i) = p^z\left\{1+q\left[i+z-ip^z-\frac{i(i-1)}{2}q\,p^z\right]\right\} \end{aligned}\right\} \tag{9}$$

etc. $(i = 1,2,\dots z)$

The dependence of the information processing probability $\mathcal{P}_\tau(v, t)$ on time t at various values τ is shown in fig.4.

For determination of mathematical expectation of information processing time (τ_n and τ_b are taken here as random values) the following random value is introduced:

$$X_k = x_k - x_{k-1} \quad (k = 1, 2, \dots),$$

where x_k is the instant of the completion of the device restorating after the k-th failure, or the instant of the completion of information processing (in the latter case $X_{k+1} = x_{k+2} = \dots = 0$)

Then the random time of the task execution is:

$$T = \sum_{k=1}^{\infty} X_k ,$$

and the mathematical expectation of the task execution time is:

$$M[T] = \sum_{k=1}^{\infty} M[X_k]$$

Using the known methods of finding the mathematical expectation, we easily get:

$$M[T] = \frac{(1-e^{-\lambda v})}{(1-e^{-\frac{\lambda v}{\tau}}) \cdot e^{-\lambda v}} \left[e^{-\frac{\lambda v}{\tau}} d + (1-e^{-\frac{\lambda v}{\tau}}) h + \frac{v}{\tau} \right] \qquad (10)$$

The nature of mathematical expectation $M[T]$ change depending on stage number τ is shown in fig.5. Optimal stage number τ_0 is derived from condition of finding a mathematical expectation minimum, which is determined through the formula (10):

$$\frac{\lambda v}{\sqrt{2\lambda d}} \leq \tau_0 \leq \frac{2\lambda v}{\sqrt{2\lambda d}} e^{\frac{1}{2}\sqrt{2\lambda d}} \qquad (11)$$

Should λd be small, then with a good accuracy can be accepted, that

$$\tau_0 \approx \frac{\lambda v}{\sqrt{2\lambda d}} \qquad (12)$$

b) The results of the information processing at each stage are checked by means of repetition of processing at the stage till two coincident results appear (not necessarily in succession).

The coincident results are considered as correct ones. Special monitoring devices are not used. In case of a failure at the stage the information processing results of only this atage depreciate (availability of information storages is supposed). The organization of each test - data output, comparison and restoration, if necessary - takes time, equal to d (Fig.3b).

As shown in fig.3b, the whole information processing cycle consists of z intervals. At each interval the calculations at one stage are made a random number of times (up to the appearance of two coincident results).

Successful processing of information volume v for time interval t with the volume of information divided into z stages and with the results to be checked several times, is equivalent to the event that among n checks not more than $n-2z$ checks will be unsuccessful, because at each of z intervals it is necessary to have two successful checks, i.e. only $2z$ (or more) successful checks. Using the binomial law, we will get a probability $\mathcal{P}_z(v,t)$ of the following form:

$$\mathcal{P}_z(v,t) = \sum_{\kappa=2z}^{n} C_n^{\kappa} e^{-\frac{\kappa\lambda v}{z}} \left(1-e^{-\frac{\lambda v}{z}}\right)^{n-\kappa} \tag{13}$$

The information processing time expectation can be found with due regard to the fact that the mathematical expectation of test number before the appearance of a successful test with probability p is equal to $\frac{1}{p}$. The mathematical expectation to be found will be as follows:

$$M[T] = 2(v+zd)e^{\frac{\lambda v}{z}} \tag{14}$$

The optimal number of stages is:

$$z_0 \approx v\sqrt{\frac{\lambda}{d}} \tag{15}$$

b) After each stage a fully reliable test with the aid of a special monitoring device is performed (Fig.3b). The correct results of information processing at each stage are stored in storage devices. For processing information successfully for time interval t it is necessary that not more than $n-z$ stages from a possible number of the stages n are unsuccessful. Thus the probability to be found is determined in the following way:

$$\mathcal{P}_z(v,t)=\sum_{\kappa=z}^{n} C_n^{\kappa} e^{-\frac{\kappa\lambda v}{z}}\left(1-e^{-\frac{\lambda v}{z}}\right)^{n-\kappa} \tag{16}$$

The information processing time expectation in this case is equal to:

$$M[T]=(v+zd)e^{\frac{\lambda v}{z}} \tag{17}$$

The optimal stages number is:

$$z_0 \approx v\sqrt{\frac{\lambda}{d}} \tag{18}$$

CONCLUSIONS

1. In the present report is proposed a system which makes it possible to accomplish a well-founded evaluation of certain reliability parameters of restorable devices accordingly to the given conditions of use and maintenance of the devices.

2. The proposed system makes it possible to rationally organize a time diagram of device operation, which leads to a considerable rise of reliability. This advantage is obtained as a result of a time redundancy, and, secondly, as result of division of the whole information processing volume into stages, which permits rational use of the available time reserve (especially due to the fact that the successful stages results do not get lost and are used again when the device starts operating.

3. The approach described above and the results thus obtained make it possible to select some characteristics of the monitoring system - type of monitoring and its regularity which ensure preset reliability standards.

4. The results thus obtained make it possible to associate the devices reliability characteristics with such an important characteristic as efficiency (speed or throughput).

Indeed, it is quite possible to introduce such device characteristic as the real efficiency coefficient determined by a ratio of net time, required for processing information volume ν (an absolute reliability of device is supposed) to the mathematical expectation of information processing time with due regard to device unreliability, stage-by-stage operation and time required for monitoring restoration. The real efficiency coefficient must be determined for a preset information processing probability value. Knowledge of these two characteristics permits, for example, settling the question of required operation speed choice when designing digital computers.

<u>Reference</u>

1. E.S.Kochetkov,Restorable systems reliability calculation with regard to the restoration time."Automation and Telemechanics",N 5,1965.

2. J.K.Beljaev,Efficiency at two types of failures. "Cybernetics in communism service",v.2,collection of papers, edited by A.I.Berg,N.G.Bruevich and B.V.Gnedenko. Publishing House "Energy",1964.

3. G.N.Cherkessov,Efficiency of the restorable systems,Thèsès of reports made on the 3-rd scientific and technical conference on reliability,P.1,published by VSNTO,Leningrad,1967.

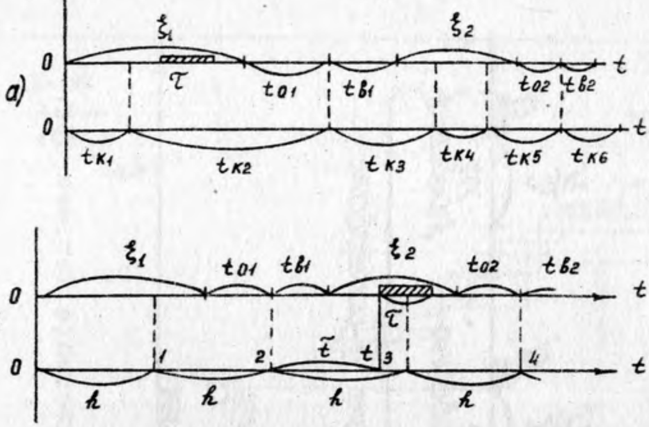


Fig.1. Random operation process of equipment under check and monitoring equipment.
a) random monitoring.
b) periodical monitoring.

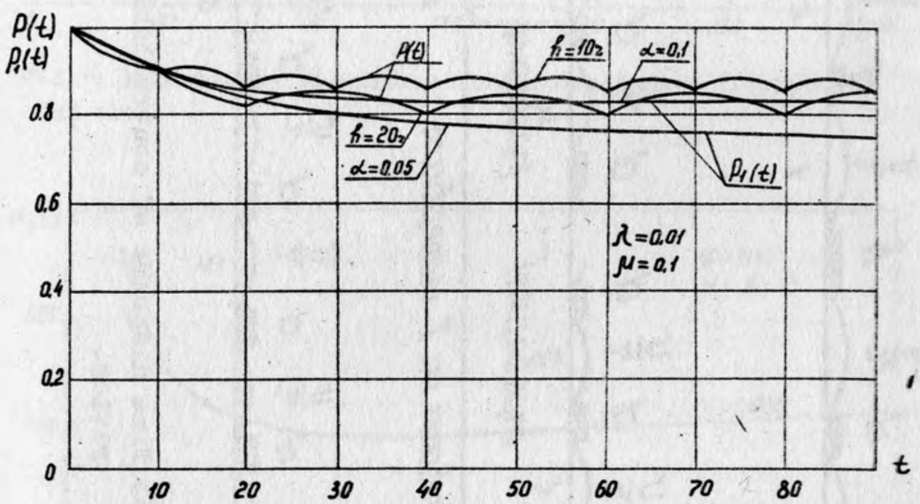


Fig.2. Dependence of probabilities $P_1(t)$ and $P_n(t)$ on time t.

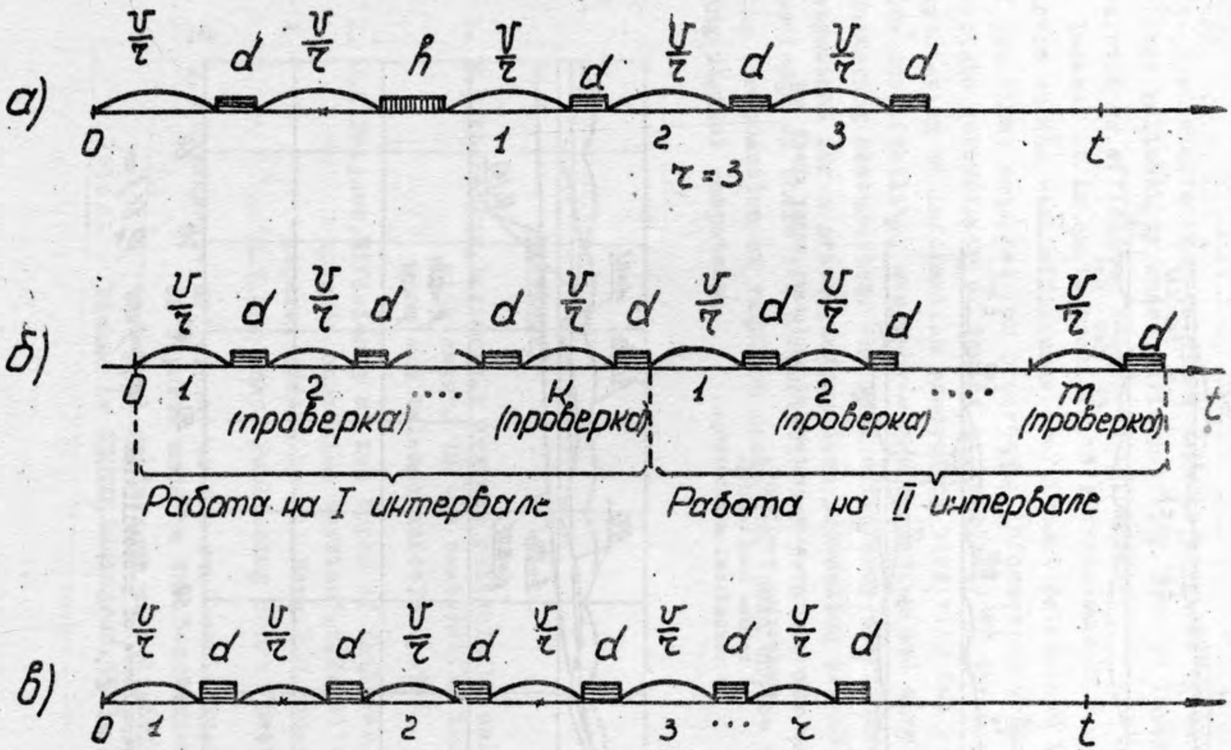


Fig.3. Time diagrams of controllable ~~xxx~~ restorable devices operation.

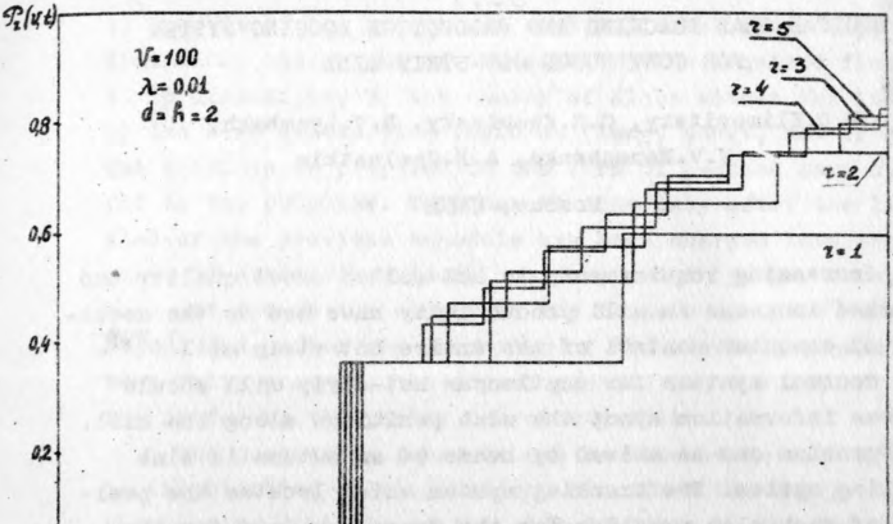


Fig.4. Probability of information processing as function of time.

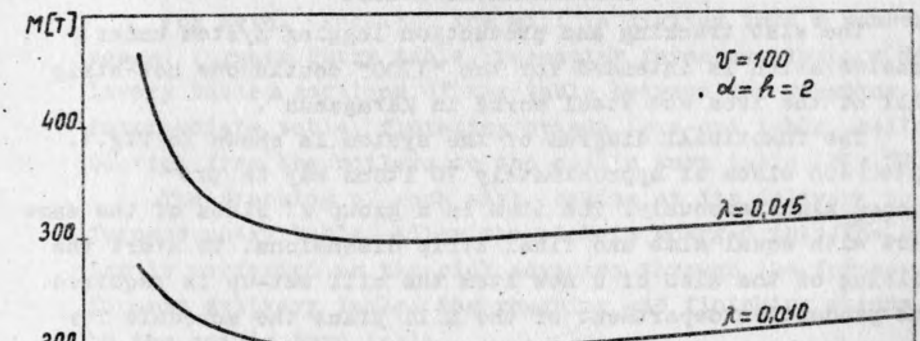


Fig.5. Information processing time expectation dependence on stages number

32.1

DIGITAL SLAB TRACKING AND PRODUCTION LOGGING SYSTEM FOR CONTINUOUS HOT-STRIP MILL

M.D.Klimovitsky, O.S.Kozhinsky, R.V.Lyambakh, V.V.Naumchenko, A.B.Chelustkin

Moscow, USSR

Increasing requirements to hot-rolled sheet quality and a marked increase in mill productivity have led to the necessity of computer control of the entire hot-strip mill[1,2,3].

Control systems for continuous hot-strip mill should possess information about the slab positions along the mill. This problem can be solved by means of an automatic slab tracking system. The tracking system which locates the position of each slab provides for the necessary data for the systems of the mill set-up, gauge control, furnace control, production logging, etc. The normal functioning of these systems requires data of the slab position with space and time.

1. Slab Tracking System

The slab tracking and production logging system under consideration is intended for the "1700" continuous hot-strip mill of the iron and steel works in Karaganda[4].

The functional diagram of the system is shown in Fig.1. Up to 400 slabs of approximately 70 items may be processed simultaneously. The item is a group of slabs of the same cast with equal slab and final strip dimensions. To start the rolling of the slab of a new item the mill set-up is required. The production department of the mill plans the schedule for the processing of different items, according to the requirements of the user's orders.

This schedule is usualy prepared for a definite time interval (not less than 3 or 4 hours) depending on the available information and includes the following data for each item:

1) numbers of the orders and cast; 2) steel grade ; 3) the dimensions and weight of the slab; 4) the required final strip dimensions; 5) the number of slabs within the item; 6) the slab distinctive features (long, short, hot or cold). The schedule is prepared in the form of punched card and then fed to the computer. This can be made only after the last slab of the previous schedule has been charged into the furnace. If a new schedule is not prepared in proper time, and thus not fed to the computer, the furnace operator receives a signal and, after the last item of the previous schedule had been fully charged, the signal prohibiting the further charging is fed to the pusher operator pulpits.

The production schedule can be altered in case of a change in the slabbing mill operation, etc. Then the charge operator changes the schedule of the slab delivery to the furnace entry table. Since the total information about the new cast is not yet available by this time, the charge operator may insert in the computer only the cast number and the number of slabs in it. According to this information, the computer assigns this cast an intrasystem item number. The rest of the cast data will be fed later by the computer operator.

For metal tracking, the mill is divided into a number of zones: furnace entry table, reheating furnaces, furnace delivery table, sections of the table between the roughing stands, intermediate table, finishing stands, run-out table, coilers, section from the coilers to the coil's turn table (Fig.2).

The tracking of each slab begins at its delivery to the furnace entry table. After the slab is charged into the furnace, the tracking is performed as the slab advances through the furnace, the furnace delivery table, the roughing and finishing stands up to the coil's turn table.

To reduce the mill set-up time, the signal is fed to the mill set-up system as soon as the last slab of the item being rolled leaves the mill unit requiring set-up. Thus, the slab of a new item may enter the first stand before the last slab of the previous item leaves the mill.

The slab tracking algorithms are developed for various process areas. Fig. 3 shows the algorithm flow-chart of the slab tracking for the furnace entry table.

In developing the above algorithm, the following conditions were adopted:

1. The slab being on a furnace entry table is charged either into a furnace P_j or into a furnace row R_q depending on its length (long or short).

2. Only one long slab or two short slabs may be on the entry table of each furnace at a given time moment.

3. A slab can move forward and backward depending on the rotation direction of the table rollers.

4. A slab delivered to the furnace loading table can be charged into this furnace only.

The following designations are accepted for the flow-chart of Fig.3:

$1MD_j(1MD_j')$ - metal detector signal;

$1TRS$ - signal of the roll-table reverse sensor;

$n_{V_i}HP_j(n_{V_i}HR_q)$ - the slab n of the item V_i is on the entry table of the furnace P_j (or in front of the furnace row R_q);

$n_{V_i}OP_j(n_{V_i}OR_q)$ - the slab n of the xx item V_i is on the loading table of the furnace P_j (or of the row R_q).

To locate the slabs moving from one mill zone to the other and within a zone, various detectors and sensors are provided, such as hot metal detectors, mill-motor load-curregnt relays, etc. For increasing the reliability of the whole system, three sensors of one and the same or different types are installed at each of the appropriate points along the mill.

A slab removed from the process for any reason is deleted from the rolling schedule by the operator of the respective mill area.

The system operation is checked by manual feeding a signal that the last slab of the given item is supplied to the furnace entry table (or discharged to the furnace delivery table). In case the signals of the operator and computer do not coincide, a corresponding signal to check the system operation is produced.

2. Production Logging System

Production logging is performed by quantitative and qualitative characteristics. All initial information is received from the tracking system and various process sensors and gauges such as thickness and width gauges, sensors of finishing and coiling temperatures, strip length gauge and coil weighbridge.

Some information (date, team number, numbers of the order and cast, steel chemical analysis, etc.) is printed without previous processing. The rest of information is printed after appropriate processing of the initial data. In particular, the coil theoretical weight G_t is determined as the product of steel specific gravity γ, required strip width S_t and thickness h_t, and actual strip length ℓ_a:

$$G_t = \gamma S_t h_t \ell_a ; \tag{2.1}$$

ε is determined as the ratio of the actual coil weight G_a and theoretical weight:

$$\varepsilon = \frac{G_a}{G_t} . \tag{2.2}$$

In addition, the total length and the location of the strip sections where the parameter U (thickness, width, finishing or coiling temperature) is outside the preset tolerances are determined. Readings of the parameters U are carried out at each half-meter length of the strip by the signals from the strip length gauge. For each reading the validity of the inequality

$$U_{min} \leqslant U_i \leqslant U_{max} \tag{2.3}$$

is checked; (U_{max} and U_{min} are the preset limits of the parameter U ; i is the reading ordinal number).

The strip length gauge is located at some distance from the process sensor or gauge, therefore the actual location of the strip section where the parameter U is outside the limits is determined by subtracting the number Δi constant for the given sensor from the number of the length gauge signals:

$$\Delta i = \frac{\Delta l}{0.5} .$$

Checking of the validity of the inequality (2.3) at arbitrary i and $i-1$ allows to distinguish the reading numbers where this inequality is violated and to indicate the violation sign. For example, if the parameter exceeds the upper limit at the point l_i (Fig.4), it is necessary to print "$0 l_i +$". For Fig.4 we obtain:

$$0 l_i + \quad + l_{i+1} 0 \quad 0 l_{i+2} - \quad - l_{i+3} 0.$$

The flow-chart of the strip classification algorithm is shown in Fig.5.

On the basis of this information, the system prints the coil report for each rolled slab.

3. Functional Part of the Tracking and Production Logging System

The slab tracking and production logging system uses the M2000 computers to cope both with the slab tracking and production logging problems and with the furnace and mill units control problems.

The tracking system discrete signals are fed by the discrete information input unit and delivered to the main frame of the computer through the communication unit. The process information from the gauges and sensors of the logging system is transmitted to the input by individual norming converters. To increase reliability, the information flows in parallel through two input units. The computer part of the system uses two processors. One, working in conjunction with the tracking system with fixed logic, has tracking and logging functions, the other, in addition to these functions, is used to solve the control problems.

The back-up of the input and functional units of the tracking system is dictated by the increased requirements to the system faultless operation. Indeed, the tracking system failure results in loss of information about the rolling schedule and the operators do not know the final dimensions of the next slab. But failure of the processor or any other device performing the control functions requires only the manual input of control-point setting for local control systems, while the tracking and logging system provides for normal functioning of the mill.

Since the amount of data which must accompany each processed slab is very large (overall information for each slab is about 80 decimal digits), the direct transmission of all information through the tracking system (by file shift) is inexpedient because it involves large equipment expenditures. Due to this reason only the slab identity data are transmitted through the tracking system with fixed logic. The rest of information essential for the mill operation is stored in the computer memory. Slab tracking, input of information and its output (on digital display panels) are carried out according to conditional (intrasystem) item numbers.

The system block-diagram is shown in Fig.6. The tracking system operates as follows.

When a new slab item arrives to the furnace entry table, all necessary primary slab data are fed into the system by the automatic data input unit B1 or by the furnace operator from its pulpit P1 (when manual input). Simultaneously, this item is given automatically its own sequential intrasystem number. All this information together with this number (which plays the part of the address) enters the logging system computer and is stored in its memory. This item number is assigned to each slab of this item. When the slab comes to the furnace entry table this number is sent to the entry table tracking unit B 2. The slab having come to the furnace loading table, this unit transmits the appropriate signal to the receiving register P of the furnace tracking unit B3. Together with the item number, the tracking system transmits the slab indication comprising three binary digits and containing information about the slab proper (hot, cold, long, short, etc.).

The slab charged into the furnace, its item number and related information are fed into the furnace tracking unit. A new item appearing at the furnace exit, the enquire to the logging system memory is made according to the item number, and necessary information is sent to the digital display panels of the pusher operators (P2, P3), furnace operator (P4), roughing mill operator (P5), shears operator (P6), and finishing mill operator (P7). The number of the slab discharged from the furnace is sent to the display panels of the pusher operators and the furnace operator, and is transmitted to the tracking unit of the furnace delivery table (B4) together with the slab indication. The first slab of a new item discharged, the information on the quantity of the slabs of this item being in the furnaces (or on the quantity of the discharged slabs) is sent to the tracking unit of the furnace delivery table as well.

This unit tracks slabs on the furnace delivery table and counts up the slabs removed from the process and the number of slabs on the delivery table; the latter is sent to the furnace operator display panel. When the first slab of the new item reaches the edger entry table, the tracking unit enquires the logging system memory, and necessary information is transmitted to display panels of the roughing mill operator, shears operator, and finishing mill operator. Number of each slab reached the edger entry table is sent to the roughing mill tracking unit B5.

This unit B5 ensures slab tracking through the roughing mill and output of information about each slab position to the roughing mill operator display panel; moreover, after the last slab of the given item left a roughing stand, the unit sends signal for the set-up of this stand. The tracking unit enquires the logging system memory (according to the numbers of the items being rolled in the roughing mill) and sends necessary data to the display panels of the roughing mill operator, shears operator, finishing mill operator, and coiler operator P8. The number of each slab having left the roughing mill is transferred to the intermediate table tracking unit B6.

This unit tracks the slab through the shears and intermediate table and deletes from the tracking system the slabs removed from the intermediate table. When the first slab of a new item reaches the roll table in front of the finishing scalebreaker, this unit enquires the logging system memory for output of necessary information to the display panels of the finishing mill operator and coiler operator.

The finishing mill tracking unit B7 tracks the slab through the finishing mill and enquires the memory for output of information to the display panels of the finishing mill and coiler operators, deletes from the tracking system the slabs removed from the finishing mill in case there were such ones, and assigns the ordinal numbers to the strips rolled during the day. When a strip leaves the finishing mill, the tracking unit B7 sends the item number, the slab number within the item and the strip ordinal number to the coiler tracking unit B8. These numbers are assigned to the data of the given slab as addresses.

The unit B8 tracks the strips according to the item numbers and slab numbers within the item up to the marker. When the first strip of a new item enters the coiler area, the memory is enquired according to the item number, and the data are delivered to the coiler operator display panel. The first coil having come to the turn table, the memory is enquired according to the item number and coil number, and the data are delivered to the marker display panel. The number of the item, of the slab within the item and of the coil are delivered by the coiler tracking unit B8 to the marker tracking unit B9.

The tracking unit B9 enquires the computer memory according to the numbers received and sends the necessary data to the display panel P9 and to the printer. At the same time all information about this coil is deleted from the computer memory. After information of the last coil of the given item is printed, all data concerning the given item are deleted from the memory.

There are four printers in the system. Two of them (with printing speed of 10 characters per second) are located near the furnaces and print data about slabs being charged and discharged. The other two (printing speed of 400 lines per minute) are installed in the computer room and print coil reports containing the quantitative and qualitative data of the coil.

The table below shows the information about slabs being charged and discharged, the data which are brought out to the operator display panels, and the coil report data.

Table

Pusher operator display panel....	34 decimal	8 binary
Charge operator display panel...	53 decimal	7 binary
Furnace operator display panel...	95 "	10 "
Roughing mill operator display panel.................	85 "	33 "
Shears operator display panel...	34 "	4 "
Finishing mill operator display panel.................	63 "	5 "
Coiler operator display panel...	63 "	5 "
Marker display panel	8 "	3 "
Coil report	From 168 to 600 decimal characters depending on the coil quality	
Two printers at the furnace operator pulpit	About 180 decimal characters per slab	

The information required for process control purposes is sent to the operators display panels installed in various areas of the mill.

The numbers of the next three items to be charged and the number of the slabs of the item being charged which are not yet delivered to the furnace entry table are displayed on the charge operator display panel. When the last slab of the given item reaches the furnace entry table, the "check-up" signal appears on the display panel. If this signal coincides actually with the arrival of the last slab of the item, the charge operator confirms this coincidence by pressing the push-button.

To ensure that all slabs of the given item are discharged from the furnaces and to prevent the slabs of another item to be discharged not in time, the number of slabs of the given item being in each furnace at the present moment is shown on the display panels of the pusher operator and furnace operator. Besides, the "check-up" signal is displayed on the furnace operator panel after the last slab of the given item is discharged.

The data relevant to the item being rolled (required final strip dimensions, the number of slabs not rolled to the present moment, the steel grade, etc.) as well as the data required for the mill set-up before the next item, are displayed on the panels of the roughing and finishing mill operators, shears operator, and coiler operator. Besides, the slab positions on the tables between the roughing stands are displayed on the roughing mill operator panel.

The numbers of the coil and cast, the date, etc., are displayed on the marker panel.

The numbers being transmitted within the tracking system are checked for parity. When the item number is introduced into the tracking system, a check digit is assigned to it.

All data are stored in input/output devices together with the address (item number) and transmitted with it upon request. Comparison of the request address with the output data allows to check the operation of the input and output devices. In case an error in the information being transmitted between any tracking units is detected, the checking unit B10 sends an alarm signal to the display panels of each control pulpit (or to the control pulpits connected with this checking unit).

References

I. Roth J.F., A computer control hierarchy in a steel plant, Iron and Steel, 1964, No. 3,4.
2. Fapiano D.J., Technical and economic considerations in plate mill process control, Iron and Steel Eng., 1966, No. 10.
3. Obelode G., Wisdika H., Automatisches Informationssystem zur Auftragsabwicklung in einem Warmbreitbandwalzwerk, Stahl und Eisen, 1967, No. 22.
4. Лямбах Р.В., Добронравов Д.Н., Ромашкевич Л.Ф., Фельдман А.В. "Комплексная автоматизация широкополосного стана горячей прокатки с применением УВМ". Доклад на IУ Всесоюзном совещании по автоматическому управлению (технической кибернетике), Тбилиси, I968.

FUNDAMENTALS OF NONLINEAR CONTROL SYSTEMS WITH THE PULSE-FREQUENCY AND PULSE-WIDTH MODULATION

V.M. Kuntsevich, Yu.N. Chekhovoi
Institute of Cybernetics
Ukrainian Academy of Sciences
Kiev
USSR

During the last years the interest in investigating the pulse-frequency and pulse-width modulation (PFM and PWM) systems has grown considerably [1-16]. From the practical point of view this interest is justified by the simplicity of such systems technical realization in comparison with the pulse relay systems and by their much higher dynamical properties [1,2,6,9]. From the standpoint of theory the systems in question are interesting by their rather original nonlinear effects due to changing the pulse repetition rate and having no analogies in the theory of continuous and amplitude pulse systems. This fact permits to generalize some classical problems of the automatic control theory (for instance, the absolute stability problems) and enriches them with new content.

1. The Motion Equations

Let us consider the nonlinear sampled data control system (Fig.1) consisting of the continuous linear part (CLP), sequential correcting filter (CF) and nonlinear pulse modulator (PM) of the first-type.

CPL consists of the linear stationary units with lumped parameters and has the fractional rational response function

$$W(s)=\frac{B_\ell(s)}{A_m(s)}=\frac{\sum_{i=0}^{\ell} b_i s^i}{s^m+\sum_{i=0}^{m-1} a_i s^i} \qquad (\ell<m), \tag{1.1}$$

CF - is the linear correcting filter defined by the equation

$$\sigma=C^T(U-X), \tag{1.2}$$

where $X=(x,x',...,x^{(m-1)})$ is the system phase coordinates column-vector; $U=(u,u',...,u^{(m-1)})$- the input signal; $C=(c_1,c_2,...,c_m)$-

– the numerical column-vector; $c_i = 0$ when $i > \kappa \, (1 \le \kappa \le m-\ell)$; „$T$" – is the operation of transformation.

Structural schemes of PM possible versions for different types of the pulse modulations are given on Fig.2. PE is the ideal sampler (pulse-amplitude modulator); the asterisk denotes the time quantizing operation made by PE: $\sigma^* = \sum_{n=0}^{\infty} \sigma_n \delta(t-t_n)$; $\delta(t)$ – is a unit delta function; $t_n = \sum_{i=0}^{n-1} T_i$ – time of n-pulse appearance ($t_0 = 0$); $T_n = t_{n+1} - t_n$ – an interval between n- and (n+1)-pulses; $\sigma_n = \lim_{0<\varepsilon \to 0} (t_n - \varepsilon)$; Φ – a zero-order hold circuit (with constant or variable hold time); RE – relay element; f and F – elements for setting a pulse duration, which control the hold circuit and pulse elements PE. The ideal pulse-frequency modulator (Fig.2, a) modulates the frequency rate and sign of the $z^*(t)$ sequence of unit δ-pulses. Real pulse-frequency modulator (Fig. 2, b) acts on the frequency rate and sign of the $y(t)$ sequence of rectangular pulses having the constant duration τ and unit amplitude. The pulse-width modulator (Fig.2, c) acts on the sign and duration of the $y(t)$ rectangular pulse sequence coming with constant frequency $1/T$. Finally, the pulse-width-frequency modulator (Fig. 2, d) acts on the sign, frequency and duration of $y(t)$. sequence.

For all PM types the modulation by sign is determined by relay function (RE-characteristic)

$$z_n = z(\sigma_n) = \begin{cases} \operatorname{sign} \sigma_n & \text{for} \quad |\sigma_n| > \Delta\,; \\ 0 & \text{for} \quad |\sigma_n| \le \Delta\,; \end{cases} \tag{1.3}$$

the frequency modulation is determined by the FPM law

$$T_n = F(\sigma_n) \tag{1.4}$$

and the pulse-width modulation – by the PWM law

$$0 \le \tau_n = f(\sigma_n) \begin{cases} < T_n & \text{for} \quad |\sigma_n| < \Delta_0\,; \\ = T_n = const & \text{for} \quad |\sigma_n| \ge \Delta_0\,. \end{cases} \tag{1.5}$$

Here $F(\sigma)$ and $f(\sigma)$–are the even simple functions defined for all σ; $F(\sigma) > 0$, $f(\sigma) \ge 0$, $f(\sigma)$ being zero only for $|\sigma| \le \Delta$; $\Delta_0 > \Delta$ – PM saturation threshold.

Difference equations of the considered systems motion are reduced to the form [10, 11, 17, 18]:

$$X_{n+1} = H_n(X_n + K_n) \tag{1.6}$$

where $X_n = (x_n, x_n', \dots, x_n^{(m-1)})$; $x_n^{(i)} = \lim_{0<\varepsilon \to 0} x^{(i)}(t_n - \varepsilon)$; $H_n = H[F(\sigma_n)] = \exp AF(\sigma_n)$ – is a transient matrix for the CLP; A – accompanying matrix for a CLP characteristic polynomial; $K_n = K[f(\sigma_n)]\, z(\sigma_n)$ – is the state changing vector of the CLP. Function $K(f)$ depends on the pulse modulation type. For the ideal PFM (Fig.2, a)

$$K(f) = G = (g, g', \dots, g^{(m-1)});\ g^{(i)} = g^{(i)}(0);\ g(t) = L^{-1}[W(s)]; \tag{1.7}$$

For the real PFM (Fig.2, b)

$$K(f) = H(-\tau)R(\tau);\ R(\tau) = (z(\tau), z'(\tau), \dots, z^{(m-1)}(\tau));\ z(t) = L^{-1}[\tfrac{1}{s}W(s)] \tag{1.8}$$

and, finally, for the PWM (Fig.2, c) and double frequency and width modulation (PFM and PWM, Fig.2, d)

$$K(f) = H(-\tau_n)R(\tau_n) = H[-f(\sigma_n)]R[f(\sigma_n)]. \tag{1.9}$$

The matrix equation (1.6) describes the system motion (Fig.1) in natural phase space $E^m = \{X_n\}$. Let us show that it is always possible to pass from (1.6) to an equation

$$\check{X}_{n+1} = \check{H}_n (\check{X}_n + \check{K}_n) \tag{1.10}$$

in difference phase space $D^m = \{\check{X}_n\}$, $\check{X}_n = (x_n, x_{n+1}, \dots, x_{n+m-1})$. Let us compose the following equation system [10, 11]:

$$\begin{Vmatrix} -H_n & I & 0 & \cdots & 0 & 0 \\ 0 & -H_{n+1} & I & \cdots & 0 & 0 \\ 0 & 0 & -H_{n+2} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -H_{n+m-1} & I \end{Vmatrix} \cdot \begin{Vmatrix} X_n \\ X_{n+1} \\ X_{n+2} \\ \cdots \\ X_{n+m-1} \\ X_{n+m} \end{Vmatrix} = \begin{Vmatrix} H_n K_n \\ H_{n+1} K_{n+1} \\ H_{n+2} K_{n+2} \\ \cdots \\ H_{n+m-1} K_{n+m-1} \end{Vmatrix} \tag{1.11}$$

where I – is a unit matrix. The system (1.11) has the rectangular matrix with $m^2 \times m(m+1)$ dimensions and of m^2 rank. Therefore (1.11) can be solved relatively to x_{n+m} assuming the variables x_{n+i} $(i = 0, 1, \dots, m-1)$ to be known. This result represented in matrix form provides the equation (1.10).

In future we shall use the motion equations in the form of (1.6) only, because its relation with system parameters is more simple and it is more convenient for investigations.

2. The Stability of the Equilibrium Points

Let us use discrete analogies of theorems of the Liapunov's direct method [11, 19, 20]. Consider the main case when CLP of the system is stable and the simplest critical case when CLP is neutral. In order the system (Fig.1) might have an equilibrium point assume $u = const$.

The main case. According to (1.6) X_∞ column-vector of the equilibrium point coordinates must satisfy the equality

$$X_\infty = (H_\infty^{-1} - I)^{-1} K_\infty, \tag{2.1}$$

where H_∞ and K_∞ - matrix H_n and vector K_n, respectively, for $X_n = X_\infty$.

For $|u| \leq \frac{\Delta}{c_1}$ ($c_1 > 0$ - the vector c element) equation (2.1) has zero solution $X_\infty = 0$, i.e. the system has the equilibrium point in the origin. This fact is easily checked by the simple substitution. For $|u| > \frac{\Delta}{c_1}$ the solution $X_\infty \neq 0$; in general case there may be several solutions[21].

By substitution $X_n = E_n^\circ + X_\infty$ we shall display the phase origin E^m in the system equilibrium point. In new coordinates instead of (1.6) we obtain

$$E_{n+1}^\circ = H_n (E_n^\circ + K_n^\circ), \tag{2.2}$$

where

$$K_n^\circ = (I - H_n^{-1}) X_\infty + K_n; \quad \sigma_n = c^T (E_\infty - E_n^\circ); \quad E_\infty = V - X_\infty \tag{2.3}$$

The Liapunov's function we shall choose as a positively determined quadratic form

$$v_n = (E_n^\circ)^T P E_n^\circ \quad ; \quad P > 0. \tag{2.4}$$

The first difference of the function (2.4) owing to (2.2) is equal to:

$$\Delta v_n = -(E_n^\circ)^T (P - M_n) E_n^\circ + 2 (E_n^\circ)^T M_n K_n^\circ + (K_n^\circ)^T M_n K_n^\circ, \quad M_n = H_n^T P H_n. \tag{2.5}$$

The system (2.2) is asymptotically stable on the whole if for any $E_n^\circ \neq 0$ function (2.5) is negative, i.e. if the inequalities are satisfied[11, 19, 20]:

$$P - M_n > 0; \tag{2.6}$$

$$(E_n^\circ)^T (P - M_n) E_n^\circ - 2 (E_n^\circ)^T M_n K_n^\circ > (K_n^\circ)^T M_n K_n^\circ. \tag{2.7}$$

It is shown that $\lim\limits_{F \to \infty} M[F(\sigma_n)] = 0$, thus there is a class of sufficiently large functions (1.4), for which the condition (2.6) is satisfied[22]. Assume that the function (1.4) belongs to this still unknown class and consider the condition (2.7).

Let us examine the equation of a surface on which the function (2.5) becomes zero:

$$(E_n^\circ)^T (P - M_n) E_n^\circ - 2 (E_n^\circ) M_n K_n^\circ = (K_n^\circ)^T M_n K_n^\circ. \tag{2.8}$$

Condition (2.7) will be satisfied if for all $F_n^\circ \neq 0$ this surface does not exist. We substitute in (2.8) σ_n defined from (2.3), by expression $\sigma_\infty - \sigma$ where $\sigma_\infty = C^T E_\infty$ and σ - is the arbitrary real parameter not depending on E_n° :

$$(E_n^\circ)^T (P-M) E_n^\circ - 2(E_n^\circ)^T M K^\circ = (K^\circ)^T M K^\circ ;\ M = M_n \Big|_{\sigma_n = \sigma_\infty - \sigma} ;\ K^\circ = K_n^\circ \Big|_{\sigma_n = \sigma_\infty - \sigma} \quad (2.9)$$

Equation (2.9) describes a family of m-dimensional ellipsoids depending on σ parameters, the surface (2.8) not existing for $E_n^\circ \neq 0$, if for all $\sigma \neq 0$ the ellipsoid (2.9) is not contiguous to $C^T E_n^\circ = \sigma$ plane (Fig.3, a) [11, 20]. Let us construct the plane

$$C^T E_n^\circ = \rho(\sigma), \quad (2.10)$$

tangent to ellipsoid (2.9). Condition (2.7) is transformed into an inequality

$$|\sigma| > |\rho(\sigma)| , \quad \sigma \neq 0 , \quad |\sigma| \leq \Delta_\sigma . \quad (2.11)$$

To determine $\rho(\sigma)$ function we write down the general equation of a plane tangent to ellipsoid (2.9) in some point [21-23]

$$(E_n^\circ)^T (P-M) E^\circ(A) - [E_n^\circ + E^\circ(A)]^T M K^\circ = (K^\circ)^T M K^\circ , \quad (2.12)$$

where $E^\circ(A)$ - radius-vector of the point of tangency (Fig.3,a). Planes (2.10) and (2.12) coincide if for some $\alpha > 0$

$$(P-M) E^\circ(A) - M K^\circ = \alpha C ; \quad (2.13)$$

$$(K^\circ)^T M K^\circ + [E^\circ(A)]^T M K^\circ = \alpha \rho(\sigma). \quad (2.14)$$

The combined solution of (2.9), (2.13) and (2.14) gives [21,22]

$$\rho(\sigma) = \left\{ (K^\circ)^T [M + M(P-M)^{-1} M] K^\circ \cdot C^T (P-M)^{-1} C \right\}^{1/2} \operatorname{sign}\sigma + C^T (P-M) M K^\circ \quad (2.15)$$

Note that when $|u| \leq \frac{\Delta}{c_1}$ function (2.15) becomes zero for $|\sigma| \leq \Delta$ and it is quite sufficiently to check the unequality (2.11) for $|\sigma| > \Delta$.

Condition (2.11) and equation (2.15) are obtained on the base of assumption that there exists (2.6) for all $\sigma \neq 0$, but it is unnecessary to check the satisfaction of this condition along all σ axis. Really let us assume that inequality (2.6) is satisfied for $\sigma = \sigma_1$ and violated for $\sigma = \sigma_2 \neq \sigma_1$; then there is such $\sigma_3 \in (\sigma_1 , \sigma_2)$ for which matrix $(P-M)$ is degenerated and function (2.15) has the gap of the continuity. It means that it is sufficiently to check (2.6) only in one arbitrary point of every interval of the function continuity. Now the final result may be formulated as follows: in the main case the equilibrium point X_∞ of system (1.6) is asymptotically stable on the whole if condition (2.11) is satisfied and inside any continuity interval of function (2.15) it is possible to point out at least one value of σ for which the inequality (2.6) is satisfied.

The simplest critical case. In this case one root of the caracteristic equation $A_m(s)=0$ is zero $(a_0=0)$, matrix A is degenerated and has $m-1$ rank. Transform (1.6) multiplying it by the matrix[22]

$$R = \begin{Vmatrix} 1 & \frac{a_2}{a_1} & \cdots & \frac{a_{m-1}}{a_1} & \frac{1}{a_1} \\ 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{Vmatrix}; \tag{2.16}$$

after simple transformations we obtain:

$$\tilde{E}^o_{n+1} = \tilde{H}_n(\tilde{E}^o_n + \tilde{K}_n), \quad \sigma_n = \tilde{C}^T(V - \tilde{X}_n) = -\tilde{C}^T\tilde{E}^o_n \tag{2.17}$$

where $\tilde{E}^o_n = \tilde{X}_n - V$; $\tilde{X}_n = RX_n$; $\tilde{K}_n = RK_n$; $\tilde{C} = (R^T)^{-1}C$; $\tilde{H}_n = RH_nR^{-1} = diag\{1, (H_n)_{1,1}\}$; $(H_n)_{1,1}$ - the matrix of $m-1$ order, obtained from H_n by crossing out the first row and first column. To equation (2.17) a set $\mathfrak{E}$ of equilibrium points corresponds over which

$$(\tilde{X}_n)_1 = (x_n)_1 = 0; \quad |u - \tilde{x}_{1,n}| = |u - x_{1,n}| \leq \frac{\Delta}{\tilde{c}_1} = \frac{\Delta}{c_1}. \tag{2.18}$$

Here $(X)_1$ is a column-vector obtained from X by crossing the first element out; $\tilde{x}_{1,n}$ and $\tilde{C}_1$ are the elements of vectors $\tilde{X}_n$ and $\tilde{C}$.

We choose the Liapunov's function in a form

$$v_n = (\tilde{E}^o_n)^T P \tilde{E}^o_n, \tag{2.19}$$

in addition demanding from matrix P that $P = diag\{1, (P)_{1,1}\}$; $(P)_{1,1} > 0$. Then

$$\Delta v_n = -(\tilde{E}^o_n)^T_1 (P - \tilde{M}_n)_{1,1} (\tilde{E}^o_n)_1 + 2(\tilde{E}^o_n)^T \tilde{M}_n K_n + \tilde{K}^T_n \tilde{M}_n \tilde{K}_n; M_n = \tilde{H}^T_n P \tilde{H}_n \tag{2.20}$$

In accordance with discrete analog of La-Sall theorem the set $\mathfrak{E}$ of equilibrium points of system (2.17) is asymptotically stable on the whole[11, 20] if

$$\Delta v_n = 0, \quad \tilde{E}^o_n \in \mathfrak{E} \quad \Delta v_n < 0, \quad \tilde{E}^o_n \notin \mathfrak{E} \tag{2.21}$$

From (2.20), (2.17) and definition of $\mathfrak{E}$ it follows that the first of two conditions (2.21) is always satisfied. The check of the second condition (2.21) is analog to the previous one and leads us to the geometrical problem of finding conditions under which m-dimensional paraboloid

$$(\tilde{E}_n^\circ)_1^T (P-\tilde{M})_{1,1} (\tilde{E}_n^\circ)_1 - 2(\tilde{E}_n^\circ)^T \tilde{M}\tilde{K} = \tilde{K}^T\tilde{M}\tilde{K};\ \tilde{M}=\tilde{M}_n\Big|_{\sigma_n=-\sigma},\ \tilde{K}=\tilde{K}_n\Big|_{\sigma_n=-\sigma} \tag{2.22}$$

is not tangent to the plane $\tilde{C}^T\tilde{E}_n^\circ = \sigma$ (Fig.3,b). As before, the problem solution leads to inequality (2.11). Omitting the intermediate calculations (analog to those realized in the main case) let us write the expression for function $\rho(\sigma)$ [22]:

$$\rho(\sigma) = -\frac{c_1}{2\tilde{k}_1}\Big\{\tilde{k}_1^2 + (\tilde{K})_1^T\big[(\tilde{M})_{1,1} + (\tilde{M})_{1,1}(P-\tilde{M})_{1,1}^{-1}(\tilde{M})_{1,1}\big](\tilde{K})_1 + \frac{\tilde{k}_1^2}{\tilde{c}_1^2}(\tilde{C})_1^T(P-\tilde{M})_{1,1}^{-1}(\tilde{C})_1\Big\} + (\tilde{C})_1^T(P-\tilde{M})_{1,1}^{-1}(\tilde{M})_{1,1}(\tilde{K})_1, \tag{2.23}$$

where $\tilde{k}_1$ - element of $\tilde{K}$ vector.

Function (2.23) becomes zero when $|\sigma| \le \Delta$, so in the simplest critical case (irrespective of u value) it is sufficient to check the inequality (2.11) only for $|\sigma| > \Delta$.

In the simplest critical case the additional condition (2.6) takes on the form:

$$(P-\tilde{M})_{1,1} > 0. \tag{2.24}$$

Taking into account the last observations the final result is formulated as follows: in the simplest critical case the set $\mathfrak{E}$ of system (1.6) equilibrium points is asymptotically stable on the whole, provided that for $|\sigma| > \Delta$ condition (2.11) is satisfied and that it is possible to indicate at least one σ value in every interval of function (2.23) continuity for which inequality (2.24) is satisfied.

<u>CLP Critical Transfer Coefficient</u>. Let us represent (1.1) as $W(s) = kW_0(s)$, where constant k (we call it CLP transfer coefficient) is $\lim_{s\to 0} W(s)$ mainly, and $\lim_{s\to 0} sW(s)$ in the simplest critical cases. Then inequality (2.11) can be reduced to the form

$$\frac{|\sigma|}{k} > |\rho_0(\sigma)|, \tag{2.26}$$

where $\rho_0(\sigma)$ is determined by (2.15) or (2.23) if instead of (1.1) we use $W_0(s)$. From (2.26) it follows that the critical transfer coefficient (or the smallest k for which the stability condition is not satisfied)corresponds to the largest k for which the function $\rho_0(\sigma)$ is in sector $[0, \frac{1}{k}]$ (Fig.4).

The Absolute Stability of the Equilibrium Points. The system consisting of a nonlinear element (NE) and a linear part (LP) is called as absolute stable if it is asymptotically stable on the whole for all NE characteristics of some class [24, 25]. In the system under investigation (Fig. 1) the nonlinear element is the PM whose properties are completely defined by (1.3), (1.4) and (1.5). Thus the absolute stability problem can be considered here in three different statements: 1) (1.4) and (1.5) are known, a class of feasable functions (1.3) is searched for; 2) (1.3) and (1.5) are known, a class of suitable PFM laws (1.4) is searched for; 3) (1.3) and (1.4) are known, a suitable class of PWM laws (1.5) is searched for [26].

To simplify the problem let us assume $u=0$ in the main case. Then the stability condition on the whole will be

$$|\sigma| > |\rho(\sigma)|, \quad \Delta < |\sigma| \leqslant \Delta_o \tag{2.27}$$

The first statement of the absolute stability problem is trivial and is solved completely by condition (2.27). Really if for some $\Delta = \Delta_*$ condition (2.27) is satisfied it will be satisfied for any $\Delta \geqslant \Delta_*$. The last inequality solves the problem completely as it defines the class of functions (1.3) searched for.

For solving the problem in its 2nd and 3rd statements let us represent the function $\rho(\sigma)$ for $\sigma>0$ in the following form:

$$\rho(\sigma) = \rho_\Delta[F(\sigma), f(\sigma)] = \rho_\Delta(F,f), \quad \sigma>0 \tag{2.28}$$

For $F \to \infty$ and $f \to 0$ the system under investigation will be broken out. In accordance with the condition the CLP is stable (or limiting stable[24]) so the feasable class of functions $F(\sigma)$ is bounded below and $f(\sigma)$ - is bounded above. Let us consider the equation

$$\sigma = \rho_\Delta(F,f), \quad \sigma>0, \tag{2.29}$$

corresponding to the boundary of the stability domain given by (2.27). If function (1.5) is set the equation (2.29) gives inexplicitly the function

$$\sigma = \rho_{\Delta,f}(F), \quad \sigma>0. \tag{2.30}$$

Function (2.30) is defined and positive over the interval

$F \in (F_0, \infty)$, where $F_0 \geq 0$ - is the highest value of F for which conditions (2.6) or (2.24) are violated. Therefore there is the positive inverse function $F_*(\sigma) = \rho^{-1}_{\Delta,f}(\sigma)$ which use makes it posssible to reduce (2.27) condition to the form:

$$F(\sigma) > F_*(\sigma) = \rho^{-1}_{\Delta,f}(\sigma); \quad \sigma \in [\Delta, \Delta_0]. \quad (2.31)$$

The obtained inequality defines the class of functions (1.4) searched for. Similarly the class of feasable functions (1.5) is defined by the inequality

$$f(\sigma) < f_*(\sigma) = \rho^{-1}_{\Delta,F}(\sigma), \quad \sigma \in (\Delta, \Delta_0). \quad (2.32)$$

In general case one fails to find inverse functions $\rho^{-1}_{\Delta,f}(\sigma)$ and $\rho^{-1}_{\Delta,F}(\sigma)$ in analytic form but it is simple to find them graphically [26].

The Asymptotic Stability Region. If condition (2.27) is violated for $\sigma = \Delta_1 \in [\Delta, \Delta_0]$, the system under investigation does not possess the asymptotic stability on the whole but it hasthe region (in E^m space) of the asymptotic stability. The largest of the open regions is the estimation of this region, bounded by the surfaces $v_n = \beta = const$ and satisfying inequality $|\sigma| < \Delta_1$. It is easy to show that in the main and the simplest critical cases, respectively, β value is equal to[26]:

$$\beta = \frac{\Delta_1^2}{C^T P^{-1} C}; \qquad \beta = \frac{\Delta_1^2}{\tilde{C}^T P^{-1} \tilde{C}}. \quad (2.33)$$

Remark about Choosing the Liapunov Function. From the stated above it follows that inequalities (2.31) and (2.32) ensure the stability region for any $P > 0$ and $(P)_{1,1} > 0$, respectively. Satisfactory results and material simplifications of calculations can be obtained when choosing

$$P = S^* S, \quad P = S^* D S, \quad (P)_{1,1} = S^* S, \quad (P)_{1,1} = S^* D S, \quad (2.34)$$

where S - matrix transforming the matrix A or $(A)_{1,1}$ respectively into the normal Jordan's form; S^* - matrix, Hermitianly conjugate to S; D - the positive elements diagonal matrix[11, 17, 18, 20-22, 26].

3. The Forced Stationar Regime Stability

Let us consider the (1.6) system in the simplest critical case for $u(t) = \omega t$ (the regime of following the linearly increasing input signal). Using the substitution $X_n = U_n - E_n$ we reduce (1.6) to the following form:

$$E_{n+1} = H_n (E_n + L_n), \quad L_n = H_n^{-1} U_{n+1} - U_n - K_n \qquad (3.1)$$

It may be shown that for $u(t) = \omega t$ this equation does not depend explicitly on t_n and there exists for it a limiting equality $\lim_{n\to\infty} E_n = E_\infty$, where E_∞ – is a numerical vector[27]. According to (3.1) vector E_∞ must satisfy the equality

$$(H_\infty^{-1} - I) \cdot E_\infty = L_\infty, \qquad L_\infty = \lim_{n\to\infty} L_n . \qquad (3.2)$$

It may be shown that there is always interval $\omega \in (\omega_1, \omega_2)$ $(\omega_2 > \omega_1 \geq 0)$ for which equation (3.2) has at least one solution[27]. By substituting $E_n = E_\infty - E_n^\circ$ we replace the origin of phase space coordinates to the point corresponding to the investigated stationary regime. In new coordinates instead of (3.1) we obtain

$$E_{n+1}^\circ = H_n (E_n^\circ + L_n^\circ), \quad L_n^\circ = -(I - H_n^{-1}) E_\infty - L_n, \; \sigma_n = C^T (E_\infty - E_n^\circ) \qquad (3.3)$$

Similarly to the previous let us transform equation (3.3) by multiplying it by the matrix (2.16)

$$\tilde{E}_{n+1}^\circ = \tilde{H}_n (\tilde{E}_n^\circ + \tilde{L}_n^\circ), \quad \sigma_n = \tilde{C}^T (\tilde{E}_\infty - E_n^\circ) = \sigma_\infty - \tilde{C}^T \tilde{E}_n^\circ, \qquad (3.4)$$

where $\tilde{L}_n^\circ = R L_n^\circ$, $\tilde{E}_\infty = R E_\infty$ (the rest of indeces correspond to these taken above). Equation (3.4) is similar to (2.17) but instead of $\mathfrak{E}$ set of equilibrium points the system (3.4) has one equilibrium point $\tilde{E}_n^\circ = 0$. When Liapunov function is chosen as (2.19) the stability condition (similarly to the previous one) will be obtained in the form of (2.11), where

$$\rho(\sigma) = -\frac{c_1}{2\tilde{\ell}_1^\circ} \Big\{ (\tilde{\ell}_1^\circ)^2 - (\tilde{L}^\circ)_1^T \big[(\tilde{M})_{1,1} + (\tilde{M})_{1,1} (P - \tilde{M})_{1,1}^{-1} (\tilde{M})_{1,1} \big] (\tilde{L}^\circ)_1 + \frac{(\tilde{\ell}_1^\circ)^2}{c_1^2} (\tilde{C})_1^T (P - M)_{1,1}^{-1} (\tilde{C})_1 \Big\} + (\tilde{C})_1^T (P - \tilde{M})_{1,1}^{-1} (\tilde{M})_{1,1} (\tilde{L}^\circ)_1 . \qquad (3.5)$$

Here $\tilde{L}^\circ = \tilde{L}_n^\circ \big|_{\sigma_n = \sigma_\infty - \sigma}$; $\tilde{M} = \tilde{M}_n \big|_{\sigma_n = \sigma_\infty - \sigma}$; $\tilde{\ell}_1^\circ$ – is vector element.

4. Limited Boundedness (Dissipativeness)

In those cases when the condition of asymptotical stability is not satisfied on the whole in (2.2) system existence of periodical regimes is possible. The exact and approximate methods for analysis of such regimes have small practical efficiency being cumbersome and requiring the apriori knowledge about the form of the periodical process (number of pulses on on a half-period, order of their alternation, etc.) or searching all possible variants[10, 11, 28]. The method of investigating a limited boundedness (dissipativeness) of automatic systems gives the best results.

System (2.2) is called a limited bounded (dissipative) on the whole if there is such a compact set $\mathfrak{S}$ (asymptotically stable set) that for any initial conditions $X_n \to \mathfrak{S}$ when $n \to \infty$. On the ground of discrete analog of the T. Yoshisava theorem system (2.2) is limited bounded, if the set $\mathfrak{P}$ over which function (2.5) is positive is limited. The estimation of the asymptotically stable set $\mathfrak{S}$ is closed domain limited by the surface $v_n = \mu$ $(0<\mu = const)$ which describes $\mathfrak{P}$ [11, 29, 30]. Let us express matrix P of the quadratic form (2.4) as $P = Q^T Q$ and transform equation (2.2) by multiplying it by Q from the left:

$$Y_{n+1} = Q H_n Q^{-1} (Y_n + Q K_n^{\circ}), \quad Y_n = Q E_n^{\circ}. \qquad (4.1)$$

In the space $E_Q^m = \{Y_n\}$ the surface $v_n = \mu$ is a sphere and, consequently, the sphere circumscribed the set $\mathfrak{S}$ is a boundary of asymptotically stable set $\mathfrak{P}$ [11,29,30]. For the second order systems this conclusion leads to the simple graphical procedure (Fig.5) that sufficiently simplifies the investigation. The examples of investigating the stability and limited boundedness of the concrete systems can be found in the works[11,17,18,20-22,26-30].

REFERENCES

1. Цыпкин Я.З. Дискретные автоматические системы,их теория и перспективы развития.В сб."Теория и применение дискретных автоматических систем".Изд-во АН СССР,1960.

2. Кунцевич В.М.Исследование переходных и установившихся процессов в релейно-импульсных и экстремальных системах с постоянным и переменным периодом регулирования.Изв.АН СССР,ОТН, Энергетика и автоматика, № 5,1961.

3. Dorf R.C., Farren M.C., Phillips C.A. Adaptive Sampling Frequency for Sampled-Data Control Systems. IRE Transactions, V. AC-7, No.1, 1962.

4. Murphy G., West R.L. The Use of Pulse-Frequency Modulation for Adaptive Control. Proceedings of the National

Electronics Conference, V. XVIII, Pt I, Chicago, 1962.

5. Ли С.С., Джонс Р.В. Интегральные системы управления с частотно-импульсной модуляцией. Докл. на 2-м межд. конгр. ИФАК, М., 1963.

6. Кунцевич В.М. Исследование импульсных самонастраивающихся систем и способов повышения их качества. Дисс. на соискание уч. ст. доктора техн. наук, Ин-т кибернетики АН УССР, Киев, 1964.

7. Murphy G.J., Wu S.H. A Stability Criterion for Pulse-Width Modulated Feedback Control Systems. IEEE Transactions, V. AC-9, No. 4, 1964.

8. Pavlidis T., Jury E.I. Analysis of a New Class of Pulse-Frequency Modulated Feedback Systems. IEEE Transactions, V. AC-10, No. 1, 1965.

9. Кунцевич В.М. Импульсные экстремальные и самонастраивающиеся системы. Изд-во "Техника", Киев, 1966.

10. Чеховой Ю.Н. Динамика импульсных систем с частотно-импульсной модуляцией. Точные методы анализа. Ч.I и II. В сб. "Теория автоматического управления", вып. I, изд. "Наукова думка", Киев, 1966.

11. Чеховой Ю.Н. Динамика нелинейных импульсных систем управления с частотно-импульсной модуляцией. Дисс. на соискание уч. ст. канд. техн. наук, Ин-т кибернетики АН УССР, Киев, 1966.

12. Pavlidis T. Stability of a Class of Discontinuous Dynamical Systems. Information and Control, V.9, No. 3, 1966.

13. Clark J.P.C., Noges E. The Stability of Pulse-Frequency Modulated Closed Loop Control Systems. IEEE International Conventional Record, V. 14, No. 6, 1966.

14. Чеховой Ю.Н. Нелинейный частотно-импульсный регулятор. Энергетика и электрификация, № I, Киев, 1967.

15. Гелиг А.Х. Стабилизация нелинейных систем с частотно-импульсной модуляцией. Автоматика и телемеханика, № 6, 1967.

16. Дымков В.И. Об абсолютной устойчивости частотно-импульсных систем. Автоматика и телемеханика, № 10, 1967.

17. Кунцевич В.М., Чеховой Ю.Н. Устойчивость систем с двойной (частотной и широтной) импульсной модуляцией. Автоматика и телемеханика, № 7, 1967.

18. Кунцевич В.М., Чеховой Ю.Н. Исследование устойчивости систем управления с двойной (частотной и широтной) импульсной модуляцией прямым методом Ляпунова. В сб. "Сложные системы управления", вып. 3, изд-во "Наукова думка", Киев, 1967.

19. Бромберг П.В. Матричные методы в теории релейного и

импульсного регулирования. Изд-во "Наука",М.,1967.

20.Кунцевич В.М.,Чеховой Ю.Н.Исследование устойчивости импульсных систем управления с частотно-импульсной модуляцией прямым методом Ляпунова. Автоматика и телемеханика, № 2, 1967.

21.Чеховой Ю.Н.Устойчивость систем управления с частотно-импульсной модуляцией при постоянных внешних воздействиях. В сб."Дискретные системы управления",вып.I.Изд-во "Наукова думка",Киев,1969, в печати.

22.Чеховой Ю.Н.К задаче об устойчивости в целом импульсных систем управления с частотно-импульсной модуляцией,Автоматика и телемеханика, № 6,1968.

23.Розенфельд Б.А. Многомерные пространства.Изд-во "Наука",М.,1966.

24.Айзерман М.А.,Гантмахер Ф.Р. Абсолютная устойчивость регулируемых систем. Изд-во АН СССР,М.1963.

25.Гантмахер Ф.Р.,Якубович В.А.Абсолютная устойчивость нелинейных регулируемых систем. В трудах 2-го всесоюзного съезда по теор.и прикл.механике, вып.I,изд-во "Наука",М.1965.

26.Чеховой Ю.Н.Абсолютная устойчивость импульсных систем управления с частотной и широтной импульсной модуляцией.В сб. "Теория автоматического управления",вып.I,Киев,1968.

27.Кунцевич В.М. Исследование устойчивости стационарного вынужденного режима в системах управления с частотно-импульсной модуляцией.В сб."Дискретные системы управления",вып.I, изд-во "Наукова думка",Киев,1969,в печати.

28.Чеховой Ю.Н.Динамика импульсных систем с частотно-импульсной модуляцией.Ч.II,В сб."Теория автоматического управления",вып.3,Киев,1967.

29.Чеховой Ю.Н.Об одной форме устойчивости нелинейных импульсных систем. Изв.АН СССР.Техническая кибернетика, № 4, 1967.

30.Чеховой Ю.Н.Предельная ограниченность нелинейных импульсных систем. В сб."Теория автоматического управления", вып.2,Киев,1967.

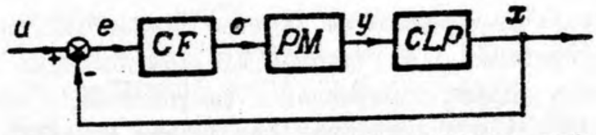


Fig.1. Block diagram of nonlinear pulse system of automatic control.

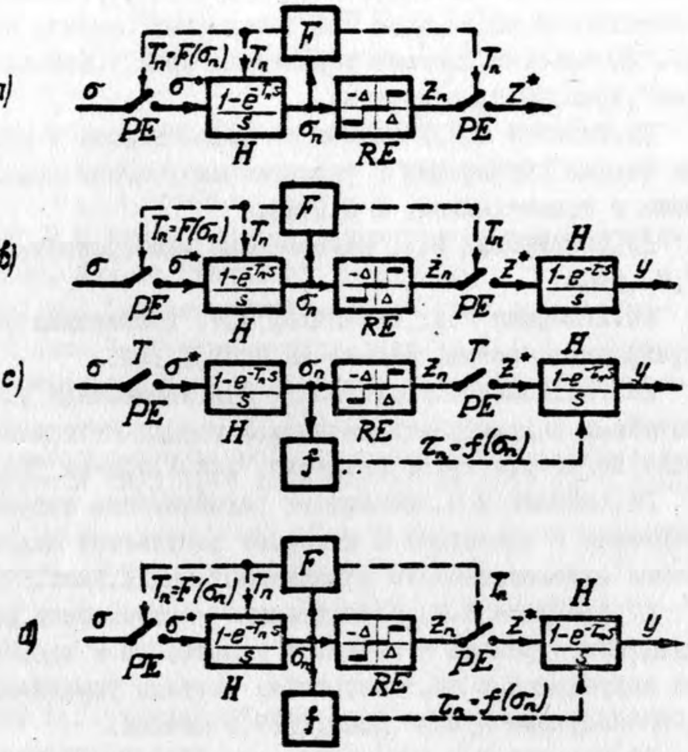


Fig.2. Block diagrams of nonlinear pulse modulators: a - ideal frequency-pulse modulator; b - real frequency-pulse modulator; c - width-pulse modulator; d - frequency-width pulse modulator.

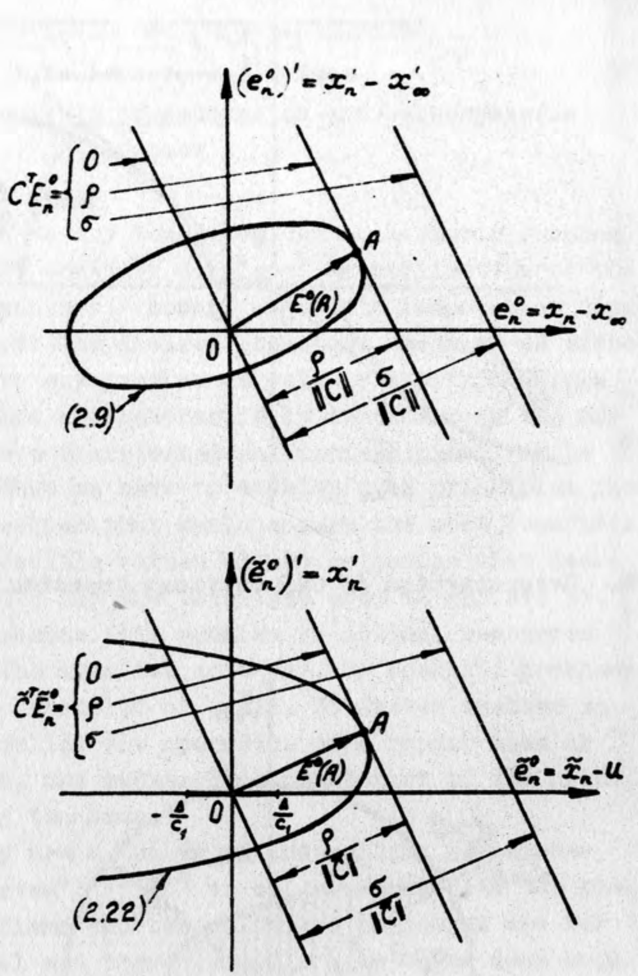


Fig.3. Graphical interpretation of stability condition (2.11) in the main (a) and simplest critical (b) cases.

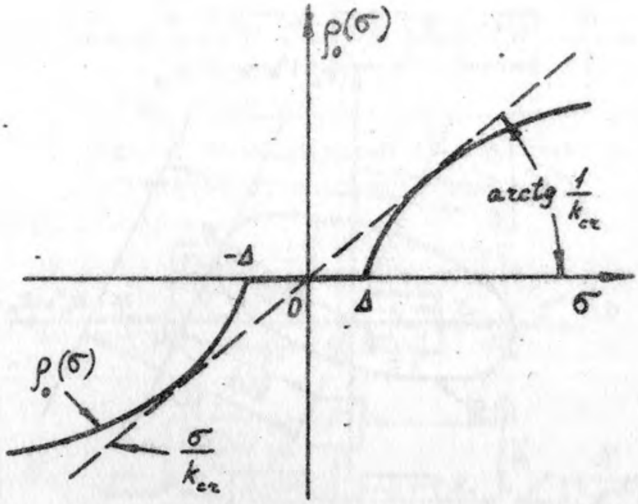


Fig.4. Determination of CLP critical transfer coefficient.

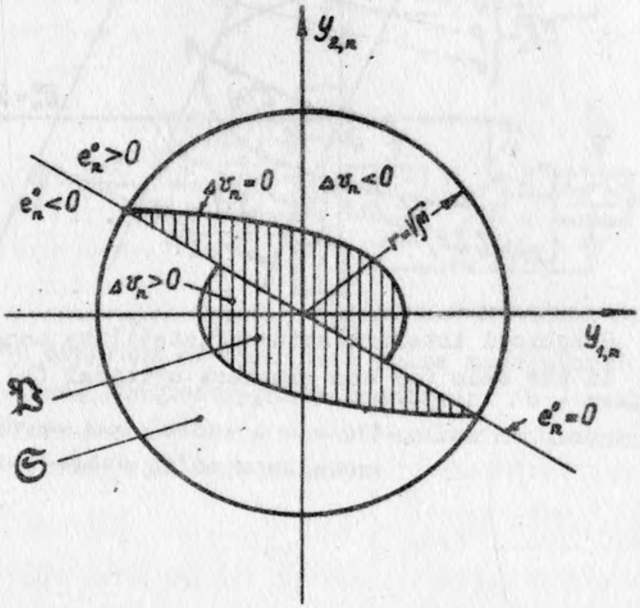


Fig.5, Determination of asymptotically stable set of limited bounded system of the 2nd order.

35.1

ON OPTIMAL RESOURCES ALLOCATION

A.Ya.Lerner, A.I.Teiman
Institute of Automation and Telemechanics
Moscow
USSR

A solution to any technological or economic problem requires certain activity displayed as utilization of the appropriate resources - human, material, temporal or financial - to achieve the desired objective. Because an effective solution to any problem as well as the expenditure of time and funds are substantially dependent on the way the resources are distributed and redistributed, it is quite natural that we have to develop such principles for resources allocation that would ensure the most beneficial value of the possible values of the criterion that describes the result and the technique used to achieve it.

In its essence this problem of optimal resources allocation is the same for most various specific problems whatever their variables or scale. No matter whether we deal with scheduling the operation of a repair team or a major project, the mathematical statement of the problem is actually the same.

Thought by now a number of interesting researches have been reported[1,2,3 etc] we believe that both the statements of problems and the solutions suggested are far from being final and there is still much to be done both in theoretical research and new applications.

This paper is an attempt to give strict definitions to basic concepts and suggest possible solutions to certain theoretical problems. The research was undertaken in the Large-Scale Systems Division of the Institute of Automation and Telemechanics[x/].

---

x/ Among the collaborators of the Division who took part in the research we would note V.N.Burkov, A.A.Voronov, S.Ye.Lovetsky and others.

A number of definitions as well as classification of problems and techniques are not conventional. The writers hope that a discussion will result in a uniform point of view on these matters which is now very much desirable.

Basic Concepts and Definitions

1.1. An activity is a process described by equation of the form

$$\omega_i(t) = \frac{dx_i(t)}{dt} = \varphi_i\left[u_{i1}(t), \ldots, u_{im}(t)\right] \tag{1}$$

where $x_i(t)$ is the status of the i-th activity at time t,

$u_{ij}(t)$ is the amount of resources of the j-th form in the i-th activity at time t,

$\omega_i(t)$ is the rate at which the i-th activity is being accomplished at time t. Since resources are involved in an activity in certain ratios the concept of resources set will be useful.

A resources set in the i-th activity is a set of resources allocations $\{u_{ij}(t)\}$ such that $u_{ij}(t) = \alpha_{ij}\, \vartheta_i(t)$ where $\{\alpha_{ij}\}$ are the parameters of the set and $\vartheta_i(t)$ is the capacity of the set at time t. Different sets are generally feasible in an activity, but for simplicity we will discuss only the case of one feasible set per activity. With the given parameters of a set the rate of each activity depends only on the capacity of the set and time elapsed (the rate will be assumed not depending on time explicitely). Thus

$$\omega_i(t) = f_i\left[\vartheta_i(t)\right] \tag{2}$$

where f_i is a non-decreasing continuous function of ϑ_i and $f_i(0) = 0$. Completion of an activity implies the change of its state from the initial one $x_i(0) = 0$ to the final one $x_i(T) = W_i$ where W_i is the volume of the activity and T is the time when all activities have been completed.

1.2. A final set of activities makes a complex of activities. This is normally shown as a network. The state of the complex is the vector $X(t) = (x_1(t), \ldots, x_n(t))$ whose components are the states of activities in the complex. The volume of the complex is the vector $W = (W_1, \ldots, W_n)$. However, the concept of an equivalent volume is sometimes more convenient as a scalar quantity depending on the volumes of activities (e.g. $W_e = (\sum_i W_i^{\alpha})^{1/\alpha}, \alpha > 1$).

1.3. Let us introduce two types of constraints on resources, constraints on capacity (instantaneous constraints)

$$\sum_i \alpha_{ij} \vartheta_i(t) \le N_j(t) \quad , \quad j = 1, 2, \ldots, m \tag{3}$$

and constraints on expenditure (integral constraints)

$$\sum_i \alpha_{ij} S_i \le S_j \quad , \quad j = 1, 2, \ldots, m. \tag{4}$$

where $S_i = \int_0^T \vartheta_i(t) dt$ are expenditures of capacity.

2. Description of the model

2.1. The statement of and solution to the problems of optimal resources allocation may be based on a model of a complex of activities including the following constituents

a) A network or a matrix which represents the composition, volumes and feasible sequences of activities in the complex.

b) A graph or a matrix representing the feasible distributions of resources among the activities of the complex, permissible expenditures of resources to complete these activities, and expenditures of time and funds for redistribution of resources.

c) A set of equations representing the activities of the complex: $\omega_i(t) = f_i[\vartheta_i(t)]$

d) A functional $\mathcal{J}$ which represents the total efficiency in completion of the activity complex.

2.2. The problem of optimal resources allocation consists in selecting programs $\varphi_{ij}^{K}(t)$ (the amount of resources of the K-th form that are reallocated from the i-th activity to the j-th activity at time t) such that with the constraints on feasible sequences of operations (2.1a) and on feasible values of resources fluxes (2.1b) would ensure completion of the complex described by eqs of (2.1b) so as the functional $\mathcal{J}$ takes the minimal value.

2.3. Depending on specific conditions the functional may depend on different arguments or their combinations. In practice the most important are cases where the objectives of optimization include the following factors:

T – the time of completing the complex;

R – resources allocated for completion of a complex;

P – the probability that the complex will be completed within the time interval

K – the quality of the result.

Generally optimizatiom of the schedule for completion of a complex should include minimization of a certain functional of the above factors

$$\mathcal{J}(T, R, P, K).$$

In specific cases some arguments of the functional $\mathcal{J}$ can be neglected. Depending on the parameters that are included either in the functional $\mathcal{J}$ or in the constraints we can distinguish the classes of problems represented in Fig. 1 (the time parameter is assumed to be always incorporated in the problem). The problems of the class T (time) are not optimization problems. Such are those of the complex analysis (finding the time when the complex has been completed, finding the time slack, etc.). The class TP (time-probability) is made chiefly by the problems of allocationg the dependent time slacks between operations qith the view to maximizing the probability of meeting the schedule. The classes which incorporate the quality are relatively little known. The remaining discussion will cover the problems of the classes time-resources and time-probability.(TR and TP).

3. Statement of the problems

3.1. Two basic classes of problems can be distinguished. In the problems of the first type the time T for completion of the complex is given, in the problems of the second type it is not.

Problem 1. Find $\vartheta_i(t), i = 1, 2, \ldots, n$ that satisfy eqs (3), (4) so that the complex be completed within the time T and the optimality criterion take the minimal value.

Problem 2. Find $\vartheta_i(t), i = 1, 2, \ldots n$ so that the complex be completed and the optimality criterion take the minimal value (time is not given).

3.2. The following optimality criteria are most widely used

a) $J_1 = T$ (minimization of the complex time)

b) $J_2 = \sum_j c_j \max \sum_i \alpha_{ij} \vartheta_i(t)$ (minimization of the resources levels).

c) $J_3 = \sum_j c_j \int_0^T (\sum_i \alpha_{ij} \vartheta_i(t))^2 dt$ (steady of resources)

(minimization of costs)

d) $J_4 = \sum_j c_j \sum_i S_i \alpha_{ij}$

e) $J_5 = \alpha T + J_2$

f) $J_6 = \alpha T + J_4$

In these criteria $c_i \geq 0$ represents the costs of resources of the j -th form,

a is the losses when the complex completion is delayed by a time unit. If there are intermidiate objectives, e.g. the earliest possible occurence of certain events, the criterion

$$J_8 = \sum_i \sigma_i (t_i - \Delta_i)$$

is used where $\sigma_i(t_i - \Delta_i) = 0$ at $t_i \leq \Delta_i$ and is a non-decreasing function t_i at $t_i > \Delta_i$, being the time t_i when the i- th activity is completed.

4. Complex aggregation

4.1. Complex aggregation is replacement of a complex or its part by one activity[4]. Aggregation is used when a complex consists of a large number of operations. Direct solution to the optimization problem is difficult due to its high dimensionality. Aggregation is made in three stages:

1. Ordering the states of the complex.
2. Finding the parameters of the set of the aggregated operation.
3. Finding the relation between the rate of the aggregated operation and the capacity of the set.

4.2. Let us consider an example of aggregating a complex of P sequential activities completed by resources of the same kind. The first and the second stages are unnecessary since the states of the complex have already been ordered and the single parameter of the set of the aggregated activity can be assumed to be equal to one. Let $\tau_i(\vartheta)$ denote the time for completion of the i-th activity at capacity ϑ of its set, i.e.

$$\tau_i(\vartheta) = \frac{W_i}{f_i(\vartheta)} .$$

Then by definition the rate of the aggregated activity

$$f(\vartheta) = \frac{W}{\sum \tau_i(\vartheta)} \tag{5}$$

Assume the volume of the aggregated activity

$$W = \sum_i \beta_i W_i .$$

Estimate the aggregation error. Let the capacity of the set of the i-th activity be ϑ_i. Then the time for completion of an aggregated activity

$$\tau(\vartheta_1, \ldots, \vartheta_p) = \sum_i \frac{\beta_i W_i}{f(\vartheta_i)} = \frac{1}{W} \sum_j \sum_i \beta_j W_j \tau_i(\vartheta_j) .$$

The actual time of the complex

$$\tau'(\vartheta_1, \ldots, \vartheta_p) = \sum_i \tau_i(\vartheta_i) = \frac{1}{W} \sum_j \sum_i \beta_j W_j \tau_i(\vartheta_i) .$$

The aggregation error

$$\varepsilon = \tau'(\vartheta_1, \ldots, \vartheta_p) - \tau(\vartheta_1, \ldots, \vartheta_p) = \frac{1}{W} \sum_j \sum_i \beta_j W_j \left[\tau_i(\vartheta_i) - \tau_i(\vartheta_j)\right] \tag{6}$$

In this case the problem of optimal aggregation is in finding $\beta_1, \ldots, \beta_p$ so as to make the aggregation error minimal.

Theorem 1. Let $\tau_i(\vartheta_i) = a_i + b_i \tau(\vartheta_i)$, $i = 1, \ldots, p$ then the aggregation error is zero at $\beta_j = \frac{c b_j}{W_j}$ ($c > 0$ is an arbitrary constant). The proof is found by direct verification. This theorem makes it possible to solve the problem of optimal aggregation by representing the relations $\tau_i(\vartheta)$ in the approximate form $a_i + b_i \tau(\vartheta)$ and assuming

$$\beta_i^0 = c b_i / W_i .$$

4.3. The problem of optimal aggregation in which the maximal relative downtime of various types of resources was minimized in Ref.[5] for the case of linear relations between the rates of activities and the capacity of resources set. That ideal aggregation is possible (where the aggregation error is zero) in a network of an arbitrary form was shown in Ref.[6] for the case of exponential relation between the rates of activities, on the one hand and the set capacity

w_i and the resources of the same type, on the other.

Let us proceed to the precise methods of solution developed for various specific statements of the problem, such as the case of independent activities, the case of ordered events, the problem of expenditure allocation and a number of problems of the combinatorial type.

5. Independent activities

5.1. Let $f_i(v_i)$ be convex upwards functions of v_i. Then it can be shown [7] that in an optimal solution all operations are performed at constant intensity and complete simultaneously. Hence

$$f_i(v_i) = W_i/T \ , \ v_i = \varphi_i(W_i/T)$$

where φ_i is a function, reverse to f_i.
The minimal time of completing a complex is defined as a minimal T for which following set of inequalities is true

$$\sum_i \alpha_{ij} \varphi_i(W_i/T) \le N_j \ , \ j = 1, 2, \dots, m. \tag{7}$$

Example 1 $f_i(v_i) = v_i \ , \ v_i \le \beta_i \ , \ i = 1, 2, \dots, n$.
We have

$$\sum_i \alpha_{ij} W_i/T \le N_j \ , \ j = 1, 2, \dots, m.$$

$$W_i/T \le \beta_i \ , \ i = 1, 2, \dots, n$$

$$T_{min} = \max\left[\max W_i/\beta_i \ ; \max_j \frac{1}{N_j} \sum_i \alpha_{ij} W_i\right] \tag{8}$$

5.2. If $f_i(v_i)$ are non-convex functions then the optimal solution generally consists of intervals of n constants. For what follows of importance is

Theorem 11. The minimal time for completion of a complex $T_{min}(W_1, \dots, W_n)$ is a convex (downwards) function of activity volumes (even if $f_i(v_i))$ are non-convex functions).

6. Ordered events

6.1. The events of a complex are ordered if the time when the i - th activity occurs is no longer than the time of the j - th event, provided that $i < j$.
Denote by Δ_S the duration of the interval between $(S-1)$ and S - th event $S = 1, 2, \dots, z$, R_S is a set of activities that can be completed within the S-th inter-

val Q is the set of intervals where the i-th activity can be completed, x_{is} is the volume of the i-th activity completed in the s-th interval. Let $z_s = \{x_{is} : i \in R_s\}$ be given. Then for each activity we have a case of independent activities and can find the minimal duration of the s-th interval of $\Delta_s(z_s)$. The time for completing a complex

$$T = \sum_s \Delta_s(z_s) \tag{9}$$

is by Theorem 11 a convex (below) function of x_{is}. Thus we have the problem of minimizing a convex function at linear constraints

$$\sum_{s \in Q_i} x_{is} = W_i, \quad i = 1, 2, \ldots, n. \tag{10}$$

solvable by any techniques of convex programming.

6.2. Let us take now the problem of steady use of resources. Assume that each activity requires resources of just one kind. Denote as $R_{sj} \subset R_s$ set of activities which require resources of the j-th kind. We have to minimize

$$J_3 = \sum_j c_j \sum_s \frac{1}{\Delta_s} \Big(\sum_{i \in R_{sj}} x_{is} \Big)^2 \tag{11}$$

at constraints (10) and

$$\sum_s \Delta_s = T. \tag{12}$$

Since conditions (10) and (12) are independent we can improve successively a certain feasible solution by varying first x_{is} at fixed Δ_s and then Δ_s at fixed x_{is}.

1st stage. Assume that feasible values of $\Delta_s \geq 0$ are given such that $\sum_s \Delta_s = T$. Then the problem of minimization decomposes into m independent problems (by the number of resourceskinds). Each of these is in minimizing

$$J_j = \sum_s \frac{1}{\Delta_s} \Big(\sum_{i \in R_{sj}} x_{is} \Big)^2 \tag{13}$$

at constraints (10). This is a problem of quadratic programming.

2nd stage. Denote by x_{is}^{o} the optimal values obtained at the 1st stage.

$$B_s^2 = \sum_j c_j \Big(\sum_{i \in R_{sj}} x_{is}^{o} \Big)^2 \qquad (14)$$

We have the minimization problem

$$J_3 = \sum_s \frac{B_s^2}{\Delta_s} \qquad (15)$$

at constraint (12). By using the technique of Lagrangian multipliers we obtain directly

$$\Delta_s^{o} = \frac{B_s T}{\sum_s B_s} \qquad (16)$$

Then the 1st and the 2nd stage are made in the reverse order.

6.3. The most difficult is the 1st stage. To perform this a successive improvement technique is suggested which is based on hydrodynamic analogy[8,9] which is a modification of the quadratic programming technique[10].

An effective algorithm for the 1st stage can be obtained from the following rule for resources allocation: the activities with the minimal number of a final event are performed in the first turn. Its application is illustrated by this example.

Example 3. Let us consider a network of 12 activities shown in Fig.2 ($A_i(W_i)$ denotes the i-th activity of the volume W_i). Let $\Delta_1 = \Delta_2 = \cdots = \Delta_6 = 7$. We find[x/] that $\lambda_o = \frac{1}{6}\sum_i W_i = 12$. Allocate the resources and denote the events with signs. The sign (–) marks only event 3. The event closest to it on the left-hand side, event 1, is denoted as (+). Since in intervals 2 and 3 no activity with a final event above 3 is completed, we make a subnetwork of events (1,2,3) and in the remaining part of the network these events are united into one. The remainder of the network is a serial connection of two networks, one of which includes event and the united event (1,2,3), and the other the united event (1,2,3) and

---

x/ The algorithm was suggested by V.N.Burkov.

and events 4,5,6. Therefore the problem can be solved separately for each network. We obtain three subnetworks shown in Fig.3.

For the network G_1, $\lambda_1 = 8$ and the feasible solution is $x_{11} = 3$, $x_{21} = 5$.

For the network G_2, $\lambda_2 = 17$ and the feasible solution $x_{32} = 14$, $x_{42} = 3$, $x_{43} = 2$, $x_{63} = 15$.

For the network G_3, $\lambda_3 = 10$ and the feasible solutions $x_{74} = 5$, $x_{84} = 2$, $x_{94} = 2$, $x_{54} = 1$, $x_{55} = 7$, $x_{10,5} = 1$, $x_{11,5} = 2$, $x_{11,6} = 2$, $x_{12,6} = 8$.

The values of $\{x_{is}\}$ obtained make the optimal solution to the 1st stage.

The value of the criterion $J = 7 \cdot \lambda_1^2 + 2\lambda_2^2 + 3\lambda_3^2 = 942$.

7. Problems of the class TP

7.1. Among the problems of the TP class one can distinguish the problems of maximal complex reliability, i.e. of scheduling the realization of the complex so as to ensure a minimal possibility of not meeting the schedule. Two types of problems can be distinguished.

A. The probability distribution $P(t)$ of completing the activity is known. In this case there are two problems.

Problem 1. For the given time T at which the complex is to be realized find a distribution of times for completion of activities $\{t_i\}$ such that $P\{t_n \le T\} \longrightarrow \max$.

Problem 2. For the desired level of reliability in complex P_0 find a distribution of $\{t_i\}$ such that

$$P\{t_n \le T\} \ge P_0$$

B. Probability distribution $P(t)$ is unknown. In this case a certain criterion $J(t,\Delta t)$ is formed that depends on the time t of completing the activities and on the amount of reserved time slack. The value of the criterion J is an estimate of the complex reliability; problems similar to Problems 1 and 2 can be formulated for it. A number of statements for problems of the A and B type as well as the algorithms for their solution were discussed in[11,12].

7.2. It is a specific feature of the TPR time probability and time resources classes, types A and B problems that the complex is assumed given. In practice this is preceded by the process of making the complex; the problem of scheduling this process optimally arises. Below we will present the results obtained by one of the writers[13].

7.3. A model of complex formation

Let the complex S_0 consist of one activity at the initial time. Then in the process of scheduling it decomposes into parts, detailed and in the long run the complex S_0 contains n activities. The activities of the complex are decomposed, specified and aggregated.

Let us define these basic concepts. It is assumed that ξ is a random duration of an activity and $0 \le a \le \xi \le b < \infty$, $c = b - a > 0$, $F(x) = P\{\xi \le x\}$, $m = M\xi$, $\sigma^2 = D\xi$. Assuming that $\alpha = \sigma/(b-a)$; $\beta = (m-a)/(b-a)$ it easily follows that $0 \le \alpha \le 1/2$, $0 \le \beta \le 1$.

Specification of activities

The activity ξ_2 is specification of ξ_1 if
$b_2 - a_2 \le b_1 - a_1$ and $\alpha_2 \le \alpha_1$.

Activities decomposition

1. Let ξ decompose into k successive activities $\xi_1, \xi_2, \dots, \xi_K$. This decomposition is regular if
   a. random values ξ_i are independent
   b. $\sum_1^K a_i = a$; $\sum_1^K b_i = b$
   c. $\alpha_i \le \alpha$
   d. $\sum_1^K \beta_i c_i / c = \beta$.

11. Let ξ decomposes into k parallel activities $\xi_1, \xi_2, \dots, \xi_K$. This decomposition is regular if
   a. random values of ξ_i are independent
   b. $\max a_i = a$; $\max b_i = b$; $c_i \le c$
   c. $\alpha_i \le \alpha$
   d. $\max[\beta_i c_i + a_i] = \beta c + a$

111. Let the networks S_i have the corresponding decomposition Ω_i while the networks S_{i+1} the decomposition Ω_{i+1}. The decomposition $\Omega_{i+1} > \Omega_i$ and regular, if obtained by regular breakdown of activities in the network S_i.

Activities aggregation

Activities aggregation is unification of several activities into one. Like decomposition it can be made in series or in parallel. Aggregation will be defined as operation reverse to decomposition.

7.4. Analysis of complex formation scheduling. Basic theorems.

Theorem 1. Let t_{i_s} be the time for occourence of the i_s event, $t^*_{i_s}$ is a fixed scheduled deadline. Then for each event i_s there is a number τ_{i_s} such that whatever probability P_0 is given, there always is a sequence of regular complex network network decomposition $\Omega_0 \prec \Omega_t \prec$ for which, starting from Ω_0,

$$P\{t_{i_s} > t^*_{i_s}\} \geq P_0 \qquad at \quad t^*_{i_s} < \tau_{i_s}$$

and

$$P\{t_i < t^*_{i_s}\} \geq P_0 \qquad at \quad t^*_{i_s} > \tau_{i_s} .$$

If separate classes of decompositions are discussed then stricter statements can be proved. E.g. there is

Theorem 11. Let $\{\Omega_i\}, i=1,2,..$ be regular successive decompositions. Theorem 1 is true then for Ω_0 and any $\Omega_i > \Omega_0$. To prove these Theorems[13] those networks are taken that consist only of serial or parallel activities and a network of a general form.

The results obtained show that there are certain crititical deadlines for a complex and if deadlines are chosen below below critical values the probability of failures will be very high.

A number of scheduling procedures, e.g. those that use averaged indices can be shown to be incorrect in this sense.

Therefore two classes of problems appear, formulation of upper-bounds for $\bar{t} \geq \tau$ and optimal choice of t^*_i.

8. Conclusion

1. Problems of optimal resources allocation found in practice of recent years are of great economic importance. However, neither basic concepts nor strict and sufficiently general statements of problems have not been defined so far.

2. The paper is an attempt to satisfy the urgent need in formulating the problems of optimal resources allocation by using models that represent sufficiently the actual conditions for implementing a complex of activities.

It has been found that no solution to this class of problems can be based on a single mathematical tool; this requires the entire range of various tools such as mathematical programming, theory of graphs, optimal control theory, etc.

3. The efforts of scientists have undoubtedly be directed at development of techniques for solution to a number of urgent resources allocation problems such as finding paths for transportation of resources optimal in terms of maximal reliability of achieving the objective, development of techniques for solution to a general problem of optimal resources allocation, both deterministically and stochastically.

4. Some problems raised in the paper are reported in more detail in References.

References

1. Altaev V.Ya., Burkov V.N., Teiman A.I. Automation and Remote Control, vol. 27, No 5, May 1966.
2. Mcgee A.A. and Markarian M.D. Optimal allocation of research (engineering manpower within a multi-project organisational structures). IRE Trans. on Eng.Manag., 1962, v.9, N 3.
3. Lambourn S. Resource allocation and multi - project scheduling (RAMPS). A new tool in planning and control. Computer J., 1963, v.5.
4. Бурков В.Н., Лернер А.Я. Новые задачи теории сетевого планирования и управления. Сборник "Вопросы управления большими системами". Изд-во "Оптиприбор", 1967.
5. Чеботарев О.Г. распределение ресурсов в многотемных разработках на основе агрегирования комплекса операций. Доклад на IУ Всесоюзном совещании по автоматическому управлению, Тбилиси, 1968.

6. Бурков В.Н. Оптимальное управление комплексами операций. Доклад на IУ Всесоюзном совещании по автоматическому управлению, Тбилиси, 1968.

7. Burkov V.N. Automation and Remote Control, vol 27, No 7, July 1966.

8. Razumikhin B.S. Automation and Remote Control, vol 26, No 7, July 1965.

9. Razumikhin B.S. Automation and Remote Control, vol.28, No.1, January 1967.

10. Voronov A.A., Petrushinin Ye.P. Automation and Remote Control, vol. 27, No 11, 1966.

II. Тейман А.И. Календарное планирование комплексов операций. Материалы Всесоюзного семинара по теоретическим проблемам управления большими системами и исследованию операций. Изд-во "Знание", Москва, 1967.

I2. Тейман А.И. Оптимальное планирование комплексов операций. Сб. "Вопросы управления большими системами". Онтиприбор", 1967.

I3. Тейман А.И. Некоторые задачи управления комплексами операцийв условиях неопределенности. Доклад на IУ Всесоюзном совещании по автоматическому управлению, Тбилиси, 1968.

35.2

ON STOCK CONTROL THEORY

V.Avdiysky, A.Voronov, S.Lovetsky

The problem considered is to define the optimum level of row materials to be maintained by a large plant when the demand is stochastic and an outside sourse of the row materials is available.

§1. Statement of Deterministic Stock Control Problem

We shall assume that the warehouse of a large plant (say in the steel industry) holds stocks of different items of row materials P_i $(i = 1, 2, \ldots, m)$ (e.g. metal bars of different sizes); from which n types of product π_j $(j = 1, 2, \ldots, n)$ (e.g. different types of rolled metal) must be produced. Generally speaking each type of product can be manufactured from one row item not necessarily always the same one. However we shall restrict ourselves to the case when having chosen the row material P_i for the production π_j we shall now only use this row item for manufacturing this product.

Furthermore we shall assume that the demand b_j for product π_j is known and not affected by type of row material used.

Let s_{ij} denote the cost of manufacturing the product π_j from row material P_i. If for some reason it is impossible to produce π_j from the row material P_i (e.g. the length of the requested rolled bar is greater than the length of the row bar) then we shall give s_{oj} the value ∞.

Apart from the manufacturing costs we have to take into account the cost of holding stock in the warehouse. This type of expence is usually difficult to estimate. However, this is usually small in comparison with production costs, so we can arrive at an approximate figure if we take the total cost of maintaining the warehouse during the period of production and divide it by the quantity of row materials held during this period. Then total cost of a unit π_j of P_i will be

$$c'_{ij} = s_{ij} + \sigma$$

where is the estimated cost of holding a unit of row material in the warehouse.

In cases where it is necessary one can take into account the expence d_i of ordering a fresh supply of row material P_i (cost of clerical work etc.).

Under some cercumstancies it may be more profitable to purchase row materials from an outside source by passing warehouse. Let h_j' denote unit cost of row source, which includs cost of row item cost of delivery and another possible articles and η_j' quantity bued. Then the following problem can be formulated:

Find x_{ij}' and η_j' minimizing total expencies expressed in functional:

$$\left[\sum c_{ij} x_{ij}' + \sum_i d_i \operatorname{sgn} \sum_j x_{ij}' + \sum h_j \eta_j'\right] \to \min \quad (1)$$

subject to

$$\sum x_{ij}' + \eta_j' = b_j \quad (2)$$

$$x_{ij}' \geq 0, \quad \eta_j' \geq 0, \quad i = 1, \dots, m; \; j = 1, \dots n; \quad (3)$$

i.e. demand must be satisfied and all variables are nonnegative.

For convenience let's make the following transformation of the variables:

let $x_{ij} = \dfrac{x_{ij}'}{b_j}$ and $\eta_j = \dfrac{\eta_j'}{b_j}$

and include a new variable

$$y_i = \operatorname{sgn} \sum_j x_{ij} = \begin{cases} 0 & \text{if } \sum_j x_{ij} = 0; \\ 1 & \text{if } \sum_j x_{ij} > 0; \end{cases}$$

Then the problem (1) - (3) will be

$$z = \left[\sum_{i,j} c_{ij} x_{ij} + \sum_i d_i y_i + \sum_j h_j \eta_j\right] \to \min \quad (4)$$

subject to

$$\sum_i x_{ij} + \eta_j = 1, \; j = 1, \dots, n; \quad (5)$$

$$y_i = \{0, 1\}, \; i = 1, \dots, m; \quad (6)$$

$$0 \leq x_{ij} \leq 1; \quad 0 \leq \eta_j \leq 1; \quad \begin{matrix} i = 1, \dots, m; \\ j = 1, \dots, n; \end{matrix} \quad (7)$$

where

$$c_{ij} = c'_{ij}\, b_j \;, \qquad \eta_j = \eta'_j\, b_j$$

Problem (4) - (7) is a programing problem with mixed variables. It may be effectively solved by the algorithms based on branch and bound techniques.

We shall consider in the next section a method for solving (4) - (7) which in fact is a generalisation of the algorithm given in (2) for the plant location problem.

§2. The algorithm for solving deterministic stock control problem.

The branch and bound algorithm with modifications for the problem (4) - (7) which will be described later, is an effective and finite method for solving combinatorial optimisation problems arising in the integer programing (3), shedul-ing theory (4), etc. which might considerably decrease the number of trials. During the last few years many papers concerned with this method and its applications were issued (1).

The branching prosses will be done as usual by integer variable y_i putting it in turn equal 0 and 1. Let's assume that it is possible to find the lower bound (estimate) of the functional (4) (way of calculating this estimate will be described later) in each vertex of the obtained decision tree, i.e. for each subset of the feasible solutions set, neglecting the integer nature of the vari-ables y_i . Then starting from the initial vertex of the decision tree $\mathcal{V}_0$ one can calculate the value of the lower bound $Z_0 \leq Z$ in this vertex. If all y_i are found to be integer then the problem is solved. In the case when some y_K is fractional putting $y_K = 0$ then $y_K = 1$ one generates two new branches ($\mathcal{V}_0, \mathcal{V}_1$) and ($\mathcal{V}_0, \mathcal{V}_2$) from the vertex $\mathcal{V}_0$ which terminates correspondingly in the new vertices $\mathcal{V}_1$ and $\mathcal{V}_2$. (If there are several fractional y_K , then an interesting problem concerned with the optimal strategy of choosing the next variable for branching arises). For each vertex obtained one calculates the lower bounds, which we denote z_1 and z_2 correspondingly. In the set of verteces $Q = \{v_1, v_2\}$ one can find the vertex with the minimal lower bound. Let it be v_1 for example, i.e.

$$z_1 = min\,(z_1, z_2), \quad z_1 \geqslant z.$$

Then branching goes on from vertex $v_1 \in Q$ when the next fractional variable, e.g. y_z ($z \neq k$) is set to 0 and to 1. As a result two new verteces v_3 and v_4 are generated for which the lower bounds z_3 and z_4 are calculated as earlier. Again in the verteces set $Q = \{v_2, v_3, v_4\}$ the vertex with the minimal lower bound is found, say v_3, i.e.

$$z_3 = min\,\{z_2, z_3, z_4\} \quad z_3 \geqslant z_1$$

From this vertex v_3 branching procedure goes on.

Prosses described continues until coming to the "terminal" vertex of the decision tree, or alternatively speaking until working out a solution with all integer y_i.

If value z obtained by functional (4) on this solution satisfies inequality

$$z \leq z_i \qquad \text{for all} \quad v_i \in Q$$

then this solution is optimal.

Before describing the procedure for calculating the estimates z_i notice that the optimal solution of the problem (4) - (7) will contain exactly n variables x_{ij} which will be equal to 1, and all the rest will be 0. It becomes obvious if one take into account that under the conditions of the problem (4) - (7) there are no restrictions on the quantity of products shipped. This means that the optimal solution for (4) - (7) must be found among x_{ij} equal to 0 or 1 ($i = 1, \ldots, m$; $j = 1, \ldots, n$).

Now let's consider the procedure for calculating lower bounds z_i in each vertex of the decision tree v_i. One should realise that this procedure will be repeated many times during the prosses of solving the problem. That means that it is desirable to have this procedure as simple as possible.

Let N_j denote the set of different types of row materials from which the product π_j may be manufactures. Let M_i denote the set of products which may be manufactured from the i-th type of row material P_i in stock and n_i number of entries M_i.

Now, for some vertex of the decision tree, let S_1 and S_0 denote the sets of y_i which, during the process of solution, were given the values 1 and 0 correspondingly and let S denote the set consisting of the remainder of y_i. Then the value of the lower bound can be calculated if one substitute in (4) the values of x_{ij}, η_j and y_i following:

$$x_{ij} = 1$$

$$\eta_j = 0, \quad \text{if} \min_{i \in S_1 \cup S} \left\{ [c_{ij} + \frac{g_i}{n_i}], \eta_j \right\} = [c_{kj} + \frac{g_k}{n_k}]; \tag{8}$$

$$x_{ij} = 0$$

$$\eta_j = 1, \quad \text{if} \min_{i \in S_1 \cup S} \left\{ [c_{ij} + \frac{g_i}{n_i}], \eta_j \right\} = \eta_j$$

in all other cases x_{ij} and η_j are equal 0.

$$y_i = \frac{1}{n_i} \sum_{j \in M_i} x_{ij}, \; i \in S, \tag{9}$$

where

$$g_i = \begin{cases} d_i, & \text{if } i \in S_0 \\ 0, & \text{if } i \in S_1 \end{cases}$$

It is obvious that expressions (8) - (9) define the optimal solution for the problem (4) - (7) without integer restriction on y_i, i.e.

$$0 \leq y_i \leq 1, \; i = 1, \ldots, m \tag{10}$$

and therefore the value of the functional (4) which obtained from it is the lower bound for the problem (4) - (7).

To prove this notice that for

$$\sum_{j\in M_i} x_{ij} \le n_i y_i$$

and for the optimal solution of problem (4) - (7), (10) we shall have equality, i.e.

$$\sum_{j\in M_i} x_{ij} = n_i y_i \quad \text{or} \quad \sum_{j\in M_i} \frac{x_{ij}}{n_i} = y_i$$

By substituting this value of y_i for $i \in S$ in (4) we shall have

$$\min Z = \sum_{i\in S_1} d_i + \min\left\{ \sum_{i\in S_1} c_{ij} x_{ij} + \sum_{i\in S} \left[c_{ij} + \frac{d_i}{n_i}\right] x_{ij} + \sum_j h_j \eta_j \right\}.$$

providing that

$$\sum_i x_{ij} + \eta_j = 1, \quad j = 1,\dots,n$$

§3. Formulation of Stock Control problem under probabilistic demand.

In the majority of practical applications the vector of demand b_i is not known in advance. We may know the probability distribution $F(b_1,\dots,b_n)$ for the vector b or sometimes only limits in which demand may change. In this case the problem (1) - (3) becomes meaningless and one must define what the solution of the problem is.

There are a number of possible approaches to the statement of the stochastic programing problems given in (5). One of the possibilities consists in finding the x_{ij} which gives minimum to the functional (1) subject to *):

$$\sum x_{ij} + \eta_j \geqslant b_j \ (q) \qquad \text{for all } q \in Q \qquad (2)$$

$$x_{ij} \geqslant 0, \ \eta_j \geqslant 0 \qquad (3)$$

where q is a random variable or set of random parameters the sample of which specifies implementation of random entries in the problem, and Q is set of q values q appeared with non-zero probability.

This type of problems usually called "rigid" or one-stage stochastic programming problems becouse this problems assumed to be solved once and any amendments of the accepted solution not allowed even if one have got some additional data about the environment.

It seems more reasonable to take two-stage stochastic programming problem. In this kind of the problems one might imagine the process of decision making in two stages.

Firstly one can choose some solution X, not necessary satisfing all constraints for all possible samples of q. Then some sample of q and therefore implementation of the vector b (q) are fixed up and the vector η which should amend the accepted decision X is included.

Assuming that we undergo by additional expense for amending the decision made on the first stage it is appropriate to state the following problem: minimize the expected value of the total expense, i.e. cost of the solution taken on the first stage plus additional cost of amendments on the second stage. Let's consider the process of decision making in the problem (1) - (3) as two-stages. The plan X is defined during the first stage. Then when the vector b becomes known and there is not enough row materials the penalty proportional to outstanding quantity

$\eta_j = b_j - \sum_i x_{ij}$ is paid (one might consider this penalty as a cost of urgent purchase of row

*) From now we shall consider x_{ij} and η_j as actual (not relative) values which in §1 was denoted as x'_{ij} and η'_j .

materials from the more expensive outside source). Then one can formulate the problem of minimizing the expected values of the total expence

$$M[\sum_{ij} c_{ij} x_{ij} + \sum_i d_i y_i + \sum_j h_j \eta_j] \to \min \quad (11)$$

subject to

$$\sum x_{ij} + \eta_j = b_j \quad (12)$$

$$x_{ij} \geq 0, \ \eta_j \geq 0 \quad (13)$$

where η_j is unsatisfied demand, h_j penalty for unit. Notice, that if h_j grows to infinitude the two-stage problem (11) - (13) becomes one-stage problem (1), (2'), (3).

In order to find the deterministic equivalent (i.e. the problem of mathematical programming which solution is also solution of the corresponding stochastic programming problem) for the problem (11) - (13) let's fix up some feasible solution X and consider (11) - (13) as the minimizing problem with variable η. This one may be written as follows:

$$\sum h_j \eta_j \to \min$$

subject to

$$\sum_i x_{ij} + \eta_j \geq b_j \quad \text{or} \quad \eta_j \geq b_j - \sum_i x_{ij}$$

$$\eta_j \geq 0$$

becouse the rest entries of (11) not affected by η.
One can easily find that the dual problem of maximizing is follows

$$\sum_j (b_j - \sum x_{ij}) \xi_j \to \max$$

subject to

$$\xi_j \leq h_j$$

$$\xi_j \geq 0$$

It is obvious that the solution of the last problem is given by the equalities

$$\xi_j = h_j$$

Therefore by wellknown duality theorem

$$\min_{\eta} \sum_j h_j \eta_j = \max_{\xi} \sum_j (b_j - \sum_i x_{ij}) \xi_j = \\ = \sum (b_j - \sum_i x_{ij}) h_j$$

Substituting the right part of the last equality in the functional (11) one get the following deterministic equivalent for the problem (11) - (13):
minimize functional (14)

$$\Big[\sum_{ij} c_{ij} x_{ij} + \sum_i d_i y_i + \sum_j h_j M[b_j - \sum_i x_{ij}]\Big] \to \min \quad (14)$$

subject to

$$x_{ij} \geqslant 0 \quad (15)$$

$$y_i = \{0, 1\} \quad (16)$$

As some of the variables in the problem (14) - (16) are integer this problem is the mixed programming problem again which might be solved by the algorithm based on branch and bound techniques described in the next section.

§4. Algorithm for solving problem (14)-(16).

Method for solving problem (14) - (16) is based on the branch and bound techniques. The branching procedure done by the integer variable y_i putting it in turn equal to 0 and 1 as described earlier.

Let's now consider the problems of estimating the lower bounds in this problem.

First of all let's pay attention to some features of the "trancated" mathematical expectation $M[b_j - \sum_i x_{ij}]$ for $\sum_i x_{ij} \leq b_j$, which will be usefull for calculating the estimates.

Consider the behavior of the function $F(x) = M[b - x]$ where b is a random variable with known probability density $\varphi(b)$ for $x \leq b$.

One can easily see that

$$F(x) = \int_x^{\infty} (b - x)\,\varphi(b)\,db$$

Using this expression it is easy to calculate the first derivative $F'(x)$

$$F'(x) = \int_0^x \varphi(b)\,db - 1$$

Similary, calculating the second derivative

$$F''(x) = \varphi(x) \geqslant 0$$

one can easily see that the function $F(x)$ is convex. Hence the function

$$F\Big(\sum_i x_{ij}\Big) = M\Big[b_j - \sum_i x_{ij}\Big]$$

is a convex function of all variables.

Now consider the problem of minimizing the functional

$$\Phi(x) = \sum_i k_i x_i + F\Big(\sum_i x_i\Big) \to \min$$

subject to $x_i \geqslant 0$, and assuming that $F(\sum_i x_i)$ is convex and all $k_i > 0$. Also we assume the existence of the continious derivative of $F(\sum_i x_i)$. The assumptions of a positive k_i and convex $F(\sum_i x_i)$ provides that the minimum exists and attained inside the cone $x_i \geqslant 0$ or on the bound.

Let's assume that minimum attained inside the cone. Then becouse the derivative is continious the coordinates of the minimum point must satisfy the system of the equalities

$$\frac{\partial \Phi}{\partial x_i} = 0$$

or write it in more details

$$k_i + F_z'\Big(\sum_i x_i\Big) = 0, \quad z = \sum_i x_i$$

It is obvious that if $K_i \neq K_j$ for $i \neq j$ this system is inconsistent, moreover, there are no two equations from this system which are compatible except case $K_i = K_j$ for $i \neq j$.

This contradiction proves that the minimum can be attained only on the bound of the cone $x_i \geqslant 0$. Then one can easily find out that the minimum is attained on a vector all components of which are zero, except one. The index of this coordinate may be defined by choosing minimal k_{i0}, i.e.

$$K_{i0} = \min_i K_i$$

and its value is a root of the equation

$$k_{i0} + F'(x_i\, 0) = 0 \qquad (\bullet)$$

In the case when there are several minimal K_i (which are all equal of course) if makes no difference what quantity of each is taken, provided their sum is equal to the quantity defined by equation (•).

Another point is that, obviously if one allowes y_i to be a continious variable between 0 and 1, then the problem of minimizing (14) might be split up on the n subproblems for each index j , and each such subproblem is a problem of minimizing the functional similar to $\Phi(x)$

All points mentioned above allows us to construct an algorithm for calculating the lower bounds of functional (14) in some vertex V_i of the decision tree.

Let S_0 and S_1 denote the sets of y_i to which during the process of solution were attached values 0 and 1 correspondingly, and let S be the set of the remaining y_i . Let n_i denote number kinds of products $\mathcal{T}_j$ which may be produced from i-th kind of row material D_i .

The value of the lower bound may be found if one substitute in (14) following values of x_{ij} and y_i . Let $C_{i_0 j} = \min_i C_{ij}$, then

$$\begin{aligned} x_{ij} &= 0 \quad \text{if} \quad i \in S_0 \\ x_{ij} &= 0 \quad \text{if} \quad i \neq i_0 \\ x_{ij} &= x_{ij}^{(0)} \quad \text{if} \quad i = i_0 \end{aligned} \qquad (17)$$

where $x_{ij}^{(0)}$ is a root of the equation

$$\frac{\partial}{\partial x_{ij}}\left[c_{ij}x_{ij}+h_j M\left[b_j-x_{ij}\right]\right]=0$$

If one remember the expression for the derivative of the "trancated" mathematical expectation, it is obvious that $x_{ij}^{(0)}$ is a root of the equation

$$\int_0^{x_{ij}} \varphi(b_j)\,db_j = 1-\frac{c_{ij}}{h_j}$$

Now

$$y_i = 0 \quad \text{if} \quad i \in S_0$$

$$y_i = 1 \quad \text{if} \quad i \in S_1 \tag{18}$$

$$y_i = \frac{\sum_j \operatorname{sgn} x_{ij}}{n_i} \quad \text{if} \quad i \in S$$

The values (17) - (18) of x_{ij} and y_i gives the solution for the minimizing problem (14) - (15) if $0 \le y_i \le 1$, and therefore the value of the functional attained by it is a lower bound for the problem (14) - (16) in some vertex v_i of the decision tree.

References

1. Lowler E.L., Wood D.E., "Branch-and-Bound Methods. A Survey". Opns. Res. v. 14, No.4, 1966.
2. Efroymson M.A., Ray T.L., "A Branch and Bound Algorithm for Plant Location". Opns. Res. v. 14, No.3, 1966.
3. Balas E., "An Additive Algorithm for Solving Linear Programs with Zero-One Variable". Opns. Res. v. 13, No. 14, 1965.
4. Ignall Edward, Schrage Linis. "Applications of the Branch-and-Bound Technique to some Flow-Shop Sheduling Problems". Opns. Res. v.13, No.3, 1965.
5. Golshtain E.G., Yudin D.B., "New Directions in The Linear Programming". "Sovetskoe Radio", ch. 6, 1966.

35:3

RESOURCE ALLOCATION IN MULTI-PROJECT BASED ON AGGREGATION OF THE PROJECT NETWORKS

Oleg G. Tchebotarev
Central Economic Mathematical Institute
(Moskow, USSR)

The problem of PERT/CPM system design and the problem of limited resource allocation among sevral projects activities is very complicated but it is of the great practical importance at the same time.

The paper is concerned with a technique for the solution of resource allocation problem in multi-project when the functions of the activity rate is a linear one.

I. Basic Definitions and Concepts

The activity is a process described by the following equation:

$$v_i(t) = \frac{dx_i(t)}{dt} = f_i\left[r_{ij}(t), t\right], \tag{1}$$

where $x_i(t)$ is the state of the i-th activity at the moment t; $x_i(t_{in})=0$, $x_i(t_{fin})=W_i$ (t_{in} and t_{fin} is the initial and the final moment of an activity correspondingly, W_i is the volume of the i-th activity); $v_i(t)$ is the rate of the i-th activity at the moment t; $r_{ij}(t)$ are the resource group parameters of the i-th activity, $j=1,2,..,m$ (If the j-th type of resources doesn't take part in the i-th activity the parameter $r_{ij}(t)=0$).

Realization of the activity is the variation of its state from some initial value $x_i(t_{in})=0$ up to the final one $x_i(t_{fin})=W_i$.

$\vec{r}_i(t) = \{r_{ij}(t)\}$ is the m-vector of the i-th activity resource group.

Let the modulus of the vector $\vec{r}_i(t)$ depends on the value of t only. So vector $\vec{\tau}(t)$ can be represented as $\vec{\tau}_i(t)=\rho_i(t)\vec{\alpha}_i$, where $\rho_i(t)$ is the power of the i-th activity resource group at the moment t, $\vec{\alpha}_i=\{\alpha_{ij}\}$ is the m- vector of resource group parameters for the i-th activity and resource group parameter α_{ij} is the value of the j-th resource when $\rho_i(t)=1$.

Project is a partially ordered set which consists of the final number of the activities. We can represent the project as a network. Realization of the project is the variation of its state $x(t)=[x_1(t),\dots,x_n(t)]$ from some initial value $x(0)=0$ up to the final one $x(T)=W$, $W=(W_1,W_2,\dots,W_n)$, where W is the volume of the project and T is the final moment of the project. The project is finished if all its activities are finished.

Multi-project is a set independent projects, which must be realized by common resources.

Resources constraints. There are two types of resources constraints

$$\sum_{p=1}^{\ell}\sum_{i=1}^{n_p}\alpha_{ij}\rho_i(t)\le N_j(t),\qquad j=1,2,\dots,m,\quad (2)$$

or $$\sum_{p=1}^{\ell}\sum_{i=1}^{n_p}\alpha_{ij}q_i\le S_j,\qquad j=1,2,\dots,m,\quad (3)$$

where $q_i=\int_0^T\rho_i(t)dt$ is the power consumption of the i-th activity $(i=1,2,\dots,n_p)$, S_j is the permissible consumption of the j-th resource in the multi-project, N_j is the given value of the j-th resource in the multi-project at the moment t, i is the number of the project activity $(i=1,2,\dots,n_p)$, p is the number of the project $(p=1,2,\dots,\ell)$.

If there are constraints of the type (2) only, a resources are considered, as a power type resources.

If there are constraints of the type (3) only, a resources are considered as a cost type resources.

The project aggregation is the representation of a project network by one activity. The values W, $\vec{\alpha}$ and function

$f[\rho(t),t]$ are defined for this activity using the given values of W_i, $\vec{\alpha}_i$ and functions $v_i(t) = f_i[\rho_i(t),t]$ for each project activity.

The definitions and concepts introduced are basically analogous to those ones used by V.Burkov and A.Lerner[1].

2. Resource Allocation Problem in Multi-Project

Let the multi-project consisting of ℓ projects each having the volume $W_1, W_2, \dots, W_\ell$ must be realized under definite constraints in resources. The problem is to allocate resources to projects activities so that to minimize a criterion under constraints (2) and/or (3).

Several functions may be used as the criterion, for example $\max_p T_p$ and $\sum_{p=1}^{\ell} \varphi_p(T_p)$, where T_p is the final moment of the p-th project and $\varphi_p(T_p)$ is a non-decreasing function of the T_p . $\varphi_p(T_p)$ is a penalty function for the delay of the p -th project (for example, $\varphi_p(T_p)=0$, if $T_p \le T'_p$ and $\varphi_p(T_p) = a_p(T_p - T'_p)$, if $T_p > T'_p$, where T'_p is the given final moment the p-th project and a_p is a constant).

The aggregation of the project networks permits to obtain the solution of the problem as a sequence of the following actions.

1. Aggregate the project networks that is given the values of $W_i, \vec{\alpha}_i$ and functions $v_i(t) = f_i[\rho_i(t),t]$ for each activity define the values of W_p, $\vec{\alpha}_p$ and functions $v_p(t) = f_p[\rho_p(t),t]$.

2. When aggregation is completed solves the resource allocation problem with ℓ independent activities. The result of this step is the values of $N_{pj}(t)$ and S_{pj} for each project.

3. Using the values found at the previouse step solves the allocation problem for each separate project.

Thus the problem of resource allocation with $n = \sum_{p=1}^{\ell} n_p$ activities is transformed to ℓ resource allocation problems with n_p activities. As an example of such the appoach consider the solution of the resource allocation problem in the multi--project with resources of cost type and with the criterion

$\max_p T_p \to \min$. It is assumed that each activity in the network is subject to a continuous upward - concave time - cost relationship.

In this case the functions $S_p(T_p)$ may be found by means of Berman's algorithm[2]. Because each $S_p(T_p)$ is the non-decreasing function of T_p, all projects have the same final moment that is $T_1 = T_2 = \cdots = T_\ell = T$, if the final moment of the multi-project is minimized[3]. Thus the value of T may be found from the equation

$$S_1(T) + S_2(T) + \cdots + S_\ell(T) = S.$$

ℓ separate the resource allocation problems may be solved by means of Berman's algorithm too[2].

3. Aggregation of the project network

The project aggregation process consist of the following steps:

1. Ordering of the project states.
2. Determination of the permissible vector of resource group parameters for the project.
3. Determination of the function $f_p[\rho_p(t), t]$, $(p = 1, 2, \dots, \ell)$.

I step. Each activity is the ordered sequence of its states that is $x_i^1 < x_i^2$ or $x_i^1 > x_i^2$ is valide for any two states: x_i^1 and x_i^2. The same condition must be valide for the project states if the project is transformed to single activity. One may easily show that to make ordering of the project states it is nesessory and sufficient:

1) to order the events of the project network
2) to find the values of y_{is} corresponding to those parts of the activities volumes, which must take place within the s-th interval (within the interval from the $(s-1)$-th to s-th event, $s = 1, 2, \dots q$). The values of y_{is} must satisfy to condition $\sum_{s \in Q_i} y_{is} = W_i$, where Q_i is the set of the intervals, within which the i-th activity takes place.

3) to define the policy of the activity realization within each interval. This may be done by means of parametric equations $x_{is} = \Psi_{is}(\beta_s)$, where $i \in R_s$ (R_s is the set of activities, which may take place in the s-th interval), β_s is a parameter, which has the same value for each activity of the set R_s, $(0 \le \beta_s \le 1)$, Ψ_{is} are continuous non-decreasing functions of the β_s; $\Psi_{is}(0) = 0$ and $\Psi_{is}(1) = y_{is}$.

II step. Suppose that the values of y_{is}, $\vec{\alpha}_i$ and functions $v_i(t) = f_i[\rho_i(t), t]$ are given and the values of α_j should be found.

Note. We shall suppose hereafter that the functions $f_i[\rho_i(t), t]$ are

$$f_i(\rho_i) = \rho_i. \tag{4}$$

Consider the case of the simultaneous realization of the activity parts y_{is} within each interval $s \in G_i$. The condition of simultaneity may be written as

$$f_i(\rho_{is}) = \beta_s y_{is}, \qquad 0 \le \beta_s \le 1, \tag{5}$$

where ρ_{is} is the power of the i-th activity resource group in the s-th interval. From (4) and (5) one has: $\rho_{is} = \beta_s y_{is}$. Let $\sum_{i \in R_s} \beta_s y_{is} \le \rho$ if the power ρ of the project is given. Then the minimum value of the Δ_s (where Δ_s is the value of the s-th interval) is equal to $1/\beta_s$ that is

$$\Delta_s(\rho) = \frac{1}{\rho} \sum_{i \in R_s} y_{is}.$$

The value of the j-th resources required in the s-th interval is

$$N_{js}(\rho) = \sum_{i \in R_s} \alpha_{ij} \rho_{is} = \rho \, \frac{\sum_{i \in R_s} \alpha_{ij} y_{is}}{\sum_{i \in R_s} y_{is}}.$$

If the project power of the resource group has the given value ρ , the value of the j-th resources required for realization of the project is equal to the maximum value of $N_{js}(\rho)$, that is

$$N_j = \max_s \rho \frac{\sum_{i\in R_s} \alpha_{ij} y_{is}}{\sum_{i\in R_s} y_{is}}.$$

Because it is defined that $N_j(\rho)=\alpha_j \rho$ one has

$$\alpha_j = \max_s \frac{\sum_{i\in R_s} \alpha_{ij} y_{is}}{\sum_{i\in R_s} y_{is}}.$$

III step. If $\Delta_s(\rho)$ and α_j are known, the rate of the project is

$$f(\rho) = \frac{W}{T(\rho)} = \frac{W}{\sum_{s=1}^{g} \Delta_s(\rho)}.$$

4. The aggregation of the project network with linear functions of the activity rates

Let each activity has a linear power-rate relationship, that is $f_i(\rho_i) = \rho_i$ and let the network events are ordered.

We shall consider the case of serial activities realization within each interval. In that case the s-th interval is splitted into several subintervals (the number of the subinterval is equal to the number of the activities $i \in R_s$) under the condition $\sum_{i\in R_s} \Delta_{is} = \Delta_s$. One has $\Delta_{is} = y_{is}/\rho$ and $\Delta_s = \frac{1}{\rho}\sum_{i\in R_s} y_{is}$. The project time is

$$T = \sum_{s=1}^{g} \Delta_s = \frac{1}{\rho}\sum_{s=1}^{g}\sum_{i\in R_s} y_{is} = \frac{1}{\rho}\sum_{i=1}^{n}\sum_{s\in Q_i} y_{is} = \frac{1}{\rho}\sum_{i=1}^{n} W_i.$$

The maximum value of the j-th resource taking part in the project is

$N_j = \max_s \max_{i\in R_s} \alpha_{ij}\rho = \rho \max_i \alpha_{ij}$.

Taking into account that $N(\rho) = \alpha_j \rho$, one has $\alpha_j = \max_i \alpha_{ij}$.

As each component of the vector $\vec{\alpha}$ is the maximum of the corresponding components of vectors $\vec{\alpha}_i$, the use of resources is irregular (if corresponding components of $\vec{\alpha}_i$ is not equal). If the j-th resource takes part in the i-th activity the value of the resource used within each subinterval $\Delta_{is} \in \theta_i$ is $\alpha_{ij}\rho$. However the value of the resource used may be $\alpha_j\rho$ if α_j and ρ are given. Thus we can define the values of the j-th resource standing idle during the time W_i/ρ of the i-th activity realization. This value is equal to $(\alpha_j - \alpha_{ij})\rho$. The value $(\alpha_j - \alpha_{ij})W_i$ will be called by the loss of the j-th resource for the i-th activity. The relative loss of the j-th resource for the project is

$$\delta_j = \frac{\sum_{i=1}^{n}(\alpha_j - \alpha_{ij})W_i}{\sum_{i=1}^{n}\alpha_j W_i} = 1 - \frac{\sum_{i=1}^{n}\alpha_{ij} W_i}{\alpha_j W}. \tag{8}$$

Because the parameters α_{ij} defines the correlation of different resources only one may take $\alpha'_{ij} = \alpha_{ij}\mu_i$ as parameters. (In this case the function of an activity rate is $f_i(\rho_i) = \mu_i \rho_i$. After substitution the equation (8) takes the form:

$$\delta_j = 1 - \min_i \frac{\sum_{i=1}^{n}\alpha_{ij} W_i}{\alpha_{ij}\mu_i \sum_{i=1}^{n} W_i/\mu_i}.$$

The the problem arises to define μ_i so that

$$\max_j \delta_j \to \min. \tag{9}$$

The optimal solution of this problem is obviously the same as the optimal solution of the problem

$$\min_j \; \min_i \frac{\sum_{i=1}^{n}\alpha_{ij} W_i}{\alpha_{ij}\mu_i \sum_{i=1}^{n} W_i/\mu_i} \to \max. \tag{10}$$

It is evident that the multiplying μ_i by a constant doesn't change the (10). So one may suppose that $\sum_{i=1}^{n} W_i/\mu_i = 1$. Denote $\eta_i = W_i/\mu_i$ and $\sigma_{ij} = \frac{1}{\alpha_{ij}W_i}\sum_{i=1}^{n}\alpha_{ij}W_i$.

Thus we have the problem:

$$\min_j \; \min_i \sigma_{ij}\,\eta_i \to \max,$$

under constrain $\sum_{i=1}^{n}\eta_i = 1$.

Having changed the order of minimizing we come to the problem $\min_i \lambda_i \eta_i \to \max$ under constrain $\sum_{i=1}^n \eta_i = 1$, where $\lambda_i = \min_j \sigma_{ij}$.
The value of $\lambda_i \eta_i$ remains constant for each activity at the optimal solution.
Denote $z = \lambda_i \eta_i$. We have $\eta_i = z_i / \lambda_i$, $\sum_{i=1}^n \eta_i = z_i \sum_{i=1}^n 1/\lambda_i = 1$, $z = \left(\sum_{i=1}^n 1/\lambda_i\right)^{-1}$, $\eta_i = \left(\lambda_i \sum_{i=1}^n 1/\lambda_i\right)^{-1}$, and $\mu_i = W_i/\eta_i = W_i \lambda_i \sum_{i=1}^n 1/\lambda_i$ that is

$$\mu_i = W_i \left(\min_j \frac{\sum_{i=1}^n \alpha_{ij} W_i}{\alpha_{ij} W_i}\right) \cdot \sum_{i=1}^n \max_j \frac{\alpha_{ij} W_i}{\sum_{i=1}^n \alpha_{ij} W_i}.$$

The relative loss of a resource doesn't exceed the value

$$\delta = 1 - z = 1 - \left(\sum_{i=1}^n \max_j \frac{\alpha_{ij} W_i}{\sum_{i=1}^n \alpha_{ij} W_i}\right)^{-1}. \tag{11}$$

The parameters of the project resource group are

$$\alpha_j = \max_i \alpha_{ij} W_i \left(\min_j \frac{\sum_{i=1}^n \alpha_{ij} W_i}{\alpha_{ij} W_i}\right) \sum_{i=1}^n \max_j \frac{\alpha_{ij} W_i}{\sum_{i=1}^n \alpha_{ij} W_i}.$$

The duration of the project is

$$T(\rho) = \sum_{i=1}^n W_i/\rho\mu_i = \frac{1}{\rho} \sum_{i=1}^n \eta_i = \frac{1}{\rho}.$$

Consider the case of simultaneous realization of activities $i \in R_s$ within each interval s.

The condition of simultaneity and values of Δ_s and α_j are defined by relationships (5),(6) and (7). In this case the relative loss of the j-th resource of the project is

$$\delta_j = 1 - \frac{\sum_{s=1}^{S} \sum_{i \in R_s} \alpha_{ij} y_{is}}{\left(\max_s \frac{\sum_{i \in R_s} \alpha_{ij} y_{is}}{\sum_{i \in R_s} y_{is}}\right) \sum_{s=1}^{S} \sum_{i \in R_s} y_{is}}. \tag{12}$$

Like in the previous case, we takes as a parameter of the resource group $\alpha'_{ij} = \alpha_{ij} \mu_i$ substituting ρ_i by $\rho'_i = \mu_i \rho_i$

at the equation (4). After that equations (4),(6),(7) and (12) are replaced by

$$f_i(\rho_i) = \rho_i \mu_i, \tag{4a}$$

$$\Delta_s(\rho) = \frac{1}{\rho} \sum_{i \in R_s} \tilde{y}_{is}, \tag{6a}$$

$$\alpha_j = \max_s \frac{z_{js}}{\sum_{i \in R_s} \tilde{y}_{is}}, \tag{7a}$$

$$\delta_j = 1 - \frac{A_j}{\left(\max_s \frac{z_{js}}{\sum_{i \in R_s} \tilde{y}_{is}}\right) \sum_{s=1}^{g} \sum_{i \in R_s} \tilde{y}_{is}}, \tag{12a}$$

where $\tilde{y}_{is} = y_{is}/\mu_i$, $z_{js} = \sum_{i \in R_s} \alpha_{ij} y_{is}$ and $A_j = \sum_{s=1}^{g} z_{js}$.

Consider the problem: to find μ_i so that $\max_j \delta_j \to \min$ or

$$\min_j \min_s \frac{A_j \sum_{i \in R_s} \tilde{y}_{is}}{z_{js} \sum_{i=1}^{n} W_i/\mu_i} \longrightarrow \max . \tag{13}$$

Because the value of (13) doesn't change when each μ_i is multiplied by the same constant one may assume that $\sum_{i=1}^{n} \eta_i = 1$, where $\eta_i = W_i/\mu_i$.

The problem (13) is equivalent to the problem

$$\min_s \left(\min_j \frac{A_j}{z_{js}}\right) \sum_{i \in R_s} \frac{y_{is}\eta_i}{W_i} \longrightarrow \max \quad \text{under } \sum_{i=1}^{n} \eta_i = 1. \tag{14}$$

Denote $B_s = \min_j \frac{A_j}{z_{js}}$, $\sigma_{is} = \frac{y_{is}}{W_i}$ and write (14) in the form

$$\min_s B_s \sum_{i \in R_s} \sigma_{is} \eta_i \to \max \qquad \text{under } \sum_{i=1}^{n} \eta_i = 1.$$

Let $C_{is} = B_s \sigma_{is}$. So the problem is

$$\min_s \sum_{i \in R_s} C_{is} \eta_i \to \max \qquad \text{under } \sum_{i=1}^{n} \eta_i = 1.$$

Denote $\min_s \sum_{i \in R_s} C_{is} \eta_i = \nu$. We have the next problem: to maximize $\Phi = \nu$ under constrains

$$\begin{cases} \sum_{i \in R_s} C_{is} \eta_i = \nu, \\ \sum_{i=1}^{n} \eta_i = 1. \end{cases} \tag{15}$$

Constrains (15) are equivalent to

$$\begin{cases} \nu - \sum_{i\in R_s} C_{is}\eta_i \le 0, \\ \sum_{i=1}^{n} \eta_i = 1. \end{cases} \tag{16}$$

Transfer the problem to a dual one. Let $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_g$ and ξ are new variables (ξ correspond to the second constrain). We have the next problem: to minimize ξ under constrains

$$-\sum_{s\in Q_i} C_{is}\varepsilon_s + \xi \ge 0 \tag{17}$$

and $$\sum_{s=1}^{g} \varepsilon_s \ge 1. \tag{18}$$

The constrain (17) may be replaced by equation

$$\xi = \max_i \sum_{s\in Q_i} C_{is}\varepsilon_s$$

and constrain (18) may be replaced by equation

$$\sum_{s=1}^{g} \varepsilon_g = 1. \tag{18a}$$

So the dual problem is

$$\max_i \sum_{s\in Q_i} C_{is}\varepsilon_s \longrightarrow \min \qquad \text{under } \sum_{s=1}^{g} \varepsilon_s = 1.$$

Take ν, η_i as basic variables of the activities that may realized within a single interval only. Denote the set of those activities by H (remark that $|H| = g$).

The basic solution is the one of the next problem:

$$\min_{i\in H} C_{is}\eta_i \longrightarrow \max \qquad \text{under } \sum_{i\in H} \eta_i = 1.$$

The all values $C_{is}\eta_i$ is readily seen to be the same in the optimal solution of the problem. So the basic solution is

$$\eta_i = \begin{cases} \left(C_{is}\sum_{i\in H}\frac{1}{C_{is}}\right)^{-1} & \text{for } i \in H, \\ 0 & \text{for } i \notin H. \end{cases} \tag{19}$$

In this case $\nu = \left(\sum_{i\in H}\frac{1}{C_{is}}\right)^{-1}$.

Prove that the solution (19) is optimal.

For this purpose define the solution of the dual problem.

It follows from the relation of duality that the strict equality in equation (17) is true for the activities belonged to set H. Taking into account that each activity from set H may be realized within a single interval only, one has

$C_{is}\varepsilon_s = \xi$ for $i \in H$. So

$$\varepsilon_s = \frac{\xi}{C_{is}}. \tag{20}$$

After substituting (20) in (18a) we have

$$\xi = \left(\sum_{s=1}^{s} \frac{1}{C_{is}}\right)^{-1} = \left(\sum_{i\in H} \frac{1}{C_{is}}\right)^{-1} = \nu. \tag{21}$$

Since the solutions of the direct and dual problems are the same, the solution (19) is optimal.

Let η_s is the η_i of the activity which realizing within the s-th interval. Correspondingly C_s is the C_{is} of the same activity. We have

$$C_s = \min_j \frac{\sum_{s=1}^{s}\sum_{i\in R_s}\alpha_{ij} y_{is}}{\sum_{i\in R_s}\alpha_{ij} y_{is}} = \min_j \frac{\sum_{i=1}^{n}\alpha_{ij} W_i}{\sum_{i\in R_s}\alpha_{ij} y_{is}};$$

$$\eta_s = \left(\sum_{s=1}^{s} \max_j \frac{\sum_{i\in R_s}\alpha_{ij} y_{is}}{\sum_{i=1}^{n}\alpha_{ij} W_i}\right)^{-1} \cdot \max_j \frac{\sum_{i\in R_s}\alpha_{ij} y_{is}}{\sum_{i=1}^{n}\alpha_{ij} W_i};$$

$$\alpha_j = \left(\max_s \frac{\sum_{i\in R_s}\alpha_{ij} y_{is}}{\max_j \frac{\sum_{i\in R_s}\alpha_{ij} y_{is}}{\sum_{i=1}^{n}\alpha_{ij} W_i}}\right) \cdot \sum_{s=1}^{s} \max_j \frac{\sum_{i\in R_s}\alpha_{ij} y_{is}}{\sum_{i=1}^{n}\alpha_{ij} W_i};$$

$$T(\rho) = \frac{1}{\rho}\sum_{s=1}^{s}\sum_{i\in R_s} \frac{y_{is}\eta_i}{W_i} = \frac{1}{\rho}\sum_{s=1}^{s}\eta_s = \frac{1}{\rho}.$$

Suppose that $W=1$ for the project. So we have $f(\rho)=\rho$

for the project. The relative loss for a resource doesn't exceed

$$\delta = 1 - \nu = 1 - \left(\sum_{s=1}^{g} \max_{j} \frac{\sum_{i \in R_s} \alpha_{ij} y_{is}}{\sum_{i=1}^{n} \alpha_{ij} W_i} \right)^{-1}.$$

The results to be found here permit us to state the next problem of the optimal aggregation: define y_{is} so that $\sum_{s=1}^{g} \max_{j} \frac{1}{D_j} \sum_{i \in R_s} \alpha_{ij} y_{is} \longrightarrow \max$ under constrain $\sum_{s \in Q_i} y_{is} = W_i$, where $D_j = \sum_{i=1}^{n} \alpha_{ij} W_i$.

In conclusion let's compare the value of the relative losses under the serial and simultaneous realization within each interval. For this purpose consider permissible solution in case of the simultaneous realization of the activities: $y_{is} = W_i$, if the activity is realized within the s-th interval and $y_{is} = 0$ other wise. In such a case the equation of relative losses is

$$\delta_{sim.} = 1 - \left(\sum_{s=1}^{g} \max_{j} \frac{\sum_{i \in R_s^*} \alpha_{ij} W_i}{\sum_{i=1}^{n} \alpha_{ij} W_i} \right)^{-1}, \quad \text{where } R_s^* = \{ i : y_{is} > 0 \}.$$

The equation of relative losses in the case of series realization of activities may be represented in the next form

$$\delta_{ser.} = 1 - \left(\sum_{s=1}^{g} \sum_{i \in R_s^*} \max_{j} \frac{\alpha_{ij} W_i}{\sum_{i=1}^{n} \alpha_{ij} W_i} \right)^{-1}.$$

So for as the maximum of the sum is larger than the sum of maxima, $\delta_{sim} \leq \delta_{ser.}$ Consequently, the case of simultaneous realization is not worse than the case of serial realization in the sense of the value of the relative resources loss. (If the values of y_{is} are optimal).

References

1. Бурков В.Н., Лернер А.Я. Новые задачи теории сетевого планирования. В сб. "Вопросы управления большими системами", 1968.
2. Berman E.B. Resource allocation in a PERT network under continuous activity time-cost functions. Management Science, Vol. 10, No 4, (July 1964), 734-745.
3. Бурков В.Н. Распределение ресурсов как задача оптимального быстродействия. Автоматика и телемеханика, № 7, 1966.

35.4

OPTIMAL PROJECT CONTROL

V.N.Burkov

Institute of Automation and Remote Control

Moscow

USSR

1. Statement of the problem.

The problems of Resource allocation in the project were set and classified Comparatively not long ago [1]. At the moment resources allocation problem is one of the main part of the network Planning and Control Theory (NPC-Theory).

A minimum-time resource allocation problem is considered in the paper. Denote

$w_i(t)$ the speed of the i-th activity at the moment t,
$v_i(t)$ the value of the i-th activity resources at the moment, t,
W_i the volume of the i-th activity,
$N(t)$ the general level of the resources at the moment t (given function),
$s_i(t)$ the i-th activity cost at the moment t,
$S(t)$ the project cost at the moment t (given function);

The following relations take place:

$$w_i(t) = f_i[v_i(t)], \tag{1}$$

$$W_i = \int_0^T f_i[v_i(t)]\,dt, \tag{2}$$

$$\sum_1^n v_i(t) \le N(t) \tag{3}$$

$$s_i(t) = \int_0^t v_i(\tau)\,d\tau, \tag{4}$$

$$\sum_1^n s_i(t) \le S(t), \tag{5}$$

$$v_i(t) \ge 0,\ W_i > 0,\ N(t) \ge 0,\ S(t) \ge 0,\ i = 1,\dots,n,\ t \in (0,T), \tag{6}$$

where f_i - nondecreasing function of v_i $(v_i \ge 0)$, $f_i(0) = 0$, T - the project time.

The problem 1. Find $v_i(t)$ satisfying (3) so that the project time is minimal.

The problem 2. Find $v_i(t)$ satisfying (5) so that the project time is minimal.

The problem 3. Find $S_i(T)$ under the Condition

$\sum_1^n s_i(T) \le S(T)$, so that the project time is minimal.

2. Solution of problem 1.

Let $N(t) = N$ = const. Suppose the events of the network are ordered; that is the moment of the s-th event t_s is less than the moment of the K-th event t_K if $s < K$. Denote

R_s - the set of activities which can be done at the S-th interval that is at the interval (t_{s-1}, t_s), $s = 1,2,...,m$, where $(m+1)$ is the number of the events.

Q_i - the set of the intervals where the i-th activity can be done $i = 1,2,..., n$,

x_{is} - the volume of the i-th activity done at the S-th interval, $i \in R_s$, $s = 1,2,...,m$.

$\Delta_s(z_s)$ - the minimal duration of the S-th interval ($z_s = \{x_{is} : i \in R_s\}$), $x_{is} \geq 0$, $S = 1,2,...,m$).

Note that

$$\sum_{s \in Q_i} x_{is} = W_i \quad , i = 1, 2, \dots, n. \tag{7}$$

Given $\{x_{is}\}$ the minimal time problem for each interval can be stated [3]. It can be shown that $\Delta_s(Z_s)$ is a convex function of x_{is}.

Then the project time

$$T(z_1, z_2, \dots, z_m) = \sum_1^m \Delta_s(z_s) \tag{8}$$

is the convex function of x_{is} as well.

So we have the problem of convex function minimization under the linear constrains (7). This problem can be solved by any method of the convex programming. If some $\Delta_s(z_s) = 0$ the solution can be improved by changing the order of the events $(s-1)$ and S and Solving the problem under new ordering [4]. However the solution obtained by this procedure is not necesserilly optimal.

3. The case of power functions.

Let $f_i(v_i) = v_i^{1/\alpha}$, $\alpha > 1$, $i = 1, 2, \dots, n$. Then [3]

$$\Delta_s(z_s) = N^{-1/\alpha} \left[\sum_{i \in R_s} x_{is}^{\alpha} \right]^{1/\alpha} \tag{9}$$

Applying the Lagrange Multipliers Method one has

$$\frac{\partial \Delta_s(z_s)}{\partial x_{is}} = N^{-1/\alpha} \left(\frac{x_{is}}{\Delta_s} \right)^{\alpha-1} = \lambda_i \tag{10}$$

Denote v_{is} - the i-th activity resource value at the S-th interval.

From (10) it follows that $v_{is} = (N\lambda_i^{\alpha})^{1/\alpha-1}$.
Consequently doesn't v_{is} depend on S . So the important property of the optimal Solution can be formulated: <u>each activity has a constant level of resoures. (property 1).</u>

Suppose that $\Delta_s(z_s) > 0$, $s = 1, 2, \dots, m$ under the optimal solution. Then taking into account property 1 property 2 can be stated as follows: <u>the resource values $\{v_i\}$ are the flow N in the network (property 2).</u>
It can be proved that property 2 takes place when some $\Delta_s = 0$.

Properties 1 and 2 are valid for any ordering of the events. Consequently they are valid without any assumption of ordering. Therefore the ordering of the events isn't supposed here after.

To get other properties consider q-dimentional region Y corresponding to the project network [2]. Define the distance ρ between any two points y^1 and y^2 of the region as

$$\rho(y^1, y^2) = \left(\sum_1^q | y_j^1 - y_j^2 |^{\alpha} \right)^{1/\alpha} \qquad (11)$$

Comparing (11) and (9) one can notice that some trajectory corresponds to any set $\{x_{is}\}$. From this fact it follows that:

<u>the shortest trajectory connecting initial and final points corresponds to the optimal solution (property 3).</u>

Denote W_e - the length of the shortest trajectory and call it the equivalent project volume.

Then

$$T_{min} = W_e\, N^{-1/\alpha} \qquad (12)$$

As it follows from (12) the phase point velocity is equal to $N^{1/\alpha}$. As the shortest traectory is the only one under any velocity it follows that <u>values $\{x_{is}\}$ don't depend on the resources level $N(t)$ under the optimal solution (property 4)</u>

The minimal project time is defined by the equation

$$\int_0^T N^{1/\alpha}(t)\, dt = W_e \qquad (13)$$

Describe an algorithm of foundation W_e . First of all represent the network in the new coordinat system where axis are parallel. The number of axis is equal to the project dimention q .

Each point of the ordinary (Decart) system is represented by a curve (front) in the parallel coordinat system. Such representation is intermediate one between network and phase-space representation. Fig:1 shows all three types of the project representation. Here $F_0 = (0,0,0)$ - the initial state of the project, $F_k = (y_1^k, y_2^k, y_3^k)$ - the final state of the project. The main idea of the algorithm is based on shortest route foundation.

In order to find the shortest route between points O and A define the line OA in the ordinary coordinat system. This line will be represented by ~~[illegible]~~ the set of fronts in the parallel coordinat System

$$y_j(t) = t\,y_j^k, \quad 0 \le t \le 1, \quad j = 1, 2, \dots, q. \tag{14}$$

If this line is complet-ly inside Y corresponding solution will be optimal. In the other case the shortest trajectory will consist of two segments OD and DA. The following relation takes place

$$\frac{\mathcal{D}C}{\mathcal{D}B} = \frac{O\mathcal{D}}{\mathcal{D}A} \tag{15}$$

Point D defines the front $F_1 = (y_1^1, y_2^1, y_3^1)$ where $y_1^1 = W_1$, $y_2^1 = W_2$,

$$y_3^1 = W_5 \frac{(W_1^{\alpha} + W_2^{\alpha})^{1/\alpha}}{(W_1^{\alpha} + W_2^{\alpha})^{1/\alpha} + (W_3^{\alpha} + W_4^{\alpha})^{1/\alpha}} \tag{16}$$

Notice that (16) is equivalent to (15). Similary one can define a number of basic fronts such that . The trajectory between them is a line. The solution obtained can be improved by correcting the position of the each front basic with respect to the position of two neighboring ones:

4. The relation between problem 1 and 3.

Define the activity time-cost function

$$S_i(\tau_i) = v_i(\tau_i)\cdot\tau_i = \tau_i f^{-1}\left(\frac{W_i}{\tau_i}\right) = \frac{W_i^{\alpha}}{\tau_i^{\alpha-1}} \tag{17}$$

taking into consideration property 1,

So the activity time-cost function is the power convex function of τ_i .

Consider minimal cost problem under the given project time T . The optimality conditions can be written as follow[5].

$$\sum_{i\in U_p^-} \frac{d s_i(\tau_i)}{d\tau_i} = \sum_{i\in U_p^+} \frac{d s_i(\tau_i)}{d\tau_i}, \; p = 2,\dots,m-1, \quad (18)$$

where U_p^- is the set of the activities for which the p -th event is the final one, U_p^+ is the set of the activities for which the p -th event is the initial one.

From (17) and (18) we obtain,

$$\sum_{i\in U_p^+} v_i = \sum_{i\in U_p^-} v_i \;, \; p = 2,\dots,m-1 \quad (19)$$

In other words (19) is the property 2 for the problem 1. Let S_0 be the minimal project cost under the project time T_0 , τ_i^0 and s_i^0 - the time and cost of the i -th activity correspondingly . Then the equivalent volume W_e ; minimal project time T , under the given resource level N the values τ_i and s_i can be found from the following formulas

$$W_e = \left(\frac{S_0}{T_0}\right)^{1/\alpha} T_0 \;, \quad T = T_0\left(\frac{S_0}{N T_0}\right)^{1/\alpha}$$
$$v_i = \frac{s_i^0}{\tau_i^0}\,\frac{N T_0}{S_0} \;, \quad \tau_i = \tau_i^0\left(\frac{S_0}{N T_0}\right)^{1/\alpha} \quad (20)$$

So knowing the solution of problem 3 it is possible to find the optimal solution of problem 1. The relation between the minimal cost problem and minimal resources level problem can be expended to the case of the arbitrary functions $f_i(v_i)$ if the solution satisfying properties 1 and 2 is obtained . In this case the optimal activity times τ_i in problem 1 are the same as in problem 3 under the same project time T if

$$s_i(\tau_i) = \int_{\tau_i}^{\infty} f_i^{-1}\left[\frac{W_i}{\tau}\right] d\tau \quad (21)$$

The proof follows immediately from the fact that property 2 is equivalent to (18) if (21) takes place.

Example . Let $s_i(\tau_i) = a_i - b_i\tau_i$,

$$d_i \le \tau_i \le D_i \;, \; i = 1,2,\dots,n.$$

From (21) we have

$$\tau_i(v_i) = \begin{cases} D_i & \text{if } v_i < b_i \\ d_i & \text{if } v_i > b_i \end{cases}$$

and $\tau_i(v_i)$ has any value from d_i to D_i if $v_i = b_i$. Consequently the minimal flow problem under the given project time T is obtained.

5. Solution of problem 2.

Consider problem 2 for the case of the power functions $f_i(v_i)$. Let W_e be " the equivalent project volume. Then $w(t) = N^{1/\alpha}(t)$ is the project speed and the constrain (5) takes form

$$\int_0^t N(\tau)\,d\tau \le S(t) \tag{22}$$

The problem is to find $N(\tau)$ satisfying (22) so that the project time is minimal.

The main point of the algorithm is to build the convex $\hat{S}_T(t)$, $t \in [0,T]$, so that $\hat{S}_T(T) = S(T)$ (Fig.4).

Step 1. Find T_0 using the equation

$$T_0 \left[S(T_0)\, T_0^{-1} \right]^{1/\alpha} = W_e .$$

If $S(T_0)\, T_0^{-1} t \le S(t)$, $t \in [0, T_0]$ then T_0 is the minimal time . If there are intervals $\mathcal{E} \in [0, T_0]$ such that $S(T_0) T_0^{-1} t > S(t)$ for $t \in \mathcal{E}$ find minimal Δ such that $S(T_0) T_0^{-1}(t - \Delta) \le S(t)$ for $t \in [0, T_0]$. To provide this find $t(\Delta)$ minimazing $S(t) - S(T_0) T_0^{-1}(t - \Delta)$. The value Δ is determined by the equation $t(\Delta) = 0$. When the value Δ is calculated, find $T_1 = T_0 + \Delta$

Step 2. Construct $\hat{S}_{T_1}(t)$, $t \in [0, T_1]$

Calculate

$$W(T_1) = \int_0^{T_1} \left[\frac{d\hat{S}_{T_1}(t)}{dt} \right]^{1/\alpha} dt$$

If $W(T_1) = W_e$ then T_1 is the minimal time.

If $W(T_1) > W_e$ then find

$$T_2 = \tfrac{1}{2}(T_0 + T_1)$$

and go to the 3-rd step.

Step i . Construct $\hat{S}_{T_{i-1}}(t)$, $t \in [0, T_{i-1}]$.

Calculate

$$W(T_{i-1}) = \int_0^{T_{i-1}} \left[\frac{dS_{T_{i-1}}(t)}{dt} \right]^{1/\alpha} dt$$

If $W(T_{i-1}) = W_e$ then T_{i-1} is the minimal time

If $W(T_{i-1}) > W_e$ then determine $T_i = T_{i-1} - \frac{1}{2}|T_i - T_{i-1}|$,

If $W(T_{i-1}) < W_e$ then determine $T_i = T_{i-1} + \frac{1}{2}|T_i - T_{i-1}|$

and go to the next step.

If $S(t)$ is the convex function the problem can be solved more simply.

If $S'(t)$ is convex (upwards) then T_{min} is determined by the equation

$$\left[\frac{S(T_{min})}{T_{min}} \right]^{1/\alpha} T_{min} = W_e$$

If $S(t)$ is convex (downwards) then T_{min} is determined by the equation

$$\int_0^{T_{min}} \left[\frac{d\,S(t)}{dt} \right]^{1/\alpha} dt = W_e$$

6. Analysis of the property 1.

The question of interest is what types of functions $f_i(v_i)$ give the optimal solution of problem 1, satisfing propertys. The optimaluty conditions for the arbitrary convex functions $f_i(v_i)$ are following

$$\frac{1}{\lambda_i} \frac{d\varphi_i(w_{is})}{d w_{is}} = \sum_{j\in R_s} w_{js} \frac{d\varphi_j(w_{js})}{d w_{js}}$$

where $\varphi_i = f_i^{-1}$ and $w_{js} = x_{js}/\Delta_s$.

If $w_{is} = w_i$ i.e. w_{is} doesn't depend on S then

$$\sum_{i\in U_s^-} w_i \frac{d\varphi_i(w_i)}{dw_i} = \sum_{i\in U_s^+} \frac{d\varphi_i(w_i)}{dw_i} \qquad (23)$$

Denote $F_i(v_i) = w_i \frac{d\varphi_i(w_i)}{dw_i} = f_i(v_i)\left[\frac{df_i(v_i)}{dv_i}\right]^{-1}$. As it follows from (23) $\{F_i(v_i)\}$ form the flow in the network.

Theorem. To provide that $\{F_i(v_i)\}$ is a flow for any flow $\{v_i\}$, the following necessary and sufficient Conditions must be satisfied

$$F_i(v_i) = a_i + \alpha v_i \ , \ i = 1, 2, \dots, n \ ,$$

where $\{a_i\}$ is a flow (It is supposed the network stayes connected after input and output are reboved).

Proof. Making variations of v_i along any path in the network we obtain $\frac{dF_i(v_i)}{dv_i} = \frac{dF_j(v_j)}{dv_j}$ for any i, j and $\frac{d^2F_i(v_i)}{dv_i^2} = 0$ for any i. It follows from the written above that

$$F_i(v_i) = a_i + \alpha v_i \ , \ i = 1, 2, \dots, n \ ,$$

where $\{a_i\}$ is a flow because $F_i = a_i$ when $v_i = 0$.

If follows from the theorem that $f_i(v_i) = c_i(a_i + \alpha v_i)^{1/\alpha}$, where $\alpha > 1$ because $f_i(v_i)$ must be a convex function when $a_i + \alpha v_i \geq 0$. Let W_e - the equivalent project volume and $\tilde{x}_{is}$ - optimal values of x_{is}. Then if

$$\tilde{x}_{is}^{\alpha} \left(\sum_{i\in R_s} \tilde{x}_{is}^{\alpha}\right)^{-1} \geq a_i (\alpha N + A)^{-1}$$

where A is the value of the flow $\{a_i\}$ then the minimal project time $$T_{min} = W_e \left(\alpha N + \beta \right)^{-1/\alpha}$$

REFERANCES

I. В.Н.Бурков, А.Я. Лернер. Новые задачи теории сетевого планирования и управления. В сб."Вопросы управления в больших системах". Изд. "Онтиприбор", 1968.

2. В.Н.Бурков. Применение теории оптимального управления к задачам распределения ресурсов. Труды Ш Всесоюзного совещания по автоматическому управлению /Одесса, 1965 г./. т.Управление производством. м., Наука, 1967.

3. В.Н.Бурков. Распределение ресурсов как задача оптимального быстродействия. Автоматика и телемеханика, т. XXУП, №7, 1966.

4. Б.С.Разумихин. Задачи об оптимальном распределении ресурсов. Автоматика и телемеханика, № I, 1967.

5. E.B.Berman. Resource allocation in a PERT network under continuous activity time-cost functions. Manag.Sci.,v.10, No. 4, 1964.

35.5

SOME QUESTIONS OF THE TESTING AND CONSTRUCTION PRINCIPLES OF AN OPTIMUM MULTILEVEL CONTROL STRUCTURE IN SYSTEMS WITH A SPECIFIC OBJECTIVE FUNCTION

M.K.Badunachvili, D.I.Golenko, S.S.Naumov

Electronic Control Computers Institute, Moscow U S S R

Let us consider a class of controlling subsystems with some objective function and let us assume that there exists an integral measure of the target vicinity in such subsystems and this integral measure is some nondecreasing time function V (t). It is evident that for each of the subsystems under discussion there exists some value of the integral measure corresponding to the target attainment V sch. Depending on a variety of factors this measure can be obtained at different time instants. For the given subsystem an earliest early target attainment time t_e is assumed to exists,objectively,also (there exists the maximum movement speed to the target)which corresponds to the maximum utilization of the subsystem internal resources and the minimum external obstacles; generally speaking, this time t_p differs from the optimistic time t_0 used in PERT Systems. At the same time there exists a deadline for the target attainment t_{cr} x). Of the target is reached later than t_{kp}, the subsystem pays a fine. Hence, with the above constraints the subsystem must reach the target within the time interval (t_e , t_{cr}).

It is necessary to note that during the movement to the given target the subsystem tries to minimize the efforts directed to the attainment of this target because repeated and lengthy operation in critical conditions (corresponding to the maximum speed of the movement to the target) can an early wear the subsystem too early, Therefore in this subsystem the scheduled time of the target attainment $t_p \leqslant t_{sch} < t_{kp}$ is chosen such that it is

x) Systems for which $t_{cr} = \infty$ obviously do not bebap to the class of systems discussed.

ahead of t_{cr} by some time interval and provides cerfain confidence boundaries simultaneously both for the reliability of the target attainment not later than t_{cr} and for the minimization of the efforts spent in this movement.

Let us assume the curve $V_{sch}(t)$ /fig.I/ is the scheduled curve of the movement to the target and the integral measure corresponding to the target attainment is

$$V_{sch} = \int_{t_i}^{t_{sch}} V_{sch}(t)\,dt, \qquad /I/$$

where t_i -the initial instant of the movement to the aim. Then in the process of the system movement, for the testing purposes it is necessary to compare the true measure

$$V(t_i) = \int_{t_i}^{t_j} V(t)\,dt, \qquad /2/$$

which is a random value with the value of the target vicinity V_{sch} (ti) at certain time instants ti, which must be befined by the choice of the quantization unit V (t) is the true curve of the movement to the target in the above formula.

The control function of this system is elimination of the divergence between the actual and the scheduled measures of the target vicinity appearing due to the effect disturbances with the view to ensuring the target attainment by the subsystem by the time t_{sch}.

Before we discuss the algorithm let us shift the curve $V_p(t)$, corresponding to the movement to the target at the time t_e, in the direction of the X axis so that the end of the curve $V_p(t)$ would coincinde with the point (V_{sch}, t_{sch}) /fig,I). In this case we shall receive the point t_I as the intersection of the X axis with V_e (t). It is easy to see that if the subsystem does not move to the target $V(t_I) = 0$ before the time t_I, obviously there is a non zero probability that we shall reach the target by the time moment t_{sch} starting from the time instant t_I, and using the possibilities of the subsystem to the utmost. We can consider t_I as the critical permissible time of the first interrogation. If the interrogation will be carried out after the time instant t_I then the deadline t_{sch} cannot be met because the highest velocity to the target is defined by the curve $V_e(t)$.

Thus, to meet the deadline t_{sch}, the first interrogation of the subsystem must be performed within the period specified by the interval $t_i < t \leqslant t_I$. Assuming the critical value $t = t_I$ and performing the interrogation, the subsystem receives the information on the target vicinity by comparing $V(t_I)$ and $V_{sch}(t_I)$. With the is information effective controlling actions aimed et suppression of the divergence must be worked out. Drawing now through the point $[t_I, V(t_I)]$ a straight line, parallel to the X- axis, to the intersection with the shifted curve $V_e(t)$, we shall receive a point with the abscissa (t_2) which, as indicated above, deferimines the critical value of the second interrogatory period.

The next times of the interrogation are found in a similar way.

We have noted above that the actual curve the movement to the target $V(t)$ is a random curve. Thus, depending on the nature of its change, the same volume of V_{sch} can be reached at a certain actual time t_{act}. Here we have three cases :

I. The case t_{sch} t_{act} can occur a) when the subsystem is uncontroll able; it is evident that for this subsystem the interrogation is generally useless; b) when, in spite of the fact that the subsystem is controlled, the condition t t_{sch} is obtained. In this case, at some interrogation step we shall be on the shifted curve $V_e(t)$, As the possibility of the movement strictly along the curve $V_e(t)$ is low, the unit controlled in this case is to obtain the least delay as far as the time t_{sch} is concerned,

2. In the case where $t_i < t_{act} < t_{sch}$ a finite number of the interrogatory steps for the control of the movement to the target aire evidently required.

3. In the case $t_{act} = t_{sch}$ the sequence of all points of the interrogation has the limit of convergenc $- t_{sch}$ and the process of approximation occurs in an number of steps. But since in practice the volume V_{sch} is to be reached by the time t_{sch} with some specified accuracy $t_{sch} \pm \Delta t_{sch}$, in this case what is needed is to get into the region $[t_{sch} - \Delta t_{sch}\ t_{sch} + \Delta t_{sch}]$

that, in contrast to the convergence to t_{sch} can be done in a finite nümber of steps. The condition $\Delta t < \tau$ must evidently be met.

An analitical record of the geometrical interpretation of the above algorithm leads us to the following relation which describes the critical value of the ($i+1$) - th interrogation depending on the information on the measure of the target vicinity at the i - th control step of the subsystem.

$$t_{j+1} = t_1 + \frac{V(t_j)}{V_{sch}}\left(t_{sch} - t_1\right) \qquad /\ 3\ /$$

here t_I is the time of the first interrogation V_{sch} and $V(ti)$ are respectively of the target vicinity expressed by formulas (I) and (2) the planned and the actual integral measures.

Let us suppose now that $V_{sch}(t)$ and $V_e(t)$ are given in the form of straight lines and the actual curve of the movement to the target coincides with $V_{sch}(t)$. In this case $t_{act} = t_{sch}$ and the interrogation of the subsystem is needed only to prevent possible troubles

By using the evident relation $\frac{V(t_j)}{V_{sch}} = \frac{t_j}{t_{sch}}$ >

formula / 3 / can be written as

$$t_{j+1} = t_1 + \left(1 - \frac{t_1}{t_{sch}}\right)t_j$$

Using this expression th easy to receive formula for the value of the ($i+1$) step is $t_{j+1} - t_j = \left(1 - \frac{t_1}{t_{sch}}\right)\left(t_j - t_{j-1}\right)$.

This relation may be recorded also as

$$t_{j+1} - t_j = \left(1 - \frac{t_1}{t_{sch}}\right)t_1 \, ,$$

and as

$$t_{sch} = \sum^{\infty}\left(t_{j+1} - t_j\right),$$

it is easy to see that

$$t_{sch} = t_1 \sum_{j=0}^{n}\left(1 - \frac{t_1}{t_{sch}}\right)^j + t\sum_{j=n+1}^{\infty}\left(1 - \frac{t_1}{t_{sch}}\right)^j. \qquad /\ 4\ /$$

Here

$$R_n = t_1 \sum_{i=n+1}^{\infty}\left(1 - \frac{t_1}{t_{sch}}\right)^i,$$

is a remanence term for which the following constraint will be reasonable

$$R_n \leqslant \Delta t_{sch}.$$

Now since the terms under the sums sign in the right - hand part of ed / 4 / are the sums of the decreasing geometrical progression, we can write that

$$t_{sch} = t_1 + t_{sch}\left[(1 - \frac{t_1}{t_{sch}}) - (1 \quad \frac{t}{t_{sch}})^{n+1} \right] + \Delta t_{sch} .$$

After simple transformations we can obtain an expression for number of the preventive interrogations required to test the subsystem with the parameters t_{sch}, t_e t_{sch}

$$n = \frac{\ln \frac{\Delta t_{sch}}{t_e}}{\ln \frac{t_e}{t_{sch}}} \qquad /5/$$

Let us consider now a complex system consisting of subsystems of the above form. Thus, we consider two fixed levels of some multilevel structure. With this we carry out the numbering of the ranks in the following way : the number K is given to the rank of the subsystems and the number K-I to the rank of the system.

As the functioning of every subsystem is directed to of a definite target, this system evidently will generally possess with some network of the targets which are parts of the overall target. The system has the integral curves of the aim nearness $V^{(c)}_{sch}$ (t), $V^{(c)}_e$ (t) and $V^{(c)}$ (t) showing the different conditions of the system movement to the overall target as well as the appropriate parameters.

The existence of the target network also means that there is some critical path with a certain number of targets on it and this means the existence of a certain number th of the critical subsystems M . Let us choose the $\dot{L}$ subsystem from all these subsystems for which the number of the preventive interrogations given by formula (5) will be minimal , i. e.

$$\min_{1 \le j \le m} N_j = \min_{1 \le j \le m} \frac{\ln \frac{\Delta t_{schj}}{t_{ej}}}{\ln \frac{t_{ej}}{t_{schj}}}$$

Here we have the obvious inequality

$$\Delta t_{schj} < t_{ej} < t_{schj} \qquad /6/$$

If every subsystem the critical path is interrogated a specified number of times, the appropriate minimal number of the interrogation required for the interrogation of all critical path will be equal to

$$\min_{1 \le j \le m} N_K = m \cdot \min_{1 \le j \le m} N_j , \qquad /7/$$

where N_K is the number of the interrogationes needed for the eritical path with the given detailing Δt_{schj}

Let us consider now the system with the parameters t_{sch} T_e and ΔT_{sch} of its movement to the overall target

$$\begin{cases} T_{sch} = \gamma_1 \cdot t_{sch}, \\ T_e = \gamma_2 \cdot t_{ej}, \\ \Delta T_{sch} = \gamma_3 \cdot \Delta t_{schj}, \end{cases}$$

where $\gamma_1, \gamma_2, \gamma_3$ are arbitrary nonnegative number above I and let us assume that

$$\gamma_1 > \gamma_2 > \gamma_3 . \qquad /9/$$

Let us quite naturally require that in the discussed multilevel structure of the number of interrogations needed for control decrease with the increase of the rank (with the decrease of the rank number) of the hierarchy. With this we shall give more rigid form to our demand

$$N_{K-1} \leqslant \min_{1\le j\le m} N_K . \qquad /10/$$

Where N_{K-1} is the number of interrogations needed for the system interrogations in terms of detailing the ($K-1$) th rank, i.e. by parameters (8). In fact, taking into consideration expressions (6) , (7),(8) and (9) we can write the following inequalities

$$N_{K-1} = \frac{\ln \frac{\gamma_3 \Delta t_{schj}}{\gamma_2 t_{ej}}}{\ln \frac{\gamma_2 t_{ej}}{\gamma_1 t_{schj}}} \leqslant \frac{\ln \frac{\gamma_3}{\gamma_2}}{\ln \frac{\gamma_2}{\gamma_1}} + \frac{\ln \frac{\Delta t_{schj}}{t_{ej}}}{\ln \frac{t_{ej}}{t_{schj}}} = \frac{\ln \frac{\gamma_3}{\gamma_2}}{\ln \frac{\gamma_2}{\gamma_1}} + \min_{1\le j\le m} N_j \leqslant$$

$$\leqslant \frac{\ln \frac{\gamma_3}{\gamma_2}}{\ln \frac{\gamma_2}{\gamma_1}} + m \cdot \min_{1\le j\le m} N_j = \frac{\ln \frac{\gamma_3}{\gamma_2}}{\ln \frac{\gamma_2}{\gamma_1}} + (m-1) \min_{1\le j\le m} N_j + \min_{1\le j\le m} N_j .$$

Hence it is easy to see that for the realization of the condition

$$N_{K-1} \leqslant \min_{1\le j\le m} N_j$$

automatically meeting condition (10) it is necessary that

$$\frac{\ln \frac{\gamma_3}{\gamma_2}}{\ln \frac{\gamma_2}{\gamma_1}} + (m-1) \min_{1\le j\le m} N_j \leqslant 0 . \qquad (11)$$

Now, taking into consideration the evident inequality

$$\gamma_1 t_{schj} \geqslant \gamma_2 t_{ej} , \qquad (12)$$

condition (11) may be written as

$$\frac{\gamma_3}{\gamma_2} \geqslant \left(\frac{\Delta t_{schj}}{t_{ej}} \right)^{m-1} . \qquad (13)$$

Let us prove that condition (I3) is also a sufficient condition. In fact, assuming that $\frac{\gamma_3}{\gamma_2} < \left(\frac{\Delta t_{schj}}{t_{ej}}\right)^{m-1}$,

then et $m \to \infty$ $\frac{\gamma_3}{\gamma_2} < \lim_{m\to\infty} \left(\frac{\Delta t_{schj}}{t_{ej}}\right)^{m-1} = 0$.

This means that either γ_3 or γ_2 is negative that contradicts the statement of the problem. This contradiction proves the assumption. However, as in fact there exists only a certain finite value of the number of subsystems along the critical path $m = m_{ch}$ evidently there exists an interval for the positive values $\frac{\gamma_3}{\gamma_2}$ $0 < \frac{\gamma_3}{\gamma_2} \leqslant \left(\frac{\Delta t_{schj}}{t_{ej}}\right)$ within which it is not clear whether the sufficiency for condition (I3).

The rate of convergence for the value $y = \left(\frac{\Delta t_{schj}}{t_{ej}}\right)^{m_{ch}-1}$ to the zero when increases may be characterized by the bangth of the interval $\mathcal{J} = \int_1^{\infty} \left(\frac{\Delta t_{schj}}{t_{ej}}\right)^{m-1} dm = \frac{1}{\ln \frac{t_{ej}}{\Delta t_{schj}}}$.

Hence follows that the value $\mathcal{J}$ changes in proportion to the detailing index Δt_{schj}. The higher the number of secondary targets, i.e. the ~~II~~ iover the values of Δt_{schj} the higher the rate converegence for to the zero the indicated function and thus the smalles the zone where the property of the sufficiency is houbted. Obviously, by selecting the parameters Δt_{schj}, t_{ej} and m this zone may be done smaller than any number given in advance.

Now let us solve inequality (II) for $\frac{\gamma_2}{\gamma_1}$ and with analogous reasonings; it is easy to see that

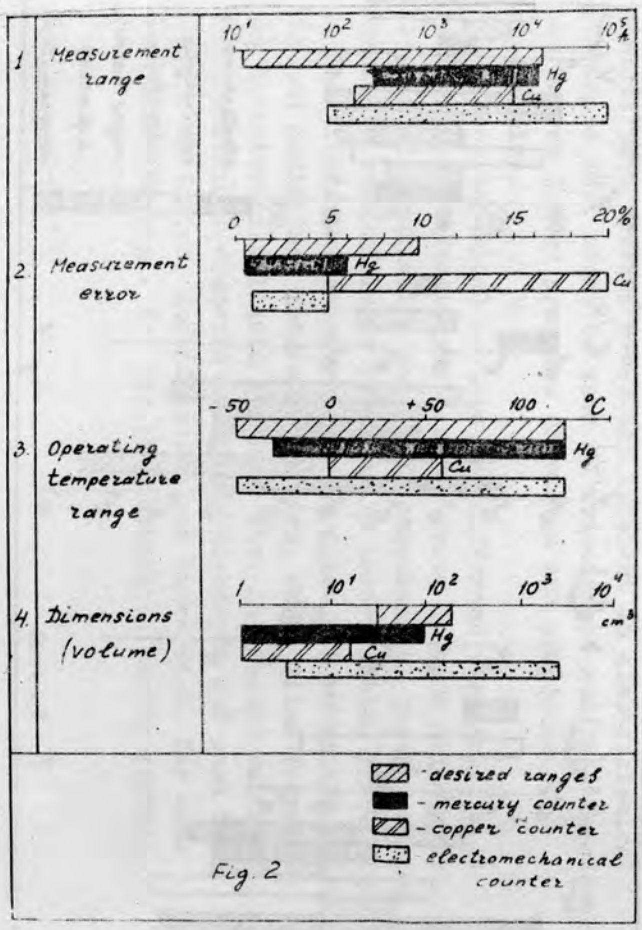
$$\frac{\gamma_2}{\gamma_1} \leqslant \left(\frac{t_{schj}}{t_{ej}}\right)^{\frac{1}{m-1}}$$

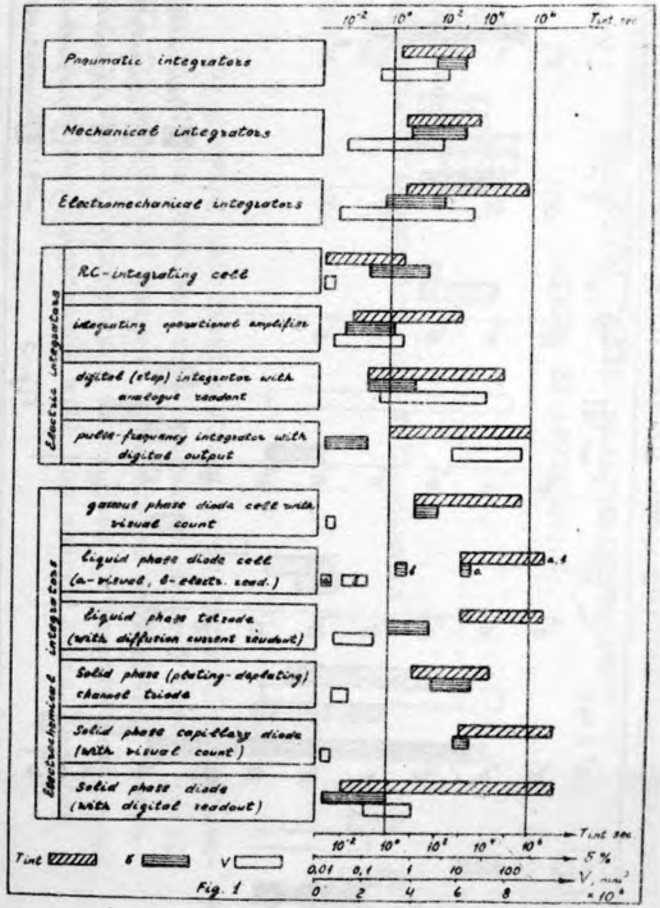
with relation (I2) as well as inequalities (II) and $\gamma_2 t_{ej} > \gamma_3 \Delta t_{schj}$ the entire system of the constrains that are put in this case on the relation $\frac{\gamma_2}{\gamma_1}$ and $\frac{\gamma_3}{\gamma_2}$ will have this form

$$\left\{ \left(\frac{\Delta t_{schj}}{t_{ej}}\right)^{m-1} \leqslant \frac{\gamma_3}{\gamma_2} \leqslant \frac{t_{ej}}{\Delta t_{schj}}, \quad \frac{t_{ej}}{t_{schj}} \leqslant \frac{\gamma_2}{\gamma_1} \leqslant \left(\frac{t_{schj}}{t_{ej}}\right)^{\frac{1}{m-1}} \right\}$$

Taking into consideration relation (8), the necessary condition for the optimality of the given multilevel structure may be finally given by.

$$\begin{cases} \left(\frac{\Delta t_{schj}}{t_{ej}}\right)^{m} \leqslant \frac{\Delta T_{sch}}{T_e} \leqslant 1 \\ \left(\frac{t_{ej}}{t_{schj}}\right)^{2} \leqslant \frac{T_e}{T_{sch}} \leqslant \left(\frac{t_{schj}}{t_{ej}}\right)^{\frac{2-m}{m-1}} \end{cases}$$


1 Measurement range
2 Measurement error
3. Operating temperature range
4. Dimensions (volume)
Hg
Cu
desired ranges
- mercury counter
- copper counter
electromechanical counter
Fig. 2


Pneumatic integrators
Mechanical integrators
Electromechanical integrators
Electric integrators
RC-integrating cell
integrating operational amplifier
digital (step) integrator with analogue readout
pulse-frequency integrator with digital output
Electrochemical integrators
gaseous phase diode cell with visual count
liquid phase diode cell (a-visual, b-electr. read.)
liquid phase tetrode (with diffusion current readout)
Solid phase (plating-deplating) channel triode
Solid phase capillary diode (with visual count)
Solid phase diode (with digital readout)
Tint
Fig. 1

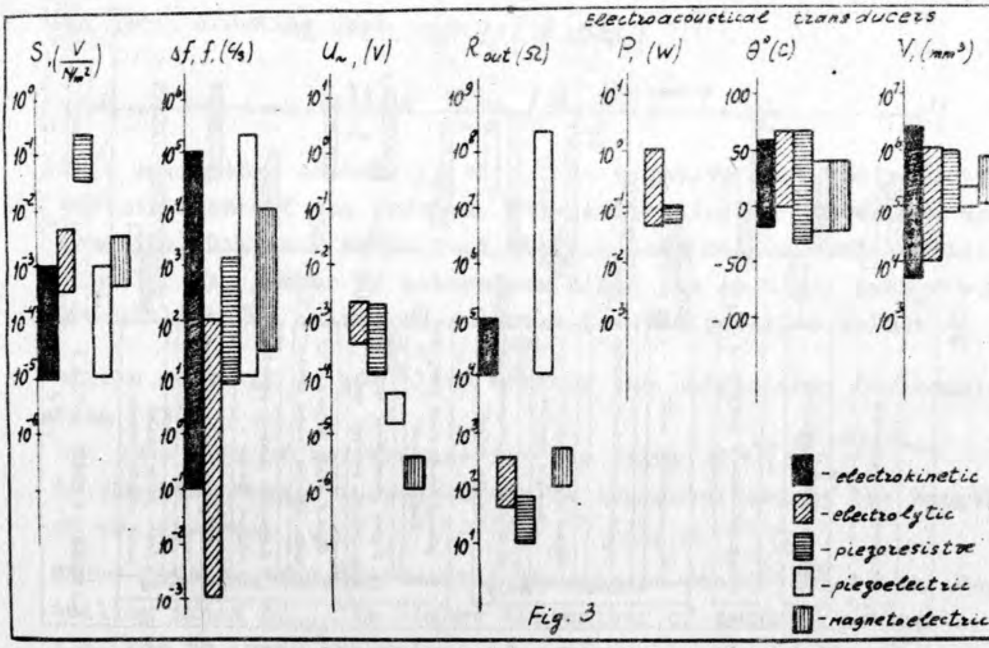

Electroacoustical transducers
S, (V/N/m²)
Δf, f, (c/s)
U~, (V)
R out (Ω)
P, (W)
θ° (C)
V, (mm³)
- electrokinetic
- electrolytic
- piezoresistive
- piezoelectric
- magnetoelectric

Fig. 3

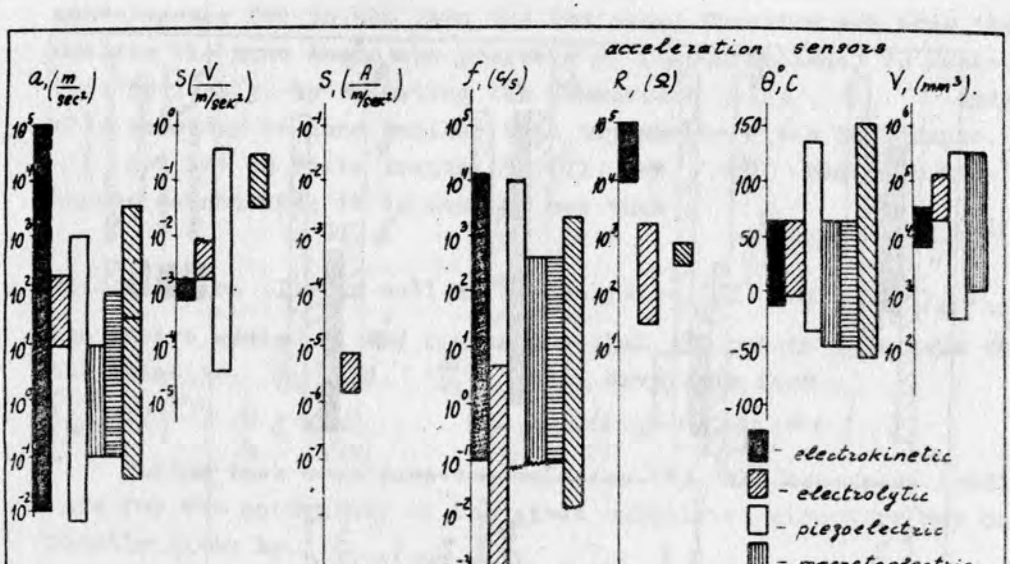

acceleration sensors
a, (m/sec²)
S (V/m/sec²)
S, (A/m/sec²)
f, (c/s)
R, (Ω)
θ°, C
V, (mm³)
- electrokinetic
- electrolytic
- piezoelectric
- magnetoelectric
- inductive
- piesoresistive
*) R_H = 1 KΩ

Fig. 4

36.1

ELECTROCHEMICAL TRANSDUCERS, COMPARATIVE PROPERTIES, BASIC CHARACTERISTICS AND FIELDS OF APPLICATION

A.P.Shorygin
Institute of Automation and Telemechanics
Moscow
USSR

The growing information flows to be controlled and complexity of the problems that automatic control has to tackle call for simple and compact information acquisition, processing and storage components to be used in lengthy continuous or discontinuous processes with low energy stimuli and/or low frequencies.

These and other requirements to components, devices and instruments made by new problems of automatic monitoring and control have aroused a certain interest in electrochemical phenomena that might be used in transducers and this despite the unprecedented progress of vacuum and semiconductor electronics.

Electrochemical devices use different reactions in liquid or solid electrolytes and/or polar liquids (electrolytic and electrokinetic cells). Electrochemical systems may be partially reversible or practically irreversible. There might be variations in concentrations of dissolved materials, plating and deplating, metals deposition, oxide films formation and deterioration, liquid transfer etc. At present various electrochemical elements are being developed and used for signals detection, amplification and conversion, for data processing, storage, for use in adaptation, remote control, automata etc. Investigations are underway on electrochemical matrices; attempts are made to use electrochemical devices for simulation of biological processes; electrochemical systems and processes that could be used in transducers are investigated.

The existing electrochemical devices can be divided into two major groups. The first includes those that already are or can be used in industry due to their clear-cut advantages or specific features.

These units are:

I) integrating units with visual, electric or photoelectric read-out;

2) electric rectifying diodes for very low currents and low frequency;

3) sensors of variable or pulsed pressures and acoustic pressure gradients of low and infralow frequency;

4) conductive concentration (salination) conductometric and meters, galvanic gas analysers, composition sensors, etc.

Of substantial interest for a number of fields are first of all the electrolytic integrating devices.

$$P_{output} = \varphi\left[\int_0^{\tau} i_{inp}(t)\,dt\right] \qquad (I)$$

where i_{inp} is the input current, τ - the duration of its flow. Electrochemical integrating devices vary widely in their characteristics and practical possibilities due to the type of redox system, electrochemical effects used and the read-put procedure (Fig. 2).

The output of "gas phase" integrating diodes is the displacement Δx of electrolyte in the capillary which connects the hydrogen-filled electrode chambers

$$\Delta x \approx \frac{RT}{k_{\partial} p S_{\kappa}} \int_0^{\tau} i_{inp}(t)\,dt \qquad (2)$$

where p is the hydrogen pressure, S_k the crosssection area of the capillary, K_{∂} - the instrument constant.

The redox reaction in liquid phase cells (solion) follows the equation

$$A_1 + ne \rightarrow A_2$$
$$A_2 - ne \rightarrow A_1 \qquad (3)$$

where A_1 is the oxided and A_2 the reduced form of the solution components. For typical conditions where the rate of the process practically depends only on diffusion towards a cathode of the form A_1, the output e.m.f. of an integrating diode is found by the equation

$$C_{01} n FS\ell \cdot th \frac{E_c}{E_T} = \int_0^{\tau} i_{inp}(t)dt \qquad (4)$$

where $E_T = RT/nF$, F the Faraday number, S and ℓ the cross-section area and the length of integral chamber separated from the remainder of the cell by a semi-permeable membrane, C_{01} the initial concentration of the form A_1 (non-basic carrier). With $E_c << E_T$ the relation becomes linear and the sensitivity of the integrator diode will be

$$S_E = \frac{E_c}{q_{inp}} = \frac{2RT}{C_{01}(nF)^2 S\ell} \qquad (5)$$

The maximal value of the input charge will depend on the capacity of integral chamber and the initial concentration of form A_1:

$$(q_{inp})_{max} = nFC_{01}V \qquad (6)$$

To ensure the required integration accuracy the upper bound of the input signal frequency spectrum should meet the condition

$$f_2 << \frac{D}{2\pi \ell^2} \qquad (7)$$

where D is the diffusion factor.

The storage time for the integral value depends on the self-discharge time constant

$$\tau_{s.d} = \frac{V\ell_K}{DS_K} \qquad (8)$$

Here S_K is the total equivalent cross-section area of the capillaries in the semi-permeable membrane,

l_k - is the equivalent thickness of the membrane.

The output of liquid state triodes and tetrodes is the diffusion current between the anode of the integrating device and the read-out electrode inserted in the anode chamber which receives the negative bias. It is described as

$$(I_\partial)_{outp} = \frac{2q_{inp} D}{\ell_{2A}^2} \cdot \frac{e^{U_2/E_T} - e^{U_A/E_T}}{e^{U_2/E_T} + e^{U_A/E_T}} \quad (9)$$

where U_2 and U_A are the electrode-electrolyte potential for the read-out electrode and the anode, while ℓ_{2A} is the distance between these electrodes. Since the current in the output circuit is limited by the load resistance R_H, if the input charge reaches the value

$$q_{inp} \geq \frac{E}{R_H} \cdot \frac{\ell_{2A}^2}{2D} \quad (10)$$

saturation occurs and the saturation diffusion current will be

$$I_{sd} = \frac{nFDC_{A1} S_{2A}}{\ell_{2A}} \quad (11)$$

Under these conditions which are used in practical integrators

$$I_{out} = S_I \int_0^{\tau} i_{inp}(t)\,dt \quad (12)$$

($S_I = \frac{k_T n D}{\ell_{2A}}$ and k_T is the solion constant). The typical triodes and tetrodes characteristics are shown in Table 1. Solid phase plate-deplate integrating devices, (Faraday coulombmeters) are now represented by a great variety of types (Table I) . The redox reaction follows the scheme

$$\begin{gathered} Me - ne \rightarrow Me^{n+} \\ Me^{n+} + ne \rightarrow Me \end{gathered} \quad (13)$$

The mass of the substance plated and/or deplated during the interval τ (e.g. during one cycle of action of integrating diode with digital read-out)

$$m_{Me} = A\eta / nF \cdot \int_0^{\tau} i_{inp}(t)\,dt \quad (14)$$

where A is the atomic weight, η the input factor ($\eta \approx 1$). In the case of a capillary cell the output is the displacement of the electrode-solution boundary

$$\Delta x_{out} = S_e \int_0^{\tau} i_{inp}(t)\,dt \tag{15}$$

where $S_e = A\eta / nF\gamma S_K$ the sensitivity (γ - is the density of the material of the electrode column, S_K - its cross-section area). When such a device is employed as the elapsed time indicator

$$t \approx \frac{\Delta x_{out}}{S_e \frac{U}{R_\partial} \cdot \frac{1}{1 + R_c / R_\partial}} \tag{16}$$

where U - is the voltage, R_∂ is the limiting resistor, R_c - the cell resistance.

Of special interest are solid phase channel triodes (memistors) with monotonous changes of linear output resistance R_{out} against an electric charge that has flown in the input circuit [2,3]. At sufficiently low R_{out} and uniform plating of the metal sheet on the storage electrode

$$\sigma_{out} = \frac{R_{out}(t)}{R_{out}(0)} = \frac{1}{1 + R_{out}(0) / \Psi \cdot \int_0^{\tau} i_{inp}(t)\,dt} \tag{17}$$

where $R_{out}(0) = (R_{out})_{max}$ is the initial value, $R_{out}(t)$ the same at the instant t and $\Psi = {}^{nF}/_{\eta A} \cdot \rho \gamma \ell_s^2$ (γ is the mean density of the deposited metal, ρ its specific resistance, ℓ_s the length of storage electrode). At sufficiently large q_{inp} value the resistance variation velocity [3] is

$$\frac{d\sigma_{out}}{dt} = \frac{\Psi}{R_{out}(0) \cdot I_{inp} \cdot t^2} \tag{18}$$

and the shortest time of complete resistance change is

$$t_{min} = \frac{\Psi [R_{out}(0) / (R_{out})_{min} - 1]}{R_{out}(0) \cdot (I_{inp})_{max}} \tag{19}$$

($(I_{inp})_{max}$ is the largest permissible input current).

Under typical conditions the resistance of the reading current for a/c will be [4]

$$Z_2 \approx 2 \frac{\sqrt{R_{S1} A_{F1}}}{f^{0,25}} \cdot th\left(\frac{\ell_s}{2} \sqrt{\frac{R_{S1}}{A_{F1}}} \cdot f^{0,25}\right), \tag{20}$$

where ℓ_s is the length of the storage electrode (resistive electrode), R_{s1} the resistance per unit length, A_{F1} is the factor which describes the electrode-electrolyte Faradey impedance per unit length of the same electrode.

By their integration time range, magnitudes of currents, accuracy and dimensions in a number of cases the electrolytic integrating devices are advantageous (see Fig. I). They have linear characteristics, very low consumption (tens or hundrends of microwatts), they can integrate fractions of microcoulombs, operate with gating photocalls, thermocouples, Hall sensors etc. Their dimensions do not exceed those of miniature electronic tubes, while their weight is 2 to 20gr. The highest accuracy is achieved in solid phase electrochemical integrating devices with digital read-out.

The processes in electrochemical elements develop in very thin layers measured in fractions of a micron; more mituature ("planar") elements are expected.

The near future will see a wide application of elapsed time indicators used on the above principles (since it is necessary to monitor the operability and longevity of instruments, devices and machinery is stipulated by state standards) (Fig. 2)[6], as well as integrating units especially with large integration time, the adjusting units of adaptive systems and optimisers etc. Of substantial interest are also electrochemical devices that process automatically the measurements data and realize an algorithm of the form

$$y_i = \frac{K_i \int_{t_i}^{t_{i+1}} [x(t) - x(t_0)]\,dt}{\sum_1^n K_j \int_{t_j}^{t_{j+1}} [x(t) - x(t_0)]\,dt} \qquad (2I)$$

In liquid phase electrochemical rectifying diodes a redox reaction represented by eq.(2). If the rate of the reaction depends practically on diffusion alone, the integral equation for current through a diode will be [5]:

$$\left[1 - \frac{1}{U_1} \cdot \frac{d}{dt} \int_0^t f_1(t-\tau) . i(\tau) . d\tau\right] \exp nFU_{inp}/RT = \qquad (22)$$

$$= \left[1 - \frac{1}{U_2} \cdot \frac{d}{dt} \int_0^t f_2(t-\tau) \cdot i(\tau)\, d\tau\right]^3,$$

where $U_{1,2} = nFSC_{1,2}$;

$$f_{1,2} = \frac{a}{\mathcal{D}_{1,2}} \left\{ 1 - \exp \frac{t\mathcal{D}_{1,2}}{a^2} \operatorname{erfc} \sqrt{\frac{t\mathcal{D}_{1,2}}{a^2}} \right\} \tag{23}$$

while a - is the radius of a spherical microelectrode; indices I and 2 denote the quantities that relate to forms A_I and A_2 respectively.

At $t \to \infty$ $\lim i(t) = i_\infty$

$$\lim \frac{1}{U_{1,2}} \cdot \frac{d}{dt} \int_0^t f_{1,2}(t-\tau) \cdot i(\tau)\, d\tau = \frac{a \cdot i_\infty}{nFC_{1,2}\mathcal{D}_{1,2}} \tag{24}$$

so that the static current-voltage characteristics of a diode will be

$$E_\infty = E_T \left[3 \ln \left(1 - \frac{a i_\infty}{nFSC_{02}\mathcal{D}_2} \right) - \ln \left(1 - \frac{a i_\infty}{nFSC_{01}\mathcal{D}_1} \right) \right] \tag{25}$$

Electrolytic rectifying diodes handle very low currents, from hundredths or tenths of a microamper to hundreds microampers which gives an advantage (at low frequencies) over diodes using other techniques. Electrolytic diodes of this type underlie elements which realize operations $\frac{1}{\sqrt{p}} x$ and $\sqrt{p} x$ [5] in a wide range of infralow frequencies as well as negative resistors.

Electrolytic and electrokinetic non-resonance sensors of low variable and discontinuous pressures, sensors of pressure gradients (in particular electroacoustic transducers of infralow and acoustic frequency bands that can operate at very high static pressures) and acceleration sensors have, as follows from Fig. 3, 4, 5, advantages in terms of possible frequency bands and coverage of low frequencies, although they are not as sensitive as piezoresistive transducers. The input/output characteristics of electrokinetic transducers are linear in a wide range of amplitudes. For free ran

$$E_{out} = \frac{K_g \xi \varepsilon \Delta p_{inp}}{4 \pi \mu \lambda_\sigma} \tag{26}$$

where ξ is electrokinetic potential, $\varepsilon, \mu, \lambda_\sigma$ . . are the dielectric factor, dynamic viscosity and the resulting electroconductivity of the polar liquid, K_g . . . is the factor which incorporates the hydrodynamic properties of the semi-permeable membrane. The amplitude-and-frequency response of that transducers is

$$E_m = \frac{\xi \varepsilon}{4\pi\mu\lambda_6} \cdot \frac{\omega \cdot \tau_M P_m}{\sqrt{1+(\omega \tau_M)^2}} \tag{27}$$

where τ_M - is the mechanical time constant of the transducer.

The input/output characteristic of liquid phase solion electrolytic transducers (which operate by mode (2)) for one of the configurations of the tiny cathode orifice

$$I_{out} = \frac{2FC_0 \delta^3 b}{3\mu a \pi^3} \cdot \Delta P_{inp} \left[1 - exp\left(- \frac{12\pi^2 \mathcal{D} \mu a^2}{\delta^4} \cdot \frac{1}{\Delta P_{inp}}\right)\right] \tag{28}$$

where C_0 is the concentration of non-basic carriers in free volume, $\mathcal{D}$ - is the diffusion factor, δ; a, b are geometric dimensions of the slot orifice, μ - is the dynamic viscosity of the solution. By changing the shape of the cathode orifice one can obtain various amplitude responces (linear, logarithmic, etc.).

Versions of these elements are employed as seismic sensors, sensors of biological parameters, acoustic receivers of infralow frequency, etc. The electrokinetic devices have highly linear amplitude responses and the widest dynamic range, while electrolytic ones have very low frequency band. However, at present both are inferior to new piezoelectric and piezoresistive sensors in terms of the operational temperature range.

Various types of electrochemical (a/c, contact and contactless, galvanic) concentration meters for solutions and gases are now manufactured on a large scale and used in laboratories and industry.

Another group is made by electrochemical devices that need further study and improvement. We can expect that scientific research will reveal new fields of application where such transducers will be aadvantageous. This group may incorporate, e.g.:

I) electrolytic sensors of vibrations;

2) electrokinetic sensors of vibrations;

3) low voltages indicators;

4) infralow frequency liquid phase electronic amplifiers;

5) electrolytic and electrokinetic data processing transducers;

6) solid phase electrolytic static switch and power amplifiers etc.

In the writers opinion this field of control needs research along these main lines:

a) new basic principles that would expand technological prospects of electrochemical devices;

b) characteristics of sensors, transducers and other units and most practical field of their applications

c) theory and calculation procedures for electrochemical devices of various types;

d) circuits with electrochemical elements to be incorporated in automatic control and monitoring;

e) most advanced technology of electrochemical elements manufacturing with a view to securing their high reliability.

Table I

Types of integrating elements	Effect used	Output physical quantity	Read-out
	Redox systems with inert electrodes		
LGas phase	a) hydrogen evolution absorption in electrochemical reactions;	a) change of volume (or pressure) difference over passive electrodes of opposite polarities;	Visual
		b) an electrolytic contact switching;	Electric
	b) change of properties of an electrode due to absorption of hydrogen formed in electrochemical reactions;	a) change in electrode electric conductivity;	Electric
		b) change in concentration e.m.f.	
2.Liquid phase (solion)	Change of the components concen9 tration distribution in the electrolyte cell.	a) change in concentration e.m.f.	Electric
		b) change in electrolyte optical density	Visual or electric
		c) change of diffusion saturation current	Electric

I	2	3	4
	<u>Redox systems with electrochemically soluble electrodes</u>		
3. Solid phase	a) change of electrode mass due to electrolytic plating or deplating;	a) change in electrode electric conductivity;	Electric
		b) change in optical density of a thin film electrode;	Photoelectric
		c) change in electrode linear dimensions;	
		d) change in electrode weight;	Electric
		e) change in electric conductivity of a cell (due to change in electrode geometric dimensions)	Electric
	b) change in equilibrium potential when the anodeactive metal dissolves over an inert support	Change in voltage drop on electrochemical cell	Electric

Conclusion

Comparison of characteristics leads one to believe that electrochemical devices based on different redox reactions can be used in sensices and data processing, in control of processes characterized with very low frequencies below tenths or hundredths of a cycle/sec, very low currents (to fractions of μA), very long time intervals (up to tens or hundreds of days and more). Very small sizes and low power consuption can be secured.

Such principles can underlie adaptive elements where microsecond pulses are used for non-destructive read-out.

References

I. А.П. Шорыгин Электрохимические элементы /основные особенности; классификация/ ЭИКА вып. 8, изд "Энергия", I967.
2. **A.P.Shorygin** Automation and Remote Control, No 8,1966.
3. В.С. Боровков, В.В.Трейер Электрохимические аналоговые запоминающие устройства /редактор А.П.Шорыгин/, изд.ВЗЭИ, М., I967.
4. А.П.Шорыгин Исследование входных, выходных и передаточных характеристик твердофазных канальных электролитических триодов. Электрохимия, т.IУ
5.Р.Ш.Нигматуллин Теория электрохимического диода, ДАН т. I50, 600, I963, №3
6.А.П.Шорыгин, Э.В.Казарян Счётчики машинного времени, Труды АН Арм.ССР I968 Вып. 4.
7. Х.Ф.Гуссейнов, М.С.Касимзаде Переходные процессы в электрокинетических сейсмоприемниках, Автометрия, стр.I9,I966, №4

39.1

RATIONAL ALGORITHM OF CONTROLLING THE THERMAL CONDITION OF BLAST FURNACE USING COMPUTERS

E.L.Suchanov, V.S.Shvidki, B.I.Kitaev, Ju.G.Yaroschenko, Ju.N.Ovchinnikov, V.G.Lisienko

(The Urals Polytechnical Institute, Sverdlovsk, USSR.)

The blast furnace is a complicated multidimensional controlled plant having distributed parameters the estimation of which is a very difficult problem. The most difficult thing here is to find out the criteria and ways of estimation of the blast furnace heat regime which is the main controlling factor for reduction processes of the blast furnace operation. The very notion of heat regime or thermal condition of the blast furnace has no uniform interpretation and needs more accurate definition.

The modern conception of the fundamental laws governing the heat exchange in the blast furnace run at combined blast is presented in recently published works[1,2]. These two and earlier studies[3] have brought to light S - shaped nature of changing temperatures over the height of the blast furnace (fig.I), the temperature of gas and burden at its middle levels varying only slightly and approximating to each other. Such an experimentally determined dependence leads to conclusion that heat exchange processes are concentrated in the upper and lower steps of intensive heat exchange. There is an intermediate region between these steps with mean temperature being practically constant under particular conditions of melting.

Under a certain stability of this temperature with respect to time and in a certain height range the intermediate region is a kind of a damper that removes the direct influence of heat exchange steps on each other. This accounts for revealed by us property of non-interaction of the thermal operation in the upper and lower sections of the blast furnace.

The property of non-interaction becomes apparent in different and sometimes opposite response to the same controlling factors displayed by the upper and lower sections of the furnace (we mean here both statics and dynamics of transient heat processes).

Relative independence and great difference in the nature and character of heat exchange in various furnace zones is reflected in the mathematical description of these phenomena[I].

As a result of the abovementioned the authors came to a conclusion that because of the property of non-interaction it would be wrong in the main to evaluate the thermal condition of the furnace as a whole. Also it would be wrong to judge of the temperature regime of the furnace according to the thermal condition of the hearth without taking into account the temperature profile of the furnace shaft, as sometimes the case may be. It goes without saying, one can't do vice versa, too.

Therefore, when dealing with a blast furnace there must be only differentiated approach to the estimation of the thermal condition in the upper and lower sections of the furnace. In carrying out such independent inspection over the thermal regime of the furnace it would be helpful to devide the entire operating volume of the furnace into two independent zones - the upper and the lower, the relative boundary between them lying on quite definite isothermic surface corresponding to the mean temperature or to some other most stable with respect to time temperature of the intermediate region.

It was assumed that the relative boundary between the upper and lower furnace zones is temperature t_0 (fig.I) which characterizes the beginning of the intensive evolution of the endothermic reaction of carbon dioxide reduction in the ore ridge area. This temperature usually ranges from 850°C to 920°C depending upon the type of iron produced, the blast composition and some other quite definite operating conditions of the furnace.

The mean temperature over the mass $\bar{t}_{ш}$ of the burden volume contained between the charging level and the assumed division line may serve as a general quantity estimation of the thermal condition of the upper zone of the furnace (in fig. I,a the volume in question is shaded). The relative value of the temperature $i_{в}=\bar{t}_{ш}/t_{o}$ was termed by the authors the index of the temperature profile of the upper section of the furnace. The revealed laws of heat exchange in the upper zone of the furnace[I] make it possible to determine index $i_{в}$ from the current information about the temperature $t_{к}$ and consumption of top gas $G_{к}$ and also the consumption of burden materials charged $G_{ш}$. Besides, it is necessary to have the information about the heat exchange coefficient α, specific heat of the burden and top gases, the average shaft cross-sectional area S and the average height of the furnace upper zone assumed for the calculation H. The calculation are made according to the following formula[4]:

$$i_{в}=1-\left[\frac{t_{к}-t_{шк}}{t_{к}-m(t_{шк}+\Delta t_{o})}\cdot\frac{1-\exp(-A)}{A}-\frac{m\cdot\Delta t_{o}}{t_{к}-m(t_{шк}+\Delta t_{o})}\right], \quad (1)$$

where $t_{к}$- and $t_{шк}$ = temperature of gas and burden on the top of the furnace, °C;

$m=0,5\left(\frac{W_{шк}}{W_{к}}+1\right)$ = average burden-gas thermal capacity flow ratio for the upper zone;

$W_{шк}$- $W_{к}$ = gas and burden flow thermal capacities on the top level, w/°C;

$\Delta t_{o}=t_{o}-t_{шo}$ = temperature difference between gas and burden on the division line between the zones, °C;

$A=\frac{\alpha\cdot H\cdot S}{W_{ш}}(1-m)$ = auxiliary coefficient.

Index $i_{в}$ is a generalized parameter of the thermal condition of the upper blast furnace zone. Recently this parameter has become more essential for the heat regime inspection and control of the blast furnace due to the wide use of oxygen for the intensification of the blast furnace production. With the oxygen enrichment of the blast the volume of hearth gases is decreased, which reduces warming up the burden, thereby slowing down the course of the main reducing reactions. Later on you will see that it is impossible to solve the problem of optimization of the blast furnace without using index $i_б$.

As a generalized parameter of the thermal condition of the lower blast furnace zone the authors suggest index $i_н$ equal to the ratio of actual quantity of heat $Q_{факт}$ in the tapped products to the heat consumption Q_0 theoretically required for physical and chemical heating of the smelting products of the predetermined composition:

$$i_н = \frac{Q_{факт}}{Q_0} = \frac{Q_{чуг} + Q_{шл} + Q_{Si,Mn,P...}}{Q_0}, \qquad (2)$$

where $Q_{чуг}$ = iron enthalpy allowing for the melting heat;
$Q_{шл}$ = slag enthalpy without the slag formation heat;
$Q_{Si,Mn,P...}$ = the heat consumed for reduction of silicon manganese, phosphorus and other elements.

Index $i_н$ which is determined according to the amount, temperature and chemical analysis of the tapped products is quite a reliable criterion of the heat condition of the lower part of the furnace and may be taken as a reference input for the automatic control system. The efficiency of the automatic control system greatly increases if we consider not only the past but also the current and even predicted thermal condition of the lower part of the furnace. Thus to determine index $i_н$ during the interval between the tappings of iron we may use the balance inspection method applying the following equation[1]:

$$i_н = \frac{1}{Q_c}\left\{\frac{V_д \cdot \sigma}{(V_{O_2})_к - (V_{O_2})_д}\Big[(c_{N_2} \cdot t_д - 1312) \cdot N_2 + (c_{O_2} \cdot t_д + 7840) \cdot O_2 + 1{,}244 \cdot 10^{-3} \cdot c_{H_2O} \cdot t_д \cdot \varphi_д - 11{,}53\,(1 + 0{,}154 \cdot \eta_{H_2}) \cdot \varphi_д\Big] + \Big[c_A \cdot G_A + c_к \cdot K \cdot (1 - 0{,}01 \cdot Л)\Big] \cdot t_o - 6912\,(1 - 0{,}411 \cdot \eta_{H_2}) \cdot Г - 31750 \cdot Fe_o \cdot \tau_d - 104500\,\frac{d}{P_{сут}} - 24600\right\}. \qquad (3)$$

For the calculation according to the above equation it is necessary to have some current information sbput the blast and top gas parameters (altogether 14 variables) and besides some recurrent data about the changes in the burden composition and some reference data. The main coefficient τ_d characterizing the degree of direct reduction can be periodically defined more precisely by comparing the results of the calculation according to the formulae (2) and (3). Index $i_н$ obtained from equation (3) may be considered a predicted value since the changes in the top gas composition to a certain degree leave behind the evolution of transient processes in the thermal condition of the lower part of the furnace.

The works of C.Staib and J.Michard (IRCID) repeated recently by P.Jourde and C.Remont[5] have shown the practical fitness of the parameter W_u similar to index $i_н$ for the inspection and control of the thermal condition in the lower part of the furnace.

The peculiarities of the blast furnace operation, namely the discrete and cyclic charging of raw materials into furnace impose certain conditions on the methods of calculation of indexes $i_в$ and $i_н$ from equations (2) and (3). The analysis of all the input information led to conclusion that charging cycle time, namely the period covering 5 - 7 chargings (depending upon the agreed charging program) should be considered an optimal interval between the successive calculations of indexes i_B and i_H . It requires averaging the information during the charging cycle. The generalized parameters i_B and i_H determined under such conditions characterize the thermal condition of the upper and lower parts of the furnace during the previous charging cycle, that is, during the last half an hour of the furnace operation.

The revealed non-interaction in thermal operation of the upper and lower zones, of the furnace as well as the above shown possibility of independent estimation of the thermal condition of the two zones enable us to consider the blast furnace to consist of two interconnected but self-contained controlled plants with their own static and dynamic properties. Such quite new treatment of the blast furnace as a controlled plant reflects the nature of the blast furnace melt more accurately thereby increasing the ways of controlling this process.

Differentiated approach to the upper and lower thermal zones of the furnace as independent controlled plants has enabled us to make an analysis of their static and dynamic properties when using different controlling factors. Fig. 2 and 3 show the static and dynamic characteristics determined by calculation, the static characteristics being calculated according to known dependences for the settled thermal conoition of the furnace while the dynamic characteristics are determined by the method of approximate calculation of transient processes.

A slight initial cooling of the upper and part of the furnace after a stopped rise of specific coke consumption is accounted for by gradual increase (fig.3) of the thermal capacity of burden materials with changing the relative proportion of coke. Warming up the shaft of the furnace begins only after the burden of a new composition reaches the tuyers and as a result of combustion of additional coke the yield of hearth gas increases.

The initial drop of index i_H during the natural gas supply (fig.3) is due to the heat consumed for converting the injected fuel. Later on due to increasing hydrogen content in reduction gases the amount of coke carbon consumed for direct reduction decreases and a slight increase of the proportion of coke burning out at the tuyers results (the predetermined fuel consumption and other controlling factors except the one under review are assumed to be unchanged). Similar results were obtained experimentally[7], the injected fuel being masont.

The revealed difference in static and dynamic properties of the controlled plants we have examined enables us to combine the controlling factors into the complex actions which have nevessary selective (local) effects on the heat regime of only one zone of the blast furnace. Such quick-response complex actions which are to correct the detected deviations in the heat regime of individual blast furnace zones are called corrective.

As an example of such simplest corrective action which is efficient.for the upper zone of the furnace and practically neutral for its lower zone we may give changing the oxygen content in the blast (in case $t_{\partial} > 1000°C$). Quite opposite is the response observed when applying another simplest corrective action, namely the change in the blast temperature (Fig.3).

The corrective actions are composed in such a way that while applying them the thermal condition of the main (limiting) zone of the furnace should return within the limits of the zone of permissible deviations from the optimum during one or two charging cycles, and afterwards should not come beyond the above limits. It should be born in mind that each carefully selected corrective action represents not only a set of necessary discrete changes of different controlling factors but also a definite programme of action (sequence and speed of change in blast parameters must be indicated). It makes possible to take into account the dynamics of transient processes and the imposed limitations when applying actions "from below".

There may be several corrective actions exerting similar influence upon the thermal condition of the furnace. The selection of the optimum variant is due to the available possibilities of using various controlling factors and depends also on such limiting conditions as the necessity to retain gas dynamics of charge column and reduction potential of gases. For particular conditions of melting in the blast furnace it is helpful to calculate in advance all the possible variants of solving problem, and to present the obtained results as ready-made tables with logical schema of selecting necessary corrective action.

Thus the application of generalized parameters $i_в$ and $i_н$ and the possibility of using quick-response corrective actions proved above makes it possible to raise the problem of stabilization and optimization of the heat regime of the blast furnace.

As it follows from the very notion of the index $(i_н)_{опт}$ = I, the value of the index $i_н$ corresponds to the optimum thermal condition of the lower section of the furnace.

To find the optimum value of index $(i_в)_{опт}$ special investigations[6] were made. These investigations have revealed the extreme character of the dependence between the thermal condition of the upper section of the furnace and its output (Fig.4), the configuration of the curves being determined by the particular conditions of furnace operation. For experimental data shown in Fig.4 the optimum values of index $(i_в)_{опт}$ corresponding to maximum output of the furnace range from 0.65 to 0.77. For other furnaces and conditions of operation the extreme character of the dependence under investigation is retained but the values of index $(i_в)_{опт}$ may be different.[6]

The results obtained affirm convincingly enough considerable influence of the thermal condition of the shafts of modern furnaces using oxygen injected fuel on the main indexes of the blast furnace process. It proves that the optimum variant of melting technique in the blast furnace (forced smooth operation of the furnace when melting iron of pre-determined composition with least amount of coke consumed) is possible only if the heat regime is stabilized at the optimum level not only in the lower section of the furnace but at the same time in the upper one as well.

The suggested algorithm of controlling thermal condition of the blast furnace is as follows: The optimum value of index $(i_{в})_{опт}$ is determined statistically or by some other method and the value Q_0 is specified in the same way. The current values of indexes $i_{в}$ and $i_{н}$ are recurrently calculated once during the charging cycle. Then the deviations $\Delta i_{в} = (i_{в})_{опт} - i_{в}$ and $\Delta i_{н} = (i_{н})_{опт} - i_{н}$ are defined, and according to the sign and the value of the latter the necessary corrective actions are calculated (or selected from ready solutions) by a computer allowing for additional conditions.

The agreed decision about the value and sequence of changing various controlling factors holds true until the steady deviations from the heat regime rates of the upper or lower sections of the furnace are fixed again. In this case the pre-selected controlling action is repeated or replaced by the new one according to the particular conditions of furnace operation.

In order to increase reliability of the taken decisions about the selection of controlling actions it was decided:

1. Definite zones of insensitiveness whose range somewhat exceeds permissible miscalculations in measuring input values must be established just the same as in the case of the discrete control systems. If the o permissible deviations of measurements from the range of controlled values of these indexes are $\pm$ 2,5 % for index $i_{в}$ and $\pm$ 6 % for index $i_{н}$ (for the index $i_{в}$ in the range of 0.65–0.85 and for index $i_{н}$ in the range of 0.75–1.25), the zones of insensitiveness may be assumed to be equal $\pm 0{,}005 \cdot i_{в}$ and $\pm 0{,}03 \cdot i_{н}$.

2. To eliminate the influence of occasional variations in the heat regime of the furnace and to secure the necessary holding for the estimation of the results of

steps taken, the authors suggest to take into account only those deviations $\pm \Delta i_{\delta} > 0.005$ and $\pm \Delta i_{H} > 0.03$, which are retained at two successive calculations of these values. Only such steady deviations may be considered discrete controlling signals on the basis of which one may estimate the changes in the thermal condition of the furnace and re-consider the decisions taken before.

Therefore, the rate of the work of a computer can't be pre-determined but is defined by technological process itself. For example, for the blast furnace of 1513 m^3 the charging cycle which was assumed to be basic period for all main calculating operations is 15-30 minutes if the output of the furnace is high and 40-50 minutes if it is low. The possible replacement of controlling commands by others in 15-50 minutes is quite acceptable in controlling blast furnace process and requires no quick-acting computers.

It is known that the main controlling factor which determines the heat regime of the melting in the blast furnace is the specific consumption of coke or the amount of ore loads. Considerable persistence and some other properties of this action, however, make it difficult to use it for stabilization of thermal condition in different zones of the furnace. In connection with this the authors suggest that the actial influence of the fixed coke consumption (or ore loads) on the heat regime of different zones of the furnace be corrected through small changes of the blast temperature Δt_{∂} , its humidity $\Delta \varphi_{\partial}$, the content of oxygen in the blast ΔO_2 and the consumption of natural gas. $\Delta \Gamma$. It is not the suggestion of advisability of applying these controlling factors, which has long been proved in pracce, that is new but the possibility to arrange them into purposeful controlling actions exerting necessary influence on the thermal condition of different zones of the furnace determined by us.

Below we give the example explaining the selection of corrective actions:

Assume that under particular operating conditions of the blast furnace of 1242 m^3 (K = 550 kg/ton iron , t_∂ = 1000°C, φ_∂ = 15 g/m^3 , O_2 = 22 %, Γ = 30 m^3/ton iron , $(i_\ell)_{опт}$ = 0.78) the control system detected in succession deviations + $\Delta i_\delta > 0.005$ and - $\Delta i_H > 0.03$, which shows steady indications of warming-up upper section and cooling lower section of the furnace. For correcting the revealed deviation of heat regime from optimum one can apply the corrective action providing for increase in oxygen content of the blast and at the same time decrease in blast humidity : $\Delta O_2 = + 0{,}9\%$, $\Delta\varphi_\partial = -4\, g/m^3$ of the blast. The rate of changing these controlling factors is assumed to be the same, since their dynamic properties coincide. Such corrective action, in fact, will not affect the aerodynamic and reducing processes of melting but under the influence of this action the thermal condition of upper and lower sections of the furnace will return to its optimum values.

The block diagram of the UPI (the Urals Polytechnical Institute) system (Fig.5) is suggested as one of the variants of carrying out the conception of independent inspection and local control of the thermal condition of the upper and lower sections of the furnace. Besides conventional systems of control and stabilization of different input parameters we suggest that a simple computer be used which calculates values of indexes i_ℓ and i_H , determines the deviations of these indexes from their optimum values $\pm\ \Delta i_\delta$ and $\pm\ \Delta i_H$ once during the charging cycle, and according to the sign and value of the latter selects the necessary corrective action allowing for imposed limitations. The determined decision is carried out by means of changes according to a definite programme of targets for controllers and stabilizers of natural gas consumption and blast parameters or by means of delivery of recommendations for changing specific coke consumption.

REFERENCES

1. Б.И.Китаев, Ю.Г.Ярошенко, Б.Л.Лазарев. Теплообмен в доменной печи. Изд. "Металлургия", 1966.

2. B.I.Kitaev, Yu.Yaroschenko, V.D.Suchkov. Heat Exchange in Shaft Furnaces. Pergamon Press, Oxford, 1967.

3. B.I.Kitaev, Y.C.Yarochenko, B.L.Lasarev. Etat actual de la theorie des echanges thermiques dans le haut fourneau. Troisiemes Journees Internationales de Siderurgie, Luxemburg, 1962.

4. Ю.Г.Ярошенко, Б.И.Китаев, В.С.Швыдкий, Б.Л.Лазарев, Е.Л.Суханов. Контроль теплового состояния шахты доменной печи. Сб. "Автоматизация доменного производства", Изд. "Металлургия", 1966.

5. P.Jourde, C.Remont, C.Staib, N.Jusseau, A.Schaller. Control et reglage automatique d'un fourneau par un calculateur industriel. Journees Internationales de Siderurgie, Amsterdam, 1965.

6. Ю.Г.Ярошенко, Б.И.Китаев, Б.Л.Лазарев, В.С.Швыдкий, И.А.Тациенко, Б.М.Герман. Тепловое состояние зоны непрямого восстановления доменных печей и его контроль. Черметинформация, Серия 4, инф. 15, 1966.

7. P.Guillemain, N.Jusseau, P.Bece. Conduite automatique et etude des regimes transitoires de deux hauts fourneaux alimentes et agglomere. Revue de Metallurgie, 1967, Nr 11.

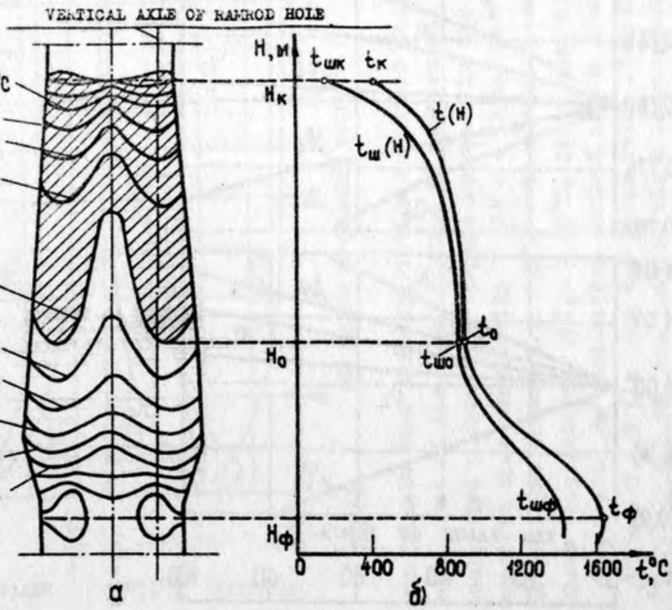


Fig.I

The nature of temperature profile of the blast furnace of 1242 m^3 when melting pig-iron and under normal operating conditions: a - the working space of the furnace is divided into upper (shaded) and lower heat zones; b - the curves of changing temperatures of countercurrents of charge and gas along the axle of ramrod hole.

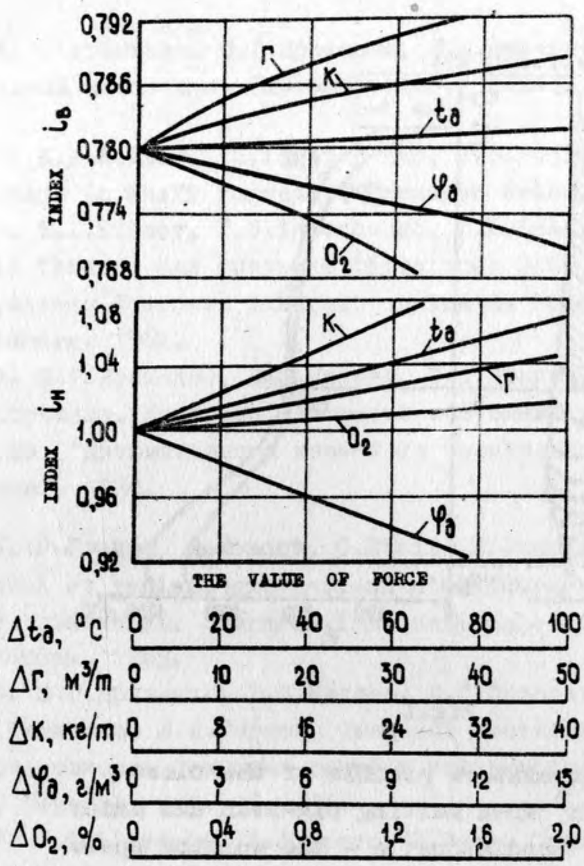


Fig. 2

Calculated statistical characteristics of individual (different) controlling actions for particular operating conditions of the furnace of 1242 m^3 when changing specific coke consumption K from 550 to 590 kg/t of iron, blast temperature from 1000 to 1100°C, blast humidity $\varphi_д$ from 7 to 22 g/m³, oxygen content in the blast O_2 from 21 to 23 % and natural gas consumption Г from 10 to 60 m³/ton iron.

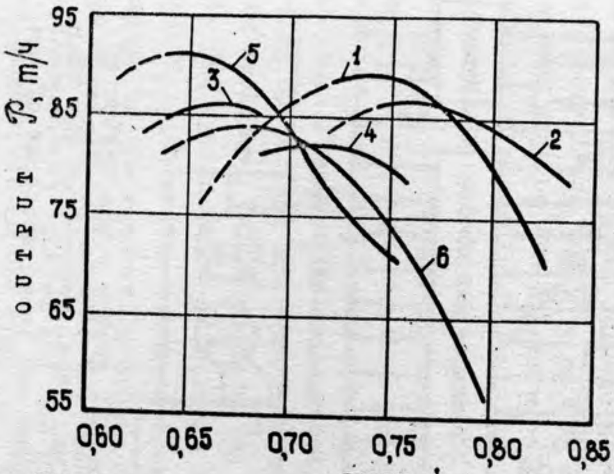


Fig.4

The relationship between index i_B and the output of the blast furnace of 1242 m^3 defined, according to average data, for 24 hours for 6 months' operation: I-, 2- and 3 - October, November and December, 1965; 4-, 5-, 6 - January, February and March, 1966.

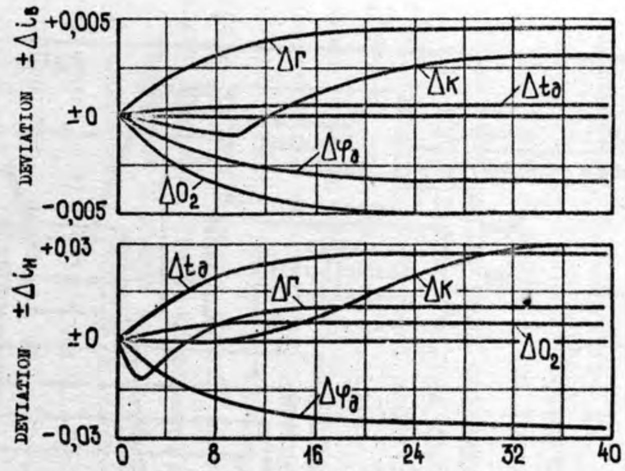


Fig.3

Approximate time characteristics for upper and lower heat zones of the blast furnace as independent controlled plants with stepped increase in the specific coke consumption $\Delta K = + 10$ kg/*ton* iron, natural gas consumption $\Delta \Gamma = + 10$ m^3/*ton* iron, blast temperature $\Delta t_д = + 40^{\circ}C$, blast humidity $\Delta \varphi_д = + 4$ /m^3 and oxygen content in the blast $\Delta O_2 = + 0.6$ % (in case the index value is optimum $(i_б)_{опт} = 0.78$, the initial blast temperature $t_д = 1000^{\circ}C$ and other particular conditions).

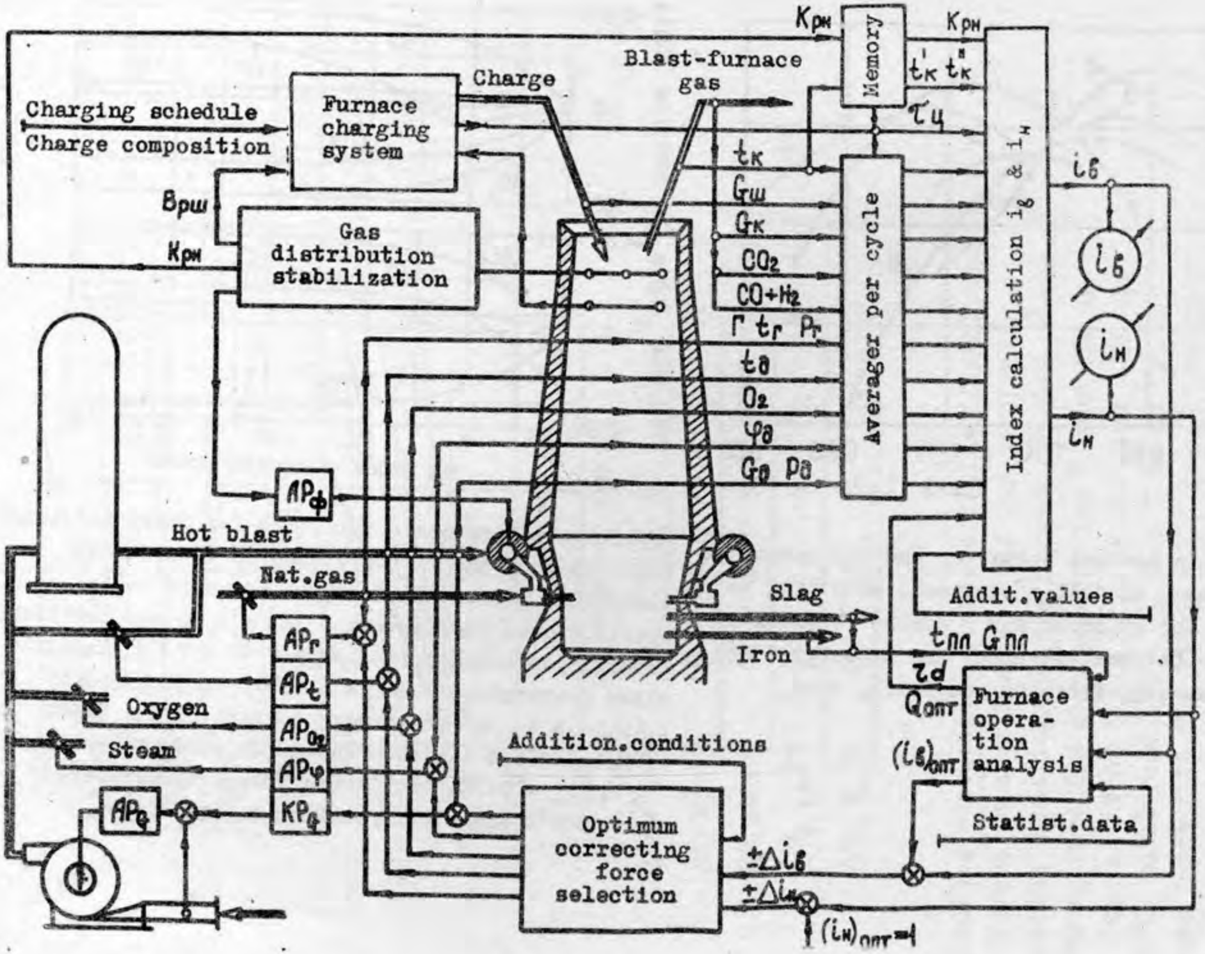


Fig.5 The block diagram of the UPI system – the system of independent inspection and local control of the thermal condition of the upper and lower sections of the blast furnace using combined blast.

Rational algorithm of controlling the thermal condition of blast furnace using computers

E.L.Suchanov, V.S.Shvidki, B.I.Kitaev, Ju.G.Yaroschenko, Ju.N.Ovchinnikov, V.G.Lisienko
(The Urals Polytechnical Institute, Sverdlovsk, USSR).

The autonomy in heat operation of upper and lower stages in a blast furnace called for the necessity of different evaluation of their heat conditions. „The index i_B" is offerred as a general variable of the heat condition in the upper part of the blast furnace and „the index i_H " is proposed for controlling heat conditions in the lower part of the furnace. These variables are calculated only once per cycle of the furnace charge according to usual information about the technological process.

The blast furnace should be considered as two connected but independent objects of controlling with their own static and dynamic characteristics. When analysing those characteristics the conclusion about roal combination of such controlling factors is obtained. The general influence of these factors on the furnace heat conditions has its necessary local action.

Both statics and dynamics of transient processes in objects of controlling are considered. Each controlling effect is a certain programme of the necessary digital changes in the temperature and humidity of the blast, in the higher oxygen concentration and in the expense of the injected fuel. The lower specific expense of the coke can be assumed because of economic consideration. All the calculating operations are performed by the information controlling machines. In the article the possibility of stabilization and optimum heat conditions in the modern blast furnace which is worked at a combined blast is proved. The bloc-diagram of the automation system which is resulted from the idea about the optimum in the blast furnace process under the independent control and the local stabilization of the heat conditions in the upper and lower stages in the furnace is given.

39.6

TIME SUB-OPTIMUM CONTROL OF THE WORK OF CRANES WITH SPECIAL REGARD TO ITS REALIZATION IN PRACTICE

M. Sc. Roman Gorecki
Academy of Mining and Metallurgy
Department of Automation and Industrial Electronics
Kraków, Poland

1. Introduction

There is already ample literature available in the field of optimum control and new items are constantly being published. As a rule these are theoretical research works and seldom concerned with a practical solution. The methods suggested in literature usually lead to very complex solutions requiring the use of computers or necessitating carrying out of numerous measurements.

The present work is an attempt to solve time sub-optimum control by means of a specific process with a special regard to practical realization.

2. Description of the process

We shall concern ourselves with the design of a system for an optimum, meaning "minimum time", control of the work of a crane. We shall limit ourselves to the investigation on the optimum control of cranes of the travelling bridge type, i.e. those in which we have at our disposal three independent drives for transporting the load. Each of them shifts the load along a different axis of the Cartesian system. A situation of this type is presented in Fig.1. The load may be shifted in the vertical plane /movement C/ and transferred to an optional position in a horizontal plane by means of drives A and B. The load may be shifted simultaneously in three directions. This is often used by the operator. In order not to set the load in excessive swinging motion the operators apply low travelling rates. In spite of this a-fter arriving above the destination the load will as a rule oscillate enough to make it impossible for the operator to lower it to the foreseen place required without damping of the oscillations.

When the crane is in constant use, e.g. in the case of harbours concerned with reloading of goods, it is worthwhile considering the possibility of modernizing the control, i.e. changing the control in order to reduce the time required for transporting loads as well as making greater use of the installed power.

To begin with a somewhat simpler case than the one described above will be considered. It is assumed that the load hangs below the trolley on a weightless rope of a constant length and has the whole mass concentrated in one point at a distance "1" from the point of suspension, otherwise it constitutes a mathematical pendulum. Controlling of the trolley motion is limited to the direction "Z" only, /Fig.2/. The trolley drive is located beyond the trolley. The drive consists of a motor with rotations controlled within certain limits, transmission and coupling through with the rope hauling the trolley is coupled with the operating drive or brake. In addition, it is assumed that the trolley mass is so small in relation to the inertia of masses in motion as to be negligible. The shifting of the point of the engagement from a certain initial position is marked by the letter "y" and the shifting of the load along the axis "x" by the symbol "x" /Fig.3./.

Several forces will be acting on the load being in motion, i.e. the force of gravity "G", force of inertia "B", resistance of dynamic friction "T", and the force of the rope reaction "R". The equilibrium conditions will be written by projecting the forces in the direction normal to the track of motion, i.e. along the line

$$G_n - R = 0 \qquad 2.1$$

where G_n is the projection of the force of gravity in a direction perpendicular to the track of motion

R - reaction in the rope.

The next condition of the equilibrium of the forces projection in the "x" direction will be given:

$$-B_x - T_x + G_s \cdot \cos \alpha - G_n \sin \alpha + R \sin \alpha = 0 \qquad 2.2$$

B_x is the component of the force of inertia in the direction of the "x" axis. It is proportional to the load mass "m" and its acceleration in the direction of this axis.

$$B_x = mx'' \qquad 2.3$$

T_x is the component of resistances to motion proportional to the rate of travel of the load in the direction of the "x" axis.

$$T_x = r\,x' \qquad 2.4$$

where "r" is the coefficient of the resistances to motion.

G_s is the component of the forces of gravity "G" tangent to the track of motion.

$$G_s = G \sin \alpha \qquad 2{,}5$$

Taking into account the relationship 2.1 to 2.5 in the formula 2.2. one may write:

$$- mx'' - rx' + G \sin\alpha \cdot \cos\alpha = 0 \qquad 2.6$$

For small angles " α " one may assume

$$\cos\alpha = 1 \qquad 2.7$$

and since from Fig. 3 one sees that:

$$\sin = \frac{y - x}{l} \qquad 2.8$$

therefore for small deviation the relation 2.6 will assume the form:

$$- mx'' - rx' + G \cdot \frac{y - x}{l} = 0 \qquad 2.9$$

After ordering one obtains:

$$x'' + \frac{r}{m} x' \frac{G}{m \cdot l} /x - y/ = 0 \qquad 2.10$$

and since

$$G = m \cdot g \qquad 2.11$$

therefore:

$$x'' + \frac{r}{m} x' + \frac{g}{l} x = \frac{g}{l} y \qquad 2.12$$

By designating:

$$\frac{r}{m} = a$$

$$\frac{g}{l} = b \qquad 2.13$$

one obtains:

$$x'' + ax' + bx = by \qquad 2.14$$

Since in fact the rate of the trolley travel is controlled therefore its shifting equals:

$$y = \int v \cdot dt \qquad 2.15$$

The rate of the trolley travel can assume values contained within the interval from 0 to $v_{max} = u > 0$, taking into account 2.15 in 2.14:

$$x'' + ax' + bx = \int_0^t bv\, dt \qquad 2.16$$

Differentiating both sides of 2.16 one obtains:

$$x''' + ax'' + bx' = bv$$

$$0 \leq v \leq v_{max} \qquad 2.17$$

The plant described by the above equation should be transported from the initial state:

$$x'' /0/ = x' /0/ = x/0/ = 0 \qquad 2.18$$

at the instant $t_o = 0$ to the terminal state:

$$x'' /t_k/ = x' /t_k/ = 0 \ ; \ x /t_k/ = x_k \qquad 2.19$$

at the instant "t_k" in the possibly shortest time.

3. <u>Application of the "maximum principle"</u>

Since equation 2.17 is a linear one the control function "v" for obtaining the time optimum course should according to the "ma-ximum principle" be constant in the intervals and assume extreme values alternately. The maximum principle gives us in this case the necessary condition of optimality in the sense of minimum time.

The third order equation 2.17 is converted into a first order system of equation which one obtains by substituting:

$$x_1 = x \ ; \ x_2 = x' \ ; \ x_3 = x'' \ . \qquad 3.1$$

a-nd now we obtain:

$$x_1' = x_2$$

$$x_2' = x_3 \qquad 3.2$$

$$x_3' = bx_2 - ax_3 + bv$$

The Hamiltonian function:

$$H = \sum_{k=1}^{3} \psi_k \frac{dx_k}{dt} \qquad 3.3$$

taking into account 3.2 we have:

$$H = \psi_1 \cdot x_2 + \psi_2 \cdot x_3 + \psi_3 /-bx_2 - ax_3 + bv/ \qquad 3.4.$$

This function will attain a maximum by a control in which

$$v = \mathrm{sign}\ /b\ \Psi_3/ \tag{3.5}$$

which in the case discussed is reduced to assuming the value:

$$v = \begin{cases} u \text{ for } \Psi_3 > 0 \\ 0 \text{ for } \Psi_3 \leq 0 \end{cases} \tag{3.6}$$

For determining Ψ_3 as a function of time we make use of the dependence:

$$\frac{d\Psi_i}{dt} = -\frac{\partial H}{\partial x_i} \qquad / \ i = 1, 2, 3 \ / \tag{3.7}$$

The following system of equations is to be solved:

$$\begin{aligned} \Psi_1' &= 0 \\ \Psi_2' &= -\Psi_1 + b\,\Psi_3 \\ \Psi_3' &= -\Psi_2 + a\,\Psi_3 \end{aligned} \tag{3.8}$$

wherefrom

$$\begin{aligned} \Psi_1 &= c_1 \\ \Psi_3'' &= -\Psi_2 + a\Psi_3' \\ \Psi_3'' &= \Psi_1 - b\Psi_3 + a\Psi_3' \end{aligned} \tag{3.9}$$

Finally the following equation is to be solved:

$$\Psi_3'' - a\,\Psi_3' + b\Psi_3 = c_1 \tag{3.10}$$

After solving equation 3.10 the function Ψ_3 is expressed by the formula:

$$\Psi_3\ /t/ = \frac{s_1^2c_3 + s_1c_2 - s_1ac_3 + c_1}{3s_1^2 - 2as_1 + b} \cdot e^{s_1t} + \frac{s_2^2c_3 + s_2c_2 - s_2ac_3 + c_1}{3s_2^2 - 2as_2 + b} \cdot e^{s_2t} + \frac{c_1}{b} \tag{3.11}$$

where:

$$\begin{aligned} c_2 &= \Psi_3'\ /0/ \\ c_3 &= \Psi_3\ /0/ \end{aligned} \tag{3.12}$$

If we assume $a = 0$, which in the case under discussion we may safely do, since the damping of the motion by the air resistance is minimum then the function ψ_3 assumes a simpler form:

$$\psi_3 /t/ = \frac{c_1}{b} /1 - \cos\sqrt{b}\ t/ + c_3 \cos\sqrt{b}\ t + c_2 \frac{1}{\sqrt{b}} \sin\sqrt{b}\ t \qquad 3.13$$

designating:

$$\sqrt{b} = \omega \qquad 3.14$$

$$\psi_3 /t/ = \frac{c_1}{\omega^2} /1 - \cos\omega t/ + c_3 \cos\omega t + c_2 \frac{1}{\omega} \sin\omega t \qquad 3.15$$

On the basis of equation 3.8 and 3.15 we may compute the remaining auxiliary functions:

$$\psi_1 = c_1$$

$$\psi_2 = /c_3\omega - \frac{c_1}{\omega} / \sin\omega t + c_2 \cos\omega t \qquad 3.16$$

$$\psi_3 = \frac{c_1}{\omega^2} + /c_3 - \frac{c_1}{\omega^2} \cos\omega t\ \ c_2 \frac{1}{\omega} \sin\omega t$$

The system of equations 3.16 may be presented as:

$$\psi_1 = c_1$$

$$\psi_2 = R \sin /\omega t + \alpha/$$

$$\psi_3 = \frac{c_1}{\omega^2} + R \cos /\omega t + \alpha/$$

where: 3.17

$$R = \frac{1}{\omega}\sqrt{/c_3\omega - \frac{c_1}{\omega}/^2 + c_2^2}$$

$$\alpha = \text{arc tg}\ \frac{c_2}{c_3\omega - \frac{c_1}{\omega}}$$

In the Cartesian system of axes ψ_1, ψ_2, ψ_3, the parametric equation 3.17 presents an ellipse laying in a plane perpendicular to axis ψ_1 which intersects this axis at point $\psi_1 = c_1$.

In the system of axes $\frac{\psi_2}{\omega}$, ψ_3 we will have a circle with a radius "R" and centre laying on axis ψ_3 at point $\psi_3 = \frac{c_1}{\omega 2}$, Fig.4. From the computations carried out the course ψ_3 as the function of time is known from which it follows that the character of the switching function "v" is known. Unfortunately, the Pontryagin method does not furnish us with information about the value of initial conditions c_1, c_2, c_3, therefore on the basis of this function we know only that function "v" will be reproducible together with period "T", but the length of the switchings is not known neither is the time after which the first and last switching will occur. From the character of the control function it results that it may be considered as a sum of shifted in time constant inputs of "u" values.

With the assumption of zero initial conditions the output signal x /t/ will be equal to the sum of responses to each input signal. From formula 2.17 one can compute the response of the system to the step input signal

$$v = \begin{cases} 0 \text{ for } t < 0 \\ u \text{ for } t \geq 0 \end{cases} \qquad 3.18$$

will be:

$$\begin{aligned} h\,/t/ &= u\,/t - \frac{1}{\omega} \sin \omega t/ \\ h'/t/ &= u\,/1 - \cos\omega t/ \\ h''\,/t/ &= u\,.\omega \sin \omega t \end{aligned} \qquad 3.19$$

The response of the system to the control signal composed of constant functions alternately equalling "u" and "0" will be:

$$x\,/t/ = u \sum_{k=0}^{n} /-1/^{k} . h\,/t - t_k\,/ \qquad 3.20$$

4. Optimum control

To determine the optimum control additional information is necessary. In the case of a second-order equation of state it is very convenient to draw the phase trajectories in the x, x' system of coordinates. In the considered case of a third-order equation the trajectories should be per analogiam considered in a special system x, x', x'' which we are unable to illustrate clearly on a plane.

The equation of the system considered is degenerated. This This is due to the fact that the value of the coefficient at "x" is equal to zero hence when observing the performance of the system in the coordinates x', x'' we shall have full informa-tion about all the derivates and the function itself.

In the x', $\frac{x''}{\omega}$ system the trajectories of the system considered will be circles with centres located on the axis "x'" at point $x' = v$.

Only the trajectories of the system around the extreme positions $v = 0$ and $v = u$ constitute the optimum control. From the initial and terminal conditions it results that the process must begin and end with a trajectory crossing through the origin of the system $\frac{x''}{\omega}$, x' shifting simultaneously x from x_0 to x_k, but this is not to be seen in Fig.4.

If the input signal $v = u$ acts on the system one will see that the leading point after time $t = T$ will return to the origin. The load will rest as at the moment of starting perpendicularly below the trolley, hence the shifting of the trolley and load will be the same and equal to:

$$x_T = x_T = u \cdot T \tag{4.1}$$

Therefore we see that for shifting the load over a distance

$$x = n \cdot x_T \qquad n = 1, 2, 3, \qquad 4.2$$

the necessary control will be limited to one switching period, This is in agreement with the result obtained by Pontryagin's method. This is a particular case in which the switching periods are equal to the period of function and the times of the trolley standstills are reduced to zero.

The number of switchings in an optimum control time are equal to:

$$i = 2\left[E^* / \frac{x_k}{x_T} / + 2 \right] \qquad 4.3$$

In Fig. 5 an example of a phase trajectory for $x_T \geq x_k \geq 0$ is shown. Owing to the relation

$$\alpha = \omega t \qquad 4.4$$

and the observation that, according to the initial and terminal conditions, the track of shifting the load is equal to the track of the trolley travel and the time of travel corresponds to the time of moving along the circular trajectory with a centre at $x' = u$ one may write:

$$x_k = u / T_1 + T_3 - T_2 / \qquad 4.5$$

and

$$\alpha = \pi - \gamma \qquad 4.6$$

and after taking into account the relation 4.4:

$$T_2 - T_1 = \frac{T}{2} - T_1 \qquad 4.7$$

$$T_2 = \frac{T}{2}$$

Reasoning thus:

$$T_k - T_1 = \frac{T}{2} \qquad 4.8$$

$E^*/z/$ designates a complete part from number z

Finally we obtain:

$$T_1 = \frac{x}{2u}$$
$$T_2 = \frac{T}{2} \qquad\qquad 4.9$$
$$T_k = \frac{x}{2u} + \frac{T}{2}$$

Determination of the optimum process for $2x_T > x_k > x_T$ can be carried out on the basis of the phase characteristics from the information already acquired which is presented in Fig.6.

The main indication when determining optimum trajectory is the fact that except the first and last intervals of time the sum of two consecutive switchings must be equal to the period T, according to Pontryagin's method.

From Fig.6 one may derive the relationship between angles α and δ :

$$\alpha = \frac{1}{2} \text{ arc ctg } \frac{2 - \cos\delta}{\sin\delta} \qquad 4.10$$

and for x from an arbitrary interval $/n - 1/\, x_T \leq x \leq n \cdot x_T$:

$$\alpha = \frac{1}{2} \text{ arc ctg } \frac{n - \cos\delta}{\sin\delta} \qquad 4.11$$

Knowing the magnitude of angle α corresponding to the intervals of the stillstand time and δ corresponding to the first and last time intervals of travel, the times of consecutive switchings may easily be completed. The operation total time will be:

$$T_{kn} = \frac{T}{360^{\circ}}\left[\, 2\,\delta + \alpha + (n - 1).360^{\circ}\right] \qquad 4.12$$

The load will be shifted to the point of destination:

$$x_k = \frac{x_T}{360^{\circ}}\left[\, 2\,\delta + (n - 1)(360^{\circ} - \alpha)\right] \qquad 4.13$$

With the increment of the assumed travel distance, i.e. with increment of "n", α aims at "0" and the operation time will tend towards a sum time of travel T_j which will be:

$$T_j = \frac{x_k}{u} \qquad 4.14$$

Since this is the shortest time during which the trolley is able to cover the assumed distance of shifting regardless of the oscillation of the load, the possible symmetric limitations of the control function $-u \leq v \leq u$ at a greater distance x_k would not shorten the time of duration of the process.

The increasingly shorter standstill periods with increment of x_k the accuracy of their realization conditioning the achievement of the final state is the reason that for greater x_k the process of time-optimum control becomes un-realizable. Moreover, considerable difficulties would be encountered if it proved necessary to introduce corrections taking into account the non-fulfillment of the assumptions made.

5. Time sub-optimum control

On practical grounds it may be profitable not to aim at obtaining optimum control, but to be satisfied with a control somewhat worse than optimum as regards the operation speed, but easier for realization.

Moreover, it is also worthwhile to consider the possibility of realizing a control which would limit the swinging of the load during travel, in particular when covering greater distances.

A control which would ensure non-oscillating travel over longer distances may easily be read from the phase plane $\frac{x''}{\omega}$, x', Fig.7.

To this purpose the load rate should be made equal with the trolley rate equalling "u". The point on the phase trajectory should be carried from the origin of the system to the point "u" on the axis "x". This is carried out most quickly ma-king use of a circular trajectory crossing through the system origin with a centre at point "u", and then from the trajectory circumscribed around the origin of the system shown in the figure and going through the point "u".

From the relationships seen in the figure, i.e. from the equilaterality of the triangle it results that angle:

$$\alpha_1 = \alpha_2 = \frac{1}{6} 2\pi \qquad 5.1$$

which corresponds to:

$$T_1 = T_2 - T_1 = \frac{1}{6} T \qquad 5.2$$

Therefore the control equivalizing the rate of the load shifting with the trolley travel rate will consist in switching the drive for a period $T_1 = \frac{1}{6} T$, then stopping the trolley for the period $T_2 - T_1 = T_1 = \frac{1}{6} T$ after which travel will be smooth during any desired length of a time interval $T_3 - T_2$. In order to stop the load an analogous operation must be carried out, i.e. the stopping of the trolley for the period $T_4 - T_5 = \frac{1}{6} T$, then again shifting it with a maximum travel rate "u" during the time $T_k - T_4 = \frac{1}{6} T$.

The load will remain motionless and the track covered will be:

$$x_k = \left[(T_k - T_4) + (T_3 - T_2) + T_1\right] u \qquad 5.3$$

or

$$x_k = (T \quad \frac{1}{7} T)\, u \qquad 5.4$$

This control may be used only for:

$$x_k = \frac{1}{3}\ T\ .\ u \qquad 5.5$$

Moreover, it may be estimated that the duration of the process in the most disadvantageous case differs from the optimum T_o by the standstill time.

$$T_k - T_o = \frac{1}{3}\,T \qquad 5.6$$

and in the case of travel over long distances and small period of oscillations the time loss will constitute a negligeable percent. In Fig.8 the time of the duration of the process is illustrated in proportional magnitudes at a time sub-optimum control and an optimum control with symmetric and asymmetric limitation of the control function for small x_k.

It is obvious that where four of the five time intervals a-re equal and are functions of only one magnitude T, and the fifth time interval is also presented by a rather uncomplicated formula

$$T_3 - T_2 = \frac{x_k}{u} + \frac{1}{3}\,T \qquad 5.7$$

the realization of a control is comparatively easy.

In the Laboratory of the Department of Automation and Industrial Electronics of the Academy of Mining and Metallurgy in Cracow a system/shown in Fig.2/ controlled by the above described programme has been realized. A load of 4 kg weight suspended on a 1.75 m coupling bar is shifted for a distance of up to 3.5 m with a maximum trolley travel rate 0.25 m/sec. After having reached the destination point the oscillation amplitude does not exceed 0.5 cm. When controlling the travel by hand it is very difficult to attain the damping of oscillations simultaneously with the determination of the process at a determined point. The loss of time as compared with the control according to an optimum time control depends on the operator's skill, but on the average it considerably exceeds the double time of the programmed control.

It should be stressed that this programme may remain uncha-nged in spite of the necessity of omitting the vertical obstacle if only the vertical motions are located in the period of the smooth travel, i.e. during the time interval from T_2 to T_3 and one returns to the previous level.

If the point aimed at is on another level it is enough to compute the two terminal time intervals as being equal to one third of the new period of proper oscillation.

If for the transporting of the load two drives, A and B, Fig.1, have to be used then the programme will continue to retain its simple form as opposed to the time optimum programme where the control with for example drive A must depend on the character of motions B and C.

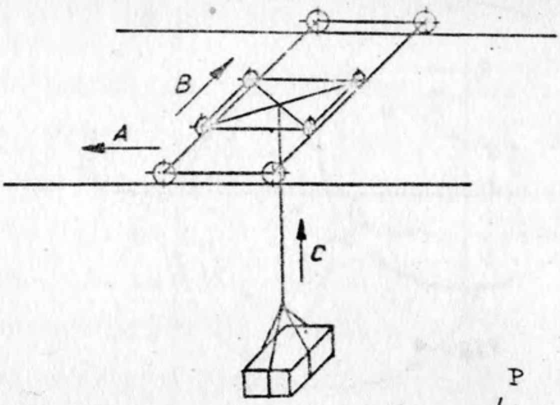


Fig. 1

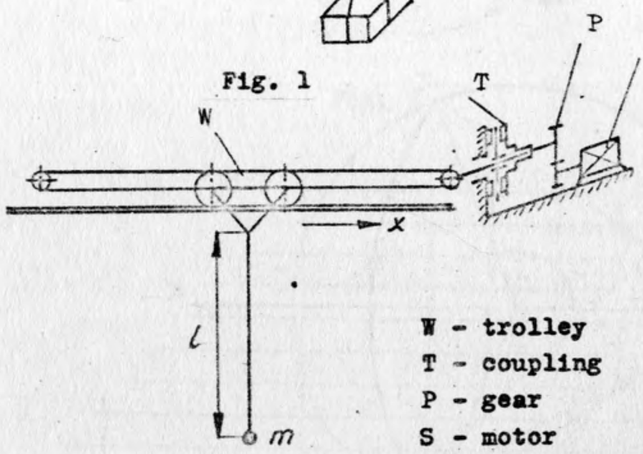


Fig. 2

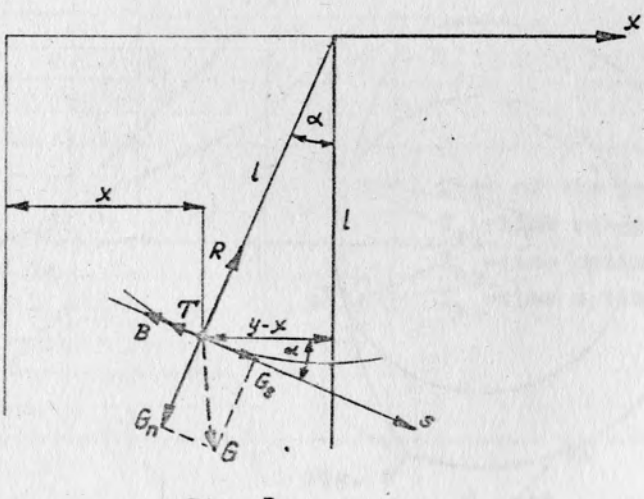


Fig. 3

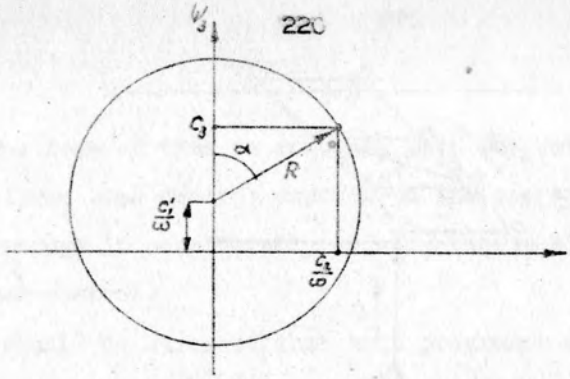


Fig. 4

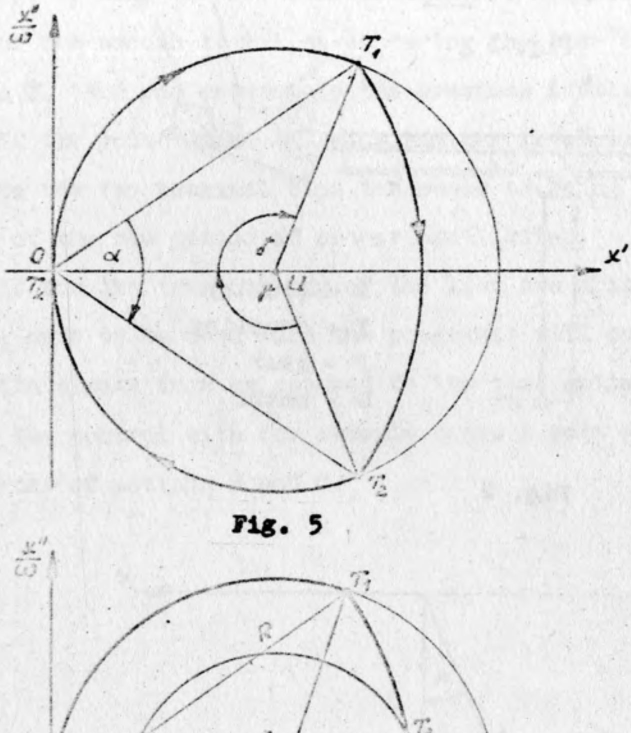


Fig. 5

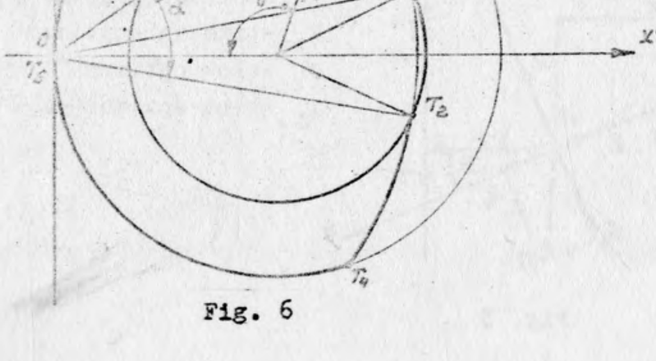


Fig. 6

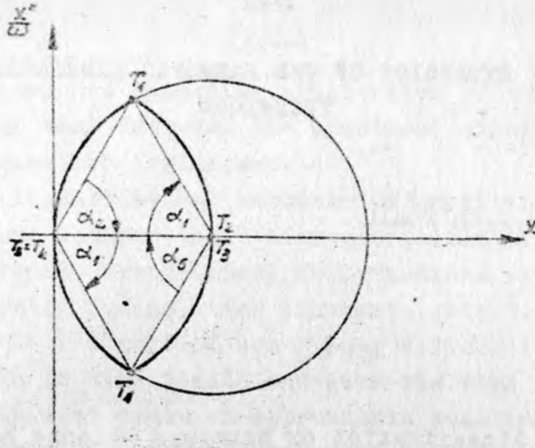

$\frac{x''}{\omega}$
T_1
T_2
T_3
T_4
x'
α_1
α_2
α_5

Fig. 7

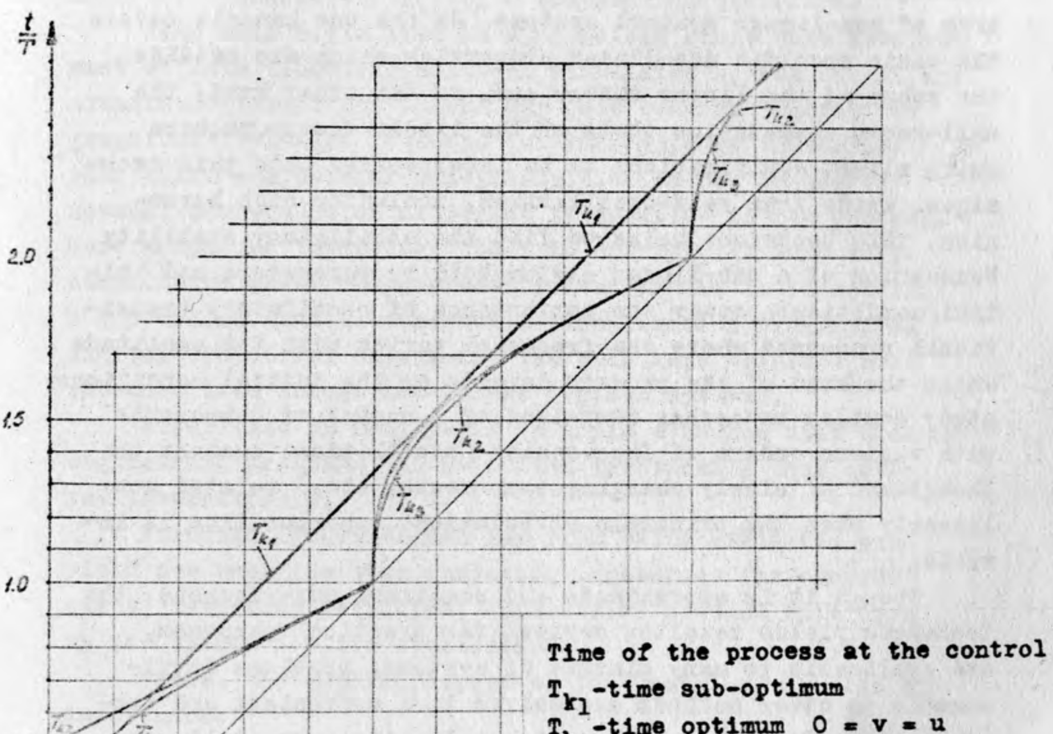

$\frac{t}{T}$
2.0
1.5
1.0
0.5
0
1.0
2.0
$\frac{x_k}{x_T}$
T_{k_1}
T_{k_2}
T_{k_3}
Time of the process at the control
T_{k_1}-time sub-optimum
T_{k_2}-time optimum 0 = v = u
T_{k_3}-time optimum -u = v = u

Fig. 8

41.3

AN EXTENSION OF THE HARMONIC LINEARIZATION TECHNIQUE

Ye.P.Popov, Ye.I.Khlypalo

Moscow
USSR

Harmonic linearization or harmonic balance has become a technique used widely in study and design of a wide spectrum of non-linear control systems. On the one hand it covers the basic specific non-linear properties which are outside the scope of the linear theory and, on the other hand, the well-known computation tools of the linear theory require quite slight modifications to be incorporated into this technique. Aside from self-oscillations, including high harmonics, this technique helps to find the oscillatory stability boundaries of a non-linear system both by parameters and initial conditions; study the performance of oscillatory transitional processes where the frequency varies with the amplitude while the kind of the process depends on the initial conditions; study complex processes comprised of a number of components with various orders of frequencies (oscillations against the background of slowly changing components, etc.) related non-linearly when the principle of solutions superposition is invalid [1],[2].

Though it is approximate and sometimes non-rigorous, the technique yields results, correct for practical purposes, and applicable to many classes of systems; problems invulnerable to other methods are solved in a convenient and comprehensible form. Hence its wide application especially because new control problems and increasing requirement to the performance and accuracy of system have attracted engineers to non-linear problems, to practical utilization of specially introduced non-linearities and non-linear control laws.

The advances and expanding application of computers increase rather than decrease the practical significance of such approximation techniques.

Harmonic linearization (harmonic balance) yields readily the relations between basic characteristics of processes such as amplitudes, frequencies, decay indices and parameters of the system (gains, time constants, etc.). No accurate solution to the initial non-linear differential and other equations in time domain can give the same results. Furthermore, in most cases an approximate solution by the harmonic linearization technique cannot be replaced by more accurate - and more involved - computation procedures.

With this basic fact in mind we can state that development of this technique has been stimulated by the advent and advances of computing technology just because "manual" or graphical treatment of these problems appears cumbersome when there are several non-linearities in the system or several components of different frequencies in the process. Here computing technology comes to our rescue. Many very complicated problems have been solved so far in industrial R & D bodies that used harmonic linearization programs on digital computers, e.g. M-20; the solutions obtained were embedded into actual non-linear control systems.

Experience of many years is ample evidence that much more engineering problems can be solved practically than covered theoretically.

Harmonic linearization far from being exhausted can yield new benefits when computing procedures for complex cases have been developed; theoretical studies will open up now fields for its application.

Hence, the applied aspect of the technique is also of importance.

This paper will describe a new representation of the harmonically linearized equivalent for non-univalent nonlinearities with hysteresis loops which practicaly workers have often to deal with. This new form corrects the existing non-rigorous notations in certain problems stated when formally this approach is utilized.

A harmonically linearized expression for non-univalent non-linearities

$$y = F(x) \tag{1}$$

with hysteresis loops is normally given by

$$y = \left[a(A) + \frac{b(A)}{\omega} p\right] x \,, \qquad b(A) < 0, \tag{2}$$

where A, ω are the oscillations amplitude and frequency, $a(A)$ and $b(A)$ are coefficients

$$a = \frac{1}{\pi A}\int_0^{2\pi} F(A \sin\psi) \sin\psi \, d\psi \,, \qquad b = \frac{1}{\pi A}\int_0^{2\pi} F(A \sin\psi) \cos\psi \, d\psi .$$

As a result the transfer function of an open-loop system has a pole in the right-hand hemi-plane (which changes its position when the amplitude of the oscillations changes). In terms of the linear control theory an equivalent linearized non-minimal phase element appears here.

Sometimes when the loop is closed no drastic changes occur and all operations proceed normally. However, there is a class of systems where this makes certain coefficients of the characteristic equation negative which would seem to prove that the oscillations in question are unstable, but actually these remain stable and the solution for the oscillations amplitude and frequency is valid.

An example of such a system is shown in Fig. 1. The dynamics of the system is described by

$$\varepsilon = g_1 - g_2 \,, \qquad u_3 = \frac{k_3}{1+T_2 p}(u_2 - u_{f8}) \,, \qquad \Omega = k_4 u_4 \,,$$

$$u_1 = k_1 \varepsilon \,, \qquad u_{f8} = k_{f8} p \Omega \,, \qquad g_2 = \frac{k_5}{p}\Omega \,,$$

$$u_2 = \frac{k_2 u_1}{1+T_1 p} \,, \qquad u_4 = F(u_3) = a(A) u_3 + \frac{b(A)}{\omega} p u_3 \,,$$

$$k = k_1 k_2 k_3 k_4 k_5 \,, \qquad k_6 = k_3 k_4 k_{f8} .$$

The transfer function of that port of the system which has a feedback loop is

$$W_1(p) = \frac{\Omega}{u_2} = \frac{k_3 k_4 \left[a(A) + \frac{b(A)}{\omega} p\right]}{1 + T_2 p + k_6 \left[a(A) + \frac{b(A)}{\omega} p\right] p} .$$

The transfer function of an open-loop system is

$$W(p)=\frac{k\left[a(A)+\frac{b(A)}{\omega}p\right]}{p(1+Tp)\left\{1+T_2p+k_6p\left[a(A)+\frac{b(A)}{\omega}p\right]\right\}}.$$

The characteristic equation of the system under consideration is

$$W(p)+1=0$$

or

$$A_0p^4+A_1p^3+A_2p^2+A_3p+A_4=0.$$

The factors of the characteristic equation will be

$$A_0=T_1k_6\frac{b(A)}{\omega},$$
$$A_1=T_1T_2+T_1k_6a(A)+k_6\frac{b(A)}{\omega},$$
$$A_2=T_1+T_2+k_6a(A),$$
$$A_3=1+k\frac{b(A)}{\omega},\qquad A_4=ka(A).$$

By eq. (2) the coefficient A_0 is negative. The coefficient A_3 can also prove negative whereas oll other coefficients are necessarily positive.

This is the result of erroneous form for equivalent harmonically linearized expression (2) which is normally employed in harmonic balance. We can preserve the same validity the solution for the amplitude and frequency and avoid the above non-rigorous intermediate expression (characteristic equation) if non-linear component (1) with an ambiguous loop non-linearity of a hysteresis type is assumed to be in the form of an inertial component equivalent to harmonically linearized transfer function given by

$$y=\frac{k_*A}{1+T_*(A)p}x. \qquad (4)$$

New harmonic linearization coefficients $k_*(A)$ and $T_*(A)$, that is the gain and the time constant of an inertial component equivalent to a non-linear component with hysteresis loops can be expressed identically through the former harmonic linearization coefficients $a(A)$ and $b(A)$.

In the case of periodic oscillations when $p = j\omega$ from the desired identity

$$\frac{k_*}{1+T_* j\omega} = a + \frac{b}{\omega} j\omega \qquad (b<0)$$

follows

$$\frac{k_*}{1+T_*^2\omega^2} = a(A), \quad \frac{k_* T_* \omega}{1+T_*^2\omega^2} = -b(A),$$

hence

$$T_* = \frac{-b(A)}{\omega a(A)}, \quad k_* = \frac{a^2(A)+b^2(A)}{a(A)}, \quad (5)$$

where $b(A)<0$; $a(A)$ and $b(A)$ are found by eq.(3).

In the case of decaying and diverging oscillations the equivalent harmonic linearized expression for non-univalent loop non-linearity is given, instead of by eq. (2), by [1]

$$y = \left[a(A) - \frac{\xi}{\omega} b(A) + \frac{b(A)}{\omega} p\right] x, \quad (6)$$

where $\xi < 0$ for decaying oscillations and $\xi > 0$ for diverging oscillations.

In this case when $p = \xi + j\omega$ from the desired identity

$$\frac{k_*}{1+T_*(\xi+j\omega)} = a - \frac{\xi}{\omega} b + \frac{b}{\omega}(\xi + j\omega)$$

follows

$$\frac{k_*(1+T_*\xi)}{(1+T_*\xi)^2 + T_*^2\omega^2} = a(A), \quad \frac{k_* T_* \omega}{(1+T_*\xi)^2 + T_*^2\omega^2} = -b(A),$$

Hence

$$\omega T_* = \frac{-b(A)}{a(A) + \frac{\xi}{\omega} b(A)}, \quad k_* = \frac{a^2(A) + b^2(A)}{a(A) + \frac{\xi}{\omega} b(A)}, \quad (7)$$

where $b(A) < 0$; $a(A)$ and $b(A)$ are also found eq. (3).

In the case of complex processes when the oscillations are superimposed on a slowly changing component, an approximate solution instead of $x = A \sin \omega t$ is found in the form $x=x^0+x^*$, $x^*=A\sin\omega t$, and the usual form of harmonic linearization instead of (2) is given by[2]

$$y=F^0(A,x^0)+\left[a(A,x^0)+\frac{b(A,x^0)}{\omega}p\right]x^*, \tag{8}$$

where

$$\left.\begin{aligned} F^0&=\frac{1}{2\pi}\int_0^{2\pi}F(x^0+A\sin\psi)\,d\psi,\\ a&=\frac{1}{\pi A}\int_0^{2\pi}F(x^0+A\sin\psi)\sin\psi\,d\psi,\\ b&=\frac{1}{\pi A}\int_0^{2\pi}F(x^0+A\sin\psi)\cos\psi\,d\psi. \end{aligned}\right\} \tag{9}$$

Then the non-linear system equation will be

$$Q(p)x+R(p)F(x)=N(p)f(t), \tag{10}$$

where $f(t)$, a external action changing slower than the ω oscillations frequency decomposes, after harmonic linearization, into two non-linearly related equations

$$Q(p)x^0+R(p)F^0(A,x^0)=N(p)f(t), \tag{11}$$

$$Q(p)x^*+R(p)\left[a(A,x^0)+\frac{b(A,x^0)}{\omega}p\right]x^*=0. \tag{12}$$

In this case the new form of an equivalent transfer function can be applied in eq. (12), that is the expression in brackets can be replaced by

$$\frac{k_*(A,x^0)}{1+T_*(A,x^0)p},$$

where

$$T_*=\frac{-b(A,x^0)}{\omega\, a(A,x^0)},\qquad k_*=\frac{a^2(A,x^0)+b^2(A,x^0)}{a(A,x^0)},$$

while a and b are found by eq. (9), $b<0$.

We can treat similarly <u>the forced oscillations</u> where in the right-hand part of the non-linear system of eq.(10) we have

$$f(t)=B\sin\omega t \quad \text{or} \quad f(t)=f^0(t)+B\sin\omega t.$$

We will describe the computations with the new form of harmonic realization for the above example of a non-linear system (fig. 1).

For the kind of non-univalent non-linearity under consideration with hysteresis loops the expressions for coefficients of harmonic linarization are given by

$$a=\frac{2h}{\pi A}\left(\sqrt{1-\frac{c^2}{A^2}}+\sqrt{1-\frac{m^2c^2}{A^2}}\right),$$

$$b=-\frac{2ch}{\pi A^2}(1-m) \quad , \quad \text{at } A \geq c ,$$

where A is the oscillation amplitude of the voltage u_3.

E.g. for numerical values of a non-linear components parameters

h = 110 v, c = 24 v, m = 0.2,

the graphs of harmonic linearization coefficients are given in Fig. 2.

These graphs with eq. (5) can easily give relations of equivalent parameters k_* and ωT_* against oscillations amplitudes represented in Fig. 3.

It whould be remembered that the formulae of harmonic linearization and therefore the graphs of Fig. 3 are meaningful only at $A > c$. Therefore the magnitude of T_* is bounded. Fig. 3 shows that with increase of the amplitude A the time constant T_* decreases; with the given non-linearity this is equivalent to a decreased effect of the hysteresis loops at large oscillation amplitudes.

With eq. (4) for the equivalent transfer function the dynamics of the system (Fig. 1) is given by the equations

$$\varepsilon = g_1 - g_2 , \quad u_3=\frac{k_3}{1+T_2 p}(u_2-u_{f b}), \quad u_{f b}=k_{f b}p\Omega,$$

$$u_1=k_1\varepsilon, \quad u_4=\frac{k_*}{1+T_* p}u_3 , \quad g_2=\frac{k_5\Omega}{p} ,$$

$$u_2=\frac{k_2}{1+T_1 p}u_1 , \quad \Omega=k_4u_4 .$$

The transfer function of the part of the circuit with a beedback is given by

$$W_{c.l.} = \frac{k_3 k_4 k_*}{(1+T_2 p)(1+T_* p)+k_6 k_*}, \qquad k_6 = k_3 k_4 k_{\text{PB}}.$$

The transfer function of an open-loop system will be

$$W = \frac{k k_*}{p(1+T_1 p)[(1+T_2 p)(1+T_* p)+k_6 k_* p]}, \qquad k = k_1 k_2 k_3 k_4 k_5.$$

The characteristic equation of a non-linear harmonically linearized system will be given by

$$A_0 p^4 + \mathcal{A}_1 p^3 + \mathcal{A}_2 p^2 + \mathcal{A}_3 p + \mathcal{A}_4 = 0,$$

where

$$A_0 = T_1 T_2 T_*,$$
$$\mathcal{A}_1 = T_1 T_2 + T_1 T_* + T_2 T_* + T_1 k_6 k_*,$$
$$\mathcal{A}_2 = T_1 + T_2 + T_* + k_6 k_*,$$
$$\mathcal{A}_3 = 1, \quad \mathcal{A}_4 = k k_*.$$

We can see that all coefficients of a characteristic equation are, as dintict from the previous one, positive.

Assume that the objective of further computation of the system is to obtain a non-linear transition processes performance diagram[1] in order to obtain the transition process decay index ξ and the oscillation frequency ω for each value of the selected parameter of the system and the amplitude.

The gain of the system linear part $k = k_1 k_2 k_3 k_4 k_5$ will be taken as the parameter selected.

To solve our problem we will introduce into the characteristic equation of a harmonically linearized system the value $p = \xi + j\omega$ instead of p. As a result we will obtain an equation of the form

$$X(A, \omega, \xi, k) + jY(A, \omega, \xi) = 0.$$

Because the selected parameter k in included only into the coefficient $\mathcal{A}_4$, the parameter k will be only in the real part of X after the real and imaginary parts have been separated. Therefore, from the equation

$$Y(A,\omega,\xi)=0$$

we can find the relations $\xi(A)$ for different values of ω = const. and by substituting these into the expression

$$X(A,\omega,\xi,k)=0,$$

find the relations $k(A)$ for different ω = const. This will immediately yield the lines of ω = const. on the desired performance diagram of non-linear transition processes with the system of coordinates (k,A) and the relations $\xi(A)$ at different ω = const. found before will make it possible to trace the major part of the diagram [1] as lines of ξ = const. in the same system of coordinates.

Similarly to this case we can easily handle other problems solved in practice by harmonic linearization at various non-univalent non-linearities with hysteresis loops by using the new, more rigorous notation suggested here for equivalent transfer function. To facilitate practical calculations for all specific kinds of non-linearities, formulae and graphs can be prepared in advance for new harmonic linearization coefficients $k_*(A)$ and $\omega T_*(A)$ similar to those that exist now for the coefficients $\alpha(A)$ and $\beta(A)$. This can also be done in more complex cases which include an additional relation with the ratio $\frac{\xi}{\omega}$ or the constant constituent x^0.

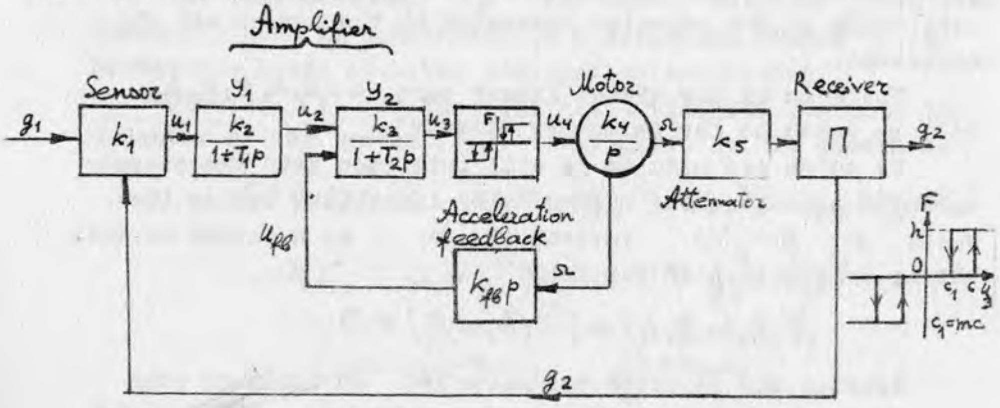


Fig. 1 System block-diagram

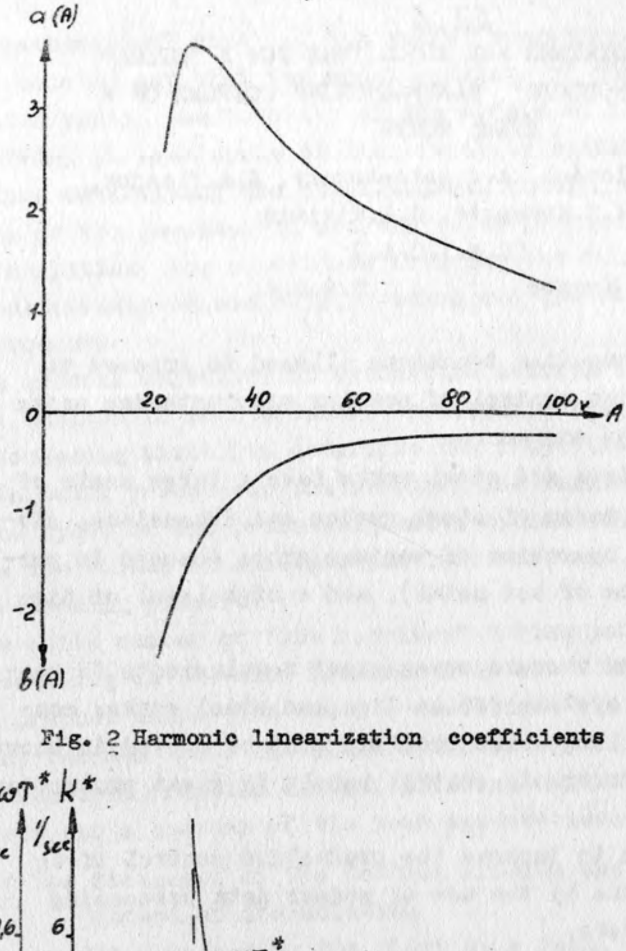


Fig. 2 Harmonic linearization coefficients .

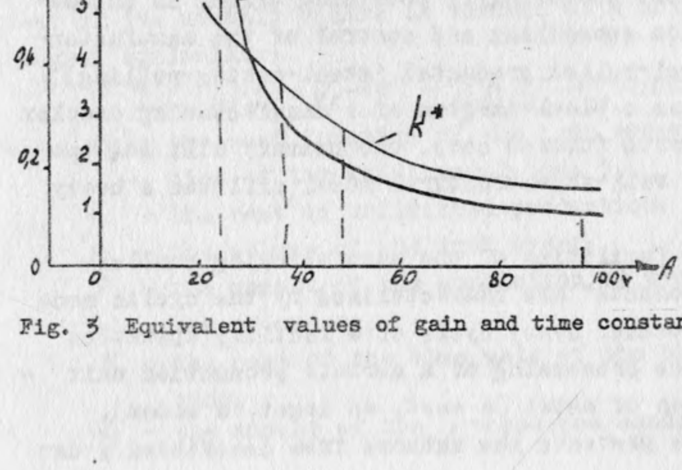


Fig. 3 Equivalent values of gain and time constant

42.4

CONTROL SYSTEMS AND ALGORITHMS FOR A "STEEL-ROLLED PRODUCTS" MANUFACTURING COMPLEX OF A STEEL WORKS

A.P.Kopelovich, A.A.Belostotsky, B.A.Vlasjuk,
V.M.Khrupkin, G.I.Nikitin
(C.R.I.C.A.)
Moscow U.S.S.R.

Progress in computing technique allowed to proceed to automatic production control of complex manufacturing units such as large shops and works.

Large modern iron and steel works have a large scale of production a wide range of steel grades and dimensions, close interaction of operation of various shops (caused in particular by the flow of hot metal), and a high level of disturbances.

These and other factors cause great requirements to control and planning systems for an iron and steel works: complex problems and large size problems must be solved in short periods of time, errors in control result in great production losses.

It is possible to improve the production control of an iron and steel works by the use of modern data processing means, i.e. computers.

One of the most economically promising trends is automation of production scheduling and control of the manufacturing complex "steel-rolled products" (steel-making-rolling)[1].

Fig.1 contains a block diagram of a manufacturing complex with one open hearth furnace shop, one primary mill and two jobbing mills (a rail-and-structural steel mill and a heavy section mill).

All the main facilities of the manufacturing complex "steel-rolled products" are characterized by the cyclic mode of production process. Every cycle of a facility operation corresponds to the processing of a certain production unit - a discrete portion of metal (a cast, an ingot, a bloom).

Lack of space prevents the authors from describing a detailed mathematical model of the works' operation. It will be only pointed out that the model consists of three main interrelated parts: the totality of the models of the processes occurring at each cycle of each facility operation; the conditions determining the metallurgical routing and the constraints on the sequence of product units processing on the works facilities; the conditions ensuring the fulfillment of the predetermined volume of production and the variety of steel products.

The general objective of production control is to choose control actions in each process, to schedule the operation of each processing plant (to determine the properties of the products being processed, the starting and termination points for each cycle of the processing units operation) and to determine the volume of discrete portions for each cycle of each processing plant.

The whole number of "the machines" taking part in the metal processing in the area considered is about 200; the number of product units (the casts processed in the area during a month) reaches 2000. The plant operates according to the orders received from the central planning office; the number of orders per a quarter of the year amounts several thousands.

The Statement of the Control Problem and the Method of its Solution

The profit obtained by the plant as a result of the fulfillment of its monthly orders is assumed as a criterion of the control optimality:

$$\sum_i C_i(t_i)\cdot\gamma_i + y_T - KT - \sum_j b_j x_j \longrightarrow max \qquad (1)$$

where $C_i(t_i)$ - the cost function of the i-th order as a function of its fulfillment time t_i;

y_T - the cost of unfinished production;

γ_i - the volume of the i-th order;

T - the moment of the completion of the total number of orders;

K - the cost of the time unit of the plant operation;

x_j - the amount of the j-resource used;

b_j - the j-resource cost.

The first term of the criterion (1) is the cost of the whole set of orders taking into account its decrease when the orders are not fulfilled by the proper time. The second criterion term contributes to the production process in the preceding areas of the metallurgical routing with the predetermined volume of the finished product units for the plant. The third and the fourth terms require minimization of the idle time of the production facilities and of the resource used.

The analysis of the control problem for the plant shows that it does not belong to any known class of the extremum problems; it involves the theory of combinations, elements of variational problems and the problems of mathematical programming. To solve accurately a problem of such complexity and of such an enormous size and to correct it according to its disturbances would require nonexisting computing devices.

The main ways of overcoming the "damnation of size" are: making the most use of the specific nature of the particular problem, i.e. the simplifying features following from the controlled plant structure and the optimality criterion; and the employment of approximate solutions assuming as a basis the previous management experience (ranking of priorities, reasonable restriction of a great number of variants etc.)

The simplifying features of the problem arise due to the fact that the ties between different areas of great manufacturing complexes give certain freedom of choice in performing separate operations. Besides there is some freedom of control within a shift, twenty four hours and so on, i.e. within separate time intervals of the total period of scheduling.

The specific features of the production control system are represented mathematically by a specific structure of the system of equations and inequalities describing the controlled plant, e.g. in a block character of this system. This allows to apply special computational algorithms which are more efficient than general methods of solution: decomposition methods of solutions of the linear programming block problems[2,3], local algorithms of integer problems solutions[4], dynamic programming of Marcovian processes[5], sequential analysis scheme of variants[6], etc.

To make use of such computational algorithms it is necessary to represent the controlled plant model structure and to select an efficient structure of the problem solution for the given model.

It is reasonable to define the structures of the model and of the solution in several steps[7]. The first step is to decompose the controlled plant into comparatively few parts or blocks. The totality of the blocks and their interdependences (the conditions combining the variables of separate blocks) represents the controlled plant structure at the first step.

Analysis of this structure allows to obtain the efficient structure of the problem solution, i.e. a set of the control subproblems and their interdependence in the process of solution.

At the second and subsequent stages, the control subproblems obtained at the preceding stages are examined; the structure of the controlled plant of each subproblem is detailed and the solution structure selected.

The works operation during a month is represented as a complex multistep process. The general model of the works operation is decomposed into blocks in accordance with the separate manufacturing areas and the time subintervals of the total planning period. A separate block of overall structure describes the work of one manufacturing area within one time subinterval.

It is reasonable to select the size of one manufacturing area covered by one block and the duration of the planning subintervals so that, firstly, the total number of the blocks were not too great allowing to derive the most efficient succession of the subproblems solution, and, secondly, the number of the parameters connecting the separate blocks were as few as possible.

The general structure of the controlled plant analysed at the first decomposition step is given in Fig.2, where a block is designated with two indices: K - the number of the time subinterval, and P - the number of the manufacturing area. Here $P = 1$ - the open-hearth furnaces area; $P = 2$ - the blooming mill area; $P = 3$ - the first N 2 jobbing mill area; $P = 4$ - the second jobbing mill area.

Each block has ties of two types: the "static" ties with the other block (area) within one planning time subinterval, and the "dynamic" ties - with the blocks describing the same manufacturing area but in other subintervals.

Decomposition of the control problem comprises the formulation of the local control subproblems of the separate blocks and determination of the method of coordination of the separate subproblems solutions with the aim to optimize solution of the general control problem.

The local control subproblems of the blocks can be derived using the optimality principle of dynamic programming: "any process occurring between two fixed finite points must be controlled optimally", i.e., at the fixed input and output coordinates of the given block, only those control variants might be optimal which provide the maximum block efficiency.

The local control subproblems are the problems of the production control; they involve the determination of the mode and succession of operations on the plants of the given area for the present initial states of all the plants, predetermined delivery schedules of the product units in the area, and at the predetermined time of processing completion of the plants.

The coordinating control subproblems are the problems of production scheduling; they involve determination of the above variables of the local control subproblems. A great size of the general coordination problem requires its decomposition into separate subproblems in each of which only a part of the ties among the blocks is solved.

Each succession of the block ties resolution may be defined by respective expenditures for the problem solution using the computer (required memory capacity, duration of solution etc.). For the problems in question, the solution expenditures depend on the number of parameters. It is reasonable to select such succession of the blocks ties resolution that the total number of parameters in all parametric coordinating subproblems would be minimal.

The structure of the solution of the steel works control problem based upon the above methods is given in fig.3. Here blocks 1-9 are the production scheduling system; blocks 10-13 - the open-hearth furnace shop control system; blocks 14-16 - the control system of conveying metal to the soaking pits; blocks 17-20 - the primary mill control system; and 21-24 are the jobbing mills control system.

The main subsystems of the "steel-rolled products" manufacturing complex for different hierarchial levels being designed at The Central Research Institute of Complex Automation are given below.

The Production Planning System

One of the most promising (so far as economy is concerned) trends of the modern high speed computers use is automation of the production planning of an iron and steel works operations.

The general iron and steel production planning problem is the following: to make optimal schedules of the main processing units operations based on the totality of orders from consumers, the current state of the processing units and unfinished processing[8].

The optimal schedule of processing units operation is such a schedule which is firstly made in accordance with the operation schedules of other processing units and areas and secondly it minimizes the plant losses through processing units idle time, undue delivery dates, hot metal cooling, etc.

Various disturbances require rather frequent corrections of the schedules formed earlier. That is why the planning automation was considered as developing the planning system, including a system of production data collection and transmission, the systems of algorithms with which the optimal schedules calculations and corrections are made and finally the systems of schedules transmission to the personnel at the production areas.

The plant operates according to monthly orders, which the central planning department sends to the plant once a quarter: the number of orders per a quarter totals several thousands. The orders are mainly for structural shape but the plant can also deliver unfinished products - steel ingots, blooms.

The quality of all the schedules was estimated according to the optimality criterion for the whole plant. The profit obtained by the plant as a result of fulfillment of the monthly book of orders was assumed as an optimality criterion.

The general problem of finding optimal schedules of all the production units operation on the basis of the monthly book of orders has a great number of variables which are connected by complex and numerous relations. That is why the centralized solution of such a problem is impossible even using modern computers and decomposition of the general problem into a set of interrelated subproblems of smaller size is necessary.

In performing this decomposition the choice of planning algorithms system is made on the basis of the following principles: the algorithms structure should be hierarchical; the optimal algorithms structure should minimize the total expenditures on the problem solution; the input data which a higher level of planning works out for a lower level should be either certain production units schedules or the prices of unfinished products.

The structure obtained envisages the use of the following basic algorithms:

1) Making up a set of lots, i.e. the totality of orders rolled on jobbing mills having one set of rolls. The problem is formulated as a combinatorial one solved by means of a heuristic algorithm based on the priority rule.

2) The determination of lots sequence rolled on jobbing mills. An algorithm combining dynamic programming and direct-

ed selection is developed to solve it.

3) Scheduling of the manufacturing complex "steel-rolled products" operations. Daily orders of open-hearth furnace steel production of rolling on the blooming mill and jobbing mills, of metal dispatching are determined. A block method of linear programming is used for solving this problem.

4) Daily scheduling of an open-hearth furnace shop: the distribution of steel grades according to furnaces and cast numbers on each furnace within a day (twenty four hours). The problem is solved as an integer linear programming problem.

5) Daily scheduling of cast processing on a blooming mill. The algorithm is formed as a result of formalization of the rules employed by the plant personnel.

6) Daily scheduling for jobbing mills. The algorithm is based on heuristic rules.

The efficiency estimation of the foregoing algorithms of scheduling showed that the system allowed to increase the productivity of jobbing mills, of the open-hearth furnace shop and the blooming mill by 2.5-3%.

To solve the problems the planning system uses computers of the iron and steel plant computing center. The production progress data are fed by means of teleprinters, placed in the open-hearth furnace shop, on the blooming mill and jobbing mills. The data are printed on the printers in the computing center and also directly for the controllers of the plant.

The Open-Hearth Furnace Shop
Control System

The automatic control system is designed for the shop having a stockyard, a mixer, several open-hearth furnaces, a furnace bay and a teeming bay[1]. The shop control problem is stated as a problem of the shop relative profit maximization:

$$\Pi = \sum_j C_j g_j - \sum_i \kappa_i \tau_i - a \sum_i g_i \rightarrow max \quad (2)$$

where g_j - the production volume of the j-th steel grade;

τ_i - the time spent by the i-th furnace for production;

g_i - the consumption of the technological oxigen by the i-th furnace;

C_j - the relative cost of the j-th steel grade;

K_i - the time unit relative cost of the i-th furnace operation;

a - the relative cost of technological oxigen.

The adopted open-hearth furnace shop criterion form is flexible enough and a proper choice of factors may reduce it to various particular forms.

The linear nature of the criterion is very convenient to solve certain optimization problems.

The control system uses the following algorithms:

1) Current scheduling of open-hearth furnaces: for each cast on the scheduling interval a certain oxigen consumption is calculated; the totality of the scheduled casts in all the furnaces forms a sequence of demands for the maintenance of the furnaces with auxiliary equipment. The schedule of meeting these demands is made for each kind of equipment; in the equipment assignment checking up the fulfillment of the schedule for the furnaces maintenance with the earlier assigned equipment is performed through a designated for each cast numerical parameter - "activity slack" (the difference between the moment of service demand arrival and the moment of releasing the equipment designated for servicing this demand).

After the assignment of all kinds of equipment reassignment of oxigen is made in such a way that the total oxigen consumption per shop at any moment would not exceed the predetermined value. The oxigen consumption of the cast period with negative activity slack increases and it decreases for periods with positive activity slack. After the oxigen reassignment the auxiliary equipment assignment is made again. In accordance with the furnaces operation ~~schedule~~ the schedule, of charge conveying to furnaces and trains arriving for teeming is calculated, the schedule of other technological operations in furnaces is calculated as well.

2) The optimal strategy determination for lack of coincidence of steel teeming while chargings in adjacent furnaces coincide; the strategy consists in determining the sequence of chargings and cast weights for every possible internal between

the chargings starting points on adjacent furnaces. The problem is formulated as a problem of optinal control search for the controlled Marcovian chain and it is solved through the modification of Horward's method.

3) The auxiliary equipment control; the problem is in assignment of the available equipment (a charger, eqipment of the teeming bay, etc.) to fulfil certain operations; the problem is also in determining the next charging of cast iron to furnaces. The equipment distribution is accomplished according to priorities of furnaces based on "activity slack".

4) The production progress control; it is performed by means of comparing the schedules of operations based on the algorithms with their actual fulfillment and the production recording.

The automatic system is developed for the centralized control shop. All the production progress information is concentrated at the shop control point. The production progress information is fed automatically and manually. The traffic control (trains and locomotives) is performed by means of relay centralization of the route and train information devices providing information about the progress and N of the trains and locomotives.

The heart of the system is a computer. The computer is to operate in real time providing multiprogram operation with interruptions for information input - output. The computer speed is about 30 000 short operations/sec; the memory capacity is about 20 000 of 24-bit words.

The benefit obtained from the implementation of the system is achieved as a result of increasing the shop productivity. Besides, the gain is expected from reduced fuel consumption and reduced ingots heat losses thanks to more uniform cast yield from the shop.

The Control System of Conveying Metal to the Soaking Pits

The control system of conveying metal to the soaking pits is intended for that area of the works which covers the teeming bay, stripper bay, mould preparation shop and cold ingots stock. The peculiarity of this area is the closest interaction of technological and transport operations[9].

The functions of the control system under consideration include:

1) The organization of conveying the trains with casts to soaking pits according to the requirements of the central planning system.

The basic criterion in organizing the traffic is the minimization of deviations from the order of casts arrival; the necessity of making up trains of moulds in the mould preparation shop in due time must be taken into account. A real traffic situation is proceeded from: the availability of empty transit rails, the possibility of the locomotive approach to the train or the necessity of shunting, the availability of vacant locomotives, etc.

The criteria of each particular control problem solution (the choice of locomotives, optimal routing, etc.) are determined by the specific features of rails in the controlled area; however, the system always tries to ensure the train movement for the shortest time.

2) Forming a sequence of trains directed to the mould preparation shop for making up. The criterion for this problem solution is the minimization of the deviation of the train arrival time for teeming from the scheduled starting point of teeming in the open-hearth furnace shop.

3) The optimal distribution of the locomotives' work which may be considered as a transport linear programming problem. The criterion for solving this problem is the minimization of the time when the trains wait for locomotives and idle running of locomotives.

The solution of these problems results in making up a schedule for the area for a certain period of time. Partial or full corrections of the schedule are made for every change of a plant state.

According to the schedule the traffic automatic devices at the controlled area receive at certain moments instructions about the automatic choice of the route (switching on proper railway points and lights, switching off the route which was completed); these traffic instructions are transmitted to locomotives.

The instructions about the order of making up and stripping are transmitted to locomotive drivers and operators in the stripper bay and mould preparation shop. The instructions about making up trains with cold ingots of certain kind are transmitted to the cold ingot stock.

The automatic traffic devices are the source of information about the traffic situation in the controlled area.

To get information about the change of situation at an area sensing devices are installed in certain points of the rails. They automatically transmit a passing train or locomotive number and also their direction into automatic traffic devices. It is possible to follow the trains and locomotive movement through the area with the aid of such devices. The numbers of the trains and locomotives are transmitted from one sensing device to the next one following the movement of the locomotives.

The data about the progress of certain technological operations (the starting and finishing points of making up trains, stripping, etc.) are fed by operators through keyboards.

The whole system is expected to increase the temperature of the metal arriving at the soaking pits and to reduce the blooming mill idle time.

REFERENCES

1. М.Д.Климовицкий,А.П.Копелович,Автоматический контроль и регулирование в черной металлургии.Изд-во "Металлургия",1967 г.

2. G.B.Dantzig, Ph.Wolfe, "Decomposition principle for linear programms", Oper.Res., 1960, vol.8, N 1.

3. Е.Г.Гольштейн. Д.Б.Юдин, Новые направления в линейном программировании. М., 1966 г.

4. Ю.М.Журавлев, Ю.Ю.Финкельштейн,Локальные алгоритмы для задач целочисленного программирования, "Проблемы кибернетики",Вып.14, 1965.

5. Р.Беллман, С,Дрейфус, Прикладные задачи динамического программирования, Изд-во "Наука", М., 1965.

6. В.Михалевич, Б.Шор,Метод последовательного анализа вариантов для численного решения задач оптимизации, "Труды конференции по вопросам применения ЭВМ в народном хозяйстве", г.Горький, 1964.

7. Г.И.Никитин, Выбор структуры автоматизированной системы управления металлургическим заводом, "Труды ЦНИИКА", вып.19, 1968.

8. Б.А.Власюк, А.П. Копелович, Г.Р.Кюсснер,Применение ЭВМ для оперативного планирования прокатного производства, "Приборы и системы управления", № 6, 1967.

9. А.А.Белостоцкий, Ю.С. Вальденберг, Система оперативно-диспетчерского управления "Импульс" для участка металлургического комбината, в сб. "Управление производством", Изд-во "Наука",1967.

FIGURES

Fig.1. The works structure:

1) Area 1; 2) Area 2; 3) Area 4; 4) Area 3; 5) Open-hearth furnace shop; 6) Dispatch; 7) Purchased ingots; 8) Purchased blooms; 9) Dispatch; 10) Stockyard; 11) Furnace bay; 12) Teeming bay; 13) Stripper bay; 14) Soaking pits; 15) Blooming mill; 16) Blooming yard; 17) Reheating furnaces; 18) Mill N 2; 19) Mill N 1; 20) Finishing; 21) Cold ingots stock.

Fig.2. The works model structure.

Fig.3. The steel works control structure:

1) Scheduling of lots sequence at the jobbing mills; 2) Making up lots at the jobbing mills; 3) Monthly production scheduling; 4) Scheduling for the open-hearth furnace shop; 5) Scheduling of casts processing; 6) Scheduling of ordered lots processing sequence; 7) Scheduling of lots processing sequence; 8) Scheduling of ingots processing sequence; 9) Scheduling for N1 jobbing mill; 10) Scheduling of open-hearth furnace operating time; 11) Control of the stockyard equipment operation; 12) Control of the furnace bay equipment operation; 13) Control of the teeming bay equipment operation; 14) Traffic control; 15) Cold ingots stock control; 15) Cold ingots stock control; 16) Stripper bay control; 17) Soaking pits control; 18) Ingot cranes control; 19) Roughing mill control; 20) Shears control; 21) Blooms stock control of N 1 mill; 22) Reheat furnaces control; 23) Compartment furnaces control; 24) Jobbing mill control.

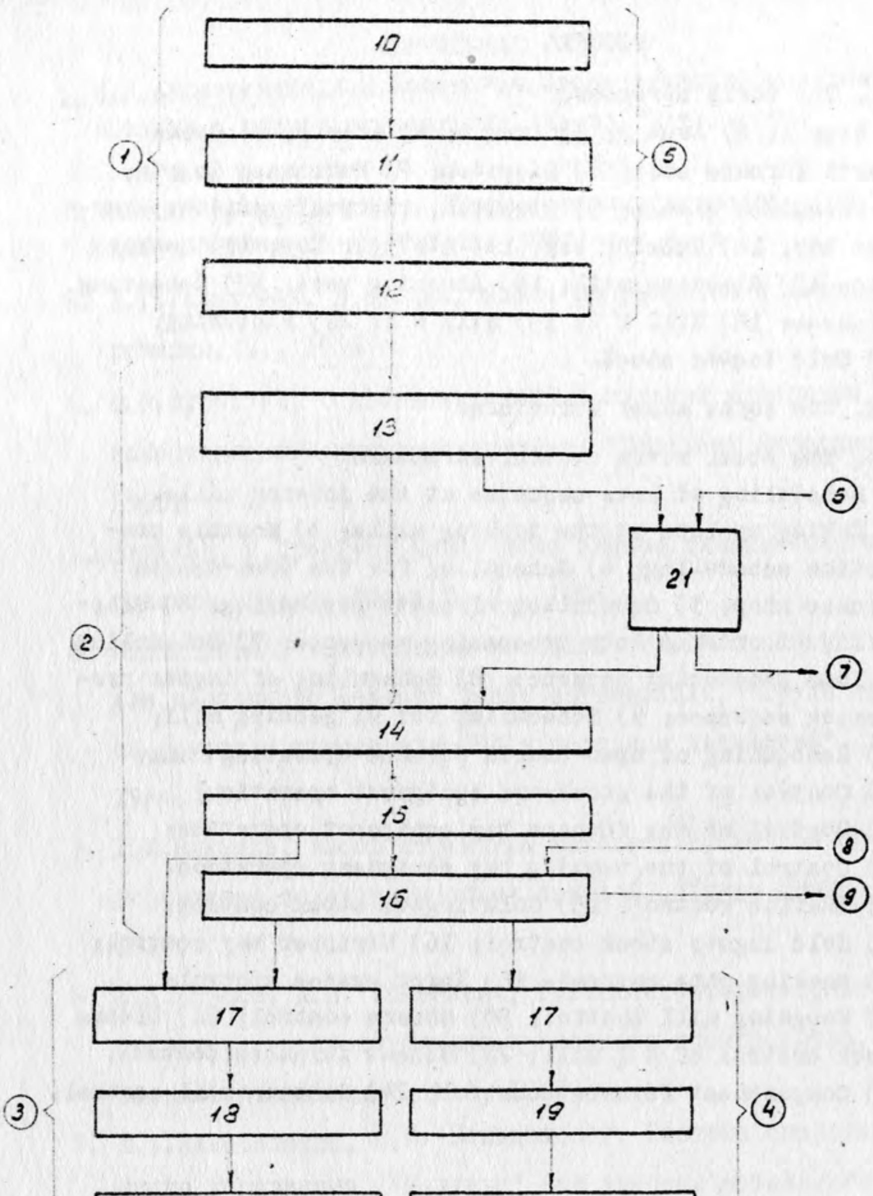


Fig. 1

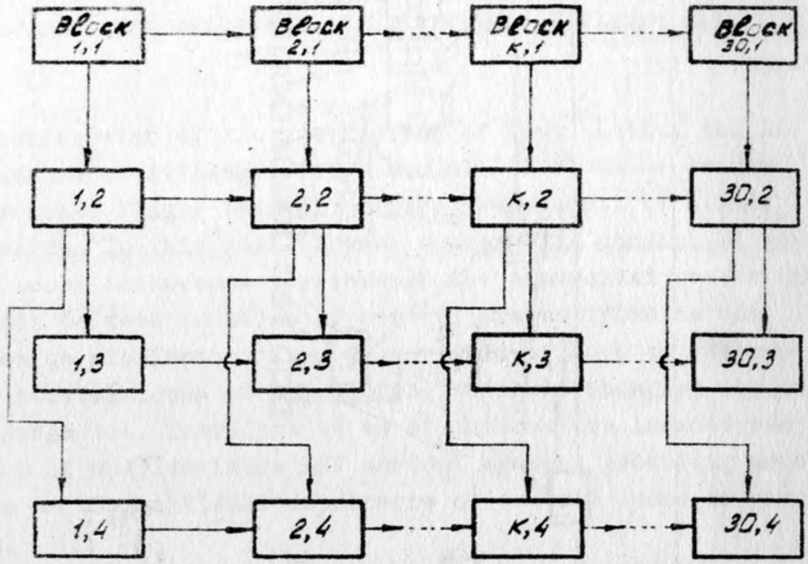

Block 1,1
Block 2,1
Block K,1
Block 30,1
1,2
2,2
K,2
30,2
1,3
2,3
K,3
30,3
1,4
2,4
K,4
30,4

Fig. 2

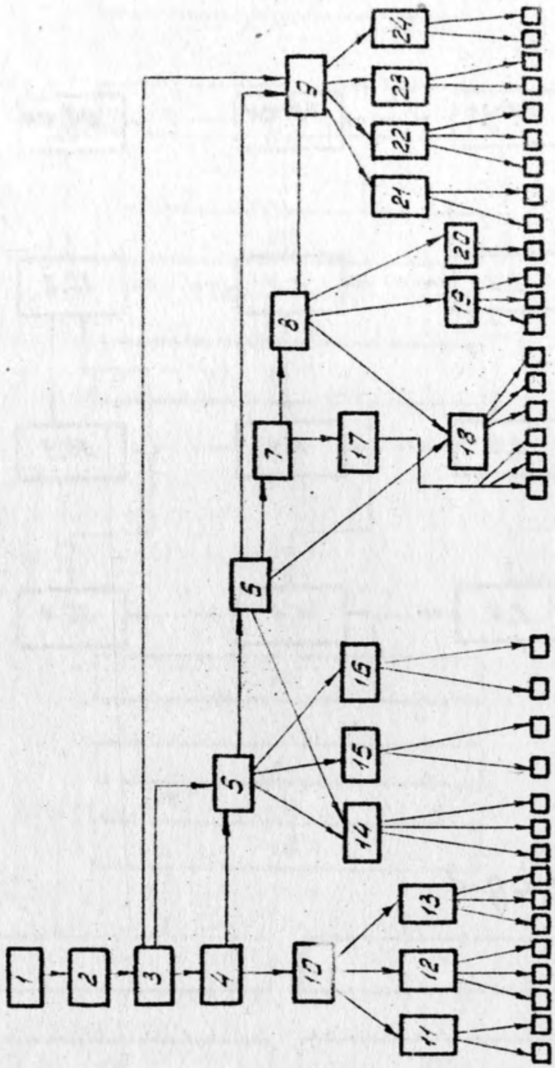


Fig. 3

43.1

CONSIDERING SYNTHESIS OF LIFTING REENTRY VEHICLE CONTROL SYSTEM STRUCTURES IN ATMOSPHERIC MANEUVER

PETROV B.N. KOLPAKOVA N.P. VASILYEV V.A. PAVLENKO A.I.
(MOSCOW)

Consideration of the interaction of longitudinal and lateral motions of lifting reentry vehicle (LRV) is necessary for hypersonic flight even at comparatively small angles of attack and sideslip. In this case control system will consist of many control loops interconnected through the controlled member LRV that leads to deterioration of control system dynamics and sometimes to its instability. As a result of such an interaction of control loops at one control variable changing the others would change too. Therefore it is of interest to investigate the set G of multivariable LRV control systems providing independence or insignificant dependence of control loops or groups of them.

In this paper the linear multivariable control system where the control loops interact through LRV have been considered and the problem of obtaining the set G {M, S, L} of structures providing selective invariance has been considered to select the best structure in the sense of control quality and simplicity of implementation.

I. EQUATIONS OF THREE-DIMENSIONAL LRV MOTION

LRV is a vehicle with comparatively high hypersonic L/D. There LRV can fulfil an effective aerodynamic maneuver [I] owing to lift force arising in its atmospheric flight. Such flight regimes at great bank angle (as high as 90°) are typical for LRV maneuver; then new problem for control system have arised.

In the LRV atmospheric flight with almost circular velocity it is impossible to neglect of Coriolis and centrifugal force as in the case of aircraft. Therefore to compute LRV trajectory we can use a set of LRV centre of mass motion equations given in velocity semi-state coordinates considering curvature and rotation of the Earth [2].

Adding the equations describing LRV rotating motion, cinematic equations of LRV cetre of mass motion converted at geographic coordinates, cinematic equations describing LRV rotating motion about Earth axes and a number of geometric correlations connecting angles at velocity semi-state coordinates with angles at vehicle state coordinates we get a set of I5 nonlinear equations.

Owing to the nonlinearity of these equalions exact parameters of LRV motion can be calculated only with methods of numerical integrating or digital computers. At the same time at the first stages of the control system synthesis we are mainly interested in analytical estimations which allow to investigate LRV dynamics.

For that purpose let's simplify the received set of equations by linearizing them about some nominal regime. But it should be noted that it is impossible to use the known techniques of separating of LRV longitudinal and lateral motions to investigate its dynamics as LRV fulfils aerodynamic maneuver with large angles of attack and usually large bank angles, which leads large angles of sideslip.

After lineariziting the set of equations is resolved as to the most important controlled variables.

The set of linearized equations of LRV three-dimensional motion is as follows:

$$\left\{\begin{aligned}
V &= a_{11}\alpha + a_{12}\beta + a_{13}\gamma + a_{14}\vartheta + a_{15}\psi + a_{17}H + a_{18}\delta_{yp} + a_{19}\delta_{B} + a_{110}\delta_{H},\\
\vartheta &= a_{21}\alpha + a_{22}\beta + a_{23}\gamma + a_{25}\psi + a_{26}V + a_{27}H + a_{28}\delta_{yp} + a_{29}\delta_{B} + a_{210}\delta_{H},\\
\psi &= a_{31}\alpha + a_{32}\beta + a_{33}\gamma + a_{34}\vartheta + a_{36}V + a_{37}H + a_{38}\delta_{yp} + a_{39}\delta_{B} + a_{310}\delta_{H},\\
\gamma &= a_{41}\alpha + a_{42}\beta + a_{44}\vartheta + a_{45}\psi + a_{46}V + a_{47}H + a_{49}\delta_{B} + a_{410}\delta_{H} + a_{411}\delta_{\text{э}},\\
\beta &= a_{53}\gamma + a_{54}\vartheta + a_{55}\psi + a_{56}V + a_{57}H + a_{510}\delta_{H} + a_{511}\delta_{\text{э}},\\
\alpha &= a_{62}\beta + a_{63}\gamma + a_{64}\vartheta + a_{65}\psi + a_{66}V + a_{67}H + a_{69}\delta_{B}\\
H &= a_{71}\alpha + a_{72}\beta + a_{73}\gamma + a_{74}\vartheta + a_{76}V
\end{aligned}\right. \qquad \ldots(1$$

where: a_{ij} - operator factors; V - velocity deviation; ϑ - angle of pitch deviation; ψ - yaw deviation; γ - bank angle deviation; β - angle of sideslip deviation; α - angle of attack deviation; H - altitude deviation, δ_{gp} - changing of thrust controller position; δ_{θ} - deflection of an altitude control; δ_{H} - deflection of an yaw rudder; δ_{ϑ} - deflection of ailerons.

We can see the interaction between the variables $V, \vartheta, \psi, \gamma, \beta, \alpha, H$ from the equations (I). The control system with four loops which would be connected with others trough the controlled member is composed at four controllers and four controlled variables. Obviously it would be difficult to develop the control system with no additional connections providing independence or insignificant dependence of control loops. Therefore a set of structures in the form of connected systems will be considered using the algorithms for composing a set of selectively invariant systems. The set of combined systems are not considered here.

2. GRAPH-A MATHEMATICAL MODEL OF MOTION EQUATIONS. SYNTHESIS OF LRV CONTROL SYSTEM.

The graphs with no loops are chosen as the basis of structural representation of LRV control systems are chosen to make easy the investigation of interconnections of LRV controlled variables and to allow choosing the controllers under the conditions of their most efficiency in control and also determining the set $G\{M,S,L\}$ of selectively invariant systems.

Then using equations (I) compose a graph-a mathematical model of LRV three-dimensional motion in earth atmosphere as of a controlled member (Fig.I).

Take V, γ, α as LRV controlled variables. So far as angle of sideslip β can be large at some LRV flight regimes and would greatly influece LRV three-dimensional trajectories it is necessary to cosider β as a controlled variable too.

Then separate the graph $G(x,\Gamma)$ (Fig.2) of controlled variables from the graph of LRV three-dimensional motion (Fig. 1) and consider a set of possible selectively invariant systems structures.

All the kinds of connections providing selective invariance of multivariable systems have been considered in detail [4]. It has been shown that autonomy in the sense of Voznesensky I.N. or invariency as to no own control commands or disturbances and both of them at the same time are possible depending upon the types of connection.

Let us begin consideration of the set of LRV control system structures with the set M in the form of great- trees. A number of the great-tree variances for a graph whose root has been chosen as bank angle γ can be determined with the theorem III (Appendix).

$$\Delta_{\gamma} = \begin{vmatrix} 2 & -1 & -1 \\ -1 & 3 & -1 \\ 0 & -1 & 3 \end{vmatrix} = 12$$

All the structures of the set M with bank angle γ as a root are shown im Fig.3. In a similar way we can determine a number of the great-trees for V $(\Delta_V = 12)$ and $\alpha(\Delta_\alpha = 16), \beta(\Delta_\beta = 8)$. Thus all the set M of control system structures in the form of great-trees would be 48. set S of the structural graphs. As well as before

Then let us begin consideration of set S with root γ will be $\Delta_\gamma = 3$.

The structural graphs for γ have been presented in Fig.4. A number of connections necessary for constructing the control system in the form of the structural graph is equal to $K=6$.

The investigation of the set S of structures gives large possibilities for choosing LRV control system satisfying to the requirements presented.

And at last let us consider the set L of autonomous groups of loops. Four versions of construction of autonomous groups are possible. ~~In the form of the symmetrix.~~ In this case a number of connections K will be:

a) for groups in the form of the symmetric subgraphs K=6,

b) for groups in the form of the great-trees K=9,

c) at triangular (or quasi-triangular) system matrix and in the presence of groups K=3.

Some structures in form of the autonomous groups or the symmetric subgraphs are presented in Fig.5.

It should be noted that in more complex cases when a number of controlled variables > 4 it is possible to construct an algorithm of looking all the set $G\{M,S,L\}$ of structures over and to obtain with digital computer all structures satisfying the requirement of implementation and given quality of control.

As an example of the above suggested method of synthe sis of LRV control system structures the synthesis of control system of hypothetical LRV has been made.

For the purpose of illustration the particular regime of LRV flight (namely thrusted fliight at constant altitude in the plane of minor circle) has been considered (Fig.6.)

Analysis of values of operator-factors a_{ij} shows that for this regime of flight some of them can be neglected because of their insignificance and then the controlled member is simplified to the graph presented in Fig.7. The reason for such a simplification have been proved with the transients presented in Fig.8. and computed by digital computer as a result of solving LRV linearised equations (I) at impulse changing of controllers $\delta_{\varphi}, \delta_{H}, \delta_{B}, \delta_{э}$.

Comparison of obtained control system structures allows to conclude that the structure (Fig.9.) in which additional connection (dotted line) is required seems to be preferable.

Peculiarity of the given structure is that the angle β is autonomous and invarient as to angles γ,α and flight velocity V. In this case the influence of the angle of sideslip β on the angles γ and α is simply a disturbance, whose calculation is very easy. The known disturbances influencing the velocity control loop will be those with the changing of the angles α, β, γ.

For example Laplace transformation of changing angle with disturbance of β loop will be as follows:

$$\alpha(s) = a_{\alpha\beta}(s)\, a_{\beta f}(s) \cdot f_\beta(s),$$

where: $f_\beta(s)$-disturbance in β control loop;
$a_{\beta f}(s)$-transfer function of β control loop at disturbance $f_\beta(s)$.

In such a way Laplace transformations of controlled variables can be written corresponding to the graph presented in Fig.9.

Let us take the following sequence of controllers synthesis assuming that α, γ and flight velocity controllers can be chosen without taking into consideration the connections through the controlled member. Obtain the transfer functions of the controllers $W_\alpha(s), W_\beta(s), W_v(s)$ for V, α, γ control and β stabilization on the basis of desired dynamics as follows:

$$\Phi_{\text{ж}}^{y_i}(s) = \Phi_i^{y_i}(s), \qquad (3)$$

where: $\Phi_{\text{ж}}^{y_i}(s)$-transfer function of closed loop, corresponding the desired dynamics;
$\Phi_i^{y_i}(s)$-transfer function of synthesised control loop.

From the equation (3) we can obtain formula of the controller transfer function for control regime:

$$W_i(s) = \frac{1}{W_{ii}(s)} \cdot \frac{A_i(s)}{\left[B_i(s) - A_i(s)\right]}, \qquad (4)$$

where: $\Phi_{\text{ж}}^{y_i} = \frac{A_i(s)}{B_i(s)}$; $W_{ii}(s)$ -transfer function of control loop for control command y_i.

Synthesis of transfer funcion $W_\zeta(s)$ will be made taking into account interconnections determined with the transfer functions of controlled variables α, γ connections.

From the graph presented in Fig.9 the controller transfer function $W_{\gamma}(s)$ according to (4) will be:

$$W_{\gamma}(s)=\frac{1}{W_{\gamma\gamma}(s)}\cdot\frac{A_{\gamma}(s)}{\left[B_{\gamma}(s)-\dot{A}_{\gamma}(s)\right]}\left[1-W_{\varepsilon}(s)\right], \qquad (5)$$

where: $$W_{\varepsilon}(s)=\frac{W_{\alpha\gamma}(s)\,W_{\gamma\alpha}(s)}{1+W_{\alpha\alpha}(s)\,W_{\alpha}(s)}$$

The coordinate transients can be computed from the graph presented in Fig.9 at control loops interaction. For example

$$\gamma_{\beta}(t)=\frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}\bar{\bar{\Phi}}_{\gamma\beta}(s)\,\beta_f(s)\,e^{st}\,ds,$$

where:

$$\bar{\bar{\Phi}}_{\gamma\beta}(s)=\frac{a_{\gamma\beta}(s)+a_{\alpha\beta}(s)\,a_{\gamma\alpha}(s)}{1-a_{\gamma\alpha}(s)\,a_{\alpha\gamma}(s)};\quad a_{\gamma\beta}(s)=\frac{W_{\gamma\beta}(s)}{1+W_{\gamma\gamma}(s)\,W_{\gamma}(s)};\quad a_{\alpha\beta}(s)=\frac{W_{\alpha\beta}(s)}{1+W_{\alpha\alpha}(s)\,W_{\alpha}(s)};$$

$$a_{\gamma\alpha}(s)=\frac{W_{\gamma\alpha}(s)}{1+W_{\gamma\gamma}(s)\,W_{\gamma}(s)};\quad a_{\alpha\gamma}(s)=\frac{W_{\alpha\gamma}(s)}{1+W_{\alpha\alpha}(s)\,W_{\alpha}(s)};\quad \beta_f(s)=\frac{W_{\beta\beta}(s)}{1+W_{\beta\beta}(s)\,W_{\beta}(s)}\,f_{\beta}(s)$$

f_{β} – disturbance;

$$V_{\beta}(t)=\frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}\bar{\bar{\Phi}}_{v\beta}(s)\,\beta_f(s)\,e^{st}\,ds,$$

where: $$\bar{\bar{\Phi}}_{v\beta}(s)=\frac{W_{v\beta}(s)}{1+W_{vv}(s)\,W_{v}(s)}$$

In a similar way other coordinate transients at control commands and disturbances $\alpha_{\beta}(t), \alpha_{\gamma}(t), \alpha_{y}(t)$ can be computed.

For example the coordinate transients have been computed with digital computer at control command y_v. As can be seen from Fig.IO the flight velocity transient satisfies to control quality requirements presented.

CONCLUSIONS

In this paper it has been shown that:

I. LRV in aerodynamic maneuver is a complex controlled member with connections of all coordinates. Therefore it is reasonable to consider a set of control system structures as multivariable systems providing independence or insignificant dependence of control.

2. Using the algorithms for composing the set of selectively

invariant system structures and decomposing a control system with the graph theory allows to choose the best structure in the sense of the quality of control and simplicity of implementation.

APPENDIX.

Some theorems used in this paper and proved in [5] are presented:

Theorem I. If a graph contains arcs of a loop the determinant $|S_n^n|$ of the incidence matrix for the arcs would be equal to zero.

Autonomous systems.

Let us consider the autonomous systems in sense of Voznesensky I.N.; first we shall transform the linear composition matrix of graph arcs and apexs as follows:

$$\left| S_i^j \right| = \begin{cases} -1 & \text{at} \quad i \neq j \quad (x_i; x_j) \in u \\ 0 & \text{at} \quad i \neq j \quad (x_i, x_j) \notin u \\ 1 & \text{at} \quad i = j \end{cases} \qquad \text{(I)}$$

Theorem II. A graph is autonomous when the square matrix determinant $|S_n^n|$ is equal to I at $\Gamma x = \Phi$(directly).

A number of connections necessary for the formation of this graph is

$$K = n(n-I)$$

Structural graph.

Define the structural graph with a root x_i [4] :

I) The structural graph is one $G(x, \Gamma)$ with a root x_i

2) The graph $G(x, \Gamma)$ has no loops .

3) No apex(root x_i) is included in the arc.

Let's begin with the consideration of the structural graphs which often appear to be the great-trees. Define a great-tree:

I) a great-tree is a graph with a root $x_i \in U$,

2) the graph (x, U) does not include loops,

3) each graph apex includes only one arc,

4) no graph root x_i includes an arc,
5) the graph (x, u) includes $(n-1)$ arcs. If n is a number of apex, m is a number of loops, then $m = n-1$.

Using the known Trent's theorem* about the number of trees and remembering that the great-tree is a tree with oriented arcs let us present the matrix as follows:

$$\|a_{ij}\| = \begin{cases} |\Gamma x_i| & \text{at } i=j \\ -1 & \text{at } i \neq j \quad (x_i, x_j) \in U \\ 0 & \text{at } i \neq j \quad (x_i, x_j) \notin U \end{cases}$$

<u>Theorem III.</u> The number of great-trees in a graph is equal to an element minor Δ_i of main diagonal in the square matrix with order n.

For symmetric graph a minor with a root x_i would be:

$$\Delta_1 = \begin{vmatrix} (n-1) & -1 & \dots & -1 \\ -1 & (n-1) & \dots & -1 \\ \cdot & \cdot & \dots & \\ \cdot & \cdot & \dots & \\ -1 & -1 & \dots & (n-1) \end{vmatrix} \tag{2}$$

As follows from (2) for symmetric graph $\Delta_1 = n^{(n-2)}$. Then:
1) the number of the connections K providing composition of a great-tree is equal to $K = n(n-1) - (n-1) = (n-1)^2$;
2) all the set M of the great-trees in symmetric graph is equal to $B = n^{(n-1)}$
3) in general the number B of possible versions of the great-trees is equal to $B = \Delta_i n$.

<u>Theorem IY.</u> The graph G is a structural one with a root x_i when and only when $\Delta = I$; in other cases $\Delta = 0$.

<u>Theorem V.</u> The number of the structural graphs is equal to a minor Δ_i of an element in the main diagonal of the square incidence matrix $\|a_{ij}\|$. In symmetric graph $G(x, \Gamma)$ the number of structural graphs with a root x_i is $(n-1)!$.

---

*) Trent H.M. A note of the enumeration and listing of all possible trees in a connected linear graph. Proc. Nat. Ac.Sciences, 40, I954, I004.

From this theorem the formula for determining number of connections providing composition of the structural graph has been obtained

$$K = \frac{n(n-1)}{2}$$

All the set of graph structures in symmetric graph is $B = n!$ In general case $b = \Delta_i n$.

Autonomous Groups, decomposition of system.

We can define the autonomous system as some control loops in the system having actual interconnections through the c ontrolled member.

Let's present the matrix $\|S^n_n\|$ in the form of:

$$\left\| S^n_n \right\| = \left\| \begin{matrix} S^K_K & 0 \\ 0 & S^{n-K}_{n-K} \end{matrix} \right\| \quad \text{or} \quad \left\| S^n_n \right\| = \left\| \begin{matrix} S^K_K & S^K_{n-K} \\ 0 & S^{n-K}_{n-K} \end{matrix} \right\| \qquad (3)$$

then as before:

$$\left\| S^n_n \right\| = \left| S^K_K \right| \times \left| S^{n-K}_{n-K} \right|$$

Four versions of the system are possible:

$$\begin{aligned} &\text{I) if } |S^K_K| = 0, \quad |S^{n-K}_{n-K}| = 0, \text{ then } \quad |S^n_n| = 0 \\ &\text{2) if } |S^K_K| = \text{I}, \quad |S^{n-K}_{n-K}| = 0, \text{ then } \quad |S^n_n| = 0 \\ &\text{3) if } |S^K_K| = 0, \quad |S^{n-K}_{n-K}| = \text{I}, \text{ then } \quad |S^n_n| = 0 \\ &\text{4) if } |S^K_K| = \text{I}, \quad |S^{n-K}_{n-K}| = \text{I}, \text{ then } \quad |S^n_n| = \text{I}. \end{aligned} \qquad (4)$$

The first version of multivariable systems corresponds to the case when all the autonomous groups are constructed in the form of the symmetric subgraphs. Some part of the groups, defined by $|S^K_K|$ or $|S^{n-K}_{n-K}|$ is constructed in form of structural graphs in the second and third versions; the forth version is a graph composed of the autonomous groups in the form of structural graphs.

In these cases the number of necessary connections (K) depends on the number of groups (ν) and controlled variables in group m:

I) for groups in the form of the symmetric subgraphs:

$$K = n(n-1) - \frac{1}{2}\sum_{z=1}^{\nu} m_z(m_z-1) \tag{5}$$

2) for groups in the form of the structural graphs:

$$K = n(n-1) - \frac{1}{2}\sum_{z=1}^{\nu} (m_z-1) \tag{6}$$

3) for groups in the form of the great-trees:

$$K = n(n-1) - \sum_{z=1}^{\nu} (m_z-1) \tag{7}$$

4) at the triangular (or quasi-triangular) system matrix and the presense of groups:

$$K = \frac{1}{2}\left[n(n-1) - \sum_{z=1}^{\nu} m_z(m_z-1)\right] \tag{8}$$

REFERENCES.

I. Bruce R.W. The combined aerodynamic propulsive orbit plane change maneuver. AIAA Paper 65-20.

2. Остославский И.В., Стражева И.В. Динамика полета, М., Оборонгиз, I963.

3. Петров Б.Н. О построении и преобразовании структурных схем, Изв.ОТН АН СССР, I945, № I2.

4. Колпакова Н.П. К теории систем с несколькими регулируемыми координатами. Труды П Всесоюзного совещания по теории инвариантности и ее применению. М., Наука, I964.

5. Колпакова Н.П. К вопросу синтеза селективно инвариантных систем. Труды Ш Всесоюзного совещания по теории инвариантности и ее применению (в печати).

6. Техническая кибернетика. Теория автоматического регулирования. Книга I, М., Машиностроение, I967.

43.3

OPTIMAL PARAMETRIC CONTROL FOR THE RE-ENTRY SPACE VEHICLE

Ponomarev V.M., Gorodezky V.I.
Leningrad State University, Leningrad, USSR.

Stochastic disturbances, acting on a space vehicle (SV) on the atmospheric part of the descent and the initial conditions scattering of the SV re-entry cause the great landing coordinate scatter, especially when the re-entry velocity is higher the first cosmic one. In this case the use of standard linear control equation of SV mass centre stabilization with isochronous parameter variations of the program motion may not give high accuracy. This tends to use more effective control methods.

So far as with the board space vehicle navigation unit and the computer: one may continuously define motion parameters and form control signals by enough complex algorithmes it makes possible to use parametric control programs when the program trajectory is given against some kinematic or dynamic trajectory parameter.

I. The problem formulation. Motion equations.

Let the mass center motion of the S.V. descenting in the atmosphere, is described by the following differential equations:

$$\frac{dV_x}{dt} = -KC_x\frac{\rho}{\rho_0}VV_x - KC_y\frac{\rho}{\rho_0}VV_y - \frac{g_z}{z}(x-x_c) - \frac{g_\omega}{\omega_3}\omega_{3x} +$$

$$+ a_{11}(x-x_c) + a_{12}(y-y_c) + a_{13}(z-z_c) + b_{12}V_y + b_{13}V_z \ ;$$

$$\frac{dV_y}{dt} = -KC_x\frac{\rho}{\rho_0}VV_y + KC_y\frac{\rho}{\rho_0}VV_x - \frac{g_z}{z}(y-y_c) - \frac{g_\omega}{\omega_3}\omega_{3y} + \qquad (1)$$

$$+ a_{21}(x-x_c)+a_{22}(y-y_c)+a_{23}(z-z_c) + b_{21}V_x + b_{23}V_z \ ;$$

$$\frac{dV_z}{dt} = -KC_x\frac{\rho}{\rho_0}VV_z + KC_z^{\beta}\frac{\rho}{\rho_0}V^2\beta - \frac{g_z}{z}(z-z_x) - \frac{g_\omega}{\omega_3}\cdot\omega_{3z} + \qquad (1)$$

$$+ a_{31}(x-x_c)+a_{32}(y-y_c)+a_{33}(z-z_c) + b_{31}V_x + b_{32}V_y \ ;$$

$$\frac{dx}{dt} = V_x \ ; \qquad \frac{dy}{dt} = V_y \ ; \qquad \frac{dz}{dt} = V_z \ .$$

These equations are given in the start coordinate system, related to the rotating Earth. The origin of the system is taken from the vertical cross-point passing through the SV mass centre at the time reference with the Earth surface. Y- axis coincides the vertical, X- axis is directed to the termination point of landing and Z- axis form the right coordinate system.

In equations (1) the following notation is given:

Vx, Vy, Vz, x,y,z - vector velocity components and the SV mass centre coordinates in the given coordinate system,

C_x, C_y, C_z^{β} - aerodynamic factors,

ρ/ρ_0 - the relative air density,

$V=\sqrt{V_x^2+V_y^2+V_z^2}$ - the velocity vector modulus,

x_c, y_c, z_c - coordinates of the Earth centre,

z - position vector of SV mass centre with the origin in the Earth centre,

$K=\frac{S\rho_0}{2m}$ - ballistic coefficient depending on the cross section area of the space vehicle S, its mass M and atmospheric density ρ_0 on Earth surface,

g_z, g_ω - components of gravitational acceleration,

α, β - vehicle angles of attack and sideslip accordingly,

$\omega_з, \omega_{зx}, \omega_{зy}, \omega_{зz}$ - angular velocity of the Earth and components of angular velocity in the start coordinate system related to the rotating Earth,

a_{ij}, b_{kl} - constants, which depend on $\omega_з, \omega_{зx}, \omega_{зy}, \omega_{зz}$ entrance latitude, initial asimuth trajectory of re-entry 1,

As it was shown above programmer trajectory should be given depending on any parameter with following stabilization of the SV mass centre in programmer trajectory according with a some control equation. From physical point of view such a control is equal feedback programmer control, that indirectly permit to take into account the stochastic disturbances, influenced over SV. In such case "the tube" of disturbanced trajectories will be more narrow and therefore the linear control quite sufficient.

The choice of the programmer argument is sufficiently complicated and now is not solved. In the case under consider programmer argument is the down range L, distance to destination along great circle rout.

Suppose that control in bank motion is performed with the angle of attack, and in side motion - with the angle of sideslip. α and β programmes are given in such a manner, that the accelerations along the nominal trajectory were less, than 3 g.

Those programmes are choosed in result of preliminary calculations

$$\alpha_{nom} = \begin{cases} 0.2 \ ; & if \ L \leqslant 2730\,km \\ 0.1 \ , & if \ 2730\,km < L \leqslant 3350\,km \\ 0.05 \ , & if \ 3350\,km < L \leqslant 3428\,km \\ 0 \ , & if \ L > 3428\,km \end{cases} \qquad (2)$$

$$\beta_{nom} = 0$$

Stability control equation of SV on the programmer trajectory should be search in such form:

$$\Delta\alpha = K_1(V_x - V_{x\,nom}) + K_2(V_y - V_{y\,nom}) + K_3(h - h_{nom}) + K_4(Z - Z_{nom}) \quad (3)$$

$$\Delta\beta = K_5(V_z - V_{z\,nom}) + K_6(Z - Z_{nom}) \quad (4)$$

where $V_{x\,nom}(L), V_{y\,nom}(L), V_{z\,nom}(L), h_{nom}(L), z_{nom}(L)$ — are the nominal values of velocity components.

altitude and side deflection accordingly,

$K_1, \ldots, K_6$ - the unknown (searched) coefficients of linear control equation,

$\Delta\alpha, \Delta\beta$ - controlling variations of angles of attack and sideslip.

Then

$$\alpha = \alpha_{nom} + \Delta\alpha \quad (5)$$

$$\beta = \Delta\beta \quad (6)$$

There are control boundaries of values α and β in this problem which are firstly because of limited control effectiveness, and secondly for limiting of maximum accelerations acting on SV:

$$|\alpha| \leq \bar{\alpha} \quad (7)$$
$$|\beta| \leq \bar{\beta}$$

The basic aim of the re-entry SV control is to reduce the terminal scattering. Therefore as minimization functional was taken the such one.

$$I = D[L] + D[z] + (M[L] - L_{nom})^2 + (M[z] - z_{nom})^2, \quad (8)$$

where $D[L], D[z], M[L], M[z]$ - dispersions and means L and Z in point of the SV landing (h = 0),

L^p and Z^p - nominal values of the corresponding parameters in point of the landing.

Minimization of the functional permits to solute the problem of SV leading to given point with the minimum scatter.

Disturbances, acting on the re-entry SV are stochastic.

The basic scattering of the SV ground-points arise because of the atmospheric dencity, cy / cx variations, and deflexion of coordinates and re-entry angle initial conditions. Suppose that initial conditions are given in form

$$x_0 = 0\ ,\quad y_0 = (100\,000 + \omega_1)[m]\ ,\quad z_0 = \omega_3\ [m]$$
$$V_0 = 7800\,[m/sec]\ ,\ \theta_0 = (0.055 + \omega_2)\ [rad]\ ,\ V_{z0} = 0 \qquad (9)$$

Here w_1, w_2, w_3 are uncorrelated and normal distributed stochastic coefficients. Their mean are zero and root-mean-square variations are known

$$\sigma(\omega_1) = 2000\ m$$
$$\sigma(\omega_2) = 0.0005\ rad$$
$$\sigma(\omega_3) = 3300\ m$$

The disturbances of atmospheric density as altitude function are given by with such canonic decomposition

$$\rho = \rho_{nom}\left(1 + \sum_{i=1}^{m} \omega_{i+3} f_i(h)\right) \qquad (10)$$

Suppose, the dispersions of the stochastic coefficients to be equal 1. Coordinate functionst can be get by calculation stochastic data about atmosphere density.

Suppose, that variations of parameter $C_y\ /\ C_x$ may be run up to 15%, and its distribution is normal.
Let us set the following problem of terminal control.

It is necessary to find such a control in form (3), which minimizes the functional (8) with given initial conditions (9) and boundaries (7).

2. Algorithm of problem solution.

When solving this problem we shall use the Consecutive Optimization Method[3]. This Method is founded on replacement of initial problem by consequences of convex quadratic programming próblem with linear boundaries.

There are stochastic functional in our problem, therefore additional difficulties arise here.

Algorithm of one step problem solving in such case consist of two independent stages: the achivement of quadratic approximation of functional and linearisation of boundadies. 2. The solving of quardratic programming problem.

Let us consider the peculiarities of each stages. The general difficulty in getting of quadratic approximation of minimized functional is connected with its stochastic character. It means, that functional's value may be calculated only approximately (by Monte-Carlo Method, Thernetsky's Interpolating Method[4], Dostupov's Method[5]. Moreover, it is not to state that the functional (8) is convex as for as differential equation (1) are nonlinear. But even it is convex it may be turn out after quadratic interpolation unconvex, that prevent using such well known methods of convexive programming[6].

Therefore when solving this problem one ought to carry out the quadratic interpolation of given initial functional with additional boundaries.

$$\lambda_i \geqslant 0, \quad i = 1, \dots, m \tag{11}$$

where λ_i - eigenvalues of quadratic matrix form got in result of interpolation the initial functional. It is evidently, that condition (12) is equal to requirement of convexity of the initial functional (8) quadratic approximation. Then interpolation problem is solved as some smoothing problem.

Let us give the short description of the best in root-mean-square sense convex approximation, which was used for solving synthesis parametric control equation to re-entry SV problem.

Let I^m - is the value of functional (8) in point K^m ($m = 1, \dots, l$) of control parameters space from permitting region. Let be some initial convex quadratic approximation of functional (8). It may be rather far from given functional, therefore its getting is not difficult.

Suppose, that its form is following

$$\bar{I}(K) = C^0 + \sum_{i=1}^{n} b_i^0 K_i + \sum_{i,j=1}^{n} a_{ij}^0 K_i K_j \tag{12}$$

All values $K^{(m)}$ are given, and the values of functional I in points K , where are given the accurate values of functional I are given may be written in form.

$$\bar{I} = C^0 + \sum_{i=1}^{n} b_i^0 K_i^{(m)} + \sum_{i,j=1}^{n} a_{ij}^0 K_i^{(m)} K_j^{(m)} \tag{13}$$

Matrix (a_{ij}^0) is symmetrical and positive. Let us set the following problem of quadratic programming. The problem is to find such values C, bi, ai , which realize the minimum of the auxiliary functional

$$\Phi = \sum_{m=1}^{\ell} (\bar{I}^m - I^m)^2 \tag{14}$$

and satisfate to boundaries

$$\lambda_i \geq 0 \tag{15}$$

$$|\bar{I}^m - I^m| \leq \delta, \quad m = 1, \dots, \ell \tag{16}$$

Boundaries (16) are made for exluding of the great difference between the functional (8) I and its convex approximation I. However, it is not essential in the most number of problems.

Let us introduce the vector

$$Y^m = (y_0^m, y_1^m, \dots, y_p^m) \tag{17}$$

by formulas

$$\begin{aligned}
&y_0^m = 1, \\
&y_1^m = K_1^m, \; y_2^m = K_2^m, \; \dots, \; y_n^m = K_n^m, \\
&y_{n+1}^m = K_1^m K_1^m, \; y_{n+2}^m = K_1^m K_2^m, \dots, \; y_{2n}^m = K_1^m K_n^m, \\
&y_{2n+1}^m = K_2^m K_2^m, \; y_{2n+2}^m = K_2^m K_3^m, \dots, \; y_{3n-1}^m = K_2^m K_n^m, \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \\
&y_p^m = K_n^m K_n^m, \\
&p = \frac{n(n+3)}{2},
\end{aligned} \tag{18}$$

It is evidently, that the components of vector Y^m are the ordered coefficients at c, b_i and a_{ij} in the equality (13).

Let us introduce also the vector to make a short note

$$\Omega = (\omega_0, \omega_1, \ldots, \omega_p) \tag{19}$$

and it's components ω_j are coefficients c, b_i and a_{ij} in the equality (13) at yi in according to designations (18). Then the minimized functional (18) is the quadratic function of components of vector Ω

$$\Phi(\Omega) = \sum_{m=1}^{\ell} \left(I^m - \sum_{i=0}^{p} y_i^m \omega_i\right)^2 \tag{20}$$

It is obviously, that the functional Φ may be made convex by choice of the magnitude y_i^m.

Boundaries (16) are linear. Boundaries (15) may be made linear:

$$\lambda_i(\Omega) \cong \lambda_i(\Omega^0) + \sum_{j=n+1}^{p} \frac{\partial \lambda_i(\Omega^0)}{\partial \omega_j}(\omega_j - \omega_j^0) \geqslant 0 \tag{21}$$

that may be written in such a way

$$-\sum_{j=n+1}^{p} \frac{\partial \lambda_i(\Omega^0)}{\partial \omega_j}\omega_j \leqslant \lambda_i(\Omega^0) - \sum_{j=n+1}^{p} \frac{\partial \lambda(\Omega^0)}{\partial \omega_j}\omega_j^0 \tag{22}$$

Thus, the problem of the best in the root-mean-square convex approximation of the initial functional is reduced to the problem of minimization quadratic functional by the linear boundaries (16), (22).

The convex programming problem was solved by Hildreth and D'Esopo's Method[6].

Let us consider some peculiarities of the second solving stage.

It is known that the several variables convex functional minimizing is quite difficult problem and for its solving it is required many of iterations. Below it is offered some methods for decreasing the number of the iterations and value of calculations.

In our practice it was used the quadratic functional optimization method widely in which on every step it was executed the optimization on the subset U, which dimension less than dimension of vector K. In that case every iteration consist of the solving the such problem:

$$\min_{\tilde{V} \subset V} I(K_1, K_2, \ldots, K_m) \tag{23}$$

where U is the region of the permitted control. In such method the general difficulty is in selection of subset V sequence which determines the number simultaneously optimized coefficients and the subset $\tilde{V}$ structure.

Sometimes the essential simplification may be get it as the quadratic functional matrix has the quasidiagonal form and in the such case it is possible the independent optimization on the coefficients of every block.

While practic calculating the quadratic form matrix elements are seldom equal zero. It is connected with the approximative character derivatives calculus as well as with correlation which take place.
However it values of some second mixed partial derivatives are far smaller as other it may consider that they are equal zero. That permit to reduce number of iterations and calculation volume as initial task is divided on several simpler problems.
However it may not be used always.

Another way may be indicated which may improve process of task solution both in regard to reduce volume of calculations and in regard to increase improving of the process convergence.

While numerical solution of optimization problem solving it may introduce some scale coefficients on different parameters of control in such a way that in the presence of limitations on parameter

$$\bar{K}_i \leqslant K_i \leqslant \bar{\bar{K}}_i, \quad i = 1, \ldots, m \tag{24}$$

differences $(\bar{\bar{K}}_i - \bar{K}_i)$ would be quantities of identical order. It was pointed out in the book[8] about it, and this fact was confirmed by the numerical optimization experience.

For convex functional it may be shown, then absolute value of a partial derivatives of functional on parameters of control $\partial I/\partial K$ are decreased while approaching to optimum. Therefore it may be stated, that on every optimization step value of gradient

$$\frac{\partial I}{\partial K} = \left(\frac{\partial I}{\partial K_1}, \frac{\partial I}{\partial K_2}, \ldots, \frac{\partial I}{\partial K_m} \right) \tag{25}$$

permits to estimate crudely possibility of functional decreasing

$$| \Delta I |^{j}_{max} \leqslant \left| \frac{\partial I}{\partial K_j} \right| | K_{j\,max} | \tag{26}$$

where $\Delta K_{j\,max}$ - as much as possible change of parameter K_j to the side, which corresponds to functional decrease with accordance to the normalized limitations.

Often the some parameter derivative are smaller than others, therefore it is carried out:

$$\left| \frac{\partial I}{\partial K_j} \right| |\Delta K_{j\,max}| >> \left| \frac{\partial I}{\partial K_i} \right| |\Delta K_{i\,max}| \tag{27}$$

so in that case

$$\left| \frac{\partial I}{\partial K_j} \right| \gg \left| \frac{\partial I}{\partial K_i} \right| \tag{28}$$

and values $\Delta K_{i\,max}$ and $\Delta K_{j\,max}$ have identical order in accordance with the normalization. It gives some grounds on r-th optimization step to carry out minimum search in the permissible subset of those parameters which have comparable derivative value and all they are more than others parameter derivatives.

It is required to repeat from step to step this method.

This algorithm simplification method leads to good results.

For the large dimension controlling vector K problem it may be given the good results with the combination of this two methods.

3. Results of calculations

The problem of optimal parametric control for the re-entry SV synthesis was solved on the computer M-20 in accordance with the described in section 2 algorithm.

There was obtained the nominal trajectory with the α - and β - program (2) and with zero disturbances ($\omega_1 = \omega_2 = \cdots = \omega_7 = 0$):

While solving this problem it was identified that the $K_1 \div K_4$ and K_5 , K_6 optimization may be done independently because of quasidiagonal form of quadratic functional. Then the sensitive analysis showed, that the sensitive of the functional (8) coefficients K_4 and K_5 variations in environment of initial approximation

$$K_1 = K_2 = \cdots = K_6 = 0$$

are enough little and the essential decrease of functional may be arise by means of coefficients $K_1 \div K_3$ and K_6 optimization at first.

While solving it was used the consecutuve Optimization Method[3].

The stochastic characteristics L and Z were calculated by Dostupov's Method[5]. The initial functional value $(K_1 = K_2 = \cdots = K_6 = 0)$ was as follows

$$I = 0.162 \cdot 10^{12} \ [m^2]$$

and the maximum variation of SV landing point was

$$\Delta R_{max} = 1298 \ [km]$$

Such solution was got after the first optimization step.

$$K_1 = -0.3 \cdot 10^{-3}, \quad K_2 = -0.35 \cdot 10^{-3}, \quad K_3 = -0.1 \cdot 10^{-4},$$
$$K_4 = 0 \ , \qquad K_5 = 0 \ , \qquad K_6 = 0.4 \cdot 10^{-4};$$
$$I^{(1)} = 0.552 \cdot 10^{8} \ [m^2],$$
$$\Delta R_{max} = 14.02 \ [km]$$

Sensitive analysis on second and third optimization steps was permitted on every step correctly the subset of optimizating coefficients.

The final solution of this problem is as follows

$$K_1 = -0.49\cdot10^{-3},\ K_2=-0.26\cdot10^{-3},\ K_3=-0.8\cdot10^{-5},$$
$$K_4=-0.3\cdot10^{-5},\ K_5=0.6\cdot10^{-4},\ K_6=0.46\cdot10^{-4},$$
$$I^{(3)}=0.407\cdot10^{8}\,[m^2]$$
$$\Delta R_{max}=1.2\,[km]$$

How the analysis of disturbanced SV centre of mass trajectory showes the transient responces on velocity vactor components and coordinates are good, and maximal value of manipulated variables $\Delta\alpha$ and $\Delta\beta$ should take place it initial variations x,z and θ are great enough.

This solution was got to SV re-entry asimuth A= 0, but as the checking showed, this solution gave the good results when SV re-entry asimuth had values from the -90° 90° range.

Small sensitivity of the criterion (8) to variations of coefficients $K_1 \div K_6$ was made clear in result of the solution character in the neighbourhood of the optimal control analysis. It have been permitted to reduce the requirements of their tasking accuracy.

Literature

I. Аппазов Р.Ф., Лавров С.С., Мишин В.П. "Баллистика управляемых ракет дальнего действия". "Наука", I966.

2. "Управление космическими летательными аппаратами" Под ред. К.Т.Леондеса. "Машиностроение", I967.

3. Пономарев В.М. "Метод последовательной оптимизации в задачах управления". Известия АН СССР "Техническая кибернетика", № 2, I967.

4. Чернецкий В.И. "Анализ точности нелинейных систем управления". "Машиностроение", I968.

5. Доступов Б.Г. "Приближенное определение вероятностных характеристик выходных координат нелинейных систем автоматического регулирования"."Автоматика и телемеханика", т.I8, I957, № II.

6. Кюнци Г.П., Крелле В. "Нелинейное программирование". "Советское Радио", I965.

7. Д.Дж.Уайльд "Методы поиска экстремума"."Наука",I967.

43.4

STOCHASTIC OPTIMIZATION OF SPACESHIP REENTRY CONTROL IN ATMOSPHERE

A.G.Vlasov, E.I.Mitroshin, I.S.Ukolov
Moscow, USSR

Abstract

The paper considers the spaceship atmospheric reentry control problem which is reduced to a problem of optimal control of the terminal state of a certain stochastic system. The control algorithm is produced by using the nominal trajectory, the acceleration measurements being considered as the information source. The problems of programming the nominal trajectory of motion and providing the conditions of its actual realization are investigated simultaneously.

- x -

Among the problems connected with the manned spaceship flight a prominent place belongs to the problem of safe reentry into the Earth atmosphere. In solving this problem there arise great difficulties, particularly in case of reentry into the atmosphere with superorbital velocities. A reentry control system must secure the spaceship landing into the given region with due regard of limitations of accelerations, aerodynamic heating etc.

Since disturbing factors in the reentry process (initial scatter of reentry parameters, fluctuations of atmospheric density ets.) are stochastic with given probability characteristics the reentry control problem should be considered as a stochastic one. Without loss of generality one may assume the latter to be reduced to the optimal control problem of the terminal state of a stochastic dynamic system described by the set of nonlinear differential equations:

$$\dot{x} = X(x, u, h, t) \tag{1}$$

where x - is a state vector of dimension $(n \times 1)$ of the system; u is a control vector of dimension $(z \times 1)$, usually belonging to a closed region U; h- is a stochastic disturbance vector; X is a known vector function; t is an independent variable (time or one of state coordinates; for simplicity later on t denotes time; $t \in [0, T]$); $\dot{}$ - is a symbol of differentiating with respect to t.

Information about the current state of the system as a result of observations usually made by using autonomous devices on board the spaceship are represented in the form

$$y = Y(x, e, t) \tag{2}$$

where y- is an observed vector of dimensions $(l \times 1;\ l \leq n)$ (e.g. acceleration vector); e- is a stochastic error vector; Y is a known vector function.

It is necessary to obtain the extremum of a terminal state function (e.g. the minimum of down range scatter or the minimum of heat supplied to the spaceship during the reentry etc.).

$$J = M\,\tilde{\omega}\left[x(T)\right] \tag{3}$$

where M - is a symbol of mathematical expectation; $\tilde{\omega}$ - is a scalar nonnegative function, with due regard of the following boundary conditions: $X(0)$ is a vector of stochastic quantities with definite probability characteristics; at the moment T :

$$T \in \{T: \Omega[x(T), T] = 0\} \quad (4)$$

where Ω is a nonlinear function, the relation

$$x(T) \in \{x: g_K(x) = 0\} \quad (5)$$

(where g_K is a nonlinear vector function) must be satisfied with a definite degree of probability.

Spaceship reentry problems are characterized by a particular requirement of meeting the current phase restriction of inequality type (e.g. maximum acceleration limit); in other words, with a definite degree of probability there must be satisfied the following condition:

$$g(t, x) \leq 0 \quad (6)$$

where g is a nonlinear vector function of dimension $(m \times 1)$.

Stochastic disturbance vectors $\mathcal{F}$ and the measured errors e in a general case include the stochastic parameters η (e.g. scatter in spaceship parameters and initial conditions of reentry) and the stochastic processes q (e.g. atmospheric density fluctuations) as well.

By using the shaping filters the stochastic processes q can be approximated in the form of solutions of differential equations of the form

$$\dot{q} = fq + \varepsilon \quad (7)$$

here f - is a known matrix; ε - is a vector of stochastic

δ -correlated processes ("white noise").

If for stochastic parameters η we write formally the equation of the shaping filter in the form

$$\dot{\eta} = 0 \tag{8}$$

and if we assume now X to be the expanded state vector $\begin{pmatrix} x \\ q \\ \eta \end{pmatrix}$ and $\mathcal{X}$ to be the expanded vector function, respectively, then without loss of generality equations (I) and (2) can be written as follows:

$$\dot{x} = \mathcal{X}(x, u, \varepsilon, t) \tag{9}$$

$$y = \mathcal{Y}(x, \xi, t) \tag{IO}$$

where ε, ξ - denote "white noise".

In such a general formulation the problem investigation is greatly complicated due to nonlinearity of the set of equations (9)-(IO).

However assuming the disturbing factors are negligible and, consequently, the disturbed trajectory is close to the theoretical (nominal) one it is possible to use a linearization method for describing the disturbed motion thus significantly simplifying the investigation of the problem.

Supposing that disturbances acting on the nominal trajectory are equal to zero one may assume the investigated dynamic system be described by the following set of differential equations:

$$\dot{x} = \mathcal{X}(x, u, t) \tag{II}$$

$$\dot{x}_1 = A(x, u, t)_{x_1} + B(x, u, t)\,v + \varepsilon(x, u, t) \tag{I2}$$

$$y_1 = H(x, u, t) x_1 + \xi(x, u, t) \qquad (13)$$

where X - is a system state vector in case of motion along the nominal trajectory; X_1-is a generalized system state vector for the case of disturbed motion; y_1-is an observed vector; ε, ξ - denote "white noise"; U - is a control vector for the case of motion along the nominal trajectory; V is a control vector for the case of motion along the disturbed trajectory; χ - is a known vector function corresponding to the motion along the nominal trajectory; A, B, H are matrices of corresponding dimensions.

The boundary conditions at the ends are written in the form:

$$X(0) \in g_0; \quad x(T) \in \{x : g_K(x) = 0\};\ T \in \{T : \Omega[x(T), T] = 0\} \qquad (14)$$

where $X_1(0)$ is a vector of stochastic quantitis with definite probability characteristics.

For the problems of spaceship descent in the atmosphere it is typical the following relation imposed on the control vector in case of motion along the nominal and disturbed trajectories:

$$U + V \in \mathcal{V} \qquad (15)$$

In accordance with the above mentioned assumptions, the current restriction of the inequality type (6) and optimized functional (3) (for simplicity later on it will be assumed as minimized) may be written as follows:

$$g(t, x) + g_x(t, x) x_1 \leq 0 \qquad (16)$$

where g_x is a matrix of partial derivatives of g with respect to X ,

$$\mathcal{J} = M\,\tilde{\omega}[X(T) + X_1(T)] \tag{17}$$

If $\tilde{\omega}$ is considered to be the function

$$\tilde{\omega}[X(T) + X_1(T)]^T \Lambda [X(T) + X_1(T)] \tag{18}$$

where $[\ ]^T$ – is a symbol of transposition; Λ is a weight matrix, then the expression for $\mathcal{J}$ can be written in the following full form:

$$\mathcal{J} = \omega[X(T)] + M\omega_1[X_1(T)] + \omega_2[X(T), MX_1(T)] \tag{19}$$

where $\omega, \omega_1, \omega_2$ are scalar functions, the first two of which being non-negative.

The problems of programming the nominal motion and securing its actual realization are usually investigated separately. The former is considered as a dereminate optimal guidance problem from the minimization conditions

$$\mathcal{J}_0 = \omega[X(T)] \tag{20}$$

The latter problem being stochastic reduces to the optimal regulator synthesis from the minimization conditions

$$\mathcal{J}_1 = M\omega_1[X_1(T)] \tag{21}$$

It is natural to solve both problems simultaneously. The importance of such an approach to solving the optimitation problem was pointed out in Reference[1] .

Consider at first the disturbed motion, since the optimal regulator, as it will be shown later on, is structurally invariant relative to the nominal motion parameters. It will be supposed that the nominal motion is prescribed and matrices A, B, H and "white noise" in equations (I2)-(I3) are only time functions.

For the optimal regulator synthesis we shall use the dynamic programming method and the concept of sufficient coordinates[3]. The sufficient coordinates are the coordinates of a space in which Bellman functional equation is being considered. Their introduction permits to formally separate the problems of information processing and optimal synthesis. In solving the first problem sufficient coordinates are defined by using linear and nonlinear optimum filtration techniques. The defined sufficient coordinates are used in the second problem for optimum synthesis by solving Bellman equation.

It is assumed that $X_1(0), \mathcal{E}(t), \xi(t)$ are independent and normally distributed:

$$
\begin{gathered}
M[x_1(0)] = 0 \qquad M[\mathcal{E}(t)] = M[\xi(t)] = 0 \\
M[x_1(0)X_1^T(0)] = K_0; \; M[\mathcal{E}(t)\mathcal{E}^T(\tau)] = Q(t)\delta(t-\tau); \\
M[\xi(t)\xi^T(\tau)] = G(t)\delta(t-\tau);
\end{gathered}
\tag{22}
$$

where K_0 is an a priori covariance matrix of the vector $X_1(0)$; and $Q(t), G(t)$ - are known intensity matrices of gaussian "white noise" $\mathcal{E}(t), \xi(t)$.

Then for the investigated dynamic system (I2)-(I3) the vector of sufficient coordinates coincides with the vector of a posteriori mathematical expectation Z of the generalized state vector X_1 that can also be defined as the vector of the optimal estimation according to the method of the a posteriori probability density maximum by solving the following set of differential equations (Kalman filter):

$$\dot{Z} = AZ + BU + PH^T G^{-1}(y - Hz) \tag{23}$$

where P - is a covariance matrix of the estimation errors and is determined from the differential equations (24) and $[\ ]^{-1}$ is an matrix inversion symbol

$$\dot{P} = AP + PA^T - PH^T G^{-1} HP + Q \tag{24}$$

Equations (23)-(24) are solved under the initial conditions

$$z(0) = M[x_r(0)] = 0; \qquad P(0) = K_0 \tag{25}$$

Accordingly Bellman equation is written as follows:

$$-\frac{\partial \varphi(t,z)}{\partial t} = \min_{v(t) \in U(t)} [\varphi_z(Az + Bv) + \tfrac{1}{2} Sp(\varphi_{zz} f)]; \quad \varphi(z,t) = \min_{v(\tau);\, \tau \in [t,T]} M\{\omega_1[x_1(T)]/z(t)\}, \tag{26}$$

where φ is a loss function; φ_z is a vector of gradient φ with respect to z ; φ_{zz} is a matrix of the second partial derivatives φ with respect to z ; Sp is a spur of matrix, $f = PH^T G^{-1} H^T P$; $M[/]$ is a symbol of conditional mathematical expectation. This equation must be solved with the boundary condition

$$\varphi[z, T] = M\{\omega_r[x_r(T)]/z(T)\} \tag{27}$$

The synthesis of an optimal regulator on the basis of the solution of Bellman equation is extremely difficult in a general case. Consider a simplified problem formulation assuming that v is a scalar quantity and the function $\omega_r[x_r(T)]$ is

$$\omega_r[x_r(T)] = \{\chi[x_r(T)]\}^2 \tag{28}$$

where χ is a linear form.

Then, adding an additional component $\chi[x_r(T)]$ to the vector $x_r(t)$, multidimensional synthesis of an optimal regulator can be reduced to a one dimensional form by transition to a new variable according to the formula [2] :

$$\rho(t) = \Phi(T,t)x_r(t) \tag{29}$$

where $\Phi(T,t)$ is a fundamental matrix of a homogeneous equation corresponding to equation (I2) and $\Phi(T,t)$ satisfies the following relations:

$$\dot{\Phi}(T,t) = -\Phi(T,t)A;\quad \Phi(T,T) = E \tag{30}$$

where E is a unity matrix.

$$\begin{aligned} &\dot{\rho} = \bar{B}v + \bar{\varepsilon};\quad y_r = \bar{H}\rho + \xi;\quad \rho(T) = x_r(T); \\ &\bar{B} = \Phi(T,t)B;\quad \bar{H} = H\Phi^{-1}(T,t);\quad \bar{\varepsilon} = \Phi(T,t)\varepsilon; \end{aligned} \tag{3I}$$

Equations for the a posteriori mathematical expectation of ρ and the a posteriori covariance matrix of estimate errors P are written respectively

$$\dot{\bar{z}} = \bar{B}v + \bar{P}\bar{H}^T G^{-1}(y_r - \bar{H}\bar{z}),\quad \dot{\bar{P}} = \bar{Q} - \bar{P}\bar{H}^T G^{-1}\bar{H}\bar{P};$$

where $\bar{z}$ is a vector of the optimal estimate ρ and $\bar{P}$ is a covariance matrix of the estimate errors ρ ;

$$\bar{Q} = \Phi(T,t)Q\Phi^T(T,t);\quad \bar{z}(0) = 0;\quad \bar{P}(0) = \Phi(T,0)K_0\Phi^T(T,0)$$

Assuming that $\chi[x_r(t)]$ is a first coordinate of the expanded state vector $x_r(t)$ and denoting by the symbol " ' ", here and later on, the first coordinates and the first lines of vectors and matrices respectively, we may write Bellman equation as:

$$-\frac{\partial \varphi(\bar{z}',t)}{\partial t} = \min_{\vartheta(t)\in U(t)} [\varphi_{\bar{z}}, \bar{B}'\vartheta + \tfrac{1}{2}\varphi_{\bar{z}\bar{z}'}, \mathcal{F}'] \tag{32}$$

where $\mathcal{F}' = (\bar{p}\bar{H}^{T}G^{-1})G(\bar{p}\bar{H}^{T}G^{-1})'^{T}$

This equation must be solved under the boundary conditions

$$\varphi(\bar{z}',T) = M\{[\rho'(T)]^2/\bar{z}'(T)\} \tag{33}$$

However, for the optimal synthesis, in this case, it is not necessary to solve Bellman equation.

Consider the case of discrete information input that is typical for atmospheric reentry problems, provided the information is processed by the digital computer on board the spaceship. Assume the information enters at the discrete time $0 < t_N < t_{N-1} < \dots < t_1 < t_0$ $(t_0 < T)$ with the discrete interval Δt in which ϑ is supposed to be constant. Then the differential equation for $\rho'(t)$ and Bellman equation are substituted by recurrent equations:[2]

$$\rho'_{K-1} = \rho'_K + \vartheta_K \int_{t_K}^{t_{K-1}} \bar{B}'d\tau + \int_{t_K}^{t_{K-1}} \mathcal{E}'d\tau \tag{34}$$

$$\varphi_K(\bar{z}'_K) = \min_{\vartheta_K \in V_K} M[\varphi_{K-1}(\bar{z}'_K + \vartheta_K \int_{t_K}^{t_{K-1}} \bar{B}'d\tau + \nu'_K)/\bar{z}'_K] \tag{35}$$

$$\nu'_K = \int_{t_K}^{t_{K-1}} \mathcal{E}'d\tau - \bar{\delta}'_K + \bar{\delta}'_{K-1}; \quad \bar{\delta}'_K = \bar{z}'_K - \rho'_K$$

where index "K" corresponds to the moment t_K. Supposing that the region of permissible values of ϑ is asymmetrical,

$$\vartheta_{K\,min} \le \vartheta_K \le \vartheta_{K\,max} \tag{36}$$

what is natural to expect taking into account that the contrls are interconnected in nominal and disturbed motions following from the condition (I5), it can be shown that the optimal control $\vartheta_{K\,opt}$ in the interval $[t_K, t_{K-1}]$ is

$$\vartheta_{K\,opt} = \begin{cases} \vartheta_{K\,min}, \text{ if } \operatorname{sign}\left\{\left[\bar{Z}'_K + \sum\limits_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau\right] \cdot \int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right\} = +1; \\ \vartheta_{K\,max}, \text{ if } \operatorname{sign}\left\{\left[\bar{Z}'_K + \sum\limits_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau\right] \cdot \int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right\} = -1; \\ \vartheta_{K\,max}\left|\int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right| \le -\left[\bar{Z}'_K + \sum\limits_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau\right] \operatorname{sign}\int_{t_K}^{t_{K-1}} \bar{B}'d\tau \le \\ \quad \le \vartheta_{K\,min}\left|\int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right|; \\ \bar{Z}'_K + \vartheta_K \int_{t_K}^{t_{K-1}} \bar{B}'d\tau + \sum\limits_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau = 0; \\ \vartheta_{K\,max}\left|\int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right| > -\left[\bar{Z}'_K + \sum\limits_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau\right] \operatorname{sign}\int_{t_K}^{t_{K-1}} \bar{B}'d\tau > \\ \quad > \vartheta_{K\,min}\left|\int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right| \end{cases} \qquad (37)$$

Hence it appears that on every segment Δt the optimal regulator with maximum speed tends to combine the optimal estimate $\bar{Z}'$ with a nonzero quantity in

$$\left[-\sum_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau\right]$$

conditioned by the asymmetry of the region of permissible values of the control except a certain region of non-unique values of the optimal control

$$\vartheta_{K\,max}\left|\int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right| > -\left[\bar{Z}'_K + \sum_{i=K-1}^{0} \frac{\vartheta_{i\,max}+\vartheta_{i\,min}}{2} \int_{t_i}^{t_{i-1}} \bar{B}'d\tau\right] \cdot \operatorname{sign}\int_{t_K}^{t_{K-1}} \bar{B}'d\tau > \vartheta_{K\,min}\left|\int_{t_K}^{t_{K-1}} \bar{B}'d\tau\right|; \qquad (38)$$

which depends on the discrete interval Δt and the permissable values of the control.

Coming to the limit $\Delta t \to 0$ in (37) we have

$$v_{opt}(t) = \begin{cases} v_{min}(t), \text{ if } \bar{x}'(t) \operatorname{sign} \bar{B}'(t) > -\int_t^T \frac{v_{max}(\tau)+v_{min}(\tau)}{2} \cdot \bar{B}'(\tau)d\tau \cdot \operatorname{sign} \bar{B}'(t); \\ 0, \quad \text{if } \bar{x}'(t) = -\int_t^T \frac{v_{max}(\tau)+v_{min}(\tau)}{2} \bar{B}'(\tau)d\tau; \\ v_{max}(t), \text{ if } \bar{x}'(t) \operatorname{sign} \bar{B}'(t) < -\int_t^T \frac{v_{max}(\tau)+v_{min}(\tau)}{2} \bar{B}'d\tau \cdot \operatorname{sign} \bar{B}'(t) \end{cases} \tag{39}$$

Thus, the algorithm of the optimal control during the continuous information input is completely determined without direct solving Bellman equation. In case of continuous information input, in contrast to the determinate case $[\varepsilon(t)=\xi(t)=0]$, where the optimal control has not unique values in the so called "complete controllability zone", in the stochastic problem the optimal control is defined uniqly on the basis of (39); at every time the optimal control with maximum speed tends to combine the optimal estimate $\bar{x}'(t)$ with the quantity

$$-\int_t^T \frac{v_{max}(\tau)+v_{min}(\tau)}{2} \bar{B}'(\tau)d\tau$$

conditioned by the asymmetry of permissible value region of the control.

Substituting the expression for the optimal control v_{opt} (39) in Bellman equation (32), from the solution of the latter we can define that part of the minimized functional $\mathcal{J}$ which is conditioned by the disturbed motion $-\mathcal{J}_1$.

$$\min \mathcal{J}_1 = \min M[\rho'(T)]^2 = \varphi[\bar{x}'(0), t=0], \tag{40}$$

Since it is very difficult to obtain the exact solution of Bellman equation we use an approximate method of solution, that is, parametric method.

The parametric method consists in approximating the function $\varphi(\bar{z}', t)$ as a function of a finite number of parameters

$$\varphi(\bar{z}', t) = \psi^T(\bar{z}')\, a(t); \tag{41}$$

where ψ is a vector function of the known expansion functions; a is a $(s \times 1)$ -dimensional vector of unknown parameters.

Imposing on the parameters a certain natural conditions for the exactness of approximating function $\varphi(\bar{z}', t)$ and using Bellman equation, we can get a set of ordinary differential equations with boundary condition at the moment T which describes the evolution of the parameters in reverse time.

These equations, respectively, for different parametric methods described in Refs.[3,4] are written as follows:

$$\text{I)}\ \dot{a} = -\Big[\int_{\bar{z}_H^*}^{\bar{z}_6^*} \psi\psi^T d\bar{z}'\Big]^{-1} \int_{\bar{z}_H^*}^{\bar{z}_6^*} \big(\bar{B}' V_{oni} \psi_{\bar{z}'}^T + \tfrac{1}{2}\psi_{\bar{z}'\bar{z}'}^T f'\big) a \cdot \psi\, d\bar{z}' \tag{42}$$

$$\text{2)}\ \big[\psi^T/_{\bar{z}'=\bar{z}_1'(t)} - \psi^T/_{\bar{z}'=\bar{z}_K'(t)}\big]\dot{a} = \big[\big(V_{ons} \bar{B}' \psi_{\bar{z}'}^T + \tfrac{1}{2}\psi_{\bar{z}'\bar{z}'}^T f'\big)/_{\bar{z}'=\bar{z}_K'(t)} \tag{43}$$

$$-\big(V_{ons} \bar{B}' \psi_{\bar{z}'}^T + \tfrac{1}{2}\psi_{\bar{z}'\bar{z}'}^T f'\big)/_{\bar{z}'=\bar{z}_1'(t)}\big] \cdot a; \quad K = 2, 3 \ldots (S+1)$$

where $[\bar{z}_H^*, \bar{z}_6^*]$ denotes the region of permissible values $\bar{z}'$; $\bar{z}_i'(t)$ are certain prescribed functions $\bar{z}'$; $\psi_{\bar{z}'}$ is a vector of gradient ψ with respect to $\bar{z}'$; $\psi_{\bar{z}'\bar{z}'}$ is a matrix of the second partial derivatives ψ with respect to $\bar{z}'$.

The equations are solved under the boundary conditions

$$a(T) = a_T \tag{44}$$

where a_{γ} follows from the conditions of securing (33).

The quantity $\min \mathcal{J}_1$ is written as follows:

$$\min \mathcal{J}_1 = \varphi[\bar{z}'(0), t=0] = \psi^T[\bar{z}'(0)]\cdot\alpha(0) \tag{45}$$

As mentioned above, because of the stochastic nature of the component of the vector x_1 we may speak about solving the inequality (I6) only with some a priori level of probability; therefore it is necessary to know the a priori probability density of vector x_1 or its substitute parameters at every time t. In this case the estimation of the exactness of the stochastic vector means obtaining the vector of an a priori mathematical expectation and an a priori covariance matrix.

As we are interested in the estimation of the exactness of the stochastic vector $g_x x_1$ (of dimension $m \times 1$) then by introducing the expanded vector of state $x_2 = \begin{pmatrix} g_x x_1 \\ x_1 \\ \frac{\rho}{z} \end{pmatrix}$ we can state that for the system of the type

$$\dot{x}_2 = S(t)x_2 + F(t, x_2) + G(t) \tag{46}$$

where the matrix $S(t)$, nonlinear vector-function $F(t,x_2)$ and "white noise" $G(t)$ correspond to the expanded state vector x_2. It is necessary to define the first m components of the vector

$$\bar{x}_2(t) = Mx_2(t) \tag{47}$$

and the matrix consisting of the first m lines and columns of the matrix

$$\bar{\bar{x}}_2(t) = M[x_2(t) - \bar{x}_2(t)][x_2(t) - \bar{x}_2(t)]^T \tag{48}$$

Supposing that $x_2(0)$ and $G(t)$ are independent and normally distributed

$$\bar{x}_2(0)=x_{20}\,;\quad \bar{\bar{x}}_2(0)=K_{20}\,;\quad MG(t)=0;\qquad (49)$$
$$M[G(t)G^T(\tau)]=K(t)\delta(t-\tau)$$

and using the method of the statistic linearization , we can show that $\bar{x}_2(t)$ and $\bar{\bar{x}}_2(t)$ approximately satisfy the following set of differential equations:

$$\dot{\bar{x}}_2=S(t)\bar{x}_2+\Psi_1(t_1,\bar{x}_2,\bar{\bar{x}}_2);\quad \bar{x}_2(0)=x_{20}\qquad (50)$$

$$\dot{\bar{\bar{x}}}_2=[S(t)+\Psi_2(t_1,\bar{x}_2,\bar{\bar{x}}_2)]\bar{\bar{x}}_2+\bar{\bar{x}}_2[S(t)+\Psi_2(t_1,\bar{x}_2,\bar{\bar{x}}_2)]^T+K\qquad (51)$$
$$\bar{\bar{x}}_2(0)=K_{20}$$

where Ψ_1, Ψ_2 are functional relations whose definition ultimately reduces to the solution of the following relation

$$\lambda(t,M_{\bar{z}'},\sigma_{\bar{z}'})=\frac{1}{2}[\bar{v}(t)+\bar{\bar{v}}(t)]+[\bar{v}(t)-\bar{\bar{v}}(t)]\cdot\Phi\left[\frac{\bar{z}'^{*}-M_{\bar{z}'}}{\sigma_{\bar{z}'}}\right]\qquad (52)$$

where $M_{\bar{z}'}$ and $\sigma^2_{\bar{z}'}$ are components of $\bar{x}_2$ and $\bar{\bar{x}}_2$ corresponding to the coordinate $\bar{z}'$:

$$\bar{v}(t)=\begin{cases}v_{max}(t), & \text{if sign } \bar{B}'=+1;\\ v_{min}(t), & \text{if sign } \bar{B}'=-1;\end{cases}\qquad \bar{\bar{v}}(t)=\begin{cases}v_{max}(t), & \text{if sign } \bar{B}'=-1;\\ v_{min}(t), & \text{if sign } \bar{B}'=+1;\end{cases}$$

$$\bar{z}'^{*}=-\int_t^{T}\frac{v_{max}(\tau)+v_{min}(\tau)}{2}\,\bar{B}'(\tau)\,d\tau;$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt\qquad (53)$$

where Φ is a probability integral.

Neglecting, for simplicity, the correlation between components of the vector $(x_2-\bar{x}_2)$ it is possible to assume, in accordance with the rule "3σ", that the requirement for solving the stochastic inequality (16) reduces to the necessity of solving the following determinate inequality

$$g(t,x) + \bar{x}_m(t,x) \pm 3\sigma_m(t,x) < 0 \tag{54}$$

where $\bar{x}_m$ is a vector including the first m components of the vector x_2 and σ_m is a vector including the square roots of the first m diagonal elements of the matrix $\bar{\bar{x}}_2$.

Now, having determined the structure of the optimal regulator, estimated that part of the functional which is conditioned by the disturbed motion and estimated the exactness of the generalized state vector in case of disturbed motion we may come to the conclusion that the joint consideration of the determinate problem of programming the nominal motion and stochastic problem of securing its actual realization reduces to the following determinate extremal problem.

Given a set of ordinary differential equations (notations are given above) prescribed on the segment $[0,T]$

$$\dot{x} = X(x,u,t); \quad \dot{\Phi}(T,x,t,u) = -\Phi(T,t,x,u)A(x,u,t);$$

$$\dot{\bar{P}}(x,u,t) = \bar{Q}(x,u,t) - \bar{P}(x,u,t)\bar{H}^{T}(x,u,t)G^{-1}(x,u,t)\bar{H}(x,u,t)\bar{P}(x,u,t);$$

$$\dot{\alpha}(x,u,t) = -\left[\int_{\bar{z}_H^*}^{\bar{z}_6^*}\Psi(\bar{z}')\Psi^{T}(\bar{z}')d\bar{z}'\right]^{-1}\int_{\bar{z}_H^*}^{\bar{z}_6^*}\Big[V_{onT}(x,u,t,\bar{z}')\bar{B}'(x,u,t)\Psi^{T}_{\bar{z}'}(\bar{z}') +$$

$$+\frac{1}{2}\Psi^{T}_{\bar{z}'\bar{z}'}(\bar{z}')\bar{A}'(x,u,t)\Big]\cdot\alpha(x,u,t)\Psi(\bar{z}')d\bar{z}';$$

$$\dot{\bar{x}}_2(x,u,t) = S(x,u,t)\bar{x}_2(x,u,t) + \Psi_1(x,u,t);$$

$$\dot{\bar{\bar{x}}}_2(x,u,t) = [S(x,u,t) + \Psi_2(x,u,t)]\bar{\bar{x}}_2(x,u,t) + \bar{\bar{x}}_2(x,u,t)[S(x,u,t) +$$

$$+\Psi_2(x,u,t)]^{T} + K(x,u,t);$$

(55)

with boundary conditions

$$x(0)\in g_0;\quad x(T)\in\{x: g_x(x)=0\};\quad T\in\{T: \Omega[x(T),T]=0\};$$

$$\Phi(T,T)=E;\quad \bar{P}(0)=\Phi(T,0)K_0\Phi^T(T,0);\quad \alpha(T)=a_T,$$

$$\bar{x}_2(0)=x_{20};\quad \bar{\bar{x}}_2(0)=K_{20}, \tag{56}$$

It is necessary to select the program $u(t)(u\in U)$ and the initial conditions $x(0)$ according to the conditions

$$\min Y=\min\{\omega[x(T)]+\Psi^T[\bar{z}'(0)]a(0)+\omega_2[x(T),\bar{x}_2(T)]\} \tag{57}$$

in meeting the current restrictions of the type of inequality

$$g(t,x)+\bar{x}_m(x,u,t)\pm 3\sigma_m(x,u,t)\leq 0 \tag{58}$$

This is a determinate problem and it pertains to the problem of optimal control having the limits for control and phase coordinates and can be solved by using well known approximate calculation methods. Thus the algorithm for defining the optimal nominal motion is completely determined.

The realization of the optimal algorithms of controlling the disturbed motion in reentry problems obtained as stated above requires having a high speed digital computer on board the spaceship.

The specific conditions of the spaceship motion in the atmosphere with superorbital velocity completely define the self-sufficiency of the spacecraft control system. Accordingly as the source of information it is possible to use time and accelerations measured in inertial or fixed axes of the spacecraft.

A high speed digital computer on board the spacecragt by statistical processing acceleration measurement results permits to obtain complete information about the spacecraft motion parameters. The necessary information about the nominal motion that is selected beforehand by the above described optimal method is stored on board the spacecraft. The deviations of the actual values of motion parameters from the nominal ones are used for obtaining the necessary input control signal.

R E F E R E N C E S

I. LETOV A.M., "Optimal Control Theory", Proceedings of the 2nd Congress of IFAC, Vol. "Optimal Systems. Statistical Methods". "Nauka" Publishing House, Moscow, 1965 (In Russian).

2. BOGUSLAVSKY I.A., On Statistical Optimal Control of Terminal State. Journal "Automatika i Telemekhanika", No.5, 1966.

3. STRATONOVICH R.L., Recent Development of Dynamic Programming Methods and their Application for Synthesis of Optimal Systems". Proceedings of the 2nd Congress of IFAC, Vol. "Optimal Systems". Statistical Methods. "Nauka" Publishing House, Moscow, 1965 (In Russian).

4. KROTOV V.F., Approximate Synthesis of Optimal Control". Journal "Automatika i Telemekhanika", No.II, 1964.

43.5

ATMOSPHERE RE-ENTRY CONTROL PROBLEM

Okhotsimski D.E., Bukharkina A.P., Golubiev Yu.F.
Institute of Applied Mathematics
Moscow
USSR

This paper deals with the problem of re-entry control at parabolic velocity. The multistep algorithm of the downrange control is described. Some results of the digital computer simulation are given. It is shown that the algorithm provides a sufficiently smooth and accurate regulation process despite large-scale variations in the atmospheric density distribution. This paper develops further the results of previous work.[I-4]

I. Statement of the problem. This paper deals with the problem of re-entry control algorithm at parabolic velocity. It is supposed that the space vehicle has three accelerometers installed on the stabilized platform. By changing the roll angle the vehicle total lift direction changes. This makes it possible to control the motion of the vehicle. The onboard digital computer calculates the needed roll angle values taking into account the measuring data processing results.

The papers[I, 2] contain an example of the downrange control algorithm. The algorithm built in the paper[3] provides lateral-range control as well. The paper[4] contains the algorithm of the re-entry initial conditions calculation and considers the effect of the systematic, instrumental and executive errors.

The above-mentioned control algorithm of the first-dip portion of the re-entry motion (I) provided a rather narrow range of kinematic parameters at the end of the skip-out portion (II) and thus gave the possibility to compensate the deviations using the restricted control potentialities of the second-dip portion (III) (fig. I). The algorithm displayed fortitude to the errors and density distribution uncertainties, and provided a sufficient accuracy in the re-entry corridor of about ± I5 km.

These results cleared the way to the more perfect algorithms. In particular, it was desirable that the roll angle should be a more usable continuous function of time instead of a step-function employed in.[I-4] It was also desirable to take into consideration the actual constraints on the roll control torques and also the calculation time for decision taking. This paper presents a step in this direction.

2. Decision logics. Let us at first neglect the time for the calculation and divide the first-dip portion into equal time intervals. Let us suppose that the control decisions are taken when passing from one time interval to another, and that the instant roll angle jumps are impossible, and that the roll rate $\dot{\gamma}$ is constant during each time interval. The function $\gamma(t)$ will be a piecewise linear function the angular points of which coincide with the interval ends.

Let us assume that the roll rate jump is constrained by the condition

$$|\dot{\gamma}_{i+1} - \dot{\gamma}_i| \leq \Gamma \tag{2.1}$$

where Γ is the increment of the roll angle caused by the control torque during a one time interval. The fulfilment of the condition (2.I) provides obtaining such a piecewise linear function $\gamma(t)$ which may be approximately realized.

During each decision-making the values of the roll rate and, consequently, the roll angle for the nearest time interval are chosen in such a way that, with the roll angles of the rest of the first-dip portion, the desired downrange is provided. Like in[I-4] two integrations of the motion equations forward up to the end of the first-dip portion are carried out. The datathus obtained give the grounds for decision taking.

It appeared reasonable to use different decision logics during the initial part of the motion up to the velocity head maximum and during the rest of the portion. Let us describe some of the variants that have been investigated and give a number of arguments for selection.

In one of simpliest variants we choose $\dot{\gamma}$ for the nearest time interval (t_i, t_{i+1}) to provide the prescribed down-range. Discontinuity at t_{i+1} is admissible, but at t_i the roll angle must be continuous. The repeating of the process gives the roll angle as a piecewise linear function of time. But the permission of a jump at the end of the time interval may result in a saw-tooth function.

It is possible to avoid this unpleasant event by appropriate choice of γ not for one, but for two nearest time intervals (fig. 3). Let the value of γ at t_i join continuously the previous values and at t_{i+2} the beforehand nominated function $\tilde{\gamma}(t)$. Let us choose the value γ at t_{i+1} or the value of γ for the interval (t_i, t_{i+1}) (which is actually the same) to provide the required down-range. The repeating of the process gives a continuous piecewise linear function without saw-teeth.

Fig. 3 shows that if all the conditions for decision-making at point t_i would also be true for point t_{i+1} and if the two-links construction built at t_i would satisfy condition (2.I), in this case the first link of the new two-links construction during the next step of decision-making at t_{i+1} would coincide with the second link of the previous step and the new second link would coincide with the portion of the function $\tilde{\gamma}(t)$. Thus, there would be a two-links passage to the function $\tilde{\gamma}(t)$ by means of the construction obtained at t_i. As a consequence, if we deviated from the function $\tilde{\gamma}(t)$ we would return to it after two time intervals. Also, if $\tilde{\gamma}(t)$ were any admissible polygon the roll angle γ would arrive at $\tilde{\gamma}(t)$ and then coincide with it.

Actually, the atmospheric density distribution is partly unknown and conditions for decision-making vary from one point to another. It is reasonable to modify the function $\tilde{\gamma}(t)$ in accordance with the density variations which are revealed during the flight. But, nevertheless, the algorithm with two-links logics appeared to be effective enough and was used at first on the second part of the first-dip portion.

Within the bounds of the first part of the portion the effectiveness of the roll angle change is small at first and increases with time. This causes large-scale variations of the roll angle and insufficient smoothness of the regulation process. In view of this it appears to be more reasonable on the initial part of the first-dip portion to vary a constant value of the roll angle within a more lengthy period of time which includes several standard time intervals, and to add two links (at its beginning and at its end) for smooth connection with the previous and consequent course of the function $\gamma(t)$ (fig. 4). The varied position of the function $\gamma(t)$ is shown by a dashed line. The closer to the region of the maximum velocity head the shorter is the interval of the constant γ value variation, and the described logics turns smoothly enough into two-links logics.

To successfully overcome the atmospheric density uncertainties it appears reasonable to take the standard time intervals shorter, and to take decisions more often. But the decrease of the time interval length causes the decrease of the regulation smoothness. It was suitable to modify two-links logics in such a way so that each link would include two standard time intervals. One time interval later, at t_{i+1} we have to build a new two-links construction with the same length of each link, but biased by one standard interval further (fig. 5).

It appears also reasonable to modify a little the logics for the initial part of the portion and to change from one to two standard intervals the length of the link connecting the previous roll angle course with the interval of the constant roll angle value, but to take decisions after each standard time interval.

When choosing $\dot{\gamma}$ for the nearest time interval it may occur that the $\dot{\gamma}$ jump will be greater than the permissible one, according to (2.I). In this case the maximum possible value must be chosen.

For the downrange control it is sufficient to vary γ within the range of 0 to π. The whole lift force is directed upwards, if $\gamma=0$ and downwards, if $\gamma=\pi$. If γ has an intermediate value, the vertical component of the total lift force is also intermediate. In particular, if $\gamma=\frac{\pi}{2}$

the vertical lift force component equals zero and the longitudinal motion of the space vehicle is like the motion without a lift force.

During the regulation the roll angle can reach the upper or the lower stop and remain there for some time. To avoid violation of (2.I) when reaching the stop it was assumed a limitation for $\dot{\gamma}$ to provide the stop reaching smooth enough.

The algorithm is able to carry out an analysis of the real density distribution. Knowing the vehicle position and velocity and using the measuring data given by the accelerometers we can calculate for each time the real density ρ and the quantity

$$\xi = \frac{\rho}{\rho_{st}}$$

where ρ_{st} is the density value at the same point according to the standard density distribution. Along the motion, ξ appear to be a function of time the previous course of which is well-known. Let us assume, as in[I-4], that ξ is a rather smooth function of coordinates. In this case $\xi(t)$ along the motion is a rather smooth function of time, and it is possible to extrapolate this function forward for a short period of time. Such an extrapolation appeared to be useful during the integrating of the motion equations forward for decision taking.

The information about the possible $\xi(t)$ course in the nearest future was also used for the appropriate choice of the function $\tilde{\gamma}(t)$ outside the interval of time where the choice was carried out to provide the desired down-range. The choice of $\tilde{\gamma}(t)$ ensured adaptation and created the controllability reserve to compensate the predicted density variations.

The original function $\tilde{\gamma}(t)$ is shown in fig. 6. The portion of the constant value $\gamma = \ell$ at t_1 turns into an inclined line and then at t_2 into a second portion of the constant γ value which was originally taken equal $\frac{\pi}{2}$. The values t_1 and t_2 are chosen beforehand and remain fixed. When varying ℓ to provide down-range the inclined portion of $\tilde{\gamma}(t)$ varies too. It is shown by a dashed line (fig. 6). In the course of time the length of the varied portion becomes shorter, and near t_1 the logics turns into a two-links one or into its modified variant.

The adaptation is carried out by changing the position of the function $\tilde{\gamma}(t)$ after the time t_2. The quantity m (fig. 6) may be calculated according to the formula

$$m = \frac{\pi}{2} + \Delta\tilde{\gamma} \qquad (2.2)$$

where $\Delta\tilde{\gamma}$, as in[I-4], is

$$\Delta\tilde{\gamma} = f(t)[A\dot{\xi} + B\ddot{\xi}] \qquad (2.3)$$

The coefficients A and B are constants found empirically; $\dot{\xi}$ and $\ddot{\xi}$ are the first and the second derivatives of the function $\xi(t)$ calculated for the decision time according to the polynomial which approximately describes the previous course of the function $\xi(t)$; $f(t)$ is the function of time which equals zero before the region of the maximum velocity head, then it increases linearly and, after reaching the prescribed value, it remains constant. The parameters of the function $f(t)$ were defined empirically. A number of reasons associated with the interoduction of such an adaptation type was discussed in papers[I, 2].

3. Taking account of the calculation time. In the previous section of this paper we neglected the calculation time. Let us take it into account. Let us assume that during one standard time interval all the calculations connected with the decision-making were completed. Let us also assume when calculating that only that information may be used which was obtained prior to the beginning of the time interval. In the course of this calculation we obtain information about the current position and velocity and about the course of the function $\xi(t)$, but these data may be employed only during the next decision-making in the next time interval. Thus we have a lag between the end of the last information inflow and the beginning of the decision execution. The lag is one standard time interval long.

Let us change a little the previous decision logics. Let us assume that everything remains true, but when choosing $\dot{\gamma}$ for the nearest time interval the information of the previous time interval cannot be used. To enter the previous algorithm logics we have to calculate all the quantities forward for the one time interval. We obtain the values of the function $\xi(t)$ by extrapolation. Let us calculate coordinates and velocity components by integrating the motion equations for a one time interval using extrapolated values of the function $\xi(t)$. Let us obtain the quantities in the formula for the function $\tilde{\gamma}(t)$ by extrapolation. After all this additional calculations we can enter the previous algorithm.

This approach appears to be very suitable for it gave a simple method to investigate the effect of the time lag using the algorithm without the time lag.

In fact, the measurement information obtained in the course of the calculation may be partly used to deminish the time lag effect. This would be useful for the fortitude of the algorithm.

4. <u>Simulation results.</u> The algorithm performance test was simulated on a digital computer. The components of the aerodynamic acceleration were calculated by integration of the motion equations which imitated the moving space vehicle. The variations of the standard density distribution and the roll angle produced by the control algorithm were inserted into this equation system. The control algorithm itself was realized as a program block for the computer. It received the imitated measurement information, integrated the navigational equations, predicted the $\xi(t)$ course and carried out the adaptation and other calculations which were necessary to take decisions for the roll angle for the nearest time interval.

The density variations were taken as in[2, 3]. These variations imitated the density deviations depending both on the altitude and on the longitudinal range. Their magnitude is apparently larger than really existing. Therefore the algorithm which is able to struggle successfully with them will have a certain reliability reserve.

Some simulation results are presented in fig. 7 and in others. The thick polygon represents the roll angle versus time during the first-dip portion. The initial point of the time is located at the altitude of 150 km. The algorithm function begins when the integral of the acceleration reachesthe prescribed value. At first, when the efficiency is small, the flight is uncontrolled and the roll angle is maintained as a constant which depends on the perigee altitude. When t = 45 sec the algorithm begins to act. The thin line represents the function $\xi(t)$ built up during the flight. The dashed line represents the quantity m (fig. 6) which depends on the $\xi(t)$ course according to the (2.2) and (2.3).

It is seen that for the variant in fig. 7 the $\gamma(t)$ course is at first rather quiet. The tendency towards the density increase is compensated by the decrease of the roll angle and causes an increase of the vertical lift force component which controls the longitudinal movement. The increase of the $\xi(t)$ changes to decrease. Performing the $\xi(t)$ forecast the algorithm increases the roll angle in advance so as to avert the space vehicle from skipping out of the atmosphere earlier than its velocity would be sufficiently braked.

The fast change of the situation regarding the $\xi(t)$ function course requires a fast response. The control torques restrictions and the calculation time lag diminish the response of the system. Therefore the fast change in the $\xi(t)$ course leads the roll angle to the upper stop and leaves it there up to the end of the first-dip portion. Despite the measures taken the time lag of the system causes the skip-out with an excessive velocity and the initial point of the second dip appears to be biased 210 km forwards. Such a deviation may be easily compensated by the second-dip control, for it is deeply within the permissible deviation range.

It should be noted that the variant in fig. 7 was the most difficult of all the atmospheric density variations investigated. This variant was very hard on the system, for the $\xi(t)$ function began a fast decrease. Therefore the controllability reserve appears to be insufficient. The investigation shows that a similar fast change from decrease to increase of the atmospheric density appears to be much easier for the control algorithm.

All other variants investigated have considerably less second-dip initial point deviations than the variant shown in fig. 7. In the cases where the regulation process ended within the stops the deviations were usually no more than some scores of kilometres.

Fig. 8 shows a smoother course of the $\xi(t)$ function when increase changes to decrease. The roll angle reaches the upper stop but leaves it after some time. The accuracy of the second-dip initial point is rather high.

Fig. 9 illustrates a longer stay on the upper stop.

Fig. IO demonstrates the density variation which causes the increase of the density mainly. The rather fast change of the $\xi(t)$ course from decrease to increase leads the roll angle to the lower stop in order to leave the atmosphere as fast as possible. The deviation is negligible.

Fig. II shows the regulation process which takes place within the stops. The deviation is small. It is possible to observe the influence of the $\xi(t)$ course on the adaptation parameter m and on the roll angle.

Fig. I2 presents the variant in which the fast decrease of the density causes the upper stop of the roll angle for some time. The downrange deviation is very small.

The simulation results demonstrate that a reasonable choice of the parameters of the algorithm makes it sufficiently resistant to the density distribution variations. The algorithm provides for the second-dip initial point parameters to be within the domain where the second-dip control ensures precise space vehicle landing in the prescribed region.

References

I. Д.Е.Охоцимский, Г.И. Бельчанский, А.П. Бухаркина, Ю.Ф. Голубев, Н.И. Золотухина, Ю.Н. Иванов. Оптимальное управление при входе в атмосферу. Космические исследования, т. УI, вып. I, I968 г.

2. Д.Е. Охоцимский, А.П. Бухаркина, Ю.Ф. Голубев, Ю.Н. Иванов. Управление продольным движением при входе космического аппарата в атмосферу. Отчет ИПМ АН СССР, I967 г.

3. А.П. Бухаркина, Ю.Ф. Голубев, Д.Е. Охоцимский. Управление пространственным движением при входе космического аппарата в атмосферу. Отчет ИПМ АН СССР, 1968 г.
4. Ю.Ф. Голубев. Определение условий входа в атмосферу и анализ влияния погрешностей. Отчет ИПМ АН СССР, 1968 г.

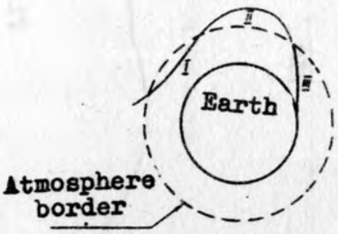


Fig.I Re-entry motion scheme

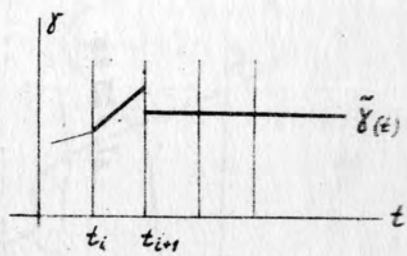


Fig.2

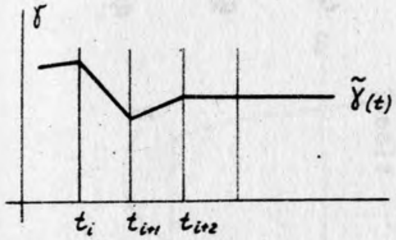


Fig.3

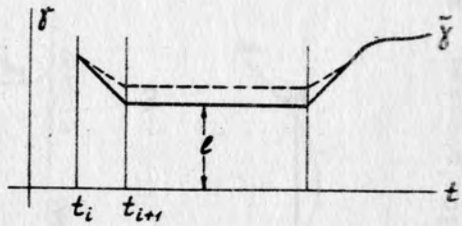


Fig.4

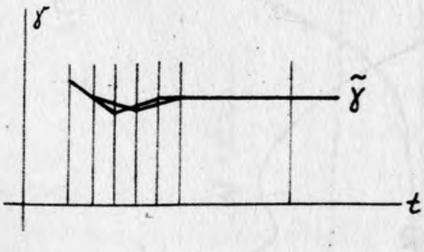


Fig.5

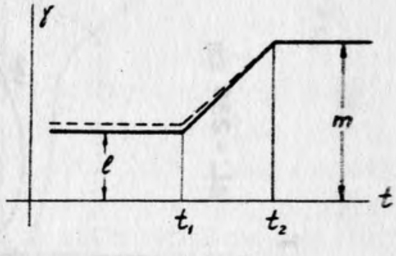


Fig.6

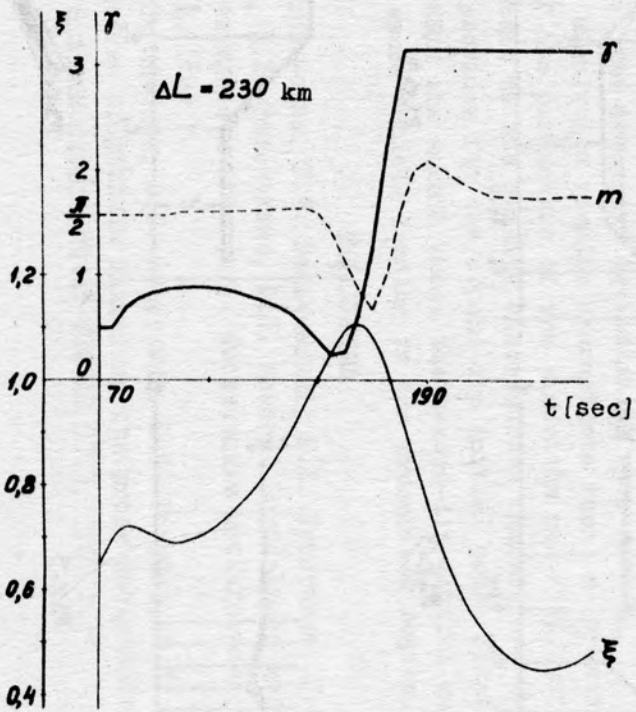

ξ
γ
3
2
π/2
1
0
1,2
1,0
0,8
0,6
0,4
ΔL = 230 km
70
190
t [sec]
γ
m
ξ

Fig.7

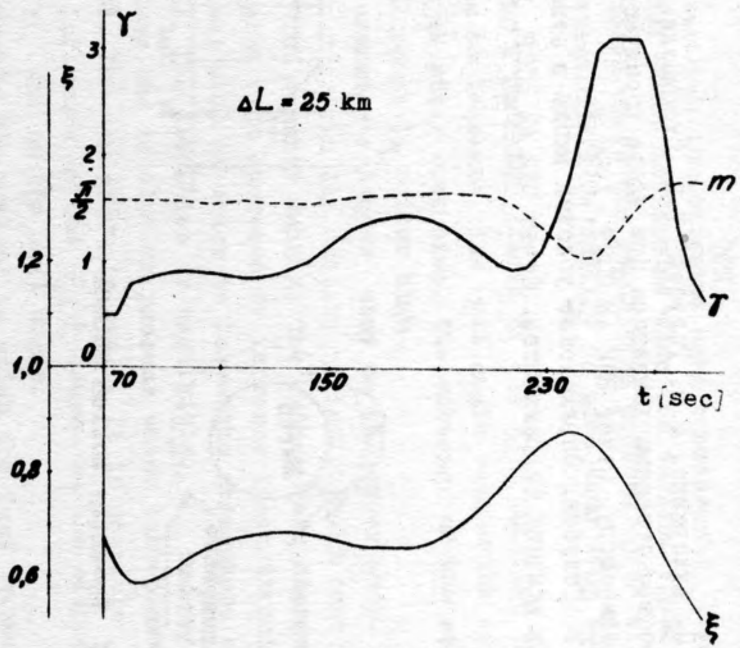

γ
ξ
3
2
π/2
1
0
1,2
1,0
0,8
0,6
ΔL = 25 km
70
150
230
t [sec]
m
γ
ξ

Fig.8

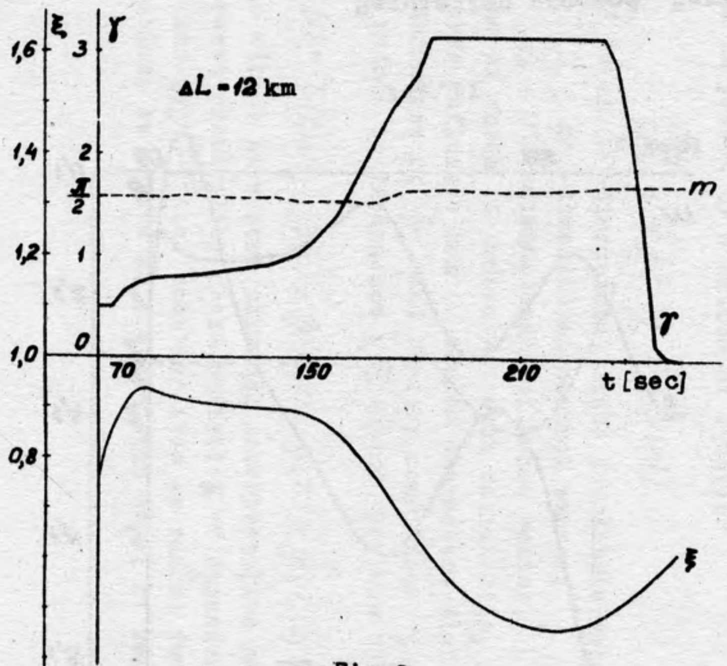


Fig.9

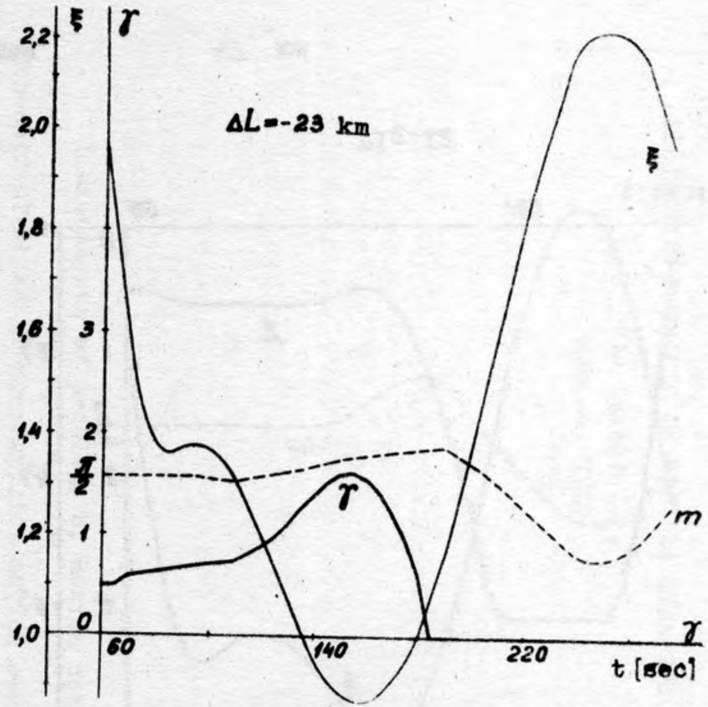


Fig.IO

Regulation process. Perigee 47 km

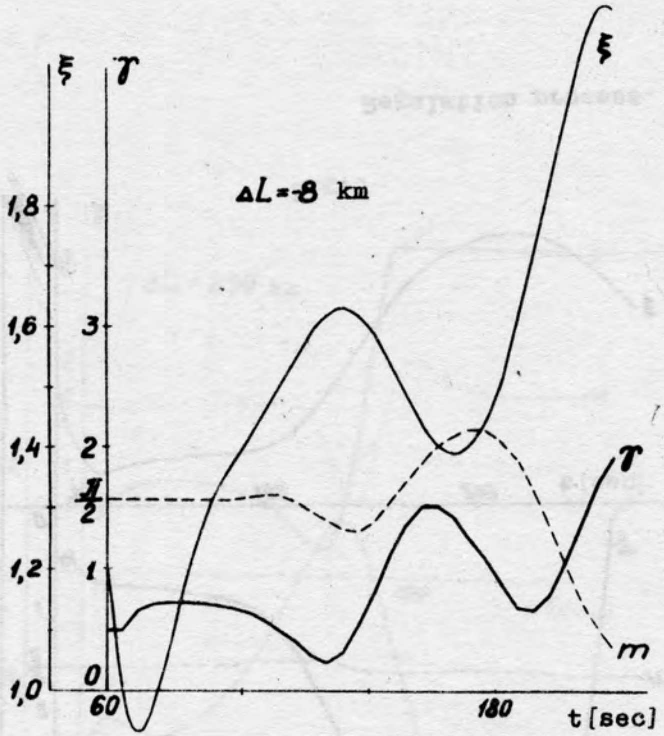


Fig.II

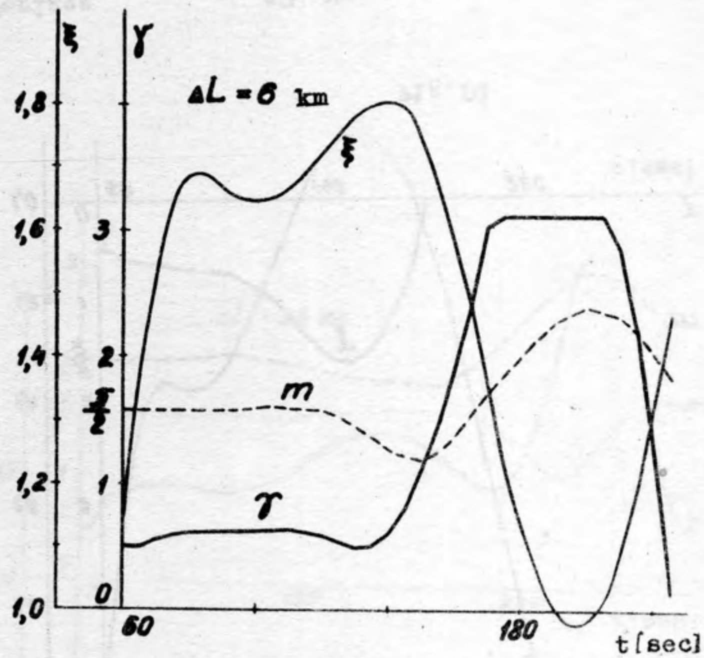


Fig.I2

Regulation process. Perigee 47 km

43.6

STOCHASTIC PROBLEMS OF MISSILE DYNAMICS

Plo tnikov Y.P.
Moscow, USSR

I. Statements of problems of motion control to be met in practice

Mathematical model of many applied motion control problems is preset by a system of equation

$$\frac{dy}{dt} = f(t, y, v, \xi(t))$$

here y is an n -dimensional state coordinate vector;
u - m -dimensional control vector;
ξ - r -dimensional influence vector,
the value of which is unknown with reliability, i.e. in other words influence has random character. Statement about randomness of ξ and bondary conditions $y(0), y(T)$ formaly reduce to dependence $\xi, y(0)$ and $y(T)$ from "chance" ω:

$$\xi(t) = \xi(t,\omega), \quad y(0) = y(0,\omega), \quad y(T) = y(T,\omega).$$

ω can either have some physical interpretation here or not. As we preset probability characteristics of dependence of $\xi(t)$ $y(0)$ and $y(T)$ on ω, i.e. consider them as random functions or variables, we arrive at a stochastic model of real control process,

$$\frac{dy}{dt} = f(t, y(t,\omega), v(t,\omega), \xi(t,\omega)) \qquad (1)$$

Vector function f is supposed to have the properties, ensuring the existence of a solution of the system (1) for the class of controls $v(t,\omega)$ and perturbations $\xi(t,\omega)$ under discussion.[7]

Not all trajectories $y(t,\omega)$ and controls $v(t,\omega)$ possible for (1) are considered on practice, but only those which with a given probability

1. belong to the region

$$g(t, y, v) \leq 0 \qquad (2)$$

where g is a given vector function of t, y and v.

2. begin under condition

$$y(0,\omega) \in \{y : g_H(y, \xi(0)) = 0\} \qquad (3)$$

3. and end, when

$$y(T,\omega) \in \{y : g_K(y, \xi(T)) = 0\} \qquad (3')$$

Performance characteristic of motion to be minimized by admissible $v(t,\omega)$ choice usually represents the expectation of some functional, including integral and terminal parts.

By now the optimization theory of stochastic control systems has not yet reached such a completness, which optimal deterministic system theory already has. This is the particular reason for the fact that the routine practice has worked out its own way of the solution of the above raised problem. This way assumes stage by stage solution of the problem, with its being divided into two basic ones.

The first of them is the problem of optimal motions programming; to be more exact, the problem of individual dynamics of a controlled object, where forces and motion are partially given and missing forces and motion are to be found so that it would offer stipulated optimal properties[2]. In that instance, physical model is described here by a deterministic system of differential equations

$$\frac{dx}{dt} = f(t, x, u, \bar{\xi}(t)) \tag{4}$$

where x means a state vector, u - control vector sought for, being function of time and initial conditions.

At any instant functions x and u are such that

$$\tilde{g}(t, x, u) \leq 0$$

when $t = 0$, $x(0) \in \{x : \bar{g}_H(x) = 0\}$

and when $t = T$, $x(T) \in \{x : \bar{g}_K(x) = 0\}$,

one of coordinates of vector $x(T)$ being optimized.

Mathematical model of program motion (deterministic as a rule) is obtained from (I) replacing $\xi(t,\omega), y(0)$ and $y(T,\omega)$ by average values of $\bar{\xi}(t), x(0)$ and $x(T)$.

The consideration of the first problem belongs to the sphere of interest of scientists and engineers engaged in a particular field of dynamics, which is concerned with the given object.

By its construction the process, treated in the first problem, differs from original one. The reason for this is other initial conditions, as well as difference in values of the right parts of motion equations in any instant of time. The idea of considering the first problem is that, knowing its solution $x(t), u(t)$, we can represent the real solution in the following form:

$$y = x + z, \quad v = u + w, \quad \xi(t,\omega) = \bar{\xi}(t) + \xi^{\circ}(t,\omega)$$

The perturbed motion z, w is defined as

$$\frac{dz}{dt} \equiv \dot{z} = f(t, y, v, \xi(t,\omega)) - f(t, x(t), u(t), \bar{\xi}(t)) =$$
$$= A(t) z + B(t) w + C(t) \xi^{\circ}(t,\omega) + R(t, z, w, \xi^{\circ}) \tag{5}$$

Also

$$g(t, x(t)+z, u(t)+w) =$$

$$z(0,\omega) \in \{z: g_H(x(0)+z, \xi)=0\} = \{z: g^{\circ}_H(z, \xi^{\circ})=0\} \qquad (6)$$

and at the end of the motion with some probability

$$z(T,\omega) \in \{z: g_K(x(T_{np})+z, \xi)=0\} = \{z: g^{\circ}_K(z, \xi^{\circ})=0\}$$

The right part of (5) is an expansion in Teylor series, so matrices A, B and C have the next form

$$A(t) = \frac{\partial f(t,y,v,\xi)}{\partial y} \Big|_{y=x(t),\, v=u(t),\, \xi=\bar{\xi}}$$

$$B(t) = \frac{\partial f(t,y,v,\xi)}{\partial v} \Big|_{y=x(t),\, v=u(t),\, \xi=\bar{\xi}}$$

$$C(t) = \frac{\partial f(t,y,v,\xi)}{\partial \xi} \Big|_{y=x(t),\, v=u(t),\, \xi=\bar{\xi}}$$

A usual linear model of perturbed motion is obtained, if it is possible to neglect the remainder R in (5).

The second basic problem is formulated for system (5). It consists in determination of law W for transformation of information about perturbed motion so that the chosen control $w(t,z)$ should put out this motion to the best advantage, with $w(t,z)$ being able to depend evidently only upon coordinates of process to be measured.

2. Shortcomings of generally used statements of the problems

A Specialist on control who has to point out the method and means for actual realization of program motion, regards it as the prescribed one, which in particular is reflected in the way of presentation of the perturbed motion in the form of (5), where matrices A, B and C are functions of time[I]. They obtain this form only after the program motion construction as $x = x(t)$, $u = u(t)$, with the system of equation (4), which determines these $x(t)$ and $u(t)$, containing no information about future pertubed motion.

This approach to original problem solution that has already become canonical in literature, is connected with the nessecity to simplity equation of motion (I) in order to obtain an answer with our restricted computing and algorithmical means. Along with that the separate consideration of the first and the second problems and original system (I) coarsening introduced with it have as a consequance the unadequcy of its properties to those of our differentiated model (4)-(5). The reason for unadequacy is in particular our neglecting the relation between the problem of program $x(t), u(t)$ choice and the one perturbed motion z, w control. But the relation - which is a bilateral one - takes place irrespective of whether $x(t)$, $u(t)$ is a solution of some variational problem or they have been chosen out of some other consideration.

On the one hand, the interrelation proves in the fact that system (5) takes its specific shape only upon the solutions of system (4), when matrices $\partial f/\partial y$, $\partial f/\partial v$ and $\partial f/\partial \xi$ become the known functions of time, corresponding to specific $y = x(t)$, $v = u(t)$, $\xi = \bar{\xi}(t)$. Therefore all perturbed motion properties, such as controllability, stabilizability and their quantative equivalents depend evidently on the choice of x and u [3].

The program motion acceptability is determined to a great extent by the accuracy, with which a prescind motion can be realized for given perturbed motion control structure[9]. To select at once the program motion extremal in some sence and satisfying the accuracy requirements, the joint consideration of (4) and (5) system is needed. The same thing is necessitated by the well founded separation of subregions of coordinate changing of perturbed and program motion from region defined by inequality (2). In the first place this refers to the region V of u and W control vectors changes[I]. It is necessary to choose the vector so that the set $V-u(t)$ might be substantial enough if we want the problem of control of perturbed motion and its potimization in particular to be have sence.

3. Some new statements of motion control problems

The above mentioned permits to consider it advisable not to fix the program motion control while describing the perturbed motion springing up in its neihbourhood. For these purposes it is necessary to consider the program mltion equation system together with perturbed motion equations, thus reformulating the problems, stated for original system (I) to the following one

$$\frac{dx}{dt} = f(t, x, u, \xi) \tag{7'}$$

$$\frac{dz}{dt} = \frac{\partial f(t, y, v, \xi)}{\partial y} z + \frac{\partial f}{\partial v} w + \frac{\partial f}{\partial \xi} \xi^{\circ}(t, \omega) + R \tag{7"}$$

($\partial f/\partial y$, $\partial f/\partial v$ and $\partial f/\partial \xi$ have been calculated for $y = x$, $v = u$ and $\xi = \xi(t)$) with

$$g(t, x + z, u + w) \leq 0 \tag{8}$$

$$x(0) \in \{x : \bar{g}_H(x) = 0\}, \quad x(T_{np}) \in \{x : \bar{g}_K(x) = 0\} \tag{9}$$

and

$$z(0) \in \{z : g_H(x(0) + z, \xi) = 0\}. \tag{IO}$$

$$z(T) \in \{z : g_K(x(T_{np}) + z, \xi) = 0\} \tag{II}$$

$T - T_{np}$ is a difference of time intervals which perturbed and program motion is exist on, caused by nonsimultaneousness fo end conditions fultilment.

Let us first of all pay attention to those problems, where perturbed motion control $w(t, z)$ have been calculated and such program control $u(t)$ is required to be found which would give to system (I) minimal (or maximal) value of one coordinate of vector $x(T_{np})$ under conditions that (8), (I0) and (II) is kept with the presscribed probability.

A relative problem to that is the following one: to choose the control in the program movement (7') so that maximaize the probebility of the event (II) with the conditions (8) - (IO) kept.

In this wording the problem of program motion control defining is a new one which is not be met either when constructing the program motion or stabilizing the perturbed one. At the first turn an practical prototype of this problem occurs where perturbed motion control $w(t,z)$ owing to the restricted composition of magnitudes measured during the motion, admits the existence side by syde with perturbations, compensated by this control, noncompensated ones also. The practical value of such statement of the problem consists in the reduction of influence of the latter.

Purposes point out above are achieved not only by the choice of program control $u(t)$. It is permittable to consider that $w(t,z)$ is not give completely, but as s structure only, for example,

$$w(t,z) = K(t)\hat{z}$$

where $\hat{z}$ is part of z , that can be measured. To define the control w it is necessary to set a matrix $K(t)$, choice of it and control $u(t)$ permitting to achieve the purposes in this statement of original problem.

If we don't confine to give structure of perturbed motion control thesolution of raised problem (to maximize the probability of event (II) with the rest of restrictions on system (7) being fulfiled). means joint choic of $u(t)$ - control on program trajectory, as well as control syntesis $w(t,z)$ in perturbed motion, that is finding it in terms of time and coordinates measured.

In missile dynamics problems those of them, where necessity of joint program and perturbed systems investigation is obvious (the latter being stochastic one), constitute the balk of problem which may be called stochastic missile dynamics problems.

The way on which we look for the solution of such stochastic problems is rested on the basic lemma stated in[4].

4. Sufficient conditions of absolute minimum for stochastic systems

The starting-point of the given direction can be found in[4]. Results obtained in this paper[x)] are true for a

x) These results constitute a part of investigation, carried out by the auther and V.F.Krotov vast class of stochastic processes $\xi(t,\omega)$. Limits to that class are given below.

Let the right parts of (I) or (7) are

$$\frac{dy(t,\omega)}{dt} = f(t, y(t,\omega), v(t,\omega), \xi(t,\omega)), \quad t \in [0,T]$$

where random vector function $\xi(t,\omega)$, given on probability space $(\Omega, \mathcal{B}, P)$, with the prescribed control $v(t,\omega)$ satisfies the conditions, which define $y(t,\omega)$ as n-dimentional random process[5,6]. In a moment T we know the value of functional, calculated on solutions of system

$$J = M\Big[\int_0^T f^0(t, y(t,\omega), v(t,\omega), \xi(t,\omega))\,dt + F(y(0,\omega), y(T,\omega))\Big] =$$
$$= \int_0^T M f^0(t, y(t,\omega), v(t,\omega), \xi(t,\omega))\,dt + MF(y(0,\omega), y(T,\omega)) \qquad (I2)$$

Functions $F(y(0,\omega), y(T,\omega))$ and $f^0(t, y(t,\omega), v(t,\omega), \xi(t,\omega))$ are P--integrable here on Ω and on $\Omega x[0,T]$ respectively for functions y, v and ξ to be met below.

In any moment of time control $v(t,\omega)$ and vector function $y(t,\omega)$ belong to sets $Q(t,y)$ and $B(t)$ of spaces $R^{\Omega}_{n_1}$ and R^{Ω}_{n} [see (2),(3) or (8)-(II)]. Let D be a set of "pairs" $y(t,\omega)$ and v , satisfying to differential equations and restrictions pointed out^x).

Let's state the problem: "from a set of pairs" $y(t,\omega), v$ we have to find such one for which the functional J would have the least value (if such a pair is absent in the class D, it is necessary to look for a minimizing sequence $(y_n(t,\omega), v_n) \in D$ upon which $J_n \to \exists \inf_D J > C$).

Lemma mentioned given the opportunity to replace the problem of functional minimization on set D by the same

^x) I don't indicate here functions of what "parameters" the control v is because in each case this defines its own class D . At this general case it is important only to pint out that v belong to given set for $t \in [0,T]$.

problem on more vast set E of independent pair of vector functions $(y(t,\omega), v)$, satisfying to all conditions raised above, besides equation (I).

Let's put to consideration a functional

$$\varphi[t, y(\omega)], \quad y(\omega) \in R_n^{\Omega}, \quad t \in [0,T],$$

differentiated with respect to t and also having restricted continious Gateau - derivalive $D_y\varphi$ for arbitrary randow variable $y(\omega)$ from R_n^{Ω}.[8]

Let

$$R[t, y(\omega), v] = \frac{\partial \varphi[t, y(\omega)]}{\partial t} + D_y\varphi[t, y(\omega)] \times f(t, y(\omega), v, \xi(t,\omega)) - M f^0(t, y(\omega), v, \xi(t,\omega)) \qquad (I3)$$

and

$$\Phi[y(0,\omega), y(T,\omega)] = MF(y(0,\omega), y(T,\omega)) + \varphi[T, y(T,\omega)] - \varphi[0, y(0,\omega)] \qquad (I4)$$

<u>Theorem A.</u> Let us have the "Pair" of functions $\bar{y}(t,\omega), \bar{v}$, then in order that this "pair" would minimize the functional J on D the existence of such functional $\varphi[t, y(\omega)]$ is sufficient with properties mentioned that

$$1.\ R[t, \bar{y}(t,\omega), \bar{v}] = \sup_{\substack{y(\omega) \in B(t) \\ v \in Q(t,y)}} R[t, y(\omega), v] = \mu(t) \qquad (I5)$$

$$2.\ \Phi[\bar{y}(0,\omega), \bar{y}(T,\omega)] = \inf_{y(0,\omega), y(T,\omega)} \Phi[y(0,\omega), y(T,\omega)], \qquad (I6)$$

$y(0,\omega)$ and $y(T,\omega)$ belonging to the sets (3)$^{x)}$. When absolute minimal does not exist on D theorem terms I and 2, defining a minimizing sequence $y_n(t,\omega), v_n$, coincide with cited ones, if the sign of equality is replaced by the symbol: "$\longrightarrow$ as $n \longrightarrow \infty$"

To prove the theorem let us define on set E functional

$$L = MF(y(0,\omega), y(T,\omega)) + \varphi[T, y(T,\omega)] - \varphi[0, y(0,\omega)] - \int_0^T \left\{ \frac{\partial \varphi[t, y(\omega)]}{\partial t} + D_y\varphi[t, y(\omega)] \times f(t, y(\omega), v, \xi(t,\omega)) - M f^0(t, y(\omega), v, \xi) \right\} dt \qquad (I7)$$

- - - - - - - - - - - - - - - - - - -

x) The theorem is also true in that case, when and are elements of some Banach spaces.

This functional is a continuation on set E of functional J, defined on D. Indeed, on D owing to (I) and that for (I),

$$d\varphi[t,y(t,\omega)] = \left[\frac{d\varphi[t,y(t,\omega)]}{dt} + D_y\varphi[t,y(t,\omega)]\times f(t,y(t,\omega),v,\xi(t,\omega))\right]dt,$$

$$L = \Phi[y(0,\omega),y(T,\omega)] - \int_T [d\varphi - Mf^0(t,y(t,\omega),v,\xi(t,\omega))]\,dt = J$$

If functional $\varphi[t,y(\omega)]$ and "pair" of $\bar{y}(t,\omega)$ and $\bar{v}$, satistying to theorem terms I and 2, exists then from (I5)-(I7) it results that this pair minimizes L on E and by lemma the functional J on D.[4]

Let us pay attention to a case, when $y(t,\omega)$ and $v\in L_2(\Omega,B,P)$ that is when admissible control and corresponding to it (in D) vector $y(t,\omega)$ are random vectors (for any $t\in[0,T]$ the coordinate sequares of which are P--integrable ~~with respect to measure~~[5]. Let $\xi(t,\omega)\in L_2(\Omega,B,P)$ for $t\in[0,T]$ also. Then under the same conditions for the right parts of system (I), which provided the existence of solution $y(t,\omega)$ of (I), $f(t,y(t,\omega),v(t,\omega),\xi(t,\omega))\in L_2(\Omega,B,P)$ for $t\in[0,T]$[6,7]. But for a linear functional in $L_2(\Omega,B,P)$ is an integral of measure $P(\cdot)$, so

$$D_y\varphi[t,y(\omega)]\times f(t,y(\omega),v,\xi(t,\omega)) =$$
$$= \int_\Omega \varphi^y(t,y(\omega),\omega)\, f(t,y(\omega),v,\xi(t,\omega))\, P(d\omega)$$

Here row vector $\varphi^y(t,y(\omega),\omega)\in L_2(\Omega,B,P)$.

Theorem A terms, putting the constraints on φ and f at minimal $(\bar{y}(t,\omega),\bar{v})$ give an arbitrariness broad enough in setting of functional $\varphi[t,y(\omega)]$ out of it. This allows to choose the most fitted algorythm for solution of problem.

Let us dwell on two of them. We arrive at the first algorythm it we demand from $R[t,y(\omega),v(\omega)]$ to satisfy identically to some conditions in the region $[0,T]\times B$ this is an analog of Hamilton-Iacobi-Bellman formalism for deterministic case[4].

If we confine ourselves by carrying out the theorem terms on the mininal sought for only, we shall arrive at an analog of Lagrange formalism. Similar to that case we reduce the problem to a boundary problem for ordinary differential (but stochastic abready) equations.

5. An analog of Hamilton-Iacobi-Bellman formalism. Systems, linear in state coordinate

Let here and below the region of change for $y(t,\omega)$ when $t\in(0,T]$ be an open one and

$$\mathcal{P}[t,y(\omega)] = \sup_{v\in Q(t,y)} R[t,y(\omega),v] = \sup_{v\in Q}\Big[\frac{\partial\varphi[t,y(\omega)]}{\partial t} + \int_\Omega\big[\varphi^y(t,y(\omega),\omega)f(t,y(\omega),v,\xi(t,\omega)) - f^0(t,y(\omega),v,\xi(t,\omega))\big]P(d\omega)\Big] =$$
$$= \frac{\partial\varphi[t,y(\omega)]}{dt} + \sup_{v\in Q}\int_\Omega\big[\varphi^y f - f^0\big]P(d\omega) \tag{18}$$

We shall take $\varphi[t,y(\omega)]$ so that $\mathcal{P}$ does not depend on $y(\omega)$, that is

$$\mathcal{P}[t,y(\omega)] = \frac{\partial\varphi}{\partial t} + \sup_{v\in Q}\int_\Omega\big[\varphi^y f - f^0\big]P(d\omega) = C(t).$$

$C(t)$ is a function of time. Then $\mathcal{P}[t,y(\omega)] = \mu(t)$ for any $y(\omega)$.

If for $v=\bar v[t,y(\omega)]$ $R[t,y(\omega),\bar v]$ has a supremum in point $(t,y(\omega))$ that is

$$R[t,y(\omega),\bar v] = \mathcal{P}[t,y(\omega)], \tag{19}$$

then solution of system (I), $y(t,\omega)$ together with v belong to D, satisfying term I of theorem A. For fulfilment of the second term it is sufficient to demand, that for $t=T$ does not depend on $y(T,\omega)$, that is

$$\Phi[\cdot,y(T,\omega)] = \mathrm{const} \tag{20}$$

As an example for application of above mentioned we can take the systems, linear in state coordinats.

Let right parts of (I) have the form

$$f = A(t,\omega)y + h(t,v,\omega) \tag{21}$$

and

$$f^0 = a^0(t,\omega)y + h^0(t,v,\omega),\quad F=0 \text{ here} \tag{22}$$

A is an matrix here, a^c and h are vectors, h^0 is a scalar function.

We shall find the minimum of functional, putting the boundaries of control change, depending on t only.

For our functional J R have the form $R[t, y(\omega), v] = \partial\varphi/\partial t + \int_\Omega \{[\varphi^y A - a^c]y + \varphi^y h(t,v,\omega) - h^c(t,v,\omega)\} P(d\omega)$

and $\mathcal{P}[t, y(\omega)] = \frac{\partial\varphi}{\partial t} + \int_\Omega [\varphi^y A - a^c] y(\omega) P(d\omega) + \mathcal{H}(t)$,

where

$$\mathcal{H}(t) = \sup_{v \in Q(t)} \int_\Omega [\varphi^y h(t,v,\omega) - h^c(t,v,\omega)] P(d\omega)$$

To satisfy the system (I) it is necessary to choose functional $\varphi[t, y(\omega)]$ so that functional would not depend on $y(\omega)$.

$$\text{Let } \varphi[t, y(\omega)] = \int_\Omega \psi(t,\omega) y(\omega) P(d\omega) \tag{24}$$

Then
$$\mathcal{P}[t, y(\omega)] = \int_\Omega \left[\frac{d\psi(t,\omega)}{dt} + \psi A - a^c\right] y(\omega) P(d\omega) + \mathcal{H}(t)$$

$$\mathcal{H}(t) = \sup_v \int_\Omega [\psi(t,\omega) h(t,v,\omega) - h^c(t,v,\omega)] P(d\omega)$$

Given row vector function $\psi(t,\omega)$ by the system

$$\frac{d\psi(t,\omega)}{dt} + \psi A = a_0 \quad \text{a.e.,} \tag{25}$$

$$\psi(T,\omega) = 0 \quad \text{a.e.,} \tag{26}$$

we shall see, that $\mathcal{P}[t, y(\omega)] = \mathcal{H}(t)$ does not depend on $y(\omega)$, and $\varphi[T, y(\omega)] = 0$ ($\varphi[0, y(0,\omega)]$ = const here, for $y(0,\omega)$ has been taken as fixed one). Thus the two terms, put upon the functional in Hamilton algorythm are fulfiled. In this case the choice of functional φ was reduced to Cauchy problem for system of linear differential equations.

6. An analog of Lagrange formalism

Let us suppose that functional $\varphi[t, y(\omega)]$ has the second Gateau-derivative and the second mixed derivative $D_y \varphi_t[t, y(\omega)]$ is continuous. We shall look for this functiona together with the minimal $\bar{y}(t,\omega), \bar{v}$ from condition of maximum of R over y and v on this minimal. The R and Φ over $y(\omega)$ stationarity condition in $(\bar{y}(t,\omega), \bar{v})$ point can be written as equality to zero of linear part of R's increment for some $y(t,\omega) = \bar{y}(t,\omega) + h(t,\omega)$, with $h(t,\omega) \in L_2$ and sufficiently small.

Namely,

$$D_y\varphi_t[t,\bar{y}(t,\omega)]\times h(t,\omega)+\int_\Omega \frac{\partial}{\partial y}\varphi^y(t,\bar{y}(t,\omega),\omega)h(t,\omega)f(t,y(t,\omega),v,\xi)P(d\omega)+$$

$$+\int_\Omega\Big[-\frac{\partial f^0}{\partial y}h(t,\omega)+\varphi^y(t,\bar{y}(t,\omega),\omega)\frac{\partial}{\partial y}f(t,\bar{y}(t,\omega),v,\xi(t,\omega))h(t,\omega)\Big]P(d\omega)=0$$

If we designate $\varphi^y[t,\bar{y}(t,\omega)]=\Psi(t,\omega)$ the first two terms of above equality will be $\int_\Omega \frac{d\Psi(t,\omega)}{dt}h(t,\omega)P(d\omega)$ because, due to the second mixed derivative continuouness, $D_y\varphi_t=\frac{\partial}{\partial t}D_y\varphi$. Thus this equality can be rewritten as

$$\int_\Omega\Big[\frac{d\Psi(t,\omega)}{dt}+H_y(t,\bar{y}(t,\omega),v,\xi(t,\omega))\Big]h(t,\omega)P(d\omega)=0$$

where $H(t,y(\omega),v,\xi)=\Psi(t,\omega)f(t,y(\omega),v,\xi)-f^0$

Owing to arbitrariness of $h(t,\omega)$ the condition of R over $y(\omega)$ stationarity is equivalent to

$$\frac{d\Psi(t,\omega)}{dt}+H_y(t,\bar{y}(t,\omega),v,\xi(t,\omega))=0 \quad a.e \tag{27}$$

with such end condition (condition of Φ over $y(T,\omega)$ stationarity for now $y(0,\omega)$ is fixed)

$$\Psi(T,\omega)=-\frac{\partial}{\partial y}F(y(\omega))\Big|_{t=T} \quad a.e \tag{28}$$

For $\varphi_t[t,\bar{y}(t,\omega)]$ is a number, when $t\in[0,T]$, R over v supremum, the condition defining optimal v, takes the form of

$$\mathcal{H}[t,\bar{y}(t,\omega),\bar{v}]=\int_\Omega\big[\varphi^y(t,\bar{y}(t,\omega),\omega)f(t,\bar{y},\bar{v},\xi)-f(t,\bar{y},\bar{v},\xi)\big]P(d\omega)=$$

$$=\int_\Omega\big[\Psi(t,\omega)f(t,\bar{y},\bar{v},\xi)-f^0\big]P(d\omega)= \tag{29}$$

$$=\sup_{v\in Q}\int_\Omega\big[\Psi(t,\omega)f(t,\bar{y}(t,\omega),v,\xi)-f^0(t,\bar{y},v,\xi)\big]P(d\omega)=\sup_{v\in Q}\mathcal{H}[t,\bar{y},v]$$

In such a case when for $y(T,\omega)$ end conditions are given, say

$$\int_\Omega G(y(T,\omega))P(d\omega)=C \tag{30}$$

the condition of stationarity of $MF(y(T,\omega))+\varphi[T,y(T,\omega)]$ takes the form of

$$\Psi(T,\omega)=-\Big[\frac{\partial}{\partial y}F(y(\omega))+\lambda\frac{\partial}{\partial y}G(y(\omega))\Big]_{t=T} \tag{28'}$$

λ is an isoperimetric constant to be defined by equality (30).

If the functional $\varphi[t,y(\omega)]$ can be preset in neighbourhood of extremal $(\bar{y}(t,\omega), \bar{v})$ so that the sufficient condition of $R[t, y(\omega), v]$'s maximum for $t \in [0,T]$ occurs on extremal itself, the given extremal owing to theorem A is the absolute minimal. Now theorem A may be reformulated so:

Theorem B. Let the aggregate of vector functions $\bar{y}(t,\omega)$, $\bar{v}$, $\psi(t,\omega)$ be the result of systems (I), (27) and (29) solution. In order that extremal $\bar{y}(t,\omega)$, $\bar{v}$ would be the absolute minimal the existence of such functional $\varphi[t, y(\omega)]$ is sufficient with mentioned properties that

I. $\varphi^y(t, \bar{y}(t,\omega), \omega) = \psi(t,\omega)$

2. $R[t, \bar{y}(t,\omega), \bar{v}] = \sup_{y,v} R[t, y(\omega), v] = \mu(t), \; t \in [0,T]$

3. $\Phi[\bar{y}(T,\omega)] = \inf_{y(T,\omega)} \Phi[y(T,\omega)]$

7. An linear problem

As an example let us consider for system (I) with the right parts (2I) the problem of functional (22) minimization with the initial $y(0,\omega)$ and end vector $y(T,\omega)$ fixed.

Let us use the theorem B.

Solving with given end conditions the system (27) together with (I) and equation

$$\int_\Omega [\psi(t,\omega) h(t,\bar{v},\omega) - h^\circ(t,\bar{v},\omega)] P(d\omega) = \sup_{v \in Q} \int_\Omega [\psi h(t,v,\omega) - h^\circ(t,v,\omega)] P(d\omega) \tag{3I}$$

we shall obtain the extremal $\bar{y}(t,\omega)$, $\bar{v}$, $\psi(t,\omega)$.

Let us see that this extremal is the absolute minimal of functional J.

Let $\varphi[t, y(\omega)] = \int_\Omega \psi(t,\omega) y(\omega) P(d\omega)$ then term I of theorem B is fulfiled. In this case owing to (25) and (3I)

$$R[t, \bar{y}(t,\omega), \bar{v}] = \int_\Omega \Big[\Big(\frac{d\psi}{dt} + \psi A - a^\circ\Big)\bar{y}(t,\omega) - \psi h(t,\bar{v},\omega) - h^\circ(t,\bar{v},\omega)\Big] P(d\omega) =$$

$$= \int_\Omega [\psi h(t,\bar{v},\omega) - h^\circ(t,\bar{v},\omega)] P(d\omega) = \sup_{y,v} R[t, y(\omega), v]$$

i.e. the second term of theorem B is also fulfiled.

8. About possibility to solve of first and second statements of original problem

We shall use the results of section 7 for solution of problem which can be considered as equivalent of the first or second statement of original problem (see section 3)

Let us preset the process over $[0,T]$ by the system

$$\begin{aligned} \dot{x} &= f(t,x,u) \\ \dot{z} &= Q(t,x,u)\,z + R(t,x,u)\,\xi(t,\omega) \end{aligned} \qquad (32)$$

Let $x(0)=x_0$ and some coordinates of vector $x(T)$ be fixed.

The random influence $z(0,\omega)$ and $\xi_p(t,\omega)$ is given by its probability distributions. The performance characteristic of the process is

$$P\{\,|a(x(T))\,z(T,\omega)| < C\,\} \qquad (33)$$

We shall maximize this probability by the choice of $u(t)$. Let us notice, that $P\{|a\,z(T,\omega)|<C\} = -\int_\Omega F(a\,z(T,\omega))\,P(d\omega)$, where $-F(\rho)$ is the characteristic functions of $[-C,+C]$ interval. In usual sence this function is not differentiable, so we shall consider its derivative as a limit of approximate functions derivatives $\frac{dF_k(\rho)}{d\rho}$, as $K\to\infty$. Therefore it is necessary to consider all relations includinf the derivative of this function as limitary as $K\to\infty$

Theorem B allow to look for optimal among those which minimize the integral

$$\int_\Omega \{\psi(t,\omega)\,f + \mu(t,\omega)\,[Q\,z(t,\omega) + R\,\xi(t,\omega)]\}\,P(d\omega)$$

with condition that row-vector functions ψ and μ, satisfy the equations

$$\dot{\psi}_j = -[\psi f_{x_j} + \mu(Q_{x_j} z + R_{x_j}\xi)] \quad j=1,\dots,n \qquad (34)$$

$\psi_j(T,\omega) = -\frac{dF}{d\rho}\left(\frac{\partial a}{\partial x_j}\,z(T,\omega)\right)$, if coordinate $x_j(T)$ is not fixed,

$$\dot{\mu} = -\mu\,Q(t,x,u)\,, \quad \mu(T,\omega) = -\frac{dF}{d\rho}\,a \qquad (35)$$

It follows from (35) that $\mu(t,\omega) = \frac{dF}{d\rho}\,\bar{\mu}(t)$, ~~where~~

Where

$$\dot{\bar{\mu}} = -\bar{\mu}\,Q(t,x,u)\,, \quad \bar{\mu}(T) = -a \qquad (35')$$

Therefore the last integral is transformed into

$$\int_\Omega \left\{ \Psi f + \frac{dF}{dp} \bar{\mu} \left(QZ + R\xi(t,\omega) \right) \right\} P(d\omega) = \int_\Omega \Psi(t,\omega)\, P(d\omega) \cdot f +$$
$$+ \bar{\mu} \left\{ Q \int_\Omega \frac{dF}{dp} Z(t,\omega)\, P(d\omega) + R \int_\Omega \frac{dF}{dp} \xi(t,\omega)\, P(d\omega) \right\} =$$
$$= \bar{\Psi}(t) f + \bar{\mu} \left(Q \bar{Z}_c(t) + R \bar{\xi}_c(t) \right) \tag{36}$$

Here $\bar{\Psi}(t) = \int_\Omega \Psi(t,\omega)\, P(d\omega)$; $\bar{Z}_c(t) = \int_\Omega \frac{dF}{dp} Z(t,\omega) P(d\omega) =$

$= \gamma(c) M Z\big|_{aZ(T,\omega)=c} - \gamma(-c) M Z\big|_{aZ(T,\omega)=-c}$, $\bar{\xi}_c(t) = \gamma(c) M \xi\big|_{aZ(T,\omega)=c} - \gamma(-c) M \xi\big|_{aZ(T,\omega)=-c}$

where $\gamma(c)$ and $\gamma(-C)$ is the values of probability density of random variable $\chi = aZ(T,\omega)$, when $\chi = aZ(T,\omega) = c$ and $\chi = -C$
It follows from (32) that

$$\dot{\bar{\Psi}}_j = -\left[\bar{\Psi} f_{x_j} + \bar{\mu} \left(Q_{x_j} \bar{Z}_c + R_{x_j} \bar{\xi}_c(t) \right) \right] \tag{37}$$

$j = 1,2...n$

$\bar{\Psi}_j(T) = -\frac{\partial a}{\partial x_j} \bar{Z}_c(T)$, if $x_j(T)$ is not fixed, and from equations of the process

$$\dot{\bar{Z}}_c = Q \bar{Z}_c + R \bar{\xi}_c(t) \tag{38}$$

Thus we have had the possibility to calculate the components of Hamiltionian by integrating of deterministic system of equations (35'), (37) and (38) (boundary conditions for them are not stochastical also). The difficulty of its integration is that we don't know the conditional expectation $\bar{Z}_c(0)$ and $\bar{\xi}_c(t)$ from right parts of (37) and (38) due to their values depend on the choice of control $u(t)$ over all interval of motion.

Now we shall consider one of those cases when this difficulty can be overcome. Let $\xi(t,\omega) = \xi(\omega)$ and $Z(0,\omega)$ are centred random vectors (this case included such a class of perturbations met on practice, when it is possible to consider them depending on some finite set of parameters, ie. random variables) with the controur lines of joint distribution of $Z(0)$ and ξ being hyperspheres. Then due to linearity of system for $Z, \bar{Z}$ and $\bar{\mu}$ we have $\bar{\mu}(t) = -a A(t)$, $\bar{Z}_c(t) = B(t) \bar{Z}_c(0) + S(t) \bar{\xi}_c$

Here $A(t)$ and $B(t)$ are matrices of fundamental systems of solutions for (35') and (36) respectively, with $A(T)=B(0)=E$ (identity matrix). $S(t)$ is a matrix of partical solutions for (38) when initial conditions are zero ones, S^K, K-th column of S, is the solution that correspond the influence of K-th component of vector ξ. Now owing to distribution properties mentioned

$$\bar{Z}_c(0)=2\gamma(c)MZ(0)\Big|_{x=c}=\frac{2\gamma(c)aB(T)}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}}$$

$$\bar{\xi}_c=\frac{2\gamma(c)aS(T)}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}}$$

Due to it the Hamiltonian (35) is equal

$$\bar{\Psi}f-\sum_{i=1}^{n}\frac{aA(t)QB^j(t)2\gamma(c)aB^j(T)}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}}-\sum_{K=1}^{m}aA(t)\frac{QS^K(t)2\gamma aS^K(T)+R^K2\gamma aS^K(T)}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}}=$$

$$=\bar{\Psi}f-\sum_{i=1}^{n}\frac{2\gamma(c)\,aB^j(T)}{\sqrt{\cdots}}aA(t)QB^j(t)-\sum_{K=1}^{m}\frac{2\gamma(c)aS^K(T)}{\sqrt{\cdots}}\left[QS^K(t)+R^K\right]$$

If we set

$$\dot{\mu}^j=-\mu^jQ(t,x,u),\ j=1,\dots,n \qquad \mu^j(T)=\frac{-2\gamma(c)\,B^j(T)a}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}}$$

$$\dot{B}^j=Q(t,x,u)B^j\ ,\ j=1,\dots,n \qquad B^j(0)=E^j \tag{39}$$

$$\dot{S}^K=Q(t,x,u)S^K+R^K(t,x,u)\ ,\ S^K(0)=0 \tag{40}$$

$$K=1,2,\dots m \tag{41}$$

$$\dot{\gamma}^K=-\gamma^KQ(t,x,u),\quad \gamma^K(T)=\frac{-2\gamma(c)aS^K(T)a}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}} \quad K=1,\dots,m \tag{42}$$

$$\dot{\bar{\Psi}}_j=-\left[\bar{\Psi}f_{x_j}+\sum_{i=1}^{n}\mu^jQ_{x_j}B^j+\sum_{K=1}^{m}\gamma^K\left[Q_{x_j}S^K+R^K_{x_j}\right]\right]_{j=1,\dots,n}$$

$$\bar{\Psi}_j(T)=\frac{-2\gamma(c)\frac{\partial a}{\partial x_j}\left[B(T)B'(T)a'+S(T)S'(T)a'\right]}{\sqrt{aB(T)B'(T)a'+aS(T)S'(T)a'}}\ ,\quad if\ x_j(T) \tag{43}$$

is not fixed, it follows the final expression for (35)

$$\mathcal{H}=\bar{\Psi}f+\sum_{i=1}^{n}\mu^jQB^j+\sum_{K=1}^{m}\gamma^K\left[QS^K+R^K\right] \tag{35'}$$

with optimal control being maximized this Hamiltonian for each fixed t.[10]

Now we restrict ourselves to the solution of such applied problems mathematical model of which correspond the two first formulations of original one.

We reduce them to the solution of some boundary problem for the systems (39) - (43), (35').

References

I. Летов А.М. (Letov А.М.) Теория оптимальных систем. Труды П Когресса ИФАС . Изд. Наука, 1965.

2. Летов А.М. (Letov А.М.) Аналитическое конструирование регуляторов. Автоматика и телемеханика, т.XXI, № 4,5,6 ; т. XXП, № 4; т.XXШ, № II

3. Красовский Н.Н. (Krasovsky N.N.). Теория управления движением. Наука, 1968.

4. Кротов В.Ф. (Krotov V.F.). Докторская диссертация, 1963.

5. Колмогоров А.Н. (Kolmogorov A.N). Основные понятия теории вероятностей. ОНТИ, 1936.

6. Гихман И.И. (Gichman I.I.), Скороход А.В. (Scorohod A.V.) Введение в теорию случайных процессов. Изд. "Наука", 1965.

7. Зубов В.И. (Zoobov V.I.). Теория оптимального управления. "Судостроение", 1966.

8. Вайнберг М.М. (Vienberh M.M.). Вариационные методы исследования нелинейных операторов, ГИТТЛ, 1956.

9. Пономарёв В.М. (Ponomarev V.M.). Теория управления движением космических аппаратов. Изд. "Наука", 1965.

I0.Кожевников Ю.В. (Kogevnikov J.B.). ПММ т. 30 , № 4, 1966

321

43.7

STATISTICAL SYNTHESIS OF OPTIMAL PULSE CONTROL SYSTEMS WITH REGARD TO SYSTEM'S STRUCTURE CONSTRAINTS

A.Ya. Andrienko

Institute of Automation and Telemechanics

(Engineering Cybernetics)

Moscow

U S S R

Introduction

In development of automatic closed-loop systems with a digital control computer the designers try sometimes to reproduce control laws chosen earlier for continuous systems. These laws usually foresee the formation of signals action over instantaneous derivative of the controlled variable. As a result the requirements to the speed of a digital computer become more strict and the efficiency of use of computing technique is less, especially in cases when the controlled variables are measured with random errors. At the same time digital computers enable one to realize specifically discrete algorithms of control formation on the basis of analysis of case history of the control process. Apart from this it becomes possible to additionally improve quality of control at the expense of aptimization of time sequence of quantization intervals of a pulse system.

The requirements to the characteristics of a digital control computer as well as reliability of the whole control system are specified, to a considerable extent, by the system's structure complexity realized with the aid of the digital computer. Therefore, it is expedient to synthesize optimal systems with taking into account the constraints imposed onto the structure.

The paper discusses, in connection with the terminal control systems, the statistical methods of synthesis of pulse systems with taking into account the constraints of some basic kinds.

1. Synthesis of a Pulse Control System with Memory Capacity Constraint of Control Device

A controlled plant is considered whose output variable is measured at discrete time intervals i . The equations describing the plant are considered to be known.

$$x_{\nu(i+1)} = F_{\nu(i+1)}(\bar{x}_i, \bar{V}, u_i)$$
$$(\nu = 1, 2, \dots, N; \ i = 0, 1, \dots, I), \tag{1}$$

where $\bar{x}_i = (x_{1i}, x_{2i}, \dots, x_{Ni})$ - vector of plant variables at the i -th time moment;

$\bar{V} = (V_1, V_2, \dots, V_R)$ - random vector of disturbances affecting the plant;

u_i - value of controlled variable.

The value of y_i , of the plant output variable $x_i = x_{1i}$, measured with random error f_i , is fed to the input of the control device. In this case

$$y_i = x_i + f_i \qquad (i = 1, 2, \dots, I). \tag{2}$$

The control should satisfy the inequality

$$|u_i| \le U_i \qquad (i = 0, 1, \dots, I), \tag{3}$$

where U_i - limit permissible value of the control u_i .

There are assumed to be set up the a priori densities $P(\bar{V})$, $P(\bar{x}_0)$, $P(f_i)$ of random vectors $\bar{V}$, $\bar{x}_0$ and of errors values of the measurement f_i .

The following risk function is assumed to be a criterion of optimality

$$W = M[w(x_{I+1})]. \tag{4}$$

The problem consists in defining the operation algorithm of the system's control device when a minimal value of risk is achieved; here the formation of the next control should be carried out with the use of limited volume of information on the state of control process so that control is the function of the form

$$u_i = u_i(\bar{y}_{ji}, \bar{u}_{j(i-1)}) \qquad (i = 0, 1, \dots, I; \ j = i - n), \tag{5}$$

where

$$y_{ji}=(y_j, y_{j+1}, \dots, y_i),$$
$$u_{ji}=(u_j, u_{j+1}, \dots, u_i).$$

The memory capacity limitation is realized by setting up number n which characterizes the value of observation time interval.

The solution of the given problem is made on the basis of the theory of stochastic solutions and dynamic programming.

In formation of control at the i-th time moments there are regarded to be known, firstly, the control functions $u_s=u_s(\bar{y}_{ts}, \bar{u}_{t(s-1)})$ $(s=i+1, i+2, \dots, I;\ t=s-n)$ at all subsequent time moments; secondly, control values u_s $(s=j, j+1, \dots, i-1)$ and measured variables y_s $(s=j, j+1, \dots, i)$ at preceeding time moments; and, thirdly, control functions $u_s=u_s(\bar{y}_{ts}, \bar{u}_{t(s-1)})$ $(s=1,2, \dots, j;\ j=i-n;\ t=s-n)$ at time moments preceeding to the observation time interval. The latter permits to determine[1] from the a priori densities $P(\bar{V}), P(\bar{x}_0), P(f_s)$ $(s=1,2,\dots,j)$ the distribution density of the so-called vector of the reduced disturbances $\bar{V}_s=(\bar{V}, \bar{x}_s;\ u_s=0)$.

The set of equations (1) solved with respect to the output variable x_{i+1} , the vector $\bar{V}_s$ being fixed, has the form

$$x_{i+1}=\psi_{i+1}(\bar{V}_s, \bar{u}_{si})$$
$$(i=0,1,\dots,I;\ s=0,1,\dots,j;\ j=i-n). \tag{6}$$

The totality ($I-n+1$) of the auxiliary functions is introduced

$$\gamma_{sI}(\bar{y}_{sI}, \bar{u}_{sI})=\int\limits_{\Omega(\bar{V}_s)} w[\psi_{I+1}(\bar{V}_s, \bar{u}_{sI})]\,P(\bar{V}_s)\prod_{l=s}^{I}P(y_l/\bar{V}_s, \bar{u}_{s(l-1)})\,d\Omega(\bar{V}_s) \qquad (s=0,1,\dots,I-n).$$

The optimal control $u_i = u_i^*$ is computed at sequential minimization with respect to $u_I, u_{I-1}, \ldots, u_i$ of the functions $\gamma_{jI}, \gamma_{(j-1)(I-1)}, \ldots, \gamma_{ji}$ $(j = I - n)$, so that

$$\gamma_{ji}^* = \min_{u_i \in \omega(u_i)} \gamma_{ji}(\bar{y}_{ji}, \bar{u}_{j(i-1)}; u_i) \qquad (i = 0, 1, \ldots, I),$$

where

$$\gamma_{ji} = \int_{\Omega(\bar{y}_{(i+1)I})} \gamma_{jI}(\bar{y}_{ji}, \bar{u}_{ji}; u_{i+1}^*, u_{i+2}^*, \ldots, u_I^*) d\Omega(\bar{y}_{(i+1)I})$$

$\omega(u_i)$ - region of permissible controls with respect to (3).

For $n = I$ the totality of auxiliary functions degenerates into one function and the discussed methods coincide with the procedure of defining dual control.[2]

Similar methods can be presented for the case[3] when the optimal increment $\Delta u_i = u_i - u_{i-1}$ of control is determined in the class of functions,

$$\Delta u_i = \Delta u_i(\bar{y}_{ji}, \overline{\Delta u}_{j(i-1)}) \qquad (i = 0, 1, \ldots, I; j = i - n), \quad (7)$$

where

$$\overline{\Delta u}_{ji} = (\Delta u_j, \Delta u_{j+1}, \ldots, \Delta u_i)$$

The above given relations do not permit, in a general case, to obtain in the explicit form the algorithm of operation of the system's control device. Therefore we shall discuss the method of construction of a suboptimal control system, the system with independent identification of a plant.

Let us devide the coordinates of the vector $\bar{V}_j$ into two groups.

$$\bar{V}_j = (\bar{V}_j^{(1)}, \bar{V}_j^{(2)})$$

The random vector $\bar{V}_j^{(1)}$ consists of such variables of the vector $\bar{V}_j$, whose action on the plant leads to that the plant looses the property of neutrality [2].

Generally, the vector $\bar{V}_j^{(1)}$ involves random deviations of controlled plant parameters. The vector $\bar{V}_j^{(2)}$ consists of the rest variables of the vector of reduced disturbances.

Replace the controlled plant, described by (6), by its model variable in discrete time

$$x_{i+s} = \psi_{i+s}(\bar{V}_{ji}^{*}, \bar{V}_j^{(2)}, \bar{u}_{j(i+s-1)}) \quad (i = 0,1,\dots,I;\ s = 1,2,\dots,I-i+1;\ j = i-n). \tag{8}$$

Here, vector $\bar{V}_{ji}^{*}$, the estimate of the vector $\bar{V}_j^{(1)}$, is assumed to be known but successively corrected at each interval i in the result of minimization of a certain adopted risk function.

$$W_i^{(1)} = M\{w^{(1)}[\bar{V}_j^{(1)}, \bar{V}_{ji}^{*}(\bar{y}_{ji}, \bar{u}_{j(i-1)})]\} \quad (i = 0,1,\dots,I).$$

Correspondingly, the optimality criterion (4) is replaced by the sequence of risk functions.

$$W_i = M[w(x_{i+S})] \qquad (i = 0,1,\dots,I;\ S = I-i+1).$$

The control plant model (8) refers to neutral plants. Determination of optimal, with respect to risk W_i, control of this model is made sufficiently simply according to the discussed relationships. Here the obtained control function does not depend on the function of studying the vector $\bar{V}_j^{(1)}$ at intervals $i+1, i+2, \dots, I$.

2. Approximate synthesis of Systems Limited in Number of Devices Reproducing Coefficients of Control Algorithm

The system's accuracy estimated by the risk function (4) in optimal control can be improved either at the expense of increasing the number quantization intervals I of a pulse system or at the expense of optimization of program of variation in time of quantization intervals. In certain cases the efficiency of intervals optimization turns out to be the same as that achieved at increase of the number I equal intervals by several orders. Therefore a problem may be formulated of program optimization of quantization intervals variation, which is formulated as the problem of determining the optimal "control" coordinates (i.e. intervals), independent of current values of coordinates of the pulse system. This problem has already been solved on the basis of a specific method of statistical optimization.[4]

While synthesizing a system with limited number of coefficients of the control algorithm it is necessary, in addition to the original data (1-5), to set the number L of permissible variations of algorithm's coefficients and numbers of intervals i_K $(i_1=0;\ K=1,2,\dots,L)$, at which these variations are permitted.

Introduce into consideration L controlled plants with smaller numbers of quantization intervals which are described by equations

$$x_{i+1}^{(K)}=\psi_{i+1}(\bar{V}_{j_K},u_{j_K i}^{(K)})\quad (i=j_K,j_K+1,\dots,I;\ j_K=i_K-n),\quad (9)$$

obtained from (6). The plants are affected by the vectors $\bar{V}_{j_K}$ which coincide with the vectors of reduced disturbances in the desired system. Define for each plant (9) the optimal function $u_i^{(K)}=u_i^{(K)}(\bar{y}_{ji}^{(K)},\bar{u}_{j(i-1)}^{(K)})$ without taking into account limitations with respect to the number L of permissible variations of coefficients. In a case, when quantization intervals of the original system vary according to optimal program the desired control function at intervals $(i_{K-1}\div i_K)$ approximately coincides with optimal function of control of the K-th plant (9) at the interval i_K:

$$u_{i_K}(\bar{y}_{j_K i_K},\bar{u}_{j_K(i_K-1)})\equiv u_{i_K}^{(K)}(\bar{y}_{j_K i_K}^{(K)},\bar{u}_{j_K(i_K-1)}^{(K)}).$$

This control can be practically calculated by iterative methods with the use of the previously given relationships.

3. Control System Synthesis with Taking into Account the Constraints with Respect to Operations Which Can be Realized in a Control Device

A control plant is considered which consists of serially connected nonlinear inertia-free part, described by the function

$$v_i = v_i(u_i) \qquad (i = 0, 1, \dots, I), \tag{10}$$

and that under the influence of random disturbances of the linear part described by the equations

$$x_{\vartheta(i+1)} = \sum_{\vartheta=1}^{N} a_{\vartheta i} x_{\vartheta i} + \sum_{\tau=1}^{R} b_{\tau i} V_\tau + c_{\vartheta i}(1+\delta c) v_i \quad (\vartheta = 1, 2, \dots, N), \tag{11}$$

where v_i - value of the output variable of the nonlinear part of the system,

δc - random deviation of the coefficient $c_{\vartheta i}$.

Assume that the a priori densities of the vectors $\bar{V}$, $\bar{x}_0$, errors of measurements f_i and the value of δc are described by the normal distribution laws.

As a criterion of optimality the risk function is adopted

$$W = M(x_{I+1}^2).$$

Regarding the peculiarities of digital computers it is reasonable to adopt algebraic operations as operations which can be realized in the control device. Considering at first the case when dispersion D_c of the value δc is equal to zero we shall define optimal control in the class of linear functions of the type

$$u_i = \sum_{l=j}^{i} A_{jl} y_l + \sum_{l=j}^{i-1} B_{jl} v_l \qquad (i = 1, 2, \dots, I;\ j = i - n). \tag{13}$$

For the linear part of the plant (11) the equation (6) has the form

$$x_{i+1} = \sum_{q=1}^{Q} \alpha_{siq} V_{sq} + \sum_{l=s}^{i} \beta_{il} v_l$$

$$(i = 1, 2, \dots, I;\ s = 1, 2, \dots, j;\ j = i - n).$$

The conditional mathematical expectation $M(x_{I+1}/\bar{y}_{ji}, \bar{v}_{j(i-1)}; \bar{v}_{iI} = 0)$ is the known linear function of the measured variables of the system [5]

$$M(x_{I+1}/\bar{y}_{ji}, \bar{v}_{j(i-1)}; \bar{v}_{iI} = 0) = \sum_{l=j}^{i} A^*_{jl} y_l + \sum_{l=j}^{i-1} B^*_{jl} v_l.$$

It can be shown that the desired coefficients of the control algorithm (12) must satisfy the following conditions

$$
\begin{aligned}
&2A^*_{jl} M(y_l^2) + \sum_{\substack{k=j \\ k\neq l}}^{i} A^*_{jk} M(y_l y_k) + \sum_{k=j}^{i-1} B^*_{jk} M(y_l v_k) = \\
&= -2\beta_{iI} M[v_i(u_i) y_l] \qquad (l = j, j+1, \dots, i) \\
&2B^*_{jl} M(v_l^2) + \sum_{k=j}^{i} A^*_{jk} M(v_l y_k) + \sum_{\substack{k=j \\ k\neq l}}^{i-1} B_{jk} M(v_l v_k) = \\
&= -2\beta_{iI} M[v_i(u_i) v_l] \qquad (l = j, j+1, \dots, i-1).
\end{aligned}
\tag{14}
$$

The correlation moments of these relations should be calculated with the use of specific methods.[6] The values satisfying the conditions (14) can be determined by iterative methods.

In case of $D_c \neq 0$ there can be set the problem of approximate determination of optimal control in the class of rational functions. This control can be calculated from the relation

$$u_i = k_i M(x_{I+1}/\bar{y}_{ji}, \bar{v}_{j(i-1)}; \bar{v}_{iI} = 0; \delta c = \delta c^*),$$

where the coefficients k_i are determined from the condition

$$
\begin{aligned}
&-\beta_{iI} M[M(x_{I+1}/\bar{y}_{ji}, \bar{v}_{j(i-1)}; \bar{v}_{iI} = 0; \delta c = \delta c^*) v_i(u_i)] = \\
&= M\{[M(x_{I+1}/\bar{y}_{ji}, v_{j(i-1)}; v_{iI} = 0; \delta c = \delta c^*)]^2\},
\end{aligned}
$$

and the value δc^* is calculated from the condition of minimization of the risk function

$$R_{ci} = M\{[\delta c - \delta c^*(\bar{y}_{ji}, v_{j(i-1)})]^2\}.$$

4. Statistical Synthesis of Invariant Pulse Systems with Taking into Account Structure Constraints

In a number of cases the accuracy of a system obtained in statistical synthesis with taking into account structure constraints turns out to be insufficiently high comparing with ultimate accuracy possibilities of the system designed without taking into account the constraints.

Improvement of control accuracy can be achieved either at the expense of weakening of requirements with respect to structure constraint or at the expense of use of the principle of invariancy in synthesis of the system.

Let us assume that a control plant is described by the equations (10), (11). Set the problem of finding the control function minimizing the risk function (12) under the condition of minimal influence on the risk of variation of parameters of the distribution law of separate disturbances $V_1, V_2, \dots, V_\tau$ The control should be found in the class of functions linear with respect to variables of the vectors $\bar{y}_{ji}, \bar{v}_{j(i-1)}$.

Such a control at the first τ intervals can be found from the condition

$$M_i^* = \min_{u_i \in \omega(u_i)} \Big[M_{i(I+1)}\left(x_{I+1} / \bar{y}_{1i}, \bar{v}_{1(i-1)}; \bar{v}_{iI} = 0\right) + v_i(u_i) \sum_{K=i}^{I} c_{1K} \Big] \quad (i = 1, 2, \dots, \tau), \tag{15}$$

where in calculation of the value $M_{i\ell}(x_\ell / \bar{y}_{1i}, \bar{v}_{1(i-1)}; v_{i(\ell-1)} = 0)$ it is assumed that the measurement errors at these intervals are equal to zero.

Introduce at the control interval $(\tau+1) \div (I+1)$ a new controlled variable

$$y_i^* = y_i - x_i^*,$$

where the values x_i^* are calculated from the relation

$$x_i^* = M_{\tau i}\left(x_i / \bar{y}_{1\tau}, \bar{v}_{1(\tau-1)}; v_{\tau i} = 0\right) + v_\tau^* \sum_{K=\tau}^{i-1} c_{1K}$$

and the value v_τ^* is determined so that $x_{I+1}^* = 0$.

The desired control at the intervals $i = \tau+1, \tau+2, \ldots, I$ can be defined from the condition

$$M_i^* = \min_{u_i \in \omega(u_i)} \{ M(x_{I+1} / \bar{y}_{j_1 i}, \overline{\Delta v}_{j_1(i-1)}; \overline{\Delta v}_{iI} = 0) + \tag{16}$$

$$+ [v_i(u_i) - v_\tau] \sum_{K=1}^{I} c_{1K} \qquad (i = \tau+1, \tau+2, \ldots, I;$$

where $\qquad j_1 = i - n - \tau),$

$$\overline{\Delta v}_{j_1 i} = (\Delta v_{\tau j_1}, \Delta v_{\tau(j_1+1)}, \ldots, \Delta v_{\tau i}), \qquad \Delta v_{\tau K} = v_K - v_\tau.$$

The conditions (15) and (16) lead to the realtionships of the type (14), which can be solved by iterative methods.

References

1. V.S. Pugachev. Teoriya sluchainykh funktsii i yego primenenie k zadacham avtomaticheskogo upravlenija. Fizmatgiz, 1962.
2. A.A. Fel'dbaum. Osnovy teorii optimal'nykh avtomaticheskikh sistem. Fizmatgiz, 1963.
3. A.Ya. Andrienko. Statisticheskii sintez optimal'nykh impul'snykh sistem upravleniya s ogranicheniem po yomkosti pamyati upravlyaushchego ustroistva."Avtomatika i Telemekhanika", N.7, 1968.
4. A.Ya. Andrienko. Metod statisticheskoi optimizatsii nelineinykh impul'snykh sistem avtomaticheskogo upravleniya. Technicheskaya kibernetika, N.4, 1967.
5. Yu.V. Linnik. Metod naimen'shikh kvadratov i osnovy matematicheskoi teorii obrabotki nablyudenii, Fizmatgiz, 1958.
6. A.Ya. Andrienko, B.N. Petrov, Yu.P. Portnov-Sokolov. Otsenka tochnosti sistem upravleniya kosmicheskikh ob'jectov.

Trudy II simpoziuma IFAC po upravleniyu v kosmose, 1967.

43.8

OPTIMAL CONTROL SYSTEM FOR STATIONARY ARTIFICIAL CIRCUMTERRESTRIAL SATELLITE ORBIT

A.A.Lebedev, M.N.Krasilshchicov, V.V.Malishev
Moscow, USSR

- Solution of technical problem of optimization the stationary artificial circumterrestrial satellite (SACS) process of transfer to a given position with necessary accuracy at minimum energetical loss is being considered in this report.

By SACS we mean 24-hr equatorial satellite. For an observer on Earth such satellite will seem motionless. Thanks to this particularity the satellite can be used at global communications' system creation. A number of technical reasons do not permit to transfer the satellite directly to the longitude required. To transfer SACS to a required lo longitude method of transfer ellipse is used [I] (Figure I); the essence of this method is in the fact that satellite moves at elliptical orbit with apogee (perigee) at altitude corresponding to the radius of stationary orbit, the period of revolution being less (more) than 24-hr. Because of it satellite will drift to East (West). Very high final accuracy required is particular for this system, which leads to the necessity to create control algorithm on feedback principle. This feedback is realized by definition of SACS's angle (azimuth) α_u and SACS's distance r_u in coordinates of ground tracking station situated at any point on Earth. Since final accuracy and total energy loss depend not only on control algorithm but on satellite initial insertion velocity as well, it is interesting to solve also the problem of optimal calculated velocity of transfer the satellite to the altitude of stationary orbit. The given problem is the one of control based on incomplete data and its exact solution on this data is difficult. Then the approximate method is used[2], the sinthesis problem being divided into two independent problems. The first one consists of transfer process potimization in assumption that all the necessary for control data are available. The second one is to det this information by potimal processing the data of measurement. Principal assumptions at solution of formulated problems are given below: I) flat problem is being considered, that is satellite deviations at latitude are being neglected; 2) initial transfer errors at realization of corrective impulses is supposed further on that control is realized by velocity impulses) and errors of phase coordinates determination are introduced into discussion as irritations.

I. Let us take into account vector of SACS's phase coordinates x , the components of which being x_1 - satellite deviation from given position at longitude at the moment of passing through apogee (perigee); x_2- satellite's drift velocity per one revolution at transfer ellipse (Figure I). Equations of satellite motions in terms of above given coordinates are the following:

$$x_{i+1} = \Phi x_i + V(1+\mu_i)u_i + V\delta_i , \quad (i = 0,1,\ldots N) \tag{I}$$

where x_i is state vector x at the moment of apogee i ,

$$\Phi = \begin{Vmatrix} 1 & 1 \\ 0 & 1 \end{Vmatrix}, \quad V = \begin{Vmatrix} 1 \\ 1 \end{Vmatrix}. \tag{2}$$

Here u_i is calculated value of corrective velocity impulse i , μ_i , u_i , δ_i are multiplex and additive components of control influence error, N - number of revolutions at transfer ellipse orbit. It is supposed that μ_i and δ_i are accidental centered independent Gaussian numbers and

$$E[\mu_i^2] = \sigma_\mu^2 , \quad E[\delta_i^2] = \sigma_\delta^2 \tag{3}$$

where E - denote mathematical expectation.
So μ_i and ε_i may be considered as discret white noise.

Mathematical formula of problem to search the optimal correction algorithm comes to the following. It is necessary to find such a sequence $\{u_i\}$(i= I, ..., N), which assures the final precision required, characterized by given risk.

$$R_1 = E[x^T_{N+1} \lambda x_{N+1}] \tag{4}$$

at minimum energy loss measured by risk

$$R_2 = E[\sum_{i=1}^{N} u_i^2]. \tag{5}$$

solution of problem given has been received in[3] and comes to the following. Algorithm of optimal correction loorks like

$$u_i = -L_i x_i , \tag{6}$$

where

$$L_i = \Gamma_i^{-1} V^T \lambda_{i+1} \Phi, \quad \Gamma_i = V^T \lambda_{i+1} V (1 + \sigma_\mu^2) + \alpha \tag{7}$$

Matrix λ_i is determined according to the recurrent relation

$$\lambda_i = \Phi_i^T [\lambda_{i+1} - \lambda_{i+1} V \Gamma_i^{-1} V^T \lambda_{i+1}] \Phi_i \tag{8}$$

with an initial condition $\lambda_{N+1} = \lambda$.
Multiplier α should be determined by method of successive approximations or by graphic techniques on condation

$$\bar{x}_0^T \lambda_{10} \bar{x}_0 + C_{10} + Sp(\lambda_{10} K_{x0}) = R_1^* \qquad (9)$$

where $\bar{x}_0$ is mathematical expectation of vector x_0, $K_{x0} = E[x_0 x_0^T]$ being its covariance matrix, R_1^*, being the required value of risk R_1.

Matrix λ_{10} and coefficient C_{10} are found by recurrent correlations

$$\lambda_{1i} = \Phi^T \lambda_{1i+1} \Phi - \Phi \lambda_{1i+1} V L_i - L_i^T V^T \lambda_{1i+1} \Phi + L_i^T \Gamma_{1i}^{-1} L_i, \qquad (10)$$

$$C_{1i} = C_{1i+1} + \sigma_\varepsilon^2 Sp(\lambda_{1i+1} V V^T), \qquad (11)$$

where $\Gamma_{1i} = V^T \lambda_{1i+1} V (1 + \sigma_\mu^2)$
with initial conditions $\lambda_{1N+1} = \lambda$, $C_{1N+1} = 0$.

Satellite drift optimal calculated velocity at its transfer to the altitude of stationary orbit according to[3] is equal

$$\bar{x}_{20\,opt} = -(\lambda_{22})_0^{-1} (\lambda_{21})_0 \bar{x}_{10}, \qquad (12)$$

where $(\lambda_{22})_0$, (λ_{21}) are corresponging elements of matrix λ_0.
As follows from the mentioned above, for optimal control one should know state vector x and time τ_a of SACS's passing through apogee. Since these components can't be measured, problem of their optimal estimation arises. System

(I) is not convenient for determination of optimal estimations of vector on base of ground measurements, because vector of measurements itself is not included in this system. That is why equations of satellite motion in rotating geocentric coordinate system, linearized relatively to reference circular orbit are used. On condition that measurement are discret, these equations is:

$$z_k = A(t_k - t_{k-1}) z_{k-1} + B(1+\mu_{k-1}) u_{k-1} + B\varepsilon_{k-1} , \tag{I3}$$

where z_k - state vector at the time t_k , its components being $z_1 = \Delta r$ - deviations of SACS altitude from the one corresponding to designed circular orbit, $z_2 = \Delta\lambda$ - SACS deviations at longtude from the position at calculated circular orbit, $z_3 = \Delta V_r$ - radial component of satellite velocity deviation, $z_4 = \Delta V_s$ - deviation of SACS's tangential velocity component, $A(t_k - t_{k-1})$ - state transition matrix for discrete equation of movement, corresponding to the reference circular orbit

$$A(\tau) = \begin{Vmatrix} 4+3\cos\omega_0\tau & 0 & \frac{\sin\omega_0\tau}{\omega_0} & \frac{2}{\omega_0}(1-\cos\omega_0\tau) \\ -\frac{6}{r_0}(\omega_0\tau - \sin\omega_0\tau) & 1 & -\frac{2}{\omega_0 r_0}(1-\cos\omega_0\tau) & -\frac{3}{r_0} + \frac{4\sin\omega_0\tau}{r_0\omega_0} \\ 3\omega_0\sin\omega_0\tau & 0 & \cos\omega_0\tau & 2\sin\omega_0\tau \\ -6\omega_0(1-\cos\omega_0\tau) & 0 & -2\sin\omega_0\tau & -3+4\cos\omega_0\tau \end{Vmatrix} . \tag{I4}$$

$$B = \begin{Vmatrix} 0 \\ 0 \\ 0 \\ \mathring{b} \end{Vmatrix} , \quad \mathring{b} = \frac{\omega_0 r_0}{2\pi} , \quad \tau = t_k - t_{k-1} \tag{I5}$$

Measured coordinates α_{uk} and ε_{uk} are related with components of state vector z_k by equations:

$$\alpha_{uk}=\operatorname{arctg}\left[\operatorname{tg}\left(-x_{10}-\Delta\lambda_k-\lambda_u+\operatorname{arctg}\frac{R_3\cos\varphi\sin(-x_{10}-\Delta\lambda_k-\lambda_u)}{r_0+\Delta r_k-R_3\cos\varphi\cos(-x_{10}-\Delta\lambda_k-\lambda_u)}\right)\times \right. \tag{I6}$$

$$\left. \times\cos\operatorname{arctg}\frac{R_3\sin\varphi}{r_0+\Delta r_k-R_3\cos\varphi}\right];$$

$$\varepsilon_{uk}=\frac{\sqrt{R_3\cos\varphi\sin(-x_{10}-\Delta\lambda_k-\lambda_u)^2+r_0+\Delta r_k-R_3\cos\varphi\cos(-x_{10}-\Delta\lambda_k-\lambda_u)}}{\cos\operatorname{arctg}\dfrac{R_3\sin\varphi}{r_0+\Delta r_k-R_3\cos\varphi}} \tag{I7}$$

Here R_3 , λ_u , φ_k - are respectively ragius of Earth, longitude and latitude of ground tracking site. To receive a more simple estimatic algorithm equations (I6) and (I7) are linearized. To provide a better convergence of estimations at it, one linesrization about previos optimal predicted estimation at a preyious step of measurement is being used[4]. In this case it corresponds to linearization of equations (I6), (I7) in vicinity of transfer ellipse predicted on the base of optimal estimations after the next correction. Then (I6) and (I7) may be rewritten: can be expressed by

$$y_k = H_k z_k + v_k \tag{I8}$$

where y_k is vector of measurements; H_k is 2x4 matrix of partial derivatives, elements of which depend on time and on previous estimations; v_k is vector of measurement errors, which are Gaussian independent random numbers with zero mean and $E[v_K v_K^T] = Q_v$.

Taking into account the above said, optimal estimations of state vector x_K, are optimal estimations according to Bayess rule x_K^* , $\hat{x}_K$ and are determined by the following equations[5]

$$x_K^* = \hat{P}_K (P_K')^{-1} \hat{x}_K + \hat{P}_K H_K^T Q_v^{-1} y_K \, , \tag{19}$$

$$\hat{P}_K = \left[(P_K')^{-1} + H_K^T Q_v^{-1} H_K \right]^{-1} ,$$

$$P_K' = A_{KK-1} \hat{P}_{K-1} A_{KK-1}^T + B_{K-1} \sigma_\varepsilon^2 B_{K-1}^T + B_{K-1} \sigma_\mu^2 U_K^2 B_{K-1}^T \, , \tag{20}$$

$$\hat{x}_K = A_{KK-1} x_{K-1}^* \tag{21}$$

(22)

If after cycle of measurements completed the estimation error of components x_K not exceed the additive error of corrective impulse realization than the determination of moment of satellite passing throuh apogee takes place from the condition $\Delta V_z = 0$. In that case optimal estimation of τ_α is equal accuracy of lineár approximation

$$\tau_\alpha^* = \frac{1}{\omega_0} \operatorname{arctg} \left(- \frac{\Delta V_z^*}{\omega_0 \Delta z^* + 2 \Delta V_s^*} \right) . \tag{23}$$

Estimation of components of vector x is being determined for the time τ^*_d.

2. In the numerical treatment 24-hr satellite has been considered[I,6] for which $x_{10} = -35^o$. Parameters of stationary orbit were:

$V_0 = r_0\omega_0 = 3075$ m/sec; $r_0 = 42I65$ km, $T_0 = 86I64$ sec.

The transfer process is considered completed if the following conditions are fulfilled: I) satellite is within the region of admissible longitudinal deviations $|x_1| \leqslant \Delta\lambda_m$; 2) residual drift velocity x is such that staing of satellite within the region $|x_1| \leqslant \Delta\lambda_m$ during the time t_6 is quaranteed. $\Delta\lambda_m = 2,5^o$; t_6 = 60 revolutions 60 days were assumed. Region of admissible final errors on phase plane (x_1, x_2) is shown at figure 2.

Let us approximize the boundary of region Γ by ellipse Γ'

$$\frac{1}{t_6^2}x_1^2 + \frac{1}{t_6}x_1x_2 + x_2^2 = \frac{3}{4}\frac{\Delta\lambda_m^2}{t_6^2} \qquad (24)$$

To provide the final admissible accuracy with probability 0,997 as the required value of risk R_1, the value R_1^* should be taken

$$R_1^* = \frac{1}{9}\cdot\frac{3}{4}\frac{\Delta\lambda_m^2}{t_6^2} \qquad (25)$$

As follows from (24), matrix λ in risk R_1 (4) is equal

$$\lambda = \begin{Vmatrix} \frac{1}{t_6^2} & \frac{1}{2t_6} \\ \frac{1}{2t_6} & 1 \end{Vmatrix} \qquad (26)$$

Taking into consideration (25), (26) and using equations (7) - (10), calculations were carried out for different values of λ and N and $\sigma_{\mu}^2 = 25.10^{-4}$; $\sigma_{\varepsilon}^2 = 0.003 \, \mathrm{grad}^2/\mathrm{rev}^2$ results are given in figures 3, 4,5.

The definition of the necessary interval of measurements and of their number per each revolution is given, considering the estimated error of the satellite velocity components should not exceed the additive error resulting fron the application of the corrective velocity impulses.

On the grounds of the covariance matrix numerical calculations in accordance with equations (20),(21) for $\sigma_{\mu} = 0$, the interval between measurements was taken equal to 1 hr, their number being S. Curves characterizing the convergence of state vector Z coordinate are given in fig.6. It was assumed ~~that~~ $\sigma_{\alpha u} = 0.003^{\circ}$; $\sigma_{zu} = 25$km., the initial launching errors being $\sigma_{\Delta z} = 50$km.; $\sigma_{\Delta\Omega} = 1^{\circ}$; $\sigma_{\Delta V_s} = \sigma_{\Delta V_z} = 1.7$ m/sec /1/, /6/.

3.In order to examine the actual possibilities of SACS's satisfactory transfer to a given position using the suggested control algorithm, the simulation of the closed-loop control system by Monte Carlo method was applied. When simulating,attitude and stabilization errors during the SACS correction were taken into account, assuming that they do not exceed the measurement error and the corrective impulse application error by more than $\pm 2^{\circ}$ The satellite dynamics was defined by the equations of the elleptical theory /7/. Furthermore it was assumed that α_u and z_u relations to vector Z components were linear.

When simulating the closed-loop system the following principal versions were examined:

1) supposing that the exact values of x_1, x_2 and τ_λ were known, it was possible to check the correction algorithm. The dispersion pattern for this case is given in fig.7. As may be seen from fig.7, the dispersion is near to normal and is in good agreement with the calculated dispersion;

2) if measurement errors are present, taking into consideration all the above mentioned factors. The dispersion pattern for this case is given in fig.8. As it may be seen from fig.8 the dispersion pattern is rather different from the previous one, but even in this case it is within the admissible limits.

References

I. Кенг Г.,Кенехен М.Е. Выведение на заданную долготу и управление орбитой 24-часового экваториального спутника Земли."Ракетная техника и космонавтика", русский пер., 1964, № 6.

2. Ли Р. Оптимальные оценки, определение характеристик и управление. Изд."Наука", 1966.

3. Малышев В.В. Задача об оптимальном дискретном управлении конечным состоянием линейной стохастической системы. "Автоматика и телемеханика",1967, № 5.

4. Красильщиков М.Н. Некоторые вопросы определения достаточных координат при управлении динамическими объектами. Доклад на I Всесоюзном симпозиуме по статистическим проблемам в технической кибернетике. Москва, 1967.

5. Богуславский И.А. Об уравнениях стохастического управления. "Автоматика и телемеханика", 1966, № II.

6. Боучер Р.А. Использование электрических ракетных двигателей в системе управления стационарными спутниками. "Вопросы ракетной техники",1965, № 4.

7. Балк М.Б. Элементы динамики космического полета. Изд."Наука" 1965 .

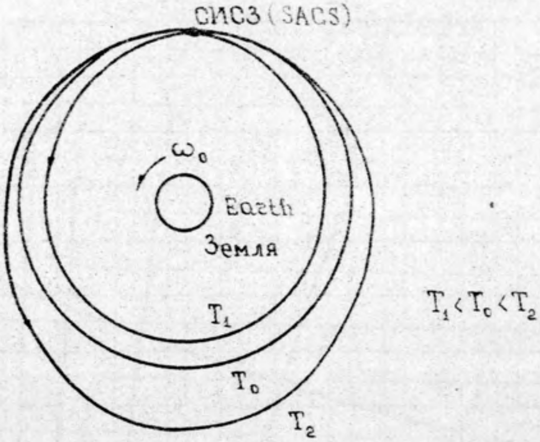


Figure 1

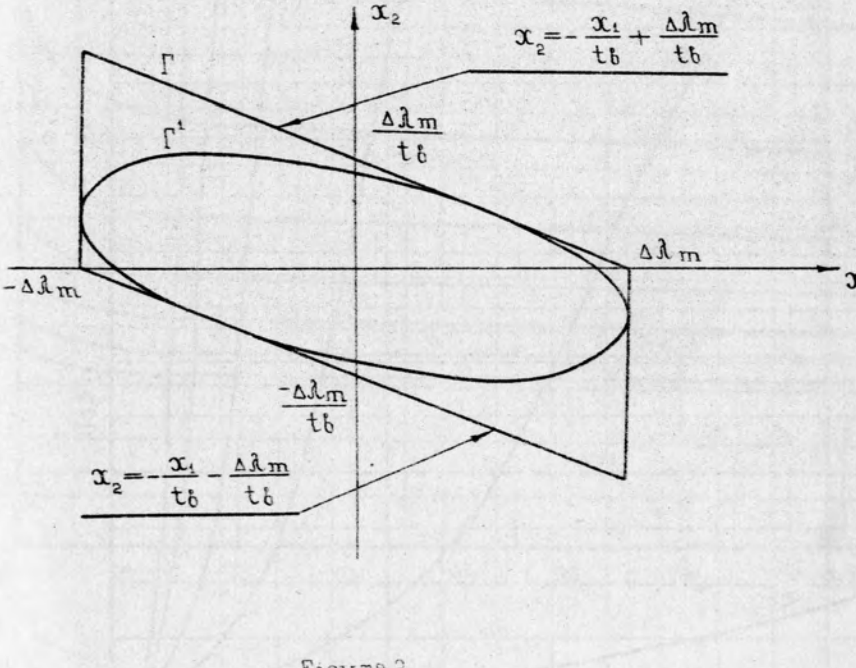


Figure 2

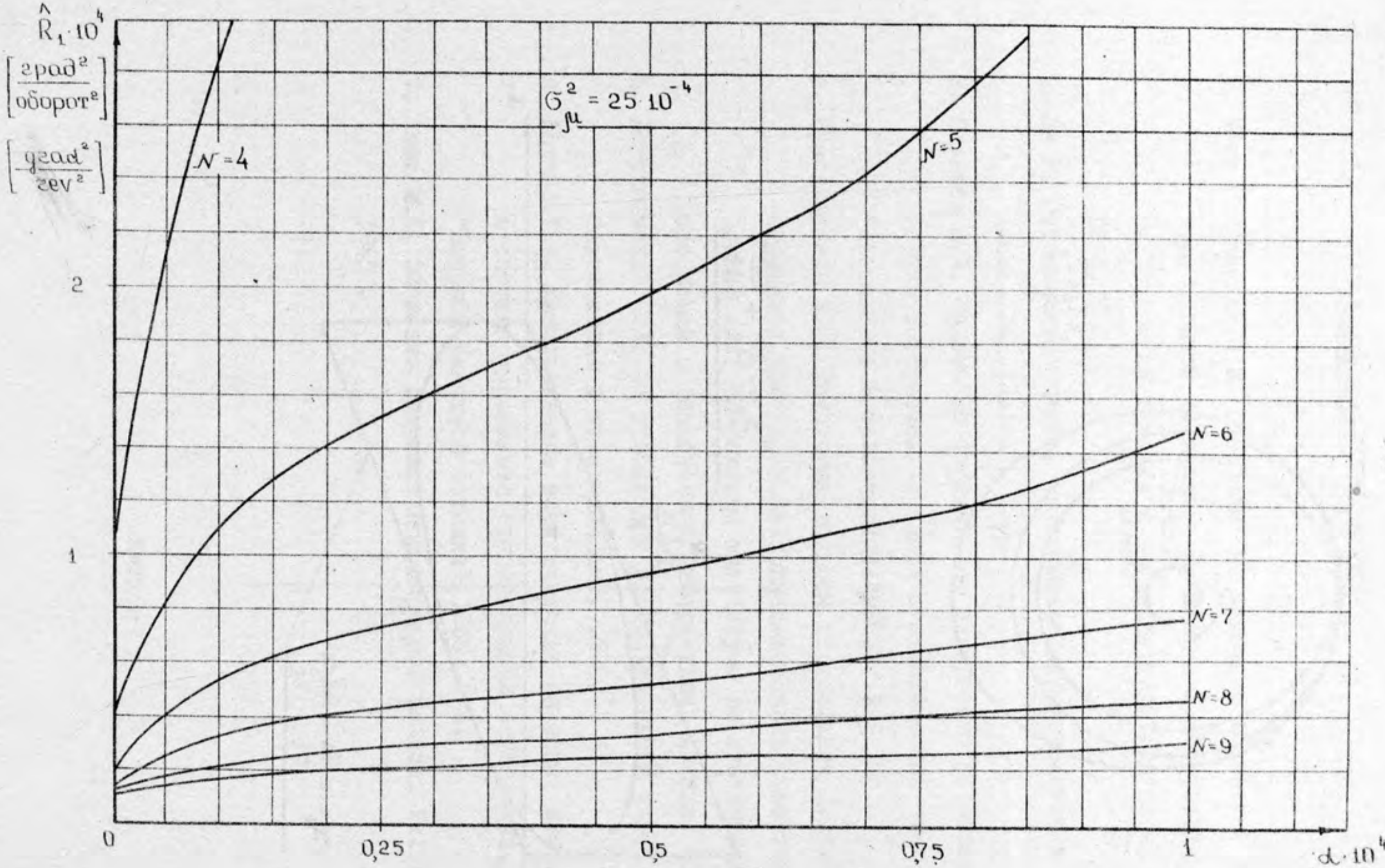
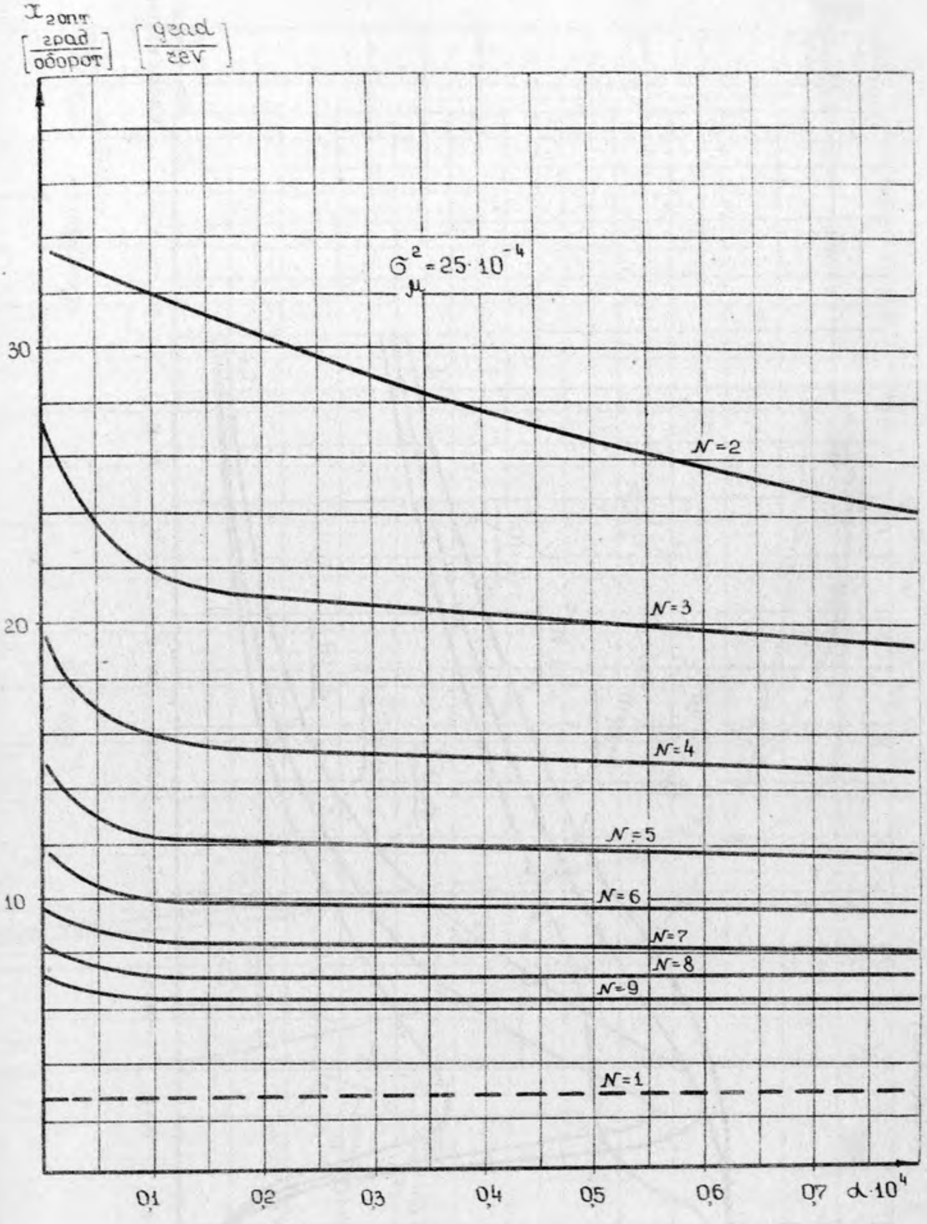

$\hat{R}_1 \cdot 10^4$
[град²/оборот²]
[grad²/rev²]
$\sigma_\mu^2 = 25 \cdot 10^{-4}$
N = 4
N = 5
N = 6
N = 7
N = 8
N = 9
0
1
2
0,25
0,5
0,75
1
$\alpha \cdot 10^4$

Figure 3



$\bar{x}_{2опт}$
[град/оборот] [grad/rev]
$\sigma_{\mu}^2 = 25 \cdot 10^{-4}$
30
20
10
N=2
N=3
N=4
N=5
N=6
N=7
N=8
N=9
N=1
0,1
0,2
0,3
0,4
0,5
0,6
0,7
$\alpha \cdot 10^4$

Figure 4

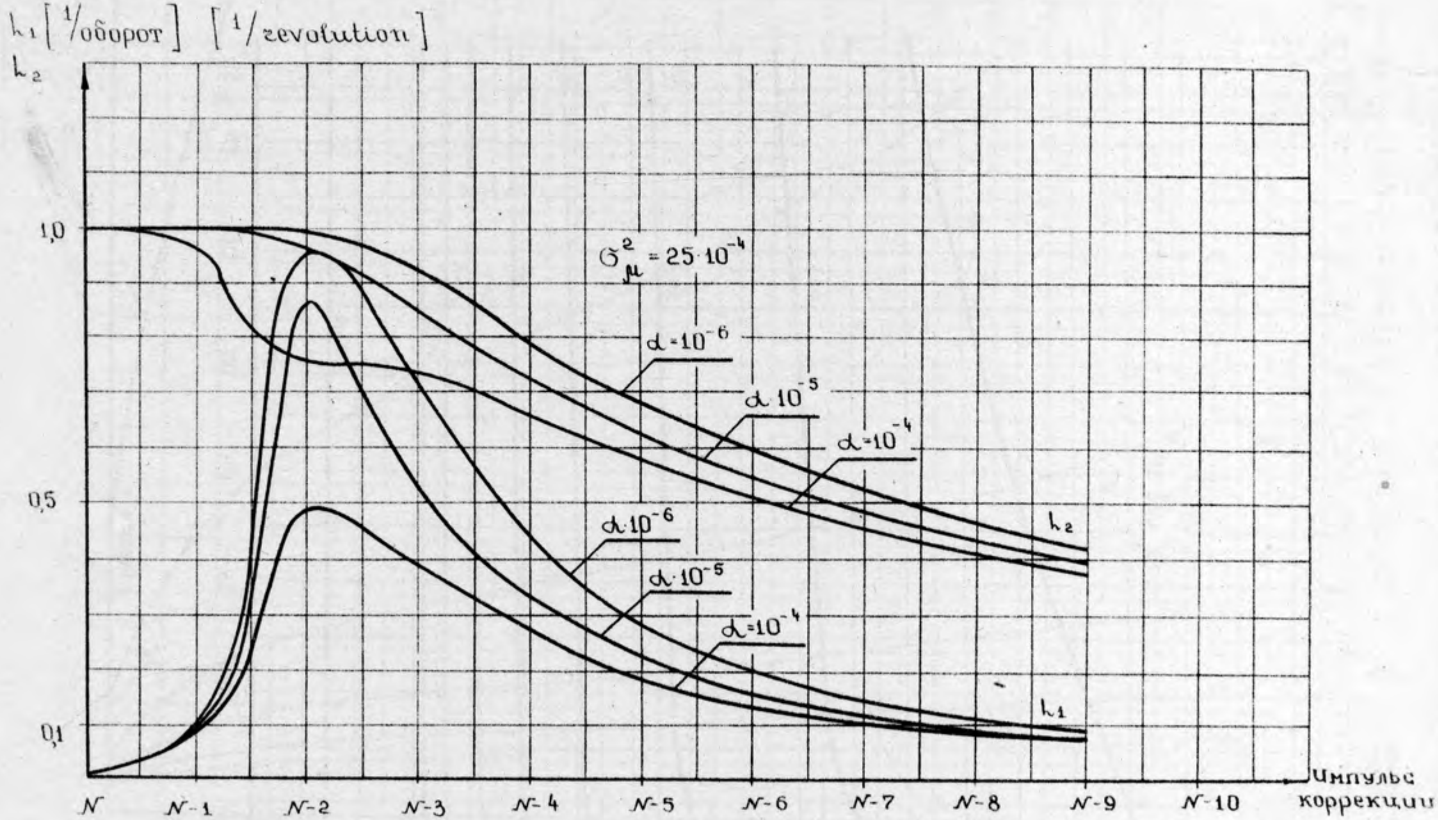
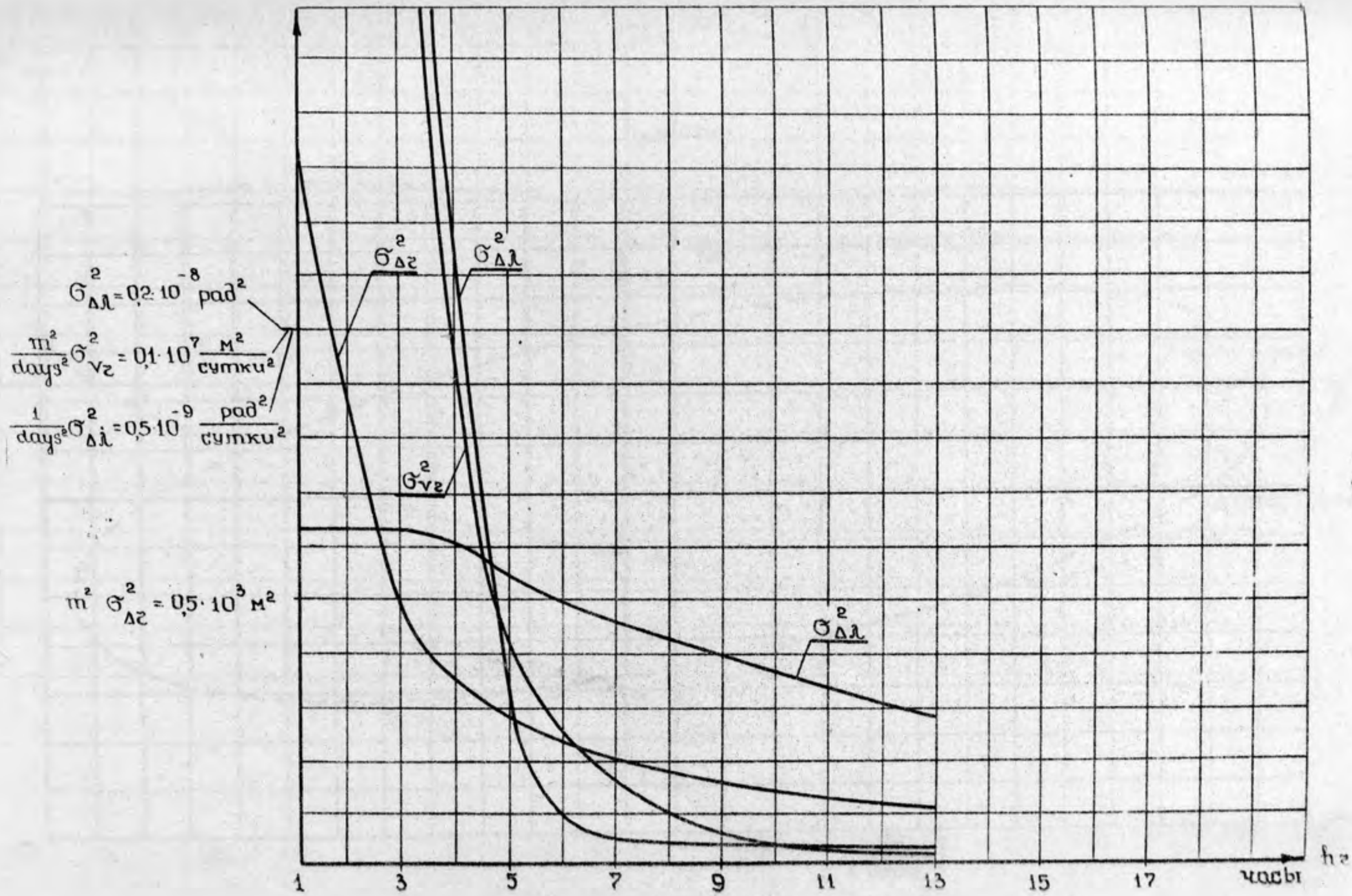

h1 [1/оборот] [1/revolution]
h2
1,0
0,5
0,1
σ²μ = 25·10⁻⁴
α=10⁻⁶
α·10⁻⁵
α=10⁻⁴
α·10⁻⁶
α·10⁻⁵
α=10⁻⁴
h2
h1
N
N-1
N-2
N-3
N-4
N-5
N-6
N-7
N-8
N-9
N-10
Импульс коррекции
number of correction impulse

Figure 5



σ²Δλ = 0,2·10⁻⁸ рад²
m²/days² σ²vz = 0,1·10⁷ м²/сутки²
1/days² σ²Δλ̇ = 0,5·10⁻⁹ рад²/сутки²
m² σ²Δz = 0,5·10³ м²
σ²Δz
σ²Δλ
σ²vz
σ²Δλ
1
3
5
7
9
11
13
15
17
часы
hr

Figure 6

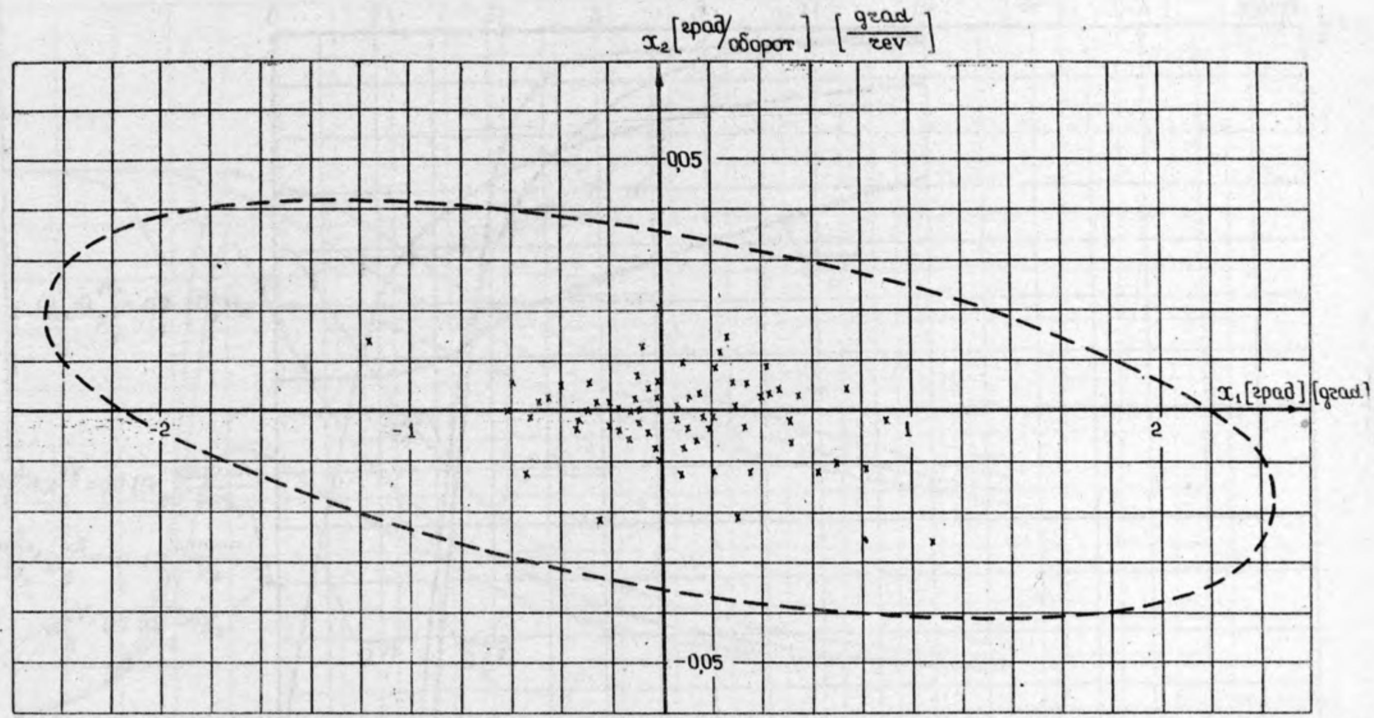


Figure 7

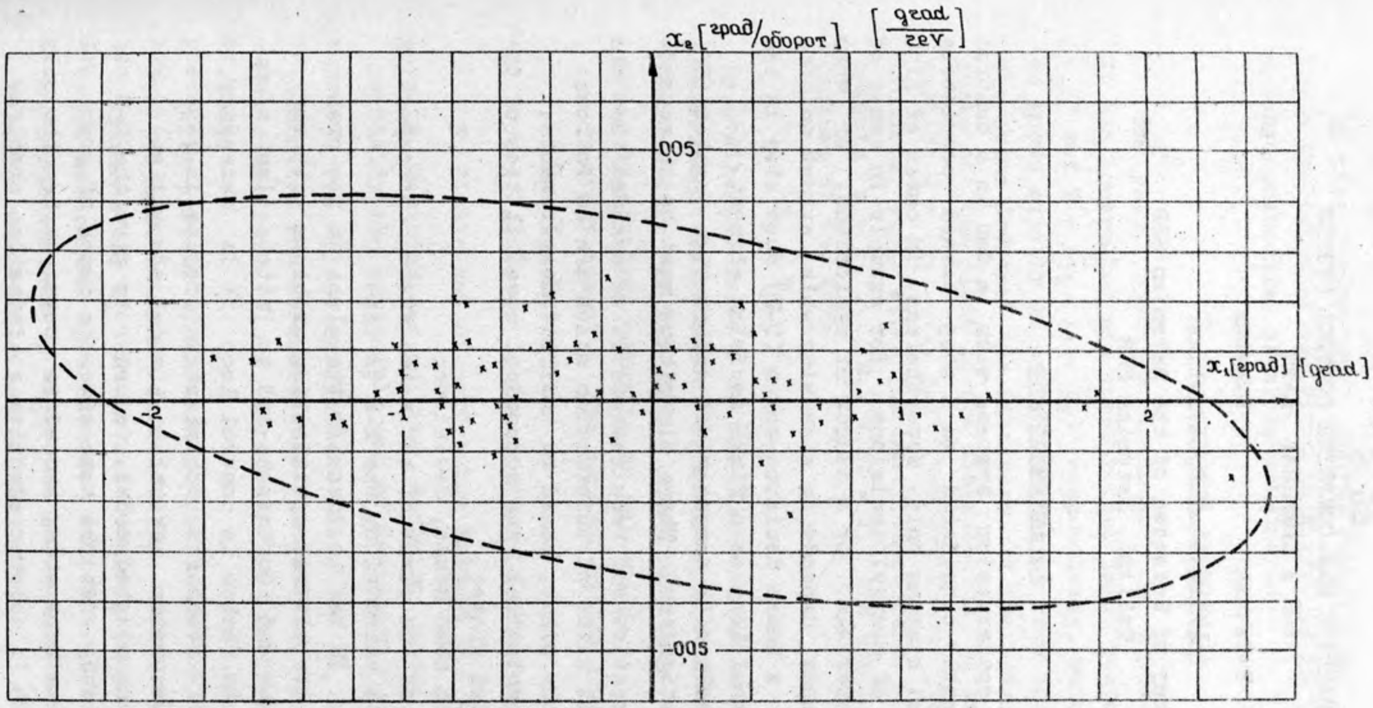


Figure 8

46.6

AN ADAPTIVE MAN-COMPUTER CONTROL SYSTEM FOR A CHEMICAL PLANT

R.Tavast L.Mytus

Institute of Cybernetics
Academy of Sciences of the Estonian SSR
Tallinn, Estonian SSR

1. Introduction

Supervisory optimization systems with a man in a control loop are frequently considered as a step toward completely automatic control systems only. Nevertheless, in cases of high responsibility of control decisions, for example in cases of high costs of materials, or a danger of explosions, or when a plant performance depends to a considerable extent on nonformal factors, a human decision-maker (D-M) must stay in the supervisory control loop as a final decision element.

The paper deals with computerized information system (IS) algorithms investigation. These algorithms must be in accordance with D-M practice and requirements. It is assumed that the D-M needs several kinds of information which are as follows:

- state of the plant, based on indirect measurements;
- limit technological and economical possibilities of the plant of estimated state;
- optimal, in some sense, decisions;
- behaviour of the plant of estimated state, corresponding to various inputs selected by the D-M to test.

Consequently, IS has to identify the plant in the presence of noises, to solve several optimization problems with various perfomance indexes and constraints and to imitate plant operating. Due to human factor in control loop it is necessary to solve the overall stochastic optimization problem in separate steps, plant parameter estimation (model adaptation) and optimization using adapted model, in order to give the D-M the results of each step. At the same time, in cases of practical importance such a separating procedure seems to be the only possible approach in solving adaptive optimization problems.

In this report the problem formulation and its solution for a typical chemical plant is given. As an example the formaldehyd production plant is considered.

2. The control problem.

The plant operates continously on a time interval $[0,T]$. It has controlled and uncontrolled input vectors $U = (u_1, \ldots, u_r)'$ and $V = (v_1, \ldots, v_s)'$ respectively, which are assumed to be known exactly at every time moment from interval considered. The prime denotes the transpose of the matrix. The physical state of the plant is the distribution of some physical properties $Q = (q_1, \ldots, q_t)'$ (temperatures, pressures, concentrations, etc.) along the space coordinate η. At discrete time moments $n = 1, \ldots, N_1$, $N_1 \Delta t = T$, the stochastic m-vector process of the form

$$Z_n = \bar{Y}_n + W_n \tag{1}$$

is available, where $\bar{Y}$ is the plant output vector and W_n is the measurement noise vector sequence with bounded variances. The time varying of the plant characteristics is caused for example by catalyst decaying or poisoning.

At every moment n the control system knows:

1. U_n, V_n, Z_n;
2. Hypothetical model of the plant $Y = F(U, V, \theta, n)$;
3. Finite set of perfomance indexes w_i, $i \in \Lambda = \{1, \ldots, e\}$;
4. Restrictions on the decision space that determines partially the admissible set of decisions Γ_U.
5. The plant environment situation S_n characterized by economical and technological time-varying, partially ill-defined conditions (plans, arrangements, technological conditions, etc.).

The purpose of the man-computer control system is to make and adjust control decisions U_n at moment n to satisfy conditions in S_n in the best manner, in some sense. In most cases economical optimization criteria (perfomance indexes) are predetermined by S_n, but there is always some freedom in choosing admissible set of decisions Γ_U. Sometimes S_n may change in a manner that demands the optimization criterion to be altered. The optimization problem formulation is the first task of the D-M. (It must be emphasized that the problem of mapping S_n into Γ_U and w_i, $i\in\Lambda$ is the psychological one and it is not discussed here). The task of IS algorithms is to solve several optimization problems formulated by the D-M for the time-varying plant. The final decision U_n at S_n must be made by the D-M again on the bases of the solutions, given by IS.

For getting additional information he can access to the computer.

<u>Optimization problem.</u> Find point U_i^* from admissible set $\Gamma_U \subset E^r$ to minimize perfomance index expectation

$$J_i = \mathcal{E}\, w_i(U_n, \bar{Y}_n),\ i\in\Lambda, \tag{2}$$

where w_i are known functions(profit, production yield, product quality, etc.). Admissible set Γ_U is of the form of joint probabilistic conditions

$$\Gamma_U = \left\{U:\ P[G(\bar{U}_n, \bar{Y}_n, V_n, \underline{G}, \bar{G}) \le 0] \ge \gamma_G\right\} \tag{3}$$

where $P[A]$ denotes probability of A, $1-\gamma_G$ is the subjectively allowable probability of violating condition $G(\cdot) \le 0$, the latter being an inequality system

$$\begin{aligned} &\underline{g}_1 \le g_1(U_n, \bar{Y}_n) \le \bar{g}_1 \\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \\ &\underline{g}_c \le g_c(U_n, \bar{Y}_n) \le \bar{g}_c . \end{aligned} \tag{4}$$

Some of the elements of real positive vectors $\underline{G} = (\underline{g}_1, \ldots, \underline{g}_c)$ $\bar{G} = (\bar{g}_1, \ldots, \bar{g}_c)$ are fixed by S_n, some can be chosen by the D-M.

<u>Optimization problem solving</u> is the task of computer algorithms, which will be considered in the following section.

<u>Final decision-making.</u> The D-M can accept one of the U^*, $i\in\Lambda$ and make decision $U_n = U_{in}^*$, or interpolate between several solutions, or formulate a new problem. Before adjusting on the plant the final decision candidate will be tested on the model.

3. The plant model.

The stationary behaviour of a distributed plant is described by lumped parameter model of the following structure. The model has input vectors U, V that coincide with those of the plant. Output vector $Y = (y_1, \ldots, y_m)'$ elements are known functions of the physical state Q at points $\eta_1, \ldots, \eta_v$

$$y_j = h_j[Q(\eta_1), \ldots, Q(\eta_v)], \quad j = 1, \ldots, m. \tag{5}$$

Differential equation which has been constructed on the hypotheses concerning the plant

$$D[Q(\eta), U, V, \Delta] = 0, \tag{6}$$

involves space derivatives $\partial Q/\partial\eta$ and unknown parameter vector Δ, but does not involve time derivatives (heat and mass transfer equations in stationary conditions together with equations of chemical kinetics). Equation (6) subject to boundary conditions

$$B[Q(\eta_b), U, V] = 0, \tag{7}$$

where η_b are coordinates of plant boundary, determines the solution

$$q_k = \varphi_k(\eta, U, V, \Delta), \; k = 1, \ldots, t, \tag{8}$$

with U, V, Δ fixed. Substituting (8) into known functions h_j, $j = 1, \ldots, m$ we have static operator as the plant model

$$Y = F(U, V, \Delta), \quad F = (f_1, \ldots, f_m)'. \tag{9}$$

In all cases of practical importance we can obtain numerical solutions of (6),(7) only. The elements of parameter p_Δ-vector Δ are [from equation (6)] heat and mass transfer coefficients, chemical reaction rate coefficients, etc. Several elements of Δ may slowly vary in time. There is extensive literature concerning chemical plant modelling in the form (6), (7) (as, for instance in[3]) but theoretical models with time dependent parameters are seldom available. Hence we shall take some formal model approach

$$\Delta_n = C(n,\theta), \tag{10}$$

where θ is an p-vector of unknown real parameters and n is a time index as before. Some simple examples of (10) are

$$\Delta_n = \theta, \quad p_\Delta = p, \tag{11}$$

$$\Delta_n = H\begin{bmatrix}1\\n\end{bmatrix}, \quad p = 2p_\Delta, \tag{12}$$

here H is an p_Δx2 matrix of unknown parameters, θ is known to lie in a domain Γ_θ:

$$\Gamma_\theta = \{\theta\colon \underline{\theta} \le \theta \le \bar{\theta}\}. \tag{13}$$

Taking (10) into consideration we have time dependent memoryless plant model

$$Y = F(U, V, \theta, n). \tag{14}$$

4. Multidimensional parameter estimation.

Consider the m-vectors residuals

$$\varepsilon_n = Z_n - F(U_n, V_n, \theta, n), \quad n = 1, \ldots, N, \tag{15}$$

which are random independent vectors with unknown multidimensional distributions. It is assumed that at the beginning of the interval $[0, T]$ N sets of vectors U_n, V_n, Z_n, $n = 1, \ldots, N$ as the basis of initial estimation, are available. Thus we are led to the problem of nonlinear simultaneous estimation of unknown parameters[1]. We introduce following notations

$$\begin{aligned}
X_n &= (U_n, V_n, n)' \\
Z_{jN} &= (z_{j1}, \ldots, z_{jN})' \\
\varepsilon_{jN} &= (\varepsilon_{j1}, \ldots, \varepsilon_{jN})' \\
F_{jN} &= [f_j(X_1, \theta), \ldots, f_j(X_N, \theta)]' \\
Z_N &= (Z'_{1N}, \ldots, Z'_{mN})' \\
\varepsilon_N &= (\varepsilon'_{1N}, \ldots, \varepsilon'_{mN})' \\
F_N &= (F'_{1N}, \ldots, F'_{mN})'
\end{aligned} \tag{16}$$

So the problem is to find estimate $\hat{\theta}$ that minimizes quadratic form

$$\Phi(\theta, \Sigma) = \varepsilon'_N \Omega^{-1} \varepsilon_N, \quad \varepsilon_N = Z_N - F_N, \tag{17}$$

with $\Omega = \Sigma \otimes I_{N\times N}$, where Σ is an unknown residual covariance matrix, $\otimes$ is the symbol of direct product, $I_{N\times N}$ is the unit matrix.

The estimate of Σ can be computed in several ways. For example, the following iteration process can be used. Denote by $\theta_N(\Sigma_i)$ the parameter estimate corresponding to minimization of $\Phi(\theta, \Sigma_i)$:

1. Take as the initial estimate $\hat{\Omega}_o = \Sigma_o \otimes I$ based on known measurement properties.

2. Minimize $\Phi_N(\theta, \Sigma_i)$ with $\Omega^{-1} = \hat{\Sigma}_i^{-1} \otimes I$ to determine $\hat{\theta}_N(\Sigma_i)$.

3. Evaluate estimates of $\hat{\sigma}_{jk}$, j, k = 1, ..., m elements of $\hat{\Sigma}_{i+1}$:

$$\hat{\sigma}_{jk} = \frac{\hat{\varepsilon}_j'\hat{\varepsilon}_k}{N}, \quad \hat{\varepsilon}_j = (\hat{\varepsilon}_j, \ldots, \hat{\varepsilon}_{jN})',$$

$$\hat{\varepsilon}_{jn} = z_{jn} - f_j[X_n, \hat{\theta}_N(\hat{\Sigma}_i)]. \tag{18}$$

4. Compute $|\det\Sigma_{i+1} - \det\Sigma_i| \le \xi$, if yes -go to 5., or else take $\Sigma_i = \Sigma_{i+1}$ go to 2.

5. Take $\hat{\theta}_N = \hat{\theta}_N(\hat{\Sigma}_{i+1})$.

The convergence properties of this procedure are not known.

Another method[1] uses each of the residual ε_j, j = 1,..., m least-square estimate $\hat{\theta}^{(j)}$ approach:

1. Minimize $\Phi^{(j)}(\theta) = \varepsilon_j'\varepsilon_j$ to determine $\hat{\theta}^{(j)}$, j = 1,..., m.

2. Compute elements $\hat{\sigma}_{jk}$ of the matrix $\hat{\Sigma}$ as in point 3. of previous method, where

$$\hat{\varepsilon}_{jn} = z_{jn} - f_j(X_n, \hat{\theta}^{(j)}), \quad j = 1, \ldots, m. \tag{19}$$

3. Minimize $\Phi(\theta, \hat{\Sigma})$ to determine initial estimate $\hat{\theta}_N$.

If we assume, due to complex interlacing of independent random effects, the normality of ε_j, the least-square estimate will be consistent and the estimate $\hat{\Sigma}$ converges in probability to Σ[1].

It is necessary to apply here minimization procedures with global minima seeking property. In[7] such an algorithm has been developed. Its brief description will be given in section 5. After the initial estimation of $\hat{\theta}_N$ the corresponding recursive estimate correction $\hat{\theta}_{N+1}$, $\hat{\theta}_{N+2}$ should be done since X_n, Z_n, n = N+1, N+2, ... arrive successively. Consider a slight modification of stochastic approximation scheme[4, 5, 6] for the case of vector output which is closely related to the "batch processing" scheme of reference[5].

Using the found matrix estimate $\hat{\Sigma}$ the unit step loss function equals to

$$\zeta = \varepsilon_n'\hat{\Sigma}^{-1}\varepsilon_n \tag{20}$$

Its gradient is the p-vector

$$\zeta = \left(\frac{\partial\zeta}{\partial\theta}\right)' = -2\left(\frac{\partial F}{\partial\theta}\right)'\hat{\Sigma}^{-1}\varepsilon_n \ , \tag{21}$$

where $\frac{\partial F}{\partial\theta}$ is an mxp matrix. Introduce an pxm matrix

$$K_n = \frac{1}{\sum\limits_{k=1}^{n}\left\|\left(\frac{\partial F}{\partial\theta}\right)'_{\hat{\theta}_k}\right\|^2}\left(\frac{\partial F}{\partial\theta}\right)_{\hat{\theta}_n}\hat{\Sigma}^{-1} \ , \tag{22}$$

where $\|B\|^2 = \lambda_{max}$ with λ_{max} equals to max eigenvalue of $B'B$. Then the truncated recursive scheme of the form

$$\hat{\theta}_{n+1} = [\hat{\theta}_n + K_n\varepsilon_n]_{\Gamma_\theta}, \tag{23}$$

with

$$[\hat{\theta}]_{\Gamma_\theta} = \begin{cases} \overline{\theta} \ , \text{ if } \hat{\theta} \geq \overline{\theta} \\ \hat{\theta} \ , \text{ if } \underline{\theta} < \hat{\theta} < \overline{\theta} \\ \underline{\theta} \ , \text{ if } \hat{\theta} \leq \underline{\theta} \end{cases} \ , \tag{24}$$

satisfies the conditions of theorem 6.4 ref[5], and hence $(\hat{\theta}_n - \theta) \to 0$, in m.sq. and w.p. one while $n \to \infty$.

5. The approximate solution of the optimization problem.

In this section, instead of plant output vector $\overline{Y}_n$, in the inequalities (3) and perfomance indexes w_i the model output function $F(U_n, \hat{V}_n, \hat{\theta}_n, n)$ at stage n would be used. So the original optimization problem is substituted by its deterministic approximation: find U^*_{in} to minimize

$$J_{id} = w_i[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)], \ i\in\Lambda \tag{25}$$

subject

$$\Gamma_{Ud} = \{U: \underline{G}_d \leq G[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] \leq \overline{G}_d\} \tag{26}$$

Here is assumed that the D-M is able to choose vectors $\underline{G}_d$, $\overline{G}_d$ to guarantee satisfaction of (3) with probability at least γ_G.

There are numerous possible (25), (26) solution methods. We shall reduce this to the seeking of unconstrained minima of the function

$$\tilde{J} = J_{id} + \chi\Pi(G, \underline{G}_d, \overline{G}_d), \tag{27}$$

where

$$\Pi = \sum_{j=1}^{c} g_{jd} - \bar{g}_{jd} + |g_j - g_{jd}| + |g_j - \bar{g}_{jd}| \qquad (28)$$

and χ is a positive constant. Here again algorithm Ru-237 is useful. It turns to account the values of $\tilde{J}$ only and does not use the derivatives, scanning on some successively decreasing domain in Γ_U with successively decreasing steps.

The IS computer algorithm is schematically on fig. 3. The initial estimation of $\hat{\Theta}_N$, $\hat{\Sigma}$ along the second method is performed using Ru-237. Further $\hat{\Sigma}$ is used in the recursive model parameter correction. Parameter estimates and the smoothed values $\hat{V}_n$ of V_n are used in the model for finding the optimal decision U_i^* with w_i, Γ_{Ud} predetermined by the D-M, and for computing $Y_k = F(U_k, \hat{V}_n, \hat{\Theta}_n, n)$ with U_k set by the D-M.

6. Formaldehyd plant control system.

The plant (fig. 2). Methanol (CH_3OH) and water are mixed in the boiler (position 1) and are converted into steam (2). The steam mixed with air is preheated (3) and flows into the reactor (5). Reactor output gas, after cooling, goes into absorbing system where formaldehyd (CH_2O) and methanol are absorbed, gas consisting of CO_2, CO, H_2, O_2, N_2 is thrown off. The plant is controlled by air flow rate (u_1), methanol-water correlation (u_2), contact zone temperature (u_3), and inlet air temperature (u_4). The controller outputs (v_i, $i = 1, 2, 3, 4$) are filtrated. Filtration error is assumed to be negligible. At discrete time moments, measured values of reaction gas and the product analysis, methanol and product flow rates are available.

Decision-maker. The D-M is the plant technologist who chooses $U = (u_1, \ldots, u_4)'$ in complex varying situations. This decision depends on catalyst preparing mode, on nominal (planned) values of product yield and quality, on season, conditions of absorbtion, in a word it depends on S_n. The main reasons for making a new decision are catalyst decaying and nominal values changing. An experienced technologist can choose for each S_n decisions that are admissible but may be far from optimality. D-M does not make a new decision if S_n

changes although decision stated previously appeared to be admissible. IS improves the D-M's decisions by complementing his conceptual model.

The control goals. The admissible decisions are such as prevent explosion and guarantee accomplishing of nominal values of product cost, quality and yield. In each state of the plant (catalyst activity, absorbtion coefficients) the D-M can choose several control goals as situation S_n changes:

1. Maximum average product yield;
2. Minimum average production cost;
3. Maximum average profit;
4. Minimum average methanol concetration in the product.

The plant model. Conversion of methanol into formaldehyd is a heterogeneous catalytic process in a fixed isothermal bed of silver catalyst. The main reactions are methanol oxydation and dehydrogenation together with side and consecutive reactions. Reaction rates are diffusion controlled, hence rate coefficients depend on gas flow rate. Model equations arose from heat and mass balance and chemical kinetics. The conversion process is an autothermal one: reaction zone temperature depends on the concentration of reagents, steam temperature and rate coefficients which in turn depend on zone temperature. Therefore the model output can be computed iteratively. The behavior of model and plant is similar (existence of stationary points, directions of gradients of Y). By identification procedure the minimum weighed mean-square model error can be obtained.

Identification. Parameters to be identified are reaction rate coefficients, heat losses, formaldehyd and methanol absortion coefficients.

Optimization problems. The admissible set Γ_U is constructed on bases of following set of inequalities:

1. $g_U \le U_n \le \bar{g}_U$, g_U, $\bar{g}_U$ - 4-vectors of technological restrictions;

2. $g_2 \le g_2[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)]$ - safety conditions;

3. $g_3[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] \le \bar{g}_3$ - normal absorbtion conditions

4. $g_4 \le g_4[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] \le \bar{g}_4$ - CH_2O

5. $g_5 \le g_5[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] \le \bar{g}_5$ - CH_3OH } concentrations in the product: quality conditions;

6. $g_6 \le g_6[U_n\ F(U_n, \hat{V}_n, \hat{\theta}_n, n)]$ - nominal yield requirement;

7. $g_7\ [U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] \le g_7$ - nominal production cost requirement, where g_2 is the CH_3OH concentration before the reactor, g_3 is gas volume flow rate, g_4 and g_5 are CH_2O and CH_3OH concentrations in the product, g_6 is yield, g_7 is production cost.

The goal number	Perfomance index	Γ_{Ud}
1	$-g_6[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] = \min\limits_{\Gamma_{Ud}}$	conditions 1.-5., 7.
2	$g_7[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] = \min\limits_{\Gamma_{Ud}}$	conditions 1.-6.
3	$-g_8 = -g_6(g_7 - \lambda) = \min\limits_{\Gamma_{Ud}}$	conditions 1.-7.
4	$g_5[U_n, F(U_n, \hat{V}_n, \hat{\theta}_n, n)] = \min\limits_{\Gamma_{Ud}}$	conditions 1.-4., 6.-7.

where g_8 is profit, λ is product price.

Control system realization. Data accumulation rate is low - 16 aweraged quantities per shift (8 hours), therefore the IS computation may be performed in an off-line mode.

The present IS algorithm enables to control several indepedent formaldehyd plants. Plants are connected with the computing center by telegraph lines. In a plant data are booked into special forms, punched and transmitted to the computer. They contain three different data sets (B1, B2, B3) and plant number label together with computation code. The latter shows what kind of information the D-M wants to acquire from computer. B1 consists of X_n, Z_n for the identification purposes, B2 presents constraint vectors and constructive parameters, B3 consist of vector U_k to be tested on the model. Information from B2 is stored and used until a new arrives. The results are punched by computer in the following forms (Fig. 4-6):

1. The means of the plant variables, Fig. 4a;

2. Estimated parameters, Fig. 4b;

3. Model outputs Y_k at the state $\hat{\theta}_n$ and $\hat{V}_n$ for decision U_k, Fig. 5;

4. Optimal decisions corresponding to goals 1.-4. and the plant output for those decisions, Fig. 6.

The algorithm is realized on a Minsk-22 computer. The program consists of $6700_{(10)}$ words in core memory and magnetic tape of 8192 + <number of plants>. 1024 words.

REFERENCE

1. Beauchamp,J.J., Cornell,R.G. Simultaneous nonlinear estimation. Technometrics, V. 8, nr. 2, May, 1966.
2. Андерсон,Т. Введение в многомерный статистический анализ. М. 1963.
3. Арис,Р. Анализ процессов в химических реакторах. Л. 1967.
4. Цыпкин,Я.З. Адаптация, обучение и самообучение в автоматических системах. Автоматика и телемеханика, т. XXYII, № I, 1966.
5. Albert,A.E., Gardner,L.A. Stochastic approximation and nonlinear regression. MIT Press, Cambridge, Mass., 1967.
6. Zhivoglyadov,V.P., Kaipov,V.Kh. Identification of distributed parameter plants in the presence of noises. Prepr. of the IFAC symp. Identification on automatic control systems, Prague, 1967.
7. Руубель,Х.В. Поиск точки минимума функции В сб. программы для ЦВМ "Минск-22", вып. 7 (to be appear).

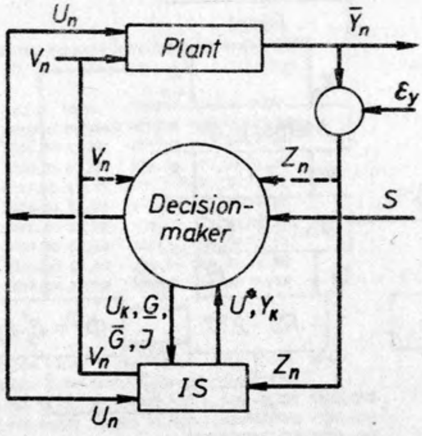

U_n
Plant
$\bar{Y}_n$
V_n
ε_y
V_n
Z_n
Decision-maker
S
$U_K, \underline{G}, \bar{G}, J$
U^*, Y_K
V_n
IS
Z_n
U_n

Fig. 1.

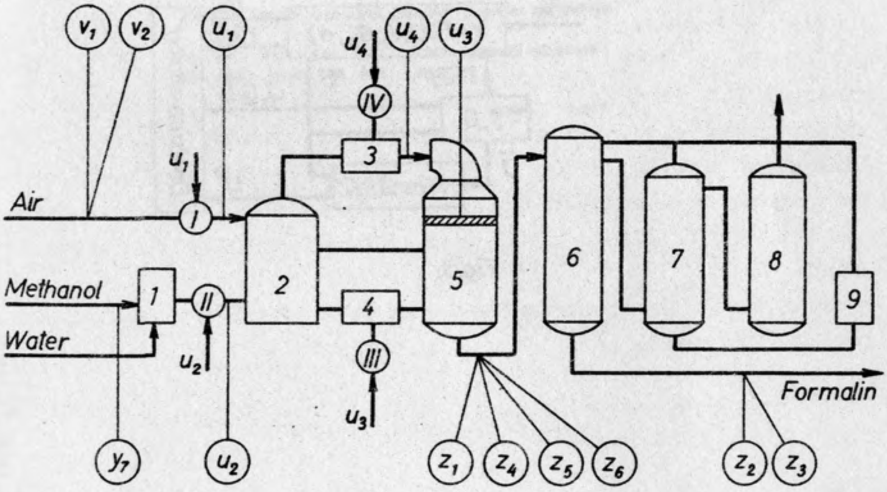
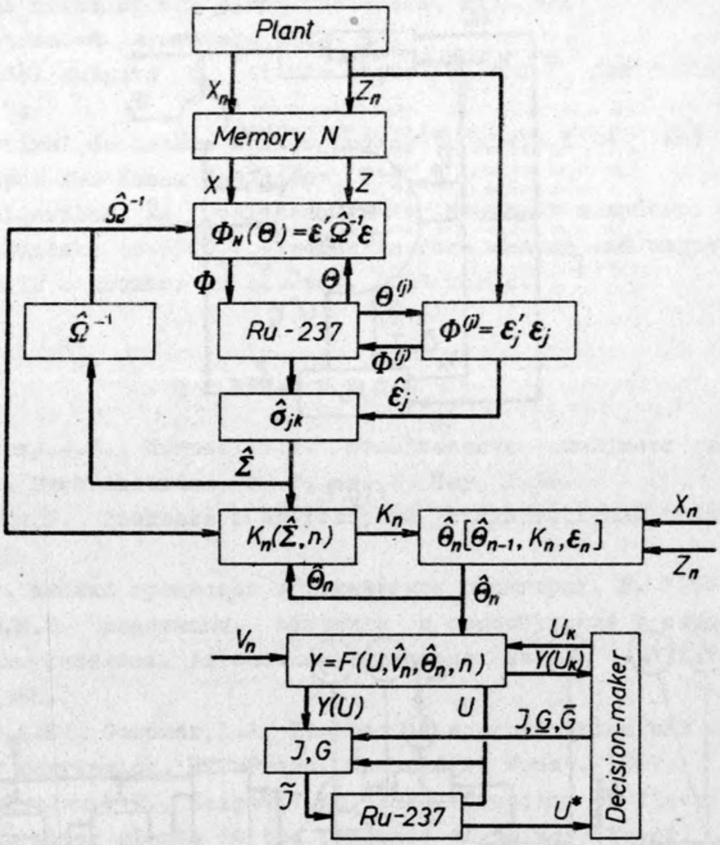

v_1
v_2
u_1
u_4
u_4
u_3
IV
3
u_1
Air
I
2
5
6
7
8
Methanol
1
II
4
9
Water
u_2
III
u_3
Formalin
y_7
u_2
z_1
z_4
z_5
z_6
z_2
z_3

Fig. 2.



Plant
X_n
Z_n
Memory N
X
Z
$\hat{\Omega}^{-1}$
$\Phi_N(\Theta) = \varepsilon' \hat{\Omega}^{-1} \varepsilon$
Φ
Θ
$\Theta^{(j)}$
$\hat{\Omega}^{-1}$
Ru - 237
$\Phi^{(j)} = \varepsilon_j' \varepsilon_j$
$\Phi^{(j)}$
$\hat{\varepsilon}_j$
$\hat{\sigma}_{jk}$
$\hat{\Sigma}$
K_n
$K_n(\hat{\Sigma}, n)$
$\hat{\theta}_n[\hat{\theta}_{n-1}, K_n, \varepsilon_n]$
X_n
Z_n
$\hat{\theta}_n$
$\hat{\theta}_n$
V_n
U_K
$Y = F(U, \hat{V}_n, \hat{\theta}_n, n)$
$Y(U_K)$
Y(U)
U
$J, \underline{G}, \bar{G}$
J, G
$\tilde{J}$
Ru-237
U^*
Decision-maker

Fig.3.

СВОДКА РЕЖИМОВ И ВЫХОДОВ АГРЕГАТА НОМЕР 2

ЧИСЛО МЕСЯЦ	НОМЕР СМЕНЫ	РАСХОД ВОЗДУХА	КОНЦ МЕТАНОЛ	ТЕМПЕР КОНТАК	ТЕМПЕР ПЕРЕГР	УСЛОВ МЕТАНОЛ
4,09	82,00	2050,00	63,40	695,00	112,00	17,42
5,09	85,00	2050,00	63,70	695,00	113,00	17,38
6,09	88,00	2050,00	63,30	695,00	112,00	17,66
7,09	91,00	2050,00	63,60	694,00	113,00	17,32
8,09	94,00	2050,00	63,50	695,00	111,00	16,59
9,09	97,00	2050,00	63,50	694,00	110,00	17,04
10,09	100,00	2050,00	63,50	695,00	110,00	16,08
11,09	103,00	2050,00	63,90	694,00	110,00	14,92

ВЫХОД ФОРМ	СЕЛЕК ТИВН	ОБЩАЯ КОНВ	РАСХ КОЭФФ	УСЛОВ ФОР-НА	ПРОЦ ФОР-ДА	ПРОЦ МЕТАН	ЭКОН МЕТАН	ДОХОД
0,756	0,883	0,856	0,523	30,387	37,400	6,600	-0,091	2513,010
0,755	0,879	0,860	0,523	30,293	37,400	6,400	-0,105	2503,836
0,755	0,885	0,853	0,525	30,674	37,500	6,700	-0,159	2529,996
0,748	0,883	0,846	0,529	29,864	37,400	7,100	-0,269	2451,801
0,744	0,874	0,852	0,531	28,477	37,400	6,800	-0,323	2331,244
0,747	0,873	0,855	0,529	29,384	37,400	7,300	-0,266	2412,255
0,745	0,870	0,856	0,526	27,888	37,100	7,200	-0,173	2297,420
0,740	0,864	0,856	0,533	25,554	37,300	7,100	-0,326	2088,409

СОСТОЯНИЕ ПРОЦЕССА

ПРЕДЬЭКСПОНЕНТЫ СКОРОСТЕЙ ПОТЕРИ СТЕПЕНИ ПОГЛОЩЕНИЯ

K00	K10	K20	K30	K40	ТЕПЛА	ФОРМ-ДА	МЕТАНОЛА
3037	4309	2476	185	593	55000		
						1,000	0,917

Figure 4

ИСПЫТАНИЕ ЗАДАННЫХ РЕЖИМОВ

1,	2050,0	63,0	695,0	113,0
2,	2050,0	63,5	695,0	114,0
3,	2100,0	64,0	635,0	110,0

	1	2	3	0	0	0	0	0
УСЛОВ	31,0	31,0	32,7	0,0	0,0	0,0	0,0	0,0
РАСХ КОЭФФ	0,512	0,517	0,609	0,000	0,000	0,000	0,000	0,000
ДОХОД	2597	2582	2422	0	0	0	0	0
СОДЕРЖ МЕТАН	6,4	6,8	16,6	0,0	0,0	0,0	0,0	0,0
СОДЕРЖ ФОРМ	49,2	49,1	40,6	0,0	0,0	0,0	0,0	0,0
ЭКОНОМИЯ	0,24	0,09	-2,90	0,00	0,00	0,00	0,00	0,00
ВЫХОД ФОРМ	0,767	0,760	0,645	0,000	0,000	0,000	0,000	0,000
СЕЛЕКТИВН	0,887	0,887	0,914	0,000	0,000	0,000	0,000	0,000
ГАЗО О2	0,20	0,20	0,24	0,00	0,00	0,00	0,00	0,00
ВЫЙ СО2	3,92	3,92	4,02	0,00	0,00	0,00	0,00	0,00
АНА СО	1,60	1,60	0,30	0,00	0,00	0,00	0,00	0,00
ЛИЗ Н2	18,0	18,0	17,1	0,0	0,0	0,0	0,0	0,0
КОНЦ О2	0,096	0,096	0,087	0,000	0,000	0,000	0,000	0,000
КОНЦ СН3ОН	0,266	0,269	0,294	0,000	0,000	0,000	0,000	

Figure 5

РЕКОМЕНДУЕМЫЕ РЕЖИМЫ

1 МАКС ВЫРАБОТКА	2050,0	61,0	692,0	129,0
2 МИН РАСХ КОЭФФ	2050,0	63,5	712,0	125,0
3 МАКС ДОХОД	2050,0	61,5	696,0	120,0
4 МИН СОДЕРЖ МЕТ	2050,0	64,0	718,0	115,0

РЕЗУЛЬТАТ ОПТИМАЛЬНЫХ РЕЖИМОВ

	1	2	3	4	0	0	0	0
УСЛОВ	31,1	30,4	31,0	30,1	0,0	0,0	0,0	0,0
РАСХ КОЭФФ	0,519	0,506	0,508	0,491	0,000	0,000	0,000	0,000
ДОХОД	2587	2568	2612	2587	0	0	0	0
СОДЕРЖ МЕТАН	7,2	5,0	6,0	3,2	0,0	0,0	0,0	0,0
СОДЕРЖ ФОРМ	46,6	51,2	48,4	53,9	0,0	0,0	0,0	0,0
ЭКОНОМИЯ	0,02	0,43	0,39	0,88	0,00	0,00	0,00	0,00
ВЫХОД ФОРМ	0,757	0,777	0,774	0,801	0,000	0,000	0,000	0,000
СЕЛЕКТИВН	0,891	0,870	0,887	0,861	0,000	0,000	0,000	0,000
ГАЗО O2	0,21	0,19	0,20	0,19	0,00	0,00	0,00	0,00
ВЫЙ CO2	3,93	3,86	3,92	3,82	0,00	0,00	0,00	0,00
АНА CO	1,41	2,37	1,60	2,80	0,00	0,00	0,00	0,00
ЛИЗ H2	17,8	18,5	18,0	18,8	0,0	0,0	0,0	0,0
КОНЦ O2	0,093	0,098	0,095	0,101	0,000	0,000	0,000	0,000
КОНЦ CH3OH	0,262	0,264	0,260	0,260	0,000	0,000	0,000	

Figure 6