

316.920

VOL. 12 • NUMBER 1
TOM 12 • НОМЕР 1

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
ПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1983

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H-1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 0BW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1983 Subscription Rate US Dollars 140.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

N. N. KRASOVSKII (USSR)
G. BOGNÁR (HUNGARY)

COORDINATING EDITORS

USSR

S. V. EMELYANOV
E. P. POPOV
V. S. PUGACHEV
V. I. SIFOROV
E. D. TERYAEV

HUNGARY

T. VÁMOS
L. VARGA
A. PRÉKOPA
S. CSIBI
I. CSISZÁR
L. KEVICZKY
J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ
V. STREJČ

РЕДАКТОРЫ ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)
Г. БОГНАР (ВНР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ
Е. П. ПОПОВ
В. С. ПУГАЧЕВ
В. И. СИФОРОВ
Е. Д. ТЕРЯЕВ

ВНР

Т. ВАМОШ
Л. ВАРГА
А. ПРЕКОПА
Ш. ЧИБИ
И. ЧИСАР
Л. КЕВИЦКИ
Я. КОЧИШ

ЧССР

И. БЕНЕШ
В. СТРЕЙЦ



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

CASCADE EQUAL-WEIGHT CODES AND MAXIMAL PACKINGS

V. A. ZINOVIEV

(Received April 30, 1982)

The general method of construction of equal-weight codes with given weight and minimal distance is presented. The power of such code differs from upper Johnson bound only by the multiplicative constant which does not depend on the length of the code. This constant tends to unity when some conditions between the weight and distance are valid.

1. Introduction

Denote by $A(n, d, w)$, $d = 2\delta$, the maximal possible power of binary equal-weight code of length n , Hamming distance d and weight of the code words w . This value is studied also in combinatorics under the name maximal packing $m(n, w, k)$, $1 \leq k \leq w \leq n$ [1–5]. Define this value. Let E^n be the set of all the binary vectors of length n and E_w^n be the subset of E^n containing all such vectors of weight w . Say that the vector $x = (x_1, \dots, x_n)$, $x \in E_w^n$, covers the vector $y = (y_1, \dots, y_n)$, $y \in E_k^n$, $k \leq w$, if the equality $x_i \cdot y_i = y_i$ is valid for all $i = 1, \dots, n$. Define

$$m(n, w, k) = \max_V |V| : V \subseteq E_w^n$$

and any vector y , $y \in E_k^n$, is covered at most by one vector x , $x \in V$. The problem consists in studying of the value $m(n, w, k)$ as a function of parameters n, w, k .

Firstly, let us note (it is known rather well) that the following equality is valid

$$m(n, w, k) = A(n, 2(w - k + 1), w). \quad (1.1)$$

It is necessary to note that in combinatorics this value is studied mainly for fixed (or slowly growing) parameters w and k and for growing n , but in the coding theory (see, for example, references in [6]) this value is considered mainly for the cases, when w and k depend linearly on n , when n grows to infinity. Nevertheless, the known results of error correcting codes allow us to obtain bounds for the value $m(n, w, k)$ for fixed w and k which had been shown in [5].

From the well-known Johnson upper bound [7]

$$m(n, w, k) \leq \binom{n}{k} / \binom{w}{k}. \quad (1.2)$$

Let us define

$$\mu(n, w, k) = m(n, w, k) \binom{w}{k} / \binom{n}{k}.$$

Erdős and Hanani [1] suggested that for fixed w and k

$$\lim_{n \rightarrow \infty} \mu(n, w, k) = 1$$

and they proved it for the case $k=2$ and all w and for the case $k=3$ and $w=q$ or $w=q+1$, where q is a prime power. Kuzurin [3] proved this conjecture for the case when $k=w-1$. Furthermore, he proved it not only for fixed w but for $w=O(n)$ (he showed that in these cases usual limits also exist).

In [5] it was shown that

$$\overline{\lim}_{n \rightarrow \infty} \mu(n, w, w-2) = 1, \quad \text{when } w=O(n)$$

and for $k \leq w-3$ when $w-k$ is fixed

$$\overline{\lim}_{n \rightarrow \infty} \mu(n, w, k) \geq 1/(w-k)!, \quad \text{when } w=O(n). \quad (1.3)$$

In terms of codes the result obtained in [5] means: for length $n=2^s-1$, $s=2, 3, \dots$,

$$A(n, 2\delta, w) \geq \binom{n}{w} / (n+1)^{\delta-1}. \quad (1.4)$$

In [8], Graham and Sloane gave several constructions for equal-weight codes, using some mapping from E_w^n to Galois fields $GF(q)$. In particular they proved, that if q is a prime power such that $q \geq n$ then

$$A(n, 2\delta, w) \geq \binom{n}{w} / q^{\delta-1}. \quad (1.5)$$

Note that bound (1.4) is coincident with (1.5) for the case $q=n+1=2^s$ and they are good for the small δ . It is possible to see from the following asymptotic expression of upper bound (1.2) and lower bounds (1.4) and (1.5) that, when w is fixed and $n \rightarrow \infty$:

$$n^{w-\delta+1}/w! \lesssim A(n, 2\delta, w) \lesssim n^{w-\delta+1}(\delta-1)!/w!. \quad (1.6)$$

We need also the result of Kuzurin [4]. Let $q=q(a)$ denote maximal prime power which is not more than a . From the number theory it is known [9] that for every $\varepsilon > 0$ it is possible to find a_ε such as for any $a > a_\varepsilon$ the following inequality is valid:

$$|q(a) - a| \leq a^{7/12 + \varepsilon}.$$

Kuzurin proved the following result. Let $w = w(n)$ and $k = k(n)$ of positive integers have the properties:

(1) $w(n) \rightarrow \infty$, when $n \rightarrow \infty$ and constant $c, c > 7/12$, exists such that for enough large n the following inequality is valid

$$w^2 + n^c - w^{1-c} - (n-w) < 0;$$

(2) $k(n)/\sqrt{w(n)} \rightarrow 0$ when $n \rightarrow \infty$;

(3) $k(n)/(n/w(n))^{1-c} \rightarrow 0$ when $n \rightarrow \infty$.

Then

$$\lim_{n \rightarrow \infty} m(n, w, k)w^k/n^k = 1. \quad (1.7)$$

Let us emphasize that all lower bounds (1.4)–(1.6) and Kuzurin's theorem are existence theorems stating that to construct corresponding codes it is necessary to make an overall choice over all possible vectors of E_w^n .

The aim of this paper is to present a general and quite simple method of constructing of equal-weight codes with given weight and minimal distance. The power of such code differs from upper Johnson bound (1.2) only in a multiplicative constant, which does not depend from the length of the code. Unlike the codes satisfying lower bounds (1.4) and (1.5), this constant grows to unity when the weight of the code words grows and some conditions between length, weight and distance are valid. On the other side the codes and equivalent maximal packings obtained are good for some finite lengths. It is also essential that this method is constructive in such a sense that for the construction of code we have no overall choice in the set E_w^n and the complexity of the construction of the code is not more than cn^3 binary operations, where the constant c does not depend on n , w , and δ , and the binary operation is the arithmetic operation in GF(2). In particular, the result of Kuzurin mentioned above is obtained constructively.

The results of this paper have been partially published in [10] without proofs and presented at the International Symposium of Information Theory (Oberwolfach, FRG, 4–10 April, 1982). The author sincerely thanks V. I. Levenshtein and L. A. Bassalygo for their useful comments which helped much to improve the present paper.

2. The main construction

Theorem 1. Let q is a prime power such that $q + 1 \geq w$. Then for any $\delta, 1 \leq \delta \leq w$ and $n = qw$

$$A(n, 2\delta, w) \geq (n/w)^{w-\delta+1}; \quad (2.1)$$

bound (2.1) is better than bounds (1.4), (1.5) when $\delta > 1/2 + w/\ln w$ and the complexity of the construction of such a code is not more than cn^3 binary operations, where c does not depend on n, w, δ .

Proof. Let w be given, δ be any integer, $1 \leq \delta \leq w$, and q be prime power such that $q + 1 \geq w$. Consider MDS code (see [6]) R over Galya field $GF(q)$ with the following parameters: the length $n' = w$, the number of information symbol $k = w - \delta + 1$, the code (Hamming) distance $d' = n' - k + 1 = \delta$. The code R with parameters n', k, d' is denote by $R = R(n', k, d')$. Now, construct the cascade constant weight code [11] using the code R as outer code and the trivial constant weight code whose code words form the identity matrix I_q of order q as inner code. In other words, we perform the following transformation over all words of code R . Denote by $\alpha_1, \alpha_2, \dots, \alpha_q$ the elements of $GF(q)$ ordered in some fixed manner. In the code word $(a_1, \dots, a_w) \in R$ we replace each element α_i by the i -th row of matrix I_q (or by bynary vector of length q and weight 1, where the unit is in the position with number i). It is clear that the resultant binary code (denote this code by C) has length $n = n'q = wq$, each code word has weight w , the distance between any two different words is not less then $2d' = 2\delta$ and the power of the code is equal $q^k = q^{w-\delta+1}$. It corresponds to lower bound (2.1). Compare values (1.4), (1.5) and (2.1). We have

$$\binom{n}{w} / q^{\delta-1} < n^{w-\delta+1} / w!,$$

when $w > 1$. The condition

$$(n/w)^{w-\delta+1} > n^{w-\delta+1} / w!$$

is equivalent to

$$w! > w^{w-\delta+1},$$

and is valid, when $\delta > 1/2 + w/\ln w$. The value of complexity of the construction follows immediatly from [12]. According to our terms, this complexity is not more than cn^3 . The theorem is proved.

Theorem 2. Let $q = q_1 \dots q_s$, where for each $i, i = 1, \dots, s, q_i$ is a prime power such that $q_i + 1 \geq w$. Then for every $\delta, 1 \leq \delta \leq w$, and $n = qw$, inequality (2.1) is valid.

Proof. Consider for each $i, i = 1, \dots, s$, the MDS code $R_i = R_i(w, k, \delta)$, $k = w + 1 - \delta$, over $GF(q_i)$. Then the direct product of codes R_1, \dots, R_s is MDS code $R = R(w, k, \delta)$ over the alphabet of size $q = q_1 \dots q_s$ and further considerations are similar to the proof of theorem 1.

In terms of maximal packings, Theorems 1 and 2 can be formulated in the following manner. For any w and $k, 1 \leq k \leq w$, and for suitable $n, n = qw$, where q

satisfies the conditions of Theorem 1 or 2,

$$\mu(n, w, k) \geq \frac{w-1}{w} \cdot \frac{w-2}{w} \cdot \dots \cdot \frac{w-k+1}{w}. \quad (2.2)$$

In the next paragraph we shall get rid of the discreteness of bounds (2.1) and (2.2) on n and obtain lower bounds for $A(n, 2\delta, w)$ and $m(n, w, k)$ which conform to n .

3. Modification of the main construction

In this paragraph, for given w let the number q satisfy the conditions of theorem 1 and k, δ are any numbers, $1 \leq k, \delta \leq w, k + \delta = w + 1$. The following simple lemmas yield the values similar to (2.1) for any length n .

Lemma 1. Let $n = qw + \gamma$. Then

$$A(n, 2\delta, w) \geq \left(\frac{n}{w}\right)^k \left(1 - \frac{\gamma}{n}\right)^k, \quad k = w + 1 - \delta. \quad (3.1)$$

Proof. For $n' = qw$ let us construct a cascade constant weight code satisfying (2.1), using Theorem 1. Addition to this code γ zero positions gives the value (3.1).

Lemma 2. Let $n = qw - \gamma$, where $\gamma < k(q-1)$ and $\gamma = kr + t$, where $0 \leq t < k$. Then

$$A(n, 2\delta, w) \geq \left(\frac{n}{w}\right)^k \left(1 + \frac{\gamma}{n}\right)^k \left(1 - \frac{r}{q}\right)^{k-t} \left(1 - \frac{r+1}{q}\right)^k. \quad (3.2)$$

Proof. Consider MDS code $R = R(w, k, \delta)$ over $\text{GF}(q)$. Let $\alpha_1, \dots, \alpha_q$ denote elements of $\text{GF}(q)$. Fix the first k positions of the code R . In the first $k-t$ positions fix $q-r$ elements $\alpha_1, \dots, \alpha_{q-r}$ and in the following t positions fix $q-r-1$ elements $\alpha_1, \dots, \alpha_{q-r-1}$. Form the new code (denote it by $R(\gamma)$) of the same length w take as code vectors all such words from R which have in the first fixed k positions only the fixed elements. As any k positions of MDS code $R(w, k, \delta)$ (see [6]) contain each vector of length k over $\text{GF}(q)$ exactly once, the power of code $R(\gamma)$ is equal to $(q-r)^{k-t}(q-r-1)^t$. Conversion of the code $R(\gamma)$ into cascade constant weight code, using the proof of Theorem 1 (considering only that the first k positions of the code $R(\gamma)$ have the alphabets of sizes $q-r$ and $q-r-1$), gives value (3.2).

Lemma 3. Let $n = vw - \gamma = (q(v) - \Delta)w - \gamma$, where $\gamma < w$ and $q(v)$ is the minimum prime power such that $q(v) \geq v$. Then

$$A(n, 2\delta, w) \geq \left(\frac{n}{w}\right)^k \left(1 - \frac{\Delta}{q(v)}\right)^{\delta-1} \left(1 - \frac{1}{q(v)-\Delta}\right)^\gamma. \quad (3.3)$$

Proof. Let $n = vw - \gamma = (q(v) - \Delta)w - \gamma$ and let $q = q(v)$. Consider the partition of space all q^w vectors of length w over $\text{GF}(q)$ on q^{w-k} cosets of MDS code $R = R(w, k, \delta)$. In the first γ positions of such vectors fix $q - \Delta - 1$ elements of $\text{GF}(q)$ and in last $w - \gamma$ positions of such vectors fix $q - \Delta$ elements of $\text{GF}(q)$. In every coset let's consider all vectors over fixed elements of $\text{GF}(q)$. As there are $(q - \Delta - 1)^\gamma (q - \Delta)^{w - \gamma}$ such vectors in all and they are distributed in q^{w-k} different cosets then there exists wittingly the code (denote it by $R(\Delta)$) of length w with code distance δ and of power $(q - \Delta - 1)^\gamma (q - \Delta)^{w - \gamma} / q^{w-k}$. This code $R(\Delta)$ has the alphabet of the size $q - \Delta - 1$ in the first γ positions and the alphabet of the size $q - \Delta$ in the last $w - \gamma$ positions. Conversion of the code $R(\Delta)$ into cascade code gives the value (3.3).

4. Asymptotic bounds of $A(n, 2\delta, w)$

In this paragraph, let $w = w(n)$, $k = k(n)$ denote the sequences of positive integers, which are growing if n grows. It is clear from (3.1) that if $\gamma k = o(n)$ when $n \rightarrow \infty$ then the lower bound $A(n, 2\delta, w)$ has the order $(n/w)^k$, $k = w + 1 - \delta$. Let us estimate at growth of γ in Lemma 1. Theorem 1 is applied, when $w(n)$ grows not more than $\sqrt{n/2}$. It means that the number v , $v = \lfloor n/w \rfloor$, grows when n grows. It is known [9] that for any $\varepsilon > 0$ as small as possible it is possible to find v_ε such that for any $v > v_\varepsilon$ the following inequality is valid: $|q(v) - v| \leq q(v)^{7/12 + \varepsilon}$ where $q(v)$ is the closest to v prime power such that $q(v) \geq v$. So for n , $n \leq v_\varepsilon w$ large enough, the number γ in Lemma 1 satisfies the inequality $\gamma \geq w \cdot q(v)^{7/12 + \varepsilon}$, and the number Δ in Lemma 3 satisfies the inequality $\Delta < q(v)^{7/12 + \varepsilon}$. Thus from this considerations and Lemmas 1 and 3 we have the following results.

Lemma 4. Let (1) $w(n) \leq \sqrt{n/2}$; (2) $k(n) = o\left(\left(\frac{n}{w(n)}\right)^{5/12 - \varepsilon}\right)$.

Then

$$\lim_{n \rightarrow \infty} A(n, 2\delta, w) \left(\frac{w}{n}\right)^k \geq 1, \quad \delta = w + 1 - k, \quad (4.1)$$

moreover the complexity of the construction for the length n is not more then cn^3 binary operations.

Lemma 5. Let (1) $w(n) = O(\sqrt{n})$; (2) $w(n) - k(n) = o\left(\left(\frac{n}{w(n)}\right)^{5/12 - \varepsilon}\right)$.

Then the inequality (4.1) is valid.

Let us consider the asymptotic of the upper Johnson bound (1.2). From (1.2) we have

$$A(n, 2\delta, w) \leq \left(\frac{n}{w}\right)^k \frac{w}{w-1} \cdot \frac{w}{w-2} \cdot \dots \cdot \frac{w}{w-k+1},$$

$$k = w + 1 - \delta. \quad (4.2)$$

So the following result is valid.

Lemma 6. Let $k(n) = O(\sqrt{w(n)})$. Then

$$\overline{\lim}_{n \rightarrow \infty} A(n, 2\delta, w) \left(\frac{w}{n} \right)^k \leq 1, \quad k = w + 1 - \delta. \quad (4.3)$$

From Lemmas 4 and 6 we have the following result, which is coincident with the result of Kuzurin [4] and which is its constructive analog.

Theorem 3. Let

- (1) $w(n) \rightarrow \infty$ if $n \rightarrow \infty$ moreover $w(n) \leq \sqrt{n/2}$;
- (2) $k(n) = O(\sqrt{w(n)})$;
- (3) $k(n) = O((n/w(n))^{5/12 - \epsilon})$. Then

$$\lim_{\rightarrow \infty} A(n, 2\delta, w) \left(\frac{w}{n} \right)^k = 1, \quad k = w + 1 - \delta, \quad (4.4)$$

moreover the complexity of the code construction for length n is not more than cn^3 binary operations.

References

1. Erdős, P., Hanani, H., On a limit theorem in combinatorial analysis. Publ. Math. Debrecen, 1963, **10**, 1, pp. 10–13.
2. Erdős, P., Spencer, J., Probabilistic methods in combinatorics. Akadémiai Kiadó, Budapest, 1974.
3. Kuzurin, N. N., On minimal coverings and maximal packings of $(k-1)$ -tuples by k -tuples. Matemat. Zametki, 1977, **21**, 4, pp. 565–571.
4. Kuzurin, N. N., Asymptotic investigation of covering problem. Problemy kibernetiki, 1980, **37**, 19–56.
5. Bassalygo, L. A., Zinoviev, V. A., Some values for combinatorial problems of packings and coverings. International Colloquium on Information Theory (August 24–28, 1981, Budapest, Hungary), Abstracts, pp. 54–55.
6. MacWilliams, F. J., Sloane, N. J. A., The theory of error-correcting codes. North-Holland Publishing Company, Amsterdam–New York–Oxford, 1977.
7. Johnson, S. M., A new upper bound for error-correcting codes. IRE Trans. Inform. Theory, 1962, **8**, 1, pp. 203–207.
8. Graham, R. L., Sloane, N. J. A., Lower bounds for constant weight codes. IEEE Trans. Inform. Theory, 1980, **26**, 1, pp. 37–43.
9. Huxley, M. N., On the difference between consecutive primes. Invent. Math., 1972, **15**, 1, pp. 164–170.
10. Zinoviev, V. A., Constant weight codes and maximal packings. In: VIIIth All-Union Conf. on Coding Theory and Inform. Trans. Abstracts of papers, part II, Moscow–Kuibyshev, 1981, pp. 75–80.
11. Zinoviev, V. A., Generalized concatenated codes. Problems of Inf. Trans., 1976, **12**, 1, pp. 5–15.
12. Zyablov, V. V., An estimate of the complexity of constructing binary linear cascade codes, Problems of Inf. Trans., 1971, **7**, 1, pp. 3–10.
13. Delsarte, P., On subfield subcodes of Reed-Solomon codes. IEEE Trans. Inform. Theory, 1975, **21**, pp. 575–576.

Каскадные равновесные коды и максимальные упаковки

В. А. ЗИНОВЬЕВ

(Москва)

В работе предложен каскадный метод построения широкого класса двоичных равновесных кодов или эквивалентных им максимальных упаковок. Мощность такого кода или соответствующей максимальной упаковки лишь на мультипликативную константу отличается от верхней границы Джонсона. Когда длина кода неограниченно возрастает, эта константа стремится к единице при соответствующих ограничениях на рост веса и расстояния.

В. А. Зиновьев

Институт проблем передачи информации АН СССР

СССР, 101447, Москва, ГСП-4

ул. Ермоловой, 19.

CONTROL OF NONSTATIONARY DYNAMIC SYSTEMS WITH QUASICONTINUOUS GENERATION OF THE CONTROL SIGNAL

S. V. YEMELYANOV, S. K. KOROVIN, B. V. ULANOV

(Moscow)

(Received February 5, 1982)

The paper is concerned with the control of nonstationary linear dynamic systems. The proposed control algorithm employs coordinate parametric and parametric feedbacks. The proposed relations lead to systems featuring the desired properties. Examples are given.

1. Introduction

Several algorithms for the control of dynamic systems where coordinate, coordinate parametric, and parametric feedback loops are used [1]. These algorithms either make the dynamic properties of the processes little dependent on the dynamic system parameters which vary within any known limited range, the dynamic system being described by one differential equation, or, generally speaking, limit the effect of variable system parameters on the dynamic properties of the processes and change these properties in the desired direction.

When the dynamic system is described by a set of differential equations of a general form which is linear in state coordinates and control, these effects are achieved without using sliding modes in the coordinate feedback loop and with finite gains in that loop. Mathematically speaking, these algorithms expand the phase space (viz. increase the dimensionality by a unity through addition of a new, parametric coordinate) of the dynamic system and so a closed-loop dynamic system (control system) in the expanded phase space is described by a set of common differential equation with a discontinuous right-hand side.

When some algorithms of [1] are used, the right-hand side of the resultant set of differential equations which describe a closed-loop dynamic system has a discontinuity along each solution of the set when the solution goes through a certain manifold and the parametric coordinate reaches certain boundary values.

Other algorithms of [1] in a closed loop lead, starting with a certain time, to a sliding mode or motion along a set on which the right-hand side of the set of differential equations undergoes discontinuities; the sliding mode starts in the coordinate-

parametric feedback loop and in actual control system an infinite number of relay-like switchings occurs in that loop.

In this paper we will design a control system which incorporates coordinate, coordinate-parametric, and parametric feedback loop; a new, parametric coordinate will be introduced to generate a continuous control signal and each solution of the resultant set of differential equations belongs to the discontinuity set of the right-hand side in the equation for variation of the parametric coordinate over a maximum of one time interval until the describing point of the control coordinates (or the error coordinate and its derivatives) reaches some vicinity of zero in the state space of the dynamic system to be controlled; then the solution can reach that set while the describing point stays in some larger vicinity of zero, which, as will be seen from the relations to be given below, can be made as small as desired by a proper choice of some parameter in the coordinate-parametric feedback loop; the equation whereby control signals are generated will be referred to as quasilinear. This control system either makes the dynamic properties of the processes little dependent on the parameters of the system to be controlled which vary within any known range or, in a general case, insures that the dynamic properties of the processes change in the desired direction.

2. Equations of a closed loop dynamic system and statement of the problem

The dynamic system is described by a differential equation (in the vector matrix form)

$$\dot{x} = A(t)x + b(t)u, \quad t \geq t_0, \quad (1)$$

where t_0 is the initial time; $x = (x_1, \dots, x_n)^T \in \mathbf{R}^n$ (hereafter T is the transposition symbol); at each fixed time t : $A(t) = (a_{ij}(t))$ is an $(n \times n)$ matrix; $b(t) = (b_1(t), \dots, b_n(t))^T \in \mathbf{R}^n$ and u is a scalar control.

It is assumed that $a_{ij}(t)$ $i, j = 1, \dots, n$ and $b_i(t)$ ($i = 1, \dots, n$) are functions on $[t_0, \infty)$, measurable for the Lebesgue measure and such that almost everywhere on $[t_0, \infty)$ the relations hold

$$\begin{aligned} a_{ij}^- \leq a_{ij}(t) \leq a_{ij}^+, \quad i, j = 1, \dots, n, \\ b_i^- \leq b_i(t) \leq b_i^+, \quad i = 1, \dots, n, \end{aligned} \quad (2)$$

where a_{ij}^- , a_{ij}^+ ($i, j = 1, \dots, n$) and b_i^- , b_i^+ ($i = 1, \dots, n$) are known constants.

The initial conditions $x(t_0) = x_0$ and control $u = u(t)$ or $u = u(x, t)$ with $t \geq t_0$ defines the solution of equation (1), $x(t)$ being the process.

It is required to obtain a control u which would solve the control problem (under any initial conditions $x(t) \rightarrow 0$ as $t \rightarrow \infty$) and constrain the dependence of properties of

equation (1) on the parameters $a_{ij}(t)$ ($i, j = 1, \dots, n$) and $b_i(t)$ ($i = 1, \dots, n$) which vary within (2). The control u should be generated continuously until, for instance, $x(t)$ reaches a certain vicinity of zero in the space of x 's which is denoted as \mathbf{R}_x^n .

For the problem to be solved the control should be generated in the following way

$$u = (\kappa^0 \mu, m\bar{x}), \tag{3}$$

$$\dot{\mu} = \begin{cases} -\alpha(\varepsilon(x) + \mu\delta\|x'\|) & \text{with } |\mu| \leq 1, \\ 0 & \text{with } |\mu| > 1, |\mu(t_0)| \leq 1, \end{cases} \tag{4}$$

where $\mu \in \mathbf{R}$, $m\bar{x} = (|x_1|, \dots, |x_n|)^T \in \mathbf{R}^l$ ($l = n-1$ or n below); $\|x'\| = \sum_{i=1}^{n-1} |x_i|$; $\kappa^0 = (\kappa_1^0, \dots, \kappa_l^0)^T \in \mathbf{R}^l$, $\alpha > 0$, and $\delta > 0$ are some constant vector and numbers; in (3) and hereafter (\cdot, \cdot) is a scalar product: $(v, w) = \sum_{i=1}^S v_i w_i$ for the vectors $v = (v_1, \dots, v_S)^T \in \mathbf{R}^S$ and $w = (w_1, \dots, w_S)^T \in \mathbf{R}^S$; $\varepsilon(x) = (c, x)$ where $c = (c_1, \dots, c_n)^T \in \mathbf{R}^n$ is some constant vector.

Equations (1) and (4) and relations (3) and (4) define a dynamic $(n+1)$ -st order system which will be denoted as (S). We will be concerned with the properties of the solutions to the system (S) for which the initial values $(x(t_0), \mu(t_0)) \in V = \{x, \mu : |\mu| \leq 1\}$; such solutions are referred to as V -solutions.

Equations (3) and (4), whereby the control u is generated, are activated when the control system employs loops of coordinate-parametric (for determining the parametric coordinate μ in equation (4)) and parametric (for determining the parameter v in the setpoint equation $g_2(t) = \varepsilon(x(t)) + v\|x'(t)\|$ of the coordinate-parametric loop $v = \delta\mu$) feedbacks (for a structural diagram, see Fig. 1). The equation

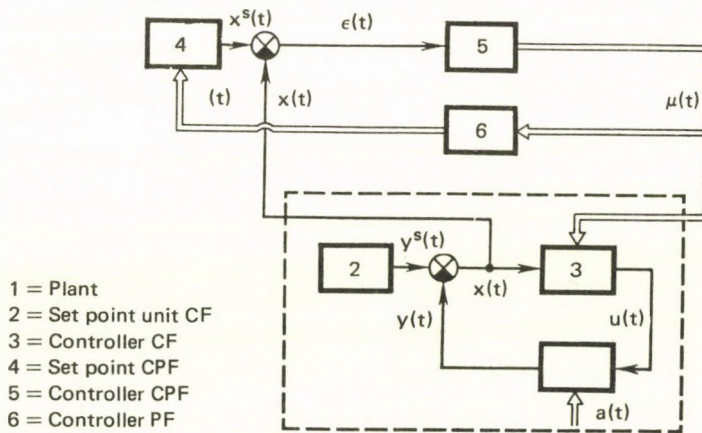


Fig. 1

for μ (4) is chosen so that, as will be shown below, if at a certain time for a V -solution of the system (S) $(x(t), \mu(t))$ with $x(t)$ outside a certain vicinity of zero in \mathbf{R}_x^n the inequality holds

$$|\varepsilon(x(t)) + \mu(t)\delta \|x'(t)\| \leq \frac{\delta}{2} \|x'(t)\|, \quad (5)$$

then (with certain relations between parameters of the system (S)) it holds also at all subsequent times until $x(t)$ reaches a specified vicinity of zero in \mathbf{R}_x^n . The right-hand side of equation (4) has a discontinuity on the sets $\{(x, \mu): \mu = -1\}$ and $\{(x, \mu): \mu = 1\}$; however, under certain conditions for each V -solution of the system (S) $(x(t), \mu(t))$ no derivative $\mu(t)$ exists (or the right-hand side of equation (4) has a discontinuity) at a maximum of one time following which the solution on a certain time interval belongs to one of these sets until $x(t)$ stays in some vicinity of zero of the space \mathbf{R}_x^n . Once this vicinity is reached, $x(t)$ stays at all subsequent times in some larger vicinity of zero in \mathbf{R}_x^n which may be made as small as desired by choosing the parameter α . Therefore the law whereby the control u is generated, i.e. (3) and (4), will be referred to as quasicontinuous.

For any V -solution of the system (S) $(x(t), \mu(t))$, it follows from (4) that $|\mu(t)| \leq 1$ and, consequently, with inequality (5) true for the V -solution of the system (S) the inclusion holds

$$x(t) \in G_{\frac{3}{2}\delta}^3 = \{x: |\varepsilon(x)| \leq \frac{3}{2} \delta \|x'\| \}, \quad (6)$$

and by choosing the parameters c_i ($i = 1, \dots, n$) and δ the dynamic properties of the vector function $x(t)$ can be influenced.

In this context the following problems are discussed: determining the conditions under which for V -solutions the inequality (5) can hold at a certain time; determining the conditions under which inequality (5) can be maintained over a certain time interval; determining V -solutions if (5) does not hold at any finite time; on feasibility of $x(t) \rightarrow 0$ as $t \rightarrow \infty$ for V -solutions if the inclusion (6) holds; finally, determining the behaviour of V -solutions of the system (S) when $x(t)$ reaches a certain vicinity of zero in \mathbf{R}_x^n .

The subsequent Section will be devoted to qualitative study of properties of V -solutions for the system (S).

3. Studying the behaviour of V -solutions in the system

The notation to be used hereafter is: $x' = (x_1, \dots, x_{n-1})^T$, $\|v\| = \sum_{i=1}^S |v_i|$ is the norm of the vector $v = (v_1, \dots, v_S)^T \in \mathbf{R}^S$, the norms of matrices coincide with those of vectors and are denoted as $\|\cdot\|$; $U_R = \{x: \|x\| < R\}$ and $B_R = \{x: \|x\| \leq R\}$ where R is a

certain positive number; $a^i(t)$ ($i = 1, \dots, n$) is the i -th column of the matrix $A(t)$. Assume that $c_n = 1$. Below the argument may be omitted in functions of time t .

Assume that for the controlled dynamic system it is true that $|(c, b(t))| \geq \text{const} > 0$ almost everywhere on $[t_0, \infty)$.

To solve the first of the problems formulated at the end of the preceding section, let us take up the equation

$$\dot{x} = A(t)x + b(t)(\kappa^0 \mu, m\bar{x}), \mu = -\text{sgn } \varepsilon(x), t \leq t_0 \quad (7)$$

(the notation is as for the system (S)) of the theory of variable structure systems [2, 3]. Solutions of equation (7) are said to hit the hyperplane $\varepsilon(x) = 0$ if for any solution $x(t)$ of (7) either there is a finite time t_1 such that $\varepsilon(x(t_1)) = 0$ or $\varepsilon(x(t)) \rightarrow 0$ as $t \rightarrow \infty$. The question whether solutions of equation (7) hit the hyperplane $\varepsilon(x) = 0$ is answered in the theory of variable structure systems [2, 3]. Solutions of equation (7) are said to feature O -hitting of the hyperplane $\varepsilon(x) = 0$ if for each solution $x(t)$ either there is time t_1 such that $\varepsilon(x(t_1)) = 0$ or $\|x(t)\| \rightarrow 0$ as $t \rightarrow \infty$.

It is easily found that for any solution of the system (S), i.e. a differential relation $(x(t), \mu(t))$ holds (almost everywhere on $[t_0, \infty)$)

$$\dot{x}'(t) = \bar{A}(t)x'(t) + h(t)\varepsilon(x(t)) + h^1(t)\dot{\varepsilon}(x(t)), \quad (8)$$

where

$$\bar{A} = A' - (a^n)'c'^T - \frac{1}{(c, b)}(b'c'^T A' - b'c'^T (a^n)'c'^T + b'a'_n - a_{nn}b'c'^T),$$

$$h = (a^n)' - \frac{1}{(c, b)}(b'c'^T (a^n)' + b'a_{nn}), \quad h^1 = \frac{b'}{(c, b)'},$$

here $A'(t) = (a_{ij}(t))_{i,j=1}^{n-1}$ is an $(n-1) \times (n-1)$ matrix;

$$(a^n)'(t) = (a_{1n}(t), \dots, a_{n-1,n}(t))^T, \quad b'(t) = (b_1(t), \dots, b_{n-1}(t))^T,$$

$$a'_n(t) = (a_{n1}(t), \dots, a_{n,n-1}(t))^T, \quad c' = (c_1, \dots, c_{n-1})^T.$$

By virtue of (2) and of the assumption $\forall t \exists \alpha_i \max_{t \geq t_0} \|h(t)\| < \infty, \forall t \exists \alpha_i \max_{t \geq t_0} h^1(t) < \infty$.

Assume furthermore that $h^1(t)$ is a function absolutely continuous on $[t_0, \infty)$ and

$$\forall t \exists \alpha_i \max_{t \geq t_0} \left\| \frac{dh^1(t)}{dt} \right\| < \infty.$$

If for the solution $x(t)$ of equation (7) $\varepsilon(x(t))$ maintains sign of $[t_0, \infty)$, then for $x(t)$ the differential relation (8) is true. Then holds the following

Theorem 1. For the Cauchy matrix $\varphi(t, t')$ of the equation $x' = \bar{A}(t)x'$ holds the inequality $\|\varphi(t, t')\| \leq C_0 \exp(-\delta_0(t-t'))$ with $t \geq t' \geq t_0$ where C_0 and δ_0 are some

positive numbers. Then, if for the absolutely continuous vector function $x(t)$, $t \geq t_0$ (8) holds and $\varepsilon(x(t)) \rightarrow 0$ as $t \rightarrow \infty$ then $x(t) \rightarrow 0$ as $t \rightarrow \infty$.

Proof. If for an absolutely continuous vector function $x(t)$, $t \geq t_0$ (8) holds and $\varepsilon(x(t)) \rightarrow 0$ as $t \rightarrow \infty$ then $x'(t)$ is a solution of some equation of the form

$$\dot{x}' = \bar{A}(t)x' + h(t)\chi(t) + h^1(t)\dot{\chi}(t), \quad t \geq t_0, \quad \chi(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad (9)$$

where $\chi(t)$ is a certain absolutely continuous function on $[t_0, \infty)$. For equation (9) write the Cauchy formula (with $t \geq t_0$)

$$x'(t) = \varphi(t, t_1)x'(t_1) + \int_{t_1}^t \varphi(t, \tau)(h(\tau)\chi(\tau) + h^1(\tau)\dot{\chi}(\tau)) d\tau,$$

where $t_1 \geq t_0$. Thence, integrating by parts, we have

$$\begin{aligned} x'(t) = & \varphi(t, t_1)x'(t_1) + \varphi(t, \tau)h^1(\tau)\chi(\tau) \Big|_{\tau=t}^{\tau=t} + \int_{t_1}^t [\varphi(t, \tau)h(\tau)\chi(\tau) - \\ & - \chi(\tau) \frac{d}{d\tau}(\varphi(t, \tau)h^1(\tau))] d\tau, \end{aligned}$$

and with due regard for the conditions of the Theorem and the fact that $\frac{d}{d\tau} \varphi(t, \tau) = -\varphi(t, \tau)\bar{A}(\tau)$ (with $t, \tau \geq t_0$), we have

$$\begin{aligned} \|x'(t)\| \leq & C_0 \|x'(t_1)\| \exp(-\delta_0(t-t_1)) + \\ & + C_0 \|h^1(t_1)\| \cdot |\chi(t_1)| \exp(-\delta_0(t-t_1)) + \\ & + \|h^1(t)\| \cdot |\chi(t)| + \int_{t_1}^t C_0 \|h(\tau) + \bar{A}(\tau)h^1(\tau) - \\ & - \frac{d}{d\tau} h^1(\tau)\| \cdot |\chi(\tau)| \exp(-\delta_0(t-\tau)) d\tau. \end{aligned} \quad (10)$$

The last four summands in the right-hand side of inequality (10) are as small as desired with $t \leq t_1$ if t_1 is large enough; the first summand in the right-hand side of (10) is as small as desired with $t \geq t_2 > t_1$ if t_2 is large enough; consequently, $\|x'(t)\|$ is as small as desired starting with some time and so $\|x'(t)\| \rightarrow 0$ as $t \rightarrow \infty$; since $\varepsilon(x(t)) \rightarrow 0$ as $t \rightarrow \infty$ then $\|x(t)\| \rightarrow 0$ as $t \rightarrow \infty$. This proves the Theorem.

It is easily seen that with the conditions of Theorem 1 true and solutions of equation (7) hitting the hyperplane $\varepsilon(x) = 0$ for these also O -hitting of the hyperplane $\varepsilon(x) = 0$ occurs.

For the system (S) the following Theorem holds.

Theorem 2. If for solutions of equation (7) O -hitting of the hyperplane $\varepsilon(x)=0$ occurs, then for any number $\rho > 0$ given in advance and for any V -solution of the system (S), $(x(t), \mu(t))$ there is time t_1 such that either the inequality

$$|\varepsilon(x(t_1)) + \mu(t_1)\delta\|x'(t_1)\| \leq \frac{\delta}{2}\|x'(t_1)\| \text{ holds or } \|x(t_1)\| \leq \rho.$$

Proof. Assume that the opposite is true, or that for a V -solution of the system (S), $(x(t), \mu(t))$ with $t \geq t_0$ it is true that

$$|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| \geq \frac{\delta}{2}\|x'(t)\| \tag{11}$$

and

$$\|x(t)\| > \rho. \tag{12}$$

Truth of inequality (11) signifies that with $t \geq t_0$ $x(t) \notin D'_{\frac{\delta}{2}}(t) = \{x: |\varepsilon(x) + \mu(t)\delta\|x'\| \leq \frac{\delta}{2}\|x'\|\}$. With $t \geq t_0$ take up the set $D_{\delta_1}(t) = \{x: |\varepsilon(x) + \mu(t)\delta\|x'\| \leq \delta_1\|x\|\}$ and choose δ_1 so that with $t \leq t_0$ $D_{\delta_1}(t) \subset D'_{\frac{\delta}{2}}(t)$.

Let $x \in D_{\delta_1}(t)$. Then $-\delta_1\|x\| \leq \varepsilon(x) + \mu(t)\delta\|x'\| \leq \delta_1\|x\|$ and since for V -solutions $|\mu(t)| \leq 1$, then $|\varepsilon(x)| \leq \delta_1\|x\| + \delta\|x'\|$ but $|\varepsilon(x)| \geq \|x\| - \|x'\| - C\|x'\|$ where $C = \max_{i=1, \dots, n-1} |c_i|$; from the latter two inequalities we have

$$(1 - \delta_1)\|x\| \leq (C + 1 + \delta)\|x'\|. \tag{13}$$

Choose $\delta_1: 0 < \delta_1 < 1$; then from (13) $\|x\| \leq \frac{C+1+\delta}{1-\delta_1}\|x'\|$ and so for $x \in D_{\delta_1}(t)$ we have

$$\|\varepsilon(x) + \mu(t)\delta\|x'\| \leq \delta_1 \frac{C+1+\delta}{1-\delta_1} \|x'\|. \tag{14}$$

Consequently, if δ_1 is determined from the inequality $0 < \delta_1 < 1$ and $\delta_1 \frac{C+1+\delta}{1-\delta_1} \leq \frac{\delta}{2}$, then from (14) with $t \geq t_0$ it follows that the inclusion $D_{\delta_1}(t) \subset D'_{\frac{\delta}{2}}(t)$ holds. Assume that a such δ_1 has been chosen. Then with $t \geq t_0$ it is true that if $x(t) \notin D'_{\frac{\delta}{2}}(t)$ then $x(t) \notin D_{\delta_1}(t)$. Therefore, if with $t \geq t_0$ the inequalities (11) and (12) hold, then

$$|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| \geq \delta_1\|x(t)\| > \delta_1\rho. \tag{15}$$

For specificity assume that

$$\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| > 0 \tag{16}$$

with $t \geq t_0$ (the case of the opposite sign of the inequality is treated in the same way).

By virtue of (15), (16) and equation (4), $\mu(t) = -1$ with $t \geq t' \geq t_0$ (here $t' - t_0 \leq \frac{2}{\alpha \delta_1 \rho}$). Therefore $\varepsilon(x(t)) - \delta \|x'(t)\| > \delta_1 \rho$ with $t \geq t'$; thence

$$\varepsilon(x(t)) > 0 \quad \text{with} \quad t \geq t'. \quad (17)$$

It is obvious that $x(t)$ with $t \geq t'$ satisfies equation (7). Consequently, by virtue of the Theorem conditions for $x(t)$ as a solution of equation (7) (with $t \geq t'$), either a time t_1 exists such that $\varepsilon(x(t_1)) = 0$ which is in conflict with (17) or $\|x(t)\| \rightarrow 0$ as $t \rightarrow \infty$ which is in conflict with (12). These conflicts show that at certain time t_1 at least one of inequalities (11) and (12) does not hold. This proves the Theorem.

Let us proceed to generation of the control u in the case of $l = n - 1$.

The relations with which inequality (5) is maintained are given by

Theorem 3. Let for the system (S) (with $l = n - 1$) the following relations hold

$$\kappa_i^0 \operatorname{sgn}(\tilde{c}, b) > \nu \tau \alpha_i \max_{t \geq t_0} \frac{1}{|(\tilde{c}, b)|} |(\tilde{c}, a^i - a^n \tilde{c}_i)|, \quad i = 1, \dots, n-1 \quad (18)$$

almost everywhere on $[t_0, \infty)$ for all $\tilde{c}: \tilde{c} = c + \sigma$, where

$$\sigma = (\sigma_1, \dots, \sigma_{n-1}, 0)^T, \quad \sigma_i = \pm \frac{3}{2} \delta, \quad i = 1, \dots, n-1;$$

$$\frac{1}{\alpha} \max_{i=1, \dots, n-1, \tilde{c} \in \Sigma, -1 \leq \lambda \leq 1} \{ \nu \tau \alpha_i \max_{t \geq t_0} |(\tilde{c}, \alpha^i - a^n \tilde{c}_i + b \kappa_i^0)| \} \leq \leq \frac{\rho \delta}{2 \left(\tilde{c} + 1 + \frac{3}{2} \delta \right)}, \quad (19)$$

where

$$\Sigma = \{ \tilde{c}: \tilde{c} = c + \lambda \sigma, \quad -1 \leq \lambda \leq 1,$$

$$\sigma = (\sigma_1, \dots, \sigma_{n-1}, 0)^T, \quad \sigma_i = \pm \frac{3}{2} \delta,$$

$$i = 1, \dots, n-1 \}, \quad C = \max_{i=1, \dots, n-1} |c_i|, \quad \rho = \operatorname{const} > 0.$$

Then, if for a V -solution $(x(t), \mu(t))$ of the system (S) at certain time t_1 the inequality holds

$$|\varepsilon(x(t_1)) + \mu(t_1) \delta \|x'(t_1)\|| \leq \frac{\delta}{2} \|x'(t_1)\|$$

$$\text{then } |\varepsilon(x(t)) + \mu(t) \delta \|x'(t)\|| \leq \frac{\delta}{2} \|x'(t)\|$$

with $t \in [t_1, t_2]$ if $x(t) \notin U_\rho$ with $t \in [t_1, t_2]$.

Proof. Consider a V -solution $(x(t), \mu(t))$ in question. Assume that the opposite is true, or that there is a time t' and an interval $(t', t' + \Delta)$ (where $\Delta > 0$) such that

$$|\varepsilon(x(t')) + \mu(t')\delta\|x'(t')\|| = \frac{\delta}{2}\|x'(t')\| \quad (20)$$

and

$$|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|| > \frac{\delta}{2}\|x'(t)\| \quad (21)$$

with $t \in (t', t' + \Delta)$ (here $(t', t' + \Delta) \subset [t_1, t_2]$). Let for specificity

$$\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| > 0 \quad \text{with } t \in [t', t' + \Delta) \quad (22)$$

(the case of opposite sign is treated in the same way). Let Δ be so small that $\|x'(t)\| \neq 0$ with $t \in [t', t' + \Delta)$ the choice of such Δ is possible since from (20) it follows that $\|x'(t')\| \neq 0$ because otherwise $x(t) = 0$ but $x(t) \notin U_\rho$ with $t \in [t_1, t_2]$). With due regard for (20) and the fact that for a V -solution $(x(t), \mu(t))$ we find that the equality holds

$$|\varepsilon(x(t))| \geq \left(1 + \frac{\omega(t)}{2}\right) \delta \|x'(t)\| \quad (23)$$

with $t \in [t', t' + \Delta)$ where $\omega(t) = \frac{2(\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|)}{\delta\|x'(t)\|} \rightarrow 1$ as $t \rightarrow t'$. But always

$$|\varepsilon(x(t))| \geq \|x(t)\| - \|x'(t)\| - C\|x'(t)\| \left(C = \max_{i=1, \dots, n-1} |c_i| \right). \quad (24)$$

From (23) and (24) we have that $\|x'(t)\| \geq \frac{\|x(t)\|}{C+1 + \left(1 + \frac{\omega(t)}{2}\right)\delta}$ (with $t \in [t', t' + \Delta)$) and

since $\|x(t)\| \geq \rho$ (with $t \in [t_1, t_2]$), then $\|x'(t)\| \geq \frac{\rho}{C+1 + \left(1 + \frac{\omega(t)}{2}\right)\delta}$ with $t \in [t', t' + \Delta)$

and consequently, by virtue of (21) and (22)

$$\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| > \frac{\delta\rho}{2\left(C+1 + \left(1 + \frac{\omega(t)}{2}\right)\delta\right)} \quad \text{with } t \in [t', t' + \Delta). \quad (25)$$

Let us consider the following possibilities:

1° Let $-1 < \mu(t') < 1$; then with Δ reasonably small $-1 < \mu(t) < 1$ with $t \in [t', t' + \Delta)$. Therefore with $t \in [t', t' + \Delta)$ $\dot{\mu}(t) = -\alpha(\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|)$ and with these t

we have

$$\begin{aligned}
 \frac{\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|}{\|x'(t)\|} &= \frac{\varepsilon(x(t')) + \mu(t')\delta\|x'(t')\|}{\|x'(t')\|} + \int_{t'}^t [(\varepsilon(x(\tau)) + \mu(\tau)\delta\|x'(\tau)\|) - \\
 &- \frac{\varepsilon(x(\tau)) + \mu(\tau)\delta\|x'(\tau)\|}{\|x'(\tau)\|}] \frac{1}{\|x'(\tau)\|} d\tau = \\
 &= \int_{t'}^t \frac{\sum_{i=1}^{n-1} [(\tilde{c}, a^i - a^n \tilde{c}_i + b\mu\kappa_i^0 \operatorname{sgn} x_i) - \alpha(\varepsilon(x) + \mu\delta\|x'\|) \operatorname{sgn} x_i] x_i}{\sum_{i=1}^{n-1} |x_i|} d\tau + \\
 &+ \frac{\varepsilon(x(t')) + \mu(t')\delta\|x'(t')\|}{\|x'(t')\|} \quad (26)
 \end{aligned}$$

where the integrand is a generalized derivative of the function in the left-hand side of equalities (26); furthermore, $\tilde{c} = c + \lambda\sigma$ where $\sigma(t) = (\sigma_1(t), \dots, \sigma_{n-1}(t), 0)^T$, $\sigma_i(t) = \delta \operatorname{sgn} x_i(t)$ ($i = 1, \dots, n-1$), $\lambda(t) = \frac{\varepsilon(x(t))}{\delta\|x'(t)\|}$ (by virtue of (23) $|\lambda(t)| \leq 1 + \frac{\omega(t)}{2}$) x_n in (26) is replaced by using the equation $x_n = -\lambda\delta\|x'\| - \sum_{i=1}^{n-1} c_i x_i = -\sum_{i=1}^{n-1} \tilde{c}_i x_i$. Since inequalities (19) and (25) hold and $\omega(t) \rightarrow 1$ as $t \rightarrow t'$, the integrand in (26) with $t \in [t', t' + \Delta]$ is nonnegative if Δ is reasonably small and so by virtue of (26) and (22) with these t

$$|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| \leq \frac{\delta}{2} \|x'(t)\|$$

which is in conflict with (21).

2° Let $\mu(t') = 1$. Then by virtue of (22) and equation (4) $-1 < \mu(t) < 1$ with $t \in (t', t' + \Delta)$ if Δ is reasonably small and, consequently, $\dot{\mu}(t) = -\alpha(\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|)$ with these t ; reasoning as in Case 1° we come in conflict with inequality (21).

3° Let $\mu(t') = -1$. Then by virtue of (22) and equation (4) $\dot{\mu}(t) = 0$ with $t \in (t', t' + \Delta)$. Then with $t \in (t', t' + \Delta)$ we have

$$\begin{aligned}
 \frac{\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|}{\|x'(t)\|} &= \frac{\varepsilon(x(t')) + \mu(t')\delta\|x'(t')\|}{\|x'(t')\|} + \\
 + \int_{t'}^t \frac{\dot{\varepsilon}(x(\tau)) + \frac{\varepsilon(x(\tau))}{\|x'(\tau)\|} \|x'(\tau)\|}{\|x'(\tau)\|} d\tau &= \frac{\varepsilon(x(t')) + \mu(t')\delta\|x'(t')\|}{\|x'(t')\|} + \\
 + \int_{t'}^t \frac{\sum_{i=1}^{n-1} (\tilde{c}, a^i - a^n \tilde{c}_i - b\kappa_i^0 \operatorname{sgn} x_i) x_i}{\sum_{i=1}^{n-1} |x_i|} d\tau &\quad (27)
 \end{aligned}$$

where the integrand is a generalized derivative of the function in the left-hand side of inequalities (27); furthermore, $\tilde{c} = c + \lambda\sigma$ where

$$\sigma(t) = (\sigma_1(t), \dots, \sigma_{n-1}(t), 0)^T, \\ \sigma_i(t) = \delta \operatorname{sgn} x_i(t) \quad (i = 1, \dots, n-1), \quad \lambda(t) = \frac{\varepsilon(x(t))}{\delta \|x'(t)\|}$$

(it is easily seen that by virtue of (20) and (22), $\lambda(t) \rightarrow \frac{3}{2}$ as $t \rightarrow t'$); with coordinate notation in the integrand of (27) x_n is replaced by $-\sum_{i=1}^{n-1} \tilde{c}_i x_i$.

Allowing for relation (18) and the fact that $\lambda(t) \rightarrow \frac{3}{2}$ as $t \rightarrow t'$, we have a negative quantity in the integrand of (27). Consequently, by virtue of (27) and (22), with these t

$$\frac{|\varepsilon(x(t)) + \mu(t)\delta \|x'(t)\||}{\|x'(t)\|} \leq \frac{\delta}{2},$$

which is conflict with the inequality (21).

The conflicts of Cases 1°–3° prove Theorem 3.

Theorem 3 provides conditions and relations with which for V -solutions of the system (S) inclusion (6) holds. Let us see whether it is possible for V -solutions that $x(t) \rightarrow 0$ as $t \rightarrow \infty$ if (6) holds. This equation is answered by

Theorem 4. Assume that for the Cauchy matrix $\varphi(t, t')$ of the equation $\dot{x}' = \bar{A}(t)x'$, $t \geq t_0$ ($\bar{A}(t)$ from (8)) it is true that $\|\varphi(t, t')\| \leq C_0 \exp(-\delta_0(t-t'))$ with $t \geq t' \geq t_0$ where C_0 and δ_0 are some positive constants. Assume also that the inequality holds

$$\frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1(t)\| < 1 \quad (28)$$

where $h^1(t)$ from (8). Then, if for an absolutely continuous vector function $x(t)$, $t_1 \leq t \leq t_2$ ($t_1 \geq t_0$) a differential relation (8) and the inclusion $x(t) \in G_{\frac{3}{2}\delta} = \left\{ x : |\varepsilon(x)| \leq \frac{3}{2} \delta \|x'\| \right\}$ with $t \in [t_1, t_2]$ hold, then for $x(t)$ with $t \in [t_1, t_2]$ the estimate is true

$$\|x(t)\| \leq \left(C + 1 + \frac{3}{2} \delta \right) \frac{C_0 \left(1 + \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\| \right)}{1 - \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|} \|x(t_1)\| \times \\ \times \exp \left(- \left(\delta_0 - \frac{3}{2} \delta M \right) (t - t_1) \right) \quad (29)$$

where

$$M = \frac{C_0 \nu \tau \alpha i \max_{t \geq t_0} \left\| h + \bar{A} h^1 - \frac{dh^1}{dt} \right\|}{1 - \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|}. \quad (30)$$

Proof. Proceeding from (8) write for $x(t)$ an integral Volterra equation ($t \in [t_1, t_2]$)

$$x'(t) = \varphi(t, t_1)x'(t_1) + \int_{t_1}^t \varphi(t, \tau) (h(\tau)\varepsilon(x(\tau)) + h^1(\tau)\dot{\varepsilon}(x(\tau))) d\tau$$

where, integrating part-wise, we have

$$x'(t) = \varphi(t, t_1)x'(t_1) + \varphi(t, \tau)h^1(\tau)\varepsilon(x(\tau)) \Big|_{\tau=t_1}^{\tau=t} + \int_{t_1}^t \varphi(t, \tau)h(\tau)\varepsilon(x(\tau)) d\tau - \int_{t_1}^t \varepsilon(x(\tau)) \frac{d}{d\tau} (\varphi(t, \tau)h^1(\tau)) d\tau$$

therefore, with due regard for the conditions of the Theorem and the fact that $\frac{d}{d\tau} \varphi(t, \tau) = -\varphi(t, \tau)\bar{A}(\tau)$ ($t, \tau \geq t_0$) we have

$$\begin{aligned} \|x'(t)\| \leq & C_0 \left(1 + \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\| \right) \|x'(t_1)\| \times \\ & \times \exp(-\delta(t-t_1)) + \frac{3}{2} \delta \|x'(t)\| \nu \tau \alpha i \max_{t \geq t_0} \|h^1\| + \\ & + \frac{3}{2} \delta C_0 \nu \tau \alpha i \max_{t \geq t_0} \left\| h(\tau) + \bar{A}(\tau)h^1(\tau) - \right. \\ & \left. - \frac{dh^1(\tau)}{d\tau} \right\| \int_{t_1}^t \|x'(\tau)\| \exp(-\delta_0(t-\tau)) d\tau. \end{aligned} \quad (31)$$

Taking into account (28) and using the Gronwall–Bellman lemma we have from (31)

$$\|x'(t)\| \leq \frac{C_0 \left(1 + \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|\right)}{1 - \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|} \times \\ \times \|x'(t_1)\| \exp\left(-\left(\delta_0 - \frac{3}{2} \delta M\right)(t-t_1)\right) \quad (32)$$

with $t \in [t_1, t_2]$ where M is as in (30) and since with $x \in G_{\frac{3}{2}\delta} \|x\| \leq \left(C+1 + \frac{3}{2}\delta\right) \|x'\|$ we have (29) from (32) and the Theorem is proved.

Remarks: 1) Theorem 4 determines the behaviour of the n -dimensional vector function $x(t)$ by the consideration of an $(n-1)$ -st order system; 2) in the case where the matrix $A(t)$ and vector $b(t)$ have a structure

$$A(t) = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \\ -a_1(t) & -a_2(t) & \dots & -a_n(t) \end{bmatrix}, \quad b(t) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \bar{b}(t) \end{bmatrix},$$

the conditions of Theorem 4 are easily verifiable and the differential relation (8) takes the form

$$\begin{cases} \dot{x}_i(t) = x_{i+1}(t), & i = 1, \dots, n-2, \\ \dot{x}_{n-1}(t) = -\sum_{i=1}^{n-1} c_i x_i(t) + \varepsilon(x(t)), \end{cases}$$

and, consequently, over a time interval where $x(t) \in G_{\frac{3}{2}\delta}$ $x'(t)$ is a solution of some system of the form

$$\begin{cases} \dot{x}_i = x_{i+1}, & i = 1, \dots, n-2 \\ \dot{x}_{n-1} = -\sum_{i=1}^{n-1} c_i x_i + \psi(x_1, \dots, x_{n-1}, t), & |\psi(x_1, \dots, x_{n-1})| \leq \frac{3}{2} \delta \|x'\| \end{cases}$$

and so over that time interval the dynamic properties of the process $x(t)$ are little dependent on parameters of the system which is described by equation (1), and these properties can be influenced by choosing the parameters c_i ($i=1, \dots, n-1$) and δ .

If the conditions of Theorems 3 and 4, and the inequality

$$\frac{3}{2} \delta \frac{C_0 \nu \tau \alpha i \max_{t \geq t_0} \left\| h + \bar{A} h^1 - \frac{dh^1}{dt} \right\|}{1 - \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|} < \delta_0 \quad (33)$$

are true when the inequality (5) holds for the V -solution of the system (S) ($x(t), \mu(t)$) at time t_1 with $x(t_1) \notin U_\rho$, it follows from the assertions of Theorems 3 and 4 that there is time t' such that $x(t') \in B_\rho$. Let us forego for a time being further (with $t > t'$) behaviour of V -solution ($x(t), \mu(t)$) and consider the times at which no derivative of $\mu(t)$ exists (following that time an equality $\mu(t) = -1$ or $\mu(t) = 1$ is maintained over a certain period in sliding mode). Let us give the following

Definition. On the solution of the system (S) ($x(t), \mu(t)$) a μ -discontinuity occurs at time $t' > t_0$ if either $\mu(t') = 1$ and $\varepsilon(x(t) + \mu(t)\sigma) \|x(t)\| < 0$ with $t \in (t', t' + \Delta')$ while with $t \in (t', t' - \Delta') \mu(t) < 1$, where Δ' is a certain positive number or $\mu(t') = -1$ and $\varepsilon(x(t) + \mu(t)\delta) \|x'(t)\| < 0$ with $t \in (t', t' + \Delta'')$ while with $t \in (t', t' - \Delta'') \mu(t) > -1$ where Δ'' is a certain positive number.

Then holds the following

Theorem 5. Assume that for the system (S) with $l = n - 1$ relations (18), (19) and

$$\frac{\kappa_i^0}{2} \operatorname{sgn}(\tilde{c}, b) > \nu \tau \alpha i \max_{t \geq t_0} \frac{1}{|(\tilde{c}, b)|} |\tilde{c}, a^i - a^n \tilde{c}_i|, \quad i = 1, \dots, n-1 \quad (34)$$

hold almost everywhere on $[t_0, \infty)$ for all $\tilde{c}: \tilde{c} = c + \sigma$ where $\sigma = (\sigma_1, \dots, \sigma_{n-1}, 0)^T$, $\sigma_i = \pm \delta$, $i = 1, \dots, n-1$.

Then on any V -solution of the system (S) ($x(t), \mu(t)$) a μ -disruption occurs maximum once in $[t_0, t_*)$ if $x(t) \notin U_\rho$ with $t \in [t_0, t_*)$.

Proof of Theorem 5 will be preceded by three lemmas.

Lemma 1. If for solution of the system (S) ($x(t), \mu(t)$) $|\varepsilon(x(t) + \mu(t)\delta) \|x'(t)\|| > \frac{\delta}{2} \|x'(t)\|$ with $t \in [t_1, t_2]$, then on that solution a μ -disruption can occur maximum once in $[t_1, t_2]$.

Proof. Assume that with $t \in [t_1, t_2]$ $\varepsilon(x(t) + \mu(t)\delta) \|x'(t)\| > 0$ (the case of opposite sign is treated in the same way). If $\mu(t) > -1$ with $t \in [t_1, t_2]$, then no μ -disruptions occur with $[t_1, t_2]$. If $\mu(t) > -1$ with $t \in [t_1, t')$ (here $[t_1, t_1) = \emptyset$) and $\mu(t') = -1$ where $t' \in [t_1, t_2]$, then by virtue of equation (4) and the above assumption $\mu(t) = -1$ with $t \in [t', t_2]$. Consequently on $[t_1, t_2]$ a μ -disruption can only occur at time t' .

Lemma 2. If for the system (S) (34) holds and for the solution

$$(x(t), \mu(t)) \quad |\varepsilon(x(t_1))| \leq \delta \|x'(t_1)\| \quad \text{and} \quad |\varepsilon(x(t)) + \mu(t)\delta \|x'(t)\|| \leq \frac{\delta}{2} \|x'(t)\| \quad \text{with} \quad t \in [t_1, t_2],$$

then with $t \in [t_1, t_2]$ there are no μ -disruptions on $(x(t), \mu(t))$.

Proof. With the conditions of the Lemma holding it is true that $|\varepsilon(x(t))| \leq \delta \|x'(t)\|$ with $t \in [t_1, t_2]$. Indeed, assume that the opposite is true. Let $t' \in [t_1, t_2]$ and $\Delta = \text{const} > 0$ exist such that $t' + \Delta \leq t_2$,

$$\begin{aligned} \varepsilon(x(t')) + \delta \|x'(t')\| &= 0 \quad \text{and} \\ \varepsilon(x(t)) + \delta \|x'(t)\| &< 0 \quad \text{with} \quad t \in (t', t' + \Delta) \end{aligned} \quad (35)$$

(the case of $\varepsilon(x(t')) - \delta \|x'(t')\| = 0$ and $\varepsilon(x(t)) - \delta \|x'(t)\| > 0$ with $t \in (t', t' + \Delta)$ is treated similarly).

By virtue of (35) and the fact that by the condition of the Lemma $-\frac{\delta}{2} \|x'(t')\| \leq \varepsilon(x(t')) + \mu(t')\delta \|x'(t')\|$ we find that $\mu(t') \geq \frac{1}{2}$. Consequently, then

$$\mu(t) \rightarrow K \quad \text{with} \quad t \rightarrow t' \quad \text{where} \quad K \geq \frac{1}{2}. \quad (36)$$

But since

$$\varepsilon(x(t)) + \delta \|x'(t)\| = \int_{t'}^t (\dot{\varepsilon}(x(\tau)) + \delta \|x'(\tau)\|) d\tau \quad (37)$$

(here the integrand is a generalized derivative of the function in the left-hand side of equality (37)), then expressing the integrand in coordinates and allowing for relation (34) and truth of (36) we have from (37) that $\varepsilon(x(t)) + \delta \|x'(t)\| \geq 0$ with $t(t > t')$ near t' contrary to (37). Consequently, $|\varepsilon(x(t))| \leq \delta \|x'(t)\|$ with $t \in [t_1, t_2]$. Then it is easily seen that there are no μ -disruption on the solution $(x(t), \mu(t))$ with $t \in [t_1, t_2]$.

Lemma 3. If for V -solution of the system (S) $(x(t), \mu(t))$ $\varepsilon(x(t)) + \delta \|x'(t)\| < 0$ or $\varepsilon(x(t)) - \delta \|x'(t)\| > 0$ with $t \in [t_1, t_2]$ then on this solution a μ -disruption occurs maximum once in $[t_1, t_2]$.

Proof. Since for V -solutions $|\mu(t)| \leq 1$, then, by virtue of the conditions of the Lemma, $\varepsilon(x(t)) + \mu(t)\delta \|x'(t)\|$ maintains sign on $[t_1, t_2]$. Then proceed as in proof of Lemma 1.

Proof of Theorem 5. Let $(x(t), \mu(t))$ be V -solution of the system (S) for which

$$x(t) \notin U_\rho \quad \text{with} \quad t \in [t_0, t_*]. \quad (38)$$

Let us show that on this V -solution $(x(t), \mu(t))$ a μ -disruption occurs maximum once in $[t_0, t_*]$.

Let us take up the following possibilities.

1° Let $|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| > \frac{\delta}{2}\|x'(t)\|$ with $t \in [t_0, t_*]$. In this case apply Lemma 1.

2° Let $|\varepsilon(x(t_0)) + \mu(t_0)\delta\|x'(t_0)\| \leq \frac{\delta}{2}\|x'(t_0)\|$. Then by virtue of (38) and the conditions of the Theorem $|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| \leq \frac{\delta}{2}\|x'(t)\|$ with $t \in [t_0, t_*]$. If $|\varepsilon(x(t))| > \delta\|x'(t)\|$ with $t \in [t_0, t_*]$, then apply Lemma 3. If $|\varepsilon(x(t'))| \leq \delta\|x'(t')\|$ and $|\varepsilon(x(t))| > \delta\|x'(t)\|$ with $t \in [t_0, t']$ (here $t' \in [t_0, t_*]$), then with $t \in [t', t_*]$ there are no μ -disruptions, by virtue of Lemma 2. On $[t_0, t']$ by virtue of Lemma 3 a μ -disruption can occur maximum once.

3° Let

$$|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| > \frac{\delta}{2}\|x'(t)\| \quad \text{with } t \in [t_0, t'] \quad (39)$$

where $t_0 < t' < t_*$ and $|\varepsilon(x(t')) + \mu(t')\delta\|x'(t')\| = \frac{\delta}{2}\|x'(t')\|$. Then by virtue of (38) and the conditions of the Theorem $|\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| \leq \frac{\delta}{2}\|x'(t)\|$ with $t \in [t', t_*]$.

By virtue of Lemma 1 a μ -disruption can occur maximum once in $[t_0, t']$. If it does not occur in $[t_0, t']$, then in considering a V -solution on $[t', t_*]$ the reasoning is as in Case 2°. Assume that at time $t'' \in (t_0, t')$ a μ -disruption occurs. Let $\mu(t'') = 1$ and $\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\| < 0$ with $t \in (t'', t'' + \Delta)$ and with $t \in (t'', t'' - \Delta)$ $\mu(t) < 1$ where $\Delta = \text{const} > 0$ (the case of opposite sign is treated similarly). Then by virtue of (39) with $t \in (t'', t')$ $\varepsilon(x(t)) + \delta\|x'(t)\| < 0$. With $t \in (t'', t')$ there are clearly no μ -disruptions. If $\varepsilon(x(t)) + \delta\|x'(t)\| < 0$ with $t \in [t', \bar{t}]$ where $t' \leq \bar{t} \leq t_*$ and $\varepsilon(x(\bar{t})) + \delta\|x'(\bar{t})\| = 0$ then there are clearly no μ -disruptions with $t \in [t', \bar{t}]$, which is also true by virtue of Lemma 2 with $t \in [\bar{t}, t_*]$. Therefore if on $[t_0, t']$ there is an μ -disruption then there is none on $[t', t_*]$.

Consequently, in all cases 1°–3° on the V -solution $(x(t), \mu(t))$ an μ -disruption occurs maximum once in $[t_0, t_*]$ if $x(t) \notin U_\rho$ with $t \in [t_0, t_*]$.

Now let us consider the behaviour of V -solutions $(x(t), \mu(t))$ following the moment of reaching the set B_ρ . The following Theorem holds.

Theorem 6. Let for the system (S) (with $l = n - 1$) the conditions of Theorems 3 and 4 and the inequality (33) and

$$\forall t \geq t_0 \quad \max (c, a^n(t)) < 0, \quad (40)$$

and inequality (18) with $\tilde{c} = c$ hold. Then for any V -solution of the system (S) $(x(t), \mu(t))$ it

follows from $x(t') \in B_\rho$ that $x(t) \in B_r$ with $t \geq t'$ where

$$r = \left(C + 1 + \frac{3}{2} \delta \right) \frac{C_0 \left(1 + \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\| \right)}{1 - \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|} \times \\ \times \frac{2\bar{C} \left(C + 1 + \frac{3}{2} \delta \right)}{3\delta} \cdot \rho \exp \left(\frac{4P}{\alpha \delta \rho} \right),$$

where $\bar{C} = \max(1, C)$, $P = \max_{i=1, \dots, n; -1 \leq \Delta \leq 1} \left\{ \nu \tau \alpha i \max_{t \geq t_0} \|a^i + b \kappa_i^0 \Delta\| \right\}$ (in the case of $l = n - 1 \kappa_n^0 = 0$).

Furthermore, for any V -solution $(x(t), \mu(t))$ there exists time t_* such that $x(t) \in B_r$ with $t \geq t_*$.

To prove Theorem 6, three Lemmas are needed.

Lemma 4. Let the conditions of Theorem 4 and inequality (33) hold. Let for the solution of the system (S) $(x(t), \mu(t))$ $x(t) \in G_{\frac{3}{2}\delta}^3$ with $t \in [t_1, t_2]$. Then for this solution

$$\|x(t)\| \leq \left(C + 1 + \frac{3}{2} \delta \right) \frac{C_0 \left(1 + \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\| \right)}{1 - \frac{3}{2} \delta \nu \tau \alpha i \max_{t \geq t_0} \|h^1\|} \|x(t_1)\|$$

with $t \in [t_1, t_2]$ where $C = \max_{i=1, \dots, n-1} |c_i|$.

Proof. Truth of the Lemma follows from the fact that under its condition (29) and (33) are true.

Lemma 5. Let for solution of the system (S) $(x(t), \mu(t))$ with $t \in [t_1, t_2]$ $|e(x(t)) + \mu(t)\delta \|x'(t)\|| > \frac{\delta}{2} \|x'(t)\|$, $\|x(t)\| \geq \rho$ and $-1 < \mu(t) < 1$.

Then for this solution $\|x(t)\| \leq \|x(t_1)\| \exp \left(\frac{4P}{\alpha \delta \rho} \right)$ with $t \in [t_1, t_2]$ where

$$P = \max_{i=1, \dots, n; -1 \leq \Delta \leq \approx} \left\{ \nu \tau \alpha i \max_{t \geq t_0} \|a^i + b \kappa_i^0 \Delta\| \right\}$$

(in the case of $l = n - 1 \kappa_n^0 = 0$).

Proof. Write for $x(t)$ an integral equation (with $t \in [t_1, t_2]$)

$$x(t) = x(t_1) + \int_{t_1}^t A(\tau)x(\tau)d\tau + \int_{t_1}^t b(\tau)(\kappa^0\mu(\tau), m\bar{x}(\tau))d\tau,$$

whence

$$\|x(t)\| \leq \|x(t_1)\| + \int_{t_1}^t P\|x(\tau)\|d\tau \quad (41)$$

where P is defined in the formulation of the Lemma. By virtue of the Lemma conditions and equation (4) for $t \in [t_1, t_2]$

$$t - t_1 \leq \frac{4}{\alpha\delta\rho}. \quad (42)$$

Using the Gronwall–Bellman lemma for the inequality (41) and bearing in mind (42), we arrive at the assertions of the Lemma.

Lemma 6. Let for the system (S) (with $l = n - 1$) inequalities (18) hold with $\bar{c} = c$ and (40). Let for the solution of the system (S) $(x(t), \mu(t) x(t) \notin G_{\frac{3}{2}\delta}^3$ and $\mu(t) = -\operatorname{sgn}(\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|)$ with $t \in [t_1, t_2]$. Then for this solution

$$\|x(t)\| \leq \frac{2\bar{C}\left(C + 1 + \frac{3}{2}\delta\right)}{3\delta} \|x(t_1)\|$$

with $t \in [t_1, t_2]$ where $C = \max_{i=1, \dots, n-1} |c_i|$, $\bar{C} = \max(1, C)$.

Proof. It is easily checked that

$$S_\delta = \{x: |\varepsilon(x)| \leq \frac{3\delta}{2\left(C + 1 + \frac{3}{2}\delta\right)} \|x\| \subset G_{\frac{3}{2}\delta}^3 = \left\{x: |\varepsilon(x)| \leq \frac{3}{2}\delta\|x'\|\right\}.$$

Therefore under the conditions of the Lemma for the solution of the system (S) $x(t) \notin S_\delta$ with $t \in [t_1, t_2]$. Since $\mu(t) = -\operatorname{sgn}(\varepsilon(x(t)) + \mu(t)\delta\|x'(t)\|)$ with $t \in [t_1, t_2]$ it follows that $\mu(t) = -\operatorname{sgn} \varepsilon(x(t))$ with $t \in [t_1, t_2]$. Consequently, with (18) at $\bar{c} = c$ and (40) true, it is easily seen that for this solution almost everywhere on $[t_1, t_2]$ $\varepsilon(x(t)) \dot{\varepsilon}(x(t)) \leq 0$ and so $|\varepsilon(x(t_1))| \geq |\varepsilon(x(t))|$ with $t \in [t_1, t_2]$. Now one can see that with $t \in [t_1, t_2]$

$$\bar{C}\|x(t_1) \geq \varepsilon(x(t_1))\| \geq |\varepsilon(x(t))| > \frac{3\delta}{2\left(C + 1 + \frac{3}{2}\delta\right)} \|x(t)\|$$

whence follows the truth of the Lemma.

The assertions of Lemma 6 follows from the fact that with (18) at $\tilde{c}=c$ (40) true, solutions of equation (7) (with $l=n-1$) reach the hyperplane $\varepsilon(x)=0$ of Theorems 1-4 and Lemmas 4-6.

Let us take up the case of control generation with $l=n$. In this case sufficient conditions for solutions of equation (7) to reach the hyperplane $\varepsilon(x)=0$ can be insured by choosing the vector κ^0 . All the results for this case are summarized in one Theorem (the results are proved almost identically with Theorems 3, 5 and 6).

Theorem 7. Let for the system (S) (with $l=n$) the relations hold

$$\begin{aligned} \kappa_i^0 \operatorname{sgn}(\tilde{c}, b) &\geq v\tau\alpha i \max_{t \geq t_0} \frac{1}{|(\tilde{c}, b)|} \times \\ &\times |(\tilde{c}, a^i - a^n \tilde{c}_i + b\kappa_n^0 \tilde{c}_i \Delta)|, \quad i=1, \dots, n-1 \end{aligned} \quad (43)$$

almost everywhere on $[t_0, \infty)$ for all $\Delta = \pm 1$ and $\tilde{c}: \tilde{c} = c + \sigma$ where $\sigma = (\sigma_1, \dots, \sigma_{n-1}, 0)^T$, $\sigma_i = \pm \frac{3}{2} \delta$, $i=1, \dots, n-1$;

$$\begin{aligned} &\frac{1}{\alpha} \max_{i=1, \dots, n-1; -1 \leq \Delta_1 \leq 1, -1 \leq \Delta_2 \leq 1, \tilde{c} \in \Sigma} \times \\ &\left\{ v\tau\alpha i \max_{t \geq t_0} |(\tilde{c}, a^i - a^n \tilde{c}_i + b\kappa_i^0 \Delta_1 + b\kappa_n^0 \Delta_2)| \right\} \leq \\ &\leq \frac{\rho\delta}{2\left(C + 1 + \frac{3}{2}\delta\right)}, \end{aligned} \quad (44)$$

where

$$\begin{aligned} \Sigma = \left\{ \tilde{c}: \tilde{c} = c + \lambda\sigma, \quad -1 \leq \lambda \leq 1, \quad \sigma = (\sigma_1, \dots, \sigma_{n-1}, 0)^T, \sigma_i = \pm \frac{3}{2} \delta, \right. \\ \left. i=1, \dots, n-1 \right\}, \quad C = \max_{i=1, \dots, n-1} |c_i|, \quad \rho = \text{const} > 0. \end{aligned}$$

Then follows from the assertion of Theorem 3.

If, furthermore, the relations hold

$$\begin{aligned} \frac{\kappa_i^0}{2} \operatorname{sgn}(\tilde{c}, b) &> v\tau\alpha i \max_{t \geq t_0} \frac{1}{|(\tilde{c}, b)|} \times \\ &\times |(\tilde{c}, a^i - a^n \tilde{c}_i + b\kappa_n^0 \tilde{c}_i \Delta)|, \quad i=1, \dots, n-1 \end{aligned}$$

almost everywhere on $[t_0, \infty)$ for all $\Delta: \frac{1}{2} \leq |\Delta| \leq 1$ and $\tilde{c}: \tilde{c} = c + \sigma$ where $\sigma = (\sigma_1, \dots, n-1, 0)^T$, $\sigma_i = \pm \delta$, $i = 1, \dots, n-1$, then follows the truth of Theorem 5.

Moreover, if relations (43) and (44), the conditions of Theorem 4, inequality (33) and the inequality

$$\kappa_i^0 \operatorname{sgn}(c, b) > \nu \tau \alpha i \max_{t \geq t_0} \frac{1}{|(c, b)|} |(c, a^i)|, \quad i = 1, \dots, n$$

almost everywhere on $[t_0, \infty)$ hold, then Theorem 6 is true.

4. Conclusions

The results of Section 3 show that in control of a dynamic system described by equation (1) through formulation of a control signal by law (3), (4), Section 2, a continuous control signal is generated and for every V -solution $(x(t), \mu(t))$, $t \geq t_0$ of the system (S) as defined in Section 2, provided that the conditions of Theorems 3-6 (with $l = n-1$) or of Theorem 7 (with $l = n$) on $[t_0, t_*)$ there is maximum one time at which no derivative of the arbitrary newly introduced parametric coordinate $\mu(t)$ exists if $x(t) \notin U_\rho$ with $t \in [t_0, t_*)$ and on $[t_0, t_*)$ the V -solution belongs to the discontinuity set of the right-hand side in the equation for variation of the parametric coordinate (4) on maximum one time interval; once the point $x(t)$ reaches the set B_ρ , which is true for any V -solution, at all subsequent times $x(t) \in B_r$ while with the parameter α properly chosen r can be made as small as desired; this way to generate a control signal is referred to as quasicontinuous.

Theorems of Section 3 provide relations for the design of a control system which insures change of the dynamic properties of the processes in the desired direction.

5. Examples

Let us consider a two-dimensional dynamic system to be controlled (the notation is as in Sections 1 and 2)

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -\alpha_1(t)x_1 - a_2(t)x_2 + u, \quad t \geq t_0 \end{cases} \quad (\text{E.1})$$

The control signal for the system (E.1) is designed by the algorithm

$$u = \kappa_i^0 \mu |x_1|, \quad (E.2)$$

$$\dot{\mu} = \begin{cases} -\alpha(c_1 x_1 + x_2 + \mu \delta |x_1|) & \text{with } |\mu| \leq 1, \\ 0 & \text{with } |\mu| > 1, |\mu(t_0)| \leq 1. \end{cases}$$

Specifying the range of the parameters $a_i(t)$ ($i = 1, 2$) let us determine the values of the parameters in (E.2) with which the algorithm of generating the control signal (E.2) is quasicontinuous and the processes to be controlled in the closed-loop system feature the desired dynamic properties which are defined as noted in Remark 2 of Section 2 by the equation

$$\dot{x}_1 = -c_1 x_1 + \psi(x_1, t), \quad |\varphi(x_1)| \leq \frac{3}{2} \delta |x_1|, \quad (E.3)$$

where ψ is some continuous function of the real argument. Note that with the conditions of Theorems in Sect. 3 valid, the process $(x_1(t), x_2(t))$ with arbitrary initial values reaches a closed sphere B_ρ and then stays inside the sphere B_r ; furthermore, following each time $t \geq t_0$ there are times in which $(x_1(t), x_2(t))$ belongs to B_ρ .

1° Let in (E.1) with almost all $t \geq t_0$:

$$-10 \leq a_1(t) \leq 10, \quad 5 \leq a_2(t) \leq 10$$

Let in (E.2) $c_1 = 2$ and $\delta = 1$; then the zero solution of (E.3) is exponentially stable. In this case the relation (40) holds and relations (18) and (34) are satisfied with $\kappa_1^0 = 79$. Then relation (19) is satisfied with $\alpha\rho \geq 1116$ and the assertion of Theorem 6 with this choice of $c_1, \delta, \kappa_1^0, \alpha$ and ρ is true with $r = 25\rho \exp\left(\frac{356}{\alpha\rho}\right)$. Assuming that $\rho = 1/10$ and $\alpha\rho = 1116$ determine $\alpha = 11160$ and have $r \leq 2.5 \exp\left(\frac{1}{3}\right)$. If $\rho = 1/25$, then assuming that $\alpha\rho = 1116$ determine $\alpha = 27.900$ and have $r \leq \exp\left(\frac{1}{3}\right)$. With $\rho = 1/50$ and assuming again that $\alpha\rho = 1116$ determine $\alpha = 55.800$ and have $r \leq 0.5 \exp\left(\frac{1}{3}\right)$. With $\rho = 1/100$ and $\alpha\rho = 1116$ we have $\alpha = 111.600$ and $r \leq 0.25 \exp\left(\frac{1}{3}\right)$.

2° Assume that in (E.1) with almost all $t \geq t_0$:

$$-100 \leq a_1(t) \leq 100, \quad 5 \leq a_2(t) \leq 10.$$

Assume that in (E.2) $c_1 = 2$ and $\delta = 1$. In this case (40) holds. Relations (18) and (34) hold with $\kappa_1^0 = 259$ and (19) hold, then Theorem 6 holds with $r = 25\rho \exp\left(\frac{1436}{\alpha\rho}\right)$.

Let $\rho = \frac{1}{10}$ then $\alpha = 35.460$ with $\alpha\rho = 3546$ and $r \leq 2.5 \exp\left(\frac{1}{2}\right)$. Let $\rho = 1/25$, then $\alpha = 88.65$ with $\alpha\rho = 3546$ and $r \leq \exp\left(\frac{1}{2}\right)$.

3° Assume that with the parameters $a_i(t)$ ($i=1, 2$) as in 2°, in (E.2) $c_1 = 3$ and $\delta = 1$. Relation (40) holds. With $\kappa_1^0 = 265$ (18) and (34) hold. Let $\alpha\rho \geq 4488$; then (19) is valid.

With such $c_1, \delta, \kappa_1^0, \alpha$ and ρ we can presume that $r = 81\rho \exp\left(\frac{1460}{\alpha\rho}\right)$. With $\rho = 1/81$ and $\alpha\rho = 4488$ we have $\alpha = 219.912$ and $r \leq \exp\left(\frac{1}{3}\right)$.

References

1. *Yemelyanov, S. V., Korovin, S. K.,* Rasshirenie mnozhestva tipov obratnykh svyazey i ikh primeneniye pri postroenii zamknutykh dinamicheskikh sistem. *Izv. AN SSSR. Tekhn. kibernet.*, No. 5, 1981.
2. *Teoriya sistem s peremennoy strukturoy. Yemelyanov, S. V. (ed.)* Nauka, Moscow, 1970.
3. *Bezvodinskaya, T. A., Sabaev, Ye. F.,* Usloviya ustoychivosti v tselom sistemy s peremennoi strukturoi. *Avtomatika i telemekhanika*, Vol. 35, No. 10, 1974.

Управление нестационарными динамическими системами при квазинепрерывном формировании управляющего воздействия

С. В. ЕМЕЛЬЯНОВ, С. К. КОРОВИН, Б. В. УЛАНОВ

(Москва)

Рассматривается задача управления динамическими системами, линейными по координатам состояния и по управлению. Предлагается алгоритм формирования управляющего воздействия, называемый квазинепрерывным, с использованием дополнительной параметрической координаты (имеющей дифференциальный закон изменения), при котором вырабатывается непрерывный сигнал управления и при некоторых соотношениях, даваемых в работе, всякое решение получаемой системы дифференциальных уравнений принадлежит множеству разрыва правой части уравнения изменения параметрической координаты не более, чем на одном промежутке времени, пока вектор регулируемых координат не достигнет некоторой окрестности нуля, после попадания в которую этот вектор во все последующие моменты времени не покидает некоторой большей окрестности нуля, которая может быть сделана как угодно малой за счет выбора некоторого параметра алгоритма управления. В системе управления формирование управляющего воздействия реализуется с использованием контуров координатно-параметрической и параметрической обратных связей.

С. В. Емельянов

Всесоюзный научно-исследовательский институт

системных исследований

СССР 119034 Москва Г-34,

ул. Рылеева, 29

UNIVERSAL CONSISTENCY RESULTS FOR WOLVERTON-WAGNER REGRESSION FUNCTION ESTIMATE WITH APPLICATION IN DISCRIMINATION

A. KRZYŻAK, M. PAWLAK

(Wrocław)

(Received January 1, 1982)

In the paper the pointwise consistency of recursive regression function estimate motivated by Wolverton and Wagner [13] was examined.

Moreover, for classification rule resulting from this estimate weak and strong Bayes risk consistencies were studied. The results obtained are universal, i.e. they do not require any assumptions about the underlying distributions.

1. Introduction

Let $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ be a sequence of independent and identically distributed random vectors from $R^d \times R$ and let μ be the probability measure of X . Estimate the regression function $m(x) = E(Y/X = x)$, by

$$m_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$$

where

$$W_{ni}(x) = \frac{h^{-d}(i) K\left(\frac{x - X_i}{h(i)}\right)}{\sum_{j=1}^n h^{-d}(j) K\left(\frac{x - X_j}{h(j)}\right)} \quad (1)$$

and $\{h(n)\}$ is a sequence of positive numbers while K is a given nonnegative function on R^d .

The recursive computation of (1) can be carried out as follows

$$m_n(x) = m_{n-1}(x) + g_n^{-1}(x) (Y_n - m_{n-1}(x)) K\left(\frac{x - X_n}{h(n)}\right)$$

$$g_n(x) = \left(\frac{h(n)}{h(n-1)}\right)^d g_{n-1}(x) + K\left(\frac{x - X_n}{h(n)}\right)$$

$$m_0(x) = g_0(x) = 0.$$

Asymptotical properties of (1) were studied by Greblicki [5], Ahmad and Lin [1], Devroye and Wagner [2] as well as Greblicki and Krzyżak [6].

Devroye and Wagner [2] assumed absolute continuity of measure μ while other authors put additional assumptions on μ .

In practice, however, we hardly have any information about underlying distributions. Therefore, the most reasonable approach is to study universal consistency because the obtained results are valid for all possible distributions. Results of this type were but recently obtained by Stone [11], Devroye and Wagner [3], Györfi [8] and Devroye [4].

In Sections 2 and 3 we will examine the pointwise weak consistency

$$m_n(x) \rightarrow m(x) \text{ in probability mod } \mu \text{ as } n \rightarrow \infty \quad (2)$$

as well as pointwise strong consistency

$$m_n(x) \rightarrow m(x) \text{ a.s. mod } \mu \text{ as } n \rightarrow \infty \quad (3)$$

of estimate (1).

In section 4 we will show how to apply these results to obtain universal weak and strong consistency of the discrimination rule derived from estimate (1).

2. Weak consistency

In what follows $S_{x,r}$ will denote the closed sphere with radius r centered at x and I_A is the indicator function of a set A .

Theorem 1. If $EY^2 < \infty$

$$h(n) \rightarrow 0, \quad n^{-2} \sum_{i=1}^n h^{-d}(i) \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (4)$$

$$\overline{\lim}_n \frac{h^{-d}(n)}{n^{-1} \sum_{i=1}^n h^{-d}(i)} < \infty \quad (5)$$

there exist positive numbers a, r such that

$$K(x) \leq a I_{\{\|x\| \leq r\}}(x). \quad (6)$$

$K(x)$ depends only on $\|x\|$ and decreases as $\|x\|$ increases, that is

$$K(x) = \varphi(\|x\|) \text{ where } \varphi(t), t > 0 \quad (7)$$

is monotone decreasing then (2) holds.

To prove Theorem 1 we introduce the following lemmas.

Lemma 1. Let K be nonnegative μ -integrable function with compact support and satisfy condition (7), then for every function $f \in L(\mu)$

$$\frac{\int K\left(\frac{x-y}{h}\right)f(y) \mu(dy)}{\int K\left(\frac{x-y}{h}\right) \mu(dy)} \rightarrow f(x) \text{ mod } \mu \quad \text{as } h \rightarrow 0. \tag{8}$$

Proof. The proof is motivated by the argument of Wheeden and Zygmund [12, pp. 156–157].

Let $x \in R^d$ and

$$E = \left\{ (y, t) : y \in R^d, t > 0, K\left(\frac{x-y}{h}\right) > t \right\}$$

be a subset of R^{d+1} , then

$$K\left(\frac{x-y}{h}\right) = \int_0^{K\left(\frac{x-y}{h}\right)} dt = \int_0^\infty I_E(y, t) dt$$

Therefore by Fubini's theorem

$$\int K\left(\frac{x-y}{h}\right) \mu(dy) = \int_0^\infty \mu(E_t) dt$$

where $E_t = \left\{ y : K\left(\frac{x-y}{h}\right) > t \right\}$, and

$$\begin{aligned} \int K\left(\frac{x-y}{h}\right)f(y)\mu(dy) &= \int_0^\infty \left(\int_{E_t} f(y)\mu(dy) \right) dt = \\ &= \int_0^\infty \mu(E_t) \left(\frac{1}{\mu(E_t)} \int_{E_t} f(y)\mu(dy) \right) dt \leq \\ &\leq \sup_{t>0} \frac{1}{\mu(E_t)} \int_{E_t} f(y)\mu(dy) \int_0^\infty \mu(E_t) dt. \end{aligned}$$

By (7) E_t is a sphere centered at x and with radius $h\varphi^{-1}(t)$. The left side of (8) is dominated by

$$\sup_{t>0} \frac{1}{\mu(E_t)} \int_{E_t} f(y) \mu(dy) = \sup_{\substack{S_{x,r} \\ 0 < r \leq ch}} \frac{1}{\mu(S_{x,r})} \int_{S_{x,r}} f(y) \mu(dy) \quad (9)$$

where c is equal to radius of support of K . (9) tends to $f(x) \bmod \mu$ if $h \rightarrow 0$ by theorem 10.49 of Wheeden and Zygmund [12, p. 189]. q.e.d.

Lemma 2. Let $h(n) \rightarrow 0$ as $n \rightarrow \infty$ and there exist positive constants β, r such that

$$K(x) \geq \beta I_{\{\|x\| \leq r\}}(x)$$

then

$$\lim_n n^{-1} \sum_{i=1}^n h^{-d}(i) EK \left(\frac{x - X_i}{h(i)} \right) > 0 \bmod \mu. \quad (10)$$

Proof.

$$n^{-1} \sum_{i=1}^n h^{-d}(i) EK \left(\frac{x - X_i}{h(i)} \right) \geq \beta n^{-1} \sum_{i=1}^n h^{-d}(i) \mu(S_{x, rh(i)}).$$

Proposition follows by measure-theoretic result of Devroye [4]:

$$\lim_{n \rightarrow \infty} \frac{h^d(n)}{\mu(S_{x, rh(n)})} \text{ exists for each } x \in R^d \bmod \mu. \quad (11)$$

q.e.d.

Conclusion. From inequality (10) it follows that

$$\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x - X_i}{h(i)} \right) \rightarrow \infty \bmod \mu \quad \text{as } n \rightarrow \infty. \quad (12)$$

Lemma 3. If

$$h(n) \rightarrow 0, n^{-2} \sum_{i=1}^n h^{-d}(i) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

$$\overline{\lim}_n \frac{h^{-d}(n)}{n^{-1} \sum_{i=1}^n h^{-d}(i)} < \infty \quad (13)$$

then

$$\frac{h^{-d}(n)}{\sum_{j=1}^n h^{-d}(j)EK\left(\frac{x-X_j}{h(j)}\right)} \rightarrow 0 \text{ mod } \mu \quad \text{as } n \rightarrow \infty.$$

Proof. Consider the quotient

$$\begin{aligned} & \frac{n^2}{\sum_{i=1}^n h^{-d}(i)} \frac{h^{-d}(n)}{\sum_{j=1}^n h^{-d}(j)EK\left(\frac{x-X_j}{h(j)}\right)} = \\ & = \frac{h^{-d}(n)}{n^{-1} \sum_{j=1}^n h^{-d}(j)} \frac{1}{n^{-1} \sum_{j=1}^n h^{-d}(j)EK\left(\frac{x-X_j}{h(j)}\right)}. \end{aligned}$$

By lemma 2 and (13) the above expression is bounded mod μ . q.e.d.

Remark 1. Condition (13) is satisfied for $h(n) = cn^{-\alpha}$, $c > 0$, $\alpha > 0$.

Let $h^{-d}(n)$ vary regularly with exponent b (see Loève [10], p. 354), i.e.

$$h^{-d}(n) = n^b V_n \text{ for } b \in R \text{ and } \frac{V_{n-l}}{V_n} \rightarrow 1 \text{ as } n \rightarrow \infty \text{ for every } l \geq 1.$$

Karamata theorem ([10], p. 356) states that

$$\lim_{n \rightarrow \infty} \frac{h^{-d}(n)}{n^{-1} \sum_{i=1}^n h^{-d}(i)} < \infty$$

if $h^{-d}(n)$ varies regularly with exponent $-1 \leq b < \infty$.

Therefore condition (13) imposes restriction on the regular variation of the sequence $\{h(n)\}$; however, it does not determine its rate of convergence to zero.

Proof of theorem 1. In the proof we use the easily verified equivalence suggested by L. Györfi

$$m_n(x) = \frac{A+B}{1+C} \tag{14}$$

where

$$A = \frac{\sum_{i=1}^n h^{-d}(i)EK\left(\frac{x-X_i}{h(i)}\right)Y_i}{\sum_{j=1}^n h^{-d}(j)EK\left(\frac{x-X_j}{h(j)}\right)},$$

$$B = \frac{\sum_{i=1}^n h^{-d}(i) \left(K \left(\frac{x-X_i}{h(i)} \right) Y_i - EK \left(\frac{x-X_i}{h(i)} \right) Y_i \right)}{\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x-X_i}{h(i)} \right)},$$

$$C = \frac{\sum_{i=1}^n h^{-d}(i) \left(K \left(\frac{x-X_i}{h(i)} \right) - EK \left(\frac{x-X_i}{h(i)} \right) \right)}{\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x-X_i}{h(i)} \right)}.$$

We prove that A tends to $m(x)$ mod μ and B, C tend to 0 in probability mod μ .

$$A = \frac{\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x-X_i}{h(i)} \right) \frac{EK \left(\frac{x-X_i}{h(i)} \right) m(X_i)}{EK \left(\frac{x-X_i}{h(i)} \right)}}{\sum_{j=1}^n h^{-d}(j) EK \left(\frac{x-X_j}{h(j)} \right)}.$$

By Toeplitz lemma (Loève, p. 250), Lemma 1 and Conclusion the above expression tends to $m(x)$ mod μ as $n \rightarrow \infty$.

By Chebyshev's inequality

$$P\{B > t\} \leq t^{-2} \frac{\sum_{i=1}^n h^{-2d}(i) EK^2 \left(\frac{x-X_i}{h(i)} \right) Y_i^2}{\left[\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x-X_i}{h(i)} \right) \right]^2} \leq$$

$$\leq \frac{at^{-2}}{\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x-X_i}{h(i)} \right)} \times$$

$$\times \sum_{j=1}^n h^{-d}(j) EK \left(\frac{x-X_j}{h(j)} \right) \frac{EK^2 \left(\frac{x-X_j}{h(j)} \right) g(X_j)}{EK^2 \left(\frac{x-X_j}{h(j)} \right)} \times$$

$$\times \frac{h^{-d}(j)}{\sum_{i=1}^n h^{-d}(i) EK \left(\frac{x-X_i}{h(i)} \right)},$$

where $g(X) = E(Y^2 | X)$.

By Toeplitz lemma as well as Lemmas 1 and 3 the above expression tends to 0 mod μ . Consistency of term C may be shown similarly to that of term B by replacing $Y_i \equiv 1$ for $i=1, 2, \dots, n$. It concludes the proof of Theorem 1.

Remark 1. Weak consistency may be also proved by considering the following type of convergence $E|m_n(x) - m(x)|^p \rightarrow 0 \text{ mod } \mu$ as $n \rightarrow \infty$ for $p \geq 1$.

It was obtained under the moment assumption $E|Y|^p < \infty, p \geq 1$ and a slightly different condition imposed on K and $\{h(n)\}$ by Krzyżak and Pawlak [9].

Remark 2. If we additionally assume in (6) that there exists a positive constant d such that $K(x) \geq dI_{\{\|x\| \leq r\}}^{(x)}$, then condition (7) may be deleted.

3. Strong consistency

Theorem 2. If $EY^2 < \infty$, the monotonic sequence $\{h(n)\}$ satisfies the following conditions

$$h(n) \rightarrow 0 \text{ as } n \rightarrow \infty, \quad \sum_{n=1}^{\infty} n^{-2} h^{-d}(n) < \infty \tag{15}$$

$$\overline{\lim}_n \frac{h^{-d}(n)}{n^{-1} \sum_{i=1}^n h^{-d}(i)} < \infty \tag{16}$$

and the kernel K satisfies conditions (6), (7) of Theorem 1, then (3) holds.

Proof. In the proof we use equality (14). It is sufficient to show that term $B \rightarrow 0$ and $C \rightarrow 0$ a.s. mod μ as $n \rightarrow \infty$.

We apply Kolmogorov's second moment version of the strong law of large numbers (Loève [10], p. 250).

To show that $B \rightarrow 0$ a.s. mod μ as $n \rightarrow \infty$ we should verify whether

$$\sum_{n=1}^{\infty} \frac{h^{-2d}(n) EK^2\left(\frac{x - X_n}{h(n)}\right) Y_n^2}{b_n^2} < \infty \text{ mod } \mu \tag{17}$$

where

$$b_n = \sum_{i=1}^n h^{-d}(i) EK\left(\frac{x - X_i}{h(i)}\right)$$

To prove that $C \rightarrow 0$ a.s. mod μ as $n \rightarrow \infty$, it is sufficient to assume in (17) $Y_n \equiv 1$.

The sum on the left side of (17) may be upper bounded as follows:

$$a \sum_{n=1}^{\infty} n^{-2} h^{-d}(n) M_n(x) u_n^2(x) \frac{h^d(n)}{EK \left(\frac{x - X_n}{h(n)} \right)} \quad (18)$$

where

$$M_n(x) = \frac{EK^2 \left(\frac{x - X_n}{h(n)} \right) g(X_n)}{EK^2 \left(\frac{x - X_n}{h(n)} \right)}$$

$$u_n(x) = \frac{h^{-d}(n) EK \left(\frac{x - X_n}{h(n)} \right)}{n^{-1} \sum_{i=1}^n h^{-d}(i) EK \left(\frac{x - X_i}{h(i)} \right)}$$

By Lemma 1, $M_n(x) \rightarrow g(x) \bmod \mu$ as $n \rightarrow \infty$.

Moreover

$$\frac{h^d(n)}{EK \left(\frac{x - X_n}{h(n)} \right)} \leq \frac{h^d(n)}{\beta \mu(S_{x, rh(n)})}$$

for some $\beta, r > 0$ and by (11) it is bounded mod μ .

$u_n(x)$ may be estimated as follows

$$u_n(x) \leq \frac{h^{-d}(n)}{n^{-1} \sum_{i=1}^n h^{-d}(i)}$$

by the monotonicity of the sequence $\{h(n)\}$ and assumption (16). It concludes the proof of Theorem 2.

Remark 3. If the measure μ is atomic then the monotonicity assumption on the sequence $\{h(n)\}$ in Theorem 2 may be deleted. Moreover, for absolutely continuous measures μ conditions (5) and (16) in Theorems 1 and 2 are redundant. In the latter case the assumptions on the sequence $\{h(n)\}$ are the same as in Devroye, Wagner [2].

4. Nonparametric discrimination

In discrimination Y is a $\{1, \dots, M\}$ valued random variable and X takes values in R^d .

Given a sequence $\{(X_1, Y_1), \dots, (X_n, Y_n)\} = V_n$ of independent random couples and X we estimate Y , say $\hat{Y} = \psi(X, V_n) = \psi_n(X)$. The probability of error for the given estimate and learning sequence V_n is

$$L(\psi_n) = L_n = P\{\psi_n(X) \neq Y | V_n\}.$$

Let

$$L(\psi^*) = L^* = \inf_{\psi: R^d \rightarrow \{1, \dots, M\}} P\{\psi(X) \neq Y\}$$

denote the Bayes probability of error. The Bayes classification rule is defined as follows:

$$\psi^*(X) = i \quad \text{if} \quad \begin{cases} P_i(X) > P_j(X) & j < i \\ P_i(X) \geq P_j(X) & j > i \end{cases}$$

where $P_i(X)$ is the a posteriori probability of the event $\{Y = i\}$ given X .

An unknown P_i can be estimated by

$$P_{in}(X) = \frac{\sum_{j=1}^n h^{-d}(j) K\left(\frac{x - X_j}{h(j)}\right) I_{\{Y_j = i\}}}{\sum_{s=1}^n h^{-d}(s) K\left(\frac{x - X_s}{h(s)}\right)}$$

and ψ_n can then be picked such that

$$\psi_n(X) = i \quad \text{if} \quad \begin{cases} P_{in}(X) > P_{jn}(X), & j < i \\ P_{in}(X) \geq P_{jn}(X), & j > i, \end{cases} \tag{19}$$

The classification rule ψ_n is called weakly (strongly) universally Bayes risk consistent if $L_n \rightarrow L^*$ in probability (a.s.) for all distributions of (X, Y) .

Using the bound of Györfi [7]

$$0 \leq L_n - L^* \leq \sum_{i=1}^M \int |P_{in}(x) - P_i(x)| \mu(dx)$$

and Theorems 1, 2 and Lebesque's Dominated Convergence Theorem, we get

Theorem 3. If kernel K satisfies conditions (6), (7) of Theorem 1 and for the sequence $\{h(n)\}$ conditions (4), (5), ((15), (16)) hold then the discrimination rule (19) is universally weakly (strongly) Bayes risk consistent.

References

1. *Ahmad, I. A., Lin, P. A.*, Nonparametric Sequential estimation of a multiple regression function, *Bull. Math. Statist.* vol. **17**, pp. 63–75, 1976.
2. *Devroye, L., Wagner, T. J.*, "On the L1 convergence of kernel estimators of regression functions with applications in discrimination". *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete.* vol. **51**, pp. 15–25, 1980.
3. *Devroye, L., Wagner, T. J.*, Distribution — free consistency results in nonparametric discrimination and regression function estimation, *Annals of Statistics*, vol. **8**, pp. 231–239, 1980.
4. *Devroye, L.*, On the almost everywhere convergence of nonparametric regression function estimates. To appear in *Annals of Statistics*, 1981.
5. *Greblicki, W.*, Asymptotically optimal probabilistic algorithms for pattern recognition and identification. Monografie No 3, Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wrocławskiej, Wrocław, Poland 1974.
6. *Greblicki, W., Krzyżak, A.*, Asymptotic properties of kernel estimates of a regression function, *Journal of Statistical Planning and Inference* vol. **4**, pp. 81–90, 1980.
7. *Györfi, L.*, On the rate of convergence of nearest neighbor rules. *IEEE Trans. on Information Theory*, vol. **IT-24**, pp. 509–512, 1978.
8. *Györfi, L.*, Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, vol. **10**, pp. 43–52, 1981.
9. *Krzyżak, A., Pawlak, M.*, Almost everywhere convergence of recursive kernel regression function estimates. Second Panonian Symposium on Mathematical Statistics, Bad Tatzmannsdorf, June 1981.
10. *Loève, M.*, *Probability Theory I.* 4th Edition, 1977.
11. *Stone, C. J.*, Consistent nonparametric regression, *Annals of Statistics*, vol. **8**, pp. 585–645, 1977.
12. *Wheeden, R. L., Zygmund, A.*, *Measure and Integral*, Marcel Dekker, 1977.
13. *Wolverton, C. T., Wagner, T. J.*, Asymptotically optimal discriminant functions for pattern classification. *IEEE Trans. on Information Theory*, vol. **IT-15**, pp. 258–265, 1969.

**Универсальная состоятельность рекурсивных оценок
типа Вольвертона–Вагнера и их применение
в классификации**

А. КРЖИЖАК, М. ПАВЛАК

(Вроцлав)

В статье исследована слабая и сильная универсальная состоятельность непараметрических оценок с особым интересом по ядерным функциям и использование полученных результатов в классификации.

A. Krzyżak

M. Pawlak

Institute of Engineering Cybernetics

Technical University of Wrocław

50-370 Wrocław

Poland

ЭКСТРЕМАЛЬНЫЕ СВОЙСТВА ЭЛЛИПСОИДОВ, АППРОКСИМИРУЮЩИХ ОБЛАСТИ ДОСТИЖИМОСТИ

А. И. ОВСЕВИЧ

(Москва)

(Поступила в редакцию 6 января 1982 г.)

Исследуется эволюция областей достижимости управляемых систем и эллипсоидов, аппроксимирующих эти области снаружи и изнутри. Найдены некоторые экстремальные свойства рассматриваемых эллипсоидов.

1. Введение

Рассмотрим линейную управляемую систему

$$\dot{x} = A(t)x + u, \quad u \in U(t), \quad x(s) \in D(s) \quad (1.1)$$

Здесь $x \in \mathbf{R}^n$ — фазовый вектор, $A(t)$ — заданная $n \times n$ -матрица, $U(t)$ — замкнутая область допустимых значений управления, $D(s)$ — замкнутое множество, откуда начинается движение в момент s (начальная область).

Множество концов $x(t)$ траекторий, начинающихся в $D(s)$ и удовлетворяющих (1.1) для некоторой измеримой вектор-функции $u(t) \in U(t)$ — допустимого управления, называется областью достижимости

$$D(t) = D(s, t) = D(s, D(s), t)$$

системы (1.1)

Знание областей достижимости необходимо в теории управления при гарантированном оценивании (фильтрации) динамических систем, в дифференциальных играх (см., например, [1, 2, 3]). Так, задача минимизации терминального функционала $F(x(T))$, где T — фиксированный момент, эквивалентна минимизации функции $F(x)$ на области достижимости $D(T)$.

Однако эффективное построение областей достижимости сталкивается с большими трудностями, связанными с необходимостью бесконечного числа параметров для задания области в \mathbf{R}^n . В ряде работ разных авторов (см., например, [4, 5]) был развит метод приближения областей достижимости областями заданной формы. В частности, в [5, 6] получены обыкновенные дифференциальные уравнения, описывающие эволюцию эллипсоидов, аппроксимирующих снаружи и изнутри области достижимости.

В начале настоящей работы приведено уравнение эволюции областей достижимости (раздел 2). На его основе ставится и решается задача о наилучшем приближении областей достижимости эллипсоидами (разделы 3, 4). Показано, что решение задачи получается интегрированием упомянутых выше дифференциальных уравнений.

2. Уравнения эволюции областей достижимости

С любым подмножеством $D \subset R^n$ связана его опорная функция

$$H(\zeta) = H_D(\zeta) = \sup_{x \in D} (x, \zeta)$$

где $\zeta \in R^n$, (\cdot, \cdot) — скалярное произведение. Если D замкнуто и выпукло, то H_D задает D однозначно [6]. Заметим, что если области управления $U(t)$ и начальная область $D(s)$ системы (1.1) замкнутые и выпуклые, то такой же является и область достижимости $D(t)$. Поэтому следующее предложение, описывающее эволюцию опорной функции области достижимости $D(t)$, определяет в этом случае эволюцию областей достижимости.

Предложение. Пусть функции $A(t)$ и множества $U(t)$ в (1.1) измеримо [7] зависят от t , $H(t, \zeta) = H_{D(t)}(\zeta)$ — опорная функция области достижимости $D(t)$, $h(t, \zeta) = H_{u(t)}(\zeta)$ — опорная функция области управления $U(t)$.

Тогда

$$\frac{\partial H}{\partial t}(t, \zeta) = \left(A(t) \frac{\partial H}{\partial \zeta}(t, \zeta), \zeta \right) + h(t, \zeta) \quad (2.1)$$

где $\partial/\partial \zeta$ означает градиент по ζ .

(Здесь $(A(\partial H/\partial \zeta), \zeta) = (\partial H/\partial \zeta, A^* \zeta)$ понимается как производная по направлению $\lim_{\varepsilon \downarrow 0} \varepsilon^{-1}(H(\zeta + \varepsilon A^* \zeta) - H(\zeta))$ при $\varepsilon \downarrow 0$). Соответствующий предел существует для любой выпуклой функции [7]).

Замечание. Аналогичная формула при нескольких иных предположениях получена в [8]. Для удобства читателя здесь, однако, приводится доказательство. Отметим, что зависимость $D(t)$ от начальной области $D(s)$ и момента s в случае, когда $D(s)$ состоит из одной точки x описывается хорошо известным уравнением Беллмана.

Доказательство. Покажем сначала, что достаточно рассмотреть случай $A(t) \equiv 0$ в (1.1), (2.1). Действительно, пусть матрица $P(t)$ (фундаментальная матрица) является решением задачи Коши:

$$\dot{P} = A(t)P, \quad P(s) = I \quad (2.2)$$

где I — единичная матрица. Тогда замена переменных $x = P(t)y$ преобразует систему (1.1) в

$$\dot{y} = v, \quad v \in V(t) = P(t)^{-1}U(t), \quad y(s) \in D(s) \quad (2.3)$$

Области достижимости $D(t)$ для (1.1) переходят при этом в области достижимости $D^0(t) = P(t)^{-1}D(t)$ для (2.3), а соответствующие опорные функции $H(t, \xi)$ в

$$H^0(t, \xi) = \sup_{y \in D^0(t)} (y, \xi) = \sup_{x \in D(t)} (x, P(t)^{* -1} \xi) = H(t, P(t)^{* -1} \xi).$$

Прямая выкладка показывает, что преобразование

$$H(t, \xi) \rightarrow H^0(t, \xi) = H(t, P(t)^{* -1} \xi)$$

переводит решения (2.1) в решения

$$\frac{\partial H^0}{\partial t}(t, \xi) = h^0(t, \xi) = H_{V(t)}(\xi). \quad (2.4)$$

В силу обратимости P остается доказать формулу (2.4) для опорных функций $H^0(t, \xi)$ областей достижимости $D^0(t)$ системы (2.3).

Имеем

$$H^0(t, \xi) = \sup_s \left(\int_s^t v(\tau) d\tau, \xi \right)$$

где \sup берется по всевозможным измеримым вектор-функциям $v(t) \in V(t)$. Из теоремы об измеримом выборе [7] следует, что

$$\sup_s \left(\int_s^t v(\tau) d\tau, \xi \right) = \int_s^t \sup_{v \in V(\tau)} (v, \xi) d\tau = \int_s^t h^0(\tau, \xi) d\tau,$$

что и требовалось доказать.

В дальнейшем будем рассматривать только случай, когда области управления $U(t)$ и начальная область $D(s)$ замкнуты и выпуклы.

Определение 1. Пусть $\Omega(t), t \geq s$ — семейство замкнутых выпуклых множеств, $H(t, \xi) = H_{\Omega(t)}(\xi)$ — соответствующее семейство опорных функций. Скажем, что $\Omega(t)$ — семейство областей субдостижимости (соответственно супердостижимости) для (1.1), если выполнено дифференциальное неравенство

$$\frac{\partial H}{\partial t}(t, \xi) \leq \left(A(t) \frac{\partial H}{\partial \xi}(t, \xi), \xi \right) + h(t, \xi), \quad H(s, \xi) = H_{D(s)}(\xi) \quad (2.5)$$

(соответственно со знаком \geq).

Включение $\Omega_1 \subset \Omega_2$ замкнутых выпуклых множеств эквивалентно неравенству $H\Omega_1(\xi) \leq H\Omega_2(\xi)$ для их опорных функций, поэтому из теоремы 1 следует, что $\Omega(t) \subset D(t)$ в случае субдостижимости и $\Omega(t) \supset D(t)$ в случае супердостижимости. Более того, если $\tau \leq t$ и $D(\tau, t) = D(\tau, \Omega(\tau), t)$ — область достижимости в момент t системы

$$\dot{x} = A(t)x + u, u \in U(t), x(\tau) \in \Omega(\tau) \quad (2.6)$$

то $D(\tau, t) \supset \Omega(t)$ в случае субдостижимости и $D(\tau, t) \subset \Omega(t)$ в случае супердостижимости.

Определение 2. Пусть $V(\Omega)$ — объем множества Ω , E — некоторый класс выпуклых множеств. Скажем, что семейство $\Omega(t) \in E$ областей суб- (супер-) достижимости локально наилучшим образом приближает области достижимости $D(t)$ системы (1.1), если при $\tau \geq s$

$$\frac{d}{dt} V(\Omega(t))|_{t=\tau}$$

достигает максимума (минимума) среди всех семейств областей суб- (супер-) достижимости из E для системы (2.6).

Таким образом, семейство $\Omega(t)$ дает наилучшую локальную аппроксимацию области достижимости по критерию объема. Теперь можно точно сформулировать задачу о приближении областей достижимости областям заданной формы. Задача состоит в эффективном построении локально наилучшего семейства областей суб- и супердостижимости из заданного класса E . В следующем разделе этот вопрос решается для случая, когда E — класс всех эллипсоидов.

3. Уравнения эволюции аппроксимирующих эллипсоидов.

Обозначим эллипсоид

$$\{x \in \mathbf{R}^n, (Q^{-1}(x-a), x-a) \leq 1\}$$

через $E(a, Q)$. Здесь $a \in \mathbf{R}^n$, Q — положительно определенная симметрическая матрица. Предположим, что начальный момент $s=0$, начальная область $D(0)$ системы (1.1) есть эллипсоид $E(a_0, Q_0)$, а области управления $U(t)$ — эллипсоиды $E(b(t), G(t))$. Тогда эволюцию эллипсоидов, локально наилучшим образом приближающих области достижимости системы

$$\dot{x} = A(t)x + u, u \in E(b(t), G(t)), x(0) \in E(a_0, Q_0) \quad (3.1)$$

можно описать дифференциальными уравнениями (3.3), (3.4) из теоремы 1.

Предварительно приведем две леммы.

Лемма 1. [3]. Опорная функция $H(\xi)$ эллипсоида $E(a, Q)$ задается формулой

$$H(\xi) = (a, \xi) + (Q\xi, \xi)^{1/2}$$

Лемма 2 [9]. Пусть $A(t)$ — семейство обратимых матриц, гладко зависящих от t . Тогда

$$\frac{d}{dt} \log \det A(t) = \text{Tr } A(t)^{-1} \dot{A}(t)$$

Следствие. Пусть $E(a(t), Q(t))$ — семейство эллипсоидов, $V(t)$ — объем $E(a(t), Q(t))$. Тогда

$$\frac{d}{dt} V(t) = \frac{1}{2} V(t) \text{Tr } Q(t)^{-1} \dot{Q}(t) \quad (3.2)$$

Действительно, $V(t) = \omega_n [\det Q(t)]^{1/2}$, где ω_n — объем единичного шара $E(0, I)$, и формула (3.2) получается из леммы 2. Введем обозначение $\{\alpha, \beta\} = \alpha\beta + \beta^*\alpha^*$ для матриц α, β .

Теорема 1. 1) Пусть $a_-(t), Q_-(t)$ — решение задачи Коши.

$$\begin{aligned} \dot{a}_- &= A(t)a_- + b(t), \quad a_-(0) = a_0 \\ \dot{Q}_- &= \{A(t), Q_-\} + 2R^{-1}(RQR^*)^{1/2}(RG(t)R^*)^{1/2}R^{*-1} \\ Q_-(0) &= Q_0, \end{aligned} \quad (3.3)$$

где R — такая невырожденная матрица, что $Q^0 = RQ_-R^*$ и $G^0 = RG(t)R^*$ — диагональные матрицы. Тогда $E(a_-(t), Q_-(t))$ — локально наилучшее семейство эллипсоидальных областей субдостижимости для системы (3.1).

2) Пусть $a_+(t), Q_+(t)$ — решение задачи Коши

$$\begin{aligned} \dot{a}_+ &= A(t)a_+ + b(t), \quad a_+(0) = a_0 \\ \dot{Q}_+ &= \{A(t), Q_+\} + hQ_+ + h^{-1}G(t), \quad Q_+(0) = Q_0 \\ h &= [n^{-1} \text{Tr } (Q_+^{-1}G(t))]^{1/2} \end{aligned} \quad (3.4)$$

где Tr обозначает след. Тогда $E(a_+(t), Q_+(t))$ — локально наилучшее семейство эллипсоидов супердостижимости для системы (3.1).

Замечание. Хорошо известно (см., например, [9]), что если A и B две симметрические матрицы, одна из которых положительно определенная, то существует такая невырожденная матрица R , что RAR^* и RBR^* — диагональные. Выражение $R^{-1}(RAR^*)^{1/2}(RBR^*)^{1/2}R^{*-1}$ несмотря на возможную неоднозначность R , зависит только от A и B и поэтому правая часть (3.3) корректно определена. Доказательство теоремы разобьем на сколько шагов.

Шаг 1. Упрощение системы (3.1). Пусть матричная функция $P(t)$ (фундаментальная матрица) является решением задачи Коши:

$$\dot{P} = A(t)P, \quad P(0) = I. \quad (3.5)$$

Сделаем замену переменных $x = P(t)y$. Тогда управляемая система (3.1) преобразуется в

$$\begin{aligned} \dot{y} &= v, \quad v \in V(t) = P(t)^{-1}U(t) = E(b_1(t), G_1(t)) \\ b_1(t) &= P(t)^{-1}b(t), \quad G_1(t) = P(t)^{-1}G(t)P(t)^{*^{-1}} \\ y(0) &\in E(a_0, Q_0). \end{aligned} \quad (3.6)$$

Очевидно, $P(t)$ переводит эллипсоиды в эллипсоиды, а области суб- (супер-) достижимости системы (3.6) в соответствующие области для (3.1). (эллипсоид $E(a(t), Q(t))$ получается под действием $P(t)$ из эллипсоида $E(P^{-1}a, P^{-1}QP^{*^{-1}})$. Нетрудно проверить прямыми выкладками, что если $Q(t), a(t)$ удовлетворяют дифференциальным уравнениям (3.3) или (3.4), то параметры $P(t)^{-1}a(t), P(t)^{-1}Q(t)P(t)^{*^{-1}}$ эллипсоида $P^{-1}E(a, Q)$ удовлетворяют системе, аналогичной (3.3) или (3.4), в которой $A(t) \equiv 0$, а $G = G_1$ задано в (3.6). Следовательно, в силу обратимости $P(t)$, дальнейшее доказательство теоремы достаточно провести, полагая $A(t) \equiv 0$ в (3.1). Замена переменных $x = y + r(t)$ где

$$\dot{r} = b(t), \quad r(0) = a_0,$$

позволяет после этого проводить доказательство только для системы вида

$$\dot{x} = u, \quad u \in E(0, G(t)), \quad x(0) \in E(0, Q_0). \quad (3.7)$$

Шаг 2. Переформулировка задачи. Пусть $E(a(t), Q(t))$ — семейство эллипсоидов субдостижимости для системы (3.7). Согласно лемме 1 и определению 1, это означает

$$\frac{d}{dt} [(a(t), \xi) + (Q(t)\xi, \xi)^{1/2}] \leq (G(t)\xi, \xi)^{1/2}, \quad a(0) = 0, \quad (3.8)$$

$$Q(0) = Q_0,$$

иначе говоря,

$$(\dot{a}, \xi) + \frac{1}{2} (Q\xi, \xi)^{-1/2} (\dot{Q}\xi, \xi) \leq (G\xi, \xi)^{1/2}, \quad \forall \xi \in R^n. \quad (3.9)$$

Формула (3.2) позволяет переписать условие локальной оптимальности (определение 2) семейства $E(a(t), Q(t))$ в виде:

$$\text{Tr } Q^{-1}\dot{Q} \rightarrow \max \quad \text{по } a, \dot{Q} \quad (3.10)$$

при фиксированном $Q = Q(t)$ среди всевозможных векторов \dot{a} и симметрических матриц \dot{Q} , для которых выполнено неравенство (3.9). Для случая супердостижимости \max в (3.10) нужно заменить на \min , а также изменить знак \geq на \leq в (3.9). Заменяя в (3.9) ξ на ζ и складывая полученное неравенство с (3.9), имеем

$$\frac{1}{2} (Q\zeta, \zeta)^{-1/2} (\dot{Q}\zeta, \zeta) \leq (G\zeta, \zeta)^{1/2}. \quad (3.11)$$

Поэтому, если пара (\dot{a}, \dot{Q}) доставляет максимум в (3.10), то пара $(0, \dot{Q})$ также доставляет этот максимум. В дальнейшем полагаем $a \equiv 0$.

Пусть R — такая невырожденная матрица, что $Q_1 = RQR^*$, $G_1 = RGR^*$ — диагональные матрицы. Положим $C = R\dot{Q}R^*$, $\zeta = R\eta$. Тогда $\text{Tr } Q^{-1}\dot{Q} = \text{Tr } Q_1^{-1}C$ и экстремальная задача (3.9), (3.10) сводится к следующей. Найти симметрическую матрицу C , доставляющую максимум

$$\text{Tr } Q_1^{-1}C \rightarrow \max \quad (3.12)$$

$$(C\eta, \eta) \leq 2(G_1\eta, \eta)^{1/2}(Q_1\eta, \eta)^{1/2}, \forall \eta \in \mathbf{R}^n$$

где G_1, Q_1 — заданные диагональные матрицы. В случае супердостижимости приходим к задаче

$$\text{Tr } Q_1^{-1}C \rightarrow \min \quad (3.13)$$

$$(C\eta, \eta) \geq 2(G_1\eta, \eta)^{1/2}(Q_1\eta, \eta)^{1/2}, \forall \eta \in \mathbf{R}^n.$$

Для доказательства теоремы нужно установить, что одно из решений задачи (3.12) дается формулой

$$C = 2Q_1^{1/2}G_1^{1/2} \quad (3.14)$$

а задачи (3.13) — формулой

$$C = hQ_1 + h^{-1}G_1, h = [h^{-1} \text{Tr } (Q_1^{-1}G_1)]^{1/2}. \quad (3.15)$$

В самом деле, из (3.14), (3.15) с учетом связи $C = R\dot{Q}R^*$ и аналогичных формул для Q_1, G_1 вытекают уравнения (3.3), (3.4) для Q_{\pm} , в которых $A(t) \equiv 0$. Из того факта, что $a \equiv 0$, вытекает также справедливость уравнений (3.3), (3.4) для a_-, a_+ при $A \equiv 0, b \equiv 0, a_0 = 0$.

Шаг 3. Решение преобразованной задачи. Покажем, что решение C экстремальной задачи (3.12) или (3.13) можно искать среди диагональных матриц. Действительно, пусть C — некоторое решение, Γ — группа диагональ-

ных матриц с диагональными элементами ± 1 . Тогда, если $g \in \Gamma$, то $g^* C g$ — также, как и C , является решением, поскольку

$$g^* = g^{-1} = g, \quad g^{-1} G_1 g = G, \quad g^{-1} Q_1 g = Q_1 \\ \text{Tr } Q_1^{-1} C = \text{Tr } (g^{-1} Q_1^{-1} g^{-1} C g) = \text{Tr } (Q_1^{-1} g^{-1} C g).$$

Неравенство (3.12) и (3.13) линейны по C , поэтому комбинация

$$C_1 = 2^{-n} \sum_{g \in \Gamma} g^* C g \quad (3.16)$$

решений $g^* C g$ также есть решение. Но C_1 — это диагональная матрица с теми же диагональными элементами, что и C . В самом деле, матричный элемент

$$(C_1)_{ij} = 2^{-n} \sum_{g \in \Gamma} g_{ii} g_{jj} C_{ij} = C_{ij} (2^{-n} \sum_{g \in \Gamma} g_{ii} g_{jj}) = C_{ij} \varphi_{ij}.$$

Если $i=j$, то $\varphi_{ij} = 2^{-n} \sum_{g \in \Gamma} 1 = 1$. Если же $i \neq j$, то поставим в соответствие матрице $g = \text{diag}(g_{kk}) \in \Gamma$ матрицу $g^0 = \text{diag}(g_{kk}^0) \in \Gamma$, заданную формулой $g_{kk}^0 = g_{kk}$ при $k \neq j$, $g_{jj}^0 = -g_{jj}$. Очевидно, разбивая $\sum_{g \in \Gamma} g_{ii} g_{jj}$ на суммы по парам соответствующих элементов, получим $\varphi_{ij} = 0$.

Поэтому в дальнейшем можно ограничиться поиском диагональной матрицы $C = (c_i \delta_{ij})$. Обозначим через p, q, r векторы с компонентами $p_i = (G_1)_{ii}$, $q_i = (Q_1)_{ii}$, $r_i = (Q_1^{-1})_{ii} = q_i^{-1}$. Тогда задачи (3.12), (3.13) сводятся к следующим двум задачам:

$$(r, c_-) \rightarrow \max \quad (3.17)$$

$$(c_-, x) \leq 2(p, x)^{1/2} (q, x)^{1/2}$$

где $x \in R^n$ — произвольный вектор с неотрицательными компонентами (квадратами компонент вектора η из (3.12)),

$$(r, c_+) \rightarrow \min \quad (3.18)$$

$$(c_+, x) \geq 2(p, x)^{1/2} (q, x)^{1/2}, \quad \forall x \in R^n, x_i \geq 0.$$

Подставим в (3.17) $x = e_i$ — i -й орт. Получим $(c_-)_i \leq 2p_i^{1/2} q_i^{1/2}$. Так как r — вектор с положительными компонентами, то вектор c_- с компонентами $(c_-)_i = 2p_i^{1/2} q_i^{1/2}$, удовлетворяющий ограничению (3.17), является решением задачи (3.17). Таким

образом, нужно проверить неравенство

$$\sum_{i=1}^n p_i^{1/2} q_i^{1/2} x_i \leq \left(\sum_{i=1}^n p_i x_i \right)^{1/2} \left(\sum_{i=1}^n q_i x_i \right)^{1/2}, \forall x, x_i \geq 0.$$

Но это есть неравенство Коши-Буняковского для векторов с компонентами $(p_i x_i)^{1/2}, (q_i x_i)^{1/2}$.

Таким образом, задача (3.17) решена и ее решение c_- приводит в точности к формуле (3.14).

Перейдем к решению задачи (3.18). Полагая в (3.18) $x=r$, получим

$$(r, c_+) \geq 2(p, r)^{1/2}(q, r)^{1/2} = 2n^{1/2}(p, r)^{1/2}. \quad (3.19)$$

С другой стороны, полагая $h = n^{-1/2}(p, r)$ и $c_+ = hq + h^{-1}p$, имеем

$$(r, c_+) = 2n^{1/2}(p, r)^{1/2}. \quad (3.20)$$

Проверим ограничение (3.18) для c_+

$$\begin{aligned} (c_+, x) &= h(q, x) + h^{-1}(p, x) \geq 2[h(q, x)]^{1/2}[h^{-1}(p, x)]^{1/2} = \\ &= 2(q, x)^{1/2}(p, x)^{1/2}. \end{aligned}$$

Из полученного неравенства и (3.19), (3.20) следует, что c_+ — решение задачи (3.18). Нетрудно проверить, что оно приводит к формуле (3.15).

Теорема доказана.

Замечание. Уравнения (3.3), (3.4) были получены ранее в работах [5, 6] с помощью конечноразностной аппроксимации управляемой системы (3.1). Однако в них не было выяснено, в каком смысле решения этих уравнений дают наилучшее приближение к областям достижимости. Теорема 1 решает этот вопрос, устанавливая экстремальные свойства эллипсоидов, описываемых уравнениями (3.3), (3.4).

4. Единственность локально оптимальной эллипсоидальной аппроксимации

В теореме 1 было показано, что уравнения (3.3), (3.4) дают локально оптимальные семейства эллипсоидов. Покажем теперь, что других локально оптимальных семейств эллипсоидов нет.

Теорема 2. Пусть $E(a(t), Q(t))$ — локально оптимальное семейство эллипсоидов суб- (супер-) достижимости для управляемой системы (3.1). Тогда $(a(t), Q(t))$ — интегральная кривая системы (3.3) (соответственно (3.4)).

Доказательство. Рассуждения из шагов 1 и 2 в доказательстве предыдущей теоремы показывают, что в случае субдостижимости достаточно установить единственность решения экстремальной задачи:

$$\begin{aligned} \text{Tr } Q^{-1}C \rightarrow \max \text{ по } d, C \\ (d, \xi) + \frac{1}{2}(Q\xi, \xi)^{-1/2}(C\xi, \xi) \leq (G\xi, \xi)^{1/2}, \quad \forall \xi \in R^n, \end{aligned} \quad (4.1)$$

где Q, G — заданные положительные диагональные матрицы, d и C — переменные вектор и симметрическая матрица. В случае супердостижимости нужно установить единственность в аналогичной задаче

$$\begin{aligned} \text{Tr } Q^{-1}C \rightarrow \min \text{ по } d, C \\ (d, \xi) + \frac{1}{2}(Q\xi, \xi)^{-1/2}(C\xi, \xi) \geq (G\xi, \xi)^{1/2}, \quad \forall \xi. \end{aligned} \quad (4.2)$$

Как было отмечено при доказательстве теоремы 1 (см. (3.11)), решения задач (4.1), (4.2) являются также решениями задач

$$\begin{aligned} \text{Tr } Q^{-1}C \rightarrow \max \\ (C\xi, \xi) \leq 2(Q\xi, \xi)^{1/2}(G\xi, \xi)^{1/2} = \Phi(\xi) \end{aligned} \quad (4.3)$$

$$\begin{aligned} \text{Tr } Q^{-1}C \rightarrow \min \\ (C\xi, \xi) \geq 2(Q\xi, \xi)^{1/2}(G\xi, \xi)^{1/2} = \Phi(\xi) \end{aligned} \quad (4.4)$$

Докажем вначале единственность решения задач (4.3), (4.4). Пусть C — решение (4.3). Тогда получим из (4.3) при $\xi = e_i$, где e_i — i -й орт, что $C_{ii} = (Ce_i, e_i) \leq \Phi(e_i)$. Поскольку C реализует максимум величины $\sum Q_{ii}^{-1}C_{ii}$, $Q_{ii} > 0$, а для решения $C_- = 2Q^{1/2}G^{1/2}$ задачи (4.3), указанного в (3.14), имеем $\sum Q_{ii}^{-1}(e_-)_{ii} = \sum Q_{ii}^{-1}\Phi(e_i)$, то необходимо $(Ce_i, e_i) = \Phi(e_i)$. Из (4.3) следует тогда, что

$$(Cx, x) - (Ce_i, e_i) \leq \Phi(x) - \Phi(e_i) \quad \forall x \in R^n \quad (4.5)$$

Положим $x = e_i + \varepsilon y$, где $y \in R^n$, ε — вещественное число, и устремим ε к нулю. Тогда из (4.5) следует, что

$$2\varepsilon(e_i, y) \leq \varepsilon(\text{grad } \Phi(e_i), y) + o(\varepsilon) \quad (4.6)$$

Поскольку ε может быть любого знака, а y — произвольный вектор, то из (4.6) получаем, что $Ce_i = (1/2) \text{grad } \Phi(e_i)$ (здесь использован «прием Минти», известный в теории монотонных операторов [10]). Тем самым матрица C определена однозначно, $C = C_-$. Пусть теперь C — некоторое решение задачи (4.4), $C_+ = hQ + h^{-1}G$ — решение указанное в (3.15), X — множество таких векторов $x \in R^n$, што $x_i^2 = Q_{ii}^{-1}$, т.е. $x_i = \pm Q_{ii}^{-1/2}$. Тогда из (4.4) и (3.20) имеем для $x \in X$:

$$(Cx, x) \geq \Phi(x) = (C_+x, x) = \text{Tr } Q^{-1}C_+. \quad (4.7)$$

Положим в обозначениях из (3.16)

$$C^0 = 2^{-n} \sum_{g \in \Gamma} g^* C g. \quad (4.8)$$

Тогда из (4.7) следует, что

$$(C^0x, x) \geq \text{Tr } Q^{-1}C_+, \forall x \in X. \quad (4.9)$$

Из (4.8), рассуждая аналогично шагу 3 доказательства теоремы 1, получаем

$$\text{Tr } Q^{-1}C^0 = \text{Tr } Q^{-1}C. \quad (4.10)$$

Кроме того, поскольку C^0 — диагональная матрица, то

$$(C^0x, x) = \text{Tr } Q^{-1}C^0, \forall x \in X. \quad (4.11)$$

Из (4.9), (4.10) и (4.11) получаем, что

$$\text{Tr } Q^{-1}C \geq \text{Tr } Q^{-1}C_+ \quad (4.12)$$

причем, имеет место строгое неравенство, если $(Cx, x) \neq \Phi(x)$ хотя бы для одного $x \in X$. Но $\text{Tr } Q^{-1}C_+ = \text{Tr } Q^{-1}C$, поскольку C, C_+ — решения задачи (4.4). Следовательно, $(Cx, x) = \Phi(x) \forall x \in X$. То же рассуждение, что и в доказательстве единственности решения задачи (4.3), показывает, что

$$Cx = (1/2) \text{grad } \Phi(x), \quad \forall x \in X.$$

Поскольку из векторов $x \in X$ можно выделить базис R^n , то C определена однозначно, $C = C_+$.

Остается показать, что векторы d в (4.1), (4.2) должны быть нулевыми. Из приведенных выше рассуждений и ограничения (4.1) получим, что $(d, \pm e_i) \leq 0$, откуда $d = 0$. Аналогично в задаче (4.2) имеем $(d, \pm x) \leq 0, \forall x \in X$, и следовательно $d = 0$.

Теорема доказана.

Автор благодарит Ф. Л. Черноусько за полезные обсуждения.

Литература

1. Красовский Н. Н. Теория управления движением. М., «Наука», 1968.
2. Красовский Н. Н. Игровые задачи о встрече движений. М., «Наука», 1970.
3. Куржанский А. Б. Управление и наблюдения в условиях неопределенности. М., «Наука», 1977.
4. Schweppe, F. C., Recursive state estimation: unknown but bounded errors and system inputs. IEEE Trans. Automat. Control, 1968, AC-13, № 1.
5. Черноусько Ф. Л. Оптимальные гарантированные оценки неопределенностей с помощью эллипсоидов, I, II, III. Изв. АН СССР, Техническая кибернетика, I: № 3, с. 3; II: № 4, с. 3; III: № 5, с. 5.
6. Черноусько Ф. Л. Эллипсоидальные оценки области достижимости управляемой системы. ПММ, т. 45, вып. II, 1981.
7. Иоффе А. Д., Тихомиров В. М. Теория экстремальных задач. М. «Наука», 1974, 479 с.
8. Панасюк А. И. Расчет множеств достижимости систем автоматического управления с помощью уравнения для опорной функции. Депонирована в Белорусском НИИТИ № 214, 36 с. 1980.
9. Гантмахер Ф. Р. Теория матриц. М. «Наука», 1967, 575 с.
10. Лионс Ж.-Л. Некоторые методы решения нелинейных краевых задач. М. «Мир», 1972, 587 с.

Extremal properties of ellipsoids approximating attainability sets

A. I. OVSEEVICH

(Moscow)

The paper is devoted to approximate construction of attainable sets of the control dynamic systems. It continues investigations by different authors (see, e.g. [4, 5, 6]) connected with approximation of attainable sets by ellipsoids.

The main result consists of statement and solution of the problem on best local approximation of attainability sets of linear control system by ellipsoids. In particular the system of ordinary differential equations for approximating ellipsoid's parameters is obtained. The corresponding equations were obtained before in [5, 6] by finite-difference approximation of the dynamic system. It was not revealed however in [5, 6] in what sense the resulting ellipsoids were optimal.

А. И. Овсеевич
Институт проблем механики
СССР, 117526 Москва В-526,
просп. Вернадского, 101

CONTRIBUTION TO SIMULATION OF DISTRIBUTED PARAMETER SYSTEMS

Z. VOSTRÝ

(Prague)

(Received February 2, 1982)

In this paper new approach to digital simulation of some distributed parameter systems with continuous control law and restrictions is discussed from the point of view of boundary conditions determination and transformation. The main idea of this paper is the recomputation of simple boundary conditions into points with complicated boundary conditions in order to simplify the simulation.

It is not possible to describe many real systems as systems with lumped parameters. Their dynamic behaviour is then usually described in a way of mathematical-physical analysis, by partial differential equations. The control of such systems is not easy, especially if they are nonlinear. The existence of a corresponding simulation model is an efficient aid for the designer in this case. Considering more complicated systems or so-called large scale systems, digital simulation models will be used. The developing of a suitable model brings many problems to be solved: time and space discretisation, frequency analysis, simplification of partial differential equations, choice of integration method, etc. Work on space discretisation problem includes decision making about the state of model and about choice of input and output variables.

It is necessary to join initial and boundary conditions to the system of equations describing the system dynamics, to make the simulation model complete.

The initial conditions determination makes usually no problems. They are given as some equilibrium solution depending on the physical and technological sense of the solved problem.

Considering the boundary conditions, quite a different situation exists. Their importance follows from the tasks of simulation. Two main purposes of simulation are as follow:

- i) to observe and study the dynamics of a system when the boundary conditions are given. Let us note that some of real controlled systems with continuous controllers may be, from the simulation point of view, considered as systems with static boundary conditions. The last ones define relations between input and output variables together with restrictions following from the system's dynamic behaviour and technology.

ii) to determine boundary conditions with the aim to obtain desired behaviour. Especially time optimal control may be solved using simulation model. Special boundary conditions and restrictions determination follow then from maximum principle.

These two tasks may be interpreted in terms of control analysis if a system including controller is given, or in terms of system synthesis if a system and control requirements are given

The boundary conditions represent that part of the described real system which is available by the designer for experiments on the computer.

The boundary conditions can be given either as input variables function of time, or as a relation between input and output variables. The former case corresponds to the input variables representing the influence of uncontrolled surroundings (e.g. weather), the latter one the input variables used for control.

In the first case the introduction of boundary conditions in a simulation process makes no difficulties.

But in the second case if the relation between input and output variables is given in complicated form including nonequalities, hard difficulties may arise. For illustration let us note that in solution of time optimal control the input variable is usually defined as a maximum of some input variable function, output variable function and function of both.

Digital simulation model of any real dynamic system can be represented in a form corresponding to the structure in Fig. 1, where \mathbf{u}_1 , \mathbf{u}_2 are arbitrarily chosen input variables and \mathbf{y}_1 , \mathbf{y}_2 are output variables. How did we separated input and output variables into two groups denoted \mathbf{u}_1 , \mathbf{y}_1 and \mathbf{u}_2 , \mathbf{y}_2 ?

The vectors \mathbf{u}_1 and \mathbf{y}_1 include such input and output variables for which it holds

$$\mathbf{u}_1 + \mathbf{a}(t)\mathbf{y}_1 = \mathbf{c}(t) \quad (1)$$

where $\mathbf{a}(t)$ is time dependent matrix,

$\mathbf{c}(t)$ is time dependent vector.

The vectors \mathbf{u}_2 and \mathbf{y}_2 consist of such input and output variables for which it holds

$$f(\mathbf{u}_2, \mathbf{y}_2) = 0, \quad (2)$$

$$g(\mathbf{u}_2, \mathbf{y}_2) < 0 \quad (3)$$

where the functions f and g are nonlinear.

This second case could be, naturally, transformed to Eq. (1) if the function f is in the full range of its arguments well linearisable and inequality (3) does not exist.

Why do we separate the input and output variables into two groups \mathbf{u}_1 , \mathbf{y}_1 and \mathbf{u}_2 , \mathbf{y}_2 ?

Two different combinations of the above-mentioned boundary condition classes can be met:

- i) the vectors \mathbf{u}_2 and \mathbf{y}_2 are empty. In this case there are no difficulties in the boundary conditions implementation.
- ii) the vectors $\mathbf{u}_1, \mathbf{y}_1, \mathbf{u}_2, \mathbf{y}_2$ or $\mathbf{u}_2, \mathbf{y}_2$ only exist in a solved simulation problem. For this most frequent case our method was elaborated. It is based on the idea how to decrease the difficulties arising from (2), (3) using boundary condition of type (1). Here the reason for classification of boundary conditions into two groups can be seen.

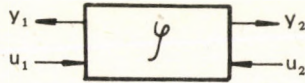


Fig. 1

System \mathcal{S} in Fig. 1 represents for example a gas pipe line, river channel, electric line, etc. A dynamic system with distributed parameters is usually described by partial differential equation. It follows from practical experience with solution and simulation of such systems that implicit integration method and linearisation is the adequate choice to obtain a system of linear algebraic equations in each integration step. The model of the system in Fig. 1 is then described by the following set of algebraic equations

$$\mathbf{A} \begin{bmatrix} \mathbf{y}_1 \\ \hat{\mathbf{y}} \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{B} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \mathbf{C} \quad (4)$$

where $\hat{\mathbf{y}}$ is a vector of inner variables (state) in time $t + \Delta t$

$\mathbf{y}_1, \mathbf{y}_2$ are output variable vectors in time

$\mathbf{u}_1, \mathbf{u}_2$ are input variable vectors in time

Δt is integration step

\mathbf{A}, \mathbf{B} are matrices of coefficients (they are computed for nonlinear system in each integration step)

The vector \mathbf{C} is given as

$$\mathbf{C} = \mathbf{E} \begin{bmatrix} \mathbf{y}_1^* \\ \hat{\mathbf{y}}^* \\ \mathbf{y}_2^* \end{bmatrix} + \mathbf{F} \begin{bmatrix} \mathbf{u}_1^* \\ \mathbf{u}_2^* \end{bmatrix} + \mathbf{G} \quad (5)$$

where \mathbf{E}, \mathbf{F} (similar \mathbf{A}, \mathbf{B}) are matrices of coefficients computed in each time-step,

\mathbf{G} is a vector given by terms from linearisation,

y_1^*, y_2^*, \hat{y}^* are variables like y_1, y_2, \hat{y} but in time t ,

u_1^*, u_2^* are variables like u_1, u_2 but in time t .

Let us rearrange (4) by dividing matrix \mathbf{B} into $[\mathbf{B}_1, \mathbf{B}_2]$ and matrix \mathbf{A} into $[\mathbf{A}_1, \hat{\mathbf{A}}, \mathbf{A}_2]$, the dimensions being given by u_1, u_2 and y_1, \hat{y}, y_2 respectively. We can then write

$$[-\mathbf{B}_1, \mathbf{A}_1, \hat{\mathbf{A}}, \mathbf{A}_2] \begin{bmatrix} u_1 \\ y_1 \\ \hat{y} \\ y_2 \end{bmatrix} = \mathbf{C} + \mathbf{B}_2 u_2. \quad (6)$$

Adding (1) and (6) we obtain

$$\begin{bmatrix} \mathbf{1}, & \mathbf{a}, & \mathbf{0}, & \mathbf{0} \\ -\mathbf{B}_1, & \mathbf{A}_1, & \hat{\mathbf{A}}, & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} u_1 \\ y_1 \\ \hat{y} \\ y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{C} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{B}_2 \end{bmatrix} u_2. \quad (7)$$

It would be now necessary to solve this system of equations together with boundary conditions of type (2), (3). Such a solution will be cumbersome due to the high dimension of the matrix on the left-hand side of Eq. (7). (In real cases the number of variables may be several hundreds — because of space discretisation.)

The situation will be easier when interpreting Eq. (7) like a system of linear equations with parameter u_2 .

Note

The left-hand side matrix in (7) has to have full rank. It is not full either in cases of wrong elaboration of this matrix, or due to discontinuity of the solution (e.g. in the case of critical flow in river channel or pipeline).

Compute the last n rows of the matrix

$$\begin{bmatrix} \mathbf{1}, & \mathbf{a}, & \mathbf{0}, & \mathbf{0} \\ -\mathbf{B}_1, & \mathbf{A}_1, & \hat{\mathbf{A}}, & \mathbf{A}_2 \end{bmatrix}^{-1},$$

where n is the dimension of the vector y_2 .

Multiplying (7) by those rows we obtain the linear function

$$y_2 = \mathbf{c}_2 + \mathbf{b}_2 u_2. \quad (8)$$

In this way the boundary condition in the form (1) was recomputed into points with boundary condition of the form (2), (3). Solution of boundary conditions (2), (3) together with condition (8) gives the values u_2, y_2 . Putting then u_2 into (7) we can compute u_1, y_1, \hat{y} . Thus the integration step is ended.

Point out how efficient the simulation is with complicated boundary conditions if our approach is used.

The efficiency is achieved by transformation of the whole dynamic simulation model (4) with joined boundary conditions of type (1) to a simple and much smaller system of linear equation (8) with variables u_2, y_2 only. In this way the dynamics of the simulated distributed parameter system is in each integration step recomputed to the small system of linear equations (8).

Examples of application

1. The water flow in a river channel represents a classical example of distributed parameter system. The unsteady one-dimension shallow-water flow in river channel is described by Saint-Venant equations

$$\frac{\partial Q}{\partial x} + \frac{\partial S}{\partial t} = 0$$

$$\frac{\partial Z}{\partial x} + \frac{1}{g} \frac{\partial U}{\partial t} + \frac{U}{g} \frac{\partial U}{\partial x} = - \frac{U|U|}{C^2 R}$$
(9)

where S cross section [m^2]

Q flow rate [m^3s^{-1}]

x the distance of the cross section from a given point

t time [s]

U mean value of the speed of flow [ms^{-1}]

Z water level elevation [m]

C speed coefficient

R hydraulic diameter [m]

g acceleration of gravity [ms^{-2}].

To obtain the system of equations for only two variables (Q, Z), it is necessary to join complementary equations describing other variables as a function of Q, Z and geometrical parameters of real channel.

The practical task to simulate and control the system of water channels with 27 built-in gates on river Labe was solved.

That part of the whole complex of solved problems which illustrate the use of our method will be given now.

The basic problem is to simulate the section of river between two gates. This system is schematically shown in Fig. 2. The boundary condition u_1 , which means flow rate at the beginning of the simulated section, depends on the situation of the foregoing (upper) section. For the purpose of simulation of this section there is no other possibility than formulate the input variable u_1 as a function of time, to be able to

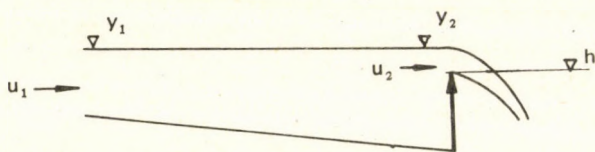


Fig. 2

prepare different simulation experiments under different conditions. Another situation is in downstream flow rate u_2 . In this case the boundary condition is given as

$$u_2 = k(y_2 - h)^{\frac{3}{2}}, \quad \text{for } y_2 \geq h$$

$$u_2 = 0, \quad \text{for } y_2 < h$$
(11)

where y_2 is level elevation,
 h is gates elevation.

This is a real example of a complicated boundary condition with restriction.

The first problem is to simulate dynamic behaviour of the described section, the gates level being fixed.

For the upstream boundary condition, according to the above described method, we write

$$u_1 + a_1 y_1 = c_1$$
(12)

where $a_1 = 0$ and $c_1 = f(t)$.

By time and space discretisation and using implicit integration method, Eqs. (9), (10) were transformed to form (4). Following the above described procedure we obtain finally

$$y_2 = c_2 + b_2 u_2$$
(13)

Hence, we see that the whole problem was reduced to solution of Eqs. (11) and (13). The simple form of Eq. (13) makes it easy to check the conditions given in (11). In

the case when $u_2 > 0$, Eq. (13) and the first of (11) will be solved using, e.g. Newton method. Moreover the described procedure ensures the convergence of the solution.

2. The second example concerns the same system as in Example 1. Now the gate is controlled by a real controller working in pulse regime and maintaining the water elevation y_2 on a constant value. As soon as the gate position is higher than the water level ($h > y_2$), that is when $u_2 = 0$, the controlled value of the water level cannot be maintained by any controller action.

The simulation of the mentioned controller action including different technical restrictions will be cumbersome. Moreover, it leads to extremely short integration step.

Seeking some way how to define boundary conditions one could try to use the boundary condition $y_2 = \text{const}$. However, there exists such a real boundary condition u_1 that the gate stops the water flow. The maintaining of $y_2 = \text{const}$. leads to a solution giving negative water flow over the gate. It means fictive water flow from lower to higher position! Such result, of course, is physically impossible.

Therefore it is necessary to formulate the boundary condition as a relation between u_2 and y_2 respecting the restriction. Such formulation is shown in Fig. 3.

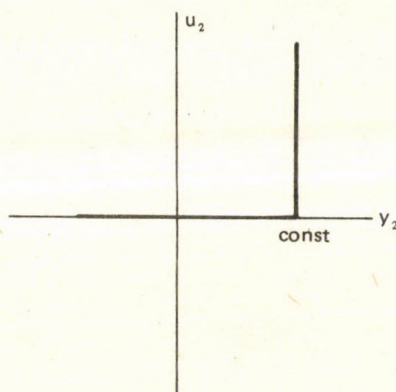


Fig. 3

The solution proceeds in the same manner as in the previous example. The final solution, using the boundary condition in Fig. 3, is then simple because it is given as intersection of straight lines.

3. The described method was used in the solution of dynamic description of large intricate pipe line networks for gas distribution [1]. The use of the method allowed us to work out special decomposition technics for connected subsystems with distributed parameters where the graph of connection is tree.

Reference

1. Králik, J., Stiegler, P., Vostřý, Z., Závorka, J., Modelling the dynamics of large scale systems with application to gas distribution networks. Academia (to be appeared).

Вклад в моделирование систем с распространенными параметрами

З. ВОСТРЫ

(Прага)

В статье описан подход к решению некоторых сложных краевых задач с неравенствами для моделирования динамических систем с распространенными параметрами. Предлагается метод интегрирования и линеаризации на каждом шаге.

Z. Vostřý

Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Praha 8, Pod vodárenskou věží 4
Czechoslovakia

PRINTED IN HUNGARY

Akadémiai Nyomda, Budapest

EXTREMAL PROPERTIES OF ELLIPSOIDS, APPROXIMATING ATTAINABILITY SETS

A. I. OVSEEVICH

(Moscow)

The evolution of attainability sets and its approximating ellipsoids is investigated. Some extremal properties of ellipsoids under consideration are found.

Introduction

Consider the linear control system

$$\dot{x} = A(t)x + u, \quad u \in U(t), \quad x(s) \in D(s) \quad (1.1)$$

Here $x \in \mathbf{R}^n$ is a phase vector, $A(t)$ is a given $n \times n$ -matrix, $U(t)$ is a closed domain of admissible control values, $D(s)$ is a closed set, where motion begins at the moment s (initial domain).

The set of ends $x(t)$ of trajectories beginning in $D(s)$ and satisfying (1.1) for some measurable vector-function $u(t) \in U(t)$ (admissible control), is called the attainability set

$$D(t) = D(s, t) = D(s, D(s), t)$$

for the system (1.1)

One needs to know attainability sets for the purposes of control, filtration of dynamic systems, differential games (see, e.g. [1, 2, 3]). For example, the minimisation of the terminal functional $F(x(T))$, where T is a fixed moment, is equivalent to minimisation of the function $F(x)$ on the attainability set $D(T)$.

The effective construction of attainability sets encounters, however, with considerable difficulties, due to the infinity of parameters identifying domain in \mathbf{R}^n .

The method of approximating attainability sets by prescribed shape domains was developed in a number of papers by different authors (see, e.g. [4, 5]). In particular, ordinary differential equations describing evolution of ellipsoids approximating attainability sets either inside or outside were obtained in [5, 6].

The present paper begins with an equation of evolution of attainability set (Section 2). On this ground the problem of the best approximation of attainability sets by ellipsoids is posed and solved (Sections 3 and 4). It is shown that the solution is obtained by integration of the differential equations mentioned above.

2. Evolution equation of attainability sets

To any subset $D \subset R^n$ one may correspond the support function

$$H(\xi) = H_D(\xi) = \sup_{x \in D} (x, \xi)$$

where $\xi \in R^n$, (\cdot, \cdot) is the scalar product. For D being convex and closed, H_D determines D uniquely [6]. We note that if the domains $U(t)$ and the initial domain $D(s)$ of system (1.1) are closed and convex, then so are the attainability sets $D(t)$. Therefore, the following proposition describing the evolution of the support function of the attainability set determines in this case the evolution of the attainability set itself.

Proposition. Let the function $A(t)$ and set $U(t)$ in (1.1) be measurable [7] in t , $H(t, \xi) = H_{D(t)}(\xi)$ be the support function of the attainability set $D(t)$, $h(t, \xi) = H_{U(t)}(\xi)$ be the support function of the domain $U(t)$.

Then

$$\frac{\partial H}{\partial t}(t, \xi) = \left(A(t) \frac{\partial H}{\partial \xi}(t, \xi), \xi \right) + h(t, \xi) \quad (2.1)$$

where $\partial/\partial \xi$ denotes the gradient with respect to ξ . (Here $(A(\partial H/\partial \xi), \xi) = (\partial H/\partial \xi, A^* \xi)$ is the directional derivative $\lim_{\varepsilon \downarrow 0} \varepsilon^{-1}(H(\xi + \varepsilon A^* \xi) - H(\xi))$ as $\varepsilon \downarrow 0$.)

The corresponding limit does exist for any convex function [7]).

Remark. An analogous formula is obtained in [8] under slightly different assumptions. We give the proof, however, for the reader's convenience. Note that the well-known Bellman equation describes the dependence of $D(t)$ on the initial domain $D(s)$ and the moment s , provided $D(s)$ consist of a single point x .

Proof. We show first, that it is sufficient to consider the case $A(t) \equiv 0$ in (1.1), (2.1). Indeed, let $P(t)$ (fundamental matrix) be the solution of Cauchy problem:

$$\dot{P} = A(t)P, \quad P(s) = I \quad (2.2)$$

where I is the unit matrix. Then the change of variables transforms the system (1.1) into

$$\dot{y} = v, \quad v \in V(t) = P(t)^{-1}U(t), \quad y(s) \in D(s) \quad (2.3)$$

The attainability set $D(t)$ for (1.1) changes then to the attainability set $D^0(t) = P(t)^{-1}D(t)$ for (2.3) and the corresponding support function $H(t, \xi)$ to $H^0(t, \xi) = H(t, P(t)^* \xi)$.

The straightforward calculation shows, that the substitution

$$H(t, \xi) \rightarrow H^0(t, \xi) = H(t, P(t)^* \xi)$$

transforms the solution of (2.1) into the solution of

$$\frac{\partial H^0}{\partial t}(t, \xi) = h^0(t, \xi) = H_{V(t)}(\xi). \quad (2.4)$$

In view of invertibility of P it remains to prove formula (2.4) for the support function $H^0(t, \xi)$ of the attainability set $D^0(t)$ of system (2.3).

In fact

$$H^0(t, \xi) = \sup \left(\int_s^t v(\tau) d\tau, \xi \right)$$

where sup is taken over all measurable vector-functions $v(t) \in V(t)$. It follows from the measurable choice theorem [6] that

$$\sup \left(\int_s^t v(\tau) d\tau, \xi \right) = \int_s^t \sup_{v \in V(\tau)} (v, \xi) d\tau = \int_s^t h^0(\tau, \xi) d\tau$$

q.e.d.

We consider in what follows the case of convex closed domains $U(t)$ and initial domain $D(s)$.

Definition 1. Let $\Omega(t)$, $t \geq s$ be a family of closed convex sets, $H(t, \xi) = H_{\Omega(t)}(\xi)$ be the corresponding family of support functions.

We say that $\Omega(t)$ is a family of subattainability sets (resp. superattainability sets) for (1.1) if the differential inequality

$$\frac{\partial H}{\partial t}(t, \xi) \leq \left(A(t) \frac{\partial H}{\partial \xi}(t, \xi), \xi \right) + h(t, \xi), \quad H(s, \xi) = H_{D(s)}(\xi) \quad (2.5)$$

holds (resp. with the sign \geq).

The inclusion $\Omega_1 \subset \Omega_2$ of closed convex sets is equivalent to the inequality $H_{\Omega_1}(\xi) \leq H_{\Omega_2}(\xi)$ for their support functions. Therefore, it follows from proposition 1, that $\Omega(t) \subset D(t)$ in the subattainability case and $\Omega(t) \supset D(t)$ in the superattainability case. Moreover, if $\tau \leq t$ and $D(\tau, t) = D(\tau, \Omega(\tau), t)$ is the attainability set at the moment t for the system

$$\dot{x} = A(t)x + u, \quad u \in U(t), \quad x(\tau) \in \Omega(\tau) \quad (2.6)$$

then $D(\tau, t) \supset \Omega(t)$ in the subattainability case and $D(\tau, t) \subset \Omega(t)$ in the superattainability case.

Definition 2. Let $V(\Omega)$ be the volume of the set Ω , E be a class of convex sets. We say that the family $\Omega(t) \in E$ of sub (super) attainability sets is the best local approximation of the attainability sets $D(t)$ of system (1.1) if for $\tau \geq s$

$$\frac{d}{dt} V(\Omega(t))|_{t=\tau}$$

is maximal (minimal) over all the sub- (super-) attainability families belonging to E for the system (2.6).

3. Evolution equations for approximating ellipsoids

Denote the ellipsoid

$$\{x \in \mathbf{R}^n, (Q^{-1}(x-a), x-a) \leq 1\}$$

by $E(a, Q)$. Here $a \in \mathbf{R}^n$, Q is a positive definite symmetric matrix. Suppose that the initial moment $s=0$, the initial domain $D(s)$ of system (1.1) is the ellipsoid $E(a_0, Q_0)$ and the control domains $U(t)$ are the ellipsoids $E(b(t), G(t))$. Then the evolution of ellipsoids approximating the attainability set of the system

$$\dot{x} = A(t)x + u, u \in E(b(t), G(t)), x(0) \in E(a_0, Q_0) \quad (3.1)$$

in the locally best way may be described by equations (3.3), (3.4) in Theorem 1.

At first we present two lemmas.

Lemma 1 [3]. The support function $H(\xi)$ of the ellipsoid $E(a, Q)$ is given by the formula

$$H(\xi) = (a, \xi) + (Q\xi, \xi)^{1/2}$$

Lemma 2 [9]. Let $A(t)$ be a family of invertible matrices smoothly depending on t . Then

$$\frac{d}{dt} \log \det A(t) = \text{Tr } A(t)^{-1} \dot{A}(t)$$

Corollary. Let $E(a(t), Q(t))$ be the family of ellipsoids, $V(t)$ be the volume of $E(a(t), Q(t))$. Then

$$\frac{d}{dt} V(t) = \frac{1}{2} V(t) \text{Tr } Q(t)^{-1} \dot{Q}(t) \quad (3.2)$$

Indeed, $V(t) = \omega_n [\det Q(t)]^{1/2}$, ω_n being the volume of unit sphere $E(0, I)$, and formula (3.2) follows from Lemma 2. Introduce the notation $\{\alpha, \beta\} = \alpha\beta + \beta^*\alpha^*$, for matrices α, β .

Theorem 1. 1) Let $a_-(t), Q_-(t)$ be the solution of the Cauchy problem

$$\begin{aligned} \dot{a}_- &= A(t)a_- + b(t), \quad a_-(0) = a_0 \\ \dot{Q}_- &= \{A(t), Q_-\} + 2R^{-1}(RQ_-\dot{R}^*)^{1/2}(RG(t)R^*)^{1/2}R^{*-1} \\ Q_-(0) &= Q_0 \end{aligned} \quad (3.3)$$

where R is a nonsingular matrix such that $G^0 = RG(t)R^*$ and $Q^0 = RQ_-\dot{R}^*$ are diagonal matrices. Then $E(a_-(t), Q_-(t))$ is the subattainable best local ellipsoidal approximation for system (3.1).

2) Let $a_+(t), Q_+(t)$ be the solution of the Cauchy problem

$$\begin{aligned} \dot{a}_+ &= A(t)a_+ + b(t), \quad a_+(0) = a_0 \\ \dot{Q}_+ &= \{A(t), Q_+\} + hQ_+ + h^{-1}G(t), \quad Q_+(0) = Q_0 \\ h &= [n^{-1} \text{Tr}(Q_+^{-1}G(t))]^{1/2} \end{aligned} \quad (3.4)$$

where Tr denotes the trace. Then $E(a_+(t), Q_+(t))$ is the best local ellipsoidal superattainable approximation for the system (3.1).

Remark. It is well known that (see, e.g. [9]) if A and B are two symmetric matrices, one of which being positive definite, then there exist a nonsingular matrix R such that RAR^* and RBR^* are diagonal. The expression $R^{-1}(RAR^*)^{1/2}(RBR^*)^{1/2}R^{*-1}$ in spite of possible nonuniqueness of R , depends only on A and B and is equal to $B^{1/2}(B^{-1/2}AB^{-1/2})^{1/2}B^{1/2}$, if B is positive definite. We divide the proof into several steps.

Step 1. Simplification of system (3.1). Let the matrix function (fundamental matrix $P(t)$ be the solution of the Cauchy problem

$$\dot{P} = A(t)P, \quad P(0) = I. \quad (3.5)$$

Under the change of variables $x = P(t)y$ control system (3.1) takes the form:

$$\begin{aligned} \dot{y} &= v, \quad v \in V(t) = P(t)^{-1}U(t) = E(b_1(t), G_1(t)) \\ b_1(t) &= P(t)^{-1}b(t), \quad G_1(t) = P(t)^{-1}G(t)P(t)^{*^{-1}} \\ y(0) &\in E(a_0, Q_0). \end{aligned} \quad (3.6)$$

Obviously $P(t)$ maps each ellipsoid onto an ellipsoid and sub- (super-) attainability domains for system (3.6) onto the corresponding domains for (3.1). It is not difficult to verify by direct calculations that if $Q(t), a(t)$ satisfy equations (3.3) or (3.4), then the parameters $P(t)^{-1}a(t), P(t)^{-1}Q(t)P(t)^{*^{-1}}$ of the ellipsoid $P^{-1}E(a, Q)$ satisfy a system, similar to (3.3) or (3.4), where $A(t) \equiv 0$, and $G = G_1$ is given by (3.6). Consequently it is sufficient to prove the theorem, assuming $A(t) \equiv 0$ in (3.1).

Setting $x = y + r(t)$, where $\dot{r} = b(t), r(0) = a_0$, we can restrict ourselves to the system of the following form

$$\dot{x} = u, \quad u \in E(0, G(t)), \quad x(0) \in E(0, Q_0). \quad (3.7)$$

Step 2. Reformulation of the problem. Let $E(a(t), Q(t))$ be a subattainability family of ellipsoids for system (3.7). According to lemma 1 and definition 1, this means that

$$\frac{d}{dt} [a(t), \xi) + (Q(t)\xi, \xi)^{1/2}] \leq (G(t)\xi, \xi)^{1/2}. \quad (3.8)$$

Otherwise

$$(\dot{a}, \xi) + \frac{1}{2} (Q\xi, \xi)^{-1/2} (\dot{Q}\xi, \xi) \leq (G\xi, \xi)^{1/2}. \quad (3.9)$$

Formula (3.2) allows us to rewrite the condition of the best local approximation of the family $E(a(t), Q(t))$ (definition 2) in the form

$$\text{Tr } Q^{-1} \dot{Q} \rightarrow \max, \quad (3.10)$$

where maximum is taken over all symmetric matrices \dot{Q} and vectors \dot{a} , satisfying (3.9), provided $Q = Q(t)$ is fixed. Max in (3.10) has to be replaced by min and the sign \geq in (3.9) by \leq in the super-attainability case.

Replacing ξ in (3.9) by $-\xi$ and adding the resulting inequality to (3.9), one gets

$$\frac{1}{2} (Q\xi, \xi)^{-1/2} (\dot{Q}\xi, \xi) \leq (G\xi, \xi)^{1/2}. \quad (3.11)$$

Therefore, if the pair (\dot{a}, \dot{Q}) provides maximum in (3.10), then the pair $(0, \dot{Q})$ also provides this maximum. In what follows we suppose $a \equiv 0$.

Let R be a nonsingular matrix such that $Q_1 = RQR^*$, $G_1 = RGR^*$ are diagonal matrices. Then $\text{Tr } Q^{-1} \dot{Q} = \text{Tr } Q_1^{-1} C$, where $C = R\dot{Q}R^*$ and the extremal problem (3.9) is reduced to the following one. Find the symmetric matrix C , providing maximum

$$\text{Tr } Q_1^{-1} C \rightarrow \max \quad (3.12)$$

$$(C\eta, \eta) \leq 2(G_1\eta, \eta)^{1/2} (Q_1\eta, \eta)^{1/2}, \forall \eta \in \mathbf{R}^n$$

where G_1, Q_1 are given diagonal matrices. The superattainability case leads to the problem

$$\text{Tr } Q_1^{-1} C \rightarrow \min \quad (3.13)$$

$$(C\eta, \eta) \geq 2(G_1\eta, \eta)^{1/2} (Q_1\eta, \eta)^{1/2}, \forall \eta \in \mathbf{R}^n.$$

To prove the theorem one needs to establish that there is a solution of problem (3.12) of the form

$$C = 2Q_1^{1/2} G_1^{1/2} \quad (3.14)$$

and there is a solution of problem (3.13) of the form

$$C = hQ_1 + h^{-1}G_1. \quad (3.15)$$

Indeed, in view of $C = R\dot{Q}R^*$ and similar formulae for Q_1, G_1 it follows from (3.14) and

(3.15) that equations (3.3), (3.4) for \dot{Q} with $A(t) \equiv 0$ hold. Since $a \equiv 0$ equations (3.3), (3.4) hold for a_+ , where $A \equiv 0$, $b \equiv 0$, $a_0 \equiv 0$.

Step 3. The solution of the transformed problem. We show that solution C of extremal problem (3.12) or (3.13) may be found among diagonal matrices. Indeed, let C be a solution and Γ be the group of diagonal matrices, with diagonal elements ± 1 . If $g \in \Gamma$, then $g^* C g^*$ is the solution together with C , since

$$g^* = g^{-1} = g, \quad g^{-1} G_1 g = G, \quad g^{-1} Q_1 g = Q_1$$

$$\text{Tr } Q_1^{-1} C = \text{Tr} (g^{-1} Q_1^{-1}) (g^{-1} C g) = \text{Tr} (Q_1^{-1} g^{-1} C g)$$

Inequalities (3.12) and (3.13) are linear in C , that is why the combination

$$C_1 = 2^{-n} \sum_{g \in \Gamma} g^* C g \quad (3.16)$$

of solutions $g^* C g$ is also the solution.

But C_1 is diagonal matrix with the same diagonal elements as C . Indeed, a matrix element

$$(C_1)_{ij} = 2^{-n} \sum_{g \in \Gamma} g_{ii} g_{jj} C_{ij} = C_{ij} (2^{-n} \sum_{g \in \Gamma} g_{ii} g_{jj}) = C_{ij} \varphi_{ij}$$

If $i=j$, then $\varphi_{ij} = 2^{-n} \sum_{g \in \Gamma} 1 = 1$. If $i \neq j$, then we associate to the matrix $g = \text{diag} (g_{kk}) \in \Gamma$ the matrix $g^0 = \text{diag} (g_{kk}^0)$, given by $g_{kk}^0 = g_{kk}$, for $k \neq j$, $g_{jj}^0 = -g_{jj}$. Dividing $\sum_{g \in \Gamma} g_{ii} g_{jj}$ into sums over the corresponding pairs, one gets $\varphi_{ij} = 0$.

We may therefore restrict ourselves to the search of the diagonal matrix $C = (c_i \delta_{ij})$. Denote by p, q, r the vectors with the components $p_i = (G_1)_{ii}$, $q_i = (Q_1)_{ii}$, $r_i = (Q_1^{-1})_{ii} = q_i^{-1}$. Then the problems (3.12), (3.13) are reduced to the following:

$$(r, c_-) \rightarrow \max \quad (3.17)$$

$$(c_-, x) \leq 2(p, x)^{1/2} (q, x)^{1/2}$$

where $x \in \mathbf{R}^n$ is an arbitrary vector with nonnegative components (squares of the components of η in (3.12))

$$(r, c_+) \rightarrow \min \quad (3.18)$$

$$(c_+, x) \geq 2(p, x)^{1/2} (q, x)^{1/2}, \quad \forall x \in \mathbf{R}^n, x_i \geq 0$$

Substituting $x = e_i$ the i -th unit vector into (3.17) we have $(c_-)_i \leq 2p_i^{1/2} q_i^{1/2}$. Since the vector r has positive components, vector c_- satisfying restriction (3.17) with the

components $(c_-)_i = 2p_i^{1/2}q_i^{1/2}$ is the solution of problem (3.17). Therefore, one has to check the inequality

$$\sum p_i^{1/2}q_i^{1/2}x_i \leq (\sum p_i x_i)^{1/2} (\sum q_i x_i)^{1/2}.$$

But it is the Cauchy–Buniakovsky inequality for vectors with the components $(p_i x_i)^{1/2}, (q_i x_i)^{1/2}$.

Thus, problem (3.17) is solved and its solution leads precisely to formula (3.14).

Now we pass to the solution of problem (3.18). Setting $x = r$ in (3.18) one gets

$$(r, c_+) \geq 2(p, r)^{1/2}(q, r)^{1/2} = 2n^{1/2}(p, r)^{1/2}. \quad (3.19)$$

On the other hand, putting $h = n^{-1/2}(p, r)^{1/2}$, $c_+ = hq + h^{-1}p$, one gets

$$(r, c_+) = 2n^{1/2}(p, r)^{1/2}. \quad (3.20)$$

We check the restriction (3.18) for c_+ :

$$\begin{aligned} (c_+, x) &= h(q, x) + h^{-1}(p, x) \geq 2[h(q, x)]^{1/2}[h^{-1}(p, x)]^{1/2} = \\ &= 2(q, x)^{1/2}(p, x)^{1/2}. \end{aligned}$$

By the resulting inequality and (3.19), (3.20) it follows that c_+ is the solution of problem (3.18). It is not difficult to verify that it leads to formula (3.18).

The theorem is proved.

Remark. Equations (3.3), (3.4) were obtained in [3, 4, 5] with the help of finite-difference approximation of the control system (3.1). It was not revealed, however, in what sense the solutions of these equations lead to the best approximations of attainability domains. Theorem 1 answers this question, by establishing extremal properties of ellipsoids, described by equations (3.3), (3.4).

4. Uniqueness of the best local ellipsoidal approximation

It was shown in theorem 1 that equations (3.3), (3.4) give the locally optimal ellipsoidal families. Next, we show that there is no other locally optimal ellipsoidal family.

Theorem 2. Let $E(a(t), Q(t))$ be a locally optimal ellipsoidal sub- (super-) attainability family for the control system (3.1). Then $(a(t), Q(t))$ is the integral curve of system (3.3) (resp. (3.4)).

Proof. The arguments from steps 1, 2 used in the proof of the preceding theorem show that in the subattainability case it is sufficient to establish the uniqueness of the

solution of the extremal problem

$$\begin{aligned} & \text{Tr } Q^{-1}C \rightarrow \max \\ & (d, \xi) + \frac{1}{2}(Q\xi, \xi)^{-1/2}(C\xi, \xi) \leq (G\xi, \xi)^{1/2}, \forall \xi \in \mathbf{R}^n \end{aligned} \quad (4.1)$$

where Q, G are given positive diagonal matrices. In the superattainability case one needs to establish the uniqueness in the similar problem

$$\begin{aligned} & \text{Tr } Q^{-1}C \rightarrow \min \\ & (d, \xi) + \frac{1}{2}(Q\xi, \xi)^{-1/2}(C\xi, \xi) \geq (G\xi, \xi)^{1/2}. \end{aligned} \quad (4.2)$$

As it was mentioned in the proof of theorem 1 (cf. (3.11)), the solutions of problems (4.1), (4.2) are also the solution of problems

$$\begin{aligned} & \text{Tr } Q^{-1}C \rightarrow \max \\ & (C\xi, \xi) \leq 2(Q\xi, \xi)^{1/2}(G\xi, \xi)^{1/2} = \Phi(\xi) \end{aligned} \quad (4.3)$$

$$\begin{aligned} & \text{Tr } Q^{-1}C \rightarrow \min \\ & (C\xi, \xi) \geq 2(Q\xi, \xi)^{1/2}(G\xi, \xi)^{1/2} = \Phi(\xi). \end{aligned} \quad (4.4)$$

We prove first the uniqueness of solutions of problems (4.3), (4.4).

Let C be a solution of (4.3). Then by (4.3), with $\xi = e_i$, where e_i is the i -th unit vector, we have $C_{ii} = (Ce_i, e_i) \leq \Phi(e_i)$. Since the maximum of $\sum Q_{ii}^{-1}C_{ii}$, $Q_{ii} > 0$ is attained at C and since $\sum Q_{ii}^{-1}(C_{-})_{ii} = \sum Q_{ii}^{-1}\Phi(e_i)$ for the solution $C_{-} = 2Q^{1/2}G^{1/2}$ (pointed in (3.14)) of problem (4.3) it is necessary $(Ce_i, e_i) = \Phi(e_i)$. It follows then from (4.3) that

$$(Cx, x) - (Ce_i, e_i) \leq \Phi(x) - \Phi(e_i) \quad \forall x \in \mathbf{R}^n. \quad (4.5)$$

Put $x = e_i + \varepsilon y$, where $y \in \mathbf{R}^n$, ε is a real number. We have by (4.5) that

$$2\varepsilon(e_i, y) \leq \varepsilon(\text{grad } \Phi(e_i), y) + o(\varepsilon) \quad (4.6)$$

as ε tends to zero.

Since ε is of any sign and y is an arbitrary vector one gets from (4.6) that $Ce_i = 1/2 \text{ grad } \Phi(e_i)$ (here we use "Minty trick" known in the monotone operators's theory).

Thus the matrix C is determined uniquely, $C = C_-$. Next let C be a solution of problem (4.4), $C_+ = hQ + h^{-1}G$ be the solution mentioned in (3.15), X be the set of vectors $x \in \mathbf{R}^n$ such that $x_i^2 = Q_{ii}^{-1}$, i.e. $x_i = \pm Q_{ii}^{-1/2}$. Then by (4.4) and (3.20) we have for $x \in X$:

$$(Cx, x) \geq \Phi(x) = (C_+x, x) = \text{Tr } Q^{-1}C_+ \quad (4.7)$$

Put

$$C^0 = 2^{-n} \sum_{g \in \Gamma} g^* C g \quad (4.8)$$

in the notations of (3.16)

Then it follows from (4.7) that

$$(C^0x, x) \geq \text{Tr } Q^{-1}C_+ \quad (4.9)$$

By (4.8), arguing similarly as in step 3 in the proof of theorem 1, we obtain

$$\text{Tr } Q^{-1}C_1^0 = \text{Tr } Q^{-1}C. \quad (4.10)$$

Moreover, since C^0 is a diagonal matrix,

$$(C^0x, x) = \text{Tr } Q^{-1}C^0, \forall x \in X. \quad (4.11)$$

One gets from (4.9), (4.10) and (4.11) that

$$\text{Tr } Q^{-1}C \geq \text{Tr } Q^{-1}C_+ \quad (4.12)$$

and moreover, the strict inequality takes place if $(Cx, x) \neq \Phi(x)$ for some $x \in X$. But $\text{Tr } Q^{-1}C_+ = \text{Tr } Q^{-1}C$, since C, C_+ are solutions of problem (4.4). Consequently $(Cx, x) = \Phi(x) \forall x \in X$. The same arguments as in the proof of the uniqueness of solution of problem (4.3), show that $Cx = (1/2) \text{grad } \Phi(x), \forall x \in X$. Since the basis of \mathbf{R}^n may be picked out of X , C is determined uniquely, $C = C_+$.

It remains to show that the vectors d in (4.1), (4.2) are equal to zero. By the above arguments and constraint (4.1), we obtain that $(d, \pm e_i) \leq 0$, hence $d = 0$. Similarly in the problem (4.2) we have $(d, \pm x) \leq 0, \forall x \in X$, and consequently $d = 0$.

The theorem is proved.

The author thanks F. L. Chernousko for useful discussions.

Reference

1. Красовский Н. Н. Теория управления движением. М., "Наука", 1968.
2. Красовский Н. Н. Игровые задачи о встрече движений. М., "Наука", 1970.
3. Куржанский А. Б. Управление и наблюдения в условиях неопределенности. М., "Наука", 1977.
4. Schweppe, F. C. Recursive state estimation: unknown but bounded errors and system inputs. IEEE Trans. Automat. Control, 1968, AC-13, № 1.

5. Черноусько Ф. Л. Оптимальные гарантированные оценки неопределенностей с помощью эллипсоидов. I, II, Изв. АН СССР, Техническая кибернетика, I: №3, с. 3; II: № 4, с. 3; III: № 5, с. 5.
6. Черноусько Ф. Л. Эллипсоидальные оценки области достижимости управляемой системы. ПММ, т. 45, вып. II, 1981.
7. Иоффе А. Д., Тихомиров В. М. Теория экстремальных задач. М. "Наука", 1974, 479 с.
8. Панасюк А. И. Расчет множеств достижимости систем автоматического управления с помощью уравнения для опорной функции. Деп. в Бел. НИИНТИ 20XI. 80, № 214, 36 с.
9. Гантмахер Ф. Р. Теория матриц. М. "Наука", 1967, 575 с.
10. Лионс Ж.-Л. Некоторые методы решения нелинейных краевых задач. М. "Мир", 1972, 587 с.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H-1111 Budapest XI., Stoczek u. 2, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper with wide margins (4–5 cm), should carry the title of the contribution, the author(s) name, and the name of the country. At the end of the typescript the name and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice-versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме – реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в трех экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициях. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 отисков их статей.

Рукописи непринятых статей возвращаются авторам.

052.312

CONTENTS · СОДЕРЖАНИЕ

<i>Zinoviev, V. A.:</i> Cascade equal-weight codes and maximal packings (<i>Зиновьев В. А. Каскадные равновесные коды и максимальные упаковки</i>)	3
<i>Yemelyanov, S. V., Korovin, S. K., Ulanov, B. V.:</i> Control of nonstationary dynamic systems with quasicontinuous generation of the control signal (<i>Емельянов С. В., Коровин, С. К., Уланов Б. В. Управление нестационарными динамическими системами при квазинепрерывном формировании управляющего воздействия</i>)	11
<i>Krzyżak, A., Pawlak, M.:</i> Universal consistency results for Wolwerton–Wagner regression function estimate with application in discrimination (<i>Кржижак А., Павлак М. Универсальная состоятельность рекурсивных оценок типа Вольвертона–Вагнера и их применение</i>)	33
<i>Овсеевич А. И.</i> Экстремальные свойства эллипсоидов, аппроксимирующих области достижимости	43
<i>Vostrý, Z.:</i> Contribution to simulation of distributed parameter systems (<i>Востры З. Вклад в моделирование систем с распространенными параметрами</i>)	55

316.920

VOL. 12 • NUMBER 2
TOM HOMEP

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF

CONTROL AND

INFORMATION

THEORY

ПРОБЛЕМЫ

УПРАВЛЕНИЯ И

ТЕОРИИ

ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1983

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H-1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 0BW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1983 Subscription Rate US Dollars 140.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

N. N. KRASOVSKII (USSR)
G. BOGNÁR (HUNGARY)

COORDINATING EDITORS

USSR

S. V. EMEL'YANOV
E. P. POPOV
V. S. PUGACHEV
V. I. SIFOROV
E. D. TERYAEV

HUNGARY

T. VÁMOS
L. VARGA
A. PRÉKOPA
S. CSIBI
I. CSISZÁR
L. KEVICZKY
J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ
V. STREJČ

РЕДАКТОРЫ ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)
Г. БОГНАР (ВНР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

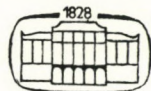
С. В. ЕМЕЛЬЯНОВ
Е. П. ПОПОВ
В. С. ПУГАЧЕВ
В. И. СИФОРОВ
Е. Д. ТЕРЯЕВ

ВНР

Т. ВАМОШ
Л. ВАРГА
А. ПРЕКОПА
Ш. ЧИБИ
И. ЧИСАР
Л. КЕВИЦКИ
Я. КОЧИШ

ЧССР

И. БЕНЕШ
В. СТРЕЙЦ



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

ΕΡΕΥΝΗ
ΤΥΠΟΓΡΑΦΕΙΟΣ ΑΚΑΔΕΜΙΑΣ
ΚΟΝΥΒΤΑΡΑ

APPLICATION OF NEW FEEDBACK TYPES IN THE PROBLEM OF SIGNAL DIFFERENTIATION

S. V. YEMEL'YANOV, A. A. SOLOVIEV

(Moscow)

(Received January 20, 1982)

Difficulties of obtaining signal derivative information are well known in control system design. In case when the object's inner coordinates and parameters are not accessible for measuring, the derivative value can be obtained only by direct signal differentiation. In the paper the possibility of application of new feedback types, i.e. coordinate-parametric (CPF) and parametric (PF), for differentiating device design is considered.

1. Introduction

The general structure of the differentiating element is shown in Fig. 1. The element input is signal $f(t)$. The output signal $z(t)$ is

$$z(t) = K(f(t) - y(t)), \quad (1)$$

where $y(t)$ varies in accordance with

$$\dot{y} = \frac{1}{T_1} z. \quad (2)$$

Using the equation we obtain after differentiating (1) the equation which connects the input and output signals of the element

$$\dot{z} = \frac{K}{T_1} (T_1 \dot{f} - z).$$

So the element transmissive function is

$$W_1(p) = \frac{z(p)}{f(p)} = \frac{T_1 p}{\frac{T_1}{K} p + 1}. \quad (3)$$

If $K \rightarrow \infty$, then $W_1(p) \rightarrow T_1 p$ and for the output signal $z(t)$ the relation

$$z(t) \approx T_1 \dot{f}(t) \quad (4)$$

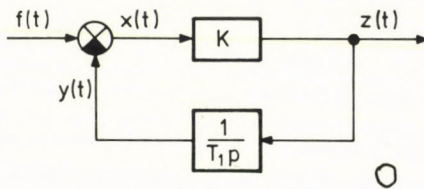


Fig. 1

is valid. But there is always limitation on the module of the output in the practical realization of an amplification section. Because of this the size of the amplifier linear zone decreases when the amplification factor increases ($K \rightarrow \infty$). As a result, the amplification section begins working as a relay and its output signal has the information only about the input signal sign. In this case a sliding mode is possible [1], the signal $z(t)$ will consist of high-frequency oscillations and relation (4) will be valid only for its average value. Real amplifiers have also their own noise.

Due to this fact the use of the element shown in Fig. 1 with the output signal $z(t)$ as differentiating element is limited.

In the paper the problem of differentiating device design is formulated and the attempt to solve it by using coordinate-parametric and parametric feedbacks [2] is made.

2. The formulation of the problem

By differentiating device we understand an element with the transmissive function which can be made as approximating the transmissive function of differentiating section as we wish. The problem of designing such a device consists in the synthesis of the structure scheme of the element with relationship between the input and the output signals, which can be represented by the transmissive function

$$W^*(p) = \frac{v(p)}{f(p)} = \frac{T_D^* p}{\delta p + 1}. \quad (5)$$

The structure scheme has to contain only amplification, integration and relay sections. Choosing finite values of amplification factors of amplification sections and time constants of integration sections there must be a possibility to obtain in (5) any value including quantum libet small of δ ($\delta \ll 1$) and any value including quantum libet large of T_D ($T_D \gg 1$). The element output signal must not be the output signal of the amplification section if it is formed of some signals; then none of these signals may be the output signals of amplification sections.

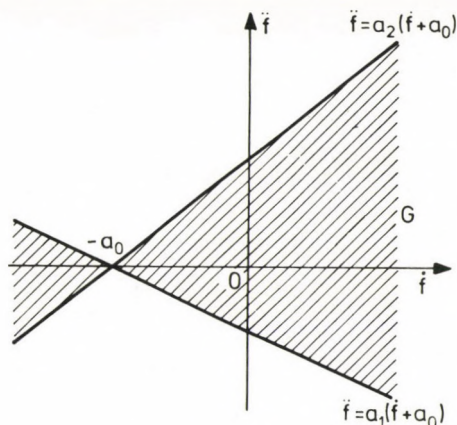


Fig. 2

The problem formulated is solved by the next restrictions.

1. The input signals $f(t)$ to which the differentiating element is applied depend on time in such a way that for all $t \geq t_0$ (t_0 is the moment when the signal begins to act as input of the element) the inequalities are valid

$$-\frac{a_2 - a_1}{2} |f + a_0| < \dot{f} + \frac{a_1 + a_2}{2} (f + a_0) < \frac{a_2 - a_1}{2} |f + a_0|, \quad (6)$$

where $a_0, a_1 < a_2$ are constants; the parameter values of the element sections are defined by values of these constants. The region G of the plane (\dot{f}, \ddot{f}) in which inequalities (6) are valid is shown in Fig. 2.

2. The input and output signals of the element are connected in accordance with the transmissive function $W^*(p)$ (5) not for all $t \geq t_0$ but for $t \geq t^*$ where $t^* \geq t_0$. The moment t^* can both be equal and non-equal to the moment t_0 .

As a result of solving the formulated problem not only the structure scheme will be constructed but also the method of calculation of the element sections' parameters by values of $T_D^*, \delta^*, a_0, a_1$ and a_2 where $\delta^* \geq \delta$ will be given.

3. Stages of the synthesis of the differentiating element structure scheme

The scheme in Fig. 1 with the output signal $z(t)$ does not give the solution of the problem formulated. For this scheme the signal $x(t) = f(t) - y(t)$ (Fig. 1) could suite the differentiation problem output signal. In this case the element's transmissive function is

$$W_2(p) = \frac{x(p)}{f(p)} = \frac{\frac{T_1}{K} p}{\frac{T_1}{K} p + 1}. \quad (7)$$

As a result of the comparison of (7) with (5) we observe that it is impossible to choose the constants T_1 and K so that the transmissive function $W_2(p)$ coincides with $W^*(p)$ when the values of δ and T_D^* are not equal and the value δ is close to zero and the value of T_D^* is large.

3.1. Introduction of coordinate-parametric feedback (CPF)

Let us modify the scheme of Fig. 1 by introducing the CPF loop [2] to control the factor K (Fig. 3). In accordance with the scheme of Fig. 3, the factor K can be seen as

$$K = K_c + K_v(t), \quad (8)$$

where $K_c = \text{const}$, and

$$K_v(t) = \begin{cases} \Delta K, & \text{if } \varepsilon(t)x(t) > 0, \\ -\Delta K, & \text{if } \varepsilon(t)x(t) < 0. \end{cases} \quad (9)$$

Here $\Delta K = \text{const}$. and

$$\varepsilon(t) = x(t) - w(t) \quad (10)$$

($\varepsilon(t)$ is an error in the CPF loop, $w(t)$ is a setpoint signal in the CPF loop).

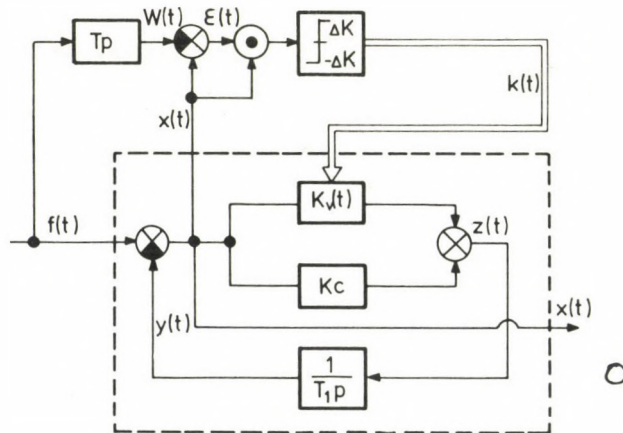


Fig. 3

In the case shown in Fig. 3, there is

$$w(t) = Tf'(t). \quad (11)$$

By differentiating the relation $x(t) = f(t) - y(t)$ we obtain the following equation, taking into account (1), (2) and (8),

$$\dot{x} = f' - \left(\frac{K_c}{T_1} + \frac{K_v(t)}{T_1} \right) x. \quad (12)$$

From (9) it follows that (12) is a non-linear differential equation with a breaking right-hand part. The right-hand part of (12) has a break on the plane $\varepsilon = 0$. Thus there may arise a specific type of movement, the sliding mode [3]. In this case

$$\varepsilon(t) = 0,$$

from (10) and (11) we get

$$x(t) = w(t) = Tf'(t).$$

The last correlation which approximates the movement in the sliding mode is a linear differential equation which connects the input ($f(t)$) and output ($x(t)$) signals of the element. Therefore when the sliding mode arises the connection between $x(t)$ and $f(t)$ can be represented by the transmissive function

$$W_3(p) = \frac{x(p)}{f(p)} = Tp.$$

But the structure scheme in Fig. 3 is not the solution of the formulated problem as it has a differentiating section with the transmissive function Tp .

Let us replace the differentiating section by the section with the transmissive function in the scheme of Fig. 3:

$$W_4(p) = \frac{Tp + B}{T_2 p + C}. \quad (13)$$

From (13) it follows that now the change of the signals $w(t)$ will depend on the equation

$$\dot{w} = \frac{1}{T_2} (Tf' + Bf - Cw). \quad (14)$$

In general, the movement will be formed by a non-linear system of two differential equations, (12) and (14), and correlations (9) and (10).

As it was mentioned above, due to the break of the right-hand part of (12) a sliding mode may occur on the plane $\varepsilon = 0$. Then $\varepsilon = 0$ and $\dot{\varepsilon} = 0$ [3]. It follows from (10)

and (14) that the connection between the signals $x(t)$ and $f(t)$ can be approximated by the equations

$$x(t) = w(t)$$

and

$$\dot{x} = \dot{w} = \frac{1}{T_2} (T\dot{f} + Bf - Cx). \quad (15)$$

Since (15) is a linear differential equation, in this case the connection between the signals $x(t)$ and $f(t)$ can be represented by a transmissive function $\frac{x(p)}{f(p)}$ which is the same as that of $W_4(p)$ (13). So the whole element will function as a section with the transmissive function $W_4(p)$. It is connected with the signal $w(t)$ depending on time which is defined only by parameters of the section with the transmissive function $W_4(p)$ (13) if the dependence on time of $f(t)$ is set. In order to change the time signal $w(t)$ depending on other element parameter values, it is necessary to introduce a parametric feedback (PF) [2].

3.2. Introduction of the parametric feedback (PF)

Let us introduce the parametric feedback as it is shown in Fig. 4.

Now the parameter value C is not constant and it depends on time in accordance with the change of $K_v(t)$, i.e.

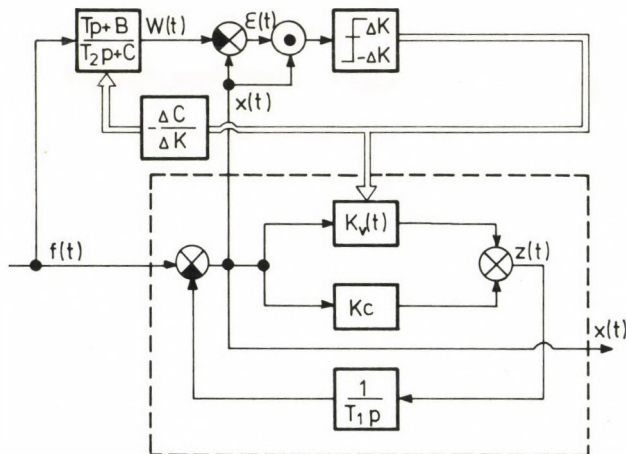


Fig. 4

$$C = C_0 + \frac{\Delta C}{\Delta K} K_v(t), \quad (16)$$

where $\Delta C = \text{const}$, $C_0 = \text{const}$.

The movement in this case may be described by a system of nonlinear differential equations

$$\begin{aligned} \dot{x} &= f' - \left(\frac{K_c}{T_1} + \frac{K_v(t)}{T_1} \right) x, \\ \dot{\varepsilon} &= -\frac{1}{T_2} \left(C_0 - \frac{\Delta C}{\Delta K} K_v(t) \right) \varepsilon + \left(1 - \frac{T}{T_2} \right) f' - \\ &\quad - \frac{B}{T_2} f - \left[\frac{K_c}{T_1} - \frac{C_0}{T_2} + \left(\frac{1}{T_1} + \frac{\Delta C}{T_2 \Delta K} \right) K_v(t) \right] x, \end{aligned} \quad (17)$$

where $K_v(t)$ is determined from (9).

The right-hand parts of system (17) have a break on the plane $\varepsilon = 0$. As a reason of this, a sliding mode may arise on the plane. Let us determine $K_v(t)$ from the equations $\varepsilon = 0$ and $\dot{\varepsilon} = 0$ and substitute it into the first equation of (17); then we see that if such type of movement arises, the connection between $x(t)$ and $f(t)$ signals is approximated by the equation

$$\dot{x} = \frac{1}{T_2 + \frac{T_1 \Delta C}{\Delta K}} \left[\left(T + \frac{T_1 \Delta C}{\Delta K} \right) f' + Bf - \left(C_0 + K_c \frac{\Delta C}{\Delta K} \right) x \right]. \quad (18)$$

Since (18) is a linear differential equation the connection between $x(t)$ and $f(t)$ signals may be presented by a transmissive function

$$W_5(p) = \frac{x(p)}{f(p)} = \frac{\left(T + \frac{T_1 \Delta C}{\Delta K} \right) p + B}{\left(T_2 + \frac{T_1 \Delta C}{\Delta K} \right) p + C_0 + K_c \frac{\Delta C}{\Delta K}}. \quad (19)$$

So if the sliding mode arises as a result of the introduction of the parametric feedback, it is possible to obtain the transmissive function $W_5(p)$ approximating connection between $x(t)$ and $f(t)$ which differs from the transmissive function $W_4(p)$ (13). As the connection between $x(t)$ and $f(t)$ signals correspond to the transmissive function $W_5(p)$ (19) only when $\varepsilon = 0$, it is necessary to ensure the existence of the time moment when $\varepsilon = 0$.

3.3. Ensuring of hit on the plane $\varepsilon=0$

The hit of the imaging point of system (17) on the plane $\varepsilon=0$ can be obtained by using the sum of $f(t)$ and $u(t)=\varepsilon(t)-f(t)$ as an input signal of the section with the transmissive function $W_4(p)$ instead of $f(t)$. The same is valid for the substitution of the input signal of the section, $f(t)$, by $\varepsilon(t)$. In this case the second equation of system (17) is as follows

$$\begin{aligned} \dot{\varepsilon} = & \frac{1}{1 + \frac{T}{T_2}} \left[- \left[\frac{B}{T_2} + \frac{1}{T_2} \left(C_0 - \frac{\Delta C}{\Delta K} K_v(t) \right) \right] \varepsilon + \right. \\ & \left. + \dot{f} - \left[\frac{K_c}{T_1} - \frac{C_0}{T_2} + \left(\frac{1}{T_1} + \frac{\Delta C}{T_2 \Delta K} \right) K_v(t) \right] x \right]. \end{aligned} \quad (20)$$

When the sliding mode arises the connection between the output $x(t)$ and input $f(t)$ is approximated by the linear differential equation

$$\dot{x} = \frac{1}{T_2 + \frac{T_1 \Delta C}{\Delta K}} \left[\frac{T_1 \Delta C}{\Delta K} \dot{f} - \left(C_0 + K_c \frac{\Delta C}{\Delta K} \right) x \right]$$

and can be represented by the transmissive function

$$W_6(p) = \frac{x(p)}{f(p)} = \frac{T_1 \frac{\Delta C}{\Delta K} p}{\left(T_2 + \frac{T_1 \Delta C}{\Delta K} \right) p + C_0 + K_c \frac{\Delta C}{\Delta K}}. \quad (21)$$

It follows from (20) that by a suitable choice of the factor B the hit of system (17) of the imaging point on the plane $\varepsilon=0$ can be ensured without influencing the transmissive function (21) which does not depend on B . Note that $W_6(p)$ does not depend on T , therefore if $T=0$ in $W_4(p)$ (13) then the connection between $x(t)$ and $f(t)$ will be the same as in the case when T has any non-zero value when the sliding mode arises. Comparing (21) with (5) we observe a coincidence between $W_6(p)$ (21) and $W^*(p)$ (5). It cannot be obtained for any positive values of T_B^* and δ if we limit ourselves to only positive values of T_1 , T_2 , ΔC , ΔK , C_0 and K_c . Therefore for a successful solution of the differentiation problem it is necessary to change the CPF's sign and make it positive.

3.4. Replacement of negative CPF by positive one

In Fig. 5 the structural scheme of the element is shown which enables us to solve the formulated signal differentiation problem.

In this case the movement can be described by a system of nonlinear differential equations

$$\begin{aligned} \dot{x} &= f + \frac{V_0}{T_1} - \left(\frac{K_c}{T_1} + \frac{K_v(t)}{T_1} \right) x, \\ \dot{\varepsilon} &= - \left[\frac{B}{T_2} + \frac{1}{T_2} \left(C_0 + \frac{\Delta C}{\Delta K} K_v(t) \right) \right] \varepsilon + f + \frac{V_0}{T_1} - \\ &\quad - \left[\frac{K_c}{T_1} - \frac{C_0}{T_2} + \left(\frac{1}{T_1} - \frac{\Delta C}{T_2 \Delta K} \right) K_v(t) \right] x, \end{aligned} \tag{22}$$

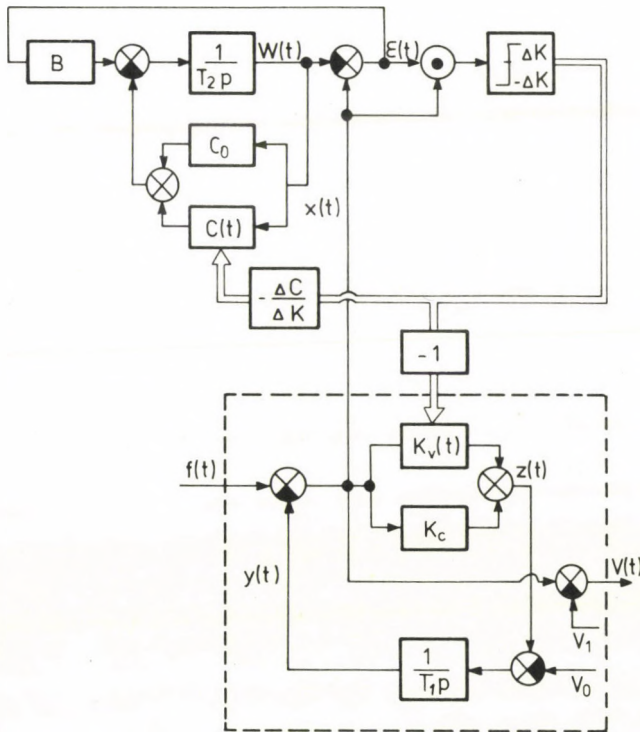


Fig. 5

where V_0 is a constant signal, the necessity of its introduction will be explained below,

$$K_v(t) = \begin{cases} -\Delta K, & \text{if } x(t)\dot{\varepsilon}(t) > 0, \\ \Delta K, & \text{if } x(t)\dot{\varepsilon}(t) < 0. \end{cases} \quad (23)$$

So, as compared with the schemes examined above, the sign of the CPF has been changed.

The right-hand parts of (22) have a break on the plane $\varepsilon=0$. Therefore a sliding mode can arise in the system on this plane. The conditions of its appearance look like [3].

$$\lim_{\varepsilon \rightarrow 0^+} \dot{\varepsilon} < 0, \quad (24)$$

$$\lim_{\varepsilon \rightarrow 0^-} \dot{\varepsilon} > 0.$$

Using (22) and (23) the inequalities (24) are transformed into

$$-\left(\frac{\Delta C}{T_2} - \frac{\Delta K}{T_1}\right)|x| < \dot{f} + \frac{V_0}{T_1} - \left(\frac{K_c}{T_1} - \frac{C_0}{T_2}\right)x < \left(\frac{\Delta C}{T_2} - \frac{\Delta K}{T_1}\right)|x|. \quad (25)$$

In such a way, if at any moment of time $t^* \geq t_0$ there is $\varepsilon(t^*)=0$ and the values $x(t^*)$ and $\dot{f}(t^*)$ satisfy inequalities (25), then a sliding mode arises. A law of time dependence for the signal $x(t)$ in this case can be obtained by determining $K_v(t)$ from the equations $\varepsilon=0$ and $\dot{\varepsilon}=0$ and substituting it into the first equation of system (22). After the necessary transformations we have

$$\dot{x} = \frac{1}{\alpha T_D} \left(T_D \dot{f} - x + \frac{T_D}{T_1} V_0 \right), \quad (26)$$

where

$$T_D = \frac{1}{\frac{K_c}{T_1} - C_0 - \frac{\Delta K}{T_1 \Delta C}}, \quad (27)$$

$$\alpha = 1 - \frac{T_2 \Delta K}{T_1 \Delta C}. \quad (28)$$

Obviously, the moment t^* may exist only in the case when the region in the space (\dot{f}, x) determined by the inequalities (25) is not empty. A necessary and sufficient condition for this is the inequality

$$\frac{\Delta C}{T_2} - \frac{\Delta K}{T_1} > 0. \quad (29)$$

Equation (26) approximates the law of time dependence for $x(t)$ as long as conditions (25) of the sliding mode existence are correct. It can be proved that inequalities (25) will be correct for $x(t)$, which satisfies equation (26) for all $t \geq t^*$ if the next inequalities are correct

$$\frac{K_c}{T_1} - \frac{C_0}{T_2} - \left(\frac{\Delta C}{T_2} - \frac{\Delta K}{T_1} \right) > 0 \quad (30)$$

and

$$-\frac{\Delta C}{T_2} \left| f^i + \frac{V_0}{T_1} \right| < \ddot{f} + \frac{C_0}{T_2} \left(f^i + \frac{V_0}{T_1} \right) < \frac{\Delta C}{T_2} \left| f^i + \frac{V_0}{T_1} \right|. \quad (31)$$

Inequalities (31) define some region G_1 on the plane (f^i, \ddot{f}) which is the same as the region G shaded in Fig. 2 and defined by inequalities (6). Obviously the region G_1 is not empty only in the case when

$$\Delta C > 0. \quad (32)$$

For the input signals under consideration, inequalities (6) are correct for all $t \geq t_0$. Therefore, inequalities (31) are also correct for these signals for all $t \geq t_0$ if the region G coincides with the region G_1 or lies inside the region G_1 . Note that if $a_0 \neq 0$ such a mutual arrangement of the regions G and G_1 is only possible if there is a non-zero signal V_0 . This explains its introduction.

If the input signal $f(t)$ satisfies inequalities (31) for all $t \geq t_0$, the existence of the moment $t^* \geq t_0$ for which $\varepsilon(t^*) = 0$ and inequalities (25) are correct is ensured by the inequality

$$B > \frac{T_2}{\alpha T_D} - C_0 + \Delta C, \quad (33)$$

where α and T_D are defined by (27) and (28).

The signal

$$v(t) = x(t) - V_1 \quad (34)$$

is used as the output signal, where V_1 is the constant signal.

From (26) and (34) we obtain that when $V_1 = \frac{V_0 T_D}{T_1}$ for $t \geq t^*$ the signals $v(t)$ and $f(t)$ are connected by the linear differential equation

$$\dot{v} = \frac{1}{\alpha T_D} (T_D \dot{f} - v). \quad (35)$$

Therefore, for $t \geq t^*$ the connection between the output and input signals $f(t)$ can be represented by the transmissive function

$$W_6(p) = \frac{v(p)}{f(p)} = \frac{T_D p}{\alpha T_D p + 1}. \quad (36)$$

The element with the structure scheme shown in Fig. 5 solves the formulated signal differentiation problem. Actually, for the values of the constants a_0, a_1 and a_2 it is possible to choose values of $C_0, \Delta C, T_2$ and V_0 . So for any signal for which inequalities (6) are correct for $t \geq t_0$ inequalities (31) will be correct. Then the values of $K_c, \Delta K, T_1$ and B are chosen so that $T_D = T_D^*, \alpha T_D \leq \delta^*$ and the moment t^* exists. The scheme in Fig. 5 has only amplification, integration and relay sections. The algorithm for calculation of the values of the constants $C_0, \Delta C, T_2, V_0, K_c, \Delta K, T_1$ and B according to the values a_0, a_1, a_2, T_D^* and δ^* is given below. It follows from the algorithm that all the values of the amplification factors of the amplification sections ($C_0, \Delta C, K_c, \Delta K$ and B) are limited even in the case when $T_D^* \rightarrow \infty$ and $\delta^* \rightarrow 0$.

4. Method for calculation of values of the differentiating element constants

Values of the differentiating element constants can be calculated successively with the help of these formulas

$$\delta = \begin{cases} \min \{ \delta^*, T_D^* \}, & \text{if } a_1 > -\frac{1}{\min \{ \delta^*, T_D^* \}}, \\ -\frac{1}{a_1} \gamma, & \text{if } a_1 \leq -\frac{1}{\min \{ \delta^*, T_D^* \}}, \end{cases} \quad (37)$$

where γ is any positive number less than 1

$$T_2 = \delta, \quad (38)$$

$$C_0 = -\frac{a_1 + a_2}{2} T_2, \quad (39)$$

$$\Delta C = \frac{a_2 - a_1}{2} T_2, \quad (40)$$

$$V_0 = a_0 T_1, \quad (41)$$

$$\Delta K = \frac{\Delta C}{T_2} \left(1 - \frac{\delta}{T_D^*}\right) T_1 = \frac{a_2 - a_1}{2} \left(1 - \frac{\delta}{T_D^*}\right) T_1, \quad (42)$$

$$K_c = \left[\frac{1}{T_D^*} \left(1 - \frac{C_0}{T_2} \delta\right) + \frac{C_0}{T_2} \right] T_1 = \left[\frac{1}{T_D^*} + \frac{a_1 + a_2}{2} \left(\frac{\delta}{T_D^*} - 1\right) \right] T_1, \quad (43)$$

the value B is chosen in such a way that the inequality is correct

$$B > 1 - C_0 + \Delta C. \quad (44)$$

Mind that the constant signal value of V_1 is defined by

$$V_1 = \frac{V_0 T_D^*}{T_1}. \quad (45)$$

Note that as a result the constant T_1 may be given any positive value. Therefore, if necessary, it is possible to decrease K_c (43) and ΔK (42) as required by decreasing the value of T_1 .

The constant values defined by formulas (37)–(43) solve the formulated problem. Actually, it follows from (39)–(41) that inequalities (31) follow from inequalities (6), i.e. if inequalities (6) are correct for $f(t)$ for $t \geq t_0$ then inequalities (31) are correct, too. From (27), (28), (42) and (43) we obtain

$$T_D = T_D^*, \quad T_D \alpha = \delta, \quad (46)$$

from (37) it follows that $0 < \delta \leq \delta^*$, i.e. the transmissive function $W_6(p)$ (36) coincides with the transmissive function $W^*(p)$ (5) required. Inequality (32) follows from (40) and also from the inequality $a_2 > a_1$. Inequality (29) follows from (40), (42) and the inequality $\delta > 0$. Using (39) (40), (42) and (43), we obtain

$$\frac{K_c}{T_1} - \frac{C_0}{T_2} - \left(\frac{\Delta C}{T_2} - \frac{\Delta K}{T_1} \right) = \frac{1}{T_D^*} (1 + \delta a_1), \quad (47)$$

and inequality (30) follows from (37) and (47).

For the choice of constant values, inequality (33) follows from inequality (44) and (38), therefore the choice of B in accordance with (44) ensures existence of the time moment t^* .

It follows from the formulas (42) and (43) that even if $T_D^* \rightarrow \infty$ and $\delta \rightarrow 0$ simultaneously the values of ΔK and K_c remain limited. It follows from (38)–(40) and (44) that the values of C_0 , ΔC and B do not depend on T_D^* when $\delta \rightarrow 0$ remains limited.

Conclusion

The suggested differentiating element can be used for solving the problem of differentiation of signals from rather a wide class. Note that if for example the considered signals which have the values of $\dot{f}(t)$ and $\ddot{f}(t)$ belong to some limited set in space (\dot{f}, \ddot{f}) , then it is possible to choose the constants a_0 , a_1 and a_2 in such a way that for any signal of the kind inequalities (6) are correct and consequently we may differentiate any of such signals having the same values of the constants of the differentiating element.

Mind that for solving the signal differentiating problem the positive coordinate-parametric feedback was used, as compared to the object control problem, where negative coordinate-parametric feedback was used [2]. In the case of the signal differentiation problem the existence of the movement along the plane $\varepsilon = 0$ is achieved by the parametric feedback which is negative as is object control.

References

1. Yemel'yanov, S. V., Sliding mode application in the problems of multifold optimal differentiation. Problems of Control and Information Theory, vol. 9 (1), pp. 47-55 (1980).
2. Yemel'yanov, S. V., Korovin, S. K., Development of feedback types and their application to design of closed-loop dynamic systems. Problems of Control and Information Theory, vol. 10 (3), pp. 161-174 (1981).
3. Yemel'yanov S. V. (Ed), Theory of variable structure systems. Nauka, Moscow, 1970 (in Russian).

Применение новых типов обратных связей для решения задачи дифференцирования сигнала

С. В. ЕМЕЛЬЯНОВ, А. А. СОЛОВЬЕВ

(Москва)

В статье рассматривается задача дифференцирования сигнала и делается попытка ее решения на основе использования координатно-параметрических и параметрических обратных связей. Синтезируется структурная схема элемента, связь между входным $f(t)$ и выходным $v(t)$ сигналами которого может быть аппроксимирована с помощью передаточной функции $W(p) = \frac{v(p)}{f(p)} = \frac{T_D p}{\delta p + 1}$. Структурная схема элемента содержит только усилительные, интегрирующие и релейные звенья, причем за счет выбора конечных значений коэффициентов усиления усилительных звеньев и постоянных времени интегрирующих звеньев существует возможность получить в указанной передаточной функции любое, в том числе сколь угодно малое значение δ и любое, в том числе сколь угодно большое значение T_D . Задача решена при ограничениях на входной сигнал $f(t)$ вида

$$\frac{a_1 - a_2}{2} |\dot{f} + a_0| < \ddot{f} - \frac{a_1 + a_2}{2} (\dot{f} + a_0) < \frac{a_2 - a_1}{2} |\dot{f} + a_0|,$$

где $a_0, a_1 < a_2$ — постоянные. Приводится метод расчета значений параметров звеньев, входящих в структурную схему элемента по значениям постоянных a_0, a_1, a_2 , задающих ограничения на входной сигнал, и требуемым значениям T_D и δ , определяющим передаточную функцию элемента.

С. В. Емельянов

Всесоюзный НИИ системных исследований
СССР, 119034, Москва Г-34, ул. Рылеева, 29

СТОХАСТИЧЕСКИЙ ПРОГРАММНЫЙ СИНТЕЗ ОДНОГО ГАРАНТИРУЮЩЕГО УПРАВЛЕНИЯ

Н. Н. КРАСОВСКИЙ, В. Е. ТРЕТЬЯКОВ

(Свердловск)

(Поступила в редакцию 14 апреля 1982 г.)

В работе описывается построение оптимального гарантирующего управления для линейной системы при условии, что показатель качества процесса управления складывается из терминального члена и интеграла от квадратичной формы относительно управляющего воздействия и помехи. Цель работы — продемонстрировать на этой задаче возможности метода стохастического программного синтеза, предложенного в работах [1–2].

1. Рассмотрим систему, описываемую уравнением

$$\dot{x} = A(t)x + B(t)u + C(t)v, \quad u \in \mathcal{P}, \quad v \in \mathcal{Q}, \quad (1.1)$$

где x — n -мерный фазовый вектор, u — p -мерный вектор управления, v — q -мерный вектор неопределенной помехи, \mathcal{P} и \mathcal{Q} суть некоторые компакты; $A(t)$, $B(t)$, $C(t)$ — непрерывные матрицы-функции; время t меняется в пределах $t_0 \leq t \leq \vartheta$.

Будем называть стратегией управления функцию $u(\cdot) = \{u(t, x, \varepsilon) \in \mathcal{P}\}$, где $\varepsilon > 0$ — параметр точности. Пусть реализовалась исходная позиция $\{t_*, x_*\}$, $t_* \in [t_0, \vartheta]$. Выберем $\varepsilon > 0$ и какое-нибудь разбиение $\Delta_\delta\{\tau_i\}$ полуинтервала $[t_*, \vartheta]$ на полуинтервалы $[\tau_i, \tau_{i+1})$, $\tau_0 = t_*$, $\tau_m = \vartheta$, $\tau_{i+1} - \tau_i \leq \delta$. Движением $x[\cdot] = x[\cdot, t_*, x_*, u(\cdot), \varepsilon, \Delta_\delta] = \{x[t, t_*, x_*, u(\cdot), v[\cdot], \varepsilon, \Delta_\delta], t_* \leq t < \vartheta\}$, порожденным при этих условиях стратегией $u(\cdot)$ будем называть решение пошагового уравнения

$$\dot{x}[t] = A(t)x[t] + B(t)u(\tau_i, x[\tau_i], \varepsilon) + C(t)v[t], \quad (1.2)$$

$$x[t_*] = x_*, \quad \tau_i \leq t < \tau_{i+1}, \quad i = 0, \dots, m-1.$$

Здесь реализацией помехи $v[\cdot] = \{v[t] \in \mathcal{Q}, t_* \leq t \leq \vartheta\}$ может быть любая измеримая функция со значениями в \mathcal{Q} .

Пусть задан показатель качества процесса управления

$$\gamma(x[\cdot], u[\cdot], v[\cdot]) = \int_{t_*}^{\vartheta} R(t, u[t], v[t]) dt + |x[\vartheta]|, \quad (1.3)$$

$$R(t, u, v) = \langle \Phi(t)u \cdot u \rangle + \langle \Psi(t)v \cdot v \rangle. \quad (1.4)$$

Здесь $\Phi(t) = \{\varphi_{ij}(t), ij=1, 2, \dots, p\}$, $\Psi(t) = \{\psi_{ij}(t), ij=1, 2, \dots, q\}$ — непрерывные симметричные матрицы-функции; $|x|$ — евклидова норма, $\langle \cdot \cdot \rangle$ — скалярное произведение, $u[\cdot] = \{u[t] \in \mathcal{P}, t_* \leq t < \vartheta\}$ — реализация стратегии $u(\cdot)$ вдоль порожденного ею движения $x[\cdot]$.

Гарантированным результатом [3, 4] $\rho(t_*, x_*, u(\cdot))$ для стратегии $u(\cdot)$ будем называть величину

$$\rho(t_*, x_*, u(\cdot)) = \overline{\lim}_{\varepsilon \rightarrow 0} \lim_{\delta \rightarrow 0} \sup_{v[\cdot], \Delta_\delta} \gamma(x[\cdot], u[\cdot], v[\cdot]). \quad (1.5)$$

Требуется найти оптимальное гарантирующее управление — стратегию $u^0(\cdot)$, которая для всякой возможной исходной позиции $\{t_*, x_*\}$ обеспечивает минимальный гарантированный результат

$$\rho^0(t_*, x_*) = \rho(t_*, x_*, u^0(\cdot)) = \min_{u(\cdot)} \rho(t_*, x_*, u(\cdot)). \quad (1.6)$$

Таким образом, стратегия $u^0(\cdot)$ обладает следующим свойством. Какова бы ни была исходная позиция $\{t_*, x_*\}$ и каким бы числом $\zeta > 0$ ни задались, найдется число $\varepsilon(\zeta) > 0$ и затем функция $\delta(\zeta, \varepsilon) > 0$ так, что для всякого движения $x[\cdot] = x[\cdot, t_*, x_*, u^0(\cdot), v[\cdot], \varepsilon, \Delta_\delta]$ будет гарантировано неравенство

$$\int_{t_*}^{\vartheta} R(t, u^0[t], v[t]) dt + |x[\vartheta]| \leq \rho^0(t_*, x_*) + \zeta \quad (1.7)$$

какова бы ни была помеха $v[\cdot]$, если только $\varepsilon \leq \varepsilon(\zeta)$ и $\delta \leq \delta(\zeta, \varepsilon)$. И никакая стратегия $u(\cdot)$ не может гарантировать неравенство $\gamma(x[\cdot, t_*, x_*, u(\cdot), v[\cdot], \varepsilon, \Delta_\delta], u[\cdot], v[\cdot]) < \rho^0(t_*, x_*) - \zeta$ при всех возможных реализациях помехи $v[\cdot]$ и разбиениях Δ_δ с достаточно малым шагом $\delta > 0$.

Оптимальное гарантирующее управление $u^0(t, x, \varepsilon)$ существует. Оно строится по величине $\rho^0(t, x)$ следующим образом [2]. В позиции $\{\tau, x_\tau\}$, $t_* \leq \tau \leq \vartheta$ искомое значение $u^0(\tau, x_\tau, \varepsilon)$ определяется из условия

$$\begin{aligned} \max_{v \in \mathcal{Z}} [\langle (x_\tau - w_\tau) \cdot (B(\tau)u^0(\tau, x_\tau, \varepsilon) + C(\tau)v) \rangle + \\ + c_\tau \cdot R(\tau, u^0(\tau, x_\tau, \varepsilon), v)] = \min_{u \in \mathcal{P}} \max_{v \in \mathcal{Z}} [\langle x_\tau - w_\tau \cdot \\ \cdot (B(\tau)u + C(\tau)v) \rangle + c_\tau R(\tau, u, v)], \end{aligned} \quad (1.8)$$

где $\{w_\tau, c_\tau\}$ — сопутствующая точка, удовлетворяющая условию

$$\rho^0(\tau, w_\tau) - c_\tau = \min_{w, c} [\rho^0(\tau, w) - c] \quad (1.9)$$

при $|x_\tau - w|^2 + c^2 \leq \varepsilon(1 + [\tau - t_*]) \exp 2L[\tau - t_*]$.

Здесь $L = \max \|A(t)\|$ при $t_0 \leq t \leq \vartheta$, $\|A\|$ — евклидова норма матрицы A .

Таким образом, построение оптимального гарантирующего управления $u^0(t, x, \varepsilon)$ сводится к вычислению функции $\rho^0 = \rho^0(t, x)$.

2. Опишем процедуру вычисления оптимального гарантированного результата $\rho^0(t, x)$ (1.6), основанную на методе стохастического программного синтеза [1, 2].

Рассмотрим w -модель [2], отвечающую уравнению (1.1) и показателю качества (1.3), (1.4). Состояние этой модели в текущий момент времени t будем характеризовать $(n+1)$ -мерным вектором $y[t] = \{w[t], y_{n+1}[t]\}$, где w — n -мерный вектор. Изменение фазового вектора $y[t]$, $t_* \leq t \leq \vartheta$ в w -модели будет определяться уравнениями

$$\dot{w} = A(t)w + f + C(t)v, \quad (2.1)$$

$$\dot{y}_{n+1} = g_{n+1} + \langle \Psi(t)v \cdot v \rangle,$$

где управление v стеснено условием $v \in \mathcal{Q}$, а $(n+1)$ -мерный вектор управления $g = \{f, g_{n+1}\}$ — условием

$$g \in F(t); \quad F(t) = \overline{\text{co}}\{h : h = \{B(t)u, \langle \Phi(t)u \cdot u \rangle, u \in P\}\}. \quad (2.2)$$

Здесь $\overline{\text{co}}\{\dots\}$ — замкнутая выпуклая оболочка множества $\{\dots\}$.

Выберем в качестве базового вероятностного элемента стандартный процесс броуновского движения $z[t, \omega]$, $t_* \leq t \leq \vartheta$, где ω есть элементарное событие из соответствующего вероятностного пространства $\{\Omega, \mathcal{A}, P\}$ ([5], стр. 39).

Будем называть неупреждающими программами $g(t_*[\cdot]\vartheta)$, $v(t_*[\cdot]\vartheta)$ функции $g(t, \omega) \in F(t)$, $v(t, \omega) \in \mathcal{Q}$, $t_* \leq t \leq \vartheta$, $\omega \in \Omega$, неупреждающие по отношению к процессу $z[t, \omega]$ (т.е. по отношению к семейству δ -алгебр $\{F_t^z\}$, связанных с процессом $z[\tau, \omega]$, $t_* \leq \tau \leq t$ ([5], стр. 100)). Иначе говоря, неупреждающие программы $g(t_*[\cdot]\vartheta)$ и $v(t_*[\cdot]\vartheta)$ — это функции двух аргументов t и ω , которые почти наверное определяются равенствами $g(t, \omega) = g_*(t, z(t_*[\cdot]t, w))$,

$v(t, \omega) = v_*(t, z(t_*[\cdot]t, \omega))$, где символом $z(t_*[\cdot]t, \omega)$ при фиксированном $\omega \in \Omega$ обозначена реализация броуновского движения $z[\tau, \omega]$ на отрезке $t_* \leq \tau \leq t$. Содержательно это означает, что величины воздействий g и v в момент времени t определяются программами $g(t_*[\cdot]t)$ и $v(t_*[\cdot]t)$ лишь на основании истории броуновского движения $z[\tau, \omega]$, $t_* \leq \tau \leq t$, которая реализовалась к моменту времени t .

При заданной исходной позиции $\{t_*, y_* = \{w_* y_{n+1*}\}\}$ пара программ $\{g(t_*[\cdot]t), v(t_*[\cdot]t)\}$ определяет в w -модели (2.1) случайное движение $y[t, \omega] = y[t, \omega, t_*, y_*, g(t_*[\cdot]t), v(t_*[\cdot]t)]$, которое при $t_* \leq t \leq t$ определяется как решение интегрального уравнения

$$y[t, \omega] = y_* + \int_{t_*}^t \left[\begin{array}{c} A(\tau)w[\tau, \omega] + f(\tau, \omega) + C(\tau)v(\tau, \omega) \\ g_{n+1}(\tau, \omega) + \langle \Psi(\tau)v(\tau, \omega) \cdot v(\tau, \omega) \rangle \end{array} \right] d\tau. \quad (2.3)$$

Рассмотрим величину

$$\rho^*(t_*, y_*) = \sup_{v(t_*[\cdot]t)} \inf_{g(t_*[\cdot]t)} M\{y_{n+1}[t, \omega] + |w[t, \omega]|\}, \quad (2.4)$$

где M — математическое ожидание. Справедливо следующее утверждение, которое доказывается подобно тому, как обосновывается аналогичное утверждение в статье [1].

Теорема 2.1. Для всякой возможной исходной позиции $\{t_*, x_*\}$ оптимальный гарантированный результат $\rho^0(t_*, x_*)$ (1.6) определяется равенством

$$\rho^0(t_*, x_*) = \rho^*(t_*, \{x_*, 0\}). \quad (2.5)$$

Переход к задаче, двойственной по отношению к задаче (2.4), приводит к следующей конструкции. Пусть $(n+1)$ -мерная вектор-функция $r[t, \omega] = \{s[t, \omega], r_{n+1}[t, \omega]\}$, $r[t_*, \omega] = \{s[t_*], r_{n+1}[t_*]\}$ есть неупреждающее случайное решение диффузионного уравнения

$$dr = \begin{bmatrix} ds \\ dr_{n+1} \end{bmatrix} = \begin{bmatrix} -A'(t)sdt + m(t, \omega)dz[t, \omega] \\ q_{n+1}(t, \omega)dz[t, \omega] \end{bmatrix}, \quad (2.6)$$

где $\{m(t, \omega), q_{n+1}(t, \omega)\} = q(t, \omega)$ есть некоторая неупреждающая по отношению к процессу $z[t, \omega]$ $(n+1)$ -мерная вектор-функция; верхний индекс штрих означает транспонирование. Пусть $\|y(\cdot)\|_*$ и $\|r(\cdot)\|_*$ — нормы в пространствах $(n+1)$ -

мерных случайных величин $y(\omega)$ и $r(\omega)$, $\omega \in \Omega$, определенные соотношениями

$$\|y(\cdot)\|_* = M\{|y_{n+1}(\omega)|\} + (M\{|w(\omega)|^2\})^{1/2}, \quad (2.7)$$

$$\|r(\cdot)\|^* = \max \{ \text{vraisup } |r_{n+1}(\omega)|, (M\{|s(\omega)|^2\})^{1/2} \}, \quad (2.8)$$

где $|x|$ — по-прежнему евклидова норма вектора x . Справедливо утверждение.

Лемма 2.1. Величина $\rho^*(t_*, y_*)$ (2.4) равна точной верхней грани тех чисел β , для которых справедливо неравенство

$$\begin{aligned} & \sup_{\|r[\vartheta, \cdot]\|^* \leq 1} [\langle r[t_*] \cdot y_* \rangle + M\{ \int_{t_*}^{\vartheta} \max_{v \in \mathcal{Z}} \min_{u \in \mathcal{P}} [\langle s[\tau, \omega] \cdot \\ & \cdot (B(\tau)u + C(\tau)v) \rangle + r_{n+1}[\tau, \omega] \cdot R(\tau, u, v)] d\tau \} - \\ & - \max_{y(\cdot) \in Y_\beta} M\{\langle r[\vartheta, \omega] \cdot y(\omega) \rangle\} > 0, \end{aligned} \quad (2.9)$$

где Y_β — множество в пространстве случайных величин $y(\cdot)$, определенное соотношением

$$Y_\beta = \{y(\cdot) : M\{|y_{n+1}(\omega) + |w(\omega)|\} \leq \beta\}. \quad (2.10)$$

В неравенстве (2.9) $r[\vartheta, \cdot] = \{r[\vartheta, \omega], \omega \in \Omega\}$ есть элемент решения $r[t, \cdot]$, $t_* \leq t \leq \vartheta$ уравнения (2.6), порожденного некоторым начальным условием $r[t_*]$ и некоторой неупреждающей функцией (управлением) $q(t, \omega)$. Стало быть, искомыми в двойственной задаче (2.9) являются начальное условие $r[t_*]$ и неупреждающее управление $q(t, \omega)$ для стохастической системы (2.6).

Справедливость леммы 2.1 доказывается подобно тому, как это делается при аналогичных обстоятельствах в математической теории управления [4] на основании теорем о разделении выпуклых множеств. Здесь, однако, соответствующие построения рассматриваются в функциональном пространстве случайных величин $y(\cdot)$ с нормой $\|y(\cdot)\|_*$ (2.7).

Из вида последнего в (2.9) слагаемого

$$\begin{aligned} & \max_{y(\cdot) \in Y_\beta} M\{\langle r[\vartheta, \omega] \cdot y(\omega) \rangle\} = \\ & = \max_{y(\cdot) \in Y_\beta} M\{r_{n+1}[\vartheta, \omega] \cdot y_{n+1}(\omega) + \langle s[\vartheta, \omega] \cdot w(\omega) \rangle\} \end{aligned} \quad (2.11)$$

и определения множества Y_β (2.10) следует, что для максимизирующих элементов $r[\vartheta, \omega] = \{s[\vartheta, \omega], r_{n+1}[\vartheta, \omega]\}$ при почти всех ω должно выполняться

неравенство $|s[\vartheta, \omega]| \leq 1$. В самом деле, если бы на множестве $\mathcal{E} \subset \Omega$ ненулевой меры имело бы место неравенство $|s[\vartheta, \omega]| > 1$, то, поскольку с учетом (2.8) $\text{vraisup } r_{n+1}[\vartheta, \omega] \leq 1$, величины $w(\omega)$ и $y_{n+1}(\omega)$ на множестве \mathcal{E} можно было бы выбрать так, что величина (2.11) делается сколь угодно большой. Следовательно, элементы $r[\vartheta, \omega]$, для которых $|s[\vartheta, \omega]| > 1$ на \mathcal{E} , не могут рассматриваться в качестве максимизирующих для (2.9).

Дальнейший анализ неравенства (2.9) показывает, что двойственная задача (2.9) редуцируется к случаю $r_{n+1}[\tau, \omega] \equiv 1$, $t_* \leq \tau \leq \vartheta$ (в уравнении (2.6) $q_{n+1}(t, \omega) \equiv 0$, $r_{n+1}[t_*] = 1$), $\max_{y(\cdot) \in Y_\beta} M\{\langle r[\vartheta, \omega] \cdot y(\omega) \rangle\} = \beta$, и тогда величина $\rho^*(t_*, y_*)$ (2.4) определяется равенством

$$\begin{aligned} \rho^*(t_*, y_*) = & \sup_{\text{vraisup } |s[\vartheta, \omega]| \leq 1} [\langle s[t_*] \cdot w_* \rangle + y_{n+1} + \\ & + M\{ \int_{t_*}^{\vartheta} \max_{v \in \mathcal{Q}} \min_{u \in \mathcal{P}} [\langle s[\tau, \omega] \cdot (B(\tau)u + C(\tau)v) \rangle + \\ & + R(\tau, u, v)] d\tau \}]. \end{aligned} \quad (2.12)$$

Сделаем теперь одно предположение о квадратичной форме $R(t, u, v)$. Примем, что задача на максимум под знаком интеграла в (2.12) имеет решение $\{u^0_\mathcal{P}(\tau, \omega), v^0_\mathcal{Q}(\tau, \omega)\}$, которое совпадает с решением $\{u^0(\tau, \omega), v^0(\tau, \omega)\}$ такой же задачи на максимум, но уже без априорных ограничений $u \in \mathcal{P}$, $v \in \mathcal{Q}$. Это предположение существенно упрощает дальнейшие выкладки и облегчает получение эффективного решения. Оно выполняется, если в (1.4) при $t_* \leq \tau \leq \vartheta$ матрица $\Phi(\tau)$ определенно-положительна, $\Psi(\tau)$ — определенно-отрицательна и абсолютные значения определителей $|\Phi(\tau)|$ и $|\Psi(\tau)|$ не менее положительного числа N , достаточно большого по сравнению с размерами \mathcal{P} и \mathcal{Q} . При этом решение $\{u^0(\tau, \omega), v^0(\tau, \omega)\}$ задачи на максимум по u и v в выражении (2.12) без априорных ограничений на u и v имеет вид

$$\begin{aligned} u^0(\tau, \omega) &= -\frac{1}{2} \Phi^{-1}(\tau) B'(\tau) s[\tau, \omega], \\ v^0(\tau, \omega) &= -\frac{1}{2} \Psi^{-1}(\tau) C'(\tau) s[\tau, \omega], \end{aligned} \quad (2.13)$$

и после подстановки (2.13) в (2.12) для $\rho^*(t_*, y_*)$ (2.4) получаем выражение

$$\rho^*(t_* y_*) = \sup_{\text{vraisup}_{|s[\vartheta, \omega]| \leq 1}} [\langle s[t_*] \cdot w_* \rangle + y_{n+1*} + \int_{t_*}^{\vartheta} M\{G(\tau, s[\tau, \omega])\} d\tau], \quad (2.14)$$

где $G(\tau, s)$ есть квадратичная форма

$$G(\tau, s) = \langle K(\tau)s \cdot s \rangle \quad (2.15)$$

$$K(\tau) = -\frac{1}{4} [B(\tau)\Phi^{-1}(\tau)B'(\tau) + C(\tau)\Psi^{-1}(\tau)C'(\tau)]. \quad (2.16)$$

Итак, задача о вычислении оптимального гарантированного результата $\rho^0(t_* x_*)$ согласно (2.5), (2.6), (2.14), (2.15) сводится при сделанном предположении о характере квадратичной формы $R(t, u, v)$ (1.4) к задаче об оптимальном программном управлении $m(t, \omega)$ стохастической системой

$$ds = -A'(t)sdt + m(t, \omega) dz[t, \omega], \quad s[t_*, \omega] = s[t_*] \quad (2.17)$$

при условии максимума (2.14) при ограничении $\text{vraisup}_{|s[\vartheta, \omega]| \leq 1}$. В такой задаче искомыми являются начальное условие $s[t_*]$ и неупреждающее управление $m(\cdot) = \{m(t, \omega), t_* \leq t < \vartheta, \omega \in \Omega\}$. Строго говоря, здесь речь идет лишь о максимизирующей для (2.14) последовательности управлений $m^{(j)}(\cdot)$, $j = 1, 2, \dots$, так как не утверждается, что верхняя грань в (2.14) достигается на некотором решении $s^0(\cdot)$ уравнения (2.17) при каком-то оптимальном управлении $m^0(\cdot)$.

3. Положим $s[\tau, \omega] = X'[\vartheta, \tau]l(\tau, \omega)$, где $X[t, \tau]$ — фундаментальная матрица решений для уравнения $\dot{x} = A(t)x$. Тогда задача об отыскании оптимального гарантированного результата $\rho^0(t_* x_*)$ с учетом (2.5), (2.14), (2.17) трансформируется в задачу

$$\rho^0(t_* x_*) = \sup_{\text{vraisup}_{|l[\vartheta, \omega]| \leq 1}} [\langle l[t_*] \cdot X[\vartheta, t_*]x_* \rangle + \int_{t_*}^{\vartheta} M\{G_*(\tau, l(\tau, \omega))\} d\tau, \quad (3.1)$$

$$dl = a(t, \omega)dz[t, \omega], \quad l[t_*, \omega] = l[t_*], \quad (3.2)$$

где $a(t, \omega) = [X'[\vartheta, t]]^{-1}m(t, \omega)$ и

$$G_*(\tau, l) = \langle K_*(\tau)l \cdot l \rangle, \quad K_*(\tau) = X[\vartheta, \tau]K(\tau)X'[\vartheta, \tau]. \quad (3.3)$$

Применяя формулу замены переменных Ито ([5], стр. 141) при фиксированном τ к квадратичной форме $G_*(\tau, l(\xi, \omega))$, $t_* \leq \xi \leq \tau$ с последующим усреднением получаем

$$M\{G_*(\tau, l(\tau, \omega))\} = G_*(\tau, l[t_*]) + \int_{t_*}^{\tau} M\{G_*(\tau, a(\xi, \omega))\}d\xi. \quad (3.4)$$

Подставляя (3.4) в (3.1) и меняя порядок интегрирования по τ и ξ , окончательно находим, что

$$\begin{aligned} \rho^0(t_*x_*) &= \sup_{\text{vraisup}_{|l[\vartheta, \omega]| \leq 1}} [\langle l[t_*] \cdot X[\vartheta, t_*]x_* \rangle + \\ &+ H(t_*, l[t_*]) + \int_{t_*}^{\vartheta} M\{H(\tau, a(\tau, \omega))\}d\tau], \end{aligned} \quad (3.5)$$

где

$$H(\tau, a) = \langle \Gamma(\tau)a \cdot a \rangle, \quad \Gamma(\tau) = \int_{\tau}^{\vartheta} K_*(\xi)d\xi. \quad (3.6)$$

Таким образом, задача о вычислении величины $\rho^0(t_*x_*)$ оказывается эквивалентной задаче определения максимизирующей последовательности $a^{(j)}(\cdot) = \{a^{(j)}(t, \omega), t_* \leq t < \vartheta, \omega \in \Omega\}$, $j = 1, 2, \dots$ для неупреждающего управления $a(\cdot)$ и начального условия $l^0[t_*]$ для стохастической системы (3.2) при условии максимизации величины (3.5).

Решение этой задачи (3.5), (3.6), (3.2) таково. Если квадратичная форма $H(\tau, a)$ при всех $\tau \in [t_*, \vartheta]$ является знакоотрицательной, то оптимальным будет тривиальное управление $a(t, \omega) \equiv 0$, и тогда $l(\tau, \omega) \equiv l[t_*]$ при всех $\tau \in [t_*, \vartheta]$ и $\omega \in \Omega$. Эта ситуация отвечает регулярному случаю [4]. Здесь задача о вычислении $\rho^0(t_*x_*)$ разрешается на основе детерминированной программной конструкции, т.е. сводится к отысканию максимизирующего вектора $l^0 = l^0[t_*]$ при $|l[t_*]| \leq 1$ из условия (3.5), где положено $a(\tau, \omega) \equiv 0$.

Пусть, однако, квадратичная форма $H(\tau, a)$ (3.6) при некоторых значениях $\tau \in [t_*, \vartheta]$ и a может принимать положительные значения. Пусть при этом $\tau_* = \tau_*[t_*] \in [t_*, \vartheta]$ есть такое значение τ , при котором максимальное собственное значение $\lambda(\tau)$ квадратичной формы $H(\tau, a)$ достигает максимума, т.е.

$$\lambda(\tau_*[t_*]) = \max_{t_* \leq \tau \leq \vartheta} \lambda(\tau), \quad \lambda(\tau) = \max_{|a| \leq 1} H(\tau, a). \quad (3.7)$$

Тогда задача (3.5) при ограничении

$$\text{vraisup} |I[\vartheta, \omega]| = \text{vraisup} |I[t_*] + \int_{t_*}^{\vartheta} a(t, \omega) dz[t, \omega]| \leq 1 \quad (3.8)$$

имеет следующую максимизирующую последовательность $a^{(j)}(\cdot)$. Заметим, что $\tau_* < \vartheta$, поскольку $\lambda(\vartheta) = 0$, и зададим последовательность значений $\tau_j = \tau_* + \varepsilon_j \in [t_*, \vartheta]$, $\varepsilon_j > 0$, $\lim \varepsilon_j = 0$ при $j \rightarrow \infty$. Полагаем $a^{(j)}(t, \omega) \equiv 0$ при $t_* \leq t < \tau_*$ и $\tau_* + \varepsilon_j \leq t < \vartheta$. При $\tau_* \leq t < \tau_* + \varepsilon_j$ управление $a^{(j)}(t, \omega)$ выбирается следующим образом. Рассмотрим соответствующую максимизирующую последовательность $I^{(j)}[t_*]$. Выбирая из нее сходящуюся подпоследовательность (пусть таковой является уже сама последовательность $I^{(j)}[t_*]$), обозначим $\lim I^{(j)}[t_*] = I^0[t_*]$ при $j \rightarrow \infty$. В рассматриваемом случае $|I^0[t_*]| < 1$. Пусть e — единичный собственный вектор, который в соответствии с (3.7) удовлетворяет условию

$$H(\tau_*, e) = \lambda(\tau_*) \quad (3.9)$$

Определим положительные числа a^+ и a^- из условий

$$|I^0[t_*] + a^+ e| = 1, \quad |I^0[t_*] - a^- e| = 1. \quad (3.10)$$

Разобьем все пространство Ω на два подмножества B_j^+ и B_j^- следующим образом

$$\begin{aligned} B_j^+ &= \{\omega : z[\tau_* + \varepsilon_j, \omega] - z[\tau_*, \omega] \geq z_*\}, \\ B_j^- &= \{\omega : z[\tau_* + \varepsilon_j, \omega] - z[\tau_*, \omega] < z_*\}, \end{aligned} \quad (3.11)$$

где число z_* выбрано из условия

$$\frac{1}{\sqrt{2\pi\varepsilon_j z_*}} \int_0^\infty e^{-\xi^2/2\varepsilon_j} d\xi = a^- / (a^+ + a^-).$$

Имеем $P(B_j^+) = a^- / (a^+ + a^-)$, $P(B_j^-) = a^+ / (a^- + a^+)$. Выберем векторную случайную величину

$$\begin{aligned} I^{(j)}(\omega) &= I^0[t_*] + a^+ e, \quad \omega \in B_j^+, \\ I^{(j)}(\omega) &= I^0[t_*] - a^- e, \quad \omega \in B_j^-. \end{aligned} \quad (3.12)$$

Имеем $M\{I^{(j)}(\omega)\} = I^0[t_*]$. Согласно ([5], стр. 186) существует случайный процесс $I^{(j)}[t, \omega]$, который удовлетворяет условиям (3.2) и условию

$$I^{(j)}[\vartheta, \omega] = I^{(j)}(\omega). \quad (3.13)$$

При этом для соответствующего управления $a^{(j)}(t, \omega)$ справедливо равенство

$$M\{|l^{(j)}[\vartheta, \omega]|^2\} = |l^0[t_*]|^2 + \int_{t_*}^{t_* + \varepsilon_j} M\{|a^{(j)}[\tau, \omega]|^2\} d\tau = 1. \quad (3.14)$$

Отсюда вытекает, что управление $a^{(j)}[\tau, \omega]$ дает величину

$$\begin{aligned} \kappa^{(j)} = & \langle l^0[t_*] \cdot X[\vartheta, t_*] x_* \rangle + H(t_*, l^0[t_*]) + \\ & + \lambda(\tau_*[t_*])(1 - |l^0[t_*]|^2) + \varepsilon_j \varphi_j, \end{aligned} \quad (3.15)$$

где $\lim_{j \rightarrow \infty} \varphi_j = 0$ при $j \rightarrow \infty$.

Стало быть, для построенной последовательности $a^{(j)}(\cdot)$, $j = 1, 2, \dots$ имеет место соотношение

$$\begin{aligned} \lim_{j \rightarrow \infty} \kappa^{(j)} = & \langle l^0[t_*] \cdot X[\vartheta, t_*] x_* \rangle + \\ & + H(t_*, l^0[t_*]) + \lambda(\tau_*[t_*])(1 - |l^0[t_*]|^2) = \kappa^0. \end{aligned} \quad (3.16)$$

Эта величина κ^0 (3.16) и равна верхней грани в соотношении (3.5). В самом деле, для того, чтобы убедиться в том, что при данном значении $l^0[t_*]$ никакая возможная последовательность $a^{(j)}(\cdot)$ не может для (3.5) дать величину, большую, чем κ^0 (3.16), достаточно заметить, что и при условии только (3.14), более широком, нежели условие (3.8), нельзя построить последовательность $a^{(j)}(\cdot)$, $j = 1, 2, \dots$, дающую значение большее, чем κ^0 (3.16). Это последнее утверждение проверяется уже непосредственно по (3.5) и (3.14).

Итак, величина оптимального гарантированного результата $\rho^0(t_*, x_*)$ определяется равенством

$$\begin{aligned} \rho^0(t_*, x_*) = & \max_{|l| \leq 1} [\langle l \cdot X[\vartheta, t_*] x_* \rangle + \\ & + H(t_*, l) - \lambda(\tau_*[t_*])|l|^2] + \lambda(\tau_*[t_*]), \end{aligned} \quad (3.17)$$

где с учетом (2.16), (3.3), (3.6)

$$\begin{aligned} H(t_*, l) = & \langle \Gamma(t_*) l \cdot l \rangle, \\ \Gamma(t_*) = & -\frac{1}{4} \int_{t_*}^{\vartheta} X[\vartheta, \tau] \cdot (B(\tau) \Phi^{-1}(\tau) B'(\tau) + \\ & + C(\tau) \Psi^{-1}(\tau) C'(\tau)) X'[\vartheta, \tau] d\tau, \end{aligned} \quad (3.18)$$

и число $\lambda(\tau_*[t_*])$ определяется из условий (3.7).

Заметим в заключение, что при $\lambda(\tau_*[t_*]) > 0$ максимум в (3.5) не достигается на каком-либо допустимом управлении $a(\cdot)$. Из приведенного выше построения максимизирующей последовательности $a^{(j)}(t, \omega)$ видно, что в данном случае более соответствует духу задачи выбор в качестве базового вероятностного элемента не броуновского процесса $z[t, \omega]$, а некоторого процесса $z[t, \omega]$ с независимыми приращениями, допускающими разрывы. Однако, такой выбор привел бы к более громоздким промежуточным выкладкам.

4. Постановка исходной задачи (1.1)–(1.6) не исключает того, что помеха $v[\cdot]$ в объекте (1.1) может быть, в частности, той или иной реализацией какого-то управления v , формируемого по принципу обратной связи. Поэтому, называя стратегией управления v функцию $v(\cdot) = \{v(t, x, \varepsilon) \in \mathcal{Q}\}$, можно поставить задачу о выборе оптимальной стратегии $v^0(\cdot)$, которая для всякой возможной исходной позиции $\{t_*, x_*\}$ обеспечивает максимальный гарантированный результат

$$\rho_0(t_* x_*) = \rho(t_* x_* v^0(\cdot)) = \max_{v(\cdot)} \rho(t_* x_* v(\cdot)), \quad (4.1)$$

где $\rho(t_* x_* v(\cdot))$ есть гарантированный результат для выбранной стратегии $v(\cdot)$, определяемый с понятными изменениями подобно (1.5). При этом движение $x[\cdot]$, порожденное стратегией $v(\cdot)$, формируется здесь как решение соответствующего пошагового уравнения по схеме, подобной (1.2).

Известно ([2], стр. 581), что стратегия $v^0(\cdot)$ существует и величина $\rho_0(t_* x_*)$ (4.1) совпадает с величиной $\rho^0(t_* x_*)$ (1.6), которая, как показано, вычисляется по формуле (3.17). Оптимальное гарантирующее управление $v^0(t, x, \varepsilon)$ определяется по величине $\rho_0(t_* x_*) = \rho^0(t_* x_*)$ (3.17), как и управление $u^0(t, x, \varepsilon)$, путем экстремального наведения объекта на сопутствующую точку, найденную из условия (1.9), в котором только \min заменяется теперь на \max .

Таким образом, наряду с неравенством (1.7)

$$\gamma(x[\cdot], u^0[\cdot], v[\cdot]) \leq \rho^0(t_* x_*) + \zeta, \quad (4.2)$$

которое обеспечивается стратегией $u^0(\cdot)$, имеем неравенство

$$\gamma(x[\cdot], u[\cdot], v^0[\cdot]) \geq \rho^0(t_* x_*) - \zeta, \quad (4.3)$$

гарантированное стратегией $v^0(\cdot)$ для любой измеримой реализации $u[\cdot] = \{u[t] \in \mathcal{P}, t_* \leq t \leq \vartheta\}$, если только $\varepsilon \leq \delta(\zeta)$ и $\delta \leq \delta(\zeta, \varepsilon)$.

Но из результатов настоящей статьи вытекает, что управление v , гарантирующее результат не меньший, чем $\rho^0(t_* x_*) - \zeta$, может быть построено для данной исходной позиции $\{t_* x_*\}$ и как двухшаговая $Q\{t_* x_*\}$ — процедура [2]. Именно, на первом шаге $(t_*, \tau_* = \tau_*[t_*])$ эта процедура назначает управление

v как функцию времени по формуле $v^0(\tau) = (-1/2)\Psi^{-1}(\tau) \cdot C'(\tau)X'[\vartheta, \tau]l^0[t_*]$. На втором шаге $[\tau_*, \vartheta]$ в зависимости от реализовавшейся позиции $\{\tau_*, x[\tau_*]\}$ назначается одно из двух управлений $v^0(\tau) = (-1/2)\Psi^{-1}(\tau)C'(\tau)X'[\vartheta, \tau] \cdot (l^0[t_*] + a^+e)$ или $v^0(\tau) = (-1/2)\Psi^{-1}(\tau) \cdot X[\vartheta, \tau] \cdot (l^0[t_*] - a^-e)$, где числа a^+ и a^- определены условиями (3.10), а единичный собственный вектор e — условием (3.9).

Оба указанных способа построения оптимального гарантирующего управления v^0 основаны на принципе обратной связи. Однако, результаты, связанные с программной стохастической конструкцией (2.3), (2.4), дают основание для попытки строить и в реальном объекте (1.1) оптимальное гарантирующее управление v^0 для данной исходной позиции $\{t_*, x_*\}$, как максимизирующую в (2.4) (или близкую к ней по результату) неупреждающую программу $v^0(t_*[\cdot], \vartheta)$. Заметим при этом, что в основе построения управления $v^0[t, \omega]$ на основе программы $v^0(t_*[\cdot], \vartheta)$ лежат лишь сведения о реализации винеровского процесса $z(t_*[\cdot], t, \omega)$, которые можно черпать из некоторого источника случайных процессов, никак не связанного с эволюцией объекта (1.1).

Из теоремы 2.1 вытекает, что каким бы способом, не зависящим от будущего реализации винеровского процесса, ни формировалось управление u , неупреждающая программа $v^0(t_*[\cdot], \vartheta)$ (или близкая к ней по результату) гарантирует выполнение неравенства

$$M\{\gamma(x[\cdot, \omega], v^0[\cdot, \omega], u[\cdot, \omega])\} \geq \rho^0(t_*, x_*) - \zeta. \quad (4.4)$$

Такая статистическая оценка получится, если, например, $u[\cdot, \omega]$ суть реализации любой позиционной стратегии $u(\cdot)$, или $u[\cdot, \omega]$ — реализации любой неупреждающей программы $u(t_*[\cdot], \vartheta)$ на том же самом процессе $z[t, \omega]$, или $u[\cdot, \omega^*]$ — реализации некоторой стохастической программы по независимым от $z[t, \omega]$ случайным событиям и др.

Если же управления $u[\cdot, \omega] = u^0[\cdot, \omega]$ порождены оптимальной стратегией $u^0(\cdot)$, то наряду с (4.4) с вероятностью 1 при $v[\cdot] = v^0[\cdot, \omega]$ будет выполняться неравенство (4.2). Отсюда следует, что в таком случае для любых сколь угодно малых чисел $\eta > 0$ и $\alpha > 0$ можно для данной позиции $\{t_*, x_*\}$ указать стохастическую неупреждающую программу $v^0(t_*[\cdot], \vartheta)$, которая гарантирует выполнение неравенства

$$P(\gamma(x[\cdot, \omega], u^0[\cdot, \omega], v^0[\cdot, \omega]) \geq \rho^0(t_*, x_*) - \eta) \geq 1 - \alpha.$$

Таким образом получается, что если управление u будет формироваться оптимальным образом, то можно с вероятностью, сколь угодно близкой к единице, гарантировать результат, сколь угодно близкий к $\rho^0(t_*, x_*)$, формируя в объекте управление v не по принципу обратной связи, но только на основе сведений о реализации независимого от объекта случайного процесса $z[t, \omega]$.

5. В качестве иллюстрации рассмотрим два модельных примера. Первый из них носит формальный характер. Пусть система (1.1) и функционал (1.3) имеют вид

$$\begin{aligned} \dot{x} &= 2(u + \sqrt{1 - e^{0,1t} \sin t} v), \quad x[t_*] = x_* \\ \gamma &= \int_{t_*}^{\vartheta} (|u|^2 - |v|^2) dt + |x[\vartheta]|, \end{aligned}$$

где x, u, v — n -мерные векторы, $\vartheta = 5\pi, 0 \leq t_* \leq \vartheta$.

По формуле (3.18) находим, что

$$\Gamma(t_*) = -\frac{1}{1,01} [e^{-0,1t_*} (\cos t_* + 0,1 \sin t_*) + e^{-0,5\pi}] E,$$

где E — единичная матрица.

Из формул (3.7) получаем, что

$$\lambda(\tau) = -\frac{1}{1,01} [e^{-0,1\tau} (\cos \tau + 0,1 \sin \tau) + e^{-0,5\pi}] \quad (5.1)$$

и, следовательно,

$$\lambda(\tau_*[t_*]) = \begin{cases} \lambda(\pi), & 0 \leq t_* \leq \pi \\ \lambda(t_*), & \pi < t_* \leq \tau_1 \\ \lambda(3\pi), & \tau_1 < t_* \leq 3\pi \\ \lambda(t_*), & 3\pi < t_* \leq \tau_2, \end{cases} \quad (5.2)$$

где число $\tau_1 \in (\pi, 2\pi)$ удовлетворяет условию $e^{-0,1\tau_1} (\cos \tau_1 + 0,1 \sin \tau_1) = -e^{-0,3\pi}$, а число $\tau_2 \in (3\pi, 4\pi)$ — условию $e^{-0,1\tau_2} (\cos \tau_2 + 0,1 \sin \tau_2) = -e^{-0,5\pi}$.

При всех $\tau \in (\tau_2, 5\pi)$ квадратичная форма $H(\tau, a) = \langle \Gamma(\tau) a \cdot a \rangle = \lambda(\tau) |a|^2$ в соответствии с (5.1) является знакоотрицательной и, значит, величина $\rho^0(t_*, x_*)$ вычисляется в этом случае по формуле (3.17), где положено $\lambda(\tau_*[t_*]) = 0$. Таким образом, в рассматриваемом примере имеем

$$\begin{aligned} \rho^0(t_*, x_*) &= \max_{|l| \leq 1} [\langle l \cdot x_* \rangle + (\lambda(t_*) - \\ &- \lambda(\tau_*[t_*])) \cdot |l|^2] + \lambda(\tau_*[t_*]), \end{aligned} \quad (5.3)$$

где $\lambda(\tau_*[t_*])$ при $0 \leq t_* \leq \tau_2$ определяется соотношениями (5.2), а при $t_* \in (\tau_2, 5\pi)$ полагается $\lambda(\tau_*[t_*]) = 0$. Решая задачу (5.3), находим, что

$$\rho^0(t_*, x_*) = \begin{cases} \frac{|x_*|^2}{4(\lambda(\pi) - \lambda(t_*))} + \lambda(\pi), & 0 \leq t_* \leq \pi, \\ & |x_*| \leq 2(\lambda(\pi) - \lambda(t_*)) \\ \frac{|x_*|^2}{4(\lambda(3\pi) - \lambda(t_*))} + \lambda(3\pi), & \tau_1 < t_* \leq 3\pi, \\ & |x_*| \leq 2(\lambda(3\pi) - \lambda(t_*)) \\ -\frac{|x_*|^2}{4\lambda(t_*)} & \tau_2 < t_* \leq 5\pi, \\ & |x_*| \leq -2\lambda(t_*) \end{cases}$$

и $\rho^0(t_*, x_*) = |x_*| + \lambda(t_*)$ для всех остальных позиций.

Содержание второго примера состоит в следующем. Пусть от горизонтальной оси x_1 под действием управляющего усилия u , вырабатываемого электродвигателем, перемещается механический объект массы $m = 1$. Требуется за время $T_* = \vartheta - t_*$ перевести объект из некоторой произвольной позиции $\{t_* \geq 0, x_1[t_*] = x_{1*}, \dot{x}_1[t_*] = x_{2*}\}$ в состояние $x_1[\vartheta] = \dot{x}_1[\vartheta] = 0$.

При этом в процессе управления на выполнение задания тратится энергия стоимостью

$$W_u = \int_{t_*}^{\vartheta} \frac{1}{\vartheta - t} u^2[t] dt, \quad (5.4)$$

а за неточное выполнение задания взимается штраф $(x_1^2[\vartheta] + \dot{x}_1^2[\vartheta])^{1/2}$. Предположим, что на объект действует еще неопределенная помеха, позволяющая, однако, производить отбор мощности

$$W_v = \frac{1}{2} \int_{t_*}^{\vartheta} v^2[t] dt, \quad (5.5)$$

которая поступает в электродвигатель, вырабатывающий управляющее усилие u . Тогда естественно оценить стоимость задания величиной

$$\gamma = W_u - W_v + (x_1^2[\vartheta] + \dot{x}_1^2[\vartheta])^{1/2}. \quad (5.6)$$

Записывая уравнения движения объекта и трактуя величину γ (5.6) как общий расход средств на выполнение задания, приходим к изученной уже задаче управления с оптимальным гарантированным результатом для случая, когда система (1.1) и функционал (1.3) имеют вид

$$\begin{aligned} \dot{x}_1 &= \dot{x}_2, & \dot{x}_2 &= u + v, \\ \gamma &= \int_{t_*}^{\vartheta} \left[\frac{1}{\vartheta - t} u^2 - \frac{1}{2} v^2 \right] dt + |x[\vartheta]|. \end{aligned} \quad (5.7)$$

Пусть для определенности $\vartheta = 3$. Для вычисления оптимального гарантированного результата $\rho^0(t_*, x_*)$ (здесь — минимального гарантированного расхода средств) надлежит снова воспользоваться формулой (3.17), в которой теперь

$$\lambda(\tau_*, [t_*]) = \begin{cases} \lambda(1), & 0 \leq t_* \leq 1, \\ \lambda(t_*), & 1 < t_* \leq \vartheta = 3, \end{cases} \quad (5.8)$$

где

$$\lambda(\tau) = (-1/8)[\gamma_{11}(T) + \gamma_{22}(T) - \{\gamma_{11}(T) - \gamma_{22}(T)\}^2 + 4\gamma_{12}^2(T)]^{1/2},$$

$$T = \vartheta - \tau, \quad \gamma_{11}(T) = T^4/4 - 2T^3/3, \quad \gamma_{12}(T) = T^3/3 - T^2, \quad \gamma_{22}(T) = T^2/2 - 2T.$$

Квадратичная форма $H(t_*, l)$ для данного примера в соответствии с (3.18) имеет вид

$$H(t_*, l) = -\frac{1}{4}[\gamma_{11}(T_*)l_1^2 + 2\gamma_{12}(T_*)l_1l_2 + \gamma_{22}(T_*)l_2^2]. \quad (5.9)$$

Задача (3.17) при $\lambda(\tau_*, [t_*])$ (5.8) и $H(t_*, l)$ (5.9) есть несложная задача квадратичного программирования, которая может быть решена с привлечением скромных вычислительных средств, а, например, при $t_* = 0, x_{2*} = 0$, т.е., когда объект начинает движение из состояния покоя, имеет место простая формула

$$\rho^0(t_*, x_*) = \begin{cases} \frac{4x_{1*}^2}{(9 + 16\lambda(1))} + \lambda(1), & |x_{1*}| \leq \frac{1}{8}(9 + 16\lambda(1)), \\ |x_{1*}| - 9/16, & |x_{1*}| > \frac{1}{8}(9 + 16\lambda(1)), \end{cases}$$

где $\lambda(1) = 0,76$.

Литература

1. Красовский Н. Н., Третьяков В. Е. Стохастический программный синтез для позиционной дифференциальной игры. ДАН СССР, 1981, т. 259, № 1, с. 24–27.
2. Красовский А. Н., Красовский Н. Н., Третьяков В. Е. Стохастический программный синтез для детерминированной позиционной дифференциальной игры. Прикл. мат. и мех. 1981, т. 45, вып. 4, с. 581–588.
3. Понтрягин Л. С. К теории дифференциальных игр. Успехи мат. наук, 1966, т. 21, вып. 4, с. 219–274.

4. Красовский Н. Н., Субботин А. И. Позиционные дифференциальные игры. М., Наука, 1974, с. 455.
 5. Литцер Р. Ш., Ширяев А. Н. Статистика случайных процессов. М., Наука, 1974, с. 696.

A stochastic program synthesis of a guaranteeing control

N. N. KRASOVSKII, V. E. TRET'YAKOV

(Sverdlovsk)

In the paper the optimal guaranteeing positional control is constructed by a method of a stochastic program synthesis in the case when an equation of object motion and a quality index of a process of control have the forms:

$$\dot{x} = A(t)x + B(t)u + C(t)v, \quad u \in \mathcal{P}, \quad v \in \mathcal{Q}$$

$$\gamma(x[\cdot], u[\cdot], v[\cdot]) = \int_{t_0}^{\vartheta} R(t, u[t], v[t])dt + |x[\vartheta]|$$

$$R(t, u, v) = \langle \Phi(t)u \cdot u \rangle + \langle \Psi(t)v \cdot v \rangle.$$

Here $t_0 \leq t_* \leq \vartheta$, $t_* \leq t \leq \vartheta$, t_0 and ϑ are fixed, x is an n -dimensional phase vector, u is an p -dimensional vector of control, v is an q -dimensional vector of an undetermined noise; $A(t)$, $B(t)$, $C(t)$, $\Phi(t)$, $\Psi(t)$ are matrix-valued functions, \mathcal{P} and \mathcal{Q} are some compacts; $|x|$ is the Euclidean norm, $\langle \cdot \cdot \rangle$ is the scalar product. The symmetric matrices $\Phi(t)$ and $\Psi(t)$ satisfy some conditions being fulfilled, in particular, if for $t_0 \leq t < \vartheta$ the matrix $\Phi(t)$ is positive definite, $\Psi(t)$ is negative definite and the absolute values of determinations $|\Phi(t)|$ and $|\Psi(t)|$ are not less than a number N which is sufficiently large in comparison with the dimensions of \mathcal{P} and \mathcal{Q} .

The optimal guaranteeing control is constructed from a value of guaranteed result $\rho^0(t, x)$ by the extremal method as in [2]. According to the idea of the stochastic synthesis in [1] $\rho^0(t, x)$ is a stochastic program maximin of the mean value of the quality index of process over non-anticipatory programs.

The standard process of Brownian motion is chosen as a basic probabilistic element. By passing to the dual problem the search of X is reduced to a problem of optimal program control of a diffusive stochastic system, which is an analogy of the conjugate system from the maximum principle of L. S. Pontryagin.

The solution of the latter problem leads to the result

$$\rho^0(t_*, x_*) = \max_{|l| \leq 1} \{ \langle l \cdot X[\vartheta, t_*]x_* \rangle + H(t_*, l) - \lambda(\tau_*[t_*])|l|^2 \} + \lambda(\tau_*[t_*]),$$

$$H(t_*, l) = \langle \Gamma(t_*)l \cdot l \rangle,$$

$$\Gamma(t_*) = -\frac{1}{4} \int_{t_0}^{\vartheta} X[\vartheta, \tau](B(\tau)\Phi^{-1}(\tau)B'(\tau) + C(\tau)\Psi^{-1}(\tau)C'(\tau))X'[\vartheta, \tau]d\tau,$$

where $X[t, \tau]$ is a fundamental matrix of solutions of the equation $\dot{x} = A(t)x$ and $\lambda(\tau_*, [t_*])$ is found from the conditions

$$\lambda(\tau_*, [t_*]) = \max_{t_* \leq \tau \leq \theta} \lambda(\tau), \quad \lambda(\tau) = \max_{|a| \leq 1} H(\tau, a).$$

If the quadratic form $H(\tau, a)$ for $t_* \leq \tau < \theta$ is of negative sign, there is a regular case and then we set $\lambda(\tau_*, [t_*]) = 0$.

N. N. Krasovskii
Institute of Mathematics and Mechanics of
Ural Scientific Centre of Academy of Sciences
USSR, 620066, Sverdlovsk, K-66
S. Kovalevskaya Street, 16

ON EQUATIONS OF ELLIPSOIDS APPROXIMATING REACHABLE SETS

F. L. CHERNOUSKO

(Moscow)

(Received January 20, 1982)

The paper is devoted to the analysis of nonlinear differential equations which describe evolution of optimal ellipsoids approximating reachable sets of controlled systems. Two different simplified forms of these equations are presented. Asymptotic behaviour of ellipsoids near the initial point and at infinity is studied. Some numerical examples are given.

1. Ellipsoidal bounds for reachable sets

We consider a controlled system described by differential inclusion and initial condition

$$\dot{x} \in X(x, t), \quad x(s) \in M, \quad t \geq s. \quad (1.1)$$

Here t is time; x is an n -vector of state variables; X is a set depending on x, t ; M is a given initial set. The reachable set $D(t, s, M)$ for system (1.1) for $t \geq s$ is a set of all vectors $x(t)$ which are values of all functions $x(\tau)$ satisfying (1.1) for $\tau \in [s, t]$. Reachable sets are important in different problems of control theory [1–5].

This paper follows the approach [6–9] where two-sided ellipsoidal bounds for reachable sets were obtained which are optimal in the sense of volume of sets. At first we summarize the principal results of this approach which are used below.

We denote with $E(a, Q)$ an ellipsoid in n -space defined by the inequality

$$E(a, Q) = \{x: (x - a)^T Q^{-1} (x - a) \leq 1\}. \quad (1.2)$$

Here, a is an n -vector of the centre of an ellipsoid, Q is a symmetrical positive-definite $n \times n$ -matrix. We assume that the following two-sided ellipsoidal bounds are true for the sets X, M in (1.1)

$$\begin{aligned} E(A^-(t)x + f^-(t), G^-(t)) \subset X(x, t) \subset E(A^+(t)x + f^+(t), G^+(t)) \\ E(a_0^-, Q_0^-) \subset M \subset E(a_0^+, Q_0^+), \quad t \geq s. \end{aligned} \quad (1.3)$$

Here A^\pm, G^\pm are given $n \times n$ -matrices depending on t , f^\pm are given n -vector functions, a_0^\pm are given n -vectors, Q_0^\pm are given $n \times n$ -matrices. The matrices G^\pm, Q_0^\pm are

symmetrical and positive-definite. We introduce two systems corresponding to estimates (1.3)

$$\begin{aligned} \dot{x} &\in E(A^-(t)x + f^-(t), G^-(t)), & x(s) &\in E(a_0^-, Q_0^-), \\ \dot{x} &\in E(A^+(t)x + f^+(t), G^+(t)), & x(s) &\in E(a_0^+, Q_0^+). \end{aligned} \quad (1.4)$$

Systems (1.4) are equivalent to the linear controlled systems

$$\dot{x} = A^\pm(t)x + f^\pm(t) + u, \quad u \in E(0, G^\pm(t)), \quad x(s) \in E(a_0^\pm, Q_0^\pm) \quad (1.5)$$

with ellipsoidal bounds on control u .

We denote by $D^\pm(t, s, M)$ reachable sets for systems (1.4). It follows from (1.3)

$$D^-(t, s, E(a_0^-, Q_0^-)) \subset D(t, s, M) \subset D^+(t, s, E(a_0^+, Q_0^+)). \quad (1.6)$$

Here, the sets D^\pm are not ellipsoids in the general case. We introduce ellipsoidal approximations $E(a^-(t), Q^-(t))$, $E(a^+(t), Q^+(t))$ of these sets satisfying the following conditions:

1) initial conditions

$$a^-(s) = a_0^-, \quad Q^-(s) = Q_0^-, \quad a^+(s) = a_0^+, \quad Q^+(s) = Q_0^+, \quad (1.7)$$

2) inclusion conditions for all $\tau \in [s, t]$

$$\begin{aligned} E(a^-(t), Q^-(t)) &\subset D^-(t, \tau, E(a^-(\tau), Q^-(\tau))), \\ E(a^+(t), Q^+(t)) &\supset D^+(t, \tau, E(a^+(\tau), Q^+(\tau))). \end{aligned} \quad (1.8)$$

3) optimality conditions for the increment of the state volume

$$\dot{v}^- \rightarrow \max, \quad \dot{v}^+ \rightarrow \min. \quad (1.9)$$

Conditions (1.9) mean that the volumes v^-, v^+ of ellipsoids $E(a^-(t), Q^-(t))$, $E(a^+(t), Q^+(t))$ increase with maximal (for v^-) or minimal (for v^+) velocity possible for ellipsoids satisfying (1.7), (1.8). It was shown [6–10] that ellipsoids defined by (1.7)–(1.9) are unique and their parameters a^\pm, Q^\pm satisfy the following equations and initial conditions

$$\dot{a} = Aa + f, \quad a(s) = a_0, \quad (1.10)$$

$$\dot{Q}^- = AQ^- + Q^-A^T + 2R^{-1}(RQ^-R^T)^{1/2}(RGR^T)^{1/2}(R^{-1})^T, \quad Q^-(s) = Q_0, \quad (1.11)$$

$$\begin{aligned} \dot{Q}^+ &= AQ^+ + Q^+A^T + hQ^+ + h^{-1}G, \\ h &= \{n^{-1} \text{Tr}[(Q^+)^{-1}G]\}^{1/2}, \quad Q^+(s) = Q_0. \end{aligned} \quad (1.12)$$

Here the dependence of A, f, G, R, h on t and indices $-$ in (1.11), $+$ in (1.12) after A, G, Q_0 are not indicated. The linear equation (1.10) is valid for both a^-, a^+ if we put corresponding indices $-, +$ after A, f, a_0 . The matrix $R(t)$ in (1.11) is such a matrix that both matrices RQ^-R^T, RGR^T are diagonal for $t \geq s$.

After the initial problems (1.10)–(1.12) for vectors $a^\pm(t)$ and symmetrical positive-definite matrices $Q^\pm(t)$ are solved, we have the estimates following from (1.6), (1.8)

$$E(a^-(t), Q^-(t)) \subset D(t, s, M) \subset E(a^+(t), Q^+(t)). \quad (1.13)$$

If we have only internal or external estimate (1.3), then we can obtain only internal or external estimate (1.13), respectively. If system (1.1) is linear and similar to (1.5)

$$\dot{x} = A(t)x + f(t) + u, \quad u \in E(0, G(t)), \quad x(s) \in E(a_0, Q_0) \quad (1.14)$$

then both systems (1.4) coincide with (1.14). In this case we need not put indices $-, +$ after A, G, f, a_0, Q_0 in (1.10)–(1.12), and $a^-(t) = a^+(t) = a(t)$.

Two-sided estimates (1.13) can be useful in different problems of control and state estimation [6–9]. We consider below some properties of nonlinear equations (1.11), (1.12).

2. Transformations of equations

We substitute ($V(t)$ is a non-degenerate $n \times n$ matrix)

$$Q^\pm = V(t)Z^\pm V^T(t) \quad (2.1)$$

into systems (1.11), (1.12). Then for new variables $Z^\pm(t)$ we obtain the same equations as (1.11), (1.12) with matrices $A(t), G(t)$ replaced by

$$A_1(t) = V^{-1}(AV - \dot{V}), \quad G_1(t) = V^{-1}G(V^{-1})^T. \quad (2.2)$$

An arbitrary matrix $V(t)$ can be chosen in such a way that the equations for Z^\pm are simpler than (1.11), (1.12). We consider two possibilities.

1) In order to obtain $A_1(t) \equiv 0$ we take

$$\dot{V} = A(t)V, \quad t \geq s, \quad V(s) = I, \quad (2.3)$$

where I is an identity $n \times n$ -matrix. Therefore, $V(t)$ is the fundamental matrix of system (1.14). Then the functions Z^\pm satisfy the equations

$$\begin{aligned} \dot{Z}^- &= 2R_1^{-1}(R_1 Z^- R_1^T)^{1/2}(R_1 G_1 R_1^T)^{1/2}(R_1^{-1})^T, \\ \dot{Z}^+ &= hZ^+ + h^{-1}G_1, \quad h = \{n^{-1} \text{Tr}[(Z^+)^{-1}G_1]\}^{1/2}, \\ Z^-(s) &= Z^+(s) = Q_0. \end{aligned} \quad (2.4)$$

The matrix R_1 is such that both matrices $R_1 Z^- R_1^T, R_1 G_1 R_1^T$ are diagonal for $t \geq s$.

2) In order to obtain $G_1 \equiv I$ we take

$$V(t) = [G(t)]^{1/2}. \quad (2.5)$$

Besides we require that the matrix R_1 what arises in the equation for Z^- is orthogonal: $R_1 R_1^T = I$. Therefore, R_1 is an orthogonal matrix which transforms the symmetrical matrix Z^- into diagonal matrix: $R_1 Z^- R_1^T$ is diagonal. Under these conditions the system (1.11) takes the form

$$\dot{Z}^- = A_1 Z^- + Z^- A_1^T + 2R_1^{-1} (R_1 Z^- R_1^T)^{1/2} R_1. \quad (2.6)$$

It is known [11] that $g(RZR^{-1}) = Rg(Z)R^{-1}$ for arbitrary function g and matrices R, Z , where R is non-degenerate. Therefore we obtain from (2.6), (1.12), (2.2) in the case (2.5)

$$\begin{aligned} \dot{Z}^- &= A_1 Z^- + Z^- A_1^T + 2(Z^-)^{1/2}, \\ \dot{Z}^+ &= A_1 Z^+ + Z^+ A_1^T + hZ^+ + h^{-1}I, \\ h &= \{n^{-1} \text{Tr}[(Z^+)^{-1}]\}^{1/2}, \quad A_1(t) = G^{-1/2} \left[AG^{1/2} - \frac{d}{dt}(G^{1/2}) \right], \\ Z^-(s) = Z^+(s) &= Z_0 = G^{-1/2}(s)Q_0 G^{-1/2}(s). \end{aligned} \quad (2.7)$$

The obtained results are summarized in the following theorem.

Theorem 1. Transformations (2.1), (2.3) and (2.1), (2.5) reduce systems (1.11), (1.12) to the forms (2.4) and (2.7) respectively.

Theorem 1 makes it possible to take either $A \equiv 0$ or $G \equiv I$ in (1.11), (1.12) without loss of generality, i.e. to consider simplified but equivalent systems (2.4) or (2.7) instead of (1.11), (1.12).

3. Coincidence of ellipsoids

We shall find conditions under which both ellipsoids in (1.13) coincide and present the reachable set. It follows from (1.3) that it is necessary for such coincidence that the system (1.1) has the form (1.14), i.e.

$$\begin{aligned} X(x, t) &= E(A(t)x + f(t), G(t)), \quad t \geq s, \\ M &= E(a_0, Q_0). \end{aligned} \quad (3.1)$$

Under conditions (3.1) we obtain from (1.10) that $a^-(t) \equiv a^+(t)$ and the initial conditions in (1.11), (1.12) coincide. In order that $Q^-(t) \equiv Q^+(t)$ it is necessary and sufficient that the right-hand parts of systems (1.11), (1.12) are equal. This condition is

equivalent to equality of right-hand parts of systems (2.7) for $Z^- = Z^+ = Z$. We obtain from this condition that $2Z^{1/2} = hZ + h^{-1}I$, or

$$Z(t) = \lambda^2(t)I, \quad Q(t) = \lambda^2(t)G(t), \quad \lambda(t) \geq 0, \quad t \geq s \quad (3.2)$$

where λ is a scalar function (see also (2.1), (2.5)). The matrix Z from (3.2) must satisfy equations and initial conditions (2.7). Inserting (3.2) into (2.7) we find that equations (2.7) are satisfied if

$$\begin{aligned} A_1 + A_1^T &= \mu(t)I, \quad Q_0 = \lambda_0^2 G(s), \\ \mu(t) &= 2\lambda^{-1}(\dot{\lambda} - 1), \quad \lambda_0 = \lambda(s). \end{aligned} \quad (3.3)$$

Substituting A_1 from (2.7) into (3.3), we obtain

$$AG + GA^T - \dot{G} = \mu(t)G, \quad t \geq s. \quad (3.4)$$

The last two equations (3.3) constitute a linear initial problem for λ . By integrating it we find

$$\lambda(t) = \int_s^t \exp \left[\frac{1}{2} \int_{t_2}^t \mu(t_1) dt_1 \right] dt_2 + \lambda_0 \exp \left[\frac{1}{2} \int_s^t \mu(t_1) dt_1 \right]. \quad (3.5)$$

As a result we have the following theorem.

Theorem 2. The internal and external ellipsoids (1.13) coincide ($a^- \equiv a^+$, $Q^- \equiv Q^+$) for all $t \geq s$, if and only if conditions (3.1) (for all x), (3.4) (for some scalar function $\mu(t)$) and $Q_0 = \lambda_0^2 G(s)$ (for some constant $\lambda_0 \geq 0$) are satisfied. Under these conditions the common centre $a(t)$ of both ellipsoids (1.13) satisfies (1.10), and their matrix Q is given by (3.2), where $\lambda(t)$ is given by (3.5).

4. Asymptotic expansions near the initial point

The important particular case arises when the initial point in (1.1) is fixed: $x(s) = a_0$. In this case the initial ellipsoids $E(a_0^\pm, Q_0^\pm)$ in (1.3) degenerate into the point a_0 , and the initial conditions (1.11), (1.12) for both matrices Q^\pm are zero: $Q_0 = 0$. The initial conditions (2.7) are also zero

$$Z^-(s) = Z^+(s) = 0. \quad (4.1)$$

Equations (1.11), (1.12), (2.7) have singularities when $Q^\pm \rightarrow 0$, $Z^\pm \rightarrow 0$. We shall find asymptotic solutions of such singular initial problems using equations (2.7) and initial conditions (4.1). Let the following expansion exist for the matrix $A_1(t)$ from (2.7) in the neighbourhood of the initial point

$$A_1(t) = A_{10} + A_{11}\theta + O(\theta^2), \quad \theta = t - s \geq 0. \quad (4.2)$$

Here A_{10}, A_{11} are constant matrices. We shall seek the solutions of equations (2.7) with initial conditions (4.1) as power series

$$Z^\pm(t) = Z_1^\pm \theta + Z_2^\pm \theta^2 + Z_3^\pm \theta^3 + Z_4^\pm \theta^4 + O(\theta^5). \quad (4.3)$$

Here Z_1^\pm, Z_2^\pm, \dots are unknown constant symmetrical matrices. We insert expansions (4.2), (4.3) into equations (2.7) and expand both parts of these equations into power series in θ . Making equal coefficients of these expansions for both parts of equations we find the unknown coefficients (4.3). After straightforward but rather lengthy calculations we obtain

$$\begin{aligned} Z_1^- = Z_1^+ = 0, \quad Z_2^- = Z_2^+ = I, \quad Z_3^- = Z_3^+ = D_0, \quad Z_4^- = \frac{7}{12} D_0^2 + \frac{2}{3} D_1, \\ Z_4^+ = \frac{2}{3} D_0^2 + \frac{2}{3} D_1 + \frac{1}{12} n^{-2} (\text{Tr} D_0)^2 - \frac{1}{6} n^{-1} D_0 \text{Tr} D_0. \end{aligned} \quad (4.4)$$

Here D_0, D_1 are symmetrical matrices

$$D_0 = \frac{1}{2} (A_{10} + A_{10}^T), \quad D_1 = \frac{1}{2} (A_{11} + A_{11}^T). \quad (4.5)$$

Theorem 3. Under condition (4.2) the solutions of equations (2.7) with zero initial conditions (4.1) have asymptotic expansions (4.3) in the neighbourhood of the initial point with coefficients defined by (4.4), (4.5).

The expansions for Z^-, Z^+ coincide up to the terms $O(\theta^3)$ and differ in $O(\theta^4)$. It follows from (4.3), (4.4) that

$$Z^+(t) - Z^-(t) = \frac{1}{12} \theta^4 (D_0 - n^{-1} I \text{Tr} D_0)^2 + O(\theta^5).$$

Therefore $Z^+ - Z^-$ is a non-negative matrix for small θ . It is evident for all θ , because the external ellipsoid contains the internal one.

Using transformation (2.1), (2.5) we obtain from (4.3), (4.4) asymptotic expansions for the solutions of equations (1.11), (1.12) with zero initial conditions. The obtained solutions are useful for starting numerical integration of equations (1.11), (1.12) or (2.7) with zero initial conditions.

5. Asymptotic behaviour at infinity

We consider equations of ellipsoids in the form (2.7) for the constant diagonal matrix A_1

$$A_1(t) = \text{diag} \{ \alpha_1, \dots, \alpha_n \}, \quad \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n. \quad (5.1)$$

Here $\alpha_1, \dots, \alpha_n$ are constants. We assume that initial conditions (2.7) for the matrices Z^\pm are also diagonal. Then equations (2.7) have diagonal solutions

$$Z^\pm(t) = \text{diag} \{ [y_1^\pm(t)]^2, \dots, [y_n^\pm(t)]^2 \}. \quad (5.2)$$

Here $y_i^\pm \geq 0$ are semi-axes of ellipsoids, $i = 1, \dots, n$. Inserting (5.1), (5.2) into (2.7), we obtain

$$\dot{y}_i^- = \alpha_i y_i^- + 1, \quad i = 1, \dots, n, \quad (5.3)$$

$$\dot{y}_i^+ = \alpha_i y_i^+ + \frac{1}{2} [h y_i^+ + (h y_i^+)^{-1}], \quad (5.4)$$

$$h = \left[\frac{1}{n} \sum_{i=1}^n (y_i^+)^{-2} \right]^{1/2} > 0, \quad i = 1, \dots, n.$$

We shall study the asymptotic behaviour at $t \rightarrow \infty$ of positive ($y_i^+ > 0$) solutions of systems (5.3), (5.4). System (5.3) consists of independent linear equations. All its positive solutions are monotone functions of t and

$$\begin{aligned} y_i^- &\rightarrow +\infty && \text{for } t \rightarrow +\infty, \quad (\alpha_i \geq 0), \\ y_i^- &\rightarrow \alpha_i^{-1} && \text{for } t \rightarrow +\infty, \quad (\alpha_i < 0). \end{aligned} \quad (5.5)$$

Let some solutions of the nonlinear system (5.4) have the limits

$$y_i^+ \rightarrow y_i^* > 0, \quad i \in J; \quad y_i^+ \rightarrow +\infty, \quad i \in J', \quad (5.6)$$

when $t \rightarrow +\infty$. Here J is a set of such indices i from $\{1, \dots, n\}$ for which the limit of y_i^+ is finite when $t \rightarrow +\infty$; the set J' includes all the other indices $i \in \{1, \dots, n\}$. One of the two sets J, J' may be empty.

Substituting limits (5.6) into equations (5.4) and solving these equations with respect to y_i^* we obtain

$$y_i^* = [-h^*(2\alpha_i + h^*)]^{-1/2}, \quad i \in J, \quad (5.7)$$

where h^* is the limiting value of h for $t \rightarrow +\infty$. Substituting (5.7) into formula (5.4) for h , we obtain the algebraic equation for h^* . Its solution is

$$h^* = -\frac{2}{n+v} \sum_{j \in J} \alpha_j \geq 0, \quad (5.8)$$

where v is a number of elements in J , $0 \leq v \leq n$. From equations (5.4) for $i \in J'$ and (5.7) it follows that the limits (5.6) exist only if

$$\alpha_i < -h^*/2, \quad i \in J, \quad \alpha_i \geq -h^*/2, \quad i \in J'. \quad (5.9)$$

Therefore (see enumeration (5.1)) we have

$$J = \{1, \dots, v\}, \quad J' = \{v+1, \dots, n\}. \quad (5.10)$$

Now conditions (5.9), (5.8) can be written as

$$(n+v)\alpha_v < \sum_{j=1}^v \alpha_j \leq (n+v)\alpha_{v+1}, \quad v \geq 1. \quad (5.11)$$

For $v=0$, (5.11) must be replaced by $\alpha_1 \geq 0$. We shall prove that there exists only one integer v , $0 \leq v \leq n$, satisfying (5.11). Let the contrary be true, and inequalities (5.11) be satisfied for some v and for $v_1 = v + s$, $s \geq 1$. Then

$$\sum_{j=1}^{v+s} \alpha_j > (n+v+s)\alpha_{v+s}. \quad (5.12)$$

It follows from inequalities (5.1), (5.12) that

$$\begin{aligned} \sum_{j=1}^v \alpha_j &> (n+v)\alpha_{v+s} + s\alpha_{v+s} - (\alpha_{v+1} + \dots + \alpha_{v+s}) \geq \\ &\geq (n+v)\alpha_{v+s} \geq (n+v)\alpha_{v+1}. \end{aligned} \quad (5.13)$$

But (5.13) contradicts the right inequality (5.11). Therefore the integer v defined by conditions (5.11) is unique. We obtain the following theorem.

Theorem 4. All positive diagonal solutions (5.2) of equations (2.7), (5.1) for Z^- have the asymptotic behaviour (5.5) when $t \rightarrow +\infty$. For positive diagonal solutions (5.2) of equations (2.7), (5.1) for Z^+ , there exists a unique asymptotic behaviour of the form (5.6) when $t \rightarrow +\infty$; here y_i^* , h^* , J , J' , v are unique and defined by (5.7), (5.8), (5.10), (5.11).

It is interesting to compare these results with the asymptotic behaviour of reachable sets of the system

$$\dot{x}_i = \alpha_i x_i + u_i, \quad \sum_{i=1}^n u_i^2 \leq 1, \quad i = 1, \dots, n \quad (5.14)$$

corresponding to the case (5.1) (see (1.14) with $A = A_1$, $G = I$). For arbitrary initial conditions, there exists the limit D_∞ of the reachable set when $t \rightarrow +\infty$. D_∞ is a convex

set independent of initial conditions and symmetrical with respect to all axes x_i . The lengths y_i^0 of semi-axes contained by D_∞ are

$$y_i^0 = +\infty, \quad \alpha_i \geq 0; \quad y_i^0 = -\alpha_i^{-1}, \quad \alpha_i < 0. \quad (5.15)$$

They are equal to semi-axes of limiting internal ellipsoid (5.5). Semi-axes of the limiting external ellipsoid (5.6) are greater ($y_i^* > y_i^0$), and sometimes even $y_i^+ \rightarrow +\infty$ when y_i^0 is finite.

As an illustration we consider a two-dimensional example: $n=2, \alpha_1 \leq \alpha_2$. From (5.11) we obtain

$$\begin{aligned} v=0 & \quad \text{for } \alpha_1 \geq 0; \quad v=1 \quad \text{for } \alpha_1 < 0, \quad \alpha_1 \leq 3\alpha_2, \\ v=2 & \quad \text{for } 3\alpha_2 < \alpha_1 \leq \alpha_2. \end{aligned} \quad (5.16)$$

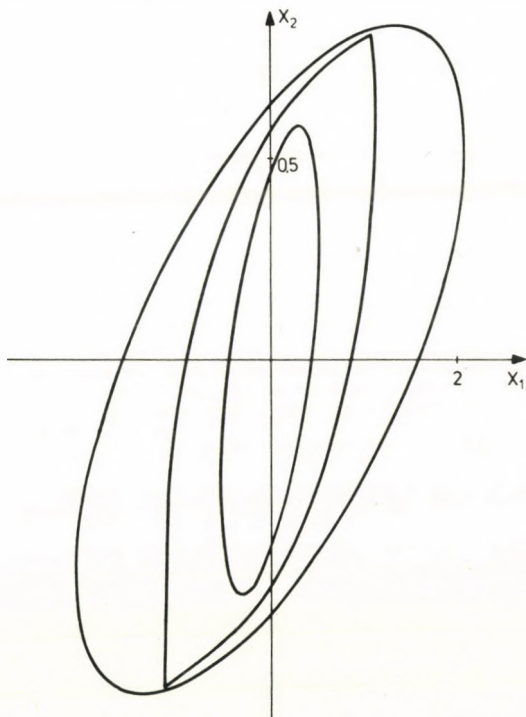


Fig. 1. Numerical example: $k=0, \beta=1, T=2$

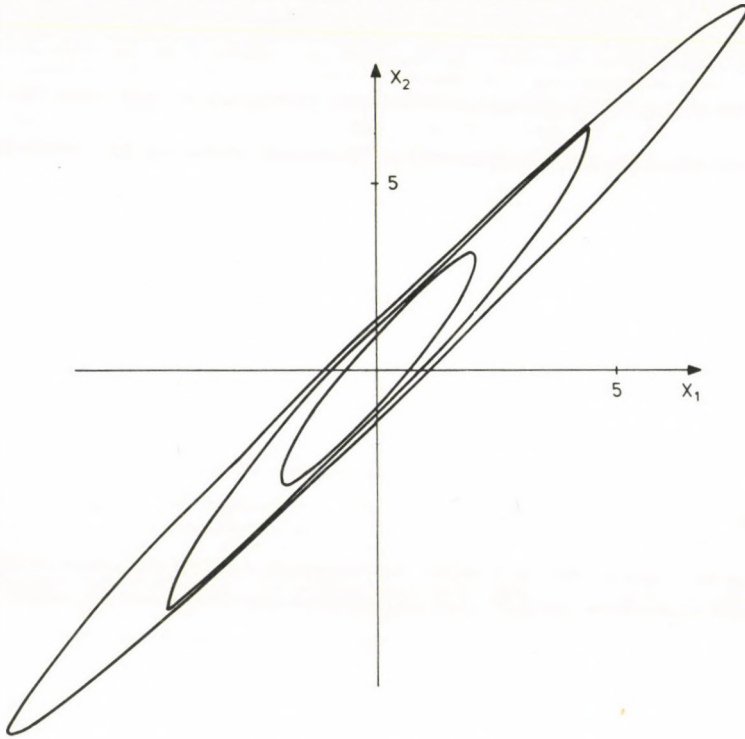


Fig. 2. Numerical example: $k=0$, $\beta=-1$, $T=2$

Limits (5.6), calculated according to (5.7), (5.8), (5.10), (5.16), are

$$\begin{aligned}
 &1) \quad y_1^+ \rightarrow +\infty, \quad y_2^+ \rightarrow +\infty (\alpha_1 \geq 0), \\
 &2) \quad y_1^+ \rightarrow -3\alpha_1^{-1}/2\sqrt{2}, \quad y_2^+ \rightarrow +\infty (\alpha_1 < 0, \quad \alpha_1 \leq 3\alpha_2), \\
 &3) \quad y_1^+ \rightarrow 2[(\alpha_1 + \alpha_2)(3\alpha_1 - \alpha_2)]^{-1/2}, \\
 &\quad y_2^+ \rightarrow 2[(\alpha_1 + \alpha_2)(3\alpha_2 - \alpha_1)]^{-1/2} (3\alpha_2 < \alpha_1 \leq \alpha_2)
 \end{aligned} \tag{5.17}$$

for three respective cases (5.16). The complete analysis of singular points and phase trajectories of system (5.4) showed that all its positive solutions have limits (5.17) when $t \rightarrow +\infty$. If $\alpha_1 < 0$, $\alpha_2 < 0$ and $|\alpha_1/\alpha_2| \geq 3$ then we have $y_2^+ \rightarrow +\infty$ from (5.17). However, in this case y_2^0 is finite, $y_2^- \rightarrow y_2^0 = -\alpha_2^{-1}$, see (5.5), (5.15).

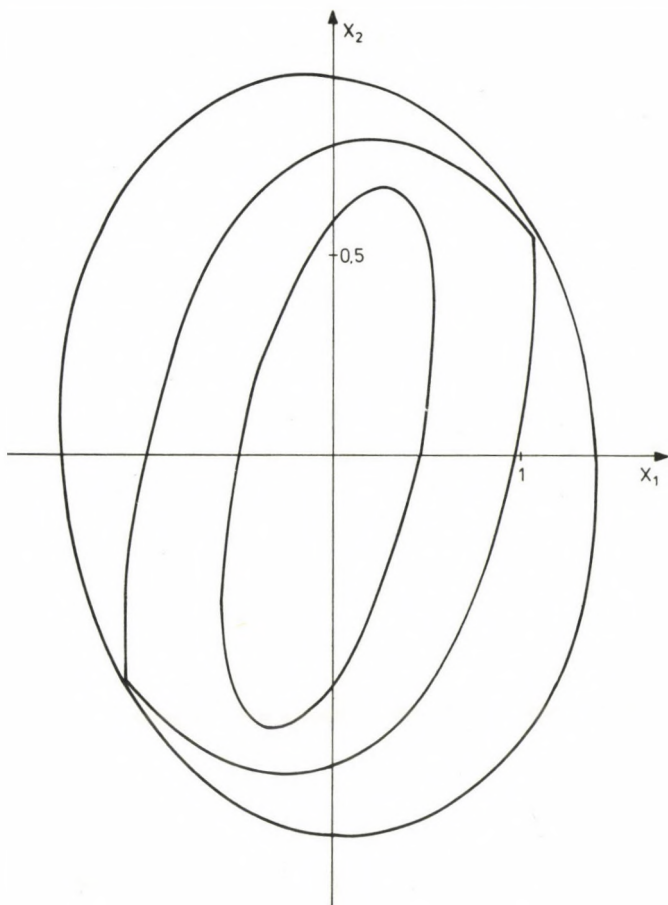


Fig. 3. Numerical example: $k=1$, $\beta=0.5$, $T=2$

6. Numerical examples

We present here some numerical examples of approximating ellipsoids for a two-dimensional system with a scalar control function

$$\begin{aligned} \dot{x}_1 &= x_2, & \dot{x}_2 &= -kx_1 - \beta x_2 + u, & |u| &\leq 1, \\ x_1(0) &= x_2(0) = 0, & 0 &\leq t \leq T. \end{aligned} \quad (6.1)$$

Here β , k , T are constants. System (1.10) for example (6.1) has a zero solution $a(t) \equiv 0$.

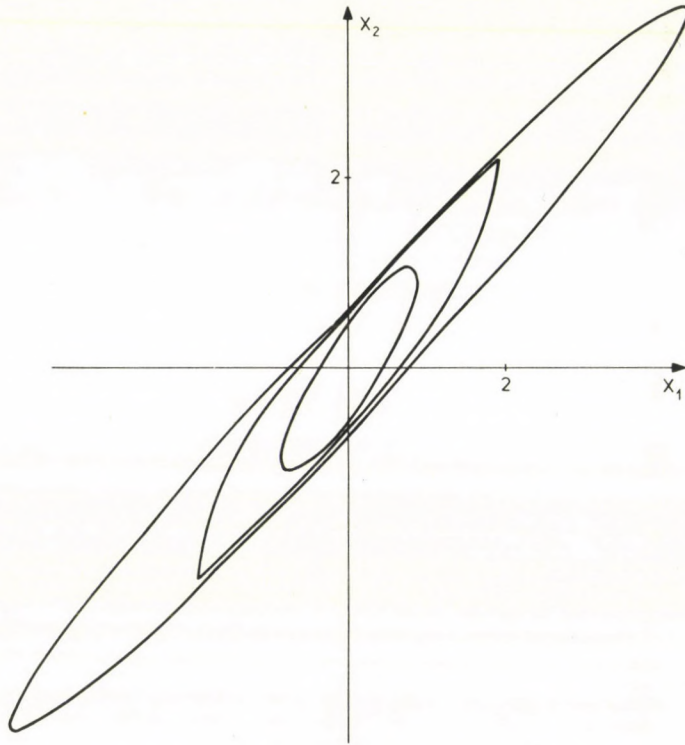


Fig. 4. Numerical example: $k = -1$, $\beta = 0.5$, $T = 2$

Systems (1.11), (1.12) for example (6.1) were integrated numerically. Some results are presented in Figs 1–4. Exact reachable sets shown here between internal and external ellipsoids were also obtained numerically.

7. Conclusions

Optimal two-sided ellipsoidal estimates of reachable sets (1.13) can be used for evaluating of these sets in the presence of control or disturbances. These estimates give rough but simple and guaranteed bounds for reachable sets. Numerical examples and asymptotic analysis show that these bounds are not too far from exact reachable sets, see also [9]. The ellipsoidal estimates of reachable sets can be used also for two-sided estimation in optimal control and differential games, for guaranteed filtering in the presence of measurement errors etc., see [6–9].

In order to obtain ellipsoids approximating reachable sets, it is necessary to integrate certain nonlinear systems of differential equations (1.11), (1.12). It was shown in this paper that these systems can be simplified essentially. The simplified versions (2.4), (2.7) of these systems depend only on one matrix (G_1 or A_1), and equation (2.7) for Z^- does not contain the matrices R, R_1 . Some general properties of equations of ellipsoids including the asymptotic behaviour of their solutions near the initial point and at infinity were established. These properties are useful for qualitative analysis and for numerical integration of equations of ellipsoids.

The author is grateful to A. I. Ovseevich for useful discussion and to B. R. Klepfish for computer programming and calculations of numerical examples.

References

1. Krasovskii, N. N., Theory of Control of Motion. Moscow, Nauka, 1968.
2. Lee, E. B., Markus, L., Foundations of Optimal Control Theory. New York, John Wiley, 1967.
3. Schweppe, F. C., Uncertain Dynamic Systems. Prentice Hall, 1973.
4. Formalskii, A. M., Controllability and stability of systems with bounded resources. Moscow, Nauka, 1974.
5. Kurzhanskii, A. B., Control and observation in conditions of uncertainty. Moscow, Nauka, 1977.
6. Chernousko, F. L., Optimal guaranteed estimates of uncertainties by means of ellipsoids, I, II, III. Izvestiya of the USSR Academy of Sciences, Engineering Cybernetics, 1980, I: No 3, 3-11; II: No. 4, 3-11; III: No. 5, 5-11.
7. Chernousko, F. L., Ellipsoidal estimates for attainable set of controlled system. Applied Mathematics and Mechanics, 45, No. 1, 11-19.
8. Chernousko, F. L., Guaranteed ellipsoidal estimates of uncertainties in control problems. 8th IFAC Triennial World Congress. Preprints, Kyoto, 1981, VI, 203-208.
9. Chernousko, F. L., Ellipsoidal bounds for sets of attainability and uncertainty in control problems. Optimal Control Applications and Methods, 3, No. 2, 1982, 187-202.
10. Ovseevich, A. I., Extremal properties of ellipsoids approximating reachable sets. Problems of Control and Information Theory, 12, No. 1, 1983.
11. Gantmakher, F. R., Theory of matrices. Moscow, Nauka, 1966.

Об уравнениях эллипсоидов, аппроксимирующих области достижимости

Ф. Л. ЧЕРНОУСЬКО

(Москва)

Множества достижимости играют важную роль во многих задачах теории управления [1-5]. В работах [6-9] предложен метод аппроксимации множеств достижимости, состоящий в построении оптимальных (в смысле объема) двусторонних эллипсоидальных оценок этих множеств. Эволюция внешнего и внутреннего аппроксимирующих эллипсоидов описывается специальными нелинейными системами обыкновенных дифференциальных уравнений, выведенными в [6, 7]. Данная работа посвящена исследованию этих дифференциальных уравнений.

В начале работы излагаются основные предположения и приводятся системы уравнений эллипсоидов. Далее указаны преобразования, позволяющие упростить уравнения эллипсоидов, и

приводятся две упрощенные формы уравнений, эквивалентные исходной. Установлены необходимые и достаточные условия того, что внешний и внутренний эллипсоиды совпадают и тем самым точно представляют множество достижимости.

Исследована асимптотика решений уравнений эллипсоидов вблизи начальной точки и на бесконечности. В первом случае построены разложения матриц внутреннего и внешнего эллипсоидов по степеням времени для случая, когда начальная точка фиксирована. Для второго случая при некоторых дополнительных предположениях указаны предельные решения для внутреннего и внешнего эллипсоидов при неограниченном возрастании времени. Выясняется, в частности, при каких условиях оси аппроксимирующих эллипсоидов остаются ограниченными или, наоборот, неограниченно возрастают. Более подробно рассмотрен случай системы второго порядка. Приводятся численные примеры построения эллипсоидов, аппроксимирующих множества достижимости.

Ф. Л. Черноусько

Институт проблем механики АН СССР

СССР, Москва, 117526,

просп. Вернадского, 101

N-PERSON NONLINEAR QUALITATIVE DIFFERENTIAL GAMES WITH INCOMPLETE INFORMATION, SURVEY OF RESULTS

ZS. LIPCSEY

(Budapest)

(Received May 3, 1982)

In this paper the author investigates nonlinear N -person differential games. Theorems 4.2 and 4.3 show that the method of synthesizing strategies as described here makes possible the evasion and capture for a coalition having approximate information about the phase point and a boundary point of the target set nearest to it. It is also shown that the coalition may evade or capture the opponent coalition without having information about its choice of strategy and together with theorems 3.1 and 3.2 theorem 4.2 and 4.3 suggest an alternative theorem.

1. Introduction

The proofs of the theorems and statements of present paper can be found in [7].

In our paper we deal with qualitative N -person cooperative differential games having dynamics such as $\dot{x} = f(x, u_1, \dots, u_N, t)$. With our approach to the problem we follow the view of Pontriagin who (instead of seeking saddle points) constructs strategy for evasion and capture under certain assumptions about the players. His work concerns the linear games. A most representative summary of this work can be found in [5] and an other paper treating the problem on the same basis is [2].

We used as a basis the formalism exposed by A. Blaquiere and P. Caussin published in [1] at the statement of the problem in Section 2. Methods of both works [2] and [5] are based on the use of the closed formulae of the solutions of the linear differential equations. As in our case such formulae cannot be given, we based our investigations on local properties of the dynamic f like Krasovskii in his book [3]. But while Krasovskii works with saddle points of local functionals formed from the distance of the phase point and the target set, we define the concept of superiority in Section 3 with the help of the same functional. A coalition may have superiority at a point even in the case if the functional does not have a saddle point.

We show that if a coalition has superiority at a certain point then a mapping V can be constructed to select the strategy $V(U)$ for the coalition if the opponent coalition chooses the strategy U . Moreover we show that this rule V for choosing the strategy is

valid in some sense in an open neighbourhood too. The pair (V, G) is called a local strategy.

In Section 4, we give a construction for the synthesis of given families of local strategies. The well-known problem of discontinuity surfaces (see e.g. [1]) does not occur. With the help of the synthesised strategy we prove one theorem about evasion and one about capture (Theorems 4.2 and 4.3).

In both cases we work with approximate information $\tilde{z}(t)$ about the phase point $z(t)$ and an estimation $y(\tilde{z}(t))$ of a boundary point of the target set nearest to $\tilde{z}(t)$. The latter is important because the superiority conditions are imposed on the boundary of the terminal set. This idea is used in the work [4] of Lagunov.

Theorems 4.2 and 4.3 together with Theorems 3.1 and 3.2 give almost a pair of alternative theorems. In subsequent papers this question will be dealt with more exactly.

2. Qualitative differential games

In this section we shall give a definition of the N -person qualitative differential games and expose the problem considered in our paper.

Let us denote by $\langle N \rangle$ the set of the N players. The letters $H, U_i, i \in \langle N \rangle$ denote real separable Hilbert spaces. Let the sets $C_i \subset U_i, i \in \langle N \rangle$ be compact. Let E denote the product space $R \times H$ and $G \subset E$ be an open set.

Let us consider the continuous function

$$f: G \times \prod_{i \in \langle N \rangle} C_i \rightarrow H \quad (2.1)$$

satisfying a local Lipschitzian condition for each fixed

$$U \in \prod_{i \in \langle N \rangle} C_i.$$

Then we can form the mapping F as follows:

$$F: G \times \prod_{i \in \langle N \rangle} C_i \rightarrow E$$

$$F: = (1, f). \quad (2.2)$$

If we denote by \bar{R} the extended real line $R: = R \cup \{+\infty\} \cup \{-\infty\}$ then let $[t_0, t_1] \subset \bar{R}, t_0 \in R$ be a closed interval.

Definition 2.1. Let us call strategies for the player $i \in \langle N \rangle$ Borel measurable mappings

$$S_i: = s_i | s_i: [t_0, t(s_i)] \subset [t_0, t_1] \rightarrow C_i \quad (2.3)$$

and let $D(s_i): = [t_0, t(s_i)]$.

Definition 2.2. The product set $S := \prod_{i \in \langle N \rangle} X S_i$ is called the set of situations. The

domain of a situation $s \in S$, $s = \prod_{i \in \langle N \rangle} X S_i$ is given by $D(s) := \bigcap_{i \in \langle N \rangle} (s_i)$.

Let $\theta \subset G$ be a closed set and the set θ is decomposed into the union $\bigcup_{i \in \langle N \rangle} \theta_i$ of not necessarily non-empty closed sets θ_i , $i \in \langle N \rangle$.

Let us suppose that a point $(t_0, x_0) = z_0 \in G \setminus \theta$ exists and let us consider the set of the solutions of the initial value problems

$$\begin{aligned} z(t_0) &= z_0 \\ \dot{z} &= F(z, s), \quad s \in S. \end{aligned} \quad (2.4)$$

Let $D(z, s)$ denote the domain of the maximal solution. Now we define a function $S \rightarrow [t_0, t_1]$ as follows: If z is the solution of (2.4) for a fixed $s \in S$ then

$$t^*(s) := \sup \{ \tau | \tau \in [t_0, t_1], z(t) \notin \theta \text{ if } t \in [t_0, \tau] \subset D(z, s) \}. \quad (2.5)$$

For $t^*(s) < \infty$ let $z(t^*(s))$ denote limit $\lim_{t \uparrow t^*(s)} z(t) \in G$ if it exists in G .

If $\chi_{\theta_i}, \chi_{G \setminus \theta}$ denote the characteristic functions of the sets $\theta_i, G \setminus \theta$, $i \in \langle N \rangle$ then

$$f_i := \begin{cases} \chi_{\theta_i} & \text{if } \theta_i \neq \emptyset \\ \chi_{G \setminus \theta} & \text{if } \theta_i = \emptyset \end{cases} \quad (2.6)$$

for $i \in \langle N \rangle$.

With the help of the functions given in (2.5) and (2.6) we can give a family $\{H_i\}_{i \in \langle N \rangle}$ of functions called payoff functions as follows:

$$H_i: S \rightarrow R$$

$$H_i(s) := \begin{cases} \lim_{t \uparrow t^*(s)} f_i(z(t)) & \text{if } z(t^*(s)) \text{ does not exist in } G \\ \text{or } t^*(s) = \infty. & \\ f_i(z(t^*(s))) & \text{if } z(t^*(s)) \text{ exists} \end{cases} \quad (2.7)$$

with the solution z of (2.4) associated to $s \in S$ for each $s \in S$.

Definition 2.3. The game $\langle \langle N \rangle, (S_1, S_2, \dots, S_N) (H_1 \dots H_N) \rangle$ given above is called qualitative differential game.

If we take $\theta_i = \theta$ for each $i \in \langle N \rangle$ then the studying of the qualitative game defined above can be considered as investigation of the existence of the playable strategies for a qualitative differential game given by the boundary value problem $z(t_0) = z_0, z(t^*(s)) \in \theta$ for the solutions of (2.4) and some additional family of payoff functions.

Definition 2.4. The non-empty subsets of the set of players $\langle N \rangle$ are called coalitions.

If $K \subset \langle N \rangle$ is a coalition then

$$U_K := \prod_{i \in K} C_i$$

$$U_K^\perp := \prod_{i \in \langle N \rangle \setminus K} C_i. \quad (2.8)$$

The payoff function \mathcal{H}_K of the coalition $K \subset \langle N \rangle$ is defined as follows:

$$\mathcal{H}_K := \sum_{i \in K} H_i. \quad (2.9)$$

The aim of the coalition K , $K \subset \langle N \rangle$ is to maximize the payoff function \mathcal{H}_K and to minimize the payoff function by a suitable choice of the strategies.

Roughly saying, the qualitative cooperative game described above is a pursuing-evading game for the coalition K with the target set $\bigcup_{i \in K} \theta_i$ for K and with the target set

$$\bigcup_{i \in K'} \theta_i \text{ for any other coalition } K' \subset \langle N \rangle, K' \cap K = \emptyset.$$

The choice of the strategies for K is based on the following informational assumptions:

- a) We have an approximate information $\tilde{z}(t)$ about the position $z(t)$ for each $t \in [t_0, t_1]$.
- b) We have approximate information about one of the nearest points to $\tilde{z}(t)$ of the target set for each $t \in [t_0, t_1]$.
- c) At the selection of the strategies in the moment t , $t \in [t_0, t_1]$ we deal with both cases as to have or have no information about the choice of the coalition $\langle N \rangle \setminus K$ in the same moment for each $t \in [t_0, t_1]$.

In the next section we define the concepts of superiorities and local strategies.

In Section 4, we show a method for the synthesis of the local strategies and using the global strategy obtained this way we prove our theorems about the capture and the evasion. The notations of this section are valid throughout the whole paper.

3. The superiority and the local strategies

In this section, first we shall define four concepts of superiority. Then using these concepts we shall construct local strategies.

From now on we use the following notations in the paper. If A, B and C are sets and Φ is a mapping $\Phi: A \times B \rightarrow C$ then

$$\Phi(V, W) := \{c \mid c = \Phi(v, w), v \in V \subset A, w \in W \subset B\} \quad (3.1)$$

with non-empty sets $V \subset A, W \subset B$. If any of V, W has only one element then it is denoted by the element itself or that in brackets $\{ \}$.

Moreover let $a \in A$ be a fixed element. Then we use the following abbreviated notation:

$$\Phi^{-1}(a, H) := \{b \in B, \Phi(a, b) \in H \subset C\}. \quad (3.2)$$

Let K, K' be two coalitions satisfying the condition $K \cap K' = \emptyset$. Let $e \in E$ and $x_0 \in G \subset E$ be a unit element and a fixed point respectively. Let us use the following notation:

$$\{e \geq c\} := \{x \mid x \in E, (x, e) \geq c\}, \quad c \in \mathbb{R}. \quad (3.3)$$

Definition 3.1. The coalition K has weak superiority over K' at the point x_0 in the direction e if there exists such $c \in \mathbb{R}, c > 0$ that an element $u(v) \in U_{K'}$ can be selected to each $v \in U_K$, satisfying the relation

$$F(x_0, (U_{K \cup K'}^\perp) \times \{u(v), v\}) \subset \{e \geq c > 0\}. \quad (3.4)$$

Definition 3.2. The coalition K has weak superiority for evasion over K' at the point x_0 in the direction e if such an element $u(v) \in U_K$ exists for each $v \in U_{K'}$ that the relation

$$F(x_0, (U_{K \cup K'}^\perp) \times \{u(v), v\}) \subset \{e \geq 0\} \quad (3.5)$$

holds.

Definition 3.3. The coalition K has strong superiority at the point x_0 in the direction e if there exist $u \in U_K$ and $c \in \mathbb{R}, c > 0$ satisfying the relation

$$F(x_0, U_K^\perp \times \{u\}) \subset \{e \geq c > 0\}. \quad (3.6)$$

Definition 3.4. The coalition K has strong superiority for evasion at the point x_0 in the direction e if an $u \in U_K$ exists satisfying the relation

$$F(x_0, U_K^\perp \times \{u\}) \subset \{e \geq 0\}. \quad (3.7)$$

First we state two important relations between two pairs of the superiority concepts defined above. Then we shall show an equivalent formulation of these concepts.

Theorem 3.1. The coalition K has strong superiority at the point x_0 in the direction e if and only if the coalition $\langle N \rangle \setminus K$ does not have weak superiority over K for evasion in that point in the direction $-e$.

Theorem 3.2. The coalition K has weak superiority over $\langle N \rangle \setminus K$ at the point x_0 in the direction e if and only if the coalition $\langle N \rangle \setminus K$ does not have strong superiority for evasion in the direction $-e$.

Now we give a characterization of the weak and strong superiorities.

Let us denote by $K_{\varepsilon, e}(0)$ the convex cone

$$K_{\varepsilon, e}(0) := \bigcup_{\lambda > 0} \lambda \{G_\varepsilon(0) + e\} \quad (3.8)$$

for $e \in E, \|e\| = 1$ and $\varepsilon \in R, 1 > \varepsilon > 0$.

Theorem 3.3. The coalition K has superiority over K' (or strong superiority respectively) at the point x_0 in the direction e if and only if

$$F(x_0, U_{K \cup K'}^\perp \times (u(v), v)) \subset K_{\varepsilon, e}(0), \quad v \in U_{K'} \quad (3.9)$$

$$(F(x_0, U_K^\perp \times \{u\}) \subset K_{\varepsilon, e}(0) \text{ respectively}) \quad (3.10)$$

for suitable $1 > \varepsilon > 0$ with $u(v)$ associated to v by definition 3.1 (u associated to K by Definition 3.3 respectively).

Now we may turn our attention to the definition of the local strategies.

The local strategies give the way for the coalitions to use the superiority they have at a certain point $x \in G$.

The construction of the rule to select strategies is based on the following theorem:

Let $K, K', K \cap K' = \emptyset$ be two coalitions.

Theorem 3.4. If such an element $v(u) \in U_{K'}$ can be selected to each element $u \in U_{K'}$ at the point $x_0 \in G$ for the $e \in E, \|e\| = 1$ that the relation

$$F(x_0, U_{K \cup K'}^\perp, (v(u), u)) \subset \{e \geq c\} \quad (3.11)$$

holds for any $u \in U_{K'}$ and a fixed $c \in R$ then a Borel measurable mapping $V_e: U_{K'} \rightarrow U_K$ can be given satisfying the condition

$$F(x_0, U_{K \cup K'}^\perp, (V_e(u), u)) \subset \{e \geq c\} \quad (3.12)$$

for each $u \in U_{K'}$.

This statement is a simple consequence of Kuratovskii's and Ryll's-Nardzewskii's theorem on selectors (see in [6]). Let us consider now the properties of the rule of choice expressed by the mapping V_e .

Theorem 3.5. If the Borel measurable mapping $V_e: U_{K'} \rightarrow U_K$ satisfies condition (3.12) at $x_0 \in G$ for $c \in R$ and arbitrary $u \in U_{K'}$, then an open neighbourhood $G_{\delta(\varepsilon)}(x_0)$ can be found with

$$F(G_{\delta(\varepsilon)}(x_0), U_{K \cup K'}^\perp, V_e(u), u) \subset \{e \geq c - \varepsilon\}. \quad (3.13)$$

Now we are in the position to define the concept of the local strategies.

Summarizing Theorems 3.4 and 3.5, we obtain a Borel measurable mapping $V_e: U_{K'} \rightarrow U_K$ and open set $G \supset \Omega^e (= G_{\delta(e)}(x_0))$ containing the point x_0 for arbitrary $\varepsilon > 0$, the pair (V_e, Ω^e) satisfying condition (3.13).

Definition 3.5. The pair (V_e, Ω^e) is called local strategy of the coalition K .

Remark 3.2. Let us suppose that the coalition K has strong or weak superiority over K' at the point x_0 in the direction $e \in E, \|e\| = 1$ (in the case of the strong superiority $K': = \langle N \rangle \setminus K$).

Then by virtue of Definitions 3.1 and 3.3 condition (3.11) is satisfied with $c > 0$, therefore using Theorems 3.4 and 3.5 a local strategy $(V_e, \Omega^{c/2})$ can be constructed.

Theorem 3.6. If the coalition K has weak superiority over K' or strong superiority at $(t_0, z_0) = x_0 \in G$ in the direction e then in the cases of the weak superiority and the strong superiority the solutions of the initial value problems

$$\begin{aligned} \dot{x}(t) &= F(x(t), v(t), V_e(u(t)), u(t)), t \in [t_0, t_1] \\ x(t_0) &= x_0 \end{aligned} \quad (3.14)$$

with $u(t) := \prod_{i \in K'} s_i(t), s_i \in S_i, i \in K', v(t) := \prod_{i \in \langle N \rangle \setminus K \cup K'} s_i(t), s_i \in S_i, i \in \langle N \rangle \setminus K \cup K'$, and

$$\begin{aligned} \dot{x}(t) &= F(x(t), V_e(u(t)), u(t)), t \in [t_0, t_1] \\ x(t_0) &= x_0 \end{aligned} \quad (3.15)$$

with $u(t) := \prod_{i \in U_K^\perp} s_i(t), s_i \in S_i, i \in \langle N \rangle \setminus K$ satisfy the condition

$$x(t) \in x_0 + K_{\varepsilon, e}(0), t \in (t_0, t_0 + \delta) \quad (3.16)$$

with suitable $\delta > 0$ and $1 > \varepsilon > 0$, respectively. (The notation $K_{\varepsilon, e}(0)$ is given in (3.8)).

In the next section we give a way to compose a "global strategy" from a given family of the local strategies and we prove our basic theorems about capture and evasion.

4. Synthesis of the local strategies for evasion and capture

a) Construction of global strategies

In this part of Section 4 we consider a given family of the local strategies $\{V_\alpha, G_\alpha\}_{\alpha \in A}$ with the family $\{G_\alpha\}_{\alpha \in A}$ covering a given open set $\Omega \subset G$. With the help of the local strategy $\{V_\alpha, G_\alpha\}$ we can choose our strategy in each time point t satisfying the relation $x(t) \in G_\alpha$. If the trajectory of the game reaches a boundary point $x(t_\alpha) \in \partial G_\alpha$ called decision point then we have to select an other local strategy $(V_{\alpha'}, G_{\alpha'})$ with $x(t_\alpha) \in G_{\alpha'}$. The aim of constructing global strategies considered here is to give a rule to

choose one of the local strategies so as to optimize in some sense the number of the decision points along the trajectories for the given family of the local strategies.

First of all we define a function characterizing the covering system:

$$h: \Omega \rightarrow R^+$$

$$h(x) := \sup \{ \delta \mid \delta > 0 \text{ and } \exists \alpha(\delta) \in A \text{ that } G_\delta(x) \subset G_{\alpha(\delta)} \}. \quad (4.1)$$

Lemma 4.1. h is a Lipschitzian function with Lipschitzian constant 1.

The synthesis of the family of the local strategies $\{V_\alpha, G_\alpha\}_{\alpha \in A}$ is based on the following theorem:

Theorem 4.1. If $0 < \mu < 1$ is an arbitrary fixed real number and $\{V_\alpha, G_\alpha\}_{\alpha \in A}$ is a family of the local strategies then a locally finite family of open sets $\{G'_\beta\}_{\beta \in B}$ covering Ω can be given with the following properties:

1. An $\alpha(\beta) \in A$ is associated to each $\beta \in B$ with $G'_\beta \subset G_{\alpha(\beta)}$ and

$$\rho(G'_\beta, \partial G_{\alpha(\beta)}) \geq (1 - \mu) \cdot \sup_{x \in G'_\beta} h(x). \quad (4.2)$$

2. For each $\beta \in B$ such a point $x \in G'_\beta$ can be found that $x \notin G'_\gamma$ if $\gamma \in B, \gamma \neq \beta$.
3. B is well ordered and if $p \in \Omega$ is an arbitrary element then

$$\beta(p) := \min \{ \beta \mid \beta \in B \text{ and } p \in G'_\beta \} \quad (4.3)$$

$$\alpha(p) := \alpha(\beta(p)).$$

Now we are able to give the definition of the global strategies.

Let us denote the set $\{g \mid g: [t_0, t(g)] \subset [t_0, t_1] \rightarrow \Omega\}^*$ of continuous curves by $C([t_0, t_1], \Omega)$.

If $g \in C([t_0, t_1], \Omega)$ then using the construction of Theorem 4.1, we associate a sequence of decision points $\{z_i\}_{i=1}^\infty$ to g as follows.

Setting $\tau_0 := t_0$ let z_0 be taken for $g(\tau_0)$. Taking $G'_{\beta(z_0)}$ let $\tau_0 < \tau_1 \in R$ be the maximal number satisfying the condition $g([\tau_0, \tau_1]) \subset G'_{\beta(z_0)}$. Then $z_1 := g(\tau_1)$. Let us suppose that (z_i, τ_i) is given for $i = 1, 2, \dots, k$. Then we define τ_{k+1} as the maximal real number $\tau_k < \tau_{k+1}$ satisfying the condition $g([\tau_k, \tau_{k+1}]) \subset G'_{\beta(z_k)}$. Then $z_{k+1} := g(\tau_{k+1})$. The sequence $\{z_i\}_{i=1}^\infty$ is called the set of decision points associated to g .

Now we define a mapping

$$V: C([t_0, t_1], \Omega) \times \Omega \times U_{K'} \rightarrow U_K \quad (4.4)$$

as follows:

If $g \in C([t_0, t_1], \Omega)$ and $x = g(t)$ for $t \in [\tau_i, \tau_{i+1}]$ then

$$V(g, x, u) := V_{\alpha(z_i)}(u), \quad u \in U_{K'}, \quad i = 0, 1, \dots \quad (4.5)$$

* The interval $[t_0, t(g)]$ can be closed or closed-open.

Definition 4.1. The mapping V is called the global strategy composed from $\{V_\alpha, G_\alpha\}_{\alpha \in A}$.

Remark 4.1. It is evident that each one of the initial value problems

$$\begin{aligned} x(t_0) &= x_0 \\ \dot{x}(t) &= F(x(t), v(t), V(x, x(t), u(t)), u(t)) \end{aligned} \quad (4.6)$$

with $v \in \prod_{i \in \langle N \rangle \setminus (K \cup K')} S_i$ and $u \in \prod_{i \in K'} S_i$ and

$$\begin{aligned} x(t_0) &= x_0 \\ \dot{x}(t) &= F(x(t), v(t), V(g, g(t), v(t)), u(t)) \end{aligned} \quad (4.7)$$

with $g \in C([t_0, t_1], \Omega)$, $v \in \prod_{i \in \langle N \rangle \setminus (K \cup K')} S_i$ and $u \in \prod_{i \in K'} S_i$ has a unique solution. In the case of the strong superiority $U_{K'} := U_{K'}^1$.

Now we may turn our attention to the theorems about evasion and capture.

b) Evasion and capture

Considering the two coalitions K and K' , $K \cap K' = \emptyset$, H_K and $H_{K'}$ denote the closed sets $\bigcup_{i \in K} \theta_i \subset G$ and $\bigcup_{i \in K'} \theta_i \subset G$, $v \in \prod_{i \in K} S_i$ that inclusion $x(t^*) \in H_{K'}^*$ be satisfied for a $t^* \in [t_0, t_1]$ and $x(t) \notin H_K$, for $t \in [t_0, t_1]$. To treat these problems we consider a closed set $H \subset G \subset E$ and by means of local and global strategies we give conditions and a method to construct strategies for both capture and evasion.

First we show a simple lemma necessary for defining the family of local strategies.

If $\rho(x, H)$ denotes the distance of x and H then an $y(x) \in \partial H$ can be selected with the property

$$\rho(x, H) \leq \|x - y(x)\| < (1 + \varepsilon \cdot \rho(x, H)) \cdot \rho(x, H) \quad (4.8)$$

for arbitrary $1 > \varepsilon > 0$.

With the help of the following definition we shall specify the now treated class of the target sets $\{H\}$.

Definition 4.2. The function ρ is smooth at the point $x \in E$ if

- an open neighbourhood $G_{\delta'(x)}(x)$,
- an $e_x \in E$, $\|e_x\| = 1$,
- $e_{x_1, x_2} \in E$, $\|e_{x_1, x_2}\| \leq C \cdot \rho(x_1, H)$ for all pair $x_1, x_2 \in G_{\delta'(x)}(x)$ and
- $\alpha_{x_1, x_2} \in R$, $x_1, x_2 \in G_{\delta'(x)}(x)$ can be given satisfying the relations

$$\rho(x_i, H) = \langle e_x + e_{x_1, x_2}, (x_i - y(x)) \rangle + \alpha_{x_1, x_2}, \quad i = 1, 2 \quad (4.9)$$

and

$$\langle e_x, x - y(x) \rangle = \|x - y(x)\| \quad (4.10)$$

for a suitable constant $c \in \mathbb{R}$, $c > 0$.

It is easy to check that if $\rho(x, H)$ is differentiable then it is smooth. This occurs in Hilbert spaces for any convex set H .

Lemma 4.2. If ρ is smooth at the point $x \in E$ then such an open neighbourhood $G_{\delta(x)}(x)$, $\delta(x) < \delta'(x)$ can be selected that

$$\delta(x) < \varepsilon \cdot \rho(x, H) \quad (4.11)$$

and

$$(1 - \varepsilon \cdot \rho(x', H)) \cdot \rho(x', H) < \langle e_x, x' - y(x) \rangle < (1 + \varepsilon \cdot \rho(x', H)) \cdot \rho(x', H),$$

$$x' \in G_{\delta(x)}(x) \quad (4.12)$$

is satisfied with $\delta'(x)$, e_x and $y(x)$ as specified in Definition 4.2 and $1 > \varepsilon > 0$.

This is evident consequence of the continuity of ρ and of the definitions.

Now we are in the position to construct the global strategies for evasion and capture.

If $\Omega_1, H \subset \Omega_1 \subset G$ is an open set then $\Omega := \Omega \setminus H$.

Let us suppose that

a) The function ρ is smooth in Ω with an uniform constant $C > 0$.

b) The coalition K has one of the four kinds of superiority at the point $y(x) \in \partial H$ in the direction e_x with $y(x)$, e_x associated to x by the smoothness of ρ for each $x \in \Omega$.

Denoting by V_x the mapping associated to the superiority at $y(x)$ by Theorem 3.4 and taking the neighbourhood $G_{\delta(x)}(x)$ given in Lemma 4.2 we obtain a family of local strategies $\{V_x, G_{\delta(x)}(x)\}_{x \in \Omega}$. Let h denote the function ordered to the open covering system $\{G_{\delta(x)}(x)\}_{x \in \Omega}$ by Definition 4.1.

Let V^μ denote the global strategy constructed in Theorem 4.1 with the constant $1 > \mu > 0$ from the family of local strategies $\{V_x, G_{\delta(x)}(x)\}_{x \in \Omega}$.

Let us suppose that $z, \tilde{z} \in C([t_0, t_1], \Omega)$ with common domains fulfil the following conditions:

$$a) \quad \|z(t) - \tilde{z}(t)\| \leq v \cdot h(\tilde{z}(t)) \quad (4.13)$$

for $t \in D(z)$ and $1 > v > 0$.

* $x(t^*)$ is given as in (2.5).

b) $z(t_0) = \tilde{z}(t_0) \in G \setminus H$ and according to the type of superiority, one of the following equations

$$\dot{z}(t) = F(z(t)) \cdot V^\mu(\tilde{z}, \tilde{z}(t), u(t)), u(t)) \quad (4.14)$$

and

$$\dot{z}(t) = F(z(t), v(t), V^\mu(\tilde{z}, \tilde{z}(t), u(t)), u(t))$$

is satisfied with $u(t), v(t)$ given in Remark 4.1.

Remark 4.2. $\tilde{z}(t)$ means our approximate information about $z(t)$ in the moment $t \in [t_0, t_1]$. From the definition of the local and the global strategies follows that in the case of strong superiorities the coalition K selects its strategy without information about $u(t) \in U_{K'}, t \in [t_0, t_1]$.

Let H be $H_{K'}$.

Theorem 4.2. If the coalition K has weak superiority over K' or strong superiority both for evasion at $y(x) \in \partial H_{K'}$ in the direction e_x for each $x \in \Omega$ and $\mu = 1 - \nu$ in (4.14)

then t^* given in (2.5) is either infinite or $\lim_{t \uparrow t^*} (z(t)) \in G \setminus H_{K'}$ holds for finite t^* .

Now we state a theorem about capture. The notations used before Theorem 4.2 are valid except that $H := H_K$.

Theorem 4.3. Let us suppose that coalition K satisfies one of the conditions (3.4) or (3.6) according to its weak superiority over K' or strong superiority at $y(x) \in \partial H_K$ in the direction $-e_x$ with a universal $c := c(p) > 0$ for each element x of a neighbourhood $G_0(p) \subset G, p \in \partial H_K$. Then such a neighbourhood $G(p)$ and a $T > t_0$ can be found that in the case of $z(t_0) \in G(p)$, using global strategy $V^{1-\nu}$ coalition K terminates the game in a suitable point of time $t^* \in [t_0, T]$; that is, the solution of (4.14) satisfies $\lim_{t \uparrow t^*} z(t) \in \partial H_K$.

References

1. *Blaquiere, A.*, Topics in Differential games. 1973. North-Holland Publ. Company, Amsterdam—London.
2. *Gamkrelidze—Haratishvili:* Differentsialnaja igra oklonenija s nelinejnym upravleniem. Trudy Ordena Lenina Math. Inst. im. V. A. Steklova A. N. SSSR. Vol. 112, 1971.
3. *Krasovskii, N. N., Subbotin, A. I.*, Pozitsionnye differentsialnye igri. Izd. Nauka, Moskva, 1974.
4. *Lagunov, V. N.*, A nonlinear differential game of evasion. DAN 202, 1972.
5. *Pontrjagin, L. S.*, Lineinaia differentsialnaia igra ubeganiya. Trudy Ordena Lenina Math. Inst. im. V. A. Steklova A. N. SSSR. Vol. 112, 1971.
6. *Ryll-Nardzewski, C.*, A general theorem on selectors. Bull. Acad. Polon. Sci. Ser. Math., 13, No. 6 (1965), 397–402.
7. *Lipcey, Zs.*, N-person nonlinear qualitative differential games with incomplete information. MTA SzTAKI Working Paper, IV/16, 1981. Preprint.

**Обзор результатов по нелинейным дифференциальным играм
качества для N сторон**

ж. ЛИПЧЕИ

(Будапешт)

В статье исследуются нелинейные дифференциальные игры N сторон. Показано, что предложенный метод синтеза стратегий делает возможным уклонение и попадание для коалиции, имеющей приближенную информацию о фазовой и граничной точке целевого множества.

Zs. Lipcsey
Computer and Automation Institute
Hungarian Academy of Sciences
Budapest 1111 Kende u. 13-17
Hungary

ROBUSTNESS OF BAYES ESTIMATION OF DYNAMIC SYSTEM PARAMETERS

J. NOVOVIČOVÁ

(Prague)

(Received April 4, 1982)

The paper is concerned with the Bayesian estimation problem in the linear dynamic system if prior knowledge about system parameters and the parameter of the error probability distribution is available. It is shown that Bayes estimator with respect to a quadratic loss function, normally distributed error and conjugate normal gamma prior distribution is robust with regard to optimality criterion, if we allow for larger classes of error and prior distributions and more general loss function.

1. Introduction

This paper deals with estimation of the dynamic system parameters of the linear regression form, when certain prior knowledge is available about the unknown parameters of the system and the parameter of the error probability distribution. The efficient use of the prior knowledge is made by the Bayesian analysis. The Bayesian estimation procedure involves the specification of a complete probabilistic model, i.e. distribution of errors and prior distribution of the unknown parameters and the specification of loss function.

In practical situations our prior knowledge are more or less incomplete. For example, we can specify only some moments of the required probability distributions and the loss function can be characterized at most by some qualitative properties such as continuity, monotonicity and convexity. Thus it becomes a natural task to look procedures which are robust under a change of the probability distributions and the loss structure.

We will consider Bayes optimality criterion for the choice of optimal estimator of the unknown parameters in the linear dynamical system. It will be demonstrated that the well-known Bayes estimator with respect to normally distributed errors, conjugate normal gamma prior distribution and quadratic loss function, is sufficiently robust with regard to the optimality criterion, if we allow for larger classes of error and prior distributions and more general loss functions.

2. Formulation of the problem; basic assumptions and notations

Consider the following problem of parameter estimation. Suppose the mathematical model of the dynamic time invariant single-input/single-output system of the linear regression form

$$y_{(t)} = \sum_{i=1}^n a_i y_{(t-i)} + \sum_{i=0}^n b_i u_{(t-i)} + e_{(t)}, \quad t=1, 2, \dots, N \quad (2.1)$$

where $u_{(t)}$ is the input variable, $y_{(t)}$ is the output variable and $e_{(t)}$ represents the errors that occur at each observation time $\tau, \tau=1, 2, \dots, N$. The variables u and y are considered to be measurable, while e is an unmeasurable one. The $p=2n+1$ parameters $a_i, i=1, 2, \dots, n$ and $b_i, i=0, 1, \dots, n$ of regression model (2.1) are unknown. The upper bound n in (2.1) is assumed to be known.

By using notation

$$\theta^T \triangleq (b_0, a_1, b_1, \dots, a_n, b_n); \quad \theta \in \Theta \subseteq R^p,$$

$$z_{(t)}^T \triangleq (u_{(t)}, y_{(t-1)}, u_{(t-1)}, \dots, y_{(t-n)}, u_{(t-n)})$$

equation (2.1) can be written in the form

$$y_{(t)} = \theta^T z_{(t)} + e_{(t)}, \quad t=1, 2, \dots, N. \quad (2.2)$$

Denote the sets of input-output pair $D_{(t)} = (y_{(t)}, u_{(t)})$ (the data) by

$$D^{(t)} = (y^{(t)}, u^{(t)}) \triangleq (y_{(t)}, u_{(t)}, \dots, y_{(1)}, u_{(1)}).$$

Suppose that the system was observed up to and including the time index t , i.e. the data $D^{(t)}$ are known.

We wish to estimate the unknown vector θ on the basis of known data $D^{(t)}$ when certain prior knowledge is available about the vector parameter θ and the parameter of the error distribution.

We can distinguish two situations [3]. In the first case, an amount of data is fixed and we want to estimate θ in one shot (one step). In the other case the amount of the data is growing and the estimation is required in real time for every t . We have an interest in the (recursive) real time estimation problem. Pilz [5] investigated robustness of one-shot Bayesian estimation in the linear static model.

For the sake of simplicity we assume, that the random errors in (2.2) are uncorrelated with the same unknown variance. For arbitrary $t > n$ denote by

$$\mathcal{P}_e = \{P_{e_{(t)}}^{(w)} \triangleq P(e_{(t)} | w, u_{(t)}, D^{(t-1)}): w \in W, w = (\sigma, \lambda); W = R^+ \times \Lambda\}$$

the class of possible conditional error distributions depending on the unknown variance parameter $\sigma^2 \in R^+ = (0, \infty)$ and a further index $\lambda \in \Lambda$, where Λ is an arbitrary index set.

Assumption 1. For any distribution $P_{e(t)}^{(w)} \in \mathcal{P}_e$ is

$$\begin{aligned} E[e_{(t)} | w, u_{(t)}, D^{(t-1)}] &= 0, \\ E[e_{(t)}^2 | w, u_{(t)}, D^{(t-1)}] &= \sigma^2 \quad (\text{independent of } \lambda). \end{aligned}$$

A typical example for $\mathcal{P}_{e(t)}^{(w)}$ would be the family of exponential power distributions with variance σ^2 and the parameter $\lambda \in \langle 1, \infty \rangle$ which can be interpreted as a measure of the deviation of the error distribution from normality, occurring for $\lambda = 2$.

Assumption 2. Natural conditions of control are satisfied (see [3]), i.e.

$$P(u_{(t)} | \theta, w, D^{(t-1)}) = P(u_{(t)} | D^{(t-1)})$$

and thus it holds

$$P(\theta, w | u_{(t)}, D^{(t-1)}) = P(\theta, w | D^{(t-1)}).$$

Assumption 3. At any time $t > n$ we are given, as a result of previous calculation and as part of the problem statement, the prior knowledge about the unknown parameter θ and $w = (\sigma, \lambda)$ described by the prior distribution $P(\theta, w | D^{(t-1)}) \triangleq P_{\theta, w}$.

By $L(\theta, w; \theta_{(t)})$ we denote the loss which occurs if θ is estimated by $\theta_{(t)}$.

Assumption 4. The loss function is of the form

$$L(\theta, w; \theta_{(t)}) = h(w)L_0(\theta - \theta_{(t)})$$

with nonnegative functions h and L_0 .

The loss function will generate the risk function

$$R(\theta, w; \theta_{(t)}) \triangleq \int_{R^1} L(\theta, w; \theta_{(t)}) dP(y_{(t)} | \theta, w, u_{(t)}, D^{(t-1)}), \quad (2.3)$$

where $P(y_{(t)} | \theta, w, u_{(t)}, D^{(t-1)})$ is the conditional distribution of $y_{(t)}$ for given θ and error distribution $P_{e(t)}^{(w)} \in \mathcal{P}_e$ and for the known past history of the input-output process including the last input.

According to these level of prior knowledge we choose the following Bayes optimality criterion.

Find the estimator $\theta_{(t)} \in \mathcal{D}$ so that it mimimizes the prior risk

$$\rho(P_{\theta, w}; \theta_{(t)}) \triangleq \int_{\theta \times w} R(\theta, w; \theta_{(t)}) dP(\theta, w | D^{(t-1)}), \quad (2.4)$$

where \mathcal{D} is the class of admissible estimators.

Definition 1. A Bayes estimator $\bar{\theta}_{(t)}$ of θ at time t with respect to the prior distribution $P_{\theta, w}$ is the estimator in \mathcal{D} which minimizes the prior risk (2.4) [6].

On the basis of the results of Ferguson [2], Pilz [5] and Peterka [4] the following assertions can be proved.

A1. The estimator $\bar{\theta}_{(t)}$ is Bayesian w.r.t. $P_{\theta, w}$ if and only if it holds

$$\begin{aligned} E_{\theta, w|D^{(t)}}[L(\theta, w; \bar{\theta}_{(t)})] &\triangleq \int_{\Theta \times W} L(\theta, w; \bar{\theta}_{(t)}) dP(\theta, w|D^{(t)}) = \\ &= \inf_{\theta_{(t)} \in \mathcal{D}} E_{\theta, w|D^{(t)}}[L(\theta, w; \theta_{(t)})]. \end{aligned}$$

A2. Let the loss function be quadratic and \mathcal{P}_e and $P(\theta, w|D^{(t-1)})$ is such that

$$\int_{\Theta} \|\theta - E(\theta|w, D^{(t)})\|^2 dP(\theta|w, D^{(t)})$$

exists. Then $\bar{\theta}_{(t)}^{(w)} = E(\theta|w, D^{(t)})$ is Bayesian estimator w.r.t. $P_{\theta, w}$.

A3. Let the loss function be quadratic. If for all $w \in W$ it holds $E(\theta|w, D^{(t)}) = E(\theta|D^{(t)})$ then

$$\bar{\theta}_{(t)} = E(\theta|D^{(t)}) \text{ is Bayes for } \theta \text{ w.r.t. } P_{\theta, w}.$$

From the results of Peterka [4], it follows that if assumption 2 is satisfied and

B1. the distribution of errors is normal, i.e.

$$P_{e_{(t)}} = N(0, \sigma^2), \quad \sigma^2 \in R^+; \quad (2.5)$$

B2. the loss function is quadratic, i.e.

$$L_0(\theta, \sigma^2; \theta_{(t)}) = \|\theta - \theta_{(t)}\|^2; \quad (2.6)$$

B3. prior knowledge about the unknown parameters (θ, ω) , $\omega = \sigma^{-2}$ is described by the prior distribution $P(\theta, \omega|D^{(t-1)})$ which is such that the conditional prior distribution $P(\theta|\omega, D^{(t-1)})$ of the regression parameters for given variance σ^2 is normal with known expectation $\bar{\theta}_{(t-1)}$ and covariance matrix $\sigma^2 C_{(t-1)}$ and the marginal prior distribution $P(\omega|D^{(t-1)})$ is gamma, i.e. $P(\theta, \omega|D^{(t-1)})$ is conjugate normal-gamma distribution, then the Bayes estimator for θ at time t with respect to the prior distribution $P_{\theta, \omega}$ is

$$\bar{\theta}_{(t)} = (z_{(t)} z_{(t)}^T + C_{(t-1)}^{-1})^{-1} (z_{(t)} y_{(t)} + C_{(t-1)}^{-1} \cdot \bar{\theta}_{(t-1)}). \quad (2.7)$$

3. Robustness of Bayes estimation against error distribution, prior distributions and the loss structure

Now we shall investigate the robustness of optimality of the Bayes estimator $\bar{\theta}_{(t)}$. We shall show that the Bayes optimality of $\bar{\theta}_{(t)}$ is preserved, if one or more of the assumptions B1–B3 are weakened.

We consider the following two cases.

(i) The loss function is quadratic and we investigate robustness of optimality of the Bayes estimator under a change of error distributions and prior distributions about unknown parameters.

(ii) The assumption B2 about the loss function is dropped.

The following theorem formulates condition on the error and prior distributions which, in the case of a quadratic loss function, assures Bayes optimality of $\bar{\theta}_{(t)}$ given by (2.7).

Theorem 1. Let $\Theta = R^p$ and assumption 2 be satisfied with L_0 given by (2.6), assume the prior distribution to be $P(\theta, w | D^{(t-1)})$ and the distributions in \mathcal{P}_e to have densities $p(\theta, w | D^{(t-1)})$ and $p(e_{(t)} | w, u_{(t)}, D^{(t-1)})$, respectively. If the densities are such that for any $w \in W$, there exists a symmetric function $f_w: R^p \rightarrow R^+$ such that for all $\theta \in R^p$ and all possible $y_{(t)}$ it holds

$$p(\theta | w, D^{(t-1)}) \cdot p(y_{(t)} | \theta, w, u_{(t)}, D^{(t-1)}) = f_w(\theta - \bar{\theta}_{(t)}),$$

then $\bar{\theta}_{(t)}$ from (2.7) is Bayesian w.r.t. $P_{\theta, w}$.

Proof. Let be $w \in W$. Under assumption 2 it holds

$$\begin{aligned} p(\theta | w, y^{(t)}, u^{(t)}) &= \\ &= \frac{p(y_{(t)} | \theta, w, y^{(t-1)}, u^{(t)}) \cdot p(\theta | w, y^{(t-1)}, u^{(t-1)})}{\int_{R^p} p(y_{(t)} | \theta, w, y^{(t-1)}, u^{(t)}) \cdot p(\theta | w, y^{(t-1)}, u^{(t-1)}) d\theta} = \\ &= f_w(\theta - \bar{\theta}_{(t)}) / \int_{R^p} f_w(\theta - \bar{\theta}_{(t)}) d\theta. \end{aligned}$$

Further

$$\begin{aligned} E(\theta | w, D^{(t)}) &= \frac{\int_{R^p} \theta f_w(\theta - \bar{\theta}_{(t)}) d\theta}{\int_{R^p} f_w(\theta - \bar{\theta}_{(t)}) d\theta} = \\ &= \int_{R^p} (\theta + \bar{\theta}_{(t)}) f_w(\theta) d\theta / \int_{R^p} f_w(\theta) d\theta = \\ &= \int_{R^p} \theta f_w(\theta) d\theta / \int_{R^p} f_w(\theta) d\theta + \bar{\theta}_{(t)}. \end{aligned}$$

Under the assumption of Theorem 1, the function $f_w(\cdot)$ is symmetric and therefore $\int_{R^p} \theta f_w(\theta) d\theta = 0$ and $E(\theta|w, D^{(t)}) = \bar{\theta}_{(t)}$ and the expectation is independent of w and the rest of the proof follows from A2.

Corollary 1. Let the assumptions B1, B2 be satisfied. Then $\bar{\theta}_{(t)}$ defined by (2.7) is Bayesian w.r.t. any prior distribution $P(\theta, \omega|D^{(t-1)})$ for which

$$P(\theta|\omega, D^{(t-1)}) = N\left(\bar{\theta}_{(t-1)}, \frac{1}{\omega} C_{(t-1)}\right).$$

Proof. If $\mathcal{P}_e = \left\{ N\left(0, \frac{1}{\omega}\right), \omega \in R^+ \right\}$ then as shown in [3]

$$p(y_{(t)}|\theta, \omega, y^{(t-1)}, u^{(t)}) \propto \exp\left\{-\frac{\omega}{2}(y_{(t)} - \theta^T z_{(t)})^2\right\}$$

and we obtain

$$p(y_{(t)}|\theta, \omega, y^{(t-1)}, u^{(t)}) \cdot p(\theta|\omega, y^{(t-1)}, u^{(t-1)}) \propto f_w^{(1)}(\cdot) \cdot f_w^{(2)}(\cdot)$$

where $f_w^{(1)} = \exp\left\{-\frac{\omega}{2}\|\theta - T(y_{(t)})\|_{z_{(t)}z_{(t)}^T}^2\right\}$ with $T(y_{(t)}) = (z_{(t)}z_{(t)}^T)^+ z_{(t)}y_{(t)}$, $(z_{(t)}z_{(t)}^T)^+$ is the pseudoinverse of $(z_{(t)}z_{(t)}^T)$ and

$$f_w^{(2)} = \exp\left\{-\frac{\omega}{2}\|\theta - \bar{\theta}_{(t-1)}\|_{c_{(t-1)}^{-1}}^2\right\}.$$

It may be shown that the function

$$f_w^{(3)} = \exp\left\{-\frac{\omega}{2}\|\theta - \bar{\theta}_{(t)}\|_{z_{(t)}z_{(t)}^T + c_{(t-1)}^{-1}}^2\right\}$$

is proportional to $f_w^{(1)} \cdot f_w^{(2)}$, therefore

$$\begin{aligned} p(y_{(t)}|\theta, \omega, y^{(t-1)}, u^{(t)}) \cdot p(\theta|\omega, D^{(t)}) &= \\ &= k(y_{(t)}, \omega) \cdot \exp\left\{-\frac{\omega}{2}\|\theta - \bar{\theta}_{(t-1)}\|_{z_{(t)}z_{(t)}^T + c_{(t-1)}^{-1}}^2\right\} \end{aligned}$$

is symmetric and the assertion follows from Theorem 1.

In other words, Corollary 1 tells us that $\bar{\theta}_{(t)}$ is not only a Bayes estimator with respect to conjugate normal gamma distribution. In the case of normally distributed errors it is sufficient that $p(\theta|\omega, D^{(t)})$ is normal and no specification of the marginal distribution $p(\omega|D^{(t)})$ is necessary.

Let us confine to the class of linear estimators, i.e. to the class

$$\mathcal{D}_L = \{\theta_{(t)} : \exists a, b \in R^p : \theta_{(t)} = ay_{(t)} + b\}.$$

Theorem 2. Let be $\Theta = R^p$ and for the class of error distributions \mathcal{P}_e it holds

$$\mathcal{P}_e = \{P_e : E(e_{(t)}|w, u_{(t)}, D^{(t-1)}) = 0 \wedge E(e_{(t)}^2|w, u_{(t)}, D^{(t-1)}) = \sigma^2\}.$$

Let the loss function satisfy assumption B2 and let assumption 2 be satisfied. Then the estimator $\bar{\theta}_{(t)}$ given by (2.7) is Bayesian among all estimators from the class \mathcal{D}_L with respect to any prior distribution $P(\theta, w|D^{(t-1)})$ with

$$E(\theta|w, D^{(t-1)}) = \bar{\theta}_{(t-1)} \wedge \text{cov}(\theta|w, D^{(t-1)}) = \sigma^2 C_{(t-1)} \forall w \in W \quad (2.8)$$

and finite expectations $E(\sigma^2 h(w))$ and $E(h(w))$.

Proof. Let us denote by $\mathcal{P}_{\theta, w}$ the class of the distributions with property (2.8). Let $P(\theta, w|D^{(t-1)}) \in \mathcal{P}_{\theta, w}$ and $\theta_{(t)} = ay_{(t)} + b$ be some linear estimations. For the risk function (2.3) it holds

$$\begin{aligned} R(\theta, w; \theta_{(t)}) &= \int_{R^1} h(w) \|\theta - ay_{(t)} - b\|^2 p(y_{(t)}|\theta, w, u_{(t)}, D^{(t-1)}) dy_{(t)} = \\ &= \int_{R^1} h(w) \|\theta - az_{(t)}^T \theta - ae_{(t)} - b\|^2 p(e_{(t)}|w, u_{(t)}, D^{(t-1)}) de_{(t)} = \\ &= \int_{R^1} h(w) \{ \|(I_p - az_{(t)}^T)\theta - b\|^2 + a^T a e_{(t)}^2 - \\ &\quad - 2a^T [(I_p - az_{(t)}^T)\theta - b] e_{(t)} \} p(e_{(t)}|w, u_{(t)}, D^{(t-1)}) de_{(t)} = \\ &= h(w) \|A_0(\theta - \bar{\theta}_{(t-1)}) + b_0\|^2 + \sigma^2 h(w) a^T a, \end{aligned}$$

where $A_0 = I_p - az_{(t)}^T$, $b_0 = A_0 \bar{\theta}_{(t-1)} - b$ and I_p is the identity matrix. Let us denote $H_1(P_w) = E_w(\sigma^2 h(w))$ and $H_2(P_w) = E_w(h(w))$. Then for the Bayes risk of $\theta_{(t)}$ w.r.t. $P_{\theta, w}$ we obtain

$$\begin{aligned} \rho(P_{\theta, w}; \theta_{(t)}) &= \int_{R^p \times W} R(\theta, w, \theta_{(t)}) dP(\theta, w|D^{(t-1)}) = \\ &= \int_W \int_{R^p} R(\theta, w, \theta_{(t)}) dP(\theta|w, D^{(t-1)}) dP(w|D^{(t-1)}) = \\ &= \int_W \{ \sigma^2 h(w) a^T a + h(w) \text{tr}(A_0 \sigma^2 C_{(t-1)} \cdot A_0^T + b_0 b_0^T) \} dP(w|D^{(t-1)}) = \\ &= H_1(P_w) a^T a + H_2(P_w) \text{tr} A_0 C_{(t-1)} A_0^T + H_2(P_w) b_0^T b_0. \end{aligned}$$

Because all distributions from $\mathcal{P}_{\theta, w}$ have the same conditional expectation and covariance matrix, the Bayes risk of estimation $\theta_{(t)} = ay_{(t)} + b$ depends only on the marginal distribution P_w . Let now $\tilde{P}_{\theta, w} \in \mathcal{P}_{\theta, w}$ be such that

$$\tilde{P}_{\theta, w} = N(\bar{\theta}_{(t-1)}, \sigma^2 C_{(t-1)}) \wedge H_1(\tilde{P}_w) = H_1(P_w) \wedge H_2(\tilde{P}_w) = H_2(P_w),$$

then for all $\theta_{(t)} \in \mathcal{D}_L$ it holds

$$\rho(\tilde{P}_{\theta, w}; \theta_{(t)}) = \rho(P_{\theta, w}; \theta_{(t)}).$$

From the fact that the class \mathcal{P}_e involves the class $\mathcal{P}_e^N = \{N(0, \sigma^2), \sigma^2 \in R^+\}$ and the linear estimation by Corollary 1 is Bayes w.r.t. $\tilde{P}_{\theta, w}$ it follows for all $\theta_{(t)} \in \mathcal{D}_L$

$$\rho(\tilde{P}_{\theta, w}; \theta_{(t)}) \geq \rho(\tilde{P}_{\theta, w}; \bar{\theta}_{(t)}).$$

Hence, since $\rho(\tilde{P}_{\theta, w}; \bar{\theta}_{(t)}) = \rho(P_{\theta, w}; \bar{\theta}_{(t)})$ we have for all $\theta_{(t)} \in \mathcal{D}_L$ and $P_{\theta, w} \in \mathcal{P}_{\theta, w}$

$$\rho(P_{\theta, w}; \theta_{(t)}) \geq \rho(P_{\theta, w}; \bar{\theta}_{(t)}).$$

This implies that we restricting ourselves to the class of linear estimators, we only need to know the variance structure of error and the first two moments of the conditional prior distribution of θ to construct the Bayes estimator.

We will now consider case 2. We first formulate an analogue to Theorem 1 assuring Bayes optimality of $\bar{\theta}_{(t)}$ under monotone loss functions.

Theorem 3. Let $\Theta = R^p$, $P_{\theta, w}$ be some prior distribution and \mathcal{P}_e some class of error distributions. Assume the conditional prior distribution $P_{\theta|w}$ and the distributions $P_{e_{(t)}^{(w)}}$ to have densities $p_{e_{(t)}}$ and $p_{\theta|w}$ and for all $w \in W$ to exist a strongly monotone decreasing function $g_w: R^+ \rightarrow R^+$ and a positive definite matrix B_w such that

$$p(y_t | \theta, w, y^{(t-1)}, u^{(t)}) \cdot p(\theta | w, D^{(t-1)}) = g_w(\|\theta - \bar{\theta}_{(t)}\|_{B_w}^2)$$

holds for any output $y_{(t)}$. If the loss function L satisfies assumption 4 with $L_0: R^+ \rightarrow R^+$ continuous and monotone increasing, then $\bar{\theta}_{(t)}$ is Bayesian w.r.t. $P_{\theta, w}$.

Proof. It holds for any $\theta \in R^p$ and $w \in W$ and given $D^{(t)}$

$$p(\theta | w, D^{(t)}) = g_w(\|\theta - \bar{\theta}_{(t)}\|_{B_w}^2) / \int_{R^p} g_w(\|\theta - \bar{\theta}_{(t)}\|_{B_w}^2) d\theta.$$

From the assumption that g_w is strongly monotonous it follows that the conditional posterior distribution $P(\theta | w, D^{(t)})$ has the unique mode $\bar{\theta}_{(t)}$ and thus

$$\begin{aligned} & \int_{R^p} L_0(\theta - \theta_{(t)}) P(\theta | w, D^{(t)}) d\theta \geq \\ & \geq \int_{R^p} L_0(\theta - \bar{\theta}_{(t)}) P(\theta | w, D^{(t)}) d\theta \end{aligned}$$

for arbitrary estimators $\theta_{(t)} \in \mathcal{D}$. It means that for any $\theta_{(t)} \in \mathcal{D}$ it holds

$$\begin{aligned} & \int_{R^p} \int_W L_0(\theta - \theta_{(t)}) p(\theta, w | D^{(t)}) d\theta dw = \\ & = \int_{R^p} \int_W L_0(\theta - \theta_{(t)}) p(\theta | w, D^{(t)}) \cdot p(w | D^{(t)}) d\theta dw \geq \\ & \geq \int_{R^p} \int_W L_0(\theta - \bar{\theta}_{(t)}) p(\theta | w, D^{(t)}) \cdot p(w | D^{(t)}) d\theta dw = \\ & = \int_{R^p} \int_W L_0(\theta - \bar{\theta}_{(t)}) p(\theta, w | D^{(t)}) d\theta dw, \end{aligned}$$

i.e. $\bar{\theta}_{(t)}$ minimizes the posterior loss and thus is Bayesian.

Remark. The assumption of strong monotonicity of the functions g_w can be dropped if the function L_0 is required to be convex (Deutsch [1]).

Corollary 2. Let be $\Theta = R^p$ and $\mathcal{P}_e = \mathcal{P}_e^N$ the class of normal error distributions. If the loss function L is continuous and monotonously increasing, then $\bar{\theta}_{(t)}$ is Bayes estimator with respect to all prior distributions $P(\theta, \sigma | D^{(t-1)})$ for which $P(\theta | \sigma, D^{(t-1)}) = N(\bar{\theta}_{(t-1)}, \sigma^2 C_{(t-1)})$.

Proof. If $\mathcal{P}_e = \mathcal{P}_e^N$ and $P(\theta, w | D^{(t-1)}) = N(\bar{\theta}_{(t-1)}, \sigma^2 C_{(t-1)})$ then

$$\begin{aligned} & p(y_{(t)} | \theta, w, y^{(t-1)}, u^{(t)}) p(\theta | w, D^{(t)}) = \\ & = k \cdot \exp \cdot \left\{ -\frac{\omega}{2} \|\theta - \bar{\theta}_{(t)}\|_{z_{(t)} z_{(t)}^T + C_{(t-1)}^{-1}}^2 \right\} \end{aligned}$$

where $k = k(D^{(t)}, w)$ is the proportional constant. The matrix $B_w = \omega(z_{(t)} z_{(t)}^T + C_{(t-1)}^{-1})$ is positive definite (because $z z^T$ is positive semidefinite and $C_{(t-1)}^{-1}$ is positive definite) and the function

$$g_w(\cdot) = k \cdot \exp \left(-\frac{\omega}{2} (\cdot) \right) = k \cdot \{ \exp(\cdot) \}^{-\frac{\omega}{2}}$$

is strongly monotonously decreasing and thus the assumption of Theorem 3 is satisfied.

References

1. *Deutsch, R.*, Estimation Theory. Prentice-Hall, Englewood Cliffs, N. J. 1969.
2. *Ferguson, T. S.*, Mathematical statistics — a decision theoretic approach. Acad. Press, New York 1967.
3. *Peterka, V.*, Bayesian system identification. Proceedings 5-th IFAC Symp. on Identification and System Parameter Estimation, Darmstadt, FRG 1979, Vol. 2, 349–356.
4. *Peterka, V.*, Experience accumulation for decision making in multivariate time series. Problems of Control and Information Theory, 1978, Vol. 7, No. 3.
5. *Pilz, J.*, Das bayessche Schätzproblem im linearem Regresionsmodell, Freiburger Forschungsheft (FF4), D117, VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1979.
6. *Zachs, Sh.*, The theory of Statistical Inference. John Wiley, New York 1971.

Робастность байесовского оценивания параметров динамической системы

Я. НОВОВИЧОВА

(Прага)

В статье рассматривается задача байесовой оценки для линейной динамической системы в случае, когда имеется апостериорная информация о параметрах системы и параметрах закона распределения вероятности ошибки.

Jana Novovičová

Institute of Information Theory and Automation

Czechoslovak Academy of Sciences

Pod vodárenskou věží 4

182 08 Praha 8

Czechoslovakia

A DERIVATION OF RESPONSE TIME DISTRIBUTION FOR A M|G|1 PROCESSOR-SHARING QUEUE

S. F. YASHKOV

(Moscow)

(Received 9 February, 1982)

The analysis of M|G|1 queueing system under processor-sharing discipline in steady-state is given. The distribution of the response time for a request of given length is derived by a method of decomposition in delay elements. The distributions of the busy period and the number of requests in the system are also obtained. We give also first two moments of the response time distribution. The asymptotic estimates of the response time variance are determined for short and long requests, and new properties of M|G|1 processor-sharing system are found. A comparison is made with other queueing disciplines.

1. Introduction

Queueing models of time-sharing computer systems are important subjects of research in modern queueing theory [1–3]. We shall consider a processor-sharing system with a general distribution of the required service times. Further, for brevity, we shall use the term “initial length” of a request instead of the “required service time”. The notion of processor-sharing (PS) queueing discipline was initially introduced by Kleinrock [4] as the limiting case of round-robin scheduling for a time-sharing system in which the time quantum is allowed to approach to zero. This discipline assumes that when there are n requests in the single-server system, each request receives service with rate $1/n$. If we call the residual amount of work on the request, measured in units of time, as the residual length (or simply the length) of the request, we shall mean by the service rate the limit of the ratio of the change in the length of the request in the time interval Δ to Δ as $\Delta \rightarrow 0$.

This extremely interesting PS discipline can be rigorously described as follows. All the requests in the system receive service simultaneously with a varying service rate depending on the state of system, i.e. if n requests are at time t in the system then the length of each of these requests is decreased during $[t, t + \Delta]$ by $\Delta/n + o(\Delta)$. A new request must begin to receive a share of the server immediately after entering the system. The request receives service until its length becomes equal to zero. Jumps of the service rates take place at the instants when a number of requests in the system changes.

The practical importance of this queueing model is presented by the fact that it reflects the most essential feature of time-sharing systems or packet switching nodes in

computer networks: the delay in service of each active customer is proportional to their total number at time t . This explains the wide use of PS system in applications.

Consider a conservative M|G|1 system under PS discipline in steady-state. The arrival stream is a homogeneous Poisson process with rate λ . The initial lengths of the requests are independent identically distributed random variables (r.v.) with a general distribution $B(x)$. Let $\beta_1 < \infty$ and $\beta(s)$ be the first moment and the Laplace–Stieltjes transform (LST), respectively, of this distribution. We assume that

$$\rho = \lambda\beta_1 < 1. \quad (1.1)$$

Let $V(\tau)$ be the stationary response time (sojourn time in the system) for a request which has an initial length τ at the arrival time. We define $v(s; \tau) = E \exp(-sV(\tau))$, the LST of the distribution for the r.v. $V(\tau)$.

A theoretical study of the response time of any queueing model can be regarded as completely concluded only when the distribution of the r.v. $V(\tau)$ has been obtained (in term of the LST as usually) and also a method for calculating $v(s; \tau)$ has been provided. The LST $v(s; \tau)$ was derived by Coffman et al. [5] only for M|M|1-PS system, i.e. for the case $B(x) = 1 - \exp(-\mu x)$. A problem of the derivation of $v(s; \tau)$ for M|G|1-PS queue was assumed until recently not to be amenable to an analytical solution. The mean response time of such a system was only known [1–3, 6, 7]. An application of known methods of queueing theory did not give important results for M|G|1 processor-sharing queue until [8–10] were published. The second moment of the distribution of the r.v. $V(\tau)$ was calculated in [8], and the LST of the response time distribution in M|G|1-PS system was given in brief form in [9, 10], but up to now, no satisfactory comprehensive derivation of $v(s; \tau)$ has been given.

This paper contains a considerably improved and somewhat generalized version of the derivation of $v(s; \tau)$ for M|G|1 processor-sharing queue, in which the possibilities of a new method of analysis of the queueing models with variable service rate is explained fairly completely. This new method, called the “method of decomposition in delay elements”, is based on the ideas put forward by the author in [11], developed further in this paper. We discuss also new achievements in an investigation of the random processes in PS system.

2. Auxiliary results

Theorem 1. The LST of the busy period distribution in a PS system satisfies the functional equation

$$\pi(s) = \beta(s + \lambda - \lambda\pi(s)) \quad (2.1)$$

where $\beta(s) = \int_0^{\infty} e^{-sx} dB(x)$.

Proof. A PS system belongs to the class of conservative systems. The property of work-conserving means that the sum of the service rates of all the requests in the system equals to 1 at each instant of time for the busy period.

We consider the unfinished work as a random process $U(t)$. It is clear that $U(t)$ is the process whose value is the sum of the lengths of all the requests. The behaviour of $U(t)$ is described as follows: $U(t)$ jumps by an amount x at the time when a new request of length x arrives, and $dU(t)/dt = -1$ if $U(t) > 0$ in the interarrival times. Since the busy period is the time interval during which $U(t) > 0$, it follows that the distribution of the busy period in the M|G|1 system is invariant with respect to any conservative discipline. It is known that LST of the busy period distribution in M|G|1 ∞ queue with first-come-first-served (FCFS) discipline satisfies equation (2.1) [12]. The validity of (2.1) follows now from the above facts for all the M|G|1 ∞ queueing systems which satisfy a law of conservation [1], in particular, for PS system. Q.E.D.

To describe the state of the PS system at time t , we introduce the random process $X(t) = \{n(t); x_i(t), i = \overline{1, n(t)}\}$. Here $n(t)$ is the number of requests in the system at time t , and supplementary variables $x_i(t)$ indicate the length of the i -th request at time t (the numeration order of the requests is of no importance). The state space of the process $X(t)$ consists of the sets $\{0\}$, $\{x_1\}$, $\{x_1, x_2\}$ etc. ($x_i \geq 0, 1 \leq i \leq n, n = 1, 2, \dots$). This process belongs to the class of piecewise-linear Markov processes subject to discontinuous changes.

We consider also a process $V_t(\tau)$ which describes the response time of the request that arrives at time t and has the length τ at this time.

Theorem 2. The processes $V_t(\tau)$ and $X(t)$ possess the unique limiting (stationary) ergodic distributions independent of the initial conditions.

Proof. A regeneration cycle of the processes $V_t(\tau)$ and $X(t)$ is the sum of the busy period and an idle period, following it. Let $\{t_i\}_{i=1}^{\infty}$ be a sequence of completion times of the i -th regeneration cycle. It follows from (1.1) that $E[t_{i+1} - t_i] < \infty$ with probability 1. The successive regeneration cycles (apart, possibly, from the first) are independent identically distributed r.v. with an absolutely continuous distribution due to (1.1), (2.1) and the fact that interarrival time distribution is a nonlattice distribution. The processes $V_t(\tau)$ and $X(t)$ are regenerative with respect to the sequence $\{t_i\}_{i \geq 1}$. These remarks guarantee that the conditions of the well-known Smith's theorem [13-15] will be satisfied, on the basis of which one can obtain the required statement. Q.E.D.

Corollary 2.1. The process $X(t)$ is bounded in state probabilities.

Theorem 2 provides the existence and uniqueness of the following limits

$$\begin{aligned} v(s; \tau) &= \lim_{t \rightarrow \infty} E \exp(-sV_t(\tau)), \\ p_n(x_1, \dots, x_n) dx_1 \dots dx_n &= \lim_{t \rightarrow \infty} p_n(t; x_1, \dots, x_n) dx_1 \dots dx_n = \\ &= \lim_{t \rightarrow \infty} P\{n(t) = n; x_i(t) \in [x_i, x_i + dx_i), i = \overline{1, n}\}. \end{aligned} \quad (2.2)$$

In (2.2) $p_n(x_1, \dots, x_n)$ denotes a state probability density associated with the state $(n; x_1, \dots, x_n)$, $n \geq 1$ of the process $X(t)$ as $t \rightarrow \infty$. Note that we have assumed here that the final distribution of the process $X(t)$ for fixed $n \geq 1$ has a density. This assumption is justified if we suppose for simplicity that $B(x)$ has a density $\beta(x)$. However, this supposition is only made to simplify the proof of the next theorem since the same result can be obtained by replacement of the densities $p_n(x_1, \dots, x_n)$ by the differentials of the corresponding distributions.

Theorem 3. The joint density of the final distribution of the process $X(t)$ has a product form

$$p_n(x_1, \dots, x_n) = (1 - \rho)\lambda^n \prod_{i=1}^n [1 - B(x_i)], \quad n \geq 1 \quad (2.3)$$

$$p_0 = P\{X(t) \in \{0\}\} = 1 - \rho. \quad (2.4)$$

Proof. Since the process $X(t)$ is Markovian, its densities must satisfy the Chapman-Kolmogorov equations. These equations are derived in the same way as the well-known equations of Sevast'yanov for a multiserver Erlang loss system [12, 16]. A derivation is based on the examination of the sample paths of the process $X(t)$ during an infinitesimal time interval Δ . One must, of course, take into account that the service rates jump at the instants of arrivals and the instants of departures.

The following relation holds

$$\begin{aligned} p_n(t; x_1, \dots, x_n) &= (1 - \lambda\Delta)p_n\left(t - \Delta; x_1 + \frac{\Delta}{n}, \dots, x_n + \frac{\Delta}{n}\right) + (1 - \lambda\Delta) \cdot \\ &\cdot \sum_{i=0}^{n+1} \int_0^{\frac{\Delta}{n+1}} p_{n+1}\left(t - \Delta; x_1 + \frac{\Delta}{n+1}, \dots, x_i, \dots, x_{n+1} + \frac{\Delta}{n+1}\right) dx_i + \\ &+ \frac{\lambda\Delta}{n} \sum_{i=1}^n p_{n-1}\left(t - \Delta; x_1 + \frac{\Delta}{n-1}, \dots, x_{n-1} + \frac{\Delta}{n-1}\right) \beta(x_i) + o(\Delta). \end{aligned} \quad (2.5)$$

The terms on the right side of (2.5) describe (to within $o(\Delta)$) the continuous evolution of the process $X(t)$, and the jumps downwards and upwards during $[t - \Delta, t]$ respectively (the details of the derivation of (2.5) are omitted). Taking the limit as $\Delta \rightarrow 0$ and taking into account (2.2), we obtain from (2.5) the following equations

$$\begin{aligned} \lambda p_0 &= p_1(0), \quad n = 0 \\ \left[-\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial x_i} + \lambda \right] p_n(x_1, \dots, x_n) &= \lambda p_{n-1}(x_1, \dots, x_{n-1}) \beta(x) + p_{n+1}(x_1, \dots, x_n, 0), \\ & n \geq 1 \end{aligned} \quad (2.6)$$

The components of the vectors $X = (x_1, \dots, x_n)$ and $[X, 0] = (x_1, \dots, x_n, 0)$ in eqs. (2.6) are assumed to be unordered, since the densities $p_n(X)$ must be symmetrical with respect to any permutation of the components of X and $[X, 0]$.

The norming conditions for eqs. (2.6) are

$$\sum_{n=0}^{\infty} p_n = 1, \quad p_n = \int_0^{\infty} \dots \int_0^{\infty} p_n(x_1, \dots, x_n) dx_1 \dots dx_n. \quad (2.7)$$

It can be shown by substitution that the stationary densities

$$p_n(x_1, \dots, x_n) = p_0 \rho^n \prod_{i=1}^n \frac{1 - B(x_i)}{\beta_1}$$

satisfy eqs. (2.6), and the constant $p_0 = 1 - \rho$ is calculated from (2.7). This solution is unique in virtue of Theorem 2. Q.E.D.

Corollary 3.1. The number of requests in the M|G|1-PS system has a geometrical distribution

$$p_n = (1 - \rho) \rho^n, \quad n = 0, 1, 2, \dots \quad (2.8)$$

independent of the various distributions $B(x)$ for a fixed mean β_1 .

Remark 1. The theorem analogous to the one above was demonstrated in [8] as an auxiliary result required to prove the Poisson character of the output process of the M|G|1 queue under PS discipline and with the state defined in terms of attained, rather than residual, service time. Similar results for a narrower class of the distributions $B(x)$ with a rational LST were obtained in [17]. More recently, using a new method Yashkov [10, 18] obtained (2.3) and proved the Poisson character of the output for a more general model in which the rules of a random selection in the queue M|G|1 were generalized for a case of variable service rate depending on a position index of the request and a number of all the requests (see also [15] for further discussion and references).

3. Main results

We call x_t -request the request which has the length x at time t . Let c and l be the arrival time and departure time, respectively, of certain τ_c -request. Thus $V(\tau) = [c, l]$ is the total time spent in the system by the tagged τ_c -request. Suppose that the tagged request finds n requests in the system at time c , i.e. the state of the process $X(t)$ is $(n; x_1, \dots, x_n)$ at time $(c-0)$. It is known that Poisson arrivals find the system in a given state with a probability equal to (2.3).

It is necessary at first to calculate the following conditional LST: $E[\exp(-sV(\tau)) | (n; x_1, \dots, x_n)]$. For this purpose it is convenient to use the concepts of

the theory of branching processes [19, Ch. 13–14]. Every requests which are in the system at time $(c - 0)$ are assumed to be the “progenitor” while the new arrivals arriving after the instant c are assumed to be “descendants” of the progenitors. The tagged request is also progenitor.

The key idea of the analysis is the decomposition of the r.v. $V(\tau)$ into a sum of independent delay elements, associated with $(n + 1)$ progenitors [8–11]. This enables one to reduce the problem of derivation of $v(s; \tau)$ to the calculation of certain functionals of the corresponding $(n + 1)$ branching processes which produce descendants.

Remark 2. We shall distinguish the direct descendants by the means of equiprobable random selection mechanism: if n requests are present in the system, then each new arrival is declared with probability $1/n$ by a direct descendant of any (but only one) of these progenitors. Each branching process is formed by one progenitor which generates only direct descendants.

Let $\Phi(x, \tau)$ be the r.v. equal to the sum of increments of attained service of a certain x_c -request and its direct descendants for the time interval during which τ_c -request will be served until completion. This r.v. is a “delay element” experienced by the tagged τ_c -request because of the fact that x_c -request and its direct descendants receive service simultaneously with τ_c -request in $[c, l]$. The direct descendants of one specific progenitor arrive for the service time of this progenitor or of requests which are its previous direct descendants, but we take into account only those descendants which arrive before the instant l . It is easily seen that $\Phi(x, \tau)$ can be interpreted as a new kind of a “broken” busy subperiod initiated by a progenitor of length x , which is broken at time l .

It should be noted that the r.v. $\Phi(x, \tau)$ is independent of x when $x \geq \tau$ because of the fact that x_c -request leaves the system after a departure of the tagged τ_c -request. Hence, the corresponding “broken” busy subperiod is initiated in fact by a τ -request (τ is the attained service of the x -progenitor at time 1). Define

$$D(\tau) = \Phi(x, \tau) \quad \text{for } x \geq \tau. \quad (3.1)$$

We now prove

Theorem 4. The stationary distribution of the response time of τ -request for M|G|1 processor-sharing queue has a LST given by the expressions:

$$v(s; \tau) = (1 - \rho)e^{-(s + \lambda)\tau} [\psi(s; \tau) - \lambda \int_0^\tau e^{-(s + \lambda)x} \psi(s; \tau - x) \bar{B}(x) dx - \lambda e^{-(s + \lambda)\tau} \int_\tau^\infty \bar{B}(x) dx]^{-1} \quad (3.2)$$

where

$$\psi(s; \tau) = \frac{1}{2\pi i} \int_{-i\infty+0}^{+i\infty+0} \frac{q+s+\lambda\beta(q+s+\lambda)}{(q+s+\lambda)[q+\lambda\beta(q+s+\lambda)]} e^{q\tau} dq, \quad (3.3)$$

$$\bar{B}(x) = 1 - B(x) \text{ and } \beta(s) = \int_0^{\infty} e^{-sx} dB(x).$$

Proof. We decompose $V(\tau)$ as

$$V(\tau) = \sum_{i=1}^n \Phi(x_i, \tau) + D(\tau) \quad (3.4)$$

under condition that the state of the process $X(t)$ is $(n; x_1, \dots, x_n)$ at time $c-0$.

Suppose that Δ is infinitesimal; then a recurrent relationship for $\Phi(x, \tau)$ can be represented within $o(\Delta)$ as

$$\Phi(x+\Delta, \tau+\Delta) = \begin{cases} \Delta + \Phi(x, \tau) & \text{if there are no direct descendants} \\ & \text{of a } (x+\Delta)\text{-request} \\ & \text{in a time interval } \eta \\ \Delta + \Phi(x, \tau) + \Phi(y, \tau) & \text{if there is a direct descendant} \\ & \text{of length } y \text{ in a time interval } \eta. \end{cases} \quad (3.5)$$

Here it is essential that the delay element generated by the progenitor should be identical in probability structure with the delay element generated by any direct descendant. The terms of the right side of (3.5) are independent, and η is a time interval during which the $(x+\Delta)$ -request is turned into the x -request. Since Δ is an infinitesimal, η is also infinitesimal due to Corollary 2.1; namely, if there are n requests then $\eta = n\Delta$. But whatever n , the probability of nonappearance or appearance of a direct descendant for an interval η is independent of n because of the fact that the probabilities of the first and second events are respectively:

$$(1 - \lambda\eta) + \lambda\eta \frac{n-1}{n} \rightarrow o(\eta) = 1 - \lambda\Delta + o(\Delta), \quad (3.6)$$

$$\frac{1}{n} [\lambda\eta + o(\eta)] = \lambda\Delta + o(\Delta)$$

due to Remark 2. Then it can be easily seen that the terms of the right side of (3.4) are independent because of the facts that a) the number of the direct descendants of every progenitor depends only on the length of the progenitor, b) the lengths of the requests are independent and c) the increments of the Poisson process are independent.

We define $\varphi(s; x, \tau) = E \exp(-s\Phi(x, \tau))$ and $\delta(s; \tau) = E \exp(-sD(\tau))$, the LST of the distributions for r.v. $\Phi(x, \tau)$ and $D(\tau)$, respectively. Since the terms of the right side of (3.4) are independent, it is easy to rewrite (3.4) in terms of the LST

$$E [\exp(-sV(\tau)) | (n; x_1, \dots, x_n)] = \delta(s; \tau) \prod_{i=1}^n \varphi(s; x_i, \tau).$$

Removing the condition by averaging over the density of distribution (2.3) and (2.4), we have

$$\begin{aligned} v(s; \tau) &= E \exp(-sV(\tau)) = \\ &= (1 - \rho) \delta(s; \tau) [1 - \lambda \int_0^{\infty} \varphi(s; x, \tau) (1 - B(x)) dx]^{-1}. \end{aligned} \quad (3.7)$$

We now proceed to obtain an expression for the $\varphi(s; x, \tau)$ and $\delta(s; \tau)$. Taking into account (3.6), eq. (3.5) may be rewritten in terms of LST by means of a theorem of a total probability as follows

$$\begin{aligned} \varphi(s; x + \Delta, \tau + \Delta) &= (1 - \lambda\Delta) (1 - s\Delta) \varphi(s; x, \tau) + \\ &+ \lambda\Delta (1 - s\Delta) \varphi(s; x, \tau) \int_0^{\infty} \varphi(s; y, \tau) dB(y) + o(\Delta) \end{aligned}$$

which yields

$$\begin{aligned} \frac{\partial \varphi(s; x, \tau)}{\partial x} + \frac{\partial \varphi(s; x, \tau)}{\partial \tau} + \\ + [s + \lambda - \lambda \int_0^{\infty} \varphi(s; y, \tau) dB(y)] \varphi(s; x, \tau) = 0 \end{aligned} \quad (3.8)$$

by taking the limit as $\Delta \rightarrow 0$.

It follows from (3.1) that

$$\varphi(s; x, \tau) = \delta(s; \tau) \quad \text{for } x \geq \tau. \quad (3.9)$$

Taking into account (3.9), we obtain from (3.8)

$$\frac{\partial \delta(s; \tau)}{\partial \tau} + [s + \lambda - \lambda \int_0^{\infty} \varphi(s; y, \tau) dB(y)] \delta(s; \tau) = 0. \quad (3.10)$$

We shall solve the set of integro-differential equations (3.8) and (3.10) with the obvious boundary conditions

$$\varphi(s; 0, \tau) = \varphi(s; x, 0) = \delta(s; 0) = 1 \quad (3.11)$$

in the region $x < \tau$. The note in passing that the following estimate holds for $\delta(s; \tau)$

$$\delta(s; \tau) > e^{-\tau(s + \lambda - \lambda\pi(s))} \quad (3.12)$$

where $\pi(s)$ is given by (2.1).

Equation (3.10) may be rewritten

$$\frac{\partial}{\partial \tau} \ln \delta(s; \tau) = -[s + \lambda - \lambda \int_0^{\infty} \varphi(s; y, \tau) dB(y)].$$

Substituting this equation into (3.8), we obtain the following first-order partial differential equation

$$\frac{\partial \varphi(s; x, \tau)}{\partial x} + \frac{\partial \varphi(s; x, \tau)}{\partial \tau} - \varphi(s; x, \tau) \frac{\partial}{\partial \tau} \ln \delta(s; \tau) = 0. \quad (3.13)$$

Equation (3.13) can be solved by the method of characteristics. This method reduces the problem of integrating eq. (3.13) to integrating an auxiliary set of ordinary differential equations:

$$\frac{dx}{1} = \frac{d\tau}{1} = \frac{d\varphi}{\varphi(\ln \delta)'}$$

where $\varphi \equiv \varphi(s; x, \tau)$ and $\delta \equiv \delta(s; \tau)$.

The first integrals of this set of equations are $C_1 = \tau - x$ and $C_2 = \varphi/\delta$. Since φ only occurs in C_2 , then the general solution of (3.13) is $C_2 = f(C_1)$, where f is an arbitrary function which, in our case, will be determined from the boundary condition (3.11). We have for $x=0$ that $f(\tau) = 1/\delta(s; \tau)$ which enables us to obtain the following relationship

$$\varphi(s; x, \tau) = \delta(s; \tau)/\delta(s; \tau - x) \quad \text{for } x < \tau. \quad (3.14)$$

Substituting (3.14) and (3.9) into (3.10), we obtain the following integro-differential equation for the unknown function $\delta(s; \tau)$

$$\frac{\partial \delta(s; \tau)}{\partial \tau} + \left[s + \lambda - \lambda \int_0^{\tau} \frac{\delta(s; \tau)}{\delta(s; \tau - y)} dB(y) - \lambda \delta(s; \tau) \int_{\tau}^{\infty} dB(y) \right] \delta(s; \tau) = 0. \quad (3.15)$$

Since every solution of the set of equations (3.8) and (3.10) must satisfy condition (3.12), $\delta(s; \tau)$ can be represented in the form

$$\delta(s; \tau) = e^{-(s + \lambda)\tau} \psi(s; \tau)^{-1} \quad (3.16)$$

where $\psi(s; \tau) < e^{-\lambda\pi(s)\tau}$ when $\text{Re } s > 0$.

Substituting (3.16) into (3.15) we obtain the following equation

$$\frac{\partial \psi(s; \tau)}{\partial \tau} + \lambda \int_0^{\tau} e^{-(s+\lambda)y} \psi(s; \tau-y) dB(y) + \lambda [1 - B(\tau)] e^{-(s+\lambda)\tau} = 0 \quad (3.17)$$

with the additional conditions $\psi(s; 0) = 1$ and $\psi(0; \tau) = \exp(-\lambda\tau)$ which are found from (3.16) and (3.11). This equation can be solved by means of a Laplace transform. We define

$$\tilde{\psi}(q, s) = \int_0^{\infty} e^{-q\tau} \psi(s; \tau) d\tau. \quad (3.18)$$

Note that $\tilde{\psi}(q, s)$ is two-dimensional Laplace transform of the function $\Psi(x, \tau)$ of two variables possessing the probability density, i.e.

$$\tilde{\psi}(q, s) = \int_0^{\infty} \int_0^{\infty} e^{-sx - q\tau} d_x \Psi(x, \tau) d\tau$$

while $\psi(s; \tau)$ is the usual Laplace transform with respect to x of the density of the function $\Psi(x, \tau)$.

After applying the transform (3.18) to each term of eq. (3.17), we have

$$q\tilde{\psi}(q, s) - \psi(s; 0) + \lambda \tilde{\psi}(q, s) \beta(q + s + \lambda) + \frac{\lambda [1 - \beta(q + s + \lambda)]}{q + s + \lambda} = 0$$

whence, after some routine algebra and using the inversion integral, we obtain the solution of eq. (3.17) represented by eq. (3.3).

Using expressions (3.9), (3.14) and (3.16) by means of which the functions φ and δ can be rewritten in terms of ψ , we obtain from (3.7) the final result (3.2). Q.E.D.

Let $v_1(\tau) = E[V(\tau)]$ and $\sigma^2[V(\tau)]$ denote the mean and variance, respectively, of the response time distribution for M|G|1-PS queue.

Corollary 4.1.

$$v_1 = \frac{\tau}{1 - \rho}, \quad (3.19)$$

$$\sigma^2[V(\tau)] = \frac{\tau^2}{(1 - \rho)^2} - \frac{2}{1 - \rho} \int_0^{\tau} \delta_1(u) du \quad (3.20)$$

where

$$\delta_1(\tau) = E[D(\tau)] = \tau + \int_0^{\tau} \sum_{n=1}^{\infty} \rho^n F^{*n}(x) dx \quad (3.21)$$

and $F^{*n}(x)$ is an n -fold convolution of a distribution of residual lengths $F(x) = \frac{1}{\beta_1} \int_0^x [1 - B(y)] dy$ with itself, i.e.

$$F^{*n}(x) = \int_0^x F^{*(n-1)}(x-u) dF(u), \quad F^{*0}(u) = 1. \quad (3.22)$$

Proof. This result is proved by taking derivatives of (3.2), using the known properties of the LST:

$$E[V(\tau)^j] = \lim_{s \rightarrow 0} (-1)^j \frac{d^j v(s; \tau)}{ds^j}, \quad j = 1, 2, \dots;$$

$$\sigma^2[V(\tau)] = E[V(\tau)^2] - v_1(\tau)^2.$$

Remark 3. The LST $v(s; \tau)$ is somewhat difficult to differentiate to zero more than once because of the fact that $v(s; \tau)$ has a fairly complex form. Therefore, $E[V(\tau)^2]$ is easier to obtain by solving a set of integro-differential equations [8] which are derived by analogy with (3.8) and (3.10).

4. Discussion of results

It is necessary at first to consider a way of simplifying the calculation of $\sigma^2[V(\tau)]$.

Define

$$R(x) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n F^{*n}(x) \quad (4.1)$$

where $F^{*n}(x)$ is given by (3.22).

We note that $R(x)$ is the limiting distribution of the unfinished work $U(t)$, i.e.

$R(x) = \lim_{t \rightarrow \infty} P\{U(t) \leq x\}$. Taking into account (4.1), it is easy to rewrite (3.20) to the form

$$\begin{aligned} \sigma^2[V(\tau)] &= \frac{2}{(1-\rho)^2} \int_0^{\tau} \int_0^u [1 - R(x)] dx du = \\ &= 2(1-\rho)^2 \int_0^{\tau} (\tau - u) [1 - R(u)] du. \end{aligned} \quad (4.2)$$

Expanding eq. (4.2) in powers of τ at the point $\tau=0$, we obtain the asymptotic estimate of the variance of $V(\tau)$ for small τ

$$\sigma^2[V(\tau)] \sim \tau^2 \rho / [(1-\rho)^2] \quad \text{for } \tau \rightarrow 0. \quad (4.3)$$

We estimate $\sigma^2[V(\tau)]$ for large τ assuming that the first three moments of the distribution $B(x)$ (β_1 , β_2 and β_3) exist. This means that the first two moments of the distribution $F(x)$ are finite, i.e. $f_j = \beta_{j+1} / [(j+1)\beta_1] < \infty$, $j=1, 2$. Define

$$r_1 = \int_0^{\infty} [1 - R(x)] dx = \frac{\rho}{1-\rho} f_1,$$

$$r_2 = \int_0^{\infty} x[1 - R(x)] dx = \frac{\rho}{2(1-\rho)} \left(f_2 + \frac{2\rho}{1-\rho} f_1^2 \right).$$

In terms of r_1 and r_2 , eq. (4.2) may be rewritten

$$\sigma^2[V(\tau)] = \frac{2}{(1-\rho)^2} \left[r_1 \tau - r_2 + \int_{\tau}^{\infty} (u - \tau) [1 - R(u)] du \right]$$

which yields

$$\sigma^2[V(\tau)] \sim \frac{2}{(1-\rho)^2} (r_1 \tau - r_2) \quad \text{for } \tau \rightarrow \infty. \quad (4.4)$$

It is useful to have in applications comparative estimates of the variance of $V(\tau)$ for the various distributions $B(x)$. If the first moments of $B(x)$ are the same, the following monotonous property of $\sigma^2[V(\tau)]$ [9] holds:

$$\sigma_{E_m}^2[V(\tau)] \leq \sigma_M^2[V(\tau)] \leq \sigma_{H_m}^2[V(\tau)]. \quad (4.5)$$

Here the subscripts M , E_m and H_m denote the type of distribution $B(x)$: negative exponential, Erlang of order m and hyperexponential of order m , respectively.

We shall list the main properties of the PS system.

1. The mean response time depends on the length of the request in a linear fashion (see (3.19)).
2. The mean response time is independent of the initial length distribution $B(x)$ for fixed β_1 .
3. The processor resource is shared "fairly" between different requests.

4. PS system has a smaller mean response time (over all the requests) than FCFS system for distributions $B(x)$ that have a coefficient of variation greater than one. The opposite is true for the distributions $B(x)$ that have a coefficient of variation less than one.

5. The number of requests in the PS system is distributed geometrically irrespective of the distribution $B(x)$ (see (2.8)).

6. The final distribution of the unfinished work process is given by (4.1)

7. The response time distribution has a coefficient of variation less than one (see (3.20)).

8. The variance of $V(\tau)$ depends on the length of a request in a quadratic fashion for small τ , and one is approximately linear with τ for large values of τ (see (4.3) and (4.4)).

9. The variance of $V(\tau)$ increases monotonously as a coefficient of variation of $B(x)$ increases (see inequality (4.5)).

10. A PS discipline converts a Poisson process at the input into a Poisson process at the output irrespective of the distribution $B(x)$ [8].

Remark 4. Properties 1–5 were known previously [1–3, 6–7]. Properties 5–10 are new.

Some properties described above are illustrated. The variance of $V(\tau)$ is plotted versus the load ρ for several values of the length τ in Fig. 1 for $M|H_2|1$ -PS system. The service distribution is $B(x) = 1 - 0.6 \exp(-3x) - 0.4 \exp(-0.5x)$ ($\beta_1 = 1.0$, $\beta_2 = 3.33$ and $\beta_3 = 19.33$). For comparison, we show the response time variance for the FCFS

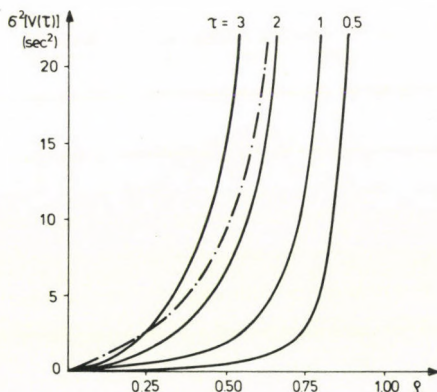


Fig. 1. Variance of response time versus ρ for request of initial length τ in $M|H_2|1$ system with PS and FCFS disciplines
(— PS, - - - FCFS)

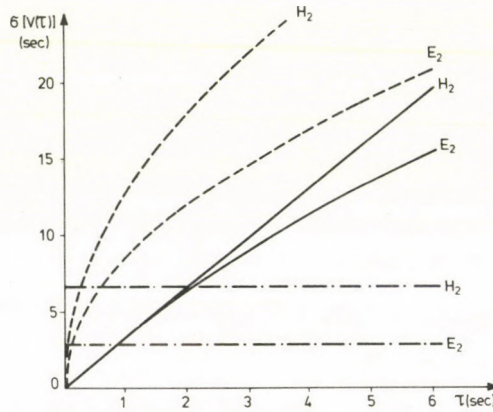


Fig. 2. Standard deviation of response time versus τ for types E_2 and H_2 service distributions under PS, FCFS and preemptive LCFS disciplines ($\rho = 0.75$; — PS, - - - FCFS, — — — preemptive LCFS)

system: $\sigma^2 = \lambda^2 \beta_2^2 / [4(1 - \rho)^2] + \lambda \beta_3 / [3(1 - \rho)]$. The curves demonstrate how over a large range ρ the variance of $V(\tau)$ for short requests (with $\tau < \beta_2 / [2\beta_1]$) is less than that in FCFS system. Thus, PS discipline benefits the short requests at the expense of the long requests in terms of the variance, as well as the mean value, of their response times. We also note that the response time variance in the PS system (over all the requests) is potentially much greater than that in the FCFS system.

Figure 2 shows the standard deviation of response time versus the length of request in $M|E_2|1$ and $M|H_2|1$ systems for PS, FCFS and preemptive LCFS (last-come-first-served) disciplines with $\rho = 0.75$. A type E_2 distribution is $B(x) = 1 - \exp(-2x) - 2x \exp(-2x)$ ($\beta_1 = 1.0$, $\beta_2 = 1.5$ and $\beta_3 = 3.0$), and type H_2 distribution was described above. We note that the response time variance in a preemptive LCFS system is: $\sigma^2[V(\tau)] = \lambda \beta_2 \tau / [(1 - \rho)^3]$ [1, 9]. The curves show that $\sigma[V(\tau)]$ varies for PS and preemptive LCFS disciplines in accordance with properties 8 and 9, and $\sigma[V(\tau)]$ is small sensitive to the service distribution for small τ . As is seen, the variance of $V(\tau)$ in preemptive LCFS system is always considerably greater than that in PS system.

5. Conclusion

We analyzed a processor-sharing system. The service requirements have a general distribution with finite mean. The main contribution of this paper is the derivation of the distribution of the conditional response time to achieve τ sec. of service (see Section 3). Theorem 4 is extremely important in advanced queueing theory

and its computer applications. The importance of this theorem is not limited solely to derivation of the LST of the response time distribution for a M|G|1 processor-sharing queue. It is even more important not the result of the analysis but a new method which enables us to obtain this result. This method is promising for a derivation of a response time distribution for queueing systems under non-standard time-sharing and processor-sharing disciplines which are impossible to investigate by such known methods of queueing theory as delay cycle analysis, a method of imbedded Markov chains, a method of collective marks, etc.

The essence of this method is a nontrivial decomposition of the response time in the sum of delay elements which are interpreted in terms of branching processes theory. The dynamic of the change of the delay elements is described by complex integro-differential equations which are derived by a new probabilistic argumentation. If the solution of these equations can be obtained, then we derive a response time distribution conditioned on a system state. Probability densities of the system states are calculated by means of the well-known supplementary variables technique. The final result is obtained after removing of the conditions.

References

1. Kleinrock, L., Queueing Systems, Vol. II: Computer Applications. Wiley-Interscience, New York-London-Sydney-Toronto, 1976.
2. Lipayev, V. V., Yashkov, S. F., Efficiency of Methods of Organizing Computing Processes in Automatic Control Systems. Statistika, Moscow, 1975 (in Russian).
3. Kobayashi, H., Konheim, A. G., Queueing Models for Computer Communications System Analysis. JEEE Trans. on Commun., Vol. Com. 25 (1977), No. 1.
4. Kleinrock, L., Time-Shared Systems: A Theoretical Treatment. J. of ACM, Vol. 14 (1967), No. 2.
5. Coffman, E. G., Muntz, R. R., Trotter, H., Waiting Time Distributions for Processor-Sharing Systems. J. of ACM, Vol. 17 (1970), No. 1.
6. Sakata, M., Noguchi, S., Oizumi, J., An Analysis of the M|G|1 Queue under Round-Robin Scheduling. Oper. Res., Vol. 19 (1971), No. 2.
7. O'Donovan, T. M., Direct Solutions of M|G|1 Processor-Sharing Models. Oper. Res., Vol. 22 (1974), No. 6.
8. Kitayev, M. Yu., Yashkov, S. F., Analysis of a Single-Server Queueing System with Egalitarian Processor-Sharing Discipline. Engrg. Cybernetics, Vol. 17 (1979), No. 6 (transl. Russian journ. Izv. Akad. Nauk SSSR, Tehn. Kibernetika).
9. Yashkov, S. F., Some Results of Analysing of a Stochastic Model of Remote Processing Systems. Automatic Control and Comput. Sci., Vol. 15 (1981), No. 4. (transl. Russian journ. Avtomat. i Vyčisl. Techn., Riga).
10. Yashkov, S. F., On New Results in M|G|1 Processor-Sharing System Analysis. In: Third Intern. Vilnius Conf. on Probability Theory and Mathem. Statistics, Vol. 3, Vilnius, 1981.
11. Yashkov, S. F., Distribution of the Conditional Waiting Time in a Time-Sharing System. Engrg. Cybernetics, Vol. 15 (1977), No. 5 (transl. Russian journ. Izv. Akad. Nauk SSSR, Tehn. Kibernetika).
12. Gnedenko, B. V., Kovalenko, J. N., Introduction to Queueing Theory. Nauka, Moscow, 1966 (in Russian).
13. Smith, W. L., Regenerative Stochastic Processes. Proc. of the Royal Soc., Ser. A: Mathem. and Phys. Sci., Vol. 232 (1955), No. 1188.
14. Cohen, J. W., On Regenerative Processes in Queueing Theory. Lect. Notes in Econ. and Mathem. Syst., 121, Springer-Verlag, Berlin-Heidelberg-New York, 1976.

15. *Yashkov, S. F.*, On Ergodicity of the Queueing Systems with a Variable Service Rate. *Engrg. Cybernetics*, Vol. 19 (1981), No. 3 (transl. Russian journ. *Izv. Akad. Nauk SSSR, Tehn. Kibernetika*).
16. *Sevast'yanov, B. A.*, An Ergodic Theory for Markov Processes and Its Application to Telephone Systems with Refusals. *Theory of Prob. and Its Appl.*, Vol. 2 (1957), No. 1 (transl. Russian journ. *Teoriya veroyatnosti i ee primeneniye*).
17. *Gelenbe, E., Muntz, R. R.*, Probabilistic Models of Computer Systems, p. 1: Exact Results. *Acta Informatica*, Vol. 7 (1976), No. 1.
18. *Yashkov, S. F.*, Invariance Properties of Probabilistic Models of Adaptive Scheduling in Time-Sharing Systems. *Automatic Control and Comput. Sci.*, Vol. 14 (1980), No. 6 (transl. Russian journ. *Avtomat. i Vychisl. Tehn., Riga*).
19. *Feller, W.*, An Introduction to Probability Theory and Its Applications, Vol. 2. Wiley-Interscience, New York-London-Sydney, 1966.

Получение распределения времени отклика для системы M|G|1 с разделением процессора

С. Ф. ЯШКОВ

(Москва)

Статья посвящена задаче анализа системы массового обслуживания M|G|1 с так называемой дисциплиной разделения процессора, при которой каждое из присутствующих в системе требований обслуживается одновременно с переменной скоростью, равной $1/n$, если система содержит n требований в текущий момент. Предполагается, что система функционирует в стационарном режиме, входящий поток — пуассоновский, а длительности обслуживания распределены произвольно. Практическое значение такой модели состоит в том, что она отражает существенные особенности процессов функционирования вычислительных систем с разделением времени, мультиплексных узлов коммутации сообщений и т.д. Теоретический интерес к исследованию системы M|G|1 с разделением процессора обусловлен тем фактом, что до недавнего времени задача вычисления распределения времени пребывания в системе требования заданной длины (распределение времени отклика) считалась не поддающейся аналитическому решению.

В статье развит новый метод анализа систем с переменной скоростью обслуживания, позволивший получить преобразование Лапласа-Стилтьеса распределения времени отклика в системе M|G|1 с разделением процессора. Суть метода состоит в нетривиальном разложении времени отклика на сумму элементов задержки требования. Динамика изменения элементов задержки описывается интегро-дифференциальными уравнениями, которые выводятся с помощью новой вероятностной аргументации. После получения решения этих уравнений, имеющего некоторую аналогию с уравнениями теории ветвящихся процессов для частиц с энергией, можно вычислить условное распределение времени отклика при условии заданного состояния системы. Плотности вероятностей состояний системы находятся с помощью хорошо известной техники введения дополнительных переменных. После снятия условий получаем окончательный результат.

Вычислены также распределения остальных вероятностных характеристик системы: периода занятости и числа требований. Эти результаты рассматриваются как вспомогательные для получения распределения времени отклика. Вычислены первые два момента распределения времени отклика. Ряд практически важных характеристик системы представлен в явном виде, удобном для расчетов, в частности, дисперсия времени отклика и ее асимптотические оценки. Установлены новые закономерности поведения системы, существенно расширяющие возможности применения полученных результатов.

С. Ф. Яшков

Вычислительный Центр АН СССР

СССР, Москва, 117333,

ул. Вавилова, 40

BOOK REVIEWS

I. CSISZÁR and J. KÖRNER: "*Information Theory: Coding Theorems for Discrete Memoryless Systems*". Akadémiai Kiadó, 1981.

Information theory, as a science for investigating the possibilities in data transmission through communication channels, came into being more than 30 years ago. Since then it has fastly been developed in close connection with communication technology and the procedures of storing and processing of information. Present day Information Theory consists of a deep foundation of its own, on the one hand, and wide ranging as well as deep applications of general mathematical results, on the other. This book by I. Csiszár and J. Körner, published in 1981, was devoted to one specific aspect of Information Theory, viz., the probabilistic coding theory.

The book contains three chapters. Scrutinizing these, it appears to be appropriate to show those peculiarities in which this book differs from others dealing with the same topic.

Chapter 1 contains fundamental ideas on constructions, used further in formulating and proving coding theorems for information sources and channels. It is well known that all methods of data compression used in Information Theory for constructing reliable communication systems for high data transmission rate are based on grouping of the source symbols to be transmitted into long sequences or "blocks", which are processed by the encoders and decoders. Therefore it follows that any study of coding theory needs a detailed examination of the properties of given sequences of symbols, i.e. the structure of the field of sequences. Actually this is the subject of Ch. 1. While what this chapter contains is fairly well known, the presentation appears to be original.

The core of this chapter is the idea of "typical sequences" for some given source rate. Some results, concerning the size of sets of typical

sequences and the measure of these sets (with respect to certain probability distribution on the source alphabet) are presented. Next the properties of mutual information are discussed. By these properties the naturalness of the notion of entropy and the amount of information is shown, and these properties are also frequently used in extremely complex analyses within the next chapters. The authors return at the end of the chapter to source coding theorems, discussing various further aspects and related topics concerning application (e.g. variable length codes, Huffman code, the connection between coding and searching strategies etc.).

Ch. 2, "Classical (two-terminal) systems", contains material which can usually be found in every book on Information Theory. Source coding theorems with admissible distortion, given in advance, and channel coding theorems for noisy channels are presented here. By using the knowledge introduced in Ch. 1, the authors succeed in a uniform presentation of all treated topics. In Sec. 1, coding for noisy channel is investigated. The channel coding theorem and its strong converse is proved.

In Sec. 2, Ch. 2, rate-distortion trade-off in source coding is examined. This leads to the notion of source coding with given fidelity criterion and to the definition of the rate-distortion function. By the rate-distortion theorem, proved in this section, the problem is completely solved.

Sec. 3 is devoted to the problem of the capacity of noisy channel and the rate-distortion function of sources. The implicit definition of the quantities includes the minimization of some multivariable functions, but the straightforward solution given in this book is possible just in special cases.

Because of this the construction of effective optimization algorithms seems to be very important. Algorithms, in which the properties of convexity of the investigated functions is used, are presented in Sec. 3.

The next two sections in Ch. 3 are devoted to the calculation of the error probability of different coding systems.

In Sec. 4 the error exponent in source coding is determined. The so-called "Covering Lemma" constitutes the geometrical base of the solution of this problem. In Sec. 5 the error exponent is determined for channel coding. The "Packing Lemma" leads to the solution of the problem.

The last section in Ch. 2 is concerned with coding for arbitrarily varying channels. In this problem channels with the same input alphabet and with the same output alphabet (i.e. the set of transmission probabilities is given) are considered, and at every use the channel is assumed to be in one of its possible states. Accordingly, various questions can be posed, depending on the knowledge of the sender and the receiver about the state of the channel at every instant. Construction of a channel code (more distinctly, a sequence of codes) is needed, having arbitrarily small error probability for all possible channel states. The solution of this problem needs new approaches, differing from those used for investigating a channel with arbitrary noise.

While in the first two chapters, after all, classical aspects of Information Theory are treated, in Ch. 3, "Multi-Terminal Systems", results obtained during the past decade are presented. In this chapter the theory of multi-terminal systems is presented. Recently, this is one of the most rapidly developing, promising branch of Information Theory. It is of interest because of the possible applications of results in the design of modern communication networks, including a number of correlated sources and channels. The authors introduce a general formulation of coding problem for multi-terminal systems and present the most general results known by now.

In Sec. 1, Ch. 3, the authors carefully examine and prove the Slepain-Wolf Theorem for arbitrary memoryless sources with two components. A general formulation of coding problem for a network of sources is introduced here, considering correlated sources, encoders and decoders.

In Sec. 2, Ch. 3, coding for a noisy channel is considered, assuming a channel with several inputs and outputs, and the simplest channel of this type, the multiple access channel with several inputs and one output is examined. The rate region for this channel is determined.

Sec. 3, Ch. 3, contains techniques. Different probabilistic problems are discussed here, which are

fundamental for coding multi-user systems. These results are used in Sec. 4, for investigating specific networks of sources and channels. The authors consider coding for a network of sources (source coding, assuming a decoder for which side information is available), and for a single specific channel (viz. asymmetric broadcast channel). The analyses of more complex networks are left to the reader. This section is based on results by the authors, many of which are published in this book for the first time.

A feature of this book is due to the great many problems, supplemented by detailed hints and discussions. By this a clear presentation of the fundamental ideas as well as an account of the theorems succeeded at the same time. The hints may appear for the mindful reader sufficient for solving the problems in question.

This very monograph by I. Csiszár and J. Körner, "Information Theory: Coding Theorems for Discrete Memoryless Systems", successfully fills up blank spaces within the literature of Information Theory and related topics. It has got good chances for attracting a broad population of readers interested just in the fundamental ideas within the theory, as well as to specialists, applying the information theory in research.

S. I. GELFAND

V. STREJC: *State Space Theory of Discrete Linear Control*. John Wiley and Sons, Chichester, New York, Brisbane, Toronto, 1982. (A Wiley-Interscience publication, published in co-edition with ACADEMIA, Praha).

The book treats the state space theory of continuous systems and its use for the solution of automatic and digital computer control problems. Besides classical methods, the state space approach is widely applied in modern control theory because of its deeper theoretical background, which enables us to solve more complex problems than classical methods do. Further the state space models are of great advantage, when models are arising from theoretical engineering fields (e.g. chemical or mechanical engineering) applying the state space description for the physical process itself.

The book is written clearly, the text proceeds from the simpler problems to the more complex ones. It contains also the basic knowledge from the field of mathematics (operators, transformations,

special matrix calculus), and from the field of system theory (reachability, controllability, stability, observability, etc.) needed for understanding the subject. It deals with continuous and discrete time state equations including the problems of sampling. It covers the identification of model parameters, the system state vector estimation, the design of computer control based on linear model and quadratic cost function and deals with single and multi-variable deterministic and stochastic systems.

The author takes every effort to treat the subject matter in such a way, as to enable the reader to

understand it without the knowledge of the classical theory. At the same time the understanding of the relationships with the classical theory and physical interpretation is also substantially facilitated. The book may serve as a course of selected studies, for a post-graduate specialization in state space theory. It is also recommended for experts and researchers not only from the field of control engineering but for those of any engineering field aiming at realizing control systems of their technological processes.

Katalin M. HANGOS

PRINTED IN HUNGARY
Akadémiai Nyomda, Budapest

A STOCHASTIC PROGRAM SYNTHESIS OF A GUARANTEEING CONTROL

N. N. KRASOVSKII, V. E. TRET'YAKOV

This paper describes the construction of an optimal guaranteeing control for a linear system with a control process quality index formed by a terminal member and quadratic integral with respect to control and disturbance. Our aim is to demonstrate due to this problem the possibilities of the method of stochastic program synthesis as suggested in [1-2].

1. Consider a system described by the equation

$$\dot{x} = A(t)x + B(t)u + C(t)v, \quad v \in \mathcal{P}, \quad v \in \mathcal{Q} \quad (1.1)$$

where x is an n -dimensional phase vector, u is a p -dimensional vector of control, v is a q -dimensional vector of uncertain disturbance; \mathcal{P} and \mathcal{Q} are taken to be compact; $A(t)$, $B(t)$, $C(t)$ are matrix-valued functions; the time t varies within the limits $t_0 \leq t \leq \vartheta$.

The function $u(\cdot) = \{u(t, x, \varepsilon) \in \mathcal{P}\}$ where $\varepsilon > 0$ is the precision parameter is said to be a control strategy. Suppose the initial position $\{t_*, x_*, t_* \in [t_0, \vartheta]\}$ to be realized. We select $\varepsilon > 0$ and some division $\Delta_\delta\{\tau_i\}$ of a semi-open interval $[t_*, \vartheta)$ into semi-open intervals $[\tau_i, \tau_{i+1})$, $\tau_0 = t_*$, $\tau_m = \vartheta$, $\tau_{i+1} - \tau_i \leq \delta$. A solution of the stepwise equation

$$\dot{x}[t] = A(t)x[t] + B(t)u(\tau_i, x[\tau_i], \varepsilon) + C(t)v[t] \quad (1.2)$$

$$x[t_*] = x_*, \quad \tau_i \leq t < \tau_{i+1}, \quad i = 0, \dots, m-1$$

is said to be a motion $x[\cdot] = x[\cdot, t_*, x_*, u(\cdot), v[\cdot], \varepsilon, \Delta_\delta] = \{x[t, t_*, x_*, u(\cdot), v[\cdot], \varepsilon, \Delta_\delta], t_* \leq t < \vartheta\}$ generated by the strategy $u(\cdot)$ under these conditions. Here any measurable function with values in \mathcal{Q} can be the realization of the disturbance $v[\cdot] = \{v[t] \in \mathcal{Q}, t_* \leq t < \vartheta\}$.

Assume the control process quality index to be as follows

$$\gamma(x[\cdot], u[\cdot], v[\cdot]) = \int_{t_*}^{\vartheta} R(t, u[t], v[t])dt + |x[\vartheta]|, \quad (1.3)$$

$$R(t, u, v) = \langle \Phi(t)u \cdot u \rangle + \langle \Psi(t)v \cdot v \rangle. \quad (1.4)$$

Here $\Phi(t) = \{\varphi_{ij}(t), i, j = 1, 2, \dots, p\}$, $\Psi(t) = \{\psi_{ij}(t), i, j = 1, 2, \dots, q\}$ are continuous symmetric matrix-valued functions; $|x|$ is the Euclidean norm, $\langle \cdot \cdot \rangle$ is the scalar

product, $u[\cdot] = \{u[t] \in \mathcal{P}, t_* \leq t \leq \vartheta\}$ is a realization of the strategy $u(\cdot)$ along the generated motion $x[\cdot]$.

The value

$$\rho(t_* x_* u(\cdot)) = \overline{\lim}_{\varepsilon \rightarrow 0} \lim_{\delta \rightarrow 0} \sup_{v[\cdot], \Delta_\delta} \gamma(x[\cdot], u[\cdot], v[\cdot]) \quad (1.5)$$

is said to be guaranteed result [3, 4] $\rho(t_* x_* u(\cdot))$ for strategy $u(\cdot)$.

One is to find an optimal guaranteeing control-strategy $u^0(\cdot)$ which provides the minimal guaranteed result

$$\rho^0(t_* x_*) = \rho(t_* x_* u^0(\cdot)) = \min_{u(\cdot)} \rho(t_* x_* u(\cdot)) \quad (1.6)$$

for any possible initial position $\{t_*, x_*\}$.

Thus, the strategy $u^0(\cdot)$ has the following property. For any initial position $\{t_*, x_*\}$ and given $\zeta > 0$ we can find a number $\varepsilon(\zeta) > 0$ and then a function $\delta(\zeta, \varepsilon) > 0$ so that for any motion $x[\cdot] = x[\cdot, t_*, x_*, u^0(\cdot), v[\cdot], \varepsilon, \Delta_\delta]$ the inequality

$$\int_{t_*}^{\vartheta} R(t, u^0[t], v[t]) dt + |x[\vartheta]| \leq \rho^0(t_* x_*) + \zeta \quad (1.7)$$

will be guaranteed whatever the disturbance $v[\cdot]$ is provided, if only $\varepsilon \leq \varepsilon(\zeta)$ and $\delta \leq \delta(\zeta, \varepsilon)$. And no strategy $u(\cdot)$ may ever guarantee the inequality $\gamma(x[\cdot, t_*, x_*, u(\cdot), v[\cdot], \varepsilon, \Delta_\delta], u[\cdot], v[\cdot]) < \rho^0(t_* x_*) - \zeta$ for all the possible realizations of the disturbance $v[\cdot]$ and the divisions Δ_δ with sufficiently small step $\delta > 0$.

The optimal guaranteeing control $u^0(t, x, \varepsilon)$ exists. It is constructed from $\rho^0(t, x)$ in the following manner [2]. In position $\{\tau, x_\tau\}$, $t_* \leq \tau \leq \vartheta$ the desired value $u^0(\tau, x, \varepsilon)$ is determined from the condition

$$\begin{aligned} & \max_{v \in \mathcal{Q}} [\langle (x_\tau - w_\tau) \cdot (B(\tau)u^0(\tau, x_\tau, \varepsilon) + C(\tau)v) \rangle + \\ & + c_\tau R(\tau, u^0(\tau, x_\tau, \varepsilon), v)] = \min_{u \in \mathcal{P}} \max_{v \in \mathcal{Q}} [\langle (x_\tau - w_\tau) \cdot \\ & \cdot (B(\tau)u + C(\tau)v) \rangle + c_\tau R(\tau, u, v)] \end{aligned} \quad (1.8)$$

where $\{w_\tau, c_\tau\}$ is a concomitant point which satisfies the condition

$$\rho^0(\tau, w_\tau) - c_\tau = \min_{w, c} [\rho^0(\tau, w) - c] \quad (1.9)$$

for $|x_\tau - w|^2 + c^2 \leq \varepsilon(1 + [\tau - t_*]) \exp 2L[\tau - t_*]$.

Here $L = \max \|A(t)\|$ at $t_0 \leq t \leq \vartheta$, $\|A\|$ is the Euclidean norm of the matrix A .

Thus the construction of the optimal guaranteeing control $u^0(t, x, \varepsilon)$ is reduced to the calculation of the function $\rho^0 = \rho^0(t, x)$.

2. We shall describe a procedure of calculating the optimal guaranteed result $\rho^0(t, x)$ (1.6) based on the method of stochastic program synthesis [1, 2].

Consider a w -model [2] corresponding to equation (1.1) and criteria (1.3), (1.4). The state of this model at the current instant of time is characterized by an $(n+1)$ -dimensional vector $y[t] = \{w[t], y_{n+1}[t]\}$ where w is an n -dimensional vector. A variation of the phase vector $y[t]$, $t_* \leq t \leq \vartheta$ in the w -model is defined by the equations

$$\begin{aligned} \dot{w} &= A(t)w + f + C(t)v \\ \dot{y}_{n+1} &= g_{n+1} + \langle \Psi(t)v \cdot v \rangle \end{aligned} \quad (2.1)$$

where the control v is constrained by the condition $v \in \mathcal{Q}$ and the $(n+1)$ -dimensional vector of control $g = \{f, g_{n+1}\}$ is constrained by the condition

$$\begin{aligned} g \in F(t); \quad F(t) &= \overline{\text{co}} \{h : h = \{B(t)u, \\ &\langle \Phi(t)u \cdot u \rangle\}, \quad u \in \mathcal{P} \} \end{aligned} \quad (2.2)$$

Here $\overline{\text{co}} \{ \dots \}$ is a closed convex hull of the set $\{ \dots \}$. As a basic probabilistic element we choose a standard process of Brownian motion $z[t, \omega]$, $t_* \leq t \leq \vartheta$, where ω is an elementary event from the corresponding probabilistic space $\{\Omega, \mathcal{A}, P\}$ ([5], p. 39).

The functions $g(t, \omega) \in F(t)$, $v(t, \omega) \in \mathcal{Q}$, $t_* \leq t \leq \vartheta$, $\omega \in \Omega$, which are non-anticipating with respect to the process $z[t, \omega]$ (i.e. with the respect to the collection of σ -algebras $\{\mathcal{F}_t^z\}$, connected with the process $z[\tau, \omega]$, $t_* \leq \tau \leq t$ ([5], p. 100)) are said to be the non-anticipatory programs $g(t_*[\cdot] \vartheta)$, $v(t_*[\cdot] \vartheta)$. In other words, the non-anticipatory programs are functions of two variables which for almost all ω are defined by the equalities $g(t, \omega) = g_*(t, z(t_*[\cdot]t, \omega))$, $v(t, \omega) = v_*(t, z(t_*[\cdot]t, \omega))$, where the symbol $z(t_*[\cdot]t, \omega)$ at fixed $\omega \in \Omega$ denotes a realization of the Brownian motion $z[\tau, \omega]$ on the segment $t_* \leq \tau \leq t$. Essentially this means that g and v at instant t are determined by the programs $g(t_*[\cdot] \vartheta)$ and $v(t_*[\cdot] \vartheta)$ only on the basis of a history of the Brownian motion $z[\tau, \omega]$, $t_* \leq \tau \leq t$, which has realized to the instant t .

At a given initial position $\{t_*, y_* = \{w_*, y_{n+1,*}\}\}$ a pair of programs $\{g(t_*[\cdot] \vartheta), v(t_*[\cdot] \vartheta)\}$ determines in the w -model (2.1) a random motion $y[t, \omega] = y[t, \omega, t_* y_* g(t_*[\cdot] \vartheta), v(t_*[\cdot] \vartheta)]$ which for $t_* \leq t \leq \vartheta$ is defined as a solution of the integral equation

$$y[t, \omega] = y_* + \int_{t_*}^t \begin{bmatrix} A(\tau)w[\tau, \omega] + f(\tau, \omega) + C(\tau)v(\tau, \omega) \\ g_{n+1}(\tau, \omega) + \langle \Psi(\tau)v(\tau, \omega) \cdot v(\tau, \omega) \rangle \end{bmatrix} d\tau. \quad (2.3)$$

Let us consider the value

$$\rho^*(t_*, y_*) = \sup_{v(t_*[\cdot], \vartheta)} \inf_{g(t_*[\cdot], \vartheta)} M\{y_{n+1}[\vartheta, \omega] + |w[\vartheta, \omega]|\} \quad (2.4)$$

where M is the mathematical expectation. The following statement holds which is proved similarly to the one is justified in [1].

Theorem 2.1. For any possible initial position $\{t_*, x_*\}$ the optimal guaranteed result $\rho^0(t_*, x_*)$ (1.6) is defined by the equality

$$\rho^0(t_*, x_*) = \rho^*(t_*, \{x_*, 0\}). \quad (2.5)$$

The transition to the problem dual to problem (2.4) leads to the following construction. Let an $(n+1)$ -dimensional vector-valued function $r[t, \omega] = \{s[t, \omega], r_{n+1}[t, \omega]\}$, $r[t_*, \omega] = \{s[t_*], r_{n+1}[t_*]\}$ be a non-anticipatory random solution of the diffusion equation

$$dr = \begin{bmatrix} ds \\ dr_{n+1} \end{bmatrix} = \begin{bmatrix} -A'(t)sd t + m(t, \omega)dz[t, \omega] \\ q_{n+1}(t, \omega)dz[t, \omega] \end{bmatrix} \quad (2.6)$$

where $\{m(t, \omega), q_{n+1}(t, \omega)\} = q(t, \omega)$ is a certain $(n+1)$ -dimensional vector-valued function non-anticipatory with respect to the process $z[t, \omega]$, and the prime denotes the transpose. Let $\|y(\cdot)\|_*$ and $\|r(\cdot)\|^*$ be the norms in spaces of $(n+1)$ -dimensional random variables $y(\omega)$ and $r(\omega)$ defined by

$$\|y(\cdot)\|_* = M\{|y_{n+1}(\omega)|\} + (M\{|w(\omega)|^2\})^{1/2} \quad (2.7)$$

$$\|r(\cdot)\|^* = \max \{v \text{raisup } |r_{n+1}(\omega)|, (M\{|s(\omega)|^2\})^{1/2}\} \quad (2.8)$$

where $|x|$ is the Euclidean norm of the vector x as before.

The following statement is valid.

Lemma 2.1. The value $\rho^*(t_*, y_*)$ of (2.4) is equal to the supremum of those number β for which the inequality

$$\begin{aligned} & \sup_{\|r[\vartheta, \cdot]\|^* \leq 1} [\langle r[t_*] \cdot y_* \rangle + M\{\int_{t_*}^{\vartheta} \max_{v \in \mathcal{L}} \min_{u \in \mathcal{P}} [\langle s[\tau, \omega] \cdot \\ & \cdot (B(\tau)u + C(\tau)v) \rangle + r_{n+1}[\tau, \omega] \cdot R(\tau, u, v)] d\tau\} - \\ & - \max_{y(\cdot) \in Y_\beta} M\{\langle r[\vartheta, \omega] \cdot y(\omega) \rangle\} > 0 \end{aligned} \quad (2.9)$$

holds, where Y_β is a set in the space of random variables defined by the relation

$$Y_\beta = \{y(\cdot) : M\{y_{n+1}(\omega) + |w(\omega)|\} \leq \beta\}. \quad (2.10)$$

In inequality (2.9) $r[\vartheta, \cdot] = \{r[\vartheta, \omega], \omega \in \Omega\}$ is an element of the solution $r[t, \cdot]$, $t_* \leq t < \vartheta$ of equation (2.6) generated by initial condition $r[t_*]$ and non-anticipatory function (control). Thus, the initial condition $r[t_*]$ and the non-anticipatory control for the stochastic system (2.6) are the values that are sought for in the dual problem (2.9). The validity of Lemma 2.1 is proved just as it is argued under similar circumstances in the mathematical theory of control [4] on the basis of the theorems on separation of convex sets. However, here the corresponding constructions are considered in a functional space of the random variables $y(\cdot)$ with $\|y(\cdot)\|_*$ (2.7).

From the form of the last term in (2.9)

$$\begin{aligned} & \max_{y(\cdot) \in Y_\beta} M\{\langle r[\vartheta, \omega] \cdot y(\omega) \rangle\} = \\ & = \max_{y(\cdot) \in Y_\beta} M\{r_{n+1}[\vartheta, \omega] \cdot y_{n+1}(\omega) + \langle s[\vartheta, \omega] \cdot w(\omega) \rangle\} \end{aligned} \tag{2.11}$$

and the definition of the set Y_β (2.1) it follows that for maximizing elements $r[\vartheta, \omega] = \{s[\vartheta, \omega], r_{n+1}[\vartheta, \omega]\}$ at almost all ω the inequality $|s[\vartheta, \omega]| \leq 1$ must be realized. Indeed, if on the set $\mathcal{E} \in \Omega$ of non-zero measure the inequality $|s[\vartheta, \omega]| > 1$ held, then, since due to (2.8) we have $\text{vraisup } r_{n+1}[\vartheta, \omega] \leq 1$, the values $w(\omega)$ and $y_{n+1}(\omega)$ on the set \mathcal{E} could have been chosen such that the value (2.11) would become as large as possible. Hence, the elements $r[\vartheta, \omega]$ for which $|s[\vartheta, \omega]| > 1$ on \mathcal{E} cannot be considered as the maximizing ones for (2.9).

A further analysis of inequality (2.9) shows that the dual problem (2.9) is reduced to the case $r_{n+1}[\tau, \omega] \equiv 1$, $t_* \leq \tau \leq \vartheta$ (in equation (2.6) $q_{n+1}(t, \omega) \equiv 0$, $r_{n+1}[t_*] = 1$),

$\max_{y(\cdot) \in Y_\beta} M\{\langle r[\vartheta, \omega] \cdot y(\omega) \rangle\} = \beta$ and then $\rho^*(t_*, y_*)$ (2.4) is determined by the equality

$$\begin{aligned} \rho^*(t_*, y_*) = & \sup_{\text{vraisup } |s[\vartheta, \omega]| \leq 1} [\langle s[t_*] \cdot w_* \rangle + y_{n+1} \cdot + \\ & + M\{\int_{t_*}^{\vartheta} \max_{v \in \mathcal{Q}} \min_{u \in \mathcal{P}} [\langle s[\tau, \omega] \cdot (B(\tau)u + C(\tau)v) \rangle + \\ & + R(\tau, u, v)] d\tau\}]. \end{aligned} \tag{2.12}$$

Now let us make an assumption about the quadratic form $R(t, u, v)$. We shall accept that the maximin problem under the integral in (2.12) has a solution $\{u_0^0(\tau, \omega), v_0^0(\tau, \omega)\}$, which coincides with a solution $\{u^0(\tau, \omega), v^0(\tau, \omega)\}$ of the same maximin problem, but already without the a priori restrictions $u \in \mathcal{P}$, $v \in \mathcal{Q}$. This assumption essentially simplifies the further calculations and facilitates the obtaining of an effective solution. The assumption is fulfilled if in (1.4) at $t_* \leq \tau \leq \vartheta$ the matrix $\Phi(\tau)$

is positive definite, the matrix $\Psi(\tau)$ is negative definite and the absolute values of the determinants $|\Phi(\tau)|$ and $|\Psi(\tau)|$ are not less than a positive number N which is sufficiently large in comparison with the sizes of \mathcal{P} and \mathcal{Q} . The solution $\{u^0(\tau, \omega), v^0(\tau, \omega)\}$ of the maximin problem over u and v in (2.12) without the a priori restrictions on u and v has the form

$$u^0(\tau, \omega) = -\frac{1}{2} \Phi^{-1}(\tau) B'(\tau) s[\tau, \omega], \quad (2.13)$$

$$v^0(\tau, \omega) = -\frac{1}{2} \Psi^{-1}(\tau) C'(\tau) s[\tau, \omega]$$

and after the substitution of (2.13) into (2.12) we have for $\rho^*(t_*, y_*)$ (2.4)

$$\begin{aligned} \rho^*(t_*, y_*) = & \sup_{\text{vraisup } |s[\vartheta, \omega]| \leq 1} [\langle s[t_*] \cdot w_* \rangle + y_{n+1} + \\ & + \int_{t_*}^{\vartheta} M\{G(\tau, s[\tau, \omega])\} d\tau] \end{aligned} \quad (2.14)$$

where $G(\tau, s)$ is a quadratic form

$$G(\tau, s) = \langle K(\tau) s \cdot s \rangle \quad (2.15)$$

$$K(\tau) = -\frac{1}{4} [B(\tau) \Phi^{-1}(\tau) B'(\tau) + C(\tau) \Psi^{-1}(\tau) C'(\tau)]. \quad (2.16)$$

Thus, the problem of calculating the optimal guaranteed result $\rho^0(t_*, x_*)$ according to (2.5), (2.6), (2.14), (2.15) for the assumption made above about the nature of the quadratic form $R(t, u, v)$ (1.4) is reduced to the problem of selecting an optimal program control $m(t, \omega)$ over the stochastic system

$$ds = -A'(t) s dt + m(t, \omega) dz[t, \omega], \quad s[t_*, \omega] = s[t_*] \quad (2.17)$$

under the condition of maximizing (2.14) and under the restriction $\text{vraisup } |s[\vartheta, \omega]| \leq 1$. In such a problem the quantities sought for are the initial condition $s[t_*]$ and the non-anticipatory control $m(\cdot) = \{m(t, \omega), t_* \leq t < \vartheta, \omega \in \Omega\}$. Strictly speaking the question is selecting only a maximizing sequence of controls $m^{(j)}(\cdot)$, $j = 1, 2, \dots$ for (2.14) since we do not state that the supremum in (2.14) is always achieved on some solution $s^0(\cdot)$ of equation (2.17) at a certain optimal control $m^0(\cdot)$.

3. We set $s[\tau, \omega] = X'[\vartheta, \tau]l(\tau, \omega)$, where $X[t, \tau]$ is a fundamental matrix of solutions for the equation $\dot{x} = A(t)x$. Thus the problem of the searching the guaranteed result $\rho^0(t_*, x_*)$ with respect to (2.5), (2.14), (2.17) is transformed into the problem

$$\rho^0(t_*, x_*) = \sup_{\text{vraisup } |l[\vartheta, \omega]| \leq 1} [\langle l[t_*] \cdot X[\vartheta, t_*]x_* \rangle + \int_{t_*}^{\vartheta} M\{G_*(\tau, l(\tau, \omega))\} d\tau \quad (3.1)$$

$$dl = a(t, \omega)dz[t, \omega], \quad l[t_* \omega] = l[t_*] \quad (3.2)$$

where $a(t, \omega) = [X'[\vartheta, t]]^{-1}m(t, \omega)$ and

$$G_*(\tau, l) = \langle K_*(\tau)l \cdot l \rangle, \quad K_*(\tau) = X[\vartheta, \tau]K(\tau)X'[\vartheta, \tau]. \quad (3.3)$$

Applying Itô's formula of interchanging the variables ([5], p. 141) at fixed τ to the quadratic form $G_*(\tau, l(\xi, \omega))$, $t_* \leq \xi \leq \tau$ with a following averaging we have

$$M\{G_*(\tau, l(\tau, \omega))\} = G_*(\tau, l[t_*]) + \int_{t_*}^{\tau} M\{G_*(\tau, a(\xi, \omega))\} d\xi. \quad (3.4)$$

Substituting (3.4) into (3.1) and changing the order of integration over τ and ξ , we finally find that

$$\rho^0(t_*, x_*) = \sup_{\text{vraisup } |l[\vartheta, \omega]| \leq 1} [\langle l[t_*] \cdot X[\vartheta, t_*]x_* \rangle + H(t_*, l[t_*]) + \int_{t_*}^{\vartheta} M\{H(\tau, a(\tau, \omega))\} d\tau \quad (3.5)$$

where

$$H(\tau, a) = \langle \Gamma(\tau)a \cdot a \rangle, \quad \Gamma(\tau) = \int_{\tau}^{\vartheta} K_*(\xi)d\xi. \quad (3.6)$$

Thus, the problem of calculating $\rho^0(t_*, x_*)$ proves to be equivalent to the problem of determining the maximizing sequence $a^{(j)}(\cdot) = \{a^{(j)}(t, \omega), t_* \leq t < \vartheta, \omega \in \Omega\}$, $j = 1, 2, \dots$ for the non-anticipatory control $a(\cdot)$ and the initial condition $l^0[t_*]$ for the stochastic system (3.2) under the condition of maximizing the value (3.5).

The solution of this problem (3.5), (3.6), (3.2) is as follows. If the quadratic form $H(\tau, a)$ at all $\tau \in [t_*, \vartheta]$ is of negative sign, the trivial control $a(t, \omega) \equiv 0$ will be optimal

and then $l[\tau, \omega] \equiv l[t_*]$ at all $\tau \in [t_* \vartheta]$ and $\omega \in \Omega$. This situation corresponds to the regular case [4]. Here the problem of calculating $\rho(t_*, x_*)$ is solved on the basis of a deterministic program construction, i.e. it is reduced to the search of the maximizing vector $l^0 = l^0[t_*]$ at $|l[t_*]| \leq 1$ from the condition (3.5) where we set $a(\tau, \omega) \equiv 0$.

However, suppose that the quadratic form $H(\tau, a)$ (3.6) for some $\tau \in [t_* \vartheta]$ and some a may attain positive values. In addition let $\tau_* = \tau_*[t_*] \in [t_* \vartheta]$ be such a τ for which the maximal eigenvalue $\lambda(\tau)$ of the quadratic form $H(\tau, a)$ achieves its maximum, i.e.

$$\lambda(\tau_*[t_*]) = \max_{t_* \leq \tau \leq \vartheta} \lambda(\tau), \quad \lambda(\tau) = \max_{|a| \leq 1} H(\tau, a). \quad (3.7)$$

Then problem (3.5) with the restriction

$$\text{vraisup } |l[\vartheta, \omega]| = \text{vraisup } |l[t_*] + \int_{t_*}^{\vartheta} a(t, \omega) dz[t, \omega]| \leq 1 \quad (3.8)$$

has the following maximizing sequence $a^{(j)}(\cdot)$. We note that $\tau_* < \vartheta$ since $\lambda(\vartheta) = 0$, and set the sequence $\tau_j = \tau_* + \varepsilon_j \in [t_* \vartheta]$, $\varepsilon_j > 0$, $\lim \varepsilon_j = 0$ as $j \rightarrow \infty$. Suppose that $a^{(j)}(t, \omega) \equiv 0$ at $t_* \leq t < \tau_*$ and $\tau_* + \varepsilon_j \leq t < \vartheta$. For $\tau_* \leq t < \tau_* + \varepsilon_j$ the control $a^{(j)}(t, \omega)$ is chosen in the following manner. We consider the corresponding maximizing sequence $l^{(j)}[t_*]$. Choosing from it a convergent subsequence (denoted as the same sequence $l^{(j)}[t_*]$), we further denote $\lim l^{(j)}[t_*] = l^0[t_*]$ as $j \rightarrow \infty$. In the considered case $|l^0[t_*]| < 1$. Suppose that e is a unit eigenvector which, according to (3.1), satisfies the condition

$$H(\tau_* e) = \lambda(\tau_*). \quad (3.9)$$

Let us determine the positive numbers a^+ and a^- from the conditions

$$|l^0[t_*] + a^+ e| = 1, \quad |l^0[t_*] - a^- e| = 1. \quad (3.10)$$

We shall divide all the space Ω into two subsets, B_j^+ and B_j^- , in the following manner

$$\begin{aligned} B_j^+ &= \{\omega : z[\tau_* + \varepsilon_j, \omega] - z[\tau_*, \omega] \geq z_*\} \\ B_j^- &= \{\omega : z[\tau_* + \varepsilon_j, \omega] - z[\tau_*, \omega] < z_*\} \end{aligned} \quad (3.11)$$

where a number z_* is chosen from the condition

$$\frac{1}{\sqrt{2\pi\varepsilon_j}} \int_{z_*}^{\infty} e^{-\xi^2/2\varepsilon_j} d\xi = a^- / (a^+ + a^-).$$

We have $P(B_j^+) = a^- / (a^+ + a^-)$, $P(B_j^-) = a^+ / (a^- + a^+)$. Let us choose the vector-valued random variable

$$\begin{aligned} l^{(j)}(\omega) &= l^0[t_*] + a^+ e, & \omega \in B_j^+ \\ l^{(j)}(\omega) &= l^0[t_*] - a^- e, & \omega \in B_j^- \end{aligned} \quad (3.12)$$

We have $M\{l^{(j)}(\omega)\} = l^0[t_*]$. According to ([5], p. 186) a random process $l^{(j)}[t, \omega]$ exists which complies with conditions (3.2) and the condition

$$l^{(j)}[\vartheta, \omega] = l^{(j)}(\omega). \quad (3.13)$$

In addition, for the corresponding control $a^{(j)}(t, \omega)$ the equality

$$M\{|l^{(j)}[\vartheta, \omega]|^2\} = |l^0[t_*]|^2 + \int_{t_*}^{t_* + \varepsilon_j} M\{|a^{(j)}[\tau, \omega]|^2\} d\tau = 1 \quad (3.14)$$

holds.

Hence it is implied that the control $a^{(j)}[\tau, \omega]$ gives the value

$$\begin{aligned} \varkappa^{(j)} &= \langle l^0[t_*] \cdot X[\vartheta, t_*]x_* \rangle + H(t_*, l^0[t_*]) + \\ &+ \lambda(\tau_*[t_*])(1 - |l^0[t_*]|^2) + \varepsilon_j \varphi_j \end{aligned} \quad (3.15)$$

where $\lim \varphi_j = 0$ as $j \rightarrow \infty$.

Thus, for the constructed sequence $a^{(j)}(\cdot)$, $j = 1, 2, \dots$ the relation

$$\begin{aligned} \lim_{j \rightarrow \infty} \varkappa^{(j)} &= \langle l^0[t_*] \cdot X[\vartheta, t_*]x_* \rangle + \\ &+ H(t_*, l^0[t_*]) + \lambda(\tau_*[t_*])(1 - |l^0[t_*]|^2) = \varkappa^0 \end{aligned} \quad (3.16)$$

is valid.

The value \varkappa^0 from (3.16) is just equal to the supremum in (3.5). Indeed, to be convinced for a given $l^0[t_*]$ that any possible sequence $a^{(j)}(\cdot)$ cannot give for (3.5) a value larger than \varkappa^0 in (3.16), it is sufficient to notice that under a broader condition (3.14), it is still impossible to construct a sequence $a^{(j)}(\cdot)$, $j = 1, 2, \dots$ that gives a value greater than \varkappa^0 in (3.16). The latter statement is directly verified with the aid of (3.5) and (3.14).

Thus, the optimal guaranteed result $\rho^0(t_*, \tau_*)$ is determined by the equality

$$\begin{aligned} \rho^0(t_*, x_*) &= \max_{|l| \leq 1} [\langle l \cdot X[\vartheta, t_*]x_* \rangle + H(t_*, l) - \\ &- \lambda(\tau_*[t_*])|l|^2] + \lambda(\tau_*[t_*]) \end{aligned} \quad (3.17)$$

where with regard to (2.16), (3.3), (3.6)

$$\begin{aligned}
 H(t_*, l) &= \langle \Gamma(t_*) l \cdot l \rangle, \\
 \Gamma(t_*) &= -\frac{1}{4} \int_{t_*}^{\theta} X[\vartheta, \tau] \cdot (B(\tau)\Phi^{-1}(\tau)B'(\tau) + \\
 &+ C(\tau)\Psi^{-1}C'(\tau)) \cdot X'[\vartheta, \tau] d\tau
 \end{aligned} \tag{3.18}$$

and the number $\lambda(\tau_*[t_*])$ is determined from (3.7).

In conclusion we shall notice that at $\lambda(\tau_*[t_*]) > 0$ the maximum in (3.5) is not achieved on some admissible control $a(\cdot)$. From the construction of the maximizing sequence $a^{(j)}(t, \omega)$ presented above it is obvious that in this case the choice of some process $z[t, \omega]$ with independent increments and possible discontinuities as a basic probabilistic element corresponds to the spirit of the problem more than a Brownian process.

However, such a choice could lead to more tedious intermediate calculations.

4. The formulation of the original problem (1.1)–(1.6) does not exclude that the disturbance $v[\cdot]$ in the object (1.1) may be, in particular, one or another realization of some control v , formed by a feedback principle. Therefore calling the function $v(\cdot) = \{v(t, x, \varepsilon) \in \mathcal{Q}\}$ to be a strategy of control v we can set a problem of selection of an optimal strategy $v^0(\cdot)$ which for any possible initial position $\{t_*, x_*\}$ ensures the maximal guaranteed result

$$\rho_0(t_*, x_*) = \rho(t_*, x_*, v^0(\cdot)) = \max_{v(\cdot)} \rho(t_*, x_*, v(\cdot)) \tag{4.1}$$

where $\rho(t_*, x_*, v(\cdot))$ is the guaranteed result for the chosen strategy $v(\cdot)$, defined similarly to (1.5) with obvious changes. In addition, the motion $x[\cdot]$ generated by the strategy $v(\cdot)$ is formed here as a solution of the corresponding stepwise equation according to the scheme similar to (1.2).

It is known ([2], p. 581) that the strategy $v^0(\cdot)$ exists and $\rho_0(t_*, x_*)$ in (4.1) coincides with $\rho^0(t_*, x_*)$ from (1.6) which, as it was shown above, is calculated according to (3.17). The optimal guaranteeing control $v^0(t, x, \varepsilon)$ is determined from $\rho_0(t_*, x_*) = \rho^0(t_*, x_*)$ (3.17) just as the control $u^0(t, x, \varepsilon)$ by means of an extremal guidance of the object at a concomitant point obtained from condition (1.9) in which min is now replaced by max.

Thus, side by side with the inequality

$$\gamma(x[\cdot], u^0[\cdot], v[\cdot]) \leq \rho^0(t_*, x_*) + \zeta \tag{4.2}$$

which is ensured by the strategy $u^0(\cdot)$ the following inequality holds

$$\gamma(x[\cdot], u[\cdot], v^0[\cdot]) \geq \rho^0(t_*, x_*) - \zeta \quad (4.3)$$

which is ensured by the strategy $v^0(\cdot)$ for any measurable realization $u[\cdot] = \{u[t] \in \mathcal{P}, t_* \leq t \leq \vartheta\}$ if only $\varepsilon \leq \varepsilon(\zeta)$ and $\delta \leq \delta(\zeta, \varepsilon)$.

But it follows from the results of the present paper that the control v , guaranteeing a result not less than $\rho^0(t_*, x_*) - \zeta$ can be constructed for the given initial position $\{t_*, x_*\}$ as a two-step $Q\{t_*, x_*\}$ -procedure [2]. Namely, on the first step $[t_*, \tau_* = \tau_*[t_*]]$ this procedure designates the control v as a function of time according to the formula $v^0(\tau) = (-1/2)\Psi^{-1}(\tau)C(\tau)X'[\vartheta, \tau]l^0[t_*]$. On the second step $[\tau_*, \vartheta]$ subject to what position $\{t_*, x[\tau_*]\}$ has been realized, one of two controls $v^0(\tau) = (-1/2)\Psi^{-1}(\tau)C(\tau)X'[\vartheta, \tau] \cdot (l^0[t_*] + a^+ e)$ or $v^0(\tau) = (-1/2)\Psi^{-1}(\tau)C(\tau)X'[\vartheta, \tau] \cdot (l^0[t_*] - a^- e)$, where a^+ and a^- are determined by the condition (3.10), and the unit eigenvector e is defined by the condition (3.9).

Both mentioned methods of constructing the optimal guaranteeing control v^0 are based on the feedback principle. However, the results connected with the program stochastic construction (2.3), (2.4) give ground for an attempt of constructing in the real object (1.1) of an optimal guaranteeing control for the given initial position $\{t_*, x_*\}$ as a maximizing non-anticipatory program $v^0(t_*[\cdot], \vartheta)$ in (2.4) (or a non-anticipatory program, approximating the optimal program $v^0(t_*[\cdot], \vartheta)$). We notice that in constructing of the control $v^0[t, \omega]$ on the basis of the program $v^0[t_*[\cdot], \vartheta]$, information on the realization of the Wiener process $z(t_*[\cdot], t, \omega)$ is only used. We can draw this information from a certain source of random processes which is not connected with the evolution of object (1.1).

It follows from Theorem 2.1 that by whatever method, independent of the future realization of the Wiener process, the control ϑ is formed, the non-anticipatory program $v^0(t_*[\cdot], \vartheta)$ (or a program approximating it) guarantees the fulfilment of the inequality

$$M\{\gamma(x[\cdot, \omega], v^0[\cdot, \omega], u[\cdot, \omega])\} \geq \rho^0(t_*, x_*) - \zeta. \quad (4.4)$$

Such a statistical estimation takes place if, for instance, $u[\cdot, \omega]$ are realizations of any positional strategy $u(\cdot)$ or $u[\cdot, \omega]$ are realizations of any non-anticipatory program $u(t_*[\cdot], \vartheta)$ on the same process $z[t, \omega]$ or $u[\cdot, \omega^*]$ are realizations of a certain stochastic program over random events independent of $z[t, \omega]$, etc.

If the controls $u[\cdot, \omega] = u^0[\cdot, \omega]$ are generated by the optimal strategy $u^0(\cdot)$, then side by side with (4.4) the inequality (4.2) will be fulfilled with the probability 1 at $v[\cdot] = v^0[\cdot, \omega]$. Hence, it follows that in such a case for any numbers $\eta > 0$ and $\alpha > 0$ as small as desired, and for the given initial position $\{t_*, x_*\}$ we can indicate the stochastic non-anticipatory program $v^0(t_*[\cdot], \vartheta)$ which guarantees the fulfilment of the inequality

$$P(\gamma(x[\cdot, \omega], u^0[\cdot, \omega], v^0[\cdot, \omega]) \geq \rho^0(t_*, x_*) - \eta) \geq 1 - \alpha.$$

Thus, we have the following fact. If the control u is formed in an optimal manner, then with probability arbitrary close to 1 a result as close to $\rho^0(t_*, x_*)$ as desired may be guaranteed by forming the control v not according to the feedback principle but only on the basis of the information on the realization of the random process $z[t, \omega]$, independent of the object.

5. As an illustration we consider two model examples. The first one bears a formal character. Let system (1.1) and functional (1.3) have the form

$$\dot{x} = 2(u + \sqrt{1 - e^{-0.1t}} \sin t v), \quad x[t_*] = x_*$$

$$\gamma = \int_{t_*}^{\vartheta} (|u|^2 - |v|^2 dt + |x[\vartheta]|)$$

where x, u, v are n -dimensional vectors, $\vartheta = 5\pi$, $0 \leq t_* \leq \vartheta$.

With the aid of (3.18) we find that

$$\Gamma(t_*) = -\frac{1}{1.01} [e^{-0.1t_*} (\cos t_* + 0.1 \sin t_*) + e^{-0.5\pi}] \cdot E$$

where E is the unit matrix.

From (3.7) we obtain that

$$\lambda(\tau) = -\frac{1}{1.01} [e^{-0.1\tau} (\cos \tau + 0.1 \sin \tau) + e^{-0.5\pi}] \quad (5.1)$$

and therefore

$$\lambda(\tau_*[t_*]) = \begin{cases} \lambda(\pi), & 0 \leq t_* \leq \pi \\ \lambda(t_*), & \pi < t_* \leq \tau_1 \\ \lambda(3\pi), & \tau_1 < t_* \leq 3\pi \\ \lambda(t_*), & 3\pi < t_* \leq \tau_2 \end{cases} \quad (5.2)$$

where the number $\tau_1 \in (\pi, 2\pi)$ satisfies the condition $e^{-0.1\tau_1} (\cos \tau_1 + 0.1 \sin \tau_1) = \sim e^{-0.3\pi}$, and the number $\tau_2 \in (3\pi, 4\pi)$ satisfies the condition $e^{-0.1\tau_2} (\cos \tau_2 + 0.1 \sin \tau_2) = e^{-0.5\pi}$.

Corresponding with (5.1) for all $\tau \in (\tau_2, 5\pi)$ the quadratic form $H(\tau, a) = \langle \Gamma(\tau)a \cdot a \rangle = \lambda(\tau)|a|^2$ is of the negative sign and, hence, $\rho^0(t_*, x_*)$ is calculated in this case from (3.17), where we put $\lambda(\tau_*[t_*]) = 0$.

Thus, in the considered example we have

$$\rho^0(t_*, x_*) = \max_{|l| \leq 1} [\langle l \cdot x_* \rangle + (\lambda(t_*) - \lambda(\tau_*[t_*]))|l|^2] + \lambda(\tau_*[t_*]) \quad (5.3)$$

where $\lambda(\tau_*[t_*])$ for $0 \leq t_* \leq \tau_2$ is determined by (5.2) and for $t_* \in (\tau_2, 5\pi)$ we set $\lambda(\tau_*[t_*]) = 0$. Solving the problem (5.3) we find that

$$\rho^0(t_*, x_*) = \begin{cases} \frac{|x_*|^2}{4(\lambda(\pi) - \lambda(t_*))} + \lambda(\pi), & 0 \leq t_* \leq \pi, \\ |x_*| \leq 2(\lambda(\pi) - \lambda(t_*)) \\ \frac{|x_*|^2}{4(\lambda(3\pi) - \lambda(t_*))} + \lambda(3\pi), & \tau_1 < t_* \leq 3\pi, \\ |x_*| \leq 2(\lambda(3\pi) - \lambda(t_*)) \\ -\frac{|x_*|^2}{4\lambda[t_*]}, & \tau_2 < t_* \leq 5\pi, \\ |x_*| \leq -2\lambda(t_*) \end{cases}$$

and $\rho^0(t_*, x_*) = |x_*| + \lambda(t_*)$ for all other positions.

The contents of the second example is as follows. Let a mechanical object of the mass $m = 1$ move along a horizontal axis x_1 under the action of a controlling force u , developed by an electric motor. It is required to transfer the object from some arbitrary position $\{t_* \geq 0, x_1[t_*] = x_{1*}, \dot{x}_1[t_*] = x_{2*}\}$ to a state $x_1[\vartheta] = \dot{x}_1[\vartheta] = 0$ during the time $T_* = \vartheta - t_*$.

During the process of control, an amount of energy costing

$$W_u = \int_{t_*}^{\vartheta} \frac{1}{\vartheta - t} u^2[t] dt \quad (5.4)$$

is spent to fulfil the task, and for the inaccurate fulfilment of the task the fine $(x_1^2[\vartheta] + \dot{x}_1^2[\vartheta])^{1/2}$ is imposed.

Suppose that in addition an uncertain disturbance v acts on the object. However, this disturbance allows us to generate the power which is received by the electric motor that develops the control effort u . The cost of this power is estimated by

$$W_v = \frac{1}{2} \int_{t_*}^{\vartheta} v^2[t] dt. \quad (5.5)$$

Then it is natural to evaluate the cost of the task by the value

$$\gamma = W_u - W_v + (x_1^2[\vartheta] + \dot{x}_1^2[\vartheta])^{1/2}. \quad (5.6)$$

Writing an equation of the object motion and treating γ in (5.6) as a total expense of the means to fulfil the task we come to the studied problem of control with optimal

guaranteed result for the case when system (1.1) and functional (1.3) have the form

$$\begin{aligned} \dot{x}_1 &= x_2, & \dot{x}_2 &= u + v \\ \gamma &= \int_{t_*}^{\vartheta} \left[\frac{1}{\vartheta - t} u^2 - \frac{1}{2} v^2 \right] dt + |x[\vartheta]|. \end{aligned} \quad (5.7)$$

To be specific, take $\vartheta = 3$. For the calculation of the optimal guaranteed result $\rho^0(t_*, x_*)$ (here $\rho^0(t_*, x_*)$ corresponds to a minimal guaranteed expense of the means) it is necessary to use (3.17) again, in which now

$$\lambda(\tau_*, [t_*]) = \begin{cases} \lambda(1), & 0 \leq t_* \leq 1 \\ \lambda(t_*), & 1 < t_* \leq \vartheta = 3 \end{cases} \quad (5.8)$$

where $\lambda(\tau) = (-1/8)[\gamma_{11}(T) + \gamma_{22}(T) - \{[\gamma_{11}(T) - \gamma_{22}(T)]^2 + 4\gamma_{12}^2(T)\}^{1/2}]$,
 $T = \vartheta - \tau$, $\gamma_{11}(T) = T^4/4 - 2T^3/3$, $\gamma_{12}(T) = T^3/3 - T^2$, $\gamma_{22}(T) = T^2/2 - 2T$.

The quadratic form $H(t_*, l)$ for the given example according to (3.18) has the form

$$H(t_*, l) = -\frac{1}{4} [\gamma_{11}(T_*)l_1^2 + 2\gamma_{12}(T_*)l_1l_2 + \gamma_{22}(T_*)l_2^2]. \quad (5.9)$$

Problem (3.17) at $\lambda(\tau_*, [t_*])$ (5.8) and $H(t_*, l)$ (5.9) is not a too complicated problem of quadratic programming, which can be solved with the attraction of modest computing means, and, for instance, at $t_* = 0$, $x_{2*} = 0$, i.e. when the object begins its motion from the state of rest the simple formula holds

$$\rho^0(t_*, x_*) = \begin{cases} \frac{4x_{1*}^2}{(9 + 16\lambda(1))} + \lambda(1), & |x_{1*}| \leq \frac{1}{8}(9 + 16\lambda(1)) \\ |x_{1*}| - 9/16, & |x_{1*}| > \frac{1}{8}(9 + 16\lambda(1)) \end{cases}$$

where $\lambda(1) = 0.76$.

References

1. *Krasovskii, N. N., Tret'jakov, V. E.*, A stochastic program synthesis for a positional differential game. Dokl. Akad. Nauk SSSR, 1981, Vol. **259**, No. *1*, pp. 24–27 (in Russian).
2. *Krasovskii, A. N., Krasovskii, N. N., Tret'jakov, V. E.*, A stochastic program synthesis for a determinate positional differential game. Prikl. Mat. Meh., 1981, Vol. **45**, No. *4*, pp. 581–588 (in Russian).
3. *Pontrjagin, L. S.*, On the theory of differential games. Uspehi Mat. Nauk, 1966, Vol. **21**, No. *4*, pp. 219–274 (in Russian); English transl. in Russian Math. Surveys **21** (1966).
4. *Krasovskii, N. N., Subbotin, A. I.*, Positional differential games, "Nauka", Moscow, 1974, p. 455 (in Russian); French transl. Jeux différentiels, "Mir", Moscow, 1977.
5. *Liptser, R. Sh., Shirjaev, A. N.*, Statistics of random processes, "Nauka", Moscow, 1974, p. 696 (in Russian).

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H-1111 Budapest XI., Stoczek u. 2, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper with wide margins (4-5 cm), should carry the title of the contribution, the author(s) name, and the name of the country. At the end of the typescript the name and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary — possibly in Russian if the paper is in English and *vice-versa* — should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10-15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50-100 слов и приложено резюме-реферат объемом не менее 10-15% объема статьи на русском и на английском языке (в трех экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициях. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 отписок их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Yemelyanov, S. V., Soloviev, A. A.</i> : Application of new feedback types in the problem of signal differentiation (<i>Емельянов С. В., Соловьев А. А.</i> Применение новых типов обратных связей для решения задачи дифференцирования сигнала)	63
<i>Красовский Н. Н., Третьяков В. Е.</i> Стохастический программный синтез одного гарантирующего управления (<i>Krasovskii, N. N., Tretyakov, V. E.</i> : A stochastic program synthesis of a guaranteeing control)	79
<i>Chernousko, F. L.</i> : On equations of ellipsoids approximating reachable sets (<i>Черноусько Ф. Л.</i> Об уравнениях эллипсоидов, аппроксимирующих области достижимости)	97
<i>Lipsey, Zs.</i> : <i>N</i> -person nonlinear qualitative differential games with incomplete information, survey of results (<i>Липсей Ж.</i> Обзор результатов по нелинейным дифференциальным играм качества для <i>N</i> сторон)	111
<i>Novovičová, J.</i> : Robustness of Bayes estimation of dynamic system parameters (<i>Нововичова Я.</i> Робастность байесовского оценивания параметров динамической системы)	123
<i>Yashkov, S. F.</i> : A derivation of response time distribution for a M G 1 processor-sharing queue (<i>Яшков С. Ф.</i> Получение распределения времени отклика для системы M G 1 с разделением процессора)	133
Book Reviews	149

316920

VOL. 12 • NUMBER 3
TOM НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES



PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1983

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H-1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 0BW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1983 Subscription Rate US Dollars 140.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

N. N. KRASOVSKII (USSR)
G. BOGNÁR (HUNGARY)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

E. D. TERYAEV

HUNGARY

T. VÁMOS

L. VARGA

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJC

РЕДАКТОРЫ ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

Г. БОГНАР (ВНР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

Е. Д. ТЕРЯЕВ

ВНР

Т. ВАМОШ

Л. ВАРГА

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

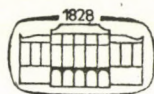
Л. КЕВИЦКИ

Я. КОЧИШ

ЧССР

И. БЕНЕШ

В. СТРЕЙЦ



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

МАСТЯН
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

СВОЙСТВА ДИФФЕРЕНЦИРУЕМОСТИ ФУНКЦИИ ЦЕНЫ ДИФФЕРЕНЦИАЛЬНОЙ ИГРЫ С ИНТЕГРАЛЬНО-ТЕРМИНАЛЬНОЙ ПЛАТОЙ

Н. Н. СУББОТИНА, А. И. СУББОТИН

(Свердловск)

(Поступила в редакцию 13 мая 1982г.)

В работе [1] Н. Н. Красовский и В. Е. Третьяков получили явное выражение для функции цены линейной дифференциальной игры с фиксированным моментом окончания и терминально-интегральной платой. Это выражение выведено методом стохастического программного синтеза [2, 3]. Полученная функция оказывается недифференцируемой. Известно, что необходимые и достаточные условия, которым должна удовлетворять недифференцируемая функция цены, можно определить в форме неравенств для производных по направлениям [4, 5]. В данной работе приведены подобные соотношения для нелинейных дифференциальных игр с терминально-интегральной платой, а для функции цены, определенной в [1], вычислены производные по направлению и показано выполнение упомянутых неравенств.

1. Введение

Рассмотрим дифференциальную игру, в которой динамика управляемого объекта описывается уравнением:

$$\frac{dx(t)}{dt} = A(t)x + B(t)u + C(t)v, \quad (1.1)$$

где $x \in R^n$ — фазовый вектор системы, $A(t)$, $B(t)$, $C(t)$ — непрерывные матрицы размерностей $(n \times n)$, $(n \times p)$, $(n \times q)$; $u \in R^p$, $v \in R^q$ — управления первого и второго игроков соответственно, стесненные геометрическими ограничениями:

$$u \in P, \quad v \in Q \quad (1.2)$$

где P и Q — выпуклые компакты.

Плата в рассматриваемой игре имеет вид:

$$\gamma(x(\cdot), u(\cdot), v(\cdot)) = \int_{t_0}^{\theta} \langle \Phi(t)u(t), u(t) \rangle + \langle \Psi(t)v(t), v(t) \rangle dt + \|x(\theta)\|, \quad (1.3)$$

где ϑ — фиксированный момент окончания игры, $x(\cdot)$ — траектория системы (1.1), выходящая из начальной позиции (t_*, x_*) под воздействием управлений $u(\cdot): [t_*, \vartheta] \rightarrow P$, $v(\cdot): [t_*, \vartheta] \rightarrow Q$. Символ $\|a\|$ обозначает евклидову норму конечного вектора a , символ $\langle a, b \rangle$ обозначает скалярное произведение векторов a и b . Предполагается, что $\Phi(t)$ и $\Psi(t)$ квадратные, симметрические матрицы размерности $(p \times p)$ и $(q \times q)$ соответственно, непрерывно зависящие от t .

Дифференциальная игра (1.1)–(1.3) ставится в классах позиционных стратегий первого и второго игроков [1, 6].

Предполагается также, что

$$\langle \Phi(t)u, u \rangle > 0, \quad \forall (t, u) \in (-\infty, \vartheta] \times R^p, \quad u \neq 0, \quad (1.4)$$

$$\langle \Psi(t)v, v \rangle < 0, \quad \forall (t, v) \in (-\infty, \vartheta] \times R^q, \quad v \neq 0, \quad (1.5)$$

а множества P и Q таковы, что для любого $l \in R^n$ выполняются соотношения

$$\begin{aligned} & \min_{u \in P} [\langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle] = \\ & = \min_{u \in R^p} [\langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle]. \end{aligned} \quad (1.6)$$

$$\begin{aligned} & \max_{v \in Q} [\langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle] = \\ & = \max_{v \in R^q} [\langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle], \end{aligned}$$

где $t \leq \vartheta$, $X[\vartheta, t]$ — фундаментальная матрица решений исходной линейной системы, т.е. $\frac{dX[\vartheta, t]}{dt} = -X[\vartheta, t]A(t)$, $X[\vartheta, \vartheta] = E$.

При указанных предположениях дифференциальная игра (1.1)–(1.3) рассматривалась в [1], где получена следующая формула цены

$$\begin{aligned} \rho^\circ(t_*, x_*) &= \max_{\|l\| \leq 1} \{ \langle l, X[\vartheta, t_*]x_* \rangle + \langle \Gamma(t_*)l, l \rangle - \lambda_{t_*} \|l\|^2 \} + \lambda_{t_*}, \\ & (t_*, x_*) \in (-\infty, \vartheta] \times R^n. \end{aligned} \quad (1.7)$$

Здесь число λ_{t_*} определяется соотношениями

$$\lambda_{t_*} = \max_{t_* \leq t \leq \vartheta} \lambda(t), \quad \lambda(t) = \max_{\|l\|=1} \langle \Gamma(t)l, l \rangle, \quad (1.8)$$

а $\Gamma(t)$ — непрерывно дифференцируемая, симметрическая матрица размерности $(n \times n)$, подобранная таким образом, что

$$-\left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle = \min_{u \in P} [\langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle] + \\ + \max_{v \in Q} [\langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle], \quad l \in R^n, \quad (1.9)$$

$$\Gamma[\vartheta] = [0], \quad (1.10)$$

т.е. при $t = \vartheta$ матрица $\Gamma(\vartheta)$ состоит из нулевых компонент.

Функция цены $\rho^0(\cdot)$ в точках (t, x) , где она дифференцируема, должна удовлетворять так называемому основному уравнению [7], которое в рассматриваемом случае имеет вид

$$\min_{u \in P} \max_{v \in Q} \left[\frac{\partial \rho^0(t, x)}{\partial t} + \left\langle \frac{\partial \rho^0(t, x)}{\partial x} (A(t)x + B(t)u + C(t)v) \right\rangle + \right. \\ \left. + \langle \Phi(t)u, u \rangle + \langle \Psi(t)v, v \rangle \right] = 0, \quad \rho^0(\vartheta, x) = \|x\|, \quad x \in R^n. \quad (1.11)$$

Отметим, что в общем случае функция $\rho^0(\cdot)$ (1.7) дифференцируема не во всякой точке (t, x) . Неовходимое условие (1.11), которому должна удовлетворять функция цены, не является достаточным. Ниже указаны неравенства (2.14), (2.15), образующие совместно с краевым равенством $\rho^0(\vartheta, x) = \|x\|$ необходимые и достаточные условия, которым должна удовлетворять функция цены. В точках (t, x) , где функция $\rho^0(\cdot)$ дифференцируема, неравенства (2.14), (2.15) обращаются в равенство (1.11).

2. Основные неравенства

В этом разделе приведено утверждение, на котором базируется дальнейший анализ функции $\rho^0(\cdot)$. Это утверждение формулируется сначала для случая нелинейной дифференциальной игры, а затем конкретизируется для игры (1.1)–(1.3).

Итак, рассмотрим дифференциальную игру

$$\dot{x} = f(t, x, u, v), \quad u \in P, \quad v \in Q, \quad (2.1)$$

$$\gamma(x(\cdot), u(\cdot), v(\cdot)) = \sigma(x(\vartheta)) + \int_{t_0}^{\vartheta} f_{n+1}(t, x(t), u(t), v(t)) dt, \quad (2.2)$$

где $x \in R^n$, $u \in R^p$, $v \in R^q$; P, Q — компакты, ϑ — фиксированный момент окончания игры, $x(\cdot) = x(\cdot, t_*, x_*, u(\cdot), v(\cdot))$ — движение системы (2.1), порождаемое из позиции $(t_*, x(t_*) = x_*)$ управлениями $u(\cdot): [t_*, \vartheta] \rightarrow P$, $v(\cdot): [t_*, \vartheta] \rightarrow Q$ — первого и второго игроков соответственно. Первый игрок стремится минимизировать значение платы, второй — максимизировать.

Предполагается, что функции $f: (-\infty, \vartheta] \times R^n \times P \times Q \rightarrow R^n$; $f_{n+1}: (-\infty, \vartheta] \times R^n \times P \times Q \rightarrow R$, $\sigma: R^n \rightarrow R$ — непрерывны по совокупности аргументов и удовлетворяют условию Липшица по x . Для любых $(t, x) \in (-\infty, \vartheta] \times R^n$, $(s, s_{n+1}) \in R^n \times R$ предполагается выполненным равенство

$$\begin{aligned} & \min_{u \in P} \max_{v \in Q} [\langle s, f(t, x, u, v) \rangle + s_{n+1} \cdot f_{n+1}(t, x, u, v)] = \\ & = \max_{v \in Q} \min_{u \in P} [\langle s, f(t, x, u, v) \rangle + s_{n+1} \cdot f_{n+1}(t, x, u, v)]. \end{aligned} \quad (2.3)$$

Вводя новую фазовую переменную x_{n+1} и новый фазовый вектор $y = (x, x_{n+1})$, перейдем от игры (2.1), (2.2) к игре

$$\dot{x} = f(t, x, u, v), \quad \dot{x}_{n+1} = f_{n+1}(t, x, u, v), \quad u \in P, \quad v \in Q, \quad (2.4)$$

$$\gamma_*(y(\cdot)) = \gamma_*(x(\cdot), x_{n+1}(\cdot)) = \sigma(x(\vartheta)) + x_{n+1}(\vartheta). \quad (2.5)$$

Известно [6], что для игры (2.4) с терминальной платой (2.5) для каждой начальной позиции $(t_*, y_*) = (t_*, x_*, x_{n+1})$ существует ситуация равновесия

$$\begin{aligned} \inf_U \sup_{y(\cdot) \in Y(t_*, y_*, U)} \gamma_*(y(\cdot)) &= \sup_V \inf_{y(\cdot) \in Y(t_*, y_*, V)} \gamma_*(y(\cdot)) = \\ &= c^0(t_*, y_*) = c^0(t_*, x_*, x_{n+1}). \end{aligned} \quad (2.6)$$

Здесь $Y(t_*, y_*, U)$, $Y(t_*, y_*, V)$ — пучки движений, порожденные позиционными стратегиями U и V соответственно [1, 6]. Заметим, что для цены дифференциальной игры (2.4), (2.5) справедливы равенства

$$c^0(t, x, x_{n+1}) = c^0(t, x, 0) - x_{n+1}, \quad (2.7)$$

$$c^0(t, x, 0) = \rho^0(t, x), \quad (2.8)$$

где $\rho^0(t, x)$ — цена дифференциальной игры (2.1), (2.2).

Символом $D\rho(t, x)|(1, f)$ будем обозначать производную функции $(t, x) \rightarrow \rho(t, x)$ в точке (t, x) по направлению $(1, f) \in R^{n+1}$, т.е.

$$D\rho^0(t, x)|(1, f) = \lim_{\delta \rightarrow +0} [\rho(t + \delta, x + \delta f) - \rho(t, x)]\delta^{-1}. \tag{2.9}$$

Введем в рассмотрение класс Dif функций $(t, x) \rightarrow \rho(t, x)$, локально-липшицевых и дифференцируемых в каждой точке $(t, x) \in (-\infty, \vartheta) \times R^n$ по любому направлению $(1, f) \in R^{n+1}$.

Используя результаты работы [5] и соотношения (2.7), (2.8), приходим к следующему утверждению.

Теорема 2.1. Для того, чтобы функция $\rho \in \text{Dif}$ совпадала с функцией цены ρ^0 дифференциальной игры (2.1), (2.2), необходимо и достаточно, чтобы

$$\rho(\vartheta, x) = \sigma(x), \quad x \in R^n, \tag{2.10}$$

$$\begin{aligned} & \max_{v \in Q} \min_{(f, f_{n+1}) \in F_1(t, x, v)} [D\rho(t, x)|(1, f) + f_{n+1}] \leq 0 \leq \\ & \leq \min_{u \in P} \max_{(f, f_{n+1}) \in F_2(t, x, u)} [D\rho(t, x)|(1, f) + f_{n+1}], \\ & t < \vartheta, \quad x \in R^n; \end{aligned} \tag{2.11}$$

где

$$\begin{aligned} F_1(t, x, v) &= \overline{\text{co}} \{(f(t, x, u, v), f_{n+1}(t, x, u, v)) : u \in P\}, \\ F_2(t, x, u) &= \overline{\text{co}} \{(f(t, x, u, v), f_{n+1}(t, x, u, v)) : v \in Q\}. \end{aligned} \tag{2.12}$$

Это утверждение для игры (1.1)–(1.3) принимает следующий вид.

Теорема 2.2. Для того, чтобы функция $\rho \in \text{Dif}$ была функцией цены ρ^0 дифференциальной игры (1.1)–(1.3), необходимо и достаточно, чтобы

$$\rho(\vartheta, x) = \|x\|, \quad x \in R^n, \tag{2.13}$$

$$\begin{aligned} & \max_{v \in Q} \min_{(f, \varphi) \in F_1(t)} [D\rho(t, x)|(t, x)|(1, A(t)x + f + C(t)v) + \varphi + \\ & + \langle \Psi(t)v, v \rangle] \leq 0, \end{aligned} \tag{2.14}$$

$$\begin{aligned} 0 & \leq \min_{u \in P} \max_{(f, \Psi) \in F_2(t)} [D\rho(t, x)|(1, A(t)x + B(t)u + f) + \\ & + \langle \Phi(t)u, u \rangle + \Psi], \end{aligned} \tag{2.15}$$

где

$$(t, x) \in (-\infty, \vartheta) \times R^n,$$

$$F_1(t) = \overline{\text{co}} \{(B(t)u, \langle \Phi(t)u, u \rangle) : u \in P\},$$

$$F_2(t) = \overline{\text{co}} \{(C(t)v, \langle \Psi(t)v, v \rangle) : v \in Q\}. \quad (2.16)$$

3. Свойства дифференцируемости функции $\rho^0(\cdot)$ (1.7)

Проверим выполнение для функции $\rho^0(\cdot)$ (1.7) условий (2.13)–(2.15).

Краевое условие. Из (1.8), (1.10) вытекает, что $\lambda_\vartheta = \lambda(\vartheta) = 0$: Используя соотношения $X[\vartheta, \vartheta] = E$, $\Gamma[\vartheta] = [0]$, $\lambda_\vartheta = 0$, получаем по формуле (1.7)

$$\rho^0(\vartheta, x) = \max_{\|l\| \leq 1} \langle l, x \rangle = \|x\|, \quad x \in R^n,$$

т.е. краевое условие (2.13) выполняется.

Основные неравенства. Проверка выполнения для функции (1.7) условий (2.14), (2.15) проводится в три этапа: I–III.

I. Вычисление производной $D\rho^0(t, x)|(1, f)$. Рассмотрим сначала функции $t \rightarrow \lambda(t)$, $t \rightarrow \lambda_t$ (см. (1.8)). Символами $\dot{\lambda}(t)$ и $\dot{\lambda}_t$ будем обозначать правые производные этих функций, т.е.

$$\dot{\lambda}(t) = \lim_{\delta \rightarrow +0} [\lambda(t+\delta) - \lambda(t)]\delta^{-1}, \quad \dot{\lambda}_t = \lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t]\delta^{-1}$$

Утверждение 3.1. Для любых $t < \vartheta$ существуют правые производные $\dot{\lambda}(t)$ и $\dot{\lambda}_t$. Причем

$$\dot{\lambda}(t) = \max_{l \in L^*(t)} \frac{d\Gamma(t)}{dt} l, l \quad ; \quad L^*(t) = \{l \in R^n : \|l\| = 1, \quad (3.1)$$

$$\langle \Gamma(t)l, l \rangle = \lambda(t)\};$$

$$\dot{\lambda}_t \leq 0, \quad t < \vartheta. \quad (3.2)$$

Доказательство. Существование производной $\dot{\lambda}(t)$ и формула (3.1) сразу следуют из (1.8) и [8] (Теорема 3.1). Докажем существование $\dot{\lambda}_t$.

В точке t , где $\lambda_t > \lambda(t)$ или $\lambda_t = \lambda(t) = \lambda(t + \delta)$ для некоторого $\delta > 0$, существует $\delta' > 0$ такое, что $\lambda_\tau \equiv \text{const}$ при $\tau \in [t, t + \delta']$ согласно (1.8). Тогда, очевидно, $\dot{\lambda}_t = 0$; в частности, справедливо и (3.2).

Если $\lambda_t = \lambda(t) > \lambda(\tau)$ при всех $\tau \in (t, \vartheta]$, то

$$\max_{\tau \leq \xi \leq \vartheta} \lambda(\xi) = \lambda_\tau < \lambda_t. \quad (3.3)$$

Из (1.8) вытекает, что $\lambda(t) \leq \lambda_\tau$ ($t < \vartheta$), и, как отмечено выше, существует $\dot{\lambda}(t)$. Тогда

$$\lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t] \delta^{-1} \geq \lim_{\delta \rightarrow +0} [\lambda(t+\delta) - \lambda(t)] \delta^{-1} = \dot{\lambda}(t), \quad t < \vartheta. \quad (3.4)$$

Пусть последовательность δ_i ($i=0, 1, \dots$) такова, что $\delta_i > 0$, $\delta_i \rightarrow 0$ при $i \rightarrow \infty$ и

$$\lim_{\delta_i \rightarrow 0} [\lambda_{t+\delta_i} - \lambda_t] \delta_i^{-1} = \overline{\lim}_{\delta \rightarrow 0} [\lambda_{t+\delta} - \lambda_t] \delta^{-1}. \quad (3.5)$$

Согласно (3.3) последовательность δ_i может быть выбрана так, что $\lambda_{t+\delta_i} > \lambda_{t+\delta_{i-1}}$ ($i=1, 2, \dots$). Отсюда, по определению λ_t , следует

$$\begin{aligned} \lambda(t+\delta_i) \leq \lambda_{t+\delta_i} &= \max_{\delta \in [\delta_i, \delta_{i-1}]} \lambda(t+\delta) < \lambda_{t+\delta_{i+1}} = \\ &= \max_{\delta \in [\delta_{i+1}, \delta_i]} \lambda(t+\delta). \end{aligned} \quad (3.6)$$

Из (3.6) вытекает, что для любого δ_i существует $\alpha_i \in (\delta_{i+1}, \delta_i)$, такое, что

$$\lambda(t+\alpha_i) = \max_{\delta \in [\delta_i, \delta_{i-1}]} \lambda(t+\delta) = \lambda_{t+\delta_i}, \quad \alpha_i \leq \delta_i. \quad (3.7)$$

Тогда, очевидно, последовательности δ_i и α_i сразу можно выбрать так, что

$$\lim_{\delta_i \rightarrow 0} [\lambda_{t+\delta_i} - \lambda_t] \delta_i^{-1} \leq \lim_{\alpha_i \rightarrow 0} [\lambda(t+\alpha_i) - \lambda(t)] \alpha_i^{-1} = \dot{\lambda}(t). \quad (3.8)$$

Из (3.4) и (3.5), (3.8) получаем, что существует $\dot{\lambda}_t$ и справедливо (согласно (1.8))

$$\dot{\lambda}_t = \dot{\lambda}(t) \leq 0. \quad (3.9)$$

Следовательно, условие (3.2) справедливо при всех $t < \vartheta$. Утверждение 3.1 доказано полностью.

Введем некоторые обозначения.

$$\begin{aligned} \varphi(t, x, l) &= \langle (\Gamma(t) - \lambda_t E)l, l \rangle + \langle l, X[\vartheta, t]x \rangle + \lambda_t; \\ L &= \{l \in R^n : \|l\| \leq 1\}. \end{aligned} \quad (3.10)$$

Тогда функция $\rho^0(\cdot)$ (1.7) может быть записана с учетом (3.10) в виде

$$\rho^0(t, x) = \max_{l \in L} \varphi(t, x, l).$$

Утверждение 3.2. Для любых $(t, x) \in (-\infty, \vartheta) \times R^n$ и $f \in R^n$ существует производная $D\rho^0(t, x)|(1, f)$, причем при $f = A(t)x + h$ справедливо равенство

$$\begin{aligned} D\rho^0(t, x)|(1, A(t)x + h) &= \\ &= \max_{l \in L^0(t, x)} \left[\left\langle \left(\frac{d\Gamma(t)}{dt} - \lambda_t E \right) l, l \right\rangle + \langle l, h \rangle \right] + \dot{\lambda}_t, \end{aligned} \quad (3.11)$$

где

$$L_0(t, x) = \{l \in R^n : \|l\| \leq 1, \quad \varphi(t, x, l) = \rho^0(t, x)\}. \quad (3.12)$$

Доказательство. Полагаем

$$\varphi^*(\tau, x, l) = \langle (\Gamma(\tau) - \lambda_\tau^* E)l, l \rangle + \langle l, X[\vartheta, \tau]x \rangle + \lambda_\tau^*, \quad (3.13)$$

$$\lambda_\tau^* = \lambda_t + \dot{\lambda}_t(\tau - t), \quad (3.14)$$

$$\rho^*(\tau, x) = \max_{l \in L} \varphi^*(\tau, x, l). \quad (3.15)$$

Заметим, что $\lambda_t^* = \lambda_t$, $\varphi^*(t, x, l) = \varphi(t, x, l)$, $\rho^*(t, x) = \rho^0(t, x)$,

$$\begin{aligned} \lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_{t+\delta}^*] \delta^{-1} &= \lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t - \dot{\lambda}_t \delta] \delta^{-1} = 0, \\ \lim_{\delta \rightarrow +0} [\varphi(t+\delta, x+\delta \cdot f, l) - \varphi^*(t+\delta, x+\delta \cdot f, l)] \delta^{-1} &= 0, \quad \forall l \in L. \end{aligned} \quad (3.16)$$

Поэтому

$$\lim_{\delta \rightarrow +0} [\rho^0(t+\delta, x+\delta f) - \rho^*(t+\delta, x+\delta \cdot f)] \delta^{-1} = 0. \quad (3.17)$$

Из утверждения 3.1, гладкости функций $\tau \rightarrow \Gamma(\tau)$, $\tau \rightarrow X[\vartheta, \tau]$ вытекает, согласно [8] (Теорема 3.1), существование $D\rho^*(t, x)|(1, f)$ ($x \in R^n$, $f \in R^n$), причем

$$\begin{aligned} D\rho^*(t, x)|(1, f) &= \max_{l \in L^0(t, x)} D\varphi^*(t, x, l)|(1, f) = \\ &= \max_{l \in L^0(t, x)} \left[\frac{\partial \varphi^*(t, x, l)}{\partial t} + \left\langle \frac{\partial \varphi^*(t, x, l)}{\partial x} \cdot f \right\rangle \right] = \\ &= \max_{l \in L^0(t, x)} \left[-\langle l, X[\vartheta, t]A(t)x \rangle + \right. \\ &\left. + \left(\frac{d\Gamma(t)}{dt} - \lambda_t E \right) l, l \right] + \dot{\lambda}_t + \langle l, X[\vartheta, t]f \rangle, \end{aligned} \quad (3.18)$$

где $L^0(t, x)$ — множество вида (3.12). Из (3.17), (3.18) следует равенство

$$\begin{aligned} & \lim_{\delta \rightarrow +0} [\rho^0(t + \delta, x + \delta \cdot f) - \rho^0(t, x)] \delta^{-1} = D\rho^0(t, x)|(1, f) = \\ & = \lim_{\delta \rightarrow +0} [\rho^*(t + \delta, x + \delta \cdot f) - \rho^*(t, x)] \delta^{-1} = D\rho^*(t, x)|(1, f). \end{aligned} \quad (3.19)$$

Из (3.18), (3.19) следует (3.11). Утверждение 3.2 доказано.

Заметим, что на любом компакте $K \subset (-\infty, \vartheta) \times R^n$, $\|f\| \leq 1$ производные $D\rho^0(t, x)|(1, f)$ равномерно ограничены. Поэтому справедливо

Следствие 3.2.1. Функция $\rho^0 \in \text{Dif}$, где $\rho^0(\cdot)$ вида (1.7).

II. Свойства множества $L^0(t, x)$. Сведения из линейной алгебры и выпуклого анализа, которые используются в дальнейших рассуждениях, вынесены в приложение.

Утверждение 3.3. При всех $(t, x) \in (-\infty, \vartheta) \times R^n$ множество $L^0(t, x)$ (3.12) — выпуклый компакт.

Доказательство. Пусть $\lambda_i(t)$ и $\mu_i(t)$, $i = \overline{1, n}$ — характеристические числа матриц $\Gamma(t)$ и $\Gamma(t) - \lambda_t E$ соответственно. Нетрудно проверить, что они связаны равенством

$$\mu_i(t) = \lambda_i(t) - \lambda_t, \quad i = \overline{1, n}. \quad (3.20)$$

Матрица $\Gamma(t)$ — действительная, симметрическая, поэтому числа $\lambda_i(t)$, $\mu_i(t)$ — действительные. По определению (1.8) из (3.20) получаем

$$\mu_i(t) \leq \max_{1 \leq i \leq n} \mu_i(t) = \max_{1 \leq i \leq n} (\lambda_i(t) - \lambda_t) = \lambda(t) - \lambda_t \leq 0. \quad (3.21)$$

Из условия (3.21) вытекает, что $\langle (\Gamma(t) - \lambda_t E)l, l \rangle \leq 0$, $l \in R^n$, а функция $l \rightarrow \varphi(t, x, l)$ вида (3.10) вогнута по l и непрерывна при любых $(t, x, l) \in (-\infty, \vartheta) \times R^n \times R^n$. Следовательно, множество $L^0(t, x)$ вида (3.12) — выпуклый компакт.

Утверждение 3.4. Если $\lambda_t = \lambda(t)$, то множество $L^0(t, x)$ содержит вектор l^0 единичной длины.

Доказательство. Пусть $\mu_i(t)$, $i \in \overline{1, n}$ — собственные числа матрицы $H = H(t) = \Gamma(t) - \lambda_t E$. Из (3.11) следует

$$\begin{aligned} \mu_{m+1}(t) = \dots = \mu_n(t) &= \max_{1 \leq i \leq n} \mu_i(t) = \lambda(t) - \lambda_t = 0, \\ & m \in \overline{0, n-1}, \end{aligned} \quad (3.22)$$

$$\mu_i(t) < 0 \quad (i = 1, \dots, m), \quad (3.23)$$

где $(n-m)$ — кратность нулевого собственного числа матрицы H . Матрицу S , удовлетворяющую соотношениям (4.5), (4.4), где $H = \Gamma(t) - \lambda_t E$ обозначим через $S(t)$. Пусть векторы S и ξ связаны равенством $l = S(t)\xi$, тогда из (4.5), (4.4), (3.22), (3.23) следует

$$\langle l, l \rangle = \langle S(t)\xi, S(t)\xi \rangle = \langle S^T(t) S(t)\xi, \xi \rangle = \langle \xi, \xi \rangle, \quad (3.24)$$

$$\begin{aligned} \varphi(t, x, l) &= \varphi(t, x, S(t)\xi) = \langle H(t) S(t)\xi, S(t)\xi \rangle + \langle S(t)\xi, X[\vartheta, t]x \rangle + \\ &+ \lambda_t = \langle (S^T(t) H(t) S(t))\xi, \xi \rangle + \langle \xi, S^T(t) X[\vartheta, t]x \rangle + \lambda_t = \\ &= \sum_{i=1}^m [\mu_i(t)\xi_i^2 + \xi_i y_i] + \sum_{i=m+1}^n \xi_i y_i + \lambda_t, \end{aligned} \quad (3.25)$$

где $y = (y_1, \dots, y_n)^T = S^T(t)X[\vartheta, t]x$ — вектор-столбец.

Если $y_i = 0, i = m+1, n$, то существование $l^0 \in L^0(t, x): \|l^0\| = 1$ очевидно.

Пусть при некотором $j \in m+1, n, y_j \neq 0$. Предположим, что

$$\forall l \in L^0(t, x): \|l^0\| < 1. \quad (3.26)$$

Пусть $\xi^0 = (\xi_1^0, \dots, \xi_n^0): l^0 = S(t)\xi^0, l^0 \in L^0(t, x)$. Из (3.24), (3.26) следует

$$\|\xi^0\| < 1, \quad \sum_{i=1}^m (\xi_i^0)^2 = c < 1, \quad \sum_{i=m+1}^n (\xi_i^0)^2 < 1 - c, \quad (3.27)$$

$$\rho^0(t, x) = \varphi(t, x, S(t)\xi^0) = \max_{\|\xi\| \leq 1} \varphi(t, x, S(t)\xi). \quad (3.28)$$

Составим вектор $\xi_{(n-m)}^0 = (\xi_{m+1}^0, \dots, \xi_n^0) \in R^{n-m}$. Согласно (3.27) точка $\xi_{(n-m)}^0$ является внутренней точкой шара в R^{n-m} радиуса $\sqrt{1-c}$. Линейная форма

$\sum_{i=m+1}^n \xi_i y_i \neq 0$ в R^{n-m} , следовательно, существует $\xi_{(n-m)}^* = (\xi_{m+1}^*, \dots, \xi_n^*)$:

$$\|\xi_{(n-m)}^*\| = \sqrt{1-c} \sum_{i=m+1}^n \xi_i^* y_i > \sum_{i=m+1}^n \xi_i^0 y_i \quad (3.29)$$

Для вектора $\xi^* = (\xi_1^0, \dots, \xi_m^0, \dots, \xi_{m+1}^*, \dots, \xi_n^*)$ из (3.27), (3.29), (3.28) получаем

$$\|\xi^*\| = \left[\sum_{i=1}^m (\xi_i^0)^2 + \sum_{i=m+1}^n (\xi_i^*)^2 \right]^{1/2} = (c + 1 - c)^{1/2} = 1 \quad (3.30)$$

$$\varphi(t, x, S(t)\xi^*) > \varphi(t, x, S(t)\xi^0) = \max_{\|\xi\| \leq 1} \varphi(t, x, S(t)\xi). \quad (3.31)$$

Полученное противоречие опровергает предположение (3.18). Утверждение 3.4 доказано.

III. Проверка условий (2.14), (2.15). Используя последовательно формулы (3.11), (2.16) и условие (1.9) и неравенство $\min \max \geq \max \min$ получим оценку

$$\begin{aligned} & \min_{u \in P} \max_{(f, \Psi) \in F_2(t)} [D\rho^0(t, x)|(1, A(t)x + B(t)u + f) + \langle \Phi(t)u, u \rangle + \Psi] \leq \\ & \geq \min_{u \in P} \max_{v \in Q} \max_{l \in L^0(t, x)} [D\varphi^*(t, x, l)|(1, A(t)x + B(t)u + C(t)v) + \\ & + \langle \varphi(t)u, u \rangle + \langle \Psi(t)v, v \rangle] \leq \max_{l \in L^0(t, x)} \max_{v \in Q} \min_{u \in P} \left\{ \left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle + \right. \\ & + \langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle + \langle l, X[\vartheta, t]C(t)v \rangle + \\ & \left. + \langle \Psi(t)v, v \rangle \right\} - \lambda_t (\langle l, l \rangle - 1) \Big\} = \max_{l \in L^0(t, x)} (-\lambda_t) (\|l\|^2 - 1). \end{aligned} \tag{3.32}$$

Используя тот факт, что функция $l \rightarrow D\varphi^*(t, x, l)|(1, f)$ вогнута по l на выпуклом компакте $L^0(t, x)$, согласно лемме 4.1, а функция $u \rightarrow D\varphi^*(t, x, l)|(1, A(t)x + B(t)u + f) + \langle \Phi(t)u, u \rangle$ — выпукла по u на выпуклом компакте P в силу (3.11) и (1.4) и применяя теорему о минимаксе [10], получим, аналогично (3.32), еще одну оценку

$$\begin{aligned} & \max_{v \in Q} \min_{(f, \varphi) \in F_1(t)} [D\rho^0(t, x)|(1, A(t)x + C(t)v + f) + \langle \Psi(t)v, v \rangle + \varphi] \leq \\ & \max_{v \in Q} \min_{u \in P} \max_{l \in L^0(t, x)} [D\varphi^*(t, x, l)|(1, A(t)x + B(t)u + C(t)v) + \\ & + \langle \Phi(t)u, u \rangle + \langle \Psi(t)v, v \rangle] = \max_{l \in L^0(t, x)} \max_{v \in Q} \min_{u \in P} \left[\left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle + \right. \\ & + \langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle + \langle l, X[\vartheta, t]C(t)v \rangle + \\ & \left. + \langle \Psi(t)v, v \rangle \right] - \lambda_t (\|l\|^2 - 1) = \max_{l \in L^0(t, x)} (-\lambda_t) (\|l\|^2 - 1). \end{aligned} \tag{3.33}$$

В случае $\lambda(t) = \lambda_t$ согласно утверждению 3.4 и условию (3.2) получаем

$$\max_{l \in L^0(t, x)} (-\lambda_t) (\|l\|^2 - 1) = 0. \tag{3.34}$$

В случае $\lambda(t) < \lambda_t$ из утверждения 3.1 извлекаем $\lambda_t = 0$, следовательно, условие (3.34) также выполняется. Из (3.34), (3.33), (3.32) следует выполнение для $\rho^0(\cdot)$ вида (1.7) условий (2.14), (2.15).

4. Приложение

Приведем некоторые сведения из теории квадратичных форм. Доказательство и подробное изложение указанных ниже утверждений можно найти, например, в [9].

Пусть H — действительная, квадратная, симметрическая $(n \times n)$ -матрица, E — единичная матрица, μ — собственное число матрицы H , т.е.

$$\det(H - \mu E) = 0. \quad (4.1)$$

Действительная, симметрическая $(n \times n)$ -матрица H имеет n собственных действительных чисел: $\mu_1 \leq \mu_2 \leq \dots \leq \mu_n$. Справедливо равенство

$$\max_{1 \leq i \leq n} \mu_i = \max_{\|s\|=1} \langle Hs, s \rangle. \quad (4.2)$$

Обозначим через $s_i \in R^n$ собственный вектор матрицы H , отвечающий собственному числу μ_i ($i \in \overline{1, n}$), т.е.

$$Hs_i = \mu_i s_i. \quad (4.3)$$

Существует ортонормированная система собственных векторов s_1, \dots, s_n , отвечающих собственным числам μ_1, \dots, μ_n матрицы H . Обозначим через S $(n \times n)$ -матрицу, столбцы которой суть собственные векторы матрицы H . Справедливо

$$\det S \neq 0, \quad S^T S = E = S S^T, \quad (4.4)$$

$$S^T H S = \text{diag} \{ \mu_1, \dots, \mu_n \}, \quad (4.5)$$

где T означает операцию транспонирования, символом $\text{diag} \{ \mu_1, \dots, \mu_n \}$ обозначена диагональная $(n \times n)$ -матрица.

Пусть $s \rightarrow \langle Hs, s \rangle$ — квадратичная форма, отвечающая матрице H .

Если $\langle Hs, s \rangle \leq 0$ для всех $s \in R^n$, квадратичная форма называется неположительно определенной, если $\langle Hs, s \rangle < 0$ для всех $s \in R^n, s \neq 0$, то квадратичная форма называется определенной отрицательно.

Для того, чтобы квадратичная форма $\langle Hs, s \rangle$ была определено неположительной (отрицательной) необходимо и достаточно, чтобы собственные числа μ_i матрицы H удовлетворяли неравенствам

$$\lambda_i \leq 0, \quad i = 1, \dots, n \quad (\mu_i < 0, \quad i = \overline{1, n}). \quad (4.6)$$

Для того, чтобы квадратичная форма $s \rightarrow \langle Hs, s \rangle$ была вогнутой (строго вогнутой) необходимо и достаточно, чтобы она была неположительно (отрицательно) определенной.

Рассмотрим функцию вида

$$(t, x) \rightarrow \rho(t, x) = \max_{l \in L} \varphi(t, x, l). \tag{4.7}$$

Пусть

$$L^0(t, x) = \{l \in L \subset R^n : \varphi(t, x, l) = \rho(t, x)\}, \tag{4.8}$$

где L — компакт, а функции $\varphi(\cdot) : (-\infty, \vartheta] \times R^n \times L \rightarrow R$, $\frac{\partial \varphi(\cdot)}{\partial t} : (-\infty, \vartheta] \times R^n \times L \rightarrow R$, $\frac{\partial \varphi(\cdot)}{\partial x} : (-\infty, \vartheta] \times R^n \times L \rightarrow R^n$ — непрерывны по совокупности.

Справедливо следующее

Лемма 4.1. Если L — выпуклый компакт, и функция $l \rightarrow \varphi(t, x, l)$ вогнута при каждом (t, x) , то при любом $(t, x) \in (-\infty, \vartheta] \times R^n$ функция $l \rightarrow D\varphi(t, x, l)|(1, f)$ вогнута на множестве $L^0(t, x)$.

Доказательство. Множество $L^0(t, x)$, очевидно, выпуклый компакт, т.е. для любых $l_1^0, l_2^0 \in L^0(t, x)$, $\lambda \in [0, 1]$ справедливо $l_\lambda^0 = \lambda l_1^0 + (1 - \lambda)l_2^0 \in L^0(t, x)$. Рассмотрим выражение

$$g(t, x, f, l, \delta) = [\varphi(t + \delta, x + \delta f, l) - \varphi(t, x, l)]\delta^{-1}. \tag{4.9}$$

При любых фиксированных $(t, x, f, l) \in (-\infty, \vartheta] \times R^n \times R^n \times L$ справедливо

$$\begin{aligned} \lim_{\delta \rightarrow +0} g(t, x, f, l, \delta) &= D\varphi(t, x, l)|(1, f) = \\ &= \frac{\partial \varphi(t, x, l)}{\partial t} + \left\langle \frac{\partial \varphi(t, x, l)}{\partial x}, f \right\rangle \end{aligned} \tag{4.10}$$

функция $l \rightarrow \varphi(t + \delta, x + \delta f, l)$ — вогнута по l на $L^0(t, x)$ и

$$\varphi(t, x, l) \equiv \rho(t, x) - \text{const}, \quad \forall l \in L^0(t, x). \tag{4.11}$$

Следовательно, функция $l \rightarrow g(t, x, f, l, \delta)$ вогнута по l на $L^0(t, x)$, т.е.

$$g(t, x, f, l_\lambda^0, \delta) \geq \lambda g(t, x, f, l_1^0, \delta) + (1 - \lambda)g(t, x, f, l_2^0, \delta) \tag{4.12}$$

и в обеих частях неравенства (4.12) можно перейти к пределу при $\delta \rightarrow +0$. Из (4.10), (4.12) получим

$$D\varphi(t, x, l_\lambda^0)|(1, f) \geq \lambda D\varphi(t, x, l_1^0)|(1, f) + (1 - \lambda)D\varphi(t, x, l_2^0)|(1, f), \tag{4.13}$$

что и требовалось доказать.

Литература

1. Красовский Н. Н., Третьяков В. Е. Стохастический программный синтез одного гарантирующего управления. Пробл. управл. и теории информ., т. 12, № 2, 1983.
2. Красовский Н. Н., Третьяков В. Е., Стохастический программный синтез для позиционной дифференциальной игры. ДАН СССР, т. 259, № 1, с. 24–27.
3. Красовский А. Н., Красовский Н. Н., Третьяков В. Е. Стохастический программный синтез для детерминированной позиционной дифференциальной игры. Прикл. мат. и мех., 1981, т. 45, вып. 4, с. 581–588.
4. Субботин А. И., Субботина Н. Н. Необходимые и достаточные условия для кусочно-гладкой цены дифференциальной игры. ДАН СССР, 1978, 243, № 4, с. 862–865.
5. Субботин А. И. Обобщение основного уравнения теории дифференциальных игр. ДАН СССР, 1980, т. 254, № 2, с. 293–297.
6. Красовский Н. Н., Субботин А. И. Позиционные дифференциальные игры, М., Наука, 1974.
7. Айзекс Р. Дифференциальные игры. М., Мир, 1967.
8. Демьянов В. Ф. Минимум: дифференцируемость по направлениям. Л., Изд-во ЛГУ, 1974.
9. Ланкастер П. Теория матриц. М., Наука, 1978.
10. Карлин С. Математические методы в теории игр, программировании и экономике. М., Мир, 1964.

**On sensitivity of the value function of the differential game
with the integral-terminal payoff**

N. N. SUBBOTINA, A. I. SUBBOTIN
(Sverdlovsk)

Using the method of the stochastic program synthesis N. N. Krasovskii and V. E. Tret'jakov have derived an expression for the value function of the feedback differential game with linear dynamic, fixed end-time and an integral-terminal payoff. In the present paper the differentiability properties of the value function are studied. The existence of the directional derivatives of the value function is proved and an expression for them is derived. Inequalities for the directional derivatives of the value function are obtained. It is shown that these inequalities together with the boundary condition constitute necessary and sufficient conditions for a nonsmooth function to be the value function of the considered game. In the region where the value function is differentiable the conditions obtained become the main equation of the theory of the differential games (the Isaacs-Bellman equation). The paper provides also necessary and sufficient conditions for a function to be the value-function of the nonlinear game with an integral-terminal payoff.

Н. Н. СУББОТИНА

А. И. СУББОТИН

Институт математики и механики УНЦ АН СССР

СССР, 620219 г. Свердловск, ГСП-384

ул. С. Ковалевской, 16

ON MINIMUM TIME CONTROL

A. K. CHAUDHURI, R. N. MUKHERJEE

(Calcutta)

(Burdwan)

(Received April 30, 1982)

Techniques of functional analysis have been applied by a number of workers [1, 2, 3, 5, 7] to tackle a variety of linear control systems. These techniques are more powerful than Pontryagin's maximum principle, because they can be successfully applied to solve many problems in linear control systems which are not amenable to Pontryagin's maximum principle. In functional analytic approach the control function is essentially required to belong to an appropriate Banach space, so the problem can be formulated as a mapping from this Banach space to another. In our papers [2, 3] it has been demonstrated how the minimum time linear control problems can be solved, where admissible control must satisfy a constraint on the norm. The application of the method developed in [2, 3] is straightforward to deal with the situation where the control function turns out to belong to a Banach space which is reflexive. In this paper we show that the same techniques can in fact be applied to the case when the Banach space is not reflexive, but is the conjugate of some appropriate Banach space. We shall illustrate the application by an example. It may be noted that this particular problem cannot be solved by the application of Pontryagin's maximum principle. No further conditions need be considered as suggested by Porter [7] to ensure the existence of the optimum control.

Introduction

Porter [7] has shown that if T is a bounded linear onto transformation defined on the Banach space B , taking values in another Banach space D , then, if B is reflexive, the minimum norm problem is always solvable, regardless of the nature of the range space D .

However, Porter has shown that if B is not reflexive, but is the conjugate of other Banach space X , i.e. $B = X^*$ and also $D = Y^*$, for some Banach space Y — then the minimum norm problem can be solved for bounded linear onto transformation $T: B \rightarrow D$, where $T = S^*$, and S is a one to one and bounded linear transformation from Y onto a closed subspace of X .

In this paper we have shown that the minimum norm problem involving any linear bounded onto transformation from B to D , where B is not reflexive but is the conjugate of some other Banach space, is indeed always solvable regardless of the nature of D . This result is then used to determine the minimum time control.

Discussion

Let us define time optimal control problem in a classical sense and to state certain theorems in a more general setting. The proofs have already been given in [2, 3]. Let B_t be a Banach space depending upon the continuous parameter t . Let D be another Banach space. Let T_t be a transformation depending upon the parameter t , mapping B_t onto D . Let $U_t \subset B_t$ be the unit ball in B_t and $\xi \in D$. The problem is to determine $u \in U_t$ such that $T_t u = \xi$ and t is minimum. We shall consider only the case when T_t is linear, bounded and onto.

In the problems which usually arise in practice, B_t is an increasing function of t in the sense that $B_{t_1} \subseteq B_{t_2}$, whenever $t_1 \leq t_2$. Also T_{t_1} can be regarded as the restriction of T_{t_2} defined on B_{t_2} , on B_{t_1} . It is easy to show that under the above conditions $U_{t_1} \subseteq U_{t_2}$.

Definition. The set of all points $\xi \in D$, such that $T_t u = \xi$ for some $u \in U_t$ will be called the Reachable Region (set) with respect to the linear transformation T_t and will be denoted by $C(t)$.

The following theorems have already been proved in [2, 3].

Th-1. The Reachable Region $C(t)$ is bounded and a convex body, symmetrical with respect to the origin of D [3].

Cor: $C(t)$ is closed, when B_t is either a reflexive space or it can be considered as a conjugate of some other Banach space.

Th-2. An admissible control which will be optimal must satisfy $\|u\|=1$ [2].

Th-3. Let $\xi \in \delta c(t)$ and $\varphi \in D^*$ determine the supporting hyperplane to $C(t)$ at ξ . Then $\langle \xi, \varphi \rangle = \|T_t^* \varphi\|$ [2].

Th-4. Let K be a weakly compact set in a Banach space D , and let $\varphi \in D^*$, the conjugate space to D . Then there exists a point $\eta_0 \in K$, such that φ defines a supporting hyperplane to K at $\eta_0 \in \delta K$ [2].

Th-5. If $\langle \xi, \varphi \rangle = \|T_t^* \varphi\|$ for some $\xi \in c(t)$ and some $\varphi \in D^*$, then $\xi \in \delta c(t)$ and φ also defines a supporting hyperplane to $C(t)$, at ξ , where B_t is either reflexive or it can be considered as the conjugate of some other Banach space [3].

Th-6. Let $\xi \in \delta c(t)$, where t is the given terminal time and $\varphi \in D^*$, define the supporting hyperplane at ξ . Let u_φ be the optimal control to reach at ξ in the above sense. Then u_φ maximizes $\langle u, T_t^* \varphi \rangle$, where T_t^* and D^* denote the adjoint transformation and adjoint space to T_t and D respectively and $\langle u_\varphi, T_t^* \varphi \rangle = \max_{\|u\|=1} \langle u, T_t^* \varphi \rangle = \|T_t^* \varphi\|$ and $\|u_\varphi\|=1$. (B_t is same as Th-5).

Proof. Since $C(t)$ is a closed convex body (by Th-1) and $\xi \in \delta c(t)$, there exists a $\varphi \in D^*$, such that $\langle \xi, \varphi \rangle \geq \langle \eta, \varphi \rangle$ for all $\eta \in c(t)$. Let $u \in U_t \subset B_t$ be such that $T_t u = \eta$. Since $C(t)$ is circled (by Th-1), it follows $\langle \xi, \varphi \rangle \geq |\langle T_t u, \varphi \rangle| = |\langle u, T_t^* \varphi \rangle|$, for all $u \in U_t \subset B_t$. Hence,

$$\langle \xi, \varphi \rangle \geq \sup_{\|u\| \leq 1, u \in B_t} |\langle u, T_t^* \varphi \rangle| = \|T_t^* \varphi\| \quad (\text{by definition}) \quad \dots (1).$$

Now, since $\xi \in \delta c(t)$, there is a $u_\varphi \in U_t$, such that $\xi = T_t u_\varphi$. Thus

$$\langle \xi, \varphi \rangle = \langle T_t u_\varphi, \varphi \rangle = \langle u_\varphi, T_t^* \varphi \rangle \leq \|u_\varphi\| \|T_t^* \varphi\| \leq \|T_t^* \varphi\| \quad \dots (2),$$

since $\|u_\varphi\| = 1$, (by Th-2).

From (1) and (2)

$$\langle u_\varphi, T_t^* \varphi \rangle = \|T_t^* \varphi\| \quad \dots (3).$$

Again, $\langle \eta, \varphi \rangle \leq \langle \xi, \varphi \rangle$, for all $\eta \in c(t)$, therefore $\langle u, T_t^* \varphi \rangle \leq \langle u_\varphi, T_t^* \varphi \rangle$ for all $u \in U_t \subset B_t$.

Hence

$$\sup_{\|u\| \leq 1, u \in B_t} \langle u, T_t^* \varphi \rangle \leq \langle u_\varphi, T_t^* \varphi \rangle = \|T_t^* \varphi\|, \text{ by (3)}. \quad \dots (4)$$

Again since $U_t \subset B_t$ is a weakly compact set, and $\langle u, T_t^* \varphi \rangle$ is a strongly continuous function of u , therefore $\sup_{\|u\| \leq 1, u \in B_t} \langle u, T_t^* \varphi \rangle = \sup_{\|u\|=1, u \in B_t} \langle u, T_t^* \varphi \rangle$ will be attained at some point $u_\varphi \in U_t \subset B_t$, $\|u_\varphi\| = 1$, which proves the theorem.

Cor: Given the conditions of the theorem, $u_\varphi = \overline{T_t^* \varphi} \in U_t$, where $\overline{T_t^* \varphi}$ is the extremal of $T_t^* \varphi$.

Proof. We already have proved that $\langle u_\varphi, T_t^* \varphi \rangle = \|T_t^* \varphi\|$, $\|u_\varphi\| = 1 \dots (1)$. Now by Hahn-Banach theorem, corresponding to $T_t^* \varphi$, there exists a $\overline{T_t^* \varphi} \in B_t^{**}$ such that $\langle \overline{T_t^* \varphi}, T_t^* \varphi \rangle = \|T_t^* \varphi\|$, $\|\overline{T_t^* \varphi}\| = 1 \dots (2)$. Comparing (1) and (2), u_φ can be selected to be equal to one of the $\overline{T_t^* \varphi} \in B_t^{**}$ which also belongs to B_t as proved in the theorem. Hence $u_\varphi = \overline{T_t^* \varphi}$.

Th-7. The N.A.S.C. for the point $\xi \in c(t)$ to be in $\delta c(t)$ at the time $t = t_f$ is that $\max_{\varphi} \frac{\langle \xi, \varphi \rangle}{\|T_t^* \varphi\|} = 1$ where $\varphi \in D^*$. (Here $T_t: B_t \rightarrow D$ is a bounded linear onto transformation and B_t is either a reflexive space or it can be considered as the conjugate of some other Banach space) [3].

Th-8. Let $\xi \in c(t_f) \cap \delta c(t_f)$ where $C(t_f)$ is the reachable region. Then $\max_{\psi} \frac{\langle \xi, \psi \rangle}{\|T_t^* \psi\|}$ is ≤ 1 or ≥ 1 according as $t \geq$ or $\leq t_f$ (Here B_t is to be considered as in Th-5) [2].

Th-9. Let $t_1 < t_2$ and $T_{t_1}: B_{t_1} \rightarrow D$ and $T_{t_2}: B_{t_2} \rightarrow D$ are bounded linear onto transformations. Then $C(t_1) \subseteq C(t_2)$ and $\delta c(t_1) \cap \delta c(t_2) = \Phi$ iff $\|T_{t_2}^* \varphi\| > \|T_{t_1}^* \varphi\|$ for all $\varphi \in D^*$, where Φ denotes the null set. (Here B_{t_1} and B_{t_2} are to be considered as in Theorem 5) [2].

$$\text{Cor:} \quad \text{If } \delta c(t_1) \cap \delta c(t_2) = \Phi \text{ then } \|T_{t_1}\| < \|T_{t_2}\|. \quad (2).$$

Th-10. Let $\xi \in \delta c(t_f) \cap c(t_f)$ and $t \geq t_f$. Then $\max_{\varphi} \frac{\langle \xi, \varphi \rangle}{\|T_t^* \varphi\|}$ is a non-increasing function for $t \geq t_f$. (Here B_t is to be considered as in Theorem 5) [3].

Th-11. Let $T_t U_t = C(t)$ for any given t and let $\eta \notin c(t)$. Let $\xi \in \delta c(t)$ be the point on the ray through η and t^* be the minimum time to reach ξ . If there exists an optimal control $u_t \in U_t$ to reach η in minimum time t^{**} , then $t^{**} > t^*$ [3].

Th-12. The sufficient conditions for the existence of minimum time control for η as in Theorem 11, are that (a). There exists a time t_1 , such that $\max_{\varphi} \frac{\langle \eta, \varphi \rangle}{\|T_{t_1}^* \varphi\|} < 1$ and (b)

$\max_{\varphi} \frac{\langle \eta, \varphi \rangle}{\|T_t^* \varphi\|}$ is a continuous function of t [3].

Th-13. Necessary condition for existence of admissible optimal control is that $\min_{\psi} \|T_t^* \psi\| = 1$ under the constraint $\langle \eta, \psi \rangle = 1$ will have at least one real positive root [3].

We shall consider the system governed by the following first order differential equations:

$$\frac{dx_1}{dt} = x_2 + u_1 \quad (1)$$

$$\frac{dx_2}{dt} = u_2$$

where, $x_1(t)$, $x_2(t)$ represent the instantaneous state of the system in the phase plane at time t and u_1 and u_2 are the control functions. $x_1(t)$ and $x_2(t)$ can be considered as deviations of the actual trajectory from the nominal trajectory. The problem is, given any initial value of the deviation $[x_1(0), x_2(0)]$ — what will be the control function required to reduce the error to the value zero in minimum time. Suppose that $u_1(t)$ and $u_2(t)$ are fuel flows which emanate from a common source and that the limiting factor in the system performance is the total instantaneous flow. Without any loss of generality we can set this value at 1. Thus the constraint imposed on $u_1(t)$ and $u_2(t)$ can be expressed as $J(u) = \text{Sup}_{t \in \tau} [|u_1(t)| + |u_2(t)|] \leq 1 \dots (2)$, where $\tau = [0, t_0]$, t_0 being the time for which the system is allowed to run. To solve this problem by applying the approach developed in (2), $u(t) = [u_1(t), u_2(t)]$ may be considered to belong to $L_{\infty}(l_1(2), \tau)$. We shall denote $L_{\infty}(l_1(2), \tau)$ by the symbol $B_{\infty,1}$. We observe that $\|u\|$ in this space coincide exactly with $J(u)$, i.e.

$$\|u\| = J(u) = \text{Sup}_{t \in \tau} [|u_1(t)| + |u_2(t)|] \quad \dots (3)$$

The solution of system (1) is given as

$$e^{-At}x(t) - x(0) = \int_0^t e^{-As}Bu(s)ds \quad \dots (4)$$

Where
$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

If the system reaches the null state at time t , equation (4) reduces to

$$-x(0) = \int_0^t e^{-As}Bu(s)ds = \int_0^t \begin{bmatrix} 1 & -s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} ds = T_t u.$$

Or putting $-x(0) = \xi$

$$\therefore \xi_1 = \int_0^t (u_1 - su_2)ds \quad \text{and} \quad \xi_2 = \int_0^t u_2 ds.$$

Where T_t is linear and onto, as can be easily verified. Thus the problem becomes one of a linear transformation of the Banach space $B_{\infty,1}$ to R^2 . Since $\|u\| \leq 1$, the above problem readily becomes one of mapping unit sphere U in $B_{\infty,1}$ into R^2 . We can consider this problem as a minimum time optimal control problem under the mapping $T_t: B_{\infty,1} \rightarrow R^2$, where $B_{\infty,1}$ is the conjugate of the Banach space $B_{1,\infty}$, i.e. $B_{1,\infty}^* = B_{\infty,1}$. If $\varphi = (\varphi_1, \varphi_2) \in R^2$, then $T_t^* \varphi \in B_{1,\infty}$. $\therefore \overline{T_t^* \varphi} \in B_{\infty,1}$ where $\overline{T_t^* \varphi}$ is the extremal of $T_t^* \varphi$. The condition of Theorem 9, i.e. $\|T_{t_1}^* \varphi\| < \|T_{t_2}^* \varphi\|$, $t_1 < t_2$ can be easily verified for this problem. The optimal control for the problem is given by [7, pp. 318-319].

$$u_1(t) = \begin{cases} \text{Sign} [\varphi_1], & t \in E \\ 0, & t \in [0, t_0] \sim E \end{cases}$$

where $E = \{t \in [0, t_0] : |\varphi_1| > |\varphi_2 - \varphi_1 t|\}$.

$$u_2(t) = \begin{cases} \text{Sign} [\varphi_2 - \varphi_1 t], & t \in [0, t_0] \sim E \\ 0, & t \in E \end{cases}$$

Since the set $C(t)$ is closed, the isochrones i.e. the boundaries of the Reachable set may be traced out by computing points $\{T_t(\overline{T_t^* \varphi}) : \varphi \in R^2\}$. To determine the shape of the set of isochrones (boundaries of the reachable sets), it will be necessary to consider the values of t lying within the ranges $0 < t < 1$, $1 < t < 2$ and $t > 2$ separately. Let us first consider the case when $0 < t < 1$. The isochrone for any $t \in (0, 1)$ will be bounded by the curves J, H, J', H' , where J' and H' are the reflections of J and H respectively as shown in Fig. 1. So it will be sufficient to trace J and H .

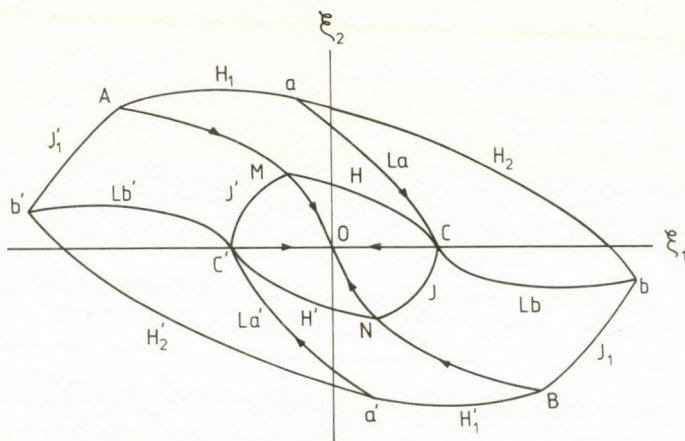


Fig. 1

J is determined by the conditions $t > \frac{\varphi_2 + \varphi_1}{\varphi_1}$ and, either (i) $\varphi_1 > 0, \varphi_2 > 0, |\varphi_1| > |\varphi_2|$ or (ii) $\varphi_1 > 0, \varphi_2 < 0, |\varphi_1| > |\varphi_2|$ as can be easily verified. Now, for $t \in (0, 1)$, the conditions (i) are not satisfied, since (i) implies $t > 1$. Hence J should be essentially determined by set of conditions (ii), i.e. $t > \frac{\varphi_2 + \varphi_1}{\varphi_1}, -1 < \frac{\varphi_2}{\varphi_1} < 0, 0 < t < 1$, or $\frac{\varphi_2}{\varphi_1} < t - 1, -1 < \frac{\varphi_2}{\varphi_1} < 0, 0 < t < 1$. Let $t = 1 - \delta$ where $0 < \delta < 1$. Hence from above $\frac{\varphi_2}{\varphi_1} < -\delta$ and this together with $-1 < \frac{\varphi_2}{\varphi_1} < 0$ implies $-1 < \frac{\varphi_2}{\varphi_1} < -\delta, 0 < \delta < 1$. It can be easily shown that the equation of J under the condition $t > \frac{\varphi_2}{\varphi_1} + 1, \varphi_1 > 0, \varphi_2 < 0, |\varphi_1| > |\varphi_2|$ is $\xi_1 - \left(t + \frac{t^2}{2}\right) = \xi_2 - 5(\xi_2 + t)^2 \therefore \frac{d\xi_2}{d\xi_1} = \frac{1}{1 - (\xi_2 + t)} \therefore \frac{\varphi_2}{\varphi_1} = \xi_2 + t - 1 = \xi_2 - \delta$. Hence $-1 < \frac{\varphi_2}{\varphi_1} < -\delta$, and $\frac{\varphi_2}{\varphi_1} = \xi_2 - \delta$ together imply that $-(1 - \delta) < \xi_2 < 0$ where $0 < \delta < 1$, i.e. $-t < \xi_2 < 0, 0 < t < 1$. Again it can be shown that H is determined by the condition $\varphi_1 > 0, \varphi_2 > 0, |\varphi_1| < |\varphi_2|, \frac{\varphi_2 - \varphi_1}{\varphi_1} < t < \frac{\varphi_2 + \varphi_1}{\varphi_1}$ and its equation is $\xi_1 - t = -\xi_2 - 5\xi_2^2 \therefore \frac{d\xi_2}{d\xi_1} = -\frac{1}{1 + \xi_2} \therefore \frac{\varphi_2}{\varphi_1} = 1 + \xi_2$. Put $t = 1 - \delta$ where $0 < \delta < 1$. Now from above $1 < \frac{\varphi_2}{\varphi_1} < t + 1$ and $\frac{\varphi_2}{\varphi_1} > t - 1 = -\delta$ and they

together imply $1 < \frac{\varphi_2}{\varphi_1} < 2 - \delta$ (\because for H , $\frac{\varphi_2}{\varphi_1} > 1$, $0 < \delta < 1$). Finally $1 < \frac{\varphi_2}{\varphi_1} < 2 - \delta$

and $\frac{\varphi_2}{\varphi_1} = 1 + \xi_2$ together imply $0 < \xi_2 < 1 - \delta$, $0 < \delta < 1$, i.e. $0 < \xi_2 < t$, $0 < t < 1$. Next let

us consider the values of $t \in (1, 2)$. The isochrone for any $t \in (1, 2)$ will be bounded by J , H , J' , and H' , where J' and H' are reflection of J and H respectively. So, as before, it will be sufficient to trace J and H . Here J is determined by both conditions (1)

$t > \frac{\varphi_2 + \varphi_1}{\varphi_1}$, $\varphi_1 > 0$, $\varphi_2 > 0$, $|\varphi_1| > |\varphi_2|$ and (2) $t > \frac{\varphi_2}{\varphi_1} + 1$, $\varphi_1 > 0$, $\varphi_2 < 0$, $|\varphi_1| > |\varphi_2|$.

Put $t = 1 + \delta$, where $0 < \delta < 1$. From (1), $0 < \frac{\varphi_2}{\varphi_1} < 1$, $\frac{\varphi_2}{\varphi_1} < t - 1 = \delta \Rightarrow 0 < \frac{\varphi_2}{\varphi_1} < \delta$. As

above $\frac{\varphi_2}{\varphi_1} = \xi_2 + t - 1 = \xi_2 + \delta$.

Finally $0 < \frac{\varphi_2}{\varphi_1} < \delta$ and $\frac{\varphi_2}{\varphi_1} = \xi_2 + \delta$ together imply $-\delta < \xi_2 < 0$ where $0 < \delta < 1$

\dots (3). From (2) we have $-1 < \frac{\varphi_2}{\varphi_1} < 0$ and $\frac{\varphi_2}{\varphi_1} < t - 1 = \delta \Rightarrow -1 < \frac{\varphi_2}{\varphi_1} < 0$.

Finally $-1 < \frac{\varphi_2}{\varphi_1} < 0$ and $\frac{\varphi_2}{\varphi_1} = \xi_2 + \delta$ together imply $-(1 + \delta) < \xi_2 < \delta \dots$ (4). From

(3) and (4), for J we have $-(1 + \delta) < \xi_2 < 0$ where $0 < \delta < 1$, i.e. $-t < \xi_2 < 0$, $1 < t < 2$.

Again H is determined by the condition $\varphi_1 > 0$, $\varphi_2 > 0$, $\frac{\varphi_2 - \varphi_1}{\varphi_1} < t <$

$< \frac{\varphi_2 + \varphi_1}{\varphi_1}$, $|\varphi_1| < |\varphi_2|$. Put $t = 1 + \delta$, where $0 < \delta < 1$. From above we have $1 < \frac{\varphi_2}{\varphi_1} <$

$< 2 + \delta$ and $\frac{\varphi_2}{\varphi_1} > \delta$ which together imply $1 < \frac{\varphi_2}{\varphi_1} < 2 + \delta$ (\because for H we must have $\frac{\varphi_2}{\varphi_1} > 1$,

$0 < \delta < 1$). Finally $1 < \frac{\varphi_2}{\varphi_1} < 2 + \delta$, $\frac{\varphi_2}{\varphi_1} = 1 + \xi_2$ together imply $0 < \xi_2 < 1 + \delta$ for $0 < \delta <$

< 1 , i.e. $0 < \xi_2 < t$, $1 < t < 2$. The set of isochrones for $t > 2$ is bounded by the curves J ,

H_1 , H_2 , J' , H'_1 , H'_2 , where J' , H'_1 , and H'_2 are reflections of J , H_1 and H_2 respectively. So it will sufficient to consider J , H_1 , and H_2 . Here J is also determined by conditions

(1) $t > \frac{\varphi_2}{\varphi_1} + 1$, $\varphi_1 > 0$, $\varphi_2 > 0$, $|\varphi_1| > |\varphi_2|$ and (2) $t > \frac{\varphi_2}{\varphi_1} + 1$, $\varphi_1 > 0$, $\varphi_2 < 0$, $|\varphi_1| > |\varphi_2|$.

From (1) we get $0 < \frac{\varphi_2}{\varphi_1} < 1$, $\frac{\varphi_2}{\varphi_1} < t - 1$. Put $t = 2 + \delta$, where $\delta > 0$. Conditions (1)

reduce to $0 < \frac{\varphi_2}{\varphi_1} < 1$ and $\frac{\varphi_2}{\varphi_1} < 1 + \delta$ which together imply $0 < \frac{\varphi_2}{\varphi_1} < 1$ (\because for J we

must have $\frac{\varphi_2}{\varphi_1} < 1$). Now, $\frac{\varphi_2}{\varphi_1} = \xi_2 + t - 1 = \xi_2 + 1 + \delta$, ($\delta > 0$). Finally $0 < \frac{\varphi_2}{\varphi_1} < 1$,

$\frac{\varphi_2}{\varphi_1} = \xi_2 + 1 + \delta$ imply $-(1 + \delta) < \xi_2 < -\delta \dots$ (3). Again from (2) we have $-1 < \frac{\varphi_2}{\varphi_1} <$

< 0 , $\frac{\varphi_2}{\varphi_1} < 1 + \delta$ which together imply $-1 < \frac{\varphi_2}{\varphi_1} < 0$ (\because for J we must have, $\frac{\varphi_2}{\varphi_1} < 0$).

But $\frac{\varphi_2}{\varphi_1} = \xi_2 + 1 + \delta$, $\delta > 0$. Finally $-1 < \frac{\varphi_2}{\varphi_1} < 0$, $\frac{\varphi_2}{\varphi_1} = \xi_2 + 1 + \delta$ together imply $-(2 + \delta) < \xi_2 < -(1 + \delta) \dots$ (4). From (3) and (4), when $\delta > 0$, for J we have $-(2 + \delta) < \xi_2 < -\delta$, i.e. $-t < \xi_2 < 2 - t$, $t > 2$. It should be noted, when $t > 2$ H is decomposed into two curves H_1 and H_2 as shown in Fig. 1. H_2 is determined by the set of

conditions $t > \frac{\varphi_2 + \varphi_1}{\varphi_1}$, $\varphi_1 > 0$, $\varphi_2 > 0$, $|\varphi_1| < |\varphi_2|$ and its equation can be easily verified as to be $\xi_1 - \left(t + \frac{t^2}{2}\right) = -25(\xi_2 + t)^2$.

$$\therefore \frac{d\xi_2}{d\xi_1} = -\frac{2}{\xi_2 + t}, \quad \text{hence} \quad \frac{\varphi_2}{\varphi_1} = \frac{\xi_2 + t}{2}.$$

Put $t = 2 + \delta$, where $\delta > 0$. Now from above we have $\frac{\varphi_2}{\varphi_1} > 1$, $\frac{\varphi_2}{\varphi_1} < t - 1 = 1 + \delta$, $\delta > 0$ which together imply $1 < \frac{\varphi_2}{\varphi_1} < 1 + \delta$. But $\frac{\varphi_2}{\varphi_1} = \frac{\xi_2 + t}{2} = \frac{\xi_2 + 2 + \delta}{2}$. Finally $1 < \frac{\varphi_2}{\varphi_1} < 1 + \delta$, $\frac{\varphi_2}{\varphi_1} = \frac{\xi_2 + 2 + \delta}{2}$ together imply $-\delta < \xi_2 < \delta$ where $\delta > 0$. Consequently when $\delta > 0$ we have for H_2 , $-\delta < \xi_2 < \delta$, $\delta > 0$, i.e. $-(t - 2) < \xi_2 < t - 2$, $t > 2$.

For H_1 we have the conditions $\frac{\varphi_2}{\varphi_1} - 1 < t < \frac{\varphi_2}{\varphi_1} + 1$, $\varphi_1 > 0$, $\varphi_2 > 0$, $|\varphi_1| < |\varphi_2|$ which yield $t - 1 < \frac{\varphi_2}{\varphi_1} < t + 1$, $\frac{\varphi_2}{\varphi_1} > 1$. Put $t = 2 + \delta$, $\delta > 0$. We have $1 + \delta < \frac{\varphi_2}{\varphi_1} <$

$< 3 + \delta$, $\frac{\varphi_2}{\varphi_1} > 1$ which together imply $1 + \delta < \frac{\varphi_2}{\varphi_1} < 3 + \delta$ (\because for H_1 we must have

$\frac{\varphi_2}{\varphi_1} > 1$ and $1 + \delta < \frac{\varphi_2}{\varphi_1}$, $\delta > 0$ imply $\frac{\varphi_2}{\varphi_1} > 1$). Finally $1 + \delta < \frac{\varphi_2}{\varphi_1} < 3 + \delta$, $\frac{\varphi_2}{\varphi_1} = 1 + \xi_2$ together imply, for H_1 , $\delta < \xi_2 < 2 + \delta$, $\delta > 0$, i.e. $t - 2 < \xi_2 < t$, $t > 2$. It can be easily verified that H_1 and H_2 are coincident upto their first order derivatives at $\xi = t - 2$, $t > 2$. The control sequence which has been used in the foregoing is needed to go from the origin of the state plane to some point on the boundary of the reachable set in minimum time.

If u is the control required for that purpose, then it can be shown that the optimal control to drive the system from some initial state to origin will be $-u$ [2].

The structure of the reachable set is depicted in Fig. 1. The reachable set is bounded by curves H, J, H' and J' for $t \leq 2$. The line OC is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ by the application of the optimal control $(-1, 0)$. The line OC' is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ by the application of the optimal control $(1, 0)$. The curve OA is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ by using the control $(0, -1)$ and its equation is $\xi_1 = -\frac{\xi_2^2}{2}, \xi_2 \geq 0$. The curve OB is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ by using the control $(0, 1)$ and its equation is $\xi_1 = \frac{\xi_2^2}{2}, \xi_2 \leq 0$. The equation of the bounding curve H is $\xi_1 - t = -\xi_2 - 5\xi_2^2$ which is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ in the same minimum time by using the control sequence $(0, -1)$ and $(-1, 0)$. The equation of the curve J is $\xi_1 - \left(t + \frac{t^2}{2}\right) = \xi_2 - 5(\xi_2 + t)^2$ which is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ by using the control sequence $(-1, 0)$ and $(0, 1)$ in the same minimum time. The equation of the bounding curves H' and J' which are the reflections of H and J can be easily determined.

We shall first obtain the law of optimal control for any initial state (ξ_1, ξ_2) such that the system can be driven to $(0, 0)$ in the same minimum time less than or equal to 2. For this purpose the reachable set ($t \leq 2$) is then divided into four regions: OMC, ONC and their reflections. Now for any point (ξ_1, ξ_2) in the region OMC the initial control should be $(0, -1)$. The system under the control would continue to trace a parabolic trajectory parallel to OM until it meets the line OC . At this instant the control switches from $(0, -1)$ to $(-1, 0)$. The system moves along CO under the control $(-1, 0)$ till it reaches the origin. Evidently OM does not belong to the region.

Again for any point (ξ_1, ξ_2) in the region ONC the initial control should be $(-1, 0)$. The system under the control would continue to trace a parabolic trajectory parallel to J until it meets the parabola ON . At this instant the control switches from $(-1, 0)$ to $(0, 1)$. The system moves along NO under the control $(0, 1)$. Evidently ON does not belong to the region. Similarly we can find the corresponding control law in the regions which are reflections of the above two regions.

Let us now describe the reachable set for $t > 2$ which is bounded by the curves H_1, H_2, J_1, H'_1 and J'_1, H'_2 (see Fig. 1).

The equation of the curve H_1 : $\xi_1 - t = -\xi_2 - 5\xi_2^2$ which is the locus of all points (ξ_1, ξ_2) that can be forced to $(0, 0)$ in same minimum time by using the control sequence $(0, -1)$ and $(-1, 0)$.

The equation of the curve H_2 : $\xi_1 - \left(t + \frac{t^2}{2}\right) = -25(\xi_2 + t)^2$ which is the locus of all points (ξ_1, ξ_2) that can be forced to $(0, 0)$ by using the control sequence $(0, -1)$,

$(-1, 0)$ and $(0, 1)$. The equation of the curve $J_1: \xi_1 - \left(t + \frac{t^2}{2}\right) = \xi_2 - 5(\xi_2 + t)^2$. J_1 is the locus of all points (ξ_1, ξ_2) which can be forced to $(0, 0)$ by using the control sequence $(-1, 0)(0, 1)$ in the same minimum time. The curve H'_1, H'_2 and J'_1 are the reflections of H_1, H_2 and J_1 respectively and the corresponding control sequences can be determined easily.

The locus of the point of first order coincidence of H_1 and H_2 is denoted by L_a and its equation is $\xi_1 = -\frac{1}{2}\xi_2^2 + 2(\xi_2 \geq 0)$. L_b denotes the locus of point of intersection of H_2 and J_1 and its equation is $\xi_1 = \frac{1}{2}(2 - \xi_2)^2, (\xi_2 \leq 0)$. L'_a and L'_b are reflections of L_a and L_b .

We shall now obtain the law of optimal control for any initial state (ξ_1, ξ_2) such that the system can be driven to $(0, 0)$ in the same minimum time greater than 2.

For this purpose the Reachable set for $t > 2$ is divided into six regions: $O A a C$, $a C b$, $O C b B$ and their reflections. For any point (ξ_1, ξ_2) in the region $O A a C$, the initial control should be $(0, -1)$. The system under this control would continue to trace a parabolic trajectory parallel to AO until it meets OC . At this instant the control switches from $(0, -1)$ to $(-1, 0)$.

The system now moves along CO under the control $(-1, 0)$ till it reaches the origin. Evidently OA does not belong to the region. Now, for any point (ξ_1, ξ_2) in the region $a C b$, the initial control would be $(0, -1)$. The system under this control would continue to trace a parabolic trajectory parallel to AO until it meets L_b . At this instant the control switches from $(0, -1)$ to $(-1, 0)$. The system then moves under the control $(-1, 0)$, till it meets BO . Again, at this instant the control switches from $(-1, 0)$ to $(0, 1)$. The system now moves along BO under the control $(0, 1)$ till it reaches to origin. Hence $a C$ and $C b$ do not belong to the origin.

Again, for any point (ξ_1, ξ_2) in the region $O c b B$, the initial control would be $(-1, 0)$. The system under the control $(-1, 0)$ would continue to move parallel to J_1 , till it meets BO . At this instant the control switches from $(-1, 0)$ to $(0, 1)$. The system now moves along BO under the control $(0, 1)$ till it finally reaches the origin. Evidently BO does not belong to the region. Similarly we can find the corresponding control law in the regions which are reflections of the above three regions.

The important features of the above regions are as follows:

- (1) L_a is a trajectory but not a switching curve. It divides the regions of single and double switching.
- (2) L_b is a switching curve but not a trajectory. It also divides the regions of single and double switching.
- (3) OC and OB are both switching curves and also trajectories. Similarly the reflections of L_a, L_b, OC and OB can be identified as above.

Thus we can formulate the control law as follows:

The time optimal control u^* , that forces the system to the origin of the phase plane is given by:

$$u^* = (0, -1) \text{ for all } (\xi_1, \xi_2) \in \\ \in \text{Region } (OAaC: -\frac{1}{2} \xi_2^2 < \xi_1 \leq -\frac{1}{2} \xi_2^2 + 2, \xi_2 > 0)$$

$$u^* = (0, -1) \text{ for all } (\xi_1, \xi_2) \in \\ \in \text{Region } (acb: \frac{1 + \text{sign } \xi_2}{2} \left\{ \xi_1 + \frac{1}{2} \xi_2^2 - 2 \right\} + \frac{1 - \text{sign } \xi_2}{2} \\ \left\{ \xi_1 - \frac{1}{2} (2 - \xi_2)^2 \right\} > 0, \text{ for } \xi_2 > 0)$$

$$u^* = (-1, 0) \text{ for all } (\xi_1, \xi_2) \in \\ \in \text{Region } (OCbB: \frac{1}{2} (2 - \xi_2)^2 \leq \xi_1 < \frac{1}{2} \xi_2^2, \xi_2 < 0)$$

References

1. Burns, J. A., „Existence theorems and necessary conditions for a general formulation of the minium effort problem”. J. Opt. Theory and appl., Vol. 15 (1975), pp. 413-440.
2. Chaudhuri, A. K., Mukherjee, R. N., “An alternative approach for solving a certain class of time optimal control problems” Indian J. pure appl. Math., 12 (2), 151-162, February 1981.
3. Chaudhuri, A. K., Mukherjee, R. N., “On the global controllability of a certain class of minimum time control problems”. Indian J. pure appl. Math., 13 (2), 163-171, February 1982.
4. Kantarovich, L. V., Akhilov, P., Functional analysis in Normed spaces. Macmillan (N. Y.) (1964).
5. Minamide, N., Nakamura, K., A minimum cost problem in Banach space. J. Math., Anal. Applic., 36 (1971), 73-85.
6. Pontryagin, L. S., Boltyanskii, V. G., Gramkrelidze, R. V., Mischenko, E. F. (1962): The Mathematical Theory of Optimal Process. Wiley & Sons, New York.
7. Porter, W. A. (1966): Modern foundations of system engineering. Macmillan & Co., New York.

О минимальном периоде управления

А. К. ЧАУДХУРИ, Р. Н. МУКХЕРИ

(Индия)

В работах [2, 3] было показано, как могут быть решены проблемы минимального периода линейного управления, причем входящее управление должно удовлетворять ограничению на норму. В этой статье показано, что такие же методы могут применяться в случае, когда исходное банахово пространство не рефлексивно, но банахово пространство, которому принадлежит функция управления, является сопряженным. Приведен пример.

A. K. CHAUDHURI
Indian Institute of Management,
Calcutta-700027, West Bengal,
INDIA

R. N. MUKHERJEE
Department of Mathematics,
University of Burdwan,
Burdwan, West Bengal,
INDIA

TOWARDS SET-THEORETIC REPRESENTATION OF NONDETERMINISTIC SYSTEMS

W. PEDRYCZ

(Gliwice)

(Received March 25, 1982)

We discuss the problem of representation of nondeterministic systems in terms of set theory taking into account a notion of relational equation (state equation of the system) linking state and input (treated as crisp or fuzzy set) by means of relation. An application of different types of set theoretic connectives in system equations makes it possible to perform analysis in very flexible manner.

An identification problem creating the background of further applications of this approach is stated and solved in analytical and numerical way. Moreover, it is shown that even the input and state are considered as crisp viz. nonfuzzy sets, a final result (the relation of the system) is fuzzy. Some numerical examples provided form numerical illustration.

Keywords: nondeterministic system, fuzzy set theory, identification.

1. Introduction

Considering various theoretical and applicational aspects of analysis and synthesis in nondeterministic systems we could easily distinguish two general and diverse concepts leading to performing uncertainty factors playing a main role in such class of systems. The first one, well known and widely implemented, is based on probability theory [3], [5], [6]; the second one uses the concept of set (instead of point) in order to express the uncertainty existing in the system which is almost impossible, or seems to form an artificial way to be processed in probabilistic manner. This form of uncertainty appears due to the complexity or ill-definedness of the concept under consideration. The idea of set-theoretic contribution in description of ambiguity could be found in some works performed in the area of automatic control and system theory (e.g. [1], [6], [16], unknown yet bounded disturbances) where nonfuzzy sets of noises were introduced. Since the establishment of fuzzy set theory [19] a lot of papers appeared where this idea was investigated as an appropriate and natural means for modelling ill-defined concepts (cf. [2], [4], [10], [11], [17], [20]).

The aim of this paper is twofold. We shall discuss several models representing different grades of uncertainty which are able to handle vague form of information,

distinguishing two classes of systems such as functional and relational where a factor of uncertainty could be treated as internal or external.

Next, one of the most important basic problems, such as identification task, is solved analytically and numerically.

First of all a brief summary of the general facts of nonfuzzy and fuzzy set theory will be presented.

2. A review of some facts of nonfuzzy and fuzzy set theory

Let X denote a universe of discourse (space); $P(X)$ express the family of all subsets (sets) of X ,

$$P(X) = \{A \mid A \subseteq X\} \quad (1)$$

(of course an empty set \emptyset and X belong to $P(X)$). Introducing set operations such as union $A \cup B$, intersection $A \cap B$, complement \bar{A} , it could be easily shown that $P(X)$ with these operations forms a Boolean algebra $\langle P(X), \cup, \cap, \bar{\ } \rangle$. It is also possible to verify that introducing the characteristic function of the set $A \in P(X)$,

$$\chi_A: X \rightarrow \{0, 1\} \quad \chi_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases} \quad (2)$$

$x \in X$, and defining in a family of all characteristic functions $CH(X)$, max, min, complement operations,

$$(\chi_A \vee \chi_B)(x) = \max(\chi_A(x), \chi_B(x)) = \chi_A(x) \vee \chi_B(x) \quad (3)$$

$$(\chi_A \wedge \chi_B)(x) = \min(\chi_A(x), \chi_B(x)) = \chi_A(x) \wedge \chi_B(x) \quad (4)$$

$$\chi_{\bar{A}}(x) = 1 - \chi_A(x) \quad (5)$$

$x \in X$, the Boolean algebra $\langle CH(X), \vee, \wedge, \bar{\ } \rangle$ and the Boolean algebra $\langle P(X), \cup, \cap, \bar{\ } \rangle$ introduced before are isomorphic [10]. It is possible to perform further discussion in terms of characteristic functions of the respective sets.

The classical notion of the set was extended and modified by introduction of the notion of fuzzy set [19] characterized by means of a membership function.

$$\mu: X \rightarrow [0, 1]. \quad (6)$$

This concept, established almost 20 years ago, was applied in various problems of system analysis as a convenient tool for handling and processing imprecise, vague, nondeterministic concepts [4], [10].

Basic operations such as union, intersection, complement are defined as before,

$$\text{union } \mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) = \mu_A(x) \vee \mu_B(x) \tag{7}$$

$$\text{intersection } \mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) = \mu_A(x) \wedge \mu_B(x) \tag{8}$$

$$\text{complement } \mu_{\bar{A}}(x) = 1 - \mu_A(x) \tag{9}$$

$x \in X$. The consequence of Eq. (6) is such that a family of all fuzzy sets $F(X)$ does not form a Boolean algebra; it is easy to check that the following properties are not preserved,

$$\mu_A(x) \vee \mu_{\bar{A}}(x) = 1, \quad \mu_A(x) \wedge \mu_{\bar{A}}(x) = 0 \tag{10}$$

$\langle F(X), \cup, \cap, \bar{\ } \rangle$ forms a so-called soft or Morgan algebra [10].

The concept of set is contained in the concept of fuzzy set, thus according to the embedding principle, every result for nonfuzzy (crisp) sets could be derived from the result obtained during an analysis where fuzzy sets were applied.

3. Nondeterministic system in set-theoretic approach

For the clarity of our considerations and without loosing the generality of investigations, let us restrict our discussion to dynamical systems of the first order, where X_k, X_{k+1} represent n -dimensional vector of state, U_k denotes m -dimensional input vector. Each element of X_k and U_k is treated as crisp (or fuzzy) set defined on the space X and U , respectively,

$$X_k = (X_1^k, X_2^k, \dots, X_n^k) \tag{11}$$

$$U_k = (U_1^k, U_2^k, \dots, U_m^k) \tag{12}$$

viz.

$$X_i^k \in P(X), \text{ or } F(X) \quad U_j^k \in P(U), \text{ or } F(U), \tag{13}$$

$$i = 1, 2, \dots, n, j = 1, 2, \dots, m.$$

The relationships between input and state are represented by a relation (which modelizes the sentence "there exists a relation between input and state"); thus we obtain,

$$(U_k X_k) R X_{k+1} \tag{14}$$

where R is a relation defined on the Cartesian product of the appropriate spaces,

$$\underbrace{U \times U \dots \times U}_m \times \underbrace{X \times X \times \dots \times X}_n \times \underbrace{X \times X \times \dots \times X}_n \tag{15}$$

viz.

$$R \in P\left(\prod_{i=1}^m U \times \prod_{j=1}^{2n} X\right) \quad \text{or} \quad R \in F\left(\prod_{i=1}^m U \times \prod_{j=1}^{2n} X\right). \quad (16)$$

Equation (14) could be rewritten in more explicit form,

$$X_{k+1} = U_k \circ X_k \circ R. \quad (17)$$

$X_k, X_{k+1} \in F(P)\left(\prod_{i=1}^n X\right), U_k \in F(P)\left(\prod_{j=1}^m U\right)$, where “ \circ ” stands for max-min composition; Eq. (17) is called nonfuzzy (fuzzy) relational equation. Making use of the characteristic (or membership) functions, we put down

$$\begin{aligned} \chi_{X_{k+1}}(y_1, y_2, \dots, y_n) = & \bigvee_{\substack{u_1 \in U \\ u_2 \in U \\ \vdots \\ u_m \in U \\ x_1 \in X \\ x_2 \in X \\ \vdots \\ x_n \in X}} (\chi_{U_k}(u_1, u_2, \dots, u_m) \wedge \chi_{X_k}(x_1, x_2, \dots, x_n) \wedge \\ & \wedge \chi_R(u_1, u_2, \dots, u_m, x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n)), \end{aligned} \quad (18)$$

$\vee = \sup(\max), \wedge = \inf(\min)$.

Now let us introduce a definition of noninteractive systems.

Definition 1. We call the nondeterministic system (Eq. (17)) noninteractive (unrelated), if it could be represented in terms of set of relational equations,

$$\begin{aligned} X_1^{k+1} &= U_1^k \circ U_2^k \circ \dots \circ U_m^k \circ X_1^k \circ X_2^k \circ \dots \circ X_n^k \circ R_1 \\ X_p^{k+1} &= U_1^k \circ U_2^k \circ \dots \circ U_m^k \circ X_1^k \circ X_2^k \circ \dots \circ X_n^k \circ R_p \\ &\vdots \\ X_n^{k+1} &= U_1^k \circ U_2^k \circ \dots \circ U_m^k \circ X_1^k \circ X_2^k \circ \dots \circ X_n^k \circ R_n \end{aligned} \quad (19)$$

viz.

$$\begin{aligned} \chi_{X_p^{k+1}}(y) = & \sup_{\substack{u_1 \in U \\ u_2 \in U \\ \vdots \\ u_m \in U \\ x_1 \in X \\ x_2 \in X \\ \vdots \\ x_n \in X}} [\min(\chi_{U_1^k}(u_1), \chi_{U_2^k}(u_2), \dots, \\ & \chi_{U_m^k}(u_m), \chi_{X_1^k}(x_1), \chi_{X_2^k}(x_2), \dots, \\ & \dots, \chi_{X_p^k}(x_p), \chi_{R_p}(u_1, u_2, \dots, u_m, x_1, x_2, \dots, x_n, y))], \end{aligned} \quad (20)$$

$p = 1, 2, \dots, n$. Equation (19) forms a general class of nondeterministic systems when the element of uncertainty is tied with the system itself (relation) and input-state data (crisp or fuzzy sets). It could be easily shown that a deterministic system described by means of the equations.

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k)) \tag{21}$$

e.g.

$$x_p(k+1) = f_p(u_1(k), u_2(k), \dots, u_m(k), x_1(k), x_2(k), \dots, x_n(k)), \tag{22}$$

$p = 1, 2, \dots, n$, forms a particular case of Eq. (19). If the factor of uncertainty appears only in the expression of input and state data, while R is a function, we speak about a nondeterministic functional system. Then we get,

$$\begin{aligned} X_1^{k+1} &= f_1(U_1^k, U_2^k, \dots, U_m^k, X_1^k, X_2^k, \dots, X_n^k) \\ X_1^k &= f_2(U_1^k, U_2^k, \dots, U_m^k, X_1^k, X_2^k, \dots, X_n^k) \\ &\vdots \\ X_n^{k+1} &= f_n(U_1^k, U_2^k, \dots, U_m^k, X_1^k, X_2^k, \dots, X_n^k). \end{aligned} \tag{23}$$

Now Eq. (20) is read as follows,

$$\begin{aligned} \chi_{X_p^{k+1}}(y) &= \sup [\min (\chi_{U_1^k}(u_1), \chi_{U_2^k}(u_2), \dots, \\ &\dots, \chi_{U_m^k}(u_m), \chi_{X_1^k}(x_1), \chi_{X_2^k}(x_2), \dots, \chi_{X_n^k}(x_n))] . \end{aligned} \tag{24}$$

Formula (24) is called max-min extension principle [20], playing a significant role in the theory of fuzzy sets. Assuming additionally that \mathbf{X} and \mathbf{U} are spaces of real numbers $\mathbf{U} = \mathbf{X} = \mathbf{R}$, and f_p is a linear function of its arguments,

$$\begin{aligned} y &= f_p(a_{p1}, a_{p2}, \dots, a_{pn}, b_{p1}, b_{p2}, \dots, b_{pn}, x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m) = \\ &= \sum_{i=1}^n a_{pi} x_i + \sum_{j=1}^m b_{pj} u_j \end{aligned} \tag{25}$$

$p = 1, 2, \dots, n$ and crisp sets $X_1^k, X_2^k, \dots, X_n^k, U_1^k, U_2^k, \dots, U_m^k$ are defined by characteristic functions equal to 1 in the intervals $[x_i^-, x_i^+]$ and $[u_j^-, u_j^+]$ respectively and 0 otherwise, $i = 1, 2, \dots, n, j = 1, 2, \dots, m$ (thus every crisp set can be treated as an interval with the center α_i and radius $r_i (\geq 0)$, where for the i -th set,

$$\alpha_i(x_i + x_i)/2 \quad \text{and} \quad r_i = |x_i - x_i|/2. \tag{26}$$

Equation (23) can be represented compactly as

$$\begin{pmatrix} X_1^{k+1} \\ X_2^{k+1} \\ \vdots \\ X_n^{k+1} \end{pmatrix} = \begin{pmatrix} a_{11}X_1^k \oplus a_{12}X_2^k \oplus \dots \oplus a_{1n}X_n^k \\ a_{21}X_1^k \oplus a_{22}X_2^k \oplus \dots \oplus a_{2n}X_n^k \\ \vdots \\ a_{n1}X_1^k \oplus a_{n2}X_2^k \oplus \dots \oplus a_{nn}X_n^k \end{pmatrix} \oplus \begin{pmatrix} b_{11}U_1^k \oplus b_{12}U_2^k \oplus \dots \oplus b_{1m}U_m^k \\ b_{21}U_1^k \oplus b_{22}U_2^k \oplus \dots \oplus b_{2m}U_m^k \\ \vdots \\ b_{n1}U_1^k \oplus b_{n2}U_2^k \oplus \dots \oplus b_{nm}U_m^k \end{pmatrix} \quad (27)$$

and further on,

$$X^{k+1} = A \circ X^k \oplus B \circ U^k \quad (28)$$

where

$$A = [a_{ij}]$$

$$B = [b_{ij}]$$

$$X^k = \begin{bmatrix} X_1^k \\ X_2^k \\ \vdots \\ X_n^k \end{bmatrix} \quad U^k = \begin{bmatrix} U_1^k \\ U_2^k \\ \vdots \\ U_m^k \end{bmatrix} \quad (29)$$

\oplus stands for Minkowski addition (Minkowski addition of sets) [9],

$$X \oplus Z = \{w \mid w = x + z, x \in X, z \in Z\} \quad (30)$$

or in terms of characteristic function is equal to,

$$\begin{aligned} \chi_{X \oplus Z}(w) &= \sup_{(X, Z): w = x + z} [\min(\chi_X(x), \chi_Z(z))] = \\ &= \sup_{x \in X} [\min(\chi_X(x), \chi_Z(w - x))] . \end{aligned} \quad (31)$$

Now we consider fuzzy systems creating a generalization of Eq. (19), introducing various forms of composition between fuzzy sets and fuzzy relations. Generally speaking, " \circ "-composition could be treated in many ways, as e.g. max-min composition as mentioned previously, max-prod, max-min_q, where min_q forms a generalized version of the min operator [13], [18]. Taking into account the p -th equation of set (19) we get,

— for sup-prod composition:

$$X_p^{k+1} = U_1^k * U_2^k * \dots * U_m^k * X_1^k * X_2^k * \dots * X_n^k * R_p \quad (32)$$

where X_p^{k+1} has the following membership function,

$$\begin{aligned} \mu_{X_p^{k+1}}(y) = & \sup (\mu_{U_1^k}(u_1) \cdot \mu_{U_2^k}(u_2) \cdot \dots \cdot \\ & \cdot \mu_{U_m^k}(u_m) \cdot \mu_{X_1^k}(x_1) \cdot \mu_{X_2^k}(x_2) \cdot \dots \cdot \\ & \cdot \dots \cdot \mu_{X_n^k}(x_n) \cdot \mu_{R_p}(u_1, u_2, \dots, u_m, x_1, x_2, \dots, x_n, y)) \end{aligned} \quad (33)$$

— for sup-min composition:

$$X_p^{k+1} = U_1^k \circ Q U_2^k \circ Q \dots \circ Q U_m^k \circ Q X_1^k \circ Q X_2^k \circ Q \dots \circ Q X_n^k \circ Q R_p \quad (34)$$

$$\min_q(a, b) = 1 - \min(1 - \sqrt[q]{(1-a)^q + (1-b)^q}) \quad (35)$$

$q \geq 1, a, b \in [0, 1]$.

It could be checked that the difference between the fuzzy state X_p^{k+1} in a sup-min and a sup-prod system for the same input and state in the k -th time moment is bounded and does not exceed the value $1 - 1/2^{n+m+1}$ [11],

$$\mu_{X_p^{k+1}}(y) - \mu_{\tilde{X}_p^{k+1}}(y) \leq 1 - 1/2^{n+m+1} \quad (36)$$

where \tilde{X}_p^{k+1} stands for the state in sup-prod fuzzy system. Sup-min_q composition generates the class of fuzzy relational systems which is more general than this is given by Eqs. (19) and (32). Putting down $q = \infty$ we get a sup-min system while the value of q equal to $1/(\log 3/\log 2 - 1)$ gives an approximation of a sup-prod system [12]. By simple inspection replacing all the membership functions by characteristic functions of sets and relations it can be verified that all the types of equations stated above are equivalent.

Similarly as before we are able to distinguish fuzzy relational systems and fuzzy functional systems described by the equation

$$\begin{aligned} \mu_{X_p^{k+1}}(y) = & \sup [\min (\mu_{U_1^k}(u_1), \mu_{U_2^k}(u_2), \dots, \\ & \dots, \mu_{U_m^k}(u_m), \mu_{X_1^k}(x_1), \mu_{X_2^k}(x_2), \dots, \mu_{X_n^k}(x_n))] \\ & (u_1, u_2, \dots, u_m, x_1, x_2, \dots, x_n) \in f_p^{-1}(y) \end{aligned} \quad (37)$$

$p = 1, 2, \dots, n$. The use of these forms of system depends on a concrete application; when we suppose that the input data are precisely described and could be treated as points (degenerated sets) while there is doubt that the input-state relationship can be expressed as a function, then this approach modelizes the situation when we deal with precise information (data), but there are some reasons (e.g. due to the complexity of the system) not permitting precise, unique and reasonable formulation of the links between input and state in terms of a smooth function. On the other hand, if the knowledge of the system is complete but there are a lot of difficulties by collecting of data having a

high noise level, resulting in imprecision in measurements, the fuzzy system with the fixed nonfuzzy structure (function) and fuzzy input and state seems to be appropriate for formulating the framework for further analysis.

Let us consider two examples explaining the concept of fuzzy relational and fuzzy functional system.

1. A subsystem of industrial process can be described by means of the equation

$$T = c/pV \quad (38)$$

where $c = \text{constant}$, $T = \text{temperature } [^{\circ}\text{K}]$, $p = \text{pressure } [\text{N}/\text{m}^2]$, $V = \text{volume } [\text{m}^3]$; p, V are treated as inputs of the system, T is an output. Here the knowledge of the system is complete (well-known physical formula), but due to the high level of unknown disturbances existing in the installation input-output data can be modelled in the framework of set theory. Thus, we deal with a fuzzy functional system, where uncertainty is an external factor which can be eliminated with additional costs. On the other hand, if we consider a complex fuzzy system in "soft" science (e.g. management science) where we are able to establish and "measure" only input-state variables (e.g. opinion of society obtained via polls), the construction of a functional model (linear regression model) seems to be very difficult; here the fuzziness (uncertainty) is built in this system.

Generally, considering fuzzy relational and fuzzy functional systems let us introduce the equivalence of these structures stating the following definition.

Definition 2. We say fuzzy relational and fuzzy functional system are (F_1, F_2) -equivalent, if the following property holds true,

$$\begin{aligned} & U_1^k \circ U_2^k \circ \dots \circ U_m^k \circ X_1^k \circ X_2^k \circ \dots \circ X_n^k \circ R_p = \\ & = f_p(U_1^k, U_2^k, \dots, U_m^k, X_1^k, X_2^k, \dots, X_n^k, A_1, A_2, \dots, A_s) \end{aligned} \quad (39)$$

stand for fuzzy parameters.

4. Identification problem in nondeterministic systems

An identification problem in systems discussed here can be splitted, as usually, into two subproblems,

- a task of determination of the structure of the system,
- a task of calculation of the parameters of the structure established before.

Dealing with the system with nonfuzzy sets and relations, we need to specify the number of inputs and states; the work with the fuzzy systems demands also additional

specification about the type of the composition operator. The problem of determination of unknown relation of the system can be solved analytically or numerically. Consider the p -th fuzzy relational equation from set (19),

$$X_p^{k+1} = U_1^k \circ U_2^k \circ \dots \circ X_1^k \circ X_2^k \circ \dots \circ X_n^k \circ R_p \tag{40}$$

with a collection of respective fuzzy data for discrete time moments $t \in \{1, 2, \dots, T\}$.

$$\begin{aligned} U_1^1, & U_2^1, \dots, U_m^1, X_1^1, X_2^1, \dots, X_n^1, X_p^2 \\ & \vdots \\ U_1^{T-1}, & U_2^{T-1}, \dots, U_m^{T-1}, X_1^{T-1}, \dots, X_n^{T-1}, X_p^T. \end{aligned} \tag{41}$$

Let Q denote a performance index measuring the sum of distances (d) between fuzzy sets $X_{pT-1}^t, t=2, 3, \dots, T$ and the fuzzy sets of state calculated from Eq. (40)

$$Q = \sum_{t=1}^T d(U_1^t \circ U_2^t \circ \dots \circ U_m^t \circ X_1^t \circ X_2^t \circ \dots \circ X_n^t \circ R_p, X_p^{t+1}). \tag{42}$$

Then we get the following optimization problem:

$$\begin{aligned} & \min Q \\ & R \in F \left(\prod_{i=1}^m U \times \prod_{j=1}^n X \right). \end{aligned} \tag{43}$$

Assuming that a collection of fuzzy data (41) fulfils Eq. (40) without any error (viz. $d=0$), the fuzzy relation R can be determined solving Eq. (40) for given $U_1^t, U_2^t, \dots, U_m^t, X_1^t, X_2^t, \dots, X_n^t$, and X_p^{t+1} . It was shown [15] that the greatest fuzzy relation \hat{R}^t such that

$$U_1^t \circ U_2^t \circ \dots \circ U_m^t \circ X_1^t \circ X_2^t \circ \dots \circ X_n^t \circ \hat{R}^t = X_p^{t+1}$$

holds true, is given by the formula,

$$\hat{R}_p^t = (U_1^t \times U_2^t \times \dots \times U_m^t \times X_1^t \times X_2^t \times \dots \times X_n^t) \otimes X_p^{t+1} \tag{44}$$

where \otimes represents an α -composition defined as

$$a \alpha b = \begin{cases} 1, & \text{if } a \leq b \\ b, & \text{if } a > b \end{cases} \tag{45}$$

$a, b \in [0, 1]$. Applying membership functions of the respective fuzzy sets of input and state we obtain

$$\begin{aligned} \mu_{\hat{R}_p^t}(u_1, u_2, \dots, u_m, x_1, x_2, \dots, x_n, y) = & \min(\mu_{U_1^t}(u_1), \mu_{U_2^t}(u_2), \dots, \\ & \dots, \mu_{U_m^t}(u_m), \mu_{X_1^t}(x_1), \mu_{X_2^t}(x_2), \dots, \mu_{X_n^t}(x_n)) \alpha \mu_{X_p^{t+1}}(y). \end{aligned} \tag{46}$$

Because of the fact that \hat{R}_p^t is contained in R_p , finally we take an intersection of partial results given by Eq. (44) [2], [11],

$$\hat{R} = \bigcap_{t=1}^{T-1} \hat{R}^t. \quad (47)$$

For a nonfuzzy set, α -composition has a very clear meaning:

$$a \alpha b = \begin{cases} 0, & \text{if } a=1, b=0 \\ 1, & \text{otherwise} \end{cases} \quad (48)$$

so it can be treated as an implication operator in two-valued logic. For another forms of composition operators presented in Section 3, analytical results can be also derived [13].

The problem remains if the data does not fulfil the equation of the system due to the noises existing during collecting the data. Then we propose to apply the modified Newton's method [3]. Assuming additionally that "d" is the Euclidean metric and \mathbf{X} and \mathbf{U} are discrete, it involves

$$\begin{aligned} Q = & T^{-1} \sum_{t=1}^{n} \sum_{l=1}^n \left(\bigvee_{\mathbf{I}, \mathbf{J}} (\mu_{U_1^t}(u_{i_1}) \wedge \mu_{U_2^t}(u_{i_2}) \wedge \dots \wedge \mu_{U_m^t}(u_{i_m}) \wedge \right. \\ & \wedge \mu_{X_1^t}(x_{j_1}) \wedge \mu_{X_2^t}(x_{j_2}) \wedge \dots \wedge \mu_{X_n^t}(x_{j_n}) \wedge \\ & \left. \wedge \mu_{R_p}(u_{i_1}, u_{i_2}, \dots, u_{i_m}, x_{j_1}, x_{j_2}, \dots, x_{j_n}, y_l) \right) - \mu_{X_{p+1}^t}(y_l) \end{aligned} \quad (49)$$

where now

$$\begin{aligned} \mathbf{I} &= \{i_1, i_2, \dots, i_m \mid 1 \leq i_1, i_2, \dots, i_m \leq m\}, \\ \mathbf{J} &= \{j_1, j_2, \dots, j_n \mid 1 \leq j_1, j_2, \dots, j_n \leq n\} \end{aligned} \quad (50)$$

card $(\mathbf{X}) = n$, card $(\mathbf{U}) = m$. Let us admit a simplified notation:

$$\begin{aligned} \mu_{U_j^t}(u_{i_j}) &= u_{i_j}, & \mu_{X_j^t}(x_{i_j}) &= x_{i_j}, & \mu_{X_{p+1}^t}(y_l) &= y_l \\ \mu_{R_p}(u_{i_1}, u_{i_2}, \dots, u_{i_m}, x_{j_1}, x_{j_2}, \dots, x_{j_n}, y_l) &= r_{i_1 i_2, \dots, i_m j_1 j_2, \dots, j_n l}. \end{aligned} \quad (51)$$

Then the modified Newton's iteration scheme takes a form,

$$r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V}^{(N+1)} = r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V}^{(N)} - \alpha_N \frac{\partial Q}{\partial r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V}^{(N)}} \quad (52)$$

$S_1 S_2 \dots S_m \in \mathbf{I}$, $W_1 W_2 \dots W_n \in \mathbf{J}$, $1 \leq V \leq n$, where superscripts stand for the number of

iteration and α_N denotes the coefficient imposing good convergence properties of the method. Calculating the derivatives standing in formula (52) we get

$$\frac{\partial Q}{\partial r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V}} = 2 \sum_{t=1}^{T-1} \bigvee_{I, J} (u_{i_1} \wedge u_{i_2} \wedge \dots \wedge u_{i_m} \wedge x_{j_1} \wedge x_{j_2} \wedge \dots \wedge \dots \wedge x_{j_n} \wedge r^{i_1 i_2 \dots i_m j_1 j_2 \dots j_n V} - y_v \cdot P_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V} \tag{53}$$

and

$$P_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V} = \left\{ \begin{array}{l} 1, \text{ if } u_{S_1} \wedge u_{S_2} \wedge \dots \wedge u_{S_m} \wedge x_{W_1} \wedge x_{W_2} \wedge \dots \wedge \\ \wedge \dots \wedge x_{W_n} \wedge r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V} \geq \\ \geq \bigvee_{\substack{I = \{S_1 S_2 \dots S_m\} \\ J = \{W_1 W_2 \dots W_n\}}} (u_{i_1} \wedge u_{i_2} \wedge \dots \wedge u_{i_m} \wedge x_{j_1} \wedge x_{j_2} \wedge \dots \wedge \\ \wedge \dots \wedge x_{j_n} \wedge x_{i_1 i_2 \dots i_m j_1 j_2 \dots j_n V}) \\ \text{and } u_{S_1} \wedge u_{S_2} \wedge \dots \wedge u_{S_m} \wedge x_{W_1} \wedge x_{W_2} \wedge \dots \wedge \\ \wedge \dots \wedge x_{W_n} \geq r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V} \\ 0, \text{ otherwise.} \end{array} \right. \tag{54}$$

In [14] α_N was adjusted according to formula,

$$\alpha_N = 1/(c + N^\kappa) \tag{55}$$

$\kappa \geq 0$, and

$$c = \max \frac{\partial Q}{\partial r_{S_1 S_2 \dots S_m W_1 W_2 \dots W_n V}} \leq 2(T-1). \tag{56}$$

5. Numerical example

Now we present a simple numerical example, illustrating the method discussed above. Let a collection of input-output data of the system described by the equation

$$Y = X \circ R \tag{57}$$

be as follows,

No. of data (k)	χ_X	χ_Y
1	1 1 1 0	0 0 1 1
2	0 1 1 0	0 1 0 0
3	1 1 0 0	0 0 1 1
4	0 0 1 1	1 1 0 0
5	0 1 1 0	0 1 1 0
6	0 0 0 1	1 1 1 0

Processing calculations (Eq. (52) with $\kappa=0.2$, initial fuzzy matrix $R=0$) we get the values of performance index displayed in Fig. 1. The final fuzzy relation R is given by means of the following matrix,

$$\mu_R = \begin{vmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & .5 & 0 \\ 0 & .66 & .46 & 0 \\ 1 & 1 & .5 & 0 \end{vmatrix}.$$

Let us note that although the input-output data (X, Y) are nonfuzzy, in general, matrix of the system is fuzzy. Moreover making use of entropy measure of fuzziness [7], [8] of R defined as,

$$H(R) = 1/\text{card}(X \times Y) \sum_{i=1}^{\text{card}(X)} \sum_{j=1}^{\text{card}(Y)} \Delta(\mu_R(x_i, y_j)) \quad (58)$$

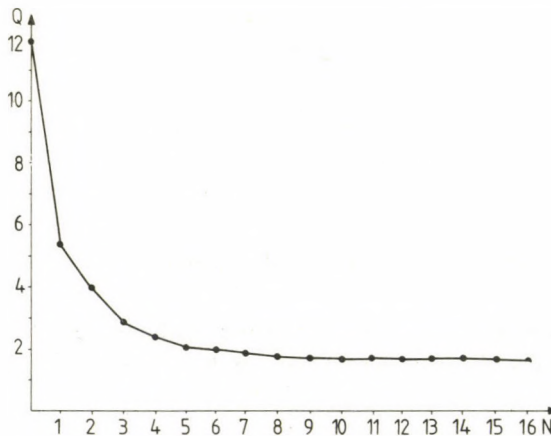


Fig. 1. Performance index Q vs. number of iterations (N)

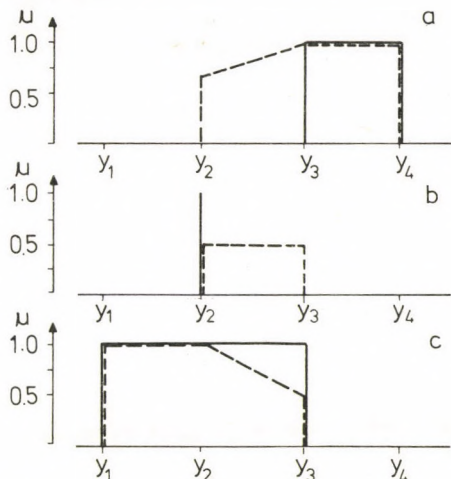


Fig. 2. Output sets of the model $X \circ R$ and output data Y , for (a) $k=1$, (b) $k=2$, (c) $k=6$

with $\Delta: [0, 1] \rightarrow [0, 1]$ such that,

$$\Delta(0) = \Delta(1) = 0 \tag{59}$$

$$\Delta(v) = \Delta(1-v) \quad v \in [0, 1]. \tag{60}$$

Δ is strictly increasing on the interval $[0, \frac{1}{2}]$ and strictly decreasing on the interval $[\frac{1}{2}, 1]$ and putting down

$$\Delta(v) = \begin{cases} v, & \text{if } v \in [0, \frac{1}{2}] \\ 1-v, & \text{if } v \in [\frac{1}{2}, 1] \end{cases} \tag{61}$$

one obtains,

$$H(R) = 1.38.$$

Figure 2 depicts the output of the system described by equation $Y = Y \circ R$ and respective collected data.

In order to obtain the nonfuzzy relation \bar{R} (viz. $\in \{0, 1\}$), let us put down

$$\chi_{\bar{R}} = \begin{cases} 1, & \text{if } \mu_R(x_i, y_j) \geq \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{62}$$

where $\varepsilon \in [0, 1]$ is a threshold level adjusted in order to minimize the following performance index,

$$V = \sum_{k=1}^T \sum_{j=1}^{\text{card}(Y)} [\mu_{Y_k}(y_j) - \mu_{X_k \circ \bar{R}}(y_j)]^2. \tag{63}$$

Then for R given above we get

$$\chi_{\bar{R}} = \begin{vmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{vmatrix} \varepsilon = 0.5$$

and $V=3$.

6. Concluding remarks

We presented an attempt for formalizing and establishing a set-theoretic approach for modelling uncertainty in system analysis. An idea of crisp and fuzzy sets was applied; a relational equation with various logical connectives was considered in details. The basic problem of identification of the relation of the system was taken into account; analytical and numerical resolution of the problem was investigated. The line of system modelling in the presence of uncertainty factor via fuzzy relational equations seems to be general enough for further theoretical investigations (e.g. control problems), as well, some numerical algorithms presented here make it possible to consider fuzzy relational equations flexible enough for practical purposes.

References

1. Bertsekas, D. P., Rhodes, I. B. (1971), Recursive state estimation for a set membership description of uncertainty, *IEEE Trans. Automatic Control*, **2**, 117-128.
2. Czogala, E., Pedrycz, W. (1981), On identification in fuzzy systems and its applications in control problems, *Fuzzy Sets and Systems*, **6**, 73-83.
3. Dorny, C. N. (1975), *A Vector Space Approach to Models and Optimization*, N. York: Wiley and Sons.
4. Dubois, D., Prade, H. (1980), *Fuzzy Sets and Systems*, N. York: Academic Press.
5. Eykhoff, P. (1971), *Parameter and State Estimation*, N. York: Wiley and Sons.
6. McGarthy, T. P. (1974), *Stochastic Systems and State Estimation*, N. York: Wiley and Sons.
7. de Luca, A., Termini, S. (1972), A definition of a non-probabilistic entropy in the setting of fuzzy set theory, *Information and Control*, **20**, 301-312.
8. de Luca, A., Termini, S. (1974), Entropy of L -fuzzy sets, *Information and Control*, **24**, 55-73.
9. Najfeld, I., Vitale, R. A., Davis, P. (1980), Minkowski iteration of sets, *Linear Algebra and Applications*, **29**, 259-291.
10. Negoita, C. V., Ralescu, D. A. (1975), *Applications of Fuzzy Sets to Systems Analysis*, Basel: Birkhauser Verlag.
11. Pedrycz, W. (1981), An approach to the analysis of fuzzy systems, *Int. J. Control*, **34**, 403-421.
12. Pedrycz, W. (1982), Some aspects of fuzzy decision-making. *Kybernetics*, **11**, 297-301.
13. Pedrycz, W., Fuzzy relational equations with generalized connectives and their applications, submitted to *Fuzzy Sets and Systems*.

14. Pedrycz, W., Numerical and applicational aspects of fuzzy relational equations, submitted to Fuzzy Sets and Systems.
15. Sanchez, E. (1976), Resolution of composite relation equations, *Information and Control*, **30**, 38–48.
16. Schweppe, F. C. (1968), Recursive state estimation: unknown but bounded errors and system inputs, *IEEE Trans. Automatic Control*, **13**, 22–28.
17. Tanaka, H., Uejima, S., Asai, K. (1981), Fuzzy linear regression model, in *Applied Systems and Cybernetics* (G. E. Lasker ed.) vol. VI, 2933–2938, N. York: Pergamon Press.
18. Yager, R. R. (1980), On general class of fuzzy connectives, *Fuzzy Sets and Systems*, **4**, 285–292.
19. Zadeh, L. A. (1965), Fuzzy sets, *Information and Control*, **8**, 338–353.
20. Zadeh, L. A. (1973), *The concept of a linguistic variable and its application to approximate reasoning*, N. York: Elsevier Publ. Comp.

К теоретико-множественному представлению недетерминистических систем

В. ПЕДРИЧ

(Гливице)

Обсуждается проблема представления недетерминистических систем в терминах теории множеств с учетом уравнения состояния системы.

Применение различных типов теоретико-множественных связей в системных уравнениях позволяет проводить анализ гибким образом.

Проблемы идентификации, составляющая основу дальнейших применений данного подхода сформулирована и решается аналитически и численно. В качестве иллюстраций приведены некоторые численные примеры.

W. PEDRYCZ

Department of Automatic Control & Computer Sciences

Silesian Technical University

Gliwice

Poland

SIMPLEX ALGORITHMS FOR UNCONSTRAINED MINIMIZATION

A. S. RYKOV

(Moscow)

(Received March 31, 1982)

Introduction

In designing complex systems, planning operations, developing technological processes the problems of maximum efficiency and quality are often formalized as extremal ones and solved by optimization techniques. However, the practical application of the latter to the optimized function value computation often involves considerable experimental and other costs, and the explicit values of the function derivatives are not obtainable. In this case an attempt can be made to estimate the derivatives by the finite differences method. The resulting errors, however, can lead either to a poor convergence of the method or to the absence thereof.

The most promising among the wide range of optimization techniques are the iterative direct search techniques employing data relating only to the values of the optimized function which makes it possible to apply them to a host of problems, both in modelling and in optimization of real-life objects. The direct search techniques are of significant importance allowing one to find an acceptable solution in a finite number of iterations and with relatively small costs involved in measuring the values of the optimized function in each iteration. The methods of this type include simplex techniques, first formulated in 1962 [1]. At present, the number of papers concerning different versions of the method and their practical application approach is one hundred. A rather good idea of the current techniques is given in [2, 3].

The present paper undertakes generalization of simplex algorithm construction principles, based on simplices of arbitrary form, introduces various types of simplex vertex reflection, employs the notion of local optimality, offers several criteria of optimal direction choice, evaluates properties of simplex algorithms.

Problem formulation and description of simplex algorithm general arrangement

Here is considered the problem of unconstrained minimization of a scalar function $f(x)$ of vector argument $x^T = (x_1, \dots, x_n) \in E^n$, E^n — n -dimensional Euclidean space. The problem solution implies construction of a minimizing sequence $\{x^N\}$, for which the following relation holds

$$\lim_{N \rightarrow \infty} f(x^N) = \inf_{x \in E^n} f(x).$$

The general simplex algorithm scheme is as follows. The first step of the simplex iterative procedure of $f(x)$ function extremum search implies construction of an arbitrary n -dimensional simplex S_1 with vertices $x^{1,i}$ ($i = \overline{1, n+1}$, where i is the vertex number)

$$x^{1,i} = x^1 + R_{1,i} r^{1,i}, \quad x^1 = \frac{1}{n+1} \sum_{i=1}^{n+1} x^{1,i},$$

where x^1 is the simplex S_1 centroid (point of simplex median intersection); $R_{1,i}$ is the distance from the simplex centroid x^1 to the i th vertex; $r^{1,i}$ is the unit n -dimensional vector directed from the center to the i th vertex. Matrix A of simplex S_1 of $(n+1) \times n$ dimension is:

$$A = \| \| r^{1,1} \dots r^{1,n+1} \| \| ^T$$

and for a special case regular simplex with $R_{1,i} = 1$ ($i = \overline{1, n+1}$).

$$A = \frac{1}{na_n} \left\| \begin{array}{cccccc} a_1 & a_2 & \dots & a_{n-1} & a_n & \\ -a_1 & a_2 & \dots & a_{n-1} & a_n & \\ \dots & \dots & \dots & \dots & \dots & \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & \dots & -(n-1)a_{n-1} & a_n & \\ 0 & 0 & \dots & 0 & -na_n & \end{array} \right\|.$$

$$a_i = \frac{1}{\sqrt{2i(i+1)}} \quad (i = \overline{1, n})$$

The values of the function $f(x^{1,i})$ ($i = \overline{1, n+1}$) are measured in vertices $x^{1,i}$ ($i = \overline{1, n+1}$), one or several S_1 simplex vertices are reflected (replaced), new vertices

(reflected) together with non-reflected ones form a new simplex S_2 . The values of the function $f(x)$ are measured in the new vertices of S_2 and the described procedure is repeated. The vertex reflection is to be done so that the generated simplex $\{S_N\}$ sequence corresponds to a non-ascending sequence of maximum function values within the simplex vertices. Let the search be halted at step N , the value $\min_{1 \leq i \leq n+1} f(x^{N,i})$ is taken as estimate of f_{\min} . This is the general algorithm.

In order to construct simplex algorithms, one has to solve a number of problems. First, determine the type of simplex vertex reflection generating possible directions of simplex displacement, hence, its centroid. Then introduce the notion of optimal direction and a rule of direction choice out of a multitude of possible centroid displacement directions. Find the step-length, the rule search halt.

To study the minimizing properties of simplex algorithms, we shall consider a sequence of simplex x^N centroids, defined as follows

$$x^{N+1} = x^N + \beta^N P^N, \quad N = 1, 2, \dots,$$

where P^N is a unit n -dimensional vector affecting the simplex centroid displacement direction when passing from simplex S_N to S_{N+1} at step N ; β^N is a positive value equal to the displacement of the simplex centroid when passing from X^N to X^{N+1} ;

$$X^N = \frac{1}{n+1} \sum_{i=1}^{n+1} x^{N,i}$$

where $x^{N,i}$ are vertices of simplex S_N .

The following definitions of reflections and optimal direction choice will be used in constructing simplex algorithms:

Definition 1. The reflection $m+k$ ($m = \overline{1, n}$; $k = \overline{0, n-m}$) of S_N simplex vertices is to be understood as the passage $m+k$ of its vertices along the vector directed from the geometrical centroid m of S_N reflected vertices toward the centroid of non-reflected ones $n+1-m-k$ of S_N , wherein the direction of the vector $x^{N+1} - x^N$ coincides with the prescribed direction and the new simplex S_{N+1} is formed by $m+k$ reflected vertices and $n+1-m-k$ vertices which have not been displaced in the mentioned direction.

Definition 2. The reflection m ($m = \overline{1, n}$) of S_N simplex vertices, retaining its form, is to be referred to the passage m of its vertices in relation to the centroid of non-reflected vertices $n+1-m$ wherein the direction of $x^{N+1} - x^N$ vector coincides with the one from the geometrical centroid m of reflected vertices toward the centroid of non-reflected vertices $n+1-m$ and the new simplex S_{N+1} is formed by m reflected vertices and $n+1-m$ vertices displaced so that the form of simplex S_{N+1} remains intact.

Definition 3. The simplex minimization algorithm is referred to locally optimal, if at each step N vector P^N and the relevant vertices are defined by the relation

$$P = \arg \max_{P \in \Omega^N} I^N, \quad (1)$$

where I^N is the direction optimally criterion whose computation implies measurement of the $f(x)$ function values in the simplex vertices.

According to the definitions, each minimization step N generates a set Ω^N of simplex P centroid displacement directions when choosing different vertices of simplex S_N for reflection whose elements depend on the type of reflection and the problem of choosing a somewhat optimal direction of simplex centroid displacement is solved.

According to Definition 1, all simplex vertices are divided into three groups having m vertices, k vertices and $n+1-m-k$ vertices. Centroids m of reflected vertices and $n+1-m-k$ of non-reflected ones determine the direction of simplex centroid displacement, and k of the reflected vertices are passed along the vector parallel to this direction. There are several versions of reflection corresponding to the introduced definitions.

Reflection 1

$$x^{N+1,j} = x^{N,j} + \alpha \Delta_N(m, k), \quad j = \overline{1, m},$$

$$x^{N+1,j} = x^{N,j} + \frac{\alpha m}{n+1-k} \Delta_N(m, k), \quad j = \overline{m+1, m+k},$$

$$x^{N+1,j} = x^{N,j}, \quad j = \overline{m+k+1, n+1},$$

$$x^{N+1} = x^N + \frac{\alpha}{n+1-k} \Delta_N(m, k),$$

$$\Delta_N(m, k) = \frac{1}{n+1-m-k} \sum_{i=m+k+1}^{n+1} x^{N,i} - \frac{1}{m} \sum_{i=1}^m x^{N,i}, \quad \alpha \in [0, \infty).$$

Reflection 2

$$x^{N+1,j} = x^{N,j} + \alpha \Delta_N^j(m, k), \quad j = \overline{1, m},$$

$$x^{N+1,j} = x^{N,j} + \frac{\alpha m}{n+1-k} \Delta_N(m, k), \quad j = \overline{m+1, m+k},$$

$$x^{N+1,j} = x^{N,j}, \quad j = \overline{m+k+1, n+1},$$

$$x^{N+1} = x^N + \frac{\alpha}{n+1-k} \Delta_N(m, k),$$

$$\Delta_N^j(m, k) = \frac{1}{n+1-m-k} \sum_{i=\overline{m+k+1}}^{n+1} x^{N,i} - x^{N,j}.$$

Reflection 3

$$x^{N+1,j} = x^{N,j} + \alpha \Delta_N(m), \quad j = \overline{1, m},$$

$$x^{N+1,j} = x^{N,j}, \quad j = \overline{m+1, n+1},$$

$$x^{N+1} = x^N - \frac{\alpha}{n+1-m} \sum_{i=1}^m (x^{N,i} - x^N),$$

$$\Delta_N(m) = \frac{1}{n+1-m} \sum_{i=\overline{m+1}}^{n+1} x^{N,i} - \frac{1}{m} \sum_{i=1}^m x^{N,i}.$$

Reflection 4

$$x^{N+1,j} = x^{N,j} + \alpha \Delta_N^j(m), \quad j = \overline{1, m},$$

$$x^{N+1,j} = x^{N,j}, \quad j = \overline{m+1, n+1},$$

$$x^{N+1} = x^N - \frac{\alpha}{n+1-m} \sum_{i=1}^m (x^{N,i} - x^N),$$

$$\Delta_N^j(m) = \frac{1}{n+1-m} \sum_{i=\overline{m+1}}^{n+1} x^{N,i} - x^{N,j}.$$

Reflection 5

$$x^{N+1,j} = x^{N,j} + \alpha \Delta_N^j(m), \quad j = \overline{1, m},$$

$$x^{N+1,j} = x^{N,j} + \alpha_1 \Delta_N^j(m), \quad j = \overline{m+1, n+1},$$

$$x^{N+1} = x^N - \frac{\alpha}{n+1-m} \sum_{i=1}^m (x^{N,i} - x^N),$$

$$\Delta_N^j(m) = \frac{1}{n+1-m} \sum_{i=m+1}^{n+1} x^{N,i} - x^{N,j}, \quad \alpha_1 = \begin{cases} \alpha & \text{at } \alpha \leq 1 \\ 2-\alpha & \text{at } \alpha > 1. \end{cases}$$

Reflection 5 corresponds to Definition 2, and reflections 1 to 4 correspond to Definition 1. Reflections 3 and 4 are special cases of reflections 1 and 2 with $k=0$. α value determines the step-length.

Let us introduce the choice criterion for the locally optimal reflection direction from the set Ω^N .

$$I_1^N = -\Delta f^N = f^*(x^N) - f^*(x^{N+1}),$$

$$I_2^N = -\frac{1}{m+k} \Delta f^N,$$

$$I_3^N = -(\text{grad } f(x^N), P),$$

$$I_4^N(m) = \sum_{i=1}^m [f(x^{N,i}) - \Phi_N],$$

$$\Phi_N = \frac{1}{2} [f(x^{N,1}) + f(x^{N,m+1})],$$

$$I_5^N(m) = \frac{1}{m} I_4^N(m),$$

$$I_6^N(m) = \sum_{i=1}^m [f(x^{N,i}) - f^*(x^N)],$$

where $f^*(x^N)$ equals either the measured value of $f(x^N)$ or

$$f^*(x^N) = \frac{1}{n+1} \sum_{i=1}^{n+1} f(x^{N,i}).$$

The first criterion value is related to the diminishing value of function $f(x)$ in the centroid of simplex S_N at the N th step of passage from simplex S_N to S_{N+1} as a result of vertex reflection. The value of the second criterion equals the diminution of $f(x)$ within the simplex centroid referred to as single function measurement. The third criterion

equals the value of the projection of vector P to the vector antigradient and estimates the proximity of simplex P centroid displacement direction to antigradient direction. The three latter criteria correspond to the case of m vertex reflections. It should be noted that criterion I_4^N can be used for minimization of non-numerical functions.

Problem (I) has to be solved at each N , hence, given the rapidly increasing number of elements in Ω^N with the growth of n , it would be of interest to find the ways of excluding the deliberately non-optimal elements from Ω^N .

Definition 4. The enumeration of simplex S vertices is correct if the following chain of inequalities holds

$$f(x^{N,1}) \geq f(x^{N,2}) \geq \dots \geq f(x^{N,n+1}).$$

The utilization of correct enumeration specifically allows one to reduce the number of elements in Ω^N .

Local properties of algorithms

Let us consider the construction and the local properties of simplex algorithms for a special case of the minimization of the linear function $f(x)$. In this case the type of reflection and optimality criterion completely determine the algorithm structure. Employing the geometric properties of n -dimensional simplex and linearity $f(x)$ we shall move to explicit dependence for criteria I_i^N ($i = 1, 2, 3$) (criteria I_i^N ($i = 4, 5, 6$) are of the required type). For reflections 1 and 2

$$I_1^N(m, k) = \frac{m\alpha}{n+1-k} \left(\frac{1}{m} \sum_{i=1}^m \Delta f_i^N - \frac{1}{n+1-m-k} \sum_{i=m+k+1}^{n+1} \Delta f_i^N \right),$$

$$I_2^N(m, k) = \frac{1}{m+k} I_1^N(m, k),$$

$$I_3^N(m, k) = \frac{1}{\|\Delta_N(m, k)\|} \left(\frac{1}{m} \sum_{i=1}^m \Delta f_i^N - \frac{1}{n+1-m-k} \sum_{i=m+k+1}^{n+1} \Delta f_i^N \right),$$

where $\Delta f_i^N = f(x^{N,i}) - f(x^N)$. For reflections 3, 4, 5

$$I_1^N(m) = \frac{\alpha}{n+1-m} \sum_{i=1}^m \Delta f_i^N,$$

$$I_2^N(m) = \frac{1}{m} I_1^N(m),$$

$$I_3^N(m) = \frac{\sum_{i=1}^m \Delta f_i^N}{\left\| \sum_{i=1}^m (x^{N,i} - x^N) \right\|}.$$

For a regular simplex the latter formula takes form of

$$I_3^N(m) = \frac{1}{R_N} \sqrt{\frac{n}{m(n+1-m)}} \sum_{i=1}^m \Delta f_i^N,$$

where $R_N = R_{N,i}$ ($i = \overline{1, n+1}$).

The presented formula shows that correct enumeration results in the maximum value of optimality criterion with each fixed m and k . Hence, to solve problem (I) for reflections 1 and 2, enumeration being correct, it suffices to compare $n(n+1)/2$ of optimality criterion values under $m = \overline{1, n}$; $k = \overline{0, n-m}$ and choose optimal m^N and k^N corresponding to the maximum value of criterion I_i^N ($i = 2, 3$). For criterion I_1^N it suffices to consider n values of criterion $I_1^N(n-j, j)$ with $j = \overline{0, n-1}$ and choose j corresponding to the maximum value of I_1^N [4]. As for reflections 3, 4 and 5 it suffices to compare less than n values of criterion I_i^N ($i = 1, 2, 3$), the quantity of these values is defined by lemma 1 and consequence 1, presented in [5]. Should m^* be the biggest of the numbers for which $f_{m^*}^N > 0$ or $f(x^{N, m^*}) - \Phi_N > 0$, then $m^N = m^*$ for reflections 3, 4 and 5 and criteria $I_4^N(m)$ and $I_6^N(m)$ respectively. For criterion $I_5^N(m)$ $m^N = 1$ (in case $I_5^N(1) = I_5^N(m)$ $2 \leq m \leq n$ we choose the smallest m).

Hence, it follows that criteria $I_4^N(m)$ and $I_6^N(m)$ make it possible to choose the number of vertices for which $\Delta f_i^N > 0$ or $f(x^{N, i}) - \Phi_N > 0$. Criterion I_5^N generates an algorithm similar to that of Nelder-Mead [6].

The combination of the introduced types of reflections and optimality criteria generates a class of simplex algorithms, for which the theorem holds true.

Theorem 1. The sequence $\{x^N\}$ generated by the simplex algorithm is minimizing and for algorithms with reflections 3, 4, 5 and criteria I_i^N ($i = 1, 6$) the following inequality holds

$$f(x^{N+1}) \leq f(x^1) - \frac{\alpha \|\text{grad } f\|}{n} \sum_{j=1}^N R_{j,1} \cos \varphi_{j,1}, \quad (2)$$

where

$$\cos \varphi_{j,1} = \frac{(x^{j,1} - x^j, \text{grad } f)}{\|x^{j,1} - x^j\| \|\text{grad } f\|} > 0, \quad \|x^{j,1} - x^j\| = R_{j,1},$$

for algorithms with reflections 1, 2 and criterion $I_1^N(m, k)$ the following inequality holds

$$f(x^{N+1}) \leq f(x^1) - \frac{\alpha}{2} \|\text{grad } f\| \sum_{j=1}^N (R_{j,1} \cos \varphi_{j,1} - R_{j,n+1} \cos \varphi_{j,n+1}), \quad (3)$$

$$\cos \varphi_{j,1} > 0, \quad \cos \varphi_{j,n+1} > 0.$$

Proof. For reflections 3, 4, 5 and criteria $I_i^N(m)$ ($i = \overline{1, 6}$) if mapping m^N vertices (for criteria $I_5^N(m)$ $m^N = 1$) due to the function linearity

$$f(x^{N+1}) - f(x^N) = - \frac{\alpha}{n+1 - m^N} \sum_{i=1}^{m^N} \Delta f_i^N \leq - \frac{\alpha}{n} \|\text{grad } f\| R_{N,1} \cos \varphi_{N,1}.$$

Hence, follows (2). Reflections 1, 2 and criterion $I_1^N(m, k)$ $m^N = n - k^N$ and

$$\begin{aligned} f(x^{N+1}) - f(x^N) &= - \frac{\alpha(n - k^N)}{n + 1 - k^N} \left(\frac{1}{n - k^N} \sum_{i=1}^{n - k^N} \Delta f_i^N - \Delta f_{n+1}^N \right) \leq \\ &\leq - \frac{\alpha}{2} (\Delta f_1^N - \Delta f_{n+1}^N). \end{aligned}$$

It follows that (3) is justified. The minimizing properties of the sequence $\{x^N\}$ arise from the obtained inequalities.

Note that the estimate for criteria $I_2^N(m, k)$ and $I_3^N(m, k)$ is not worse than (2) for $I_2^N(m)$ and $I_3^N(m)$, as the reflected vertices are chosen from the set Ω^N containing all displacement directions for criteria $I_2^N(m)$ and $I_3^N(m)$.

It follows from the theorem that the rate of the algorithm convergence depends on the form and orientation of the simplex relative to the gradient direction, in particular on the value of $\cos \varphi_{N,1}$, whose minimum varies for different criteria. One can get an exact estimate of the minimal value of $\cos \varphi_{N,1}$ only in case the simplex is regular ($R_{N,1} = R_{N,j}$, $j = \overline{2, n+1}$).

Lemma. For a regular simplex and reflections 3, 4 and 5 the following inequalities hold:

for criteria $I_1^N(m), I_3^N(m)$ $\cos \varphi_{N,1} \geq \frac{1}{n},$

for criterion $I_2^N(m)$ $\cos \varphi_{N,1} \geq \frac{3}{n+2},$

for criterion $I_3^N(m)$ $\cos \varphi_{N,1} > \frac{1}{n},$

for criteria $I_4^N(m)$ and $I_6^N(m)$

$$\cos \varphi_{N,1} \geq \begin{cases} \sqrt{\frac{2}{n+1}} & \text{with odd } n \\ \sqrt{\frac{2(n+1)}{n(n+2)}} & \text{with even } n, \end{cases}$$

for reflections 1, 2 and criterion $I_1^N(m, k)$

$$\cos \varphi_{N,1} \geq \sqrt{\frac{n+1}{2n}}, \quad \cos \varphi_{N,n+1} \leq -\sqrt{\frac{n+1}{2n}}.$$

The proof of the lemma is based on the technique described in [5]. The quoted estimates give an idea of the local properties of simplex algorithms.

Nonlinear function $f(x)$ minimization

The utilization of simplex algorithms for optimization of nonlinear functions brings about the problem of choice of steplength, size and form depending on the minimization procedure. The reflection definition required the introduction of the α parameter. By varying its value one can expand or contract the simplex in a chosen direction of simplex centroid displacement for reflections 1 to 4, adapting the simplex form and size, in the best possible manner, to the optimized function topology. When mapping 5 with α simplex size can be varied only, leaving its form intact. Thus, there is the choice problem at each step of α value minimization leading to a successful minimization of the function $f(x)$.

The α value should be chosen so that the sequence of optimized function values in the simplex $\{f(x^N)\}$ centroids was monotonically diminishing. To realize the choice we shall introduce the rule of $\{f(x^N)\}$ diminution monotony check. Each N th iteration involves a step with $\alpha = \alpha^1$ (e.g. $\alpha^1 = 2$, which retains the simplex size and form), estimation of $f^*(x^{N+1})$ value and comparison with $f^*(x^N)$. Should this step result in diminution of the function value, then an attempt is made to make a step with $\alpha > \alpha^1$ (e.g., $\alpha = 3$ and the simplex is expanded for reflections 1 to 4 or its size is increased for reflection 5) and a step is chosen leading to a greater diminution of the function value. If a step with $\alpha = \alpha^1$ has not resulted in the function value diminution within the simplex centroid, then a step with $\alpha < \alpha^1$ is made (e.g., $\alpha = 1.5$ and later $\alpha = 0.5$) and a step with α leading to the function value diminution.

For reflections 1 to 4, it is possible to choose α value for each reflected vertex separately. Initially, a step with $\alpha = \alpha^1$ is made followed by selection of α value for each reflected vertex, leading to diminution of $f(x)$. The properties of reflections 1 to 4 allow to ascertain that in case such a simplex form is adopted for reflections 1, 4 the centroid x^{N+1} will lie on the vector of chosen direction, and for reflections 2, 3 it can deviate from this direction. Other versions of simplex size variation rules are to be found in [2, 3].

The algorithm description is rounded up with the definition of the search halt rule. Its choice can result from a specific formulation of the optimization problem. Optimization can go on until the simplex reaches a certain final prescribed size, or a fixed number of steps is made, etc. Some types of halt rules are considered in [3].

Simplex algorithm convergence

The construction of optimization simplex algorithms encountered no constraints on the type of function $f(x)$. For example, there was no need in continuity of its derivatives. These techniques involve measurements of function values in separate points at each step followed by spacial displacement of independent variables leading to $f(x)$ value diminution. To prove the convergence, it is necessary to define the class of minimized functions and determine more exactly the conditions of function value diminution. There are different requirements to diminution. Assume that α is chosen so that the following condition holds

$$f^*(x^N) - f^*(x^{N+1}) \geq \varepsilon \|x^N - x^{N+1}\|^2,$$

where the constant $\varepsilon > 0$. Let us consider the convergence of simplex algorithms in the process of convex function minimization.

Theorem 2. Let the function $f(x)$ be convex and meet the conditions

$$\|\text{grad } f(x) - \text{grad } f(y)\| \leq L\|x - y\|, \quad x, y \in E^n,$$

the set $M(x) = \{x: f(x) \leq f(x^0)\}$ is limited.

Then, sequence $\{x^N\}$, generated by one of the simplex algorithms is minimizing and the following estimate holds true

$$f(x^N) - f_{\min} \leq \frac{f(x^1) - f_{\min}}{1 + (f(x^1) - f_{\min})CN/H_0}, \quad N = 1, 2, \dots,$$

where

$$f_{\min} = \min_{x \in E^n} f(x), \quad H_0 = \text{diam } M(x).$$

The proof of Theorem 2 is based on the relaxation theory methods [7]. The proof scheme coincides with that in [5]. The value of C depends on the $\text{Cos } \varphi_{N,i}$ value and is the larger, the larger this value is. The idea of $\text{Cos } \varphi_{N,i}$ values is given by the lemma.

Computational experiment

The use of the simplex algorithms is illustrated by the minimization of the function

$$f(x_1, \dots, x_n) = \sum_{i=1}^n ix_i^2, \quad n = 5, 10, 15, 20.$$

The centroid x^1 of the regular simplex S_1 in each case defined by $x^1 = (1, \dots, 1)$, the edge of simplex S_1 is equal to 1, $\varepsilon = 1$, $\alpha = (2., 1.4)$. The algorithms with regular simplexes

Table 1

Dimensionality	5							10						
Algorithm	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Number of steps	22	26	21	24	27	21	53	39	44	32	116	90	50	387
Number of measurements	141	125	118	151	122	94	88	461	422	309	1231	632	390	462
$\hat{f}_{\min} \times 10^4$	0.22	0.32	0.37	0.12	0.54	0.30	0.35	0.29	0.32	0.66	0.82	0.64	0.79	0.87
Dimensionality	15							20						
Algorithm	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Number of steps	75	53	54	171	189	76	1017	100	59	51	282	280	112	2463
Number of measurements	1246	803	737	2686	2116	765	1137	2141	1226	928	5801	3917	1382	2643
$\hat{f}_{\min} \times 10^4$	0.91	0.61	0.59	0.84	0.92	0.80	0.86	0.82	0.67	0.44	0.89	0.91	0.81	0.96

with reflection 1 and criteria $I_1(m, k)$, $I_2(m, k)$, $I_3(m, k)$ (algorithms 1, 2, 3), the algorithms with reflection 5 and criteria $I_1(m)$, $I_2(m)$, $I_3(m)$ (algorithms 4, 5, 6), the algorithms with reflection 5 and criterion $I_5(m)$ (algorithm 7) are used. Results are shown in Table 1.

References

1. Spendley, W., Hext, G. R., Himesworth, F. R., Sequential Application of Simplex Design in Optimization and Evolutionary Operation. *Technometrics*, vol. 4, 1962, pp. 441–461.
2. Дамбраускас А. П. Симплексный поиск. М., Энергия, 1979.
3. Емельянов С. В., Рыков А. С. Симплексные поисковые методы минимизации. *Итоги науки и техники, Техническая кибернетика*, т. 13, М., ВИНТИ, 1980, стр. 198–234.
4. Рыков А. С. Симплексные методы прямого поиска. *Техническая кибернетика, Известия АН СССР*, № 5, 1980, стр. 17–22.
5. Емельянов С. В., Коровин С. К., Рыков А. С. Локально оптимальные симплексные процедуры прямого поиска. *Проблемы управления и теории информации*, т. 9, № 2, 1980, стр. 83–101.
6. Nelder, I. A., Mead, R., A Simplex Method for Function Minimization. *The Computer Journal*, vol. 7, No. 1, 1965, pp. 308–313.
7. Любич Ю. И., Майстровский Г. Д. Общая теория релаксационных процессов для выпуклых функционалов. *Успехи математических наук*, т. XXV, вып. 1 (151), 1970, стр. 57–112.

Симплексные алгоритмы безусловной минимизации

А. С. РЫКОВ

(Москва)

Работа посвящена разработке и исследованию новых симплексных методов решения задачи безусловной минимизации n -мерной функции.

Рассматриваемые методы используют информацию только о значениях оптимизируемой функции в вершинах симплексов. Для конструирования алгоритмов вводятся различные виды отображения вершин симплекса, а также используются понятия локально оптимального направления смещения центра симплекса, правила выбора такого направления из множества возможных, правила выбора размера шага и останова процедуры поиска. Введено шесть типов критериев локальной оптимальности для направления шага. На каждом шаге минимизации для отображения выбираются симплексы, минимизирующие значение критерия локальной оптимальности. Описаны принципы конструирования алгоритмов, использующих симплексы произвольной формы. Введены правила, позволяющие существенно сократить объем вычислений на каждом шаге при выборе локально оптимального направления. Описаны процедуры изменения формы и размера симплекса при минимизации нелинейных функций.

Доказана сходимость методов при минимизации выпуклых функций, оценена скорость сходимости.

А. С. Рыков

Всесоюзный научно-исследовательский

институт системных исследований

СССР, 119034 Москва Г-34, ул. Рылеева, 29

DUALITY IN NONCONVEX PROBLEMS OF VECTOR OPTIMIZATION

A. YA. AZIMOV

(*Baku*)

(Received February 29, 1982)

In this paper with the aid of R. Rockafellar's perturbation method [1, 2] we obtain the duality theorems for nonconvex problems of vector optimization in linear spaces. General duality theorems are concretized for two important class problems. This work has become associated with papers [3-5]; all basis duality theorems of works [4, 5] we obtain as a result.

Introduction

In view of increased demands for the solutions of the problems of planning, economy and control at the recent time arose interest in theoretical research in vector optimization theory. The role of duality in extremal problems and their applications are well known [1]. Therefore, it is of no surprise that more often appear works devoted to the various aspects of duality for multicriterial problems. Note here the publications [3-7]. In [4] the multicriterial analogy of Fenchel duality theorem is obtained and in [5] this result is generalized for the nonconvex case.

In section 1 of the present paper, based on lemma one from [5] on the existence of support operator we find the conditions of subdifferentiability of operator with values in the order complete vector lattice.

In section 2 the infimum problem is considered. With the aid of perturbation of the problem we formulate the dual problem and also introduce the notion of stability problem. Here the theorem on the equivalence of stability of the problem with dual relations is being proved.

Then, in sections 3 and 4 the duality theorems for two important classes of extremal problems in nonconvex case are obtained.

Here the definitions of subdifferential, conjugate and double conjugate operators with values in the ordered spaces and also the definitions of infimum and supremum of the sets at such spaces we consider as known (see, for example, [8]).

1. Subdifferentiability of nonconvex map

Let Y and Y_0 be real linear spaces, moreover, Y_0 an ordered vector space with the positive cone K_0 , $K_0 \cap (-K_0) = \{0\}$. Furthermore, it is assumed that Y_0 is an order complete vector lattice, i.e., $\inf \{y_0, \bar{y}_0\}$ exists for all $y_0, \bar{y}_0 \in Y_0$, and for each nonempty subset B of Y_0 such that B is order-bounded from below, $\inf B$ exists [8]. Adding to space \bar{Y} the symbols $\pm \infty$, we denote this new set by $Y_0 \cup \{\infty\}$. The space of all linear mappings of Y into Y_0 denote by $E(Y, Y_0)$. The following lemma is proved in [5].

Lemma. Let $p: Y \rightarrow Y_0$ be an operator such that for all $y \in Y$ and $\lambda \in [0, \infty)$

$$p(\lambda y) = \lambda p(y). \quad (1.1)$$

Then, the following two properties are equivalent:

$$\exists T \in E(Y, Y_0): \forall y \in Y: Ty \leq p(y), \quad (1.2)$$

$$\forall y_0, \dots, y_n \in Y: y_0 + \dots + y_n = 0 \Rightarrow p(y_0) + \dots + p(y_n) \geq 0. \quad (1.3)$$

We apply this lemma to find conditions for subdifferentiability of mappings in nonconvex case.

Lemma 1.1. Given the mapping $h: Y \rightarrow Y_0 \cup \{\infty\}$. Assume $0 \in \text{intdom } h$, where $\text{intdom } h$ denotes the set of all algebraic interior points of the effective domain h . Then $\partial h(0) \neq \emptyset$ iff

$$h(0) \leq \sum_{i=1}^n \lambda_i h(y_i) \quad (1.4)$$

for all $y_1, \dots, y_n \in Y, \lambda_1, \dots, \lambda_n \in [0, 1]$ such that

$$\sum_{i=1}^n \lambda_i = 1, \quad \sum_{i=1}^n \lambda_i y_i = 0. \quad (1.5)$$

Proof. Without loss of generality, assume that $h(0) = 0$. Now suppose that (1.4) holds. Denote $B = \bigcup_{\lambda \geq 0} \{\lambda \text{ epi } h\}$. Define the set-valued mapping $W: Y \rightarrow 2^{Y_0}$ as $W(y) = \{y_0 \in Y_0 \mid (y, y_0) \in B\}$. W has the following properties:

- $0 \in W(0)$ and $y_0 \geq 0$ for all $y_0 \in W(0)$,
- $W(y) \neq \emptyset$ for all $y \in Y$,
- $W(y)$ is order-bounded from below for all $y \in Y$.

We prove only property c). If $y = 0$, c) follows from a). Now assume $y \neq 0$, and $y_0 \in W(y)$. Then, there are $\lambda \in (0, \infty)$ and $(\bar{y}, \bar{y}_0) \in Y \times Y_0$ such that

$$y = \lambda \bar{y}, \quad y_0 = \lambda \bar{y}_0, \quad h(\bar{y}) \leq \bar{y}_0. \quad (1.6)$$

Since $0 \in \text{intdom } h$, there is $\lambda_1 \in (0, \infty)$ such that $-\frac{y}{\lambda_1} \in \text{dom } h$. Then, there is $y_1 \in \text{dom } h$ such that $-y = \lambda_1 y_1$. Hence (1.6) implies

$$0 = \lambda \bar{y} + \lambda_1 y_1,$$

or

$$0 = \frac{\lambda}{\lambda + \lambda_1} \bar{y} + \frac{\lambda_1}{\lambda + \lambda_1} y_1.$$

Furthermore, it follows from (1.4) that

$$h(0) \leq \frac{\lambda}{\lambda + \lambda_1} h(\bar{y}) + \frac{\lambda_1}{\lambda + \lambda_1} h(y_1).$$

Since $h(0) = 0$, the last inequality together with (1.6) implies

$$0 \leq y_0 + \lambda_1 h(y_1),$$

or

$$y_0 \geq -\lambda_1 h(y_1).$$

This shows that $W(y)$ is order-bounded from below.

Hence the definition $p: Y \rightarrow Y_0$ as $p(y) = \inf W(y)$ makes sense, as Y_0 is an order complete vector lattice. It is not difficult to show that the mapping $p: Y \rightarrow Y_0$ satisfies conditions (1.1), (1.3). Hence there is $T \in E(Y, Y_0)$ such that

$$Ty \leq p(y)$$

for all $y \in Y$. Then the definition of the set $W(y)$ shows that

$$Ty \leq h(y)$$

for all $y \in Y$, i.e. $T \in \partial h(0)$, and the proof of Lemma 1.1 is completed.

2. Primal and dual problems

Let X, Y_0 be real linear spaces and, moreover, Y_0 is order complete vector lattice. Given mapping $F: X \rightarrow Y_0 \cup \{\infty\}$. Consider the problem.

$$(\mathcal{P}) \quad \inf \{F(x) \mid x \in X\}.$$

This problem is called primal. Let Y be another real linear space. Consider the mapping $\Phi: Y \times Y \rightarrow Y_0 \cup \{\infty\}$ such that $\Phi(x, 0) = F(x)$ for all $x \in X$. The mapping Φ is called

perturbation. Let $\Phi^*: E(X, Y_0) \times E(Y, Y_0) \rightarrow Y_0 \cup \{\infty\}$ be the conjugate map of Φ , i.e.,

$$\Phi^*(Q, T) = \sup \{Qx + Ty - \Phi(x, y) \mid (x, y) \in X \times Y\},$$

where $Q \in E(X, Y_0)$, $T \in E(Y, Y_0)$.

The problem

$$(\mathcal{P}^*) \quad \sup \{-\Phi^*(0, T) \mid T \in E(Y, Y_0)\}$$

is called the dual problem of Problem (\mathcal{P}) with respect to the given perturbation Φ .

It is easy to see that

$$\sup P^* \leq \inf P, \quad (2.1)$$

where $\sup P^*$ is the smallest upper bound in Problem (\mathcal{P}^*) , and $\inf P$ is the greatest lower bound in Problem (\mathcal{P}) .

For all $y \in Y$ define

$$h(y) = \inf \{\Phi(x, y) \mid x \in X\}. \quad (2.2)$$

Then $\inf P = h(0)$. Let us find $h^*(T)$ for any $T \in E(Y, Y_0)$:

$$\begin{aligned} h^*(T) &= \sup \{Ty - h(y) \mid y \in Y\} = \sup_y \{Ty - \inf \{\Phi(x, y) \mid x \in X\}\} = \\ &= \sup \{0 \cdot x + Ty - \Phi(x, y) \mid (x, y) \in X \times Y\} = \Phi^*(0, T). \end{aligned}$$

So, we have

$$h^*(T) = \Phi^*(0, T). \quad (2.3)$$

Hence, it follows from (2.3) that

$$\sup \mathcal{P}^* = h^{**}(0). \quad (2.4)$$

Definition 2.1. The Problem (\mathcal{P}) is said to be stable if $h(0) \in Y_0$ and $\partial h(0) \neq \emptyset$.

Lemma 2.1. Given the mapping $h: Y \rightarrow Y_0 \cup \{\infty\}$ and $h(0) \in Y_0$. Then

$$\text{if } \partial h(0) \neq \emptyset, \quad \text{then } h(0) = h^{**}(0),$$

$$\text{if } h(0) = h^{**}(0), \quad \text{then } \partial h(0) = \partial h^{**}(0).$$

Proof. Suppose that $\partial h(0) \neq \emptyset$ and $T \in \partial h(0)$. Then for each $y \in Y$ $h(y) - h(0) \geq Ty$. This implies $h(0) = -h^*(T)$. It follows from (2.3), (2.4) that $h^{**}(0) \geq -h^*(T)$. The last two relations imply $h^{**}(0) \geq h(0)$. So using (2.1) we obtain $h(0) = h^{**}(0)$.

Now prove the second part. Suppose $h(0) = h^{**}(0)$. We must show that $\partial h(0) = \partial h^{**}(0)$. Take $T \in \partial h(0)$. Then, $h(0) = -h^*(T)$. We have for any $y \in Y$ $h^{**}(y) \geq$

$\geq Ty - h^*(T) = Ty + h(0)$. Since, by the condition $h(0) = h^{**}(0)$, the last inequality implies $T \in \partial h^{**}(0)$. Thus we get $\partial h(0) \subseteq \partial h^{**}(0)$. The inverse inclusion is obvious.

Lemma 2.2. The set of solutions of Problem (\mathcal{P}^*) is equal to $\partial h^{**}(0)$.

Proof. Suppose $T \in E(Y, Y_0)$ is a solution of Problem (\mathcal{P}^*) . It means that $h^{**}(0) = -h^*(T)$. The definition of h^{**} implies $h^{**}(y) \geq Ty - h^*(T)$ for all $y \in Y$. Hence we have $h^{**}(y) - h^{**}(0) \geq Ty$, i.e. $T \in \partial h^{**}(0)$.

Now assume that $T \in \partial h^{**}(0)$. It means that $h^{**}(y) - h^{**}(0) \geq Ty$ for all $y \in Y$. Since $h(y) \geq h^{**}(y)$, we have $h(y) - h^{**}(0) \geq Ty$ for any $y \in Y$. This yield $-h^{**}(0) \geq h^*(T)$. On the other hand (2.3) and (2.4) imply $h^{**}(0) \geq -h^*(T)$. So we have $h^{**}(0) = -h^*(T)$. It means that the mapping T is a solution of Problem (\mathcal{P}^*) .

The next theorem follows from Lemma 2.1 and Lemma 2.2.

Theorem 2.1. The following conditions are equivalent:

1. Problem (\mathcal{P}) is stable,
2. Problem (\mathcal{P}^*) has solution, $\inf P = \sup P^*$ and this value is finite. The following theorem gives a simple criterion of the stability of Problem (\mathcal{P}) .

Theorem 2.2. Suppose $\inf P$ is finite and there are $x_0 \in X$ and neighbourhood V of the origin in Y such that $\Phi(x_0, y) < +\infty$ for all $y \in V$. Assume that for all $x_1, \dots, x_n \in X$, $y_1, \dots, y_n \in Y$ and $\lambda_1, \dots, \lambda_n \in [0, 1]$ such that $\sum_{i=1}^n \lambda_i = 1$ and $\sum_{i=1}^n \lambda_i y_i = 0$ there is $\bar{x} \in X$ such that

$$\Phi(\bar{x}, 0) \leq \sum_{i=1}^n \lambda_i \Phi(x_i, y_i). \tag{2.5}$$

Then Problem (\mathcal{P}) is stable.

Proof. From the condition of the theorem we have $h(0) \in Y_0$ and $0 \in \text{intdom } h$. Thus, for stability of Problem (\mathcal{P}) we need to show that condition (1.4) of Lemma 1.1 is satisfied. Assume $y_1, \dots, y_n \in Y$, $\lambda_1, \dots, \lambda_n \in [0, 1]$ satisfy condition (1.5). Let x_1, \dots, x_n be any elements of X . Then by condition there is $\bar{x} \in X$ such that inequality (2.5) holds. This inequality implies

$$h(0) \leq \sum_{i=1}^n \lambda_i \Phi(x_i, y_i).$$

Since vectors $x_1, \dots, x_n \in X$ are arbitrary and independent, the last inequality implies (1.4). Thus the theorem is proved.

3. Particular case (1)

In this section the problem is considered which is in the scalar case closely connected with calculus of variation [1]. Let X, Y, Y_0 be real linear spaces and moreover, Y_0 is order complete vector lattice. Given mapping $J: X \times Y \rightarrow Y_0 \cup \{\infty\}$ and

linear operator $A: X \rightarrow Y$. Suppose that minimized function F has the form

$$F(x) = J(x, Ax). \quad (3.1)$$

Thus, the primal Problem (\mathcal{P}) is the following:

$$\inf \{J(x, Ax) | x \in X\}. \quad (3.2)$$

We introduce the perturbed mapping as

$$\Phi(x, y) = J(x, Ax - y). \quad (3.3)$$

It is easy to show that the dual Problem (\mathcal{P}^*) in this case is

$$\sup \{-J^*(TA, -T) | T \in E(Y, Y_0)\}, \quad (3.4)$$

where $J^*: E(X, Y_0) \times E(Y, Y_0) \rightarrow Y_0 \cup \{\infty\}$ is the conjugate mapping of J .

We have the following stability criterion for Problem (3.2).

Theorem 3.1. Assume $\inf \{J(x, Ax) | x \in X\}$ is finite and there are $x_0 \in X$ and neighborhood V of the origin in Y such that $J(x_0, Ax_0) \in Y_0$ and $J(x_0, y) < +\infty$ for all $y \in Ax_0 + V$. Suppose for all $x_1, \dots, x_n \in X, y_1, \dots, y_n \in Y$ and $\lambda_1, \dots, \lambda_n \in [0, 1]$ such that

$$\sum_{i=1}^n \lambda_i = 1, \quad \sum_{i=1}^n \lambda_i y_i = A \sum_{i=1}^n \lambda_i x_i \quad (3.5)$$

there exists $\bar{x} \in X$ such that

$$J(\bar{x}, A\bar{x}) \leq \sum_{i=1}^n \lambda_i J(x_i, y_i). \quad (3.6)$$

Then, Problem (3.2) is stable.

Proof. Without loss of generality, we consider the case when the set V is symmetric. Assume $y \in V$. Then $Ax_0 - y \in Ax_0 + V$. We have $\Phi(x_0, y) = J(x_0, Ax_0 - y) < +\infty$.

Let $x_1, \dots, x_n \in X, y_1, \dots, y_n \in Y$ and $\lambda_1, \dots, \lambda_n \in [0, 1]$ be such that

$$\sum_{i=1}^n \lambda_i = 1, \quad \sum_{i=1}^n \lambda_i y_i = 0.$$

Denote $z_i = Ax_i - y_i$. Hence we get $\sum_{i=1}^n \lambda_i z_i = A \sum_{i=1}^n \lambda_i x_i$. Then, by the condition there is $\bar{x} \in X$ such that

$$J(\bar{x}, A\bar{x}) \leq \sum_{i=1}^n \lambda_i J(x_i, z_i).$$

Hence, we have

$$\Phi(\bar{x}, 0) \leq \sum_{i=1}^n \lambda_i J(x_i, Ax_i - y_i) = \sum_{i=1}^n \lambda_i \Phi(x_i, y_i).$$

Thus, all conditions of Theorem 2.2 are satisfied, hence Problem (3.2) is stable. Theorem 3.1 is proved.

Now assume $J(x, Ax) = F(x) + G(Ax)$, where $F: X \rightarrow Y_0 \cup \{+\infty\}$, $G: Y \rightarrow Y_0 \cup \{+\infty\}$ and $A \in E(X, Y)$. So the primal Problem (\mathcal{P}) has the form:

$$\inf \{F(x) + G(Ax) \mid x \in X\}. \quad (3.7)$$

Hence we get easily that the dual Problem (\mathcal{P}^*) is

$$\sup \{-F^*(TA) - G^*(-T) \mid T \in E(Y, Y_0)\}. \quad (3.8)$$

The next theorem is the consequence of Theorem 3.1.

Theorem 3.2. Let $\inf \{F(x) + G(Ax) \mid x \in X\}$ be finite and $x_0 \in X$ and the neighborhood V of the origin in Y such that $F(x_0) < +\infty$, $G(Ax_0) < +\infty$ and $G(y) < +\infty$ for all $y \in Ax_0 + V$. Assume that for all $x_1, \dots, x_n \in X$, $y_1, \dots, y_n \in Y$ and $\lambda_1, \dots, \lambda_n \in [0, 1]$ which satisfied condition (3.5) there exists $\bar{x} \in X$ such that

$$\sum_{i=1}^n \lambda_i [F(x_i) - F(\bar{x})] \geq \sum_{i=1}^n \lambda_i [G(A\bar{x}) - G(y_i)]. \quad (3.9)$$

Then, Problem (3.7) is stable.

If in Problem (3.7) the linear operator A is identity, then we get vector analogy of the Fenchel problem. If we concretize Theorem 3.2 for this particular case then, as a matter of fact, we obtain the duality theorem from [5].

Remark 3.1. If in (2.5), (3.6) and (3.9) we replace the vector $\Phi(\bar{x}, 0)$ with $h(0) = \inf \{\Phi(x, 0) \mid x \in X\}$, then all theorems of the present paper remain in force. In this case, for instance, condition (3.9) coincides with the analogous one from [5] and we obtain (in the case when $A = I$) the duality theorem [5] precisely.

4. Particular case (2)

Suppose X, Y_0, Y_1 and Y_2 are real linear spaces, moreover, Y_0 an order complete vector lattice. Assume that Y_1 is also an ordered space with cone K_1 . Given the nonempty set $A \subseteq X$, the mapping $F_0: A \rightarrow Y_0$, $F_1: A \rightarrow Y_1$ and $F_2: A \rightarrow Y_2$.

Denote $Y = Y_1 \times Y_2$.

Consider the problem

$$\inf \{F_0(x) \mid x \in D\}, \quad (4.1)$$

where $D = \{x \in A \mid F_1(x) \leq 0, F_2(x) = 0\}$.

For this problem the perturbed mapping is following:

$$\Phi(x, y) = \begin{cases} F_0(x), & \text{if } x \in A, F_1(x) \leq y_1, F_2(x) = y_2, \\ +\infty & \text{in other cases,} \end{cases}$$

where $y = (y_1, y_2) \in Y$.

It is not difficult to show that the dual problem associated with primal Problem (4.1) is

$$\sup_{\substack{T_1 \geq 0 \\ T_2 \in E(Y_2, Y_0)}} \inf_{x \in A} \{F_0(x) + T_1 F_1(x) + T_2 F_2(x)\}, \quad (4.3)$$

where $T_1 \geq 0$ means $T_1(K_1) \subseteq K_0$.

Now we prove the theorem of stability of Problem (4.1).

Theorem 4.1. Assume $\inf \{F_0(x) | x \in D\}$ is finite and the following conditions 1–3 are satisfied:

1) there is the set $S \in A$ such that $V_2 \equiv F_2(S)$ absorbs each element of the set $R \equiv \text{lin } F_2(A)$;

2) there exists $\bar{y}_1 \in K_1$ such that for all $x \in S$

$$F_1(x) \leq -\bar{y}_1$$

and the set $K_1 - \bar{y}_1$ absorbs each element of the set $\text{lin}(F_1(A) + K_1)$;

3) for all $x_1, \dots, x_n \in A, \lambda_1, \dots, \lambda_n \in [0, 1], \sum_{i=1}^n \lambda_i = 1$ there is $x \in A$ such that

$$F_j(\bar{x}) \leq \sum_{i=1}^n \lambda_i F_j(x_i), \quad j=0, 1,$$

$$F_2(\bar{x}) = \sum_{i=1}^n \lambda_i F_2(x_i).$$

Then, Problem (4.1) is stable.

Proof. We must show that $\partial h(0) \neq \emptyset$, where $h(y) = \inf \{\Phi(x, y) | x \in X\}$. Denote $\bar{Y} \sim \text{lin}(F_1(A) + K_1)$ and $V_1 = K_1 - \bar{y}_1$. Since, by the condition, V_1 absorbs the set \bar{Y}_1 and V_2 absorbs the space R , then $V_1 \times V_2$ is an absorbed set in the space $\bar{Y}_1 \times R$.

Let $\bar{h}: \bar{Y}_1 \times R \rightarrow Y_0 \cup \{\infty\}$ be contracted operator of $h: Y \rightarrow Y_0 \cup \{\infty\}$. Let us show that $V_1 \times V_2 \subset \text{dom } \bar{h}$. Suppose $(y_1, y_2) \in V_1 \times V_2$. From $y_2 \in V_2$ it follows that there exists $x \in S$ such that $y_2 = F_2(x)$. Moreover, condition 2) implies $F_1(x) \leq -\bar{y}_1$. Hence, for all $k_1 \in K_1$ we have $F_1(x) \leq \bar{y}_1 + k_1$. Since $y_1 \in K_1 - \bar{y}_1$, then the last inequality implies $F_1(x) \leq y_1$. Thus, for each $y = (y_1, y_2) \in V_1 \times V_2$ there exists $x \in S$ such that

$$x \in A, \quad F_1(x) \leq y_1, \quad F_2(x) = y_2.$$

Thus, due to (4.2) we have

$$\Phi(x, y) = F_0(x).$$

Hence, for all $y = (y_1, y_2) \in V_1 \times V_2$ the vector $h(y) < +\infty$. Since, obviously, $V_1 \times V_2 \subset \bar{Y}_1 \times R$, then $h(y) < +\infty$ implies $V_1 \times V_2 \subset \text{dom } h$.

Now, assume $y_1, \dots, y_n \in \bar{Y}_1 \times R, \lambda_1, \dots, \lambda_n \in [0, 1]$, where $y_i = (y_1^i, y_2^i), y_1^i \in \bar{Y}_1, y_2^i \in R$ for $i = \overline{1, n}$ is such that

$$\sum_{i=1}^n \lambda_i = 1, \quad \sum_{i=1}^n \lambda_i y_i = 0.$$

We prove that for $\bar{h}: \bar{Y}_1 \times R \rightarrow Y_0 \cup \{\infty\}$ (1.4) holds. First, we note that if for some $i = \overline{1, n}$ $\bar{h}(y_i) = +\infty$, then (1.4) trivially holds. Therefore, we consider that for all $i = \overline{1, n}$ $\bar{h}(y_i) < +\infty$. Then, for all $i = \overline{1, n}$, there exists $x_i \in A$ such that

$$F_1(x_i) \leq y_1^i, \quad F_2(x_i) = y_2^i, \quad i = \overline{1, n}. \tag{4.4}$$

These relations imply

$$\sum_{i=1}^n \lambda_i F_1(x_i) \leq 0, \quad \sum_{i=1}^n \lambda_i F_2(x_i) = 0.$$

Then, by condition 3) of the theorem there is $\bar{x} \in A$ such that

$$F_0(\bar{x}) \leq \sum_{i=1}^n \lambda_i F_0(x_i), \quad F_1(\bar{x}) \leq 0, \quad F_2(\bar{x}) = 0.$$

According to the definition of the operator h the first inequality implies

$$h(0) \leq \sum_{i=1}^n \lambda_i F_0(x_i).$$

This relation holds for all $x_i \in A$ satisfied (4.4). Since x_1, \dots, x_n are independent, hence we have

$$\bar{h}(0) \leq \sum_{i=1}^n \lambda_i \bar{h}(y_i).$$

Thus, the operator $\bar{h}: \bar{Y}_1 \times R \rightarrow Y_0 \cup \{\infty\}$ satisfies the condition of Lemma 1.1. Then by this lemma there is $\bar{T} \in E(\bar{Y}_1 \times R, Y_0)$ such that for all $y \in \bar{Y}_1 \times R$

$$\bar{h}(y) - \bar{h}(0) \geq \bar{T}y. \tag{4.5}$$

It is not difficult to see that if $y \in Y(\bar{Y}_1 \times R)$ then $h(y) = +\infty$. Therefore, if $T \in E(Y_1 \times Y_2, Y_0)$ is any linear extension of \bar{T} , then (4.5) implies

$$h(y) - h(0) \geq Ty,$$

i.e., $T \in \partial h(0)$. The proof of Theorem 4.1 is completed.

References

1. Ekeland, I., Temam, R., *Convex Analysis and Variational Problems*. (Russian). M. "Mir", 1979.
2. Rockafeller, R., *Convex Analysis*. (Russian) M. "Mir", 1973.
3. Zowe, J., *The Saddle Point Theorem of Kuhn and Tucker in Ordered Vector Spaces*. *J. Math. Anal. and Appl.*, 1977, **57**, 41–55.
4. Zowe, J., *A Duality Theorem for a Convex Programming Problem in Order Complete Vector Lattice*. *J. Math. Anal. and Appl.*, 1975, **50**, 273–287.
5. Rosinger, E., *Multiobjective Duality without Convexity*. *J. Math. Anal. and Appl.*, 1978, **66**, 442–450.
6. Rosinger, E., *Duality and Alternative in Multiobjective Optimization*. *Proc. Amer. Math. Soc.*, 1977, **64**, 307–312.
7. Ritter, K., *Optimization Theory in Linear Spaces. Part III. Mathematical Programming in Partially Ordered Banach Spaces*. *Math. Annal.*, 1970, **184**, 133–154.
8. Akilov, G. P., Kutateladze, S.S., *Ordered Vector Spaces*. (Russian). Novosibirsk, "Nauka", 1978.

Двойственность в невыпуклых задачах векторной оптимизации

А. Я. АЗИМОВ

(Баку)

В статье изучается двойственность для невыпуклых задач векторной оптимизации в линейных пространствах. Следуя методу возмущений Р. Рокафеллара, данная задача включается в семейство возмущенных задач. С помощью функции возмущения формулируется двойственная задача. Вводится понятие стабильности задачи и доказывается теорема об эквивалентности понятия стабильности двойственным соотношениям. Найден простой критерий стабильности задачи.

Полученные общие соотношения применяются к двум важным случаям. Первый случай, который тесно связан с вариационным исчислением, охватывает, в частности, задачи, изученные в скалярном случае впервые Фенхелем и Рокафелларом. Теоремы двойственности для этих задач, полученные в настоящей работе, включают ранее известные результаты.

Во втором случае рассматривается задача на инфимум с ограничениями, где инфимизируемый оператор принимает значения в частично упорядоченном пространстве. При помощи функции возмущения данной задаче сопоставляется двойственная задача, использующая функцию Лагранжа. Для этой задачи найдены условия, при выполнении которых доказывается теорема о стабильности.

А. Я. Азимов

Азербайджанский Государственный Университет
СССР, Баку 73, ул. П. Лумумбы, 23

ANALYSIS OF QUEUEING NETWORKS BY POLYNOMIAL APPROXIMATION

A. I. GERASIMOV

(Moscow)

(Received April 24, 1982)

In this paper a new approximate method is proposed based on polynomial approximation and intended for the analysis of arbitrarily-structured open, closed and mixed queueing networks with several classes of customers, several subchains, arbitrary service time distributions, priorities and blockings due to limited queues at the nodes.

Development of approximate methods for the study of stochastic networks stems from the necessity to deal with various problems occurring in design and maintenance of computer networks and systems [1-4].

The basic notions, connected with stochastic (queueing) networks and used in the paper may be found in [5, 6].

Conventionally, the existing approximate methods may be classified into four categories: those based on equivalent flows (aggregation), [7], diffusion approximation [8], mean-value analysis [9], dynamics of mean values [10].

Apart from this classification stand the method proposed by P. J. Kuchn [11] and that used in the present study which assume decomposition of the network into isolated subsystems. It should be noted that the method in [11] was developed only for open one-class networks. For closed networks the method leads to significant errors and, therefore, cannot be used for the analysis of such networks.

The present publication which is the continuation of the studies published in [12 to 16] proposes an approximate method based on polynomial approximation, intended for the analysis of open, closed and mixed queueing networks of arbitrary structure having several classes of customers, several subchains, arbitrary functions of service time distribution in nodes, priorities and blockings due to limited queues in nodes.

The principle of the approximation (stated in (4) through (5)) consists in writing a system of polynomial equations with respect to the mean intensities of flows of messages belonging to different classes and passing through nodes in closed subchains of the network under consideration. In doing so, the mean time of the message stay in nodes is estimated by means of a rational approximation which is analogous to the well

known formulas [17] for open queueing systems, and then the Little formula is applied [18].

Equations are constructed so that under small and high loads their solutions coincide asymptotically with the exact values for mean flow intensities.

Numerous verifications have born out that this method has satisfactory accuracy under arbitrary loads of the network.

Now, other networks characteristics may be estimated (mean, variance and other moments¹) through formulas similar to (2).

Queueing networks with several classes of customers. Consideration is given to a network of arbitrary structure having M nodes and circulating messages of R classes.

During waiting or service, each customer belongs to a certain class of customers.

Different customers may have different routes through the network and different service time distributions at a given node. Customers may change their classes when passing from one node to another.

Messages of the class r pass from the node i to node j and class s with probability $P_{ir,js}$.

Service time distribution functions at nodes $F_{ir}(t)$ ($i = \overline{1, M}; r = \overline{1, R}$) with finite first two moments $a_{ir} = \int_0^{\infty} t dF_{ir}(t)$ $\sigma_{ir}^2 = \int_0^{\infty} (t - a_{ir})^2 dF_{ir}(t)$ may be arbitrary. The transition probability matrix $\|P_{ir,js}\|$ defines a Markov chain (without absorbing states if messages may leave the network) whose states are marked by the pairs (i, r) . It is assumed that this Markov chain is decomposable into D ergodic subchains E_k ($k = \overline{1, D}$). Then, the following holds for each ergodic subchain E_k

$$\sum_{(i,r) \in E_k} U_{ir} P_{ir,js} + Q_{js} = U_{js}; \quad (j, s) \in E_k \quad (1)$$

where U_{ir} is the mean intensity of the flow of messages of class r leaving the i -th node; Q_{js} is the mean intensity of the flow of messages of class s coming to the j -th node from an external source.

If one introduces the coefficients $d_{ir} = U_{ir}/U_k$ ($k = \overline{1, D}$) (where $U_k = U_{i_k r_k}$ is the mean intensity of some flow $(i_k, r_k) \in E_k$) chosen in each subchain) and takes into consideration the fact in the case of closed subchain (given $Q_{js} = 0 \forall (j, s) \in E_k, k = \overline{1, D}$) the range of system (1) is equal to its dimensionality minus one, system (1) will have for closed subchains a unique solution with respect to coefficients d_{ir} .

¹ This paper gives estimates only for mean times of message stay in nodes for various classes of messages. Estimates of other moments for various network characteristics are obtained through the known formulas (e.g., see [17]) in a similar manner.

For open subchains E_k , $k = \overline{L+1, D}$ (for which $\exists(j, s) \in E_k : Q_{js} \neq 0$) mean intensity is uniquely defined by (1).

The proposed method is based on rational approximation which is an analogue of the Polyachek-Khinchin formula [17], for the estimate of the mean time of message

$$x_{ir} = a_{ir} + (1/2)\gamma \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is}' U_j \right) \cdot a_i^2 (1 + (\sigma_i/a_i)^2) / \left(1 - \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j \right) a_i \gamma \right), \quad (2)$$

where

$$a_i = \sum_{r=1}^R \left(U_{ir} a_{ir} / \sum_{r=1}^R U_{ir} \right); \quad \sigma_i^2 = \sum_{r=1}^R \left(U_{ir} (a_{ir}^2 + \sigma_{ir}^2) / \sum_{r=1}^R U_{ir} \right) - a_i^2,$$

N_k is the number of messages circulating in the closed subchain E_k ($k = \overline{1, L}$);

$N = \sum_{k=1}^L N_k$; $\gamma = (N-1)/N$ if there are closed subchains, and $\gamma = 1$, otherwise.

Note that the chosen rational approximation has the following asymptotic features:

$$\begin{aligned} x_{ir} &\rightarrow a_{ir} & \text{for } N &\rightarrow 1; \\ x_{ir} &\rightarrow \infty & \text{for } \sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j &\rightarrow 1/a_i. \end{aligned}$$

The total number of messages is represented for the closed subchain E_k by means of the Little formula [18] as

$$N_k = \sum_{i=1}^M \sum_{r \in E_k} x_{ir} d_{ir} U_k, \quad (k = \overline{1, L}). \quad (3)$$

Then the problem of approximate analysis of queueing network with L closed subchains may be posed as that of solving the following system of equations

$$P_k(U_1, \dots, U_L) = 0 \quad k = \overline{1, L}, \quad (4)$$

with respect to the variables (mean intensities of the flows U_1, \dots, U_L where $P_k(U_1, \dots, U_L)$ are polynomials of degree $M+1$ with coefficients depending on U_{L+1}, \dots, U_D and N):

$$\begin{aligned} P_k(U_1, \dots, U_L) = & U_k \left\{ \prod_{i=1}^M \left(1 - \gamma a_i \sum_{j=1}^D \sum_{r \in E_j} d_{ir} U_j \right) \right\} \left\{ \sum_{i=1}^M \sum_{r \in E_k} d_{ir} a_{ir} \right\} + \\ & + (1/2)\gamma U_k \sum_{i=1}^M \left\{ \left(\prod_{\substack{j=1 \\ j \neq i}}^M \left(1 - \gamma a_j \sum_{l=1}^D \sum_{r \in E_l} d_{jr} U_l \right) \right) \right. \\ & \cdot \left. \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j \right) (a_i^2 + \sigma_i^2) \left(\sum_{r \in E_k} d_{ir} \right) \right\} - N_k \prod_{i=1}^M \left(1 - \gamma a_i \sum_{j=1}^D \sum_{r \in E_j} d_{ir} U_j \right) \end{aligned}$$

The form of the polynomials is selected so that in the domain of definition of the variables U_j

$$0 < \gamma \sum_{j=1}^D \left(U_j \left(\sum_{r \in E_j} d_{ir} a_{ir} \right) \right) < 1 \quad (i = \overline{1, M}) \quad (5)$$

the solution of the system¹ (4) tends² at $N \rightarrow 1$ and $N \rightarrow \infty$ to the exact values of the mean intensities of some chosen flows U_i ($i = \overline{1, L}$).

The resulting system of polynomial equations (4) enables one to take into consideration the mutual influence of flows (mean intensities of flows through the nodes), the system properties of closed networks whose most prominent feature is abundance of feedbacks.

Closed networks with one class of messages form an important special case of queueing networks with several classes of messages. In this case, system (4) boils down to a single equation.

$$P(U) = 0 \quad (6)$$

provided that

$$0 < U < 1 / (\gamma \max_i \{d_i a_i\}), \quad (7)$$

where U is the expectation of the intensity of the flow of messages leaving some (e.g., last) chosen node; $P(U)$ is a polynomial of degree $M + 1$ with coefficients dependent on N :

$$P(U) = U \prod_{i=1}^M (1 - c_i U) \left(\sum_{i=1}^M d_i a_i \right) + U^2 \sum_{i=1}^M b_i d_i \prod_{\substack{j=1 \\ j \neq i}}^M (1 - c_j U) - N \prod_{i=1}^M (1 - c_i U),$$

$$b_i = (1/2) d_i a_i^2 \gamma (1 + (\sigma_i / a_i)^2), \quad c_i = d_i a_i \gamma.$$

Note that $P(U)$ is a continuous monotonous function changing sign within domain (7).

Queueing networks with priorities. Of great interest for study and practical application are queueing networks with priority service in nodes.

¹ Note that each function $P_k(U_1, \dots, U_L)$ is (i) continuous over all U_i , (ii) varies from $-\infty$ to $+\infty$ when U_k varies from 0 to the right boundary of the domain (5), and (iii) grows monotonically over all the U_i . One may readily demonstrate by induction with respect to L that the existence of solution of the equation system (4), (5) follows from the conditions (i) to (iii).

² For $N \rightarrow 1$, this follows from (3) and $x_{ir} \rightarrow a_{ir}$. For $N \rightarrow \infty$ this will be strictly proved below when estimating the method accuracy for closed networks with a single class of messages; the proof for each separate subchain with several classes of messages is similar to that used within terminology.

In the case under consideration, x_{ir} is an analogue of the well-known formulas for estimating mean time of message of class r being in node i under stationary mode and at priority service.

It is assumed that the following priority servicing disciplines may be used in network nodes: nonpreemptive policy, preemptive resume policy, preemptive-repeat-different policy, preemptive-repeat-identical policy.

For the sake of illustration, we present a rational estimate x_{ir} of the preemptive resume policy

$$x_{ir} = \left\{ (1/2)\gamma \sum_{j=1}^D \sum_{\substack{s \in E_j \\ s \neq r}} (\eta_{isr} d_{is} U_j (a_{is}^2 + \sigma_{is}^2)) \right\} / \left\{ \left(1 - \gamma \sum_{\substack{j=1 \\ s \neq r}}^D \sum_{s \in E_j} (\eta_{isr} d_{is} U_j a_{is}) \right) \cdot \left(1 - \gamma \sum_{j=1}^D \sum_{s \in E_j} (\eta_{isr} d_{is} U_j a_{is}) \right) \right\} + a_{irr} \left\{ 1 - \gamma \sum_{j=1}^D \sum_{\substack{s \in E_j \\ s \neq r}} (\eta_{isr} d_{is} U_j a_{is}) \right\},$$

where $\eta_{ir_1 r_2} = 1$ if in the i -th node the priority of messages of class r_1 is equal to or higher than that of class r_2 , otherwise $\eta_{ir_1 r_2} = 0$.

The existence condition for stationary mode in i -th node

$$0 < \gamma \left(\sum_{j=1}^D \sum_{s \in E_j} [\eta_{ir_{i,m}} s (d_{is} U_j) a_{is}] \right) < 1,$$

where $\eta_{isr_{i,m}} \geq \eta_{isr} \forall r, s$, $r_{i,m}$ being the lowest priority. Network stationarity condition is the totality of stationarity conditions of each node. Non-priority servicing disciplines may be employed in some nodes of the queueing network. In this case, x_{ir} is determined from (2).

Exponential queueing networks with several classes of customers and blockings due to limited queues in nodes. Consideration is given to a network with arbitrary structure where blockings occur because of the limited size t_j of queues in the nodes. Message service times are distributed in nodes according to the exponential law with expectations a_{ir} ($i = \overline{1, M}$; $r = \overline{1, R}$).

For open subchains E_k ($k = \overline{L+1, D}$) the mean intensity of flow U_k is uniquely defined from (1) given (8).

The time of messages of class r being in node i is estimated by the following rational approximation:

$$x_{ir} = a_{ir} + F_{ir}^{(n_*)} + \gamma \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j \right)^2 (a_i^{(n_*)})^2 / \left\{ 1 - \gamma \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j \right) a_i^{(n_*)} \right\}.$$

$F_{is}^{(n)}$ is the analogue of the well known formulas for estimating mean time of blocking messages of class s leaving node i , and is determined by rapidly converging iterative process (n is the iteration number, n_* is that of the last iteration) in the following manner:

$$F_{is}^{(0)} = 0; \quad a_i^{(n)} = \sum_{s=1}^R \left(U_{is} (a_{is} + F_{is}^{(n)}) \left/ \sum_{s=1}^R U_{is} \right. \right);$$

$$F_{is} = \sum_{(j,v) \in E_k(i,s)} P_{is,jv} f_j^{(n)} \quad (\text{if } n > 0); \quad f_i^{(n)} = 0; \quad \text{if } j \neq i,$$

$$f_j^{(n)} = \gamma b_j^{(n)} / (\eta_j^{(n)} (1 - \gamma b_j^{(n)})); \quad b_j^{(n)} = (\eta_j^{(n)} a_j^{(n)})^{t_j+1} \left/ \sum_{i=0}^{t_j+1} (\eta_j^{(n)} a_j^{(n)})^i \right.;$$

$$\eta_j^{(n)} = \sum_{i \in \Omega_j} (1/a_i^{(n)}),$$

where $E_k(i, s)$ is the subchain to which (i, s) belongs; Ω_j is the set of all the nodes from which a message of any class may go to the node j . F_{is} may be estimated non-iteratively by means of similar formulas at $\eta_j = \sum_{s=1}^R U_{js}$.

For closed subchains a system of nonlinear equations of the type of (4) may be obtained from (3) provided that

$$0 < \gamma \sum_{j=1}^D U_j \left\{ \sum_{r \in E_j} d_{ir} (a_{ir} + F_{ir}^{(n_*)}) \right\} < 1,$$

$$\gamma \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j \right)^3 (a_i^{(n_*)})^2 \leq t_i \left(1 - \gamma \left(\sum_{j=1}^D \sum_{s \in E_j} d_{is} U_j \right) a_i^{(n_*)} \right),$$

$$i = \overline{1, M}. \quad (8)$$

Estimation of polynomial approximation accuracy. Let us prove that the approximate method provides asymptotic coincidence of its results with exact values under small and great loads in arbitrary-structure closed queueing networks with one class of messages and arbitrary service time distributions in the nodes.

Obviously, Eqs (6) and (7) give exact value of U at $N = 1$. Let us demonstrate that at $N \rightarrow \infty$ the solution of Eqs (6) and (7) asymptotically coincides with the exact one. Indeed, if one chose node i_* for which $d_{i_*} a_{i_*} = \max_i d_i a_i$ holds, and if one expresses all the mean flow intensities in terms of the mean intensity of the flow leaving the node (i.e.

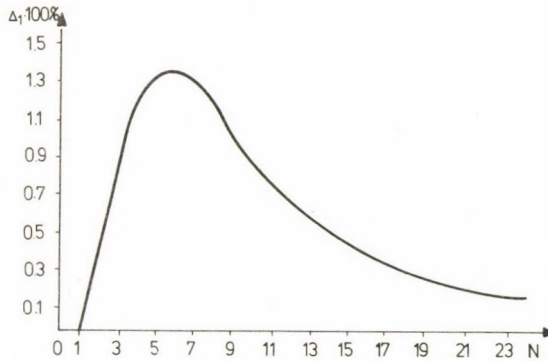


Fig. 1. Absolute error of the utilization factor (occupancy probability) of the first node in a closed exponential network with one class of messages.

Network parameters: $M = 3$, $a_1 = 28$, $a_2 = 40$, $a_3 = 280$, $P_{11} = 0.1$, $P_{12} = 0.7$, $P_{13} = 0.2$, $P_{21} = 1$, $P_{31} = 1$, the rest of $P_{ij} = 0$.

$d_{i_0} = 1$), the exact value of the mean intensity of the chosen flow is known to be $U \xrightarrow{N \rightarrow \infty} 1/a_{i_0}$. On the other hand, $U \xrightarrow{N \rightarrow \infty} 1/a_{i_0}$ follows from the form of Eqs (6), (7) if one takes into consideration that U is limited in virtue of the condition (7). Comparison of numerous applications of the proposed method with exact methods and simulation of closed queueing networks with one class of messages, with several classes of messages and several subchains, of queueing networks with priority servicing and blockings due to limited queues in nodes, has demonstrated that in all cases accuracy was satisfactory and solutions coincided asymptotically under small and great loads.

The above is illustrated in Fig. 1 by the absolute value of the first node occupancy probability error vs. number of messages (N), circulating in closed exponential network with one class of messages which may be interpreted as a model of multi-program computer systems [7, 19] where node 1 represents operation of the Central Processing Unit, and nodes 2 and 3 represent I/O operation. Exact solution for the network under consideration is determined by Buzen algorithm [19]. As may be seen from the figure, the maximal absolute error does not exceed 1.5 percent.

Computational performance of the method. The amount of computations required for the implementation of the method coincides with the number of operations required for the solution of Eqs (6) and (7) or the system of equations (4), (5). Computational complexity was estimated for linear exponential networks with one class of messages for which effective computer algorithms have been developed and number of arithmetic operations is known. One may readily obtain in this case the

upper bound of the number of arithmetic operations g required for the proposed method:

$$g = 3(M + 1) \log_2 [(M - 1)a_{i_*}/\varepsilon]$$

where ε is the desired accuracy of solution, i_* is the node number:

$$d_{i_*} a_{i_*} = \max_i d_i a_i.$$

Assuming that $N > M$ and noting that $\lim_{M \rightarrow \infty} (1/M) \log_2 [(M - 1)a_{i_*}/\varepsilon] = 0$ we obtain that $g \ll MN$.

Thus, computationally the proposed method as applied to highdimensionality networks is superior to the algorithm of Buzen [19] requiring $2MN$ arithmetic operations.

The system of non-linear equations (4), (5) may be solved by means of the steepest descent method which converges from any initial approximation in the domain (5) to the solution of the non-linear system (4). This follows from the necessary extremum condition for multi-variable function $\sum_{i=1}^L P_i^2$ and from the monotonicity of functions P_i with respect to each variable.

References

1. Chandy, K. M., Sauer, C. H., Approximate methods for analyzing queueing network models of computing systems, *Computing Surveys*, **10**, 3, 1978, pp. 281-317.
2. Sauer, C. H., Chandy, K. M., Approximate solution of queueing models, *Computer*, 1980, **4**, pp. 25-32.
3. Allen, A. O., Queueing models of computer systems, *Computer*, 1980, **4**, pp. 13-24.
4. Kienze, M. G., Sevcik, K. C., Survey of analytic queueing network models of computer systems, *Performance Evaluation Review*, **8**, 3, 1979, pp. 113-129.
5. Baskett, F., Chandy, K. M., Muntz, R. R., Palacios, F. G., Open, closed and mixed networks of queues with different classes of customers, *Journal of the Association for Computing Machinery*, **22**, 2, 1975, pp. 248-260.
6. Reiser, M., Kobayashi, H., Queueing networks with multiple closed chains: theory and computational algorithms, *IBM Journal of res. and development*, **19**, 3, 1975, pp. 283-294.
7. Chandy, K. M., Herzog, V., Woo, L. S., Parametric analysis of queueing networks, *IBM Journal of research and development*, **19**, 1, 1975, pp. 43-53.
8. Kobayashi, H., Application of the diffusion approximation to queueing networks I. Equilibrium queue distributions, *Journal of the ACM*, **21**, 2, 1974, pp. 316-328.
9. Reiser, M., Lavenberg, S., Mean-value analysis of closed multichain queueing networks, *Journal of the ACM*, **27**, 2, 1980, pp. 313-322.
10. Kagan, B. M., Karacharov, A. F., Probability models of information and computer networks based on mean-value dynamics, *Automation and Remote Control*, **3**, 1975, pp. 153-162.
11. Kuchn, P. J., Approximate analysis of general queueing networks by decomposition, *IEEE Transactions on Communications*, **COM-27**, 1, 1979, pp. 113-126.

12. Zhozhikashvili, V. A. et al., Application of the queueing theory methods to the study of hierarchical computer networks with remote terminals, in "Trudy VII Vsesoyuznogo soveshaniya po problemam upravleniya", Moscow, Inst. of Control Sciences, Minsk, Inst. of Technical Cybernetics, 1977, pp. 404-406 (in Russian).
13. Bilik, R. V. et al., On applicability of equation systems based on the Polyachek-Khinchin formulas to the study of high-dimensionality closed networks, All-Union Conference "Computer systems, networks and shared centers", Novosibirsk, Published by VTs SO AN SSSR, 1978, pp. 152-155 (in Russian).
14. Bilik, R. V., Use of the approximate method for calculation of closed terminal networks under overload, Second workshop on computerized queueing systems. Frunze, Published by the Frunze Polytechnical Inst., 1980, pp. 15-16, (in Russian).
15. Vishnevskii, V. M., Gerasimov, A. I., An approximate method for the study of queueing networks with several classes of customers. Third workshop on computerized queueing systems, Vinnitsa, Published by the Vinnitsa House of Engineering, 1981, pp. 24-25 (in Russian).
16. Gerasimov, A. I., An approximate method for the study of hierarchical queueing networks with node blocking by messages. in "Teoriya i tekhnika avtomatizirovannykh sistem massovogo obsluzhivaniya", Moscow, Moskovskii Dom Nauchno-Tekhnicheskoi Propagandy, 1982, pp. 82-85 (in Russian).
17. Kleinrock, L., Queueing systems, vol. I: Theory, Wiley, New York, 1975.
18. Little, J. D. C., A proof of the queueing formula, Oper. Res., 9, pp. 383-387.
19. Buzen, J. P., Computational algorithms for closed queueing networks with exponential servers, Communications of the ACM, 16, 9, 1973, pp. 527-531.

Исследование стохастических сетей методом полиномиальной аппроксимации

А. И. ГЕРАСИМОВ

(Москва)

В работе предложен новый приближенный метод, основанный на полиномиальной аппроксимации для анализа открытых, замкнутых и смешанных стохастических сетей произвольной структуры с несколькими классами сообщений, несколькими подцепями, произвольными функциями распределения времени обслуживания в узлах, приоритетами и блокировками из-за ограниченных очередей в узлах.

А. И. ГЕРАСИМОВ

Институт проблем управления

АН СССР, 117342 Москва, Профсоюзная, 65

PRINTED IN HUNGARY
Akadémiai Nyomda, Budapest

ON SENSITIVITY IN THE DIFFERENTIAL GAME WITH THE INTEGRAL-TERMINAL PAYOFF

N. N. SUBBOTINA, A. I. SUBBOTIN

(Sverdlovsk)

In paper [1], using the method of the stochastic program synthesis [2, 3] N. N. Krasovskii and V. E. Tret'jakov have derived an expression for the value function of the feed-back differential game with linear dynamic, fixed end-time and an integral-terminal payoff. This function is nonsmooth. Sensitivity is the study of the differentiability properties of the value function. In the present paper the existence of the directional derivatives of the value function is proved. Two inequalities for the directional derivatives are obtained. It is shown that these inequalities together with the boundary condition constitute necessary and sufficient conditions which must be satisfied by the value function.

We consider also a nonlinear differential game with an integral-terminal payoff for which we formulate necessary and sufficient conditions for a nonsmooth function to be the value function.

1. Introduction

Consider a linear system

$$\dot{x}(t) = A(t)x + B(t)u(t) + C(t)v(t) \quad (1.1)$$

Here $x \in R^n$ is the phase vector; $A(t)$, $B(t)$, $C(t)$ are continuous $(n \times n)$, $(n \times p)$, $(n \times q)$ -matrices respectively; $u(t) \in R^p$, $v(t) \in R^q$ are controls of the first and the second players constrained by the conditions

$$u(t) \in P, \quad v(t) \in Q \quad (1.2)$$

where P and Q are convex compact sets.

A payoff is given by the functional

$$\begin{aligned} \gamma(x(\cdot), u(\cdot), v(\cdot)) = & \int_{t_*}^{\vartheta} (\langle \Phi(t) u(t), u(t) \rangle + \langle \Psi(t) v(t), v(t) \rangle) dt + \\ & + \|x(\vartheta)\| \end{aligned} \quad (1.3)$$

where ϑ is the fixed end-time, $x(\cdot)$ is the motion of the system (1.1), $x(t_*) = x_*$. The symbols $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the inner product and the Euclidean norm, respectively. It is assumed that $\Phi(t)$ and $\Psi(t)$ are continuous symmetric $(p \times p)$, $(q \times q)$ -matrices.

МАТЕМАТИЧЕСКИЙ ИНСТИТУТ
УДМУРТСКОГО АКАДЕМИИ НАУК
ИМ. С. П. КОПЫЛОВА

The first player tries to minimize the payoff and the second player to maximize it. We consider this differential game in the classes of positional strategies of the first and the second players [1, 6].

It is assumed that for any $t < \vartheta$, $u \in R^p$, $v \in R^q$ ($u \neq 0, v \neq 0$) and $l \in R^n$ the following relations hold

$$\langle \Phi(t)u, u \rangle > 0, \quad \langle \Psi(t)v, v \rangle < 0 \quad (1.4)$$

$$\begin{aligned} \min_{u \in P} [\langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle] = \\ = \min_{u \in R^p} [\langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle] \end{aligned} \quad (1.5)$$

$$\begin{aligned} \max_{v \in Q} [\langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle] = \\ = \max_{v \in R^q} [\langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle] \end{aligned} \quad (1.6)$$

where $X[\vartheta, t]$ is the solution of the equation $\frac{dX[\vartheta, t]}{dt} = -X[\vartheta, t]A(t)$, $X[\vartheta, \vartheta] = E$,

E is the identity matrix.

Under these assumptions the differential game (1.1)–(1.3) was considered in [1]. There, the following expression for the value function $\rho^0(t, x)$ was obtained

$$\rho^0(t, x) = \max_{\|l\| \leq 1} [\langle l, X[\vartheta, t]x \rangle + \langle \Gamma(t)l, l \rangle - \lambda_t \|l\|^2] + \lambda_t \quad (1.7)$$

where λ_t is defined by

$$\lambda_t = \max_{t \leq \tau \leq \vartheta} \lambda(\tau), \quad \lambda(\tau) = \max_{\|l\|=1} \langle \Gamma(\tau)l, l \rangle \quad (1.8)$$

$\Gamma(t)$ is a continuously differentiable symmetric matrix such that

$$\begin{aligned} - \left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle = \min_{u \in P} [\langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle] + \\ + \max_{v \in Q} [\langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle] \end{aligned} \quad (1.9)$$

$$\Gamma(\vartheta) = [0] \quad (1.10)$$

i.e., all components of the matrix $\Gamma(\vartheta)$ are equal to zero.

If the value function $\rho^0(t, x)$ is differentiable at point (t, x) , the following equality holds

$$\min_{u \in P} \max_{v \in Q} \left[\frac{\partial \rho^0(t, x)}{\partial t} + \left\langle \frac{\partial \rho^0(t, x)}{\partial x}, (A(t)x + B(t)u + C(t)v) \right\rangle + \langle \Phi(t)u, u \rangle + \langle \Phi(t)v, v \rangle \right] = 0 \quad (1.11)$$

This well-known partial differential equation is called the main equation (or Bellman–Isaaks equation).

In the general case the value function $\rho^0(\cdot)$ (1.7) is not differentiable at every point $(t, x) \in (-\infty, \vartheta) \times R^n$. This is because the necessary condition (1.11) is not sufficient. In the next section we shall give two inequalities for the directional derivatives of a nonsmooth function $\rho(t, x)$. These inequalities (2.14), (2.15) together with the boundary condition $\rho(\vartheta, x) = \|x\|$ form necessary and sufficient conditions which must be satisfied by the value function. In the region where the function $\rho(\cdot)$ is differentiable, these inequalities become the main equation.

2. Basic inequalities

In this section we first consider a nonlinear differential game and formulate characterizing properties of the value function of this game. As a corollary we then obtain necessary and sufficient conditions that the value function of the linear differential game (1.1)–(1.3) satisfies.

Consider a differential game described by

$$\dot{x} = f(t, x, u, v), \quad u \in P, \quad v \in Q, \quad (2.1)$$

$$\gamma(x(\cdot), u(\cdot), v(\cdot)) = \sigma(x(\cdot)) + \int_{t_*}^{\vartheta} f_{n+1}(t, x(t), u(t), v(t)) dt \quad (2.2)$$

where $x \in R^n$, $u \in R^p$, $v \in R^q$, P, Q are compact sets, ϑ is the fixed end-time, $x(\cdot) = x(\cdot, t_*, x_*, u(\cdot), v(\cdot))$ is the motion of the system (2.1) corresponding to the controls $u(\cdot): [t, \vartheta] \rightarrow P, v(\cdot): [t, \vartheta] \rightarrow Q$ and starting from the initial state (t_*, x_*) , $\gamma(\cdot)$ is the payoff functional that the first player tries to minimize and the second player to maximize.

It is assumed that the functions $f: (-\infty, \vartheta] \times R^n \times P \times Q \rightarrow R^n$, $f_{n+1}: (-\infty, \vartheta] \times R^n \times P \times Q \rightarrow R$, $\sigma: R^n \rightarrow R$ are continuous and locally Lipschitz continuous in x . Assume that for any $(t, x) \in (-\infty, \vartheta] \times R^n, (s, s_{n+1}) \in R^n \times R$ the following equality holds

$$\begin{aligned} & \min_{u \in P} \max_{v \in Q} [\langle s, f(t, x, u, v) \rangle + s_{n+1} f_{n+1}(t, x, u, v)] = \\ & = \max_{v \in Q} \min_{u \in P} [\langle s, f(t, x, u, v) \rangle + s_{n+1} f_{n+1}(t, x, u, v)]. \end{aligned} \quad (2.3)$$

Introduce a new phase variable x_{n+1} and a new phase vector $y=(x, x_{n+1})$. Then games (2.1), (2.2) is reduced to the following game

$$\dot{x} = f(t, x, u, v), \quad \dot{x}_{n+1} = f_{n+1}(t, x, u, v), \quad u \in P, \quad v \in Q \quad (2.4)$$

$$\gamma_*(y(\cdot)) = \gamma_*(x(\cdot), x_{n+1}(\cdot)) = \sigma(x(\vartheta)) + x_{n+1}(\vartheta). \quad (2.5)$$

It is known (see [6]) that for any initial state $(t_*, x_*) \in (-\infty, \vartheta] \times R^n$ the differential game (2.4) with the terminal payoff (2.5) has a value c^0 :

$$c^0(t_*, x_*) = \inf_U \sup_{y(\cdot) \in Y(t_*, x_*, U)} \gamma_*(y(\cdot)) = \sup_V \inf_{y(\cdot) \in Y(t_*, x_*, V)} \gamma_*(y(\cdot)). \quad (2.6)$$

Here $Y(t_*, x_*, U)$, $Y(t_*, x_*, V)$ are sets of all motions of the system (2.4) corresponding to positional strategies U and V , respectively [1, 6]. Note that the following relations for the value functions $(t, y) \rightarrow c^0(t, y)$, $(t, x) \rightarrow \rho^0(t, x)$ of the differential games (2.4), (2.5) and (2.1), (2.2) are valid

$$c^0(t, y) = c^0(t, x, x_{n+1}) = c^0(t, x, 0) + x_{n+1} \quad (2.7)$$

$$c^0(t, x, 0) = \rho^0(t, x). \quad (2.8)$$

Denote by the symbol $D\rho(t, x)|(1, f)$ the directional derivative of the function $(t, x) \rightarrow \rho(t, x)$ at point (t, x) for the direction $(1, f) \in R^{n+1}$.

We shall denote by the symbol Dif a class of locally Lipschitz continuous functions $(t, x) \rightarrow \rho(t, x)$ for which there exists the directional derivative at any point (t, x) and for any direction $(1, f) \in R^{n+1}$.

Using the results of paper [5] and relations (2.7), (2.8) we obtain

Theorem 2.1. A function $\rho \in \text{Dif}$ coincides with the value function ρ^0 of the differential game (2.1), (2.2) iff the following conditions are satisfied

$$\rho(\vartheta, x) = \sigma(x) \quad \text{for all } x \in R^n \quad (2.9)$$

$$\max_{v \in Q} \min_{(f, f_{n+1}) \in F_1(t, x, v)} [D\rho(t, x)|(1, f) + f_{n+1}] \leq 0 \leq$$

$$\leq \min_{u \in P} \max_{(f, f_{n+1}) \in F_2(t, x, u)} [D\rho(t, x)|(1, f) + f_{n+1}] \quad (2.10)$$

for all $(t, x) \in (-\infty, \vartheta) \times R^n$, here

$$F_1(t, x, v) = \text{co} \{ (f(t, x, u, v), f_{n+1}(t, x, u, v)) : u \in P \} \quad (2.11)$$

$$F_2(t, x, u) = \text{co} \{ (f(t, x, u, v), f_{n+1}(t, x, u, v)) : v \in Q \}. \quad (2.12)$$

As a corollary of Theorem 2.1 we have the following result.

Theorem 2.2. A function $\rho \in \text{Dif}$ coincides with the value function ρ^0 of the differential game (1.1)–(1.3) iff the following conditions are satisfied

$$\rho(\vartheta, x) = \|x\| \quad \text{for all } x \in R^n \quad (2.13)$$

$$\max_{v \in Q} \min_{(f, f_{n+1}) \in F_1(t)} [D\rho(t, x)|(A(t)x + f + C(t)v) + f_{n+1} + \langle \Psi(t)v, v \rangle] \leq 0 \quad (2.14)$$

$$\min_{u \in P} \max_{(f, f_{n+1}) \in F_2(t)} [D\rho(t, x)|(A(t)x + B(t)u + f) + \langle \Phi(t)u, u \rangle + f_{n+1}] \geq 0 \quad (2.15)$$

for all $(t, x) \in (-\infty, \vartheta) \times R^n$, here

$$F_1(t) = \text{co} \{(B(t)u, \langle \Phi(t)u, u \rangle) : u \in P\},$$

$$F_2(t) = \text{co} \{(C(t)v, \langle \Psi(t)v, v \rangle) : v \in Q\}. \quad (2.16)$$

3. The differentiability properties of the function ρ^0 (1.7)

Let us prove the existence of the directional derivatives and check conditions (2.13)–(2.15) for the value function ρ^0 (1.7).

From (1.8), (1.10) it follows that $\lambda_\vartheta = \lambda(\vartheta) = 0$. Substituting the relations $X[\vartheta, \vartheta] = E$, $\Gamma(\vartheta) = [0]$, $\lambda_\vartheta = 0$ in (1.7) we obtain

$$\rho^0(\vartheta, x) = \max_{\|l\| \leq 1} \langle l, x \rangle = \|x\|, \quad x \in R^n$$

Thus the boundary condition (2.13) is satisfied. Now we shall check conditions (2.14), (2.15).

The directional derivatives $D\rho^0(t, x)|(1, f)$. Consider the functions $t \rightarrow \lambda(t)$, $t \rightarrow \lambda_t$ (see (1.8)). We denote by $\dot{\lambda}(t)$ and $\dot{\lambda}_t$ the right-handed derivatives of these functions, i.e.

$$\dot{\lambda}(t) = \lim_{\delta \rightarrow +0} [\lambda(t+\delta) - \lambda(t)]\delta^{-1}, \quad \dot{\lambda}_t = \lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t]\delta^{-1}.$$

Proposition 3.1. For any $t < \vartheta$ there exist the right-handed derivatives $\dot{\lambda}(t)$ and $\dot{\lambda}_t$ for which the following relations are valid

$$\dot{\lambda}(t) = \max_{l \in L^*(t)} \left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle, \quad L^*(t) = \{l \in R^n : \|l\| = 1, \quad (3.1)$$

$$\langle \Gamma(t)l, l \rangle = \lambda(t)\}$$

$$\dot{\lambda}_t \leq 0. \quad (3.2)$$

Proof. The existence of the derivative $\dot{\lambda}(t)$ and equality (3.1) follow from (1.8) and [8] (Theorem 3.1). We now prove the existence of $\dot{\lambda}_t$.

The following three cases are possible: (i) $\lambda_t > \lambda(t)$; (ii) there exists $\delta > 0$ such that $\lambda_t = \lambda(t) = \lambda(t + \delta)$; (iii) $\lambda_t = \lambda(t) > \lambda(\tau)$ for all $\tau \in (t, \vartheta)$. It is clear that in the cases (i), (ii) $\dot{\lambda}_t = 0$. In case (iii) we have

$$\lambda_t = \max_{\tau \leq \zeta \leq \vartheta} \lambda(\zeta) < \lambda_t, \quad \tau \in (t, \vartheta]. \quad (3.3)$$

From (1.8) we can see that $\lambda(t) \leq \lambda_t$ for all $t < \vartheta$. Hence

$$\lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t] \delta^{-1} \geq \lim_{\delta \rightarrow +0} [\lambda(t + \delta) - \lambda(t)] \delta^{-1} = \dot{\lambda}(t). \quad (3.4)$$

Let a sequence δ_i ($i = 1, 2, \dots$) be chosen so that $\delta_i > 0$, $\delta_i \rightarrow 0$ as $i \rightarrow \infty$ and

$$\lim_{\delta_i \rightarrow 0} [\lambda_{t+\delta_i} - \lambda_t] \delta_i^{-1} = \overline{\lim}_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t] \delta^{-1}. \quad (3.5)$$

Furthermore, by (3.3) and (1.8) this sequence can be chosen so that

$$\begin{aligned} \lambda(t + \delta_i) \leq \lambda_{t+\delta_i} < \lambda_{t+\delta_{i+1}} = \max_{\delta \in [\delta_{i+1}, \delta]} \lambda(t + \delta), \\ (i = 1, 2, \dots). \end{aligned} \quad (3.6)$$

By (3.6) and the continuity of the function $t \rightarrow \lambda(t)$ there exists $\alpha_i \in (\delta_{i+1}, \delta_i]$ such that

$$\lambda(t + \alpha_i) = \lambda_{t+\delta_i}, \quad \alpha_i \leq \delta_i \quad (3.7)$$

Hence

$$\lim_{\delta_i \rightarrow 0} [\lambda_{t+\delta_i} - \lambda_t] \delta_i^{-1} \leq \lim_{\alpha_i \rightarrow 0} [\lambda(t + \alpha_i) - \lambda(t)] \alpha_i^{-1} = \dot{\lambda}(t). \quad (3.8)$$

From (3.5), (3.8), (1.8) and (3.4) it follows that the derivative exists and satisfies the inequality (3.2). Proposition 3.1 is proved.

Letting

$$\begin{aligned} \varphi(t, x, l) &= \langle (\Gamma(t) - \lambda_t E)l, l \rangle + \langle l, X[\vartheta, t]x \rangle + \lambda_t, \\ L &= \{l \in R^n: \|l\| \leq 1\} \end{aligned} \quad (3.9)$$

the function ρ^0 (1.7) can be rewritten as

$$\rho^0(t, x) = \max_{l \in L} \varphi(t, x, l). \quad (3.10)$$

Proposition 3.2. For any $(t, x) \in (-\infty, \vartheta) \times R^n$, $f \in R^n$ there exists the directional derivative $D\rho^0(t, x)|(1, f)$. If $f = A(t)x + h$ then

$$\begin{aligned} D\rho^0(t, x)|(1, A(t)x + h) &= \\ &= \max_{l \in L^0(t, x)} \left[\left\langle \left(\frac{d\Gamma(t)}{dt} - \lambda_t E \right) l, l \right\rangle + \langle l, h \rangle + \lambda_t \right], \end{aligned} \quad (3.11)$$

where

$$L^0(t, x) = \{l \in R^n : \|l\| \leq 1, \quad \varphi(t, x, l) = \rho^0(t, x)\}. \quad (3.12)$$

Proof. Consider

$$\varphi^*(\tau, x, l) = \langle (\Gamma(\tau) - \lambda_\tau^* E)l, l \rangle + \langle l, X[\vartheta, \tau]x \rangle + \lambda_\tau^* \quad (3.13)$$

$$\lambda_\tau^* = \lambda_t + \lambda_t(\tau - t) \quad (3.14)$$

$$\rho^*(\tau, x) = \max_{l \in L} \varphi^*(\tau, x, l). \quad (3.15)$$

It is clear that $\lambda_t^* = \lambda_t$, $\varphi^*(t, x, l) = \varphi(t, x, l)$, $\rho^*(t, x) = \rho^0(t, x)$

$$\begin{aligned} \lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_{t+\delta}^*] \cdot \delta^{-1} &= \lim_{\delta \rightarrow +0} [\lambda_{t+\delta} - \lambda_t - \lambda_t \delta] \delta^{-1} = 0 \\ \lim_{\delta \rightarrow +0} [\varphi(t+\delta, x+\delta f, l) - \varphi^*(t+\delta, x+\delta f, l)] \delta^{-1} &= 0, \quad \forall l \in L. \end{aligned} \quad (3.16)$$

Hence

$$\lim_{\delta \rightarrow +0} [\rho^0(t+\delta, x+\delta f) - \rho^*(t+\delta, x+\delta f)] \delta^{-1} = 0. \quad (3.17)$$

Functions $\varphi^*(\cdot)$, λ_τ^* , $\rho^*(\cdot)$ satisfy the conditions of the theorem 3.1 [8]. According to this theorem we obtain the existence of the derivative $D\rho^*(t, x)|(1, f)$ and the following equality

$$\begin{aligned} D\rho^*(t, x)|(1, f) &= \max_{l \in L^0(t, x)} D\varphi^*(t, x, l)|(1, f) = \\ &= \max_{l \in L^0(t, x)} \left[\frac{\partial \varphi^*(t, x, l)}{\partial t} + \left\langle \frac{\partial \varphi^*(t, x, l)}{\partial x}, f \right\rangle \right] = \\ &= \max_{l \in L^0(t, x)} \left[-\langle l, X[\vartheta, t]A(t)x \rangle + \right. \\ &\quad \left. + \left\langle \left(\frac{d\Gamma(t)}{dt} - \lambda_t E \right) l, l \right\rangle + \lambda_t + \langle l, X[\vartheta, t]f \rangle \right] \end{aligned} \quad (3.18)$$

where $L^0(t, x)$ is defined by (3.12). From (3.17), (3.18) we have the equality

$$\begin{aligned} D\rho^0(t, x)|(1, f) &= \lim_{\delta \rightarrow +0} [\rho^0(t + \delta, x + \delta f) - \rho^0(t, x)]\delta^{-1} = \\ &= \lim_{\delta \rightarrow +0} [\rho^*(t + \delta, x + \delta f) - \rho^*(t, x)]\delta^{-1} = D\rho^*(t, x)|(1, f). \end{aligned} \quad (3.19)$$

From (3.18), (3.19) follows (3.11). Proposition 3.2 is proved. Note that the function ρ^0 (1.7) is Lipschitz continuous. Hence we have

Corollary 3.1. The function ρ^0 (1.7) belongs to the class *Dif*.

Properties of the set $L^0(t, x)$. In this section we make use of the auxiliary results presented below in Appendix.

Proposition 3.3. For any $(t, x) \in (-\infty, \vartheta) \times R^n$ the set $L^0(t, x)$ (3.12) is convex and compact in R^n .

Proof. Denote by $\lambda_i(t)$ and $\mu_i(t)$ ($i = \overline{1, n}$) the characteristic roots of the matrices $\Gamma(t)$ and $\Gamma(t) - \lambda_t E$ respectively. One can easily check that

$$\mu_i(t) = \lambda_i(t) - \lambda_t \quad i = \overline{1, n}. \quad (3.20)$$

The matrix $\Gamma(t)$ is real-valued, symmetric, hence the roots $\lambda_i(t), \mu_i(t)$ are real. From (3.20) and (1.8) we have

$$\mu_i(t) \leq \max_{1 \leq i \leq n} \mu_i(t) = \max_{1 \leq i \leq n} (\lambda_i(t) - \lambda_t) = \lambda(t) - \lambda_t \leq 0. \quad (3.21)$$

Hence, for any $l \in R^n$, $\langle (\Gamma(t) - \lambda_t E)l, l \rangle \leq 0$ and, for any $(t, x) \in (-\infty, \vartheta) \times R^n$, the function $l \rightarrow \varphi(t, x, l)$ (see (3.10)) is concave. Consequently, the set $L^0(t, x)$ (3.12) is convex and compact.

Proposition 3.4. If $\lambda_t = \lambda(t)$ then the set $L^0(t, x)$ contains a unit-norm vector l^0 .

Proof. Let $\mu = \mu_i(t)$ $i = \overline{1, n}$ be the characteristic roots of the matrix $H = H(t) = \Gamma(t) - \lambda_t E$. From (3.11) it follows that

$$\mu_{m+1}(t) = \dots = \mu_n(t) = \max_{1 \leq i \leq n} \mu_i(t) = \lambda(t) - \lambda_t = 0,$$

$$m = \overline{0, n-1} \quad (3.22)$$

$$\mu_i < 0, \quad i = \overline{1, m} \quad (3.23)$$

where $(n-m)$ is the multiplicity of the root $\mu = 0$.

We denote by $S(t) = S$ a matrix S defined by conditions (4.4), (4.5) where we put $H = \Gamma(t) - \lambda_t E$. Let the vector l be associated with the vector ξ by the equality $l = S\xi$. From (4.4), (4.5), (3.22), (2.23) it follows that

$$\langle l, l \rangle = \langle S\xi, S\xi \rangle = \langle S^T S\xi, \xi \rangle = \langle \xi, \xi \rangle \quad (3.24)$$

$$\begin{aligned} \varphi(t, x, l) &= \varphi(t, x, S(t)\xi) = \langle H(t)S(t)\xi, S(t)\xi \rangle + \langle S(t)\xi, X[\vartheta, t]x \rangle + \\ &+ \lambda_t = \langle (S^T(t)H(t)S(t))\xi, \xi \rangle + \langle \xi, S^T(t)X[\vartheta, t]x \rangle + \lambda_t = \\ &= \sum_{i=1}^m [\mu_i(t)\xi_i^2 + \xi_i y_i] + \sum_{i=m+1}^n \xi_i y_i + \lambda_t \end{aligned} \quad (3.25)$$

where we denote by the symbol T the operation of transpose and by $y = (y_1, \dots, y_n)^T = S^T(t)X[\vartheta, t]x$ the column vector.

Let $r = \sum_{i=m+1}^n (y_i)^2$. If $r = 0$ then the existence of a unit-norm vector $l^0 \in L^0(t, x)$ is obvious. Consider the case $r > 0$. Suppose the contrary:

$$\forall l \in L^0(t, x): \|l\| < 1. \quad (3.26)$$

Take arbitrary $l^0 \in L^0(t, x)$. Let $\xi^0 = S^{-1}(t)l^0$. From (3.24), (3.26) it follows that

$$\|\xi^0\| < 1, \quad \sum_{i=1}^m (\xi_i^0)^2 = c < 1, \quad \sum_{i=m+1}^n (\xi_i^0)^2 < 1 - c \quad (3.27)$$

$$\rho^0(t, x) = \rho(t, x, S(t)\xi^0) = \max_{\|\xi\| \leq 1} \varphi(t, x, S(t)\xi). \quad (3.28)$$

Let $\xi_i^* = y_i(1-c)^{1/2}r^{1/2}$ for $i = \overline{m+1, n}$. By (3.27) we have

$$\sum_{i=m+1}^n \xi_i^0 y_i < (1-c)^{1/2}r^{1/2} = \sum_{i=m+1}^n \xi_i^* y_i, \quad \sum_{i=m+1}^n (\xi_i^*)^2 = (1-c). \quad (3.29)$$

Introduce the vector $\xi^* = (\xi_1^*, \dots, \xi_n^*) = (\xi_1^0, \dots, \xi_m^0, \xi_{m+1}^*, \dots, \xi_n^*)$. By (3.27), (3.28) we obtain the following relation

$$\|\xi^*\| = \left[\sum_{i=1}^m (\xi_i^0)^2 + \sum_{i=m+1}^n (\xi_i^*)^2 \right]^{1/2} = (c + (1-c))^{1/2} = 1 \quad (3.30)$$

$$\varphi(t, x, S(t)\xi^*) > \varphi(t, x, S(t)\xi^0) = \max_{\|\xi\| \leq 1} \varphi(t, x, S(t)\xi) \quad (3.31)$$

which contradict (3.28). Proposition 3.4 is proved.

Basic inequalities. We shall prove that the function ρ^0 (1.7) satisfies inequalities (2.14), (2.15). Note that from (3.19), (3.18) it follows that $D\rho^0(t, x)|(1, f) = \max_{l \in L^0(t, x)} D\varphi^*(t, x, l)|(1, f)$ where $\varphi^*(\cdot)$ is defined by (3.13), (3.14). Using (3.18), (2.16), (1.9) and the inequality $\min \max \geq \max \min$ we obtain

$$\begin{aligned} & \min_{u \in P} \max_{(f, f_{n+1}) \in F_2(t)} [D\rho^0(t, x)|(1, A(t)x + B(t)u + f) + \langle \Phi(t)u, u \rangle + f_{n+1}] \geq \\ & \geq \min_{u \in P} \max_{v \in Q} \left[\max_{l \in L^0(t, x)} D\varphi^*(t, x, l)|(1, A(t)x + B(t)u + C(t)v) + \right. \\ & \left. + \langle \varphi(t)u, u \rangle + \langle \Psi(t)v, v \rangle \right] \geq \max_{l \in L^0(t, x)} \left\{ \max_{v \in Q} \min_{u \in P} \left(\left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle + \right. \right. \\ & \left. \left. + \langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle + \langle l, X[\vartheta, t]C(t)v \rangle + \langle \Psi(t)v, v \rangle \right) - \right. \\ & \left. - \lambda_t(\|l\|^2 - 1) \right\} = \max_{l \in L^0(t, x)} (-\lambda_t)(\|l\|^2 - 1). \end{aligned} \quad (3.32)$$

Recall that $\varphi^*(t, x, l) = \varphi(t, x, l)$ and for any (t, x) the function $l \rightarrow \varphi(t, x, l)$ is concave. Hence according to proposition 4.1 the function $l \rightarrow D\varphi^*(t, x, l)|(1, f)$ is concave on the convex compact set $L^0(t, x)$. By (3.18) and (1.4) we have that the function $u \rightarrow D\varphi^*(t, x, l)|(1, A(t)x + B(t)u + f) + \langle \Phi(t)u, u \rangle$ is convex on the convex compact set P . Using the minimax theorem [10] we obtain

$$\begin{aligned} & \max_{v \in Q} \min_{(f, f_{n+1}) \in F_1(t)} [D\rho^0(t, x)|(1, A(t)x + f + C(t)v) + f_{n+1} + \langle \Psi(t)v, v \rangle] \leq \\ & \leq \max_{v \in Q} \min_{u \in P} \left[\max_{l \in L^0(t, x)} D\varphi^*(t, x, l)|(1, A(t)x + B(t)u + C(t)v) + \right. \\ & \left. + \langle \Phi(t)u, u \rangle + \langle \Psi(t)v, v \rangle \right] = \max_{l \in L^0(t, x)} \left\{ \max_{v \in Q} \min_{u \in P} \left(\left\langle \frac{d\Gamma(t)}{dt} l, l \right\rangle + \right. \right. \\ & \left. \left. + \langle l, X[\vartheta, t]B(t)u \rangle + \langle \Phi(t)u, u \rangle + \langle l, X[\vartheta, t]C(t)v \rangle + \right. \right. \\ & \left. \left. + \langle \Psi(t)v, v \rangle \right) - \lambda_t(\|l\|^2 - 1) \right\} = \max_{l \in L^0(t, x)} (-\lambda_t)(\|l\|^2 - 1). \end{aligned} \quad (3.33)$$

If $\lambda(t) = \lambda_t$ then from (3.2) and proposition 3.4 we have

$$\max_{l \in L^0(t, x)} (-\dot{\lambda}_t) (\|l\|^2 - 1) = 0. \quad (3.34)$$

If $\lambda(t) < \lambda_t$ then $\dot{\lambda}_t = 0$ (see the proof of proposition 3.1, case (i)). Hence we again obtain (3.34). From (3.32), (3.33), (3.34) we conclude that the function ρ^0 (1.7) satisfies the inequalities (2.14), (2.15).

4. Appendix

This Appendix contains some results of the theory of the quadratic forms (see i.e. [9]) and of the convex analysis.

Let H be a real, symmetric, $(n \times n)$ -matrix, E be the identity $(n \times n)$ -matrix, μ be a characteristic root of the matrix H , i.e.

$$\det(H - \mu E) = 0. \quad (4.1)$$

A real symmetric $(n \times n)$ -matrix has n real characteristic roots: $\mu_1 \leq \dots \leq \mu_n$. The following equality holds

$$\max_{1 \leq i \leq n} \mu_i = \max_{\|s\|=1} \langle Hs, s \rangle. \quad (4.2)$$

Denote by $s_i \in R^n$ characteristic vector associated with characteristic root μ_i of the matrix H , i.e.

$$Hs_i = \mu_i s_i. \quad (4.3)$$

There is an orthonormal system of the characteristic vectors s_1, \dots, s_n associated with characteristic roots μ_1, \dots, μ_n of the matrix H . Let S be the $(n \times n)$ -matrix whose columns are the characteristic vectors s_1, \dots, s_n of the matrix H . Then we have

$$\det S \neq 0, \quad S^T S = E = S S^T \quad (4.4)$$

$$S^T H S = \text{diag} \{ \mu_1, \dots, \mu_n \} \quad (4.5)$$

where symbol $\text{diag} \{ \mu_1, \dots, \mu_n \}$ denotes the diagonal $(n \times n)$ -matrix.

Let $s \rightarrow \langle Hs, s \rangle$ be a quadratic form associated with the matrix H . If $\langle Hs, s \rangle \leq 0$ for all $s \in R^n$ they say that the quadratic form is non-positive definite; if $\langle Hs, s \rangle < 0$ for all $s \in R^n, s \neq 0$, they say that the quadratic form is negative definite.

The quadratic form $\langle Hs, s \rangle$ is non-positive (negative) definite iff the characteristic roots μ_i H satisfy the following inequalities

$$\mu_i \leq 0, \quad i = \overline{1, n} \quad (\mu_i < 0, \quad i = \overline{1, n}). \quad (4.6)$$

The function $s \rightarrow \langle Hs, s \rangle$ is concave (strongly concave) iff the quadratic form $\langle Hs, s \rangle$ is non-positive (negative) definite. We now consider a function

$$(t, x) \rightarrow \rho(t, x) = \max_{l \in L} \varphi(t, x, l). \quad (4.7)$$

Suppose that L is compact in R^n , the functions φ , $\partial\varphi/\partial t$ and $\partial\varphi/\partial x$ are continuous on $(-\infty, \vartheta) \times R^n + L$.

Proposition 4.1. If L is convex and for any $(t, x) \in (-\infty, \vartheta] \times R^n$ the function $l \rightarrow \varphi(t, x, l)$ is concave then for any $(t, x) \in (-\infty, \vartheta) \times R^n$ and $f \in R^n$ the function $l \rightarrow D\varphi(t, x, l)|(1, f)$ is concave on the set $L^0(t, x) = \{l \in L \subset R^n : \varphi(t, x, l) = \rho(t, x)\}$.

Proof. It is clear that $L^0(t, x)$ is convex. Let

$$g(t, x, f, l, \delta) = [\varphi(t + \delta, x + \delta f, l) - \varphi(t, x, l)]\delta^{-1}. \quad (4.8)$$

We have

$$\begin{aligned} \lim_{\delta \rightarrow +0} g(t, x, f, l, \delta) &= D\varphi(t, x, l)|(1, f) = \frac{\partial\varphi(t, x, l)}{\partial t} + \\ &+ \left\langle \frac{\partial\varphi(t, x, l)}{\partial x}, f \right\rangle. \end{aligned} \quad (4.9)$$

For any $(t, x, f) \in (-\infty, \vartheta) \times R^n \times R^n$, $\delta > 0$ the function $l \rightarrow \varphi(t + \delta, x + \delta f, l)$ is concave on $L^0(t, x)$. By the definition of the set $L^0(t, x)$ we have $\varphi(t, x, l) = \rho(t, x)$ for any $l \in L^0(t, x)$. Hence, the function $l \rightarrow g(t, x, f, l, \delta)$ is concave on $L^0(t, x)$, i.e.

$$g(t, x, f, l_\lambda, \delta) \geq \lambda \cdot g(t, x, f, l_1, \delta) + (1 - \lambda)g(t, x, f, l_2, \delta) \quad (4.10)$$

where $l_i \in L^0(t, x)$ ($i = 1, 2$), $\lambda \in [0, 1]$, $l_\lambda = \lambda l_1 + (1 - \lambda)l_2$. By taking limits in (4.10) as $\delta \rightarrow +0$ and making use of (4.9) we obtain

$$\begin{aligned} D\varphi(t, x, l_\lambda)|(1, f) &\geq \\ &\geq \lambda \cdot D\varphi(t, x, l_1)|(1, f) + (1 - \lambda)D\varphi(t, x, l_2)|(1, f). \end{aligned} \quad (4.11)$$

Therefore the function $l \rightarrow D\varphi(t, x, l)|(1, f)$ is concave on $L^0(t, x)$.

References

1. Krasovskii, N. N., Tret'jakov, V. E., A stochastic program synthesis of some guaranteeing control. *Problems of Control and Information Theory* (to appear).
2. Krasovskii, N. N., Tret'jakov, V. E., A stochastic program synthesis for a positional differential game, *Dokl. Acad. Nauk SSSR*, 1981, vol. **259**, No. 1, pp. 24-27 (in Russian).
3. Krasovskii, A. N., Krasovskii, N. N., Tret'jakov, V. E., A stochastic program synthesis for a determinate positional differential game, *Prikl. Math. Meh.*, 1981, vol. **45**, No. 4, pp. 581-588 (in Russian).
4. Subbotin, A. I., Subbotina, N. N., Necessary and sufficient conditions for a piecewise smooth value function of a differential game, *Dokl. Acad. Nauk SSSR*, 1978, vol. **243**, No. 4, pp. 862-865 (in Russian).
5. Subbotin, A. I., A generalization of the basic equation of the theory of differential games, *Dokl. Acad. Nauk SSSR*, 1980, vol. **254**, No. 2, pp. 293-297 (in Russian).
6. Krasovskii, N. N., Subbotin, A. I., *Positional Differential Games*, Nauka, Moscow, 1974 (in Russian).
7. Isaacs, R., *Differential Games*, J. Wiley and Sons, New York, 1965.
8. Dem'janov, V. F., *Minimax: directional differentiation*, Izdat. Leningrad Univ., Leningrad, 1974.
9. Lancaster, P., *Theory of matrices*, Academic Press, New York-London, 1969.
10. Karlin, S., *Mathematical methods and theory in games, programming and economics*, Stanford Univ., Pergamon Press, London-Paris, 1959.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H - 1111 Budapest XI., Stoczek u. 2, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper with wide margins (4-5 cm), should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary - possibly in Russian if the paper is in English and *vice-versa* - should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10-15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50-100 слов и приложено резюме - реферат объемом не менее 10-15% объема статьи на русском и на английском языке (в трех экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициях. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Subbotina N. N., Subbotin A. I.</i> Свойства дифференцируемости функции цены дифференциальной игры с интегрально-терминальной платой (<i>Subbotina, N. N., Subbotin, A. I.</i> On sensitivity of the value function of the differential game with the integral-terminal payoff)	153
<i>Chaudhuri, A. K., Mukherjee, R. N.</i> On minimum time control (<i>Чайдхури А. К., Мукхери Р. Н.</i> О минимальном времени управления)	167
<i>Pedrycz, W.</i> Towards set-theoretic representation of nondeterministic systems (<i>Педрич В.</i> К теоретико-множественному представлению недетерминистических систем)	179
<i>Rykov, A. S.</i> Simplex algorithms for unconstrained minimization (<i>Рыков А. С.</i> Симплексные алгоритмы безусловной минимизации)	195
<i>Azimov, A. Ya.</i> Duality in nonconvex problems of vector optimization (<i>Азимов А. Я.</i> Двойственность в невыпуклых задачах векторной оптимизации)	209
<i>Gerasimov, A. I.</i> Analysis of queueing networks by polynomial approximation (<i>Герасимов А. И.</i> Исследование стохастических сетей методом полиномиальной аппроксимации)	219

316.520

VOL. 12 • NUMBER 4
TOM HOMEP

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1983

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H-1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 0BW, England
or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1983 Subscription Rate US Dollars 140.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

N. N. KRASOVSKII (USSR)

G. BOGNÁR (HUNGARY)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

E. D. TERYAEV

HUNGARY

T. VÁMOS

L. VARGA

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJČ

РЕДАКТОРЫ ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

Г. БОГНАР (ВНР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

Е. Д. ТЕРЯЕВ

ВНР

Т. ВАМОШ

Л. ВАРГА

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

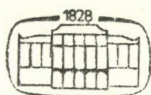
Л. КЕВИЦКИ

Я. КОЧИШ

ЧССР

И. БЕНЕШ

В. СТРЕЙЦ



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADEMIA
KÖNYVTÁRA

ОБЗОР ТЕОРИИ ДИЗЬЮНКТИВНЫХ КОДОВ

А. Г. ДЬЯЧКОВ, В. В. РЫКОВ

(Москва)

(Поступила в редакцию 14 июня 1982 г.)

Рассматривается обобщение понятия дизьюнктивного кода, введенного в 1964 году Каутсом и Синглтоном [1]. Описываются новые результаты в теории дизьюнктивных кодов. Сформулированы некоторые открытые проблемы.

1. Обозначения, определения дизьюнктивных кодов и их свойства

Пусть $1 \leq s < t$, $1 \leq L \leq t - s$, $N \geq 1$ — целые числа. Обозначим через $\mathbf{u}(j) = (u_1(j), \dots, u_N(j))$, $j = \overline{1, s}$ — двоичные (из 0 и 1) столбцы длины N . Булевой суммой

$$\mathbf{u} = \mathbf{u}(1) \vee \mathbf{u}(2) \vee \dots \vee \mathbf{u}(s)$$

столбцов $\mathbf{u}(1), \mathbf{u}(2), \dots, \mathbf{u}(s)$ называется двоичный столбец $\mathbf{u} = (u_1, u_2, \dots, u_N)$ длины N , компоненты которого определяются соотношениями

$$u_i = \begin{cases} 0, & \text{если } u_i(1) = u_i(2) = \dots = u_i(s) = 0, \\ 1, & \text{в других случаях, } i = \overline{1, N}, \end{cases}$$

Будем говорить, что столбец \mathbf{u} покрывает столбец \mathbf{v} , если $\mathbf{u} \vee \mathbf{v} = \mathbf{u}$. В противном случае будем говорить, что \mathbf{u} не покрывает \mathbf{v} .

Пусть $\mathbf{x} = \|\mathbf{x}_i(j)\|$, $i = \overline{1, N}$, $j = \overline{1, t}$ — матрица из 0 и 1 размером $N \times t$, а символ $\mathbf{x}(j)$, $j = \overline{1, t}$, обозначает j -ый столбец \mathbf{x} . Матрица \mathbf{x} в дальнейшем интерпретируется как совокупность из t двочных N -мерных столбцов $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)$.

Определение 1. Матрица \mathbf{x} называется *дизьюнктивным* (s, t, L) -кодом длины N , если булева сумма любого s -подмножества ее столбцов может покрывать не более $L - 1$ столбцов матрицы \mathbf{x} , которые не были слагаемыми данной булевой суммы.

Отметим, что в наиболее важном частном случае $L = 1$ определение 1 эквивалентно следующему условию. Булева сумма любого s -подмножества

столбцов x покрывает те и только те s столбцов, которые есть члены данной булевой суммы.

Дизъюнктивные $(s, t, 1)$ -коды были введены Каутсом и Синглетоном в работе [1], где подробно описан ряд прикладных задач, которые привели к определению $(s, t, 1)$ -кода, и даны некоторые конструкции таких кодов. Новые приложения (s, t, L) -кодов, связанные с применением кодирования в каналах с множественным доступом, изложены в недавней работе [2].

Замечание 1. Пусть $[N]$ — множество целых чисел от 1 до N . Каждый столбец матрицы x отождествим с подмножеством $[N]$, которое состоит из номеров единичных позиций этого столбца. Тогда на языке теории множеств построение (s, t, L) -кода длины N эквивалентно комбинаторной задаче построения семейства из t подмножеств множества $[N]$, для которого выполнено следующее требование: объединение (в теоретико-множественном смысле) любых s членов семейства может включать в себя (также в теоретико-множественном смысле) не более $L-1$ членов семейства, не являющихся членами данного объединения.

Пусть $N(s, t, L)$ обозначает минимально возможную длину (s, t, L) -кода. В этой работе будут употребляться также следующие стандартные обозначения:

$\lceil b \rceil$ — наименьшее целое, большее или равное b ,

$\lfloor b \rfloor$ — наибольшее целое, меньшее или равное b ,

$\log b$ — логарифм по основанию 2 числа b ,

\triangleq — равенство по определению,

$e = 2,718$ — основание натурального логарифма,

$h(u) = -u \log u - (1-u) \log (1-u)$ — двоичная энтропия.

С помощью определения 1 нетрудно проверить, что справедливы следующие три свойства.

Свойство 1. Всякий $(s+1, t, L)$ -код есть (s, t, L) -код, а всякий (s, t, L) -код есть $(s, t, L+1)$ -код. Следовательно,

$$N(s+1, t, L) \geq N(s, t, L) \geq N(s, t, L+1).$$

Свойство 2. Для любого s -подмножества столбцов (s, t, L) -кода x существует не более C_{s+L-1}^s , s -подмножеств столбцов x , таких, что булевы суммы столбцов этих s -подмножеств совпадают с булевой суммой столбцов данного s -подмножества. Отсюда вытекает

$$N(s, t, L) \geq \lceil \log C_+^s - \log C_{s+L-1}^s \rceil. \quad (1)$$

Свойство 3. Зафиксируем произвольным образом $\lfloor t/L \rfloor$ попарно непересекающихся L -подмножества из столбцов (s, t, L) -кода x . Пусть x' -матрица из 0 и 1, состоящая из N строк и $\lfloor t/L \rfloor$ столбцов, столбцы которой есть булевы

суммы столбцов данных L -подмножеств. Тогда x' является $(\lfloor s/L \rfloor, \lfloor t/L \rfloor, 1)$ -кодом, а потому

$$N(s, t, L) \geq N(\lfloor s/L \rfloor, \lfloor t/L \rfloor, 1). \tag{2}$$

Неравенство (2) позволяет получать нижние границы $N(s, t, L)$, используя соответствующие нижние границы для частного случая $L = 1$.

Определение 2. Матрица x называется *дизьюнктивным (\tilde{s}, \tilde{t}) -кодом* длины N , если булева сумма, составленная из произвольного фиксированного s -подмножества столбцов x , отлична от булевой суммы, составленной из любого другого s -подмножества столбцов x .

Через $\tilde{N}(s, t)$ обозначим минимально возможное число строк (\tilde{s}, \tilde{t}) -кода. Определение (\tilde{s}, \tilde{t}) -кода возникло в задачах планирования отсеивающих экспериментов [3]. Сформулируем еще три простых, но важных свойства введенных понятий, которые непосредственно вытекают из определений 1 и 2.

Свойство 4. Число

$$\tilde{N}(s, t) \geq \lceil \log C_t^s \rceil. \tag{3}$$

В асимптотической ситуации

$$t \rightarrow \infty, \quad s - \text{const}, \quad L - \text{const} \tag{4}$$

справедливы неравенства

$$\begin{aligned} N(s, t, L) &\geq s \log t(1 + o(1)), \\ \tilde{N}(s, t) &\geq s \log t(1 + o(1)), \end{aligned} \tag{5}$$

которые вытекают из (1) и (3), соответственно. Заметим, что матрица x размеров $N \times t$, где $t = C_{\lfloor N/2 \rfloor}^{\lfloor N/2 \rfloor}$, все t столбцов которой отличны друг от друга и содержат одинаковое число $\lfloor N/2 \rfloor$ единиц, является $(1, t, L)$ -кодом и $(\tilde{1}, t)$ -кодом одновременно. Следовательно, при $s = 1$ в (5) имеет место знак равенства.

Свойство 5. Всякий $(s, t, 1)$ -код есть (\tilde{s}, \tilde{t}) -код, а всякий (\tilde{s}, \tilde{t}) -код есть $(s - 1, t, 2)$ -код, т.е.

$$N(s, t, 1) \geq \tilde{N}(s, t) \geq N(s - 1, t, 2). \tag{6}$$

Свойство 6. Матрица x одновременно удовлетворяет определениям $(s - 1, t, 1)$ -кода и (\tilde{s}, \tilde{t}) -кода тогда и только тогда, когда каждая булева сумма, составленная из не более s столбцов x , отлична от всякой другой булевой суммы, также составленной из не более s столбцов.

Неравенства (2) и (6) позволят в дальнейшем получить нижнюю границу $\tilde{N}(s, t)$ с помощью нижней границы для $N(\lfloor s - 1/2 \rfloor, \lfloor t/2 \rfloor, 1)$. Свойство 6 имеет важное значение в теории планирования отсеивающих экспериментов. Оно дает

возможность оценить оптимальную длину кода, удовлетворяющего условию свойства 6, в терминах оптимальных длин $(s-1, t, 1)$ -кода и (\tilde{s}, t) -кода.

Далее в этой работе будут рассмотрены следующие вопросы. В разделе 2 сформулированы результаты статьи [4], которые уточняют нижние оценки (1), (3) и (5), в разделе 3 даны верхние границы оптимальных длин дизъюнктивных кодов, полученные методом случайного кодирования. В разделе 4 исследуется специальный класс дизъюнктивных кодов, введенный Каутсом и Синглтоном [1]. Эти коды являются важным частным случаем $(s, t, 1)$ -кодов, поскольку большинство известных регулярных конструкций $(s, t, 1)$ -кодов найдено в этом классе [1]. В разделе 4 получены границы оптимальных длин таких кодов.

2. Нижние границы длины дизъюнктивных кодов

Сначала рассмотрим нижние границы $N(s, t, 1)$, а затем с помощью (2) и (6) обобщим эти границы на случай (s, t, L) -кодов и (\tilde{s}, t) -кодов. Ограничимся здесь лишь формулировками теорем, (доказательства в [4]). Теорема 1 доказана в 1975 году Л. А. Бассальго.

Теорема 1. *Величина*

$$N(s, t, 1) \geq \min \left\{ \frac{(s+1)(s+2)}{2}; t-s \right\}$$

и, следовательно, для любого фиксированного α , $1/2 < \alpha < 1$, существует

$$\lim_{t \rightarrow \infty} \frac{N(\lfloor t^\alpha \rfloor, t, 1)}{t} = 1.$$

Другими словами, при $1/2 < \alpha < 1$ всякий $(\lfloor t^\alpha \rfloor, t, 1)$ -код асимптотически ($t \rightarrow \infty$) не лучше тривиального кода длины $N=t$, представляющего собой диагональную матрицу.

Теорема 2. *Пусть $s \geq 2$, а число $d = d(s-1, t)$ таково, что $N(s-1, t-1, 1) \geq d$. Тогда $N(s, t, 1) \geq d(s, t)$, где $d(s, t)$ — наименьшее из целых чисел N , удовлетворяющих неравенству*

$$t \leq N + s^2 \sum_{k=1}^{\lfloor (N-d)/s \rfloor} (C_N^{k+1} / C_{ks}^k).$$

Для каждого $s \geq 1$ введем функцию аргумента v , $0 < v < 1$,

$$f_s(v) \triangleq h(v/s) - vh(1/s),$$

где $h(u)$ — двоичная энтропия. Из теоремы 2 вытекает асимптотическая нижняя граница для $N(s, t, 1)$.

Теорема 3. В условиях (4) справедливо неравенство

$$N(s, t, 1) \geq K(s) \log t(1 + o(1)),$$

где последовательность $K(1) = 1, K(2), K(3), \dots$ задается рекуррентным образом. Число $K(s)$ при $s \geq 2$ есть единственное решение уравнения

$$K(s) = \left[\max_{(*)} f_s(v) \right]^{-1},$$

где максимум берется по всем v , удовлетворяющим условию

$$0 \leq v \leq \frac{K(s) - K(s-1)}{K(s)}. \quad (*)$$

Свойства последовательности $K(s), s \geq 2$, описывает

Теорема 4. 1) Число

$$K(2) = \left[\max_{0 \leq v \leq 1} f_2(v) \right]^{-1},$$

а при $s \geq 3$ справедливо рекуррентное равенство

$$K(s) = \left[f_s \left(\frac{K(s) - K(s-1)}{K(s)} \right) \right]^{-1}.$$

2) Для любого $s \geq 2$ имеет место неравенство

$$K(s) \geq s^2/2 \log [e(s+1)/2].$$

3) Если $s \rightarrow \infty$, то

$$K(s) = s^2 (2 \log s)^{-1} (1 + o(1)).$$

Для сравнения с асимптотической границей (5) приведем численные значения $K(s)$ при $s = \overline{2, 17}$:

s	$K(s)$	s	$K(s)$	s	$K(s)$	s	$K(s)$
2	3,10628	6	12,0482	10	24,5837	14	40,3950
3	5,01802	7	14,8578	11	28,2402	15	44,8306
4	7,11964	8	17,8876	12	32,0966	16	49,4536
5	9,46603	9	21,1313	13	36,1493	17	54,2612

Неравенства (2) и (6) позволяют обобщить результат теоремы 3 на (s, t, L) -коды и (\tilde{s}, \tilde{t}) -коды.

Следствие 1. В условиях (4)

$$N(s, t, L) \geq c(s, L) \log t(1 + o(1)),$$

$$\tilde{N}(s, t) \geq \tilde{c}(s) \log t(1 + o(1)),$$

где

$$c(s, L) = \max \{s; K(\lfloor s/L \rfloor)\},$$

$$\tilde{c}(s) = \max \{s; K(\lfloor (s-1)/2 \rfloor)\}.$$

Из теоремы 4 и числовых значений коэффициентов $K(s)$ можно сделать следующие выводы. 1) Если $L=1$, то для любого $s \geq 2$ число

$$c(s, 1) = K(s) > s.$$

2) Для любого $L \geq 2$ существует число $s(L)$, такое, что при всех $s \geq s(L)$ коэффициент $c(s, L) > s$. В частности, при $L=2$ значение $s(2) = 16$. 3) При $s \geq 19$ коэффициент $\tilde{c}(s) > s$. В терминах теории кодирования это означает, что в булевой модели планирования отсеивающих экспериментов [3] пропускная способность для средней вероятности ошибки отлична от пропускной способности при нулевой ошибке (пропускной способности для максимальной вероятности ошибки).

3. Границы случайного кодирования

Пусть $x = \|x_i(j)\|$, $i = \overline{1, N}$, $j = \overline{1, t}$ — случайная матрица, все $N \cdot t$ элементов которой являются независимыми в совокупности одинаково распределенными случайными величинами с распределением

$$P\{x_i(j)=0\} = \beta, \quad P\{x_i(j)=1\} = 1 - \beta,$$

где число β , $0 < \beta < 1$, которое выберем позже, называется параметром рандомизации. Обозначим через $\mathcal{A}_N(s, t, L)$ событие, состоящее в том, что случайная матрица x не удовлетворяет определению 1 (s, t, L) -кода. Нетрудно понять, что для вероятности события $\mathcal{A}_N(s, t, L)$ справедливо неравенство

$$P\{\mathcal{A}_N(s, t, L)\} \leq C_t^{s+L} C_{s+L}^L q(s, L, \beta)^N, \quad (7)$$

где

$$q(s, L, \beta) = 1 - \beta^s(1 - \alpha^L). \quad (8)$$

Поскольку правая часть (8) минимизируется при

$$\beta = \beta_s(L) \triangleq \left(\frac{s}{s+L} \right)^{1/L}, \quad (9)$$

то из (7), используя стандартные рассуждения метода случайного кодирования [5], нетрудно доказать следующее утверждение.

Теорема 5. Для любых $1 \leq s < t$, $L \leq t - s$ число

$$N(s, t, L) \leq \left[C(s, L) \log t + \frac{\log(s! L!)}{\log q(s, L, \beta_s(L))} \right],$$

где использованы обозначения (8), (9) и

$$C(s, L) = \frac{s+L}{-\log q(s, L, \beta_s(L))}. \quad (10)$$

Из теоремы 5 вытекает

Следствие 2. В асимптотических условиях (4) величина

$$N(s, t, L) \leq C(s, L) \log t(1 + o(1)), \quad (11)$$

где $C(s, L)$ определена (8)–(10).

При $L=1$ коэффициент задается формулой

$$C(s, 1) = \frac{s+1}{-\log [1 - s^s / (s+1)^{s+1}]},$$

которая при $s \rightarrow \infty$ означает, что

$$C(s, 1) = \frac{e}{\log e} s^2(1 + o(1)) = 1,88417 \cdot s^2(1 + o(1)).$$

Отметим, что аналогичная асимптотическая формула для коэффициента $c(s, 1)$ нижней границы следствия 1 имеет вид

$$c(s, 1) = K(s) = \frac{s^2}{2 \log s} (1 + o(1)).$$

Для сравнения с таблицей $K(s)$ из раздела 2 приведем несколько числовых значений $C(s, 1)$:

$$C(3,1) = 24,8762 \quad C(4,1) = 40,5487 \quad C(5,1) = 59,9883 \quad C(6,1) = 83,1955.$$

Если положить $L = ls$, где $l \leq 1/s$ — постоянная, не зависящая от s , то (11) можно записать в форме

$$N(s, t, ls) \leq sg(l) \log t(1 + o(1)), \quad t \rightarrow \infty, \quad (12)$$

где функция $g(l)$ параметра $l \geq 1/s$ задается формулой

$$g(l) = \frac{1+l}{-\log \left[1 - \left(\frac{1}{1+l} \right)^{1/l} \cdot \frac{l}{1+l} \right]}$$

Пусть l_0 — единственное значение параметра $l > 0$, при котором

$$g(l_0) = \min_{l > 0} g(l).$$

Численные расчеты показывают, что $l_0 = 2,235$, а

$$g(l_0) = g(2,235) = 4,269\,315.$$

Приведем для сравнения еще ряд значений функции $g(l)$:

$$\begin{array}{lll} g(1/2) = 6,484\,36 & g(2/3) = 5,616\,82 & g(3/4) = 5,339\,40 \\ g(1) = 4,818\,84 & g(2) = 4,278\,95 & g(3) = 4,335\,21. \end{array}$$

Нетрудно увидеть, что с увеличением s при $L = ls$, где l близко к l_0 , коэффициент $C(s, L) = sg(l)$ в границе (12) становится существенно меньше коэффициента $C(s, 1) = sg(1/s)$.

Для случая (\tilde{s}, t) -кодов ограничимся лишь формулировками границ случайного кодирования, которые аналогичны теореме 5 и следствию 2. Следующая теорема доказана в [6].

Теорема 6. Для любого $s < t$ число

$$\tilde{N}(s, t) \leq \lceil \tilde{C}(s) \log t + b_s \rceil,$$

где

$$\tilde{C}(s) = \frac{s+1}{-\log [1 - 2s^s/(s+1)^{s+1}]},$$

$$b_s = \begin{cases} \frac{\log(2^s/s!)}{-\log [1 - 2s^s/(s+1)^{s+1}]}, & \text{если } s = \overline{1, 3}, \\ 0, & \text{если } s \geq 4. \end{cases}$$

Следствие 3. В асимптотических условиях (4) величина

$$\tilde{N}(s, t) \leq \tilde{C}(s) \log t(1 + o(1)).$$

С целью удобства сравнения верхней границы следствия 3 с нижней границей следствия 1 приведем некоторые численные значения $\tilde{C}(s)$ и дадим асимптотические ($s \rightarrow \infty$) формулы для коэффициентов обеих границ:

$$\tilde{C}(2) = 5,92 \quad \tilde{C}(3) = 11,70 \quad \tilde{C}(4) = 19,37 \quad \tilde{C}(5) = 28,92,$$

$$\tilde{C}(s) = \frac{e}{2 \log e} s^2(1 + o(1)) = 0,94208 \cdot s^2(1 + o(1)),$$

$$\tilde{c}(s) = \frac{s^2}{8 \log s} (1 + o(1)).$$

4. Коды Каутса-Синглетона

Теоремы предыдущего раздела являются лишь теоремами существования и не дают рецептов построения конкретных кодов, длины которых удовлетворяют приведенным в них границам. Первый же вопрос, который возникает, например, при применении теоремы 5, состоит в следующем: сколько вычислительных операций Q надо произвести для того, чтобы проверить, что данная матрица x , размеры которой соответствуют границе теоремы 5, удовлетворяет определению $(s, t, 1)$ -кода? Если одной вычислительной операцией считать вычисление булевой суммы и сравнение на покрытие двух двоичных (из 0 и 1) столбцов длины N , то число Q , очевидно, имеет порядок t^{s+1} . При значениях $t \sim 10^3 - 10^4$, $s \sim 5 - 15$, которые возникают в приложениях [1], Q становится астрономически большим, т.е.

$$Q \sim 10^{18} - 10^{64}.$$

Можно ли найти какое-либо простое достаточное условие того, что матрица x есть $(s, t, 1)$ -код, проверка которого занимает существенно меньшее число Q вычислительных операций? Такое очевидное условие, которое сформулировано ниже в виде теоремы 7, предложено в [1].

Теорема 7. Пусть $x = \|x_i(j)\|$ — двоичная матрица размера $N \times t$, столбцы которой имеют одинаковое число единиц

$$w = \sum_{i=1}^N x_i(j), \quad j = \overline{1, t}. \quad (13)$$

Пусть при $k \neq j$

$$\lambda_{kj} \triangleq \sum_{i=1}^N x_i(k)x_i(j) \quad (14)$$

число строк x , в которых k -ый и j -ый столбцы содержат одновременно единицы, а

$$\lambda \triangleq \max_{k \neq j} \lambda_{kj}. \quad (15)$$

Тогда матрица x является $(s, t, 1)$ -кодом для любого s , удовлетворяющего неравенству

$$s \leq \left\lfloor \frac{w-1}{\lambda} \right\rfloor. \quad (16)$$

Проверка достаточного условия Каутса–Синглтона, т.е. вычисление числа $s = \left\lfloor \frac{w-1}{\lambda} \right\rfloor$ для матрицы x , столбцы которой имеют одинаковое число w единиц, занимает $Q = C_t^2 \sim t^2$ вычислительных операций, если под одной операцией понимать вычисление λ_{kj} . Это число при рассмотренных выше значениях параметров t и s имеет порядок $Q \sim 10^6 - 10^8$, который приемлем с точки зрения практической реализации.

Пусть $1 \leq \lambda \leq w \leq N$ — заданные натуральные числа.

Определение 3. Матрицу x размера $N \times t$, удовлетворяющую условиям (13)–(15), назовем $\{t, w, \lambda\}$ -матрицей и обозначим через $n\{t, w, \lambda\}$ минимально возможное число строк $\{t, w, \lambda\}$ -матрицы.

Определение 4. Пусть $1 \leq s < t$. Будем говорить, что $\{t, w, \lambda\}$ -матрица является (s, t) -кодом (или $(s, t, 1)$ -кодом Каутса–Синглтона), если справедливо неравенство (16).

Введем минимально возможное число строк (s, t) -кода

$$\bar{N}(s, t) \triangleq \min_{(16)} n\{t, w, \lambda\},$$

где минимум берется по параметрам w и λ , удовлетворяющим (16).

Цель этого раздела — исследование нижних и верхних границ $\bar{N}(s, t)$. Из теоремы 7 вытекает, что всякий (s, t) -код является также и $(s, t, 1)$ -кодом. Поэтому в качестве нижних границ $\bar{N}(s, t)$ можно рассматривать нижние границы $N(s, t, 1)$, описанные в разделах 1–2. Для их усиления нам понадобится

Лемма. 1) Число строк N любой $\{t, w, \lambda\}$ -матрицы удовлетворяет неравенствам

$$N \geq \frac{tw^2}{\lambda(t-1) + w}, \tag{17}$$

$$C_N^{\lambda+1} \geq C_w^{\lambda+1} \cdot t. \tag{18}$$

2) Если $\{t, w, \lambda\}$ -матрица является $(\overline{s, t})$ -кодом длины $N \leq t - 1$, то $w \geq s + 1$.

Неравенство (17) есть очевидное следствие границы Джонсона для равновесных кодов [7], неравенство (18) и утверждение 2) доказаны в [1]. Из (16), (17) и утверждения 2) леммы вытекает

Теорема 8. Для любого $s < t$ число

$$\bar{N}(s, t) \geq \min \left\{ t; \frac{s(s+1)}{1+s/t} \right\} \geq \min \{t; s^2\}.$$

Граница теоремы 8 примерно в два раза улучшает нижнюю границу $\bar{N}(s, t)$, даваемую теоремой 1. Из теоремы 8 и неравенства (1) имеем

Следствие 4. Для любого $s < t$ величина

$$\begin{aligned} \bar{N}(s, t) &\geq d(s, t) \triangleq \max \{d_1(s, t); d_2(s, t)\}, \\ d_1(s, t) &= \lceil \log C_t^s \rceil, \\ d_2(s, t) &= \min \{t, s^2\}. \end{aligned} \tag{19}$$

Из (16)–(17) следует, что во всяком $(\overline{s, t})$ -коде отношение

$$v \triangleq \frac{w}{N} \leq v(s, t) \triangleq s^{-1} + t^{-1}, \tag{20}$$

а отношение

$$u \triangleq \frac{\lambda}{N} \leq \frac{w}{sN} = \frac{v}{s}. \tag{21}$$

Если для оценок левой и правой частей (18) воспользоваться известными границами биномиальных коэффициентов (см. [5], задача 5.8) и при этом учесть (20)–(21), то нетрудно доказать следующую теорему.

Теорема 9. Пусть заданные натуральные числа s, t и d таковы, что $3 \leq s < t$,

$$v(s, t) \leq \frac{1}{2} u$$

$$\frac{v(s, t)}{s} + \frac{1}{d} < \frac{1}{es + 1}. \tag{22}$$

Тогда для произвольного $(\overline{s, t})$ -кода, длина которого $N \geq d$, имеет место неравенство

$$N \geq F(s, t, d) \triangleq \left[\frac{\log t + \frac{1}{2} \log(\pi/4)}{h(s^{-1}v(s, t) + d^{-1}) - v(s, t)h(s^{-1})} \right],$$

где $h(u)$ — двоичная энтропия.

Покажем, что с помощью теоремы 9 границу (19) можно уточнить, если определенная (19) величина $d = d(s, t)$ удовлетворяет (22). Введем рекуррентную монотонно неубывающую последовательность $D_0, D_1, D_2, \dots, D_k = D_k(s, t)$ натуральных чисел

$$D_k = \begin{cases} D_{k-1}, & \text{если } F(s, t, D_{k-1}) \leq D_{k-1}, \\ F(s, t, D_{k-1}), & \text{если } F(s, t, D_{k-1}) > D_{k-1}, \end{cases}$$

а $D_0 \triangleq d(s, t)$ определено (19). Пусть $k_0, k_0 = 0, 1, 2, \dots$, — наименьшее из чисел k , для которых $D_k = D_{k+1}$. Введем

$$D(s, t) = D_{k_0}(s, t).$$

Тогда очевидно, что $D(s, t) \geq d(s, t)$ и справедлива оценка

$$\bar{N}(s, t) \geq D(s, t). \quad (23)$$

Следствие 5. При $t \rightarrow \infty$, $s = \text{const}$ имеет место асимптотическое неравенство

$$\begin{aligned} \bar{N}(s, t) &\geq K_s \log t(1 + o(1)), \\ K_s &= [h(s^{-2}) - s^{-1}h(s^{-1})]^{-1}, \end{aligned} \quad (24)$$

которое аналогично нижней границе следствия 1 из раздела 2 для $N(s, t, 1)$.

Применяя неравенство $h(u) < u \log(e/u)$, можно проверить, что $K_s \geq s^2 / \log(se)$. Отсюда вытекает менее точная, чем (24), но зато более простая по виду асимптотическая оценка

$$\bar{N}(s, t) \geq \frac{s^2 \log t}{\log(se)} (1 + o(1)).$$

С помощью алгебраических методов теории кодирования [8] Каутс и Синглетон в [1] построили семейство $(\overline{s, t})$ -кодов, параметры которого

записываются следующим образом. Пусть $k \geq 2$ — натуральное число, а $q \geq k - 1$ — простое или степень простого числа. Тогда

$$\begin{aligned}
 t &= q^k, & s &= \lfloor q/(k-1) \rfloor, \\
 N &= q[1 + (k-1)s], & w &= q + 1, & \lambda &= k - 1.
 \end{aligned}
 \tag{25}$$

Другие известные классы регулярных $(\overline{s, t})$ -кодов даны в [1]. Для сравнения дадим небольшую числовую таблицу значений параметров семейства (25), а также границ (19), (23) и (24).

k	q	t	s	N (25)	$d(s, t)$ (19)	$D(s, t)$ (23)	K_s (24)
3	23	12 167	11	529	125	235	34,363
4	31	923 581	10	961	177	390	29,505
3	32	32 768	16	1056	256	455	63,318
3	16	4 096	8	272	81	128	20,760
5	16	1 048 576	4	272	76	120	7,437

Данный раздел завершается формулировкой верхней асимптотической границы $\overline{N}(s, t)$, получаемой с помощью метода случайного кодирования, когда

$$t \rightarrow \infty, \quad s = \text{const.} \tag{26}$$

Теорема 10. В условиях (26)

$$\overline{N}(s, t) \leq E_s \log t (1 + o(1)),$$

где

$$E_s = \left[\max_{0 < \beta < s^{-1}} E(s, \beta) \right]^{-1},$$

$$E(s, \beta) = h(\beta) - \beta h(s^{-1}) - (1 - \beta) h\left(\frac{\beta(s-1)}{(1-\beta)s}\right).$$

При $s = 2$ и $s = 3$ получены следующие числовые значения

$$E_2 = 10,6213, \quad E_3 = 28,6090.$$

Можно показать также, что

$$E_s = as^2(1 + o(1)), \quad s \rightarrow \infty,$$

где $a = 4,28127$.

5. Открытые задачи

В заключение можно сформулировать в виде вопросов три открытые проблемы теории дизъюнктивных кодов, которые представляют, на наш взгляд, наибольший интерес.

- 1) Можно ли улучшить нижнюю границу теоремы 3?
- 2) Можно ли усилить неравенство (2), а затем улучшить границу (5) при всех $L \geq 2$, $s \geq 2$?
- 3) Существует ли обобщение достаточного условия Каутса–Синглтона из раздела 4 для случая (s, t, L) -кодов при $L \geq 2$?

Литература

1. Kautz W. H., Singleton R. C. Nonrandom Binary Superimposed codes. IEEE Trans. Inform. Theory 1964, **10**, 4, 363–377.
2. Dyachkov A. S., Rykov V. V. 1981, One application of codes for multiple access channel to ALOHA-system. 6-th All-Union Seminar on Computer Networks. Papers, v. 4, pp. 18–24.
3. Дьячков А. Г. Границы вероятности ошибки для двух моделей рандомизированного планирования отсеивающих экспериментов. Проблемы передачи информации, 1979, **15**, 4, 17–31.
4. Дьячков А. Г., Рыков В. В. Границы длины дизъюнктивных кодов. Проблемы передачи информации, 1982, **18**, 3, 7–13.
5. Gallager R. G. Information theory and reliable communication J. Wiley, New York, 1968.
6. Malyutov M. B. 1976, On planning screening experiments. Proceedings of the 1975 IEEE-USSR Joint Workshop on Inform. Theory. N. Y., IEEE Inc. 144–147.
7. Johnson S. M. A new upper bound for error correcting codes, IEEE Trans. Inform. Theory, 1962, **8**, 3, 203–207.
8. Berlecamp E. R. Algebraic coding theory. McGraw-Hill Book Company, New York, 1968.

A survey of superimposed code theory

A. G. DYACHKOV, V. V. RYKOV
(Moscow)

Generalisation of superimposed code concept, introduced by Kautz and Singleton [1], is considered in this paper. New results of superimposed code theory are described. Some open problems are formulated.

А. Г. Дьячков

Московский Государственный университет им. М. В. Ломоносова,
мех.-мат. фак.

СССР, Москва В-234, Ленинские горы

ON TWO-NODE EXPONENTIAL QUEUEING NETWORK WITH INTERNAL LOSSES OR BLOCKING

P. P. BOCHAROV, F. J. ALBORES

(Moscow)

(Received June 30, 1982)

A two-node open exponential queueing network with one server on each node is studied. The first node is of unlimited capacity. The buffer length of the second node is equal to m ($m < \infty$). Three disciplines of the node conjugation which coordinate their operations with respect to limited buffer length of the second node are considered. These disciplines are:

- discipline with internal losses;
- blocking with halting the first-node server's operation;
- blocking with repeated service on the first-node server.

Necessary and sufficient ergodicity conditions and an algorithm for calculation of the queueing network stationary state probabilities are obtained.

1. Introduction

We consider an open queueing network (QN) consisting of two single-server nodes. The first node is of unlimited capacity. The buffer length of the second node is equal to m ($m < \infty$). The exogenous source (node 0) generates the flow of customers according to the Poisson law with total intensity λ . The arriving of the customers to the nodes and their routing inside the QN are scheduled by the transition matrix **T**

T	0	1	2
0	0	p	$1-p$
1	$1-\alpha-\beta$	β	α
2	$1-\gamma-\vartheta$	θ	γ

Customer service times at the first and the second nodes are independent of each other and exponentially distributed with parameters μ_1 and μ_2 respectively. Figure 1 shows the corresponding network.

In Kendall notation this network is classified as

$$(M|M|1|\infty; \quad M|1|m).$$

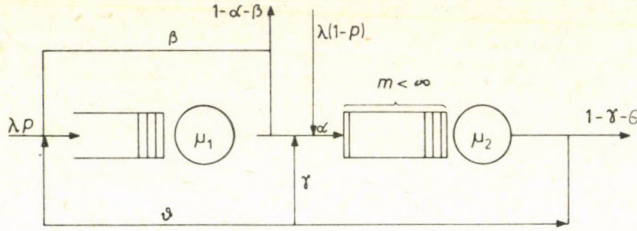


Fig. 1. Two-node queueing network

Customers at each node are served in order of their arrival. If a customer (for the sake of brevity it will be called below *c*-customer) served at the first node is routed to the second node and finds its buffer completely occupied then the further behaviour of such a customer is defined by one of the node conjugation disciplines fixed for a given model [1-3]:

- discipline *L* where the *c*-customer is lost by the system;
- discipline *B* where the *c*-customer halts the first-node server operation until the queue length in the second node falls up to $m - 1$;
- discipline *BS* where the *c*-customer is repeatedly served at the first node with the same distribution of the service times until at the end of the next service there is at least one waiting place in the second node.

The necessity for investigation of such kind of systems arises for example when modeling structure describes the computer operation [4-6].

This paper generalizes the results of publication [3] in which the case $p = 1$, $\alpha = 1$, $\gamma = 0$ was considered. For all the three conjugation disciplines necessary and sufficient system ergodicity conditions and an algorithm for calculation of the QN stationary state probabilities are obtained.

2. Discipline *L*

The considered QN may be described by a uniform Markov process $X(t)$, $t \geq 0$, over the state space

$$\mathcal{X} = \{(i, j) | i = 0, 1, \dots; \quad j = \overline{0, R}\}$$

where (i, j) means that at time t nodes one and two have i and j customers respectively, $R = m + 1$.

Let

$$p_{ij} = \lim_{t \rightarrow \infty} \mathcal{P}\{X(t) = (i, j)\}.$$

The stationary state probabilities p_{ij} satisfy the following system of equilibrium equations (SEE):

$$\begin{aligned} & [\lambda p + u(R-j)\lambda(1-p) + u(i)\mu_1(1-\beta) + u(j)\mu_2(1-\gamma)]p_{ij} = \\ & = u(i)\lambda p p_{i-1,j} + u(j)\lambda(1-p)p_{i,j-1} + u(j)\mu_1 \alpha p_{i+1,j-1} + \\ & + \mu_1(1-\alpha-\beta)p_{i+1,j} + u(R-j)\mu_2(1-\gamma-\vartheta)p_{i,j+1} + \\ & + u(i)\mu_2 \vartheta u(R-j)p_{i-1,j+1} + \mu_1 \alpha u(j-m)p_{i+1,j}, \end{aligned} \quad (2.1)$$

$$i \geq 0, \quad j = \overline{0, R},$$

where

$$u(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

In terms of generating functions (GF)

$$P_j(z) = \sum_{i=0}^{\infty} p_{ij} z^i, \quad |z| \leq 1, \quad j = \overline{0, R},$$

SEE may be written in the following form:

$$\begin{aligned} & [a(z) - \mu_2(1-\gamma)z]P_0(z) - zb(z)P_1(z) = f_0(z), \\ & -c(z)P_{j-1}(z) + a(z)P_j(z) - zb(z)P_{j+1}(z) = f_j(z), \quad j = \overline{1, m}, \\ & -c(z)P_{R-1}(z) + [a(z) - c(z)]P_R(z) = f_R(z), \end{aligned} \quad (2.2)$$

where

$$\begin{aligned} f_k(z) = \mu_1 \{ [(1-\beta)(z-1) + \alpha u(R-j)]p_{0,k} - \alpha u(j)p_{0,k-1} \}, \\ k = \overline{0, R} \end{aligned} \quad (2.3)$$

and

$$\begin{aligned} a(z) &= (1-z) [\lambda pz - \mu_1(1-\beta)] + c(z) + \mu_2(1-\gamma)z, \\ b(z) &= \mu_2(\vartheta z + 1 - \gamma - \vartheta), \\ c(z) &= \lambda(1-p)z + \mu_1 \alpha. \end{aligned} \quad (2.4)$$

Solution of (2.2) obtained by Kramer rule is

$$P_j(z) = S_j(z)/D_R(z), \quad j = \overline{0, R}, \quad (2.5)$$

where

$$S_j(z) = D_{R-j-1}(z) \sum_{k=0}^{j-1} (c(z))^{j-k} U_{k-1}(z) f_k(z) - U_{j-1}(z) \sum_{k=j}^R (zb(z))^{k-j} D_{R-k-1}(z) f_k(z) \quad (2.6)$$

and polynomials $U_k(z)$ and $D_k(z)$ are defined by the recursive relations

$$\begin{cases} U_{-1}(z) = 1, \\ U_0(z) = a(z) - \mu_2(1-\gamma)z, \\ U_k(z) = a(z)U_{k-1}(z) - zc(z)U_{k-2}(z), \quad k > 0; \end{cases} \quad (2.7)$$

$$\begin{cases} D_{-1}(z) = 1, \\ D_0(z) = a(z) - c(z), \\ D_k(z) = a(z)D_{k-1}(z) - zc(z)b(z)D_{k-2}(z), \quad k = \overline{1, m}, \\ D_R(z) = [a(z) - \mu_2(1-\gamma)z]D_{R-1}(z) - zc(z)b(z)D_{R-2}(z). \end{cases} \quad (2.8)$$

In expression (2.3) $p_{0,k}$, $k = \overline{0, R}$, are unknown. To find them we make use of the fact that by definition GF's are analytical in the domain $|z| < 1$ and continuous when $z = 1$. That is why the roots' properties of the polynomial $D_R(z)$ which is the denominator of the expression (2.5) have to be investigated.

Lemma 2.1. Let $z_{j,k}$ be the j -th real root of the polynomial $D_k(z)$, $k = \overline{0, R}$, in increasing order; $\deg(D_k(z))$ be the degree of $D_k(z)$. Then the following holds:

- 1°. $\deg(D_k(z)) = 2k + 2$.
- 2°. $D_k(0) = (-1)^{k+1} \mu_1^{k+1} (1 - \alpha - \beta)^k (1 - \beta)$,
 - 2a) $D_{2n}(0) < 0$,
 - 2b) $D_{2n+1}(0) > 0$.
- 3°. $D_k(1) > 0$, $k = \overline{0, m}$; $D_R(1) = 0$.
- 4°. $D_{2n}(\infty) < 0$, $D_{2n+1}(\infty) > 0$.
- 5°. The roots of $D_k(z)$, $k = \overline{0, R}$, are real, positive and simple except one root which is equal to unit when $k = R$.
- 6°. $D_k(z)$, $k = \overline{0, m}$, has $k + 1$ roots over $[0, 1)$.

7°. The roots of $D_k(z)$ and $D_{k-1}(z)$, $k = \overline{1, m}$, interleave as follows:

$$0 < z_{1,k} < z_{1,k-1} < z_{2,k} < \dots < z_{k,k-1} < z_{k+1,k} < 1;$$

$$1 < z_{k+2,k} < z_{k+1,k-1} < \dots < z_{2k,k-1} < z_{2k+2,k}.$$

8°. Over the intervals $[0, z_{R,R-1})$ and $(z_{R+1,R-1}, \infty)$ the roots of $D_R(z)$ and $D_{R-1}(z)$ interleave as in 7°.

9°. a) $D'_R(1) > 0 \Rightarrow z_{R+1,R} = 1$,

b) $D'_R(1) = 0 \Rightarrow z_{R+1,R} = z_{R+2,R} = 1$,

c) $D'_R(1) < 0 \Rightarrow z_{R+2,R} = 1$.

The lemma may be easily proved using publication [3].

Since unit is a root of $D_R(z)$, then if we define polynomials $g_k(z)$ so that

$$(1-z)g_k(z) = [a(z) - \mu_2(1-\gamma)z]D_{k-1}(z) - \\ -zc(z)b(z)D_{k-2}(z), \quad k \geq 0.$$

then the previous expressions and (2.8) provides us with recurrence

$$g_0(z) = \lambda pz - \mu_1(1-\beta),$$

$$g_1(z) = a(z)g_0(z) + zc(z)\mu_2\vartheta,$$

$$g_k(z) = a(z)g_{k-1}(z) - zc(z)b(z)g_{k-2}(z), \quad k \geq 2. \quad (2.9)$$

From (2.9) one may easily find that

$$g_k(1) = \{\mu_2^{k+1}(1-\gamma)^k[(1-\gamma)(\lambda p - \mu_1(1-\beta)) + (\lambda(1-p) + \mu_1\alpha)\vartheta] - \\ - (\lambda(1-p) + \mu_1\alpha)^{k+1}[\lambda p + \mu_2\vartheta - \mu_1(1-\beta)]\} / [\mu_2(1-\gamma) - \\ - (\lambda_2 + \mu_1\alpha)], \quad k \geq 0. \quad (2.10)$$

Now we demonstrate that the following holds.

Theorem 2.1. Statements

a) there exists an equilibrium state of QN $(M|M|1\infty; M|1|m)$ with discipline L ;

b) $D_R(z)$ has exactly $R+1$ roots over $[0, 1]$;

c) $D'_R(1) > 0$;

d) $g_R(1) < 0$

are equivalent.

Theorem 2.1 is proved in the way similar to the proof of the theorem 2.1 from [3].

From theorem 2.1 and (2.10) follows

Corollary 2.1. QN $(M|M|1|\infty; M|1|m)$ with discipline L is ergodic if and only if

$$\begin{aligned} & \{ \{ \mu_2^{R+1} (1-\gamma)^R [(1-\gamma)(\lambda p - \mu_1(1-\beta)) + (\lambda(1-p) + \mu_1\alpha)\vartheta] - \\ & \quad - (\lambda(1-p) + \mu_1\alpha)^{R+1} [\lambda p + \mu_2\vartheta - \mu_1(1-\beta)] \} / [\mu_2(1-\gamma) - \\ & \quad - (\lambda_2 + \mu_1\alpha)] \} \} < 0. \end{aligned} \quad (2.11)$$

Now we briefly give an algorithm to calculate the stationary state probabilities P_{ij} .

Since the matrix of system (2.2) is Jacobian, then all GF's may be expressed in terms of $P_0(z)$ which is transformed into the form

$$P_0(z) = \left\{ \mu_1 \sum_{k=0}^R (zb(z))^k [(1-\beta)D_{R-k-1}(z) - L_{R-k}(z)] p_{0,k} \right\} / g_R(z), \quad (2.12)$$

where

$$\begin{aligned} L_0(z) &= 0, \\ L_1(z) &= z\lambda p - \mu_1(1-\beta) + z\mu_2\vartheta, \\ L_k(z) &= a(z)L_{k-1}(z) - zc(z)b(z)L_{k-2}(z), \quad k \geq 2. \end{aligned} \quad (2.13)$$

Under condition (2.11), the analyticity of $P_0(z)$, (2.12) and the normalizing condition provides us with the system of equations for unknown $p_{0,k}$, $k = \overline{0, R}$

$$\begin{aligned} \sum_{k=0}^R [z_{i,R} b(z_{i,R})]^k [(1-\beta)D_{R-k-1}(z_{i,R}) - L_{R-k}(z_{i,R})] p_{0,k} &= 0, \\ & i = \overline{1, R}, \\ \sum_{j=0}^R P_j(1) &= 1. \end{aligned} \quad (2.14)$$

When probabilities $p_{0,k}$, $k = \overline{0, R}$, are determined all the GF's $P_j(z)$, $j = \overline{0, R}$, may be found. They are rational functions since

$$\deg(S_j(z)) < \deg(D_R(z)), \quad j = \overline{0, R}.$$

Let

$$\begin{aligned} S_j^*(z) &= S_j(z) / \prod_{i=1}^{R+1} (z_{i,R} - z), \quad j = \overline{0, R}, \\ D_R^*(z) &= D_R(z) / \prod_{i=1}^{R+1} (z_{i,R} - z). \end{aligned} \quad (2.15)$$

It is evident that $S_j^*(z)/D_R^*(z)$ are rational functions and hence may be decomposed into partial fractions

$$\frac{S_j^*(z)}{D_R^*(z)} = \sum_{k:z_k > 1} \frac{Q_{jk}}{z_k - z}, \quad j = \overline{0, R}, \quad (2.16)$$

where

$$Q_{jk} = - \frac{S_j^*(z_k)}{D_R^{*'}(z_k)}, \quad j = \overline{0, R}. \quad (2.17)$$

Expanding the right part of (2.17) into series by the degrees of z we get

$$p_{ij} = \sum_{k:z_k > 1} Q_{jk} z_k^{-(i+1)}, \quad i \geq 1, \quad j = \overline{0, R}. \quad (2.18)$$

Hence the problem of calculation of the stationary state probabilities $p_{i,j}$, $i \geq 0$, $j = \overline{0, R}$, for QN $(M|M|1|\infty; M|1|m)$ is solved.

3. Discipline BS

For the sake of brevity we will use the notations of Section 2. A uniform Markov process describing operation of QN with the considered node conjugation discipline is defined over the same state space as for discipline L and SEE is of the form

$$\begin{aligned} & [\lambda p + \lambda(1-p)u(R-j) + \mu_1(1-\beta)u(i) + \mu_2(1-\gamma)u(j)]p_{ij} = \\ & = \lambda p u(i)p_{i-1,j} + \lambda(1-p)u(j)p_{i,j-1} + \mu_1 \alpha u(j)p_{i+1,j-1} + \\ & + \mu_1(1-\alpha-\beta)p_{i+1,j} + \mu_2(1-\gamma-\vartheta)u(R-j)p_{i,j+1} + \\ & + \mu_2 \vartheta u(R-j)u(i)p_{i-1,j+1} + \mu_1 \alpha u(j-m)p_{i,j}, \\ & i \geq 0, \quad j = \overline{0, R}. \end{aligned} \quad (3.1)$$

In terms of GS's, SEE may be written as follows:

$$\left\{ \begin{aligned} & [a(z) - \mu_2(1-\gamma)]P_0(z) - zb(z)P_1(z) = f_0(z), \\ & -c(z)P_{j-1}(z) + a(z)P_j(z) - zb(z)P_{j+1}(z) = f_j(z), \quad j = \overline{1, m}, \\ & -c(z)P_{R-1}(z) + [(\lambda pz - \mu_1(1-\alpha-\beta))(1-z) + \\ & \quad + \mu_2(1-\gamma)z]P_R(z) = f_R(z), \end{aligned} \right. \quad (3.2)$$

where

$$f_j(z) = \mu_1 \{ [(1-\beta)(z-1) + \alpha] p_{0,j} - \alpha u(j) p_{0,j-1} \}, \quad j = \overline{0, R}. \quad (3.3)$$

Solution of (3.2) is given by (2.5)–(2.7) where $f_j(z), j = \overline{0, R}$, are defined by (3.3) and $D_j(z), j = \overline{0, R}$, are determined by recursive formulas:

$$\begin{cases} D_{-1}(z) = 1, \\ D_0(z) = [\lambda pz - \mu_1(1-\alpha-\beta)](1-z) + \mu_2(1-\gamma)z, \\ D_k(z) = a(z)D_{k-1}(z) - zc(z)b(z)D_{k-2}(z), \quad k = \overline{2, m}, \\ D_R(z) = [a(z) - \mu_2(1-\gamma)z]D_{R-1}(z) - zc(z)b(z)D_{R-2}(z). \end{cases} \quad (3.4)$$

Let us introduce such polynomials that

$$(1-z)g_k(z) = [a(z) - \mu_2(1-\gamma)z]D_{k-1}(z) - zc(z)b(z)D_{k-2}(z).$$

Polynomials $g_k(z)$ satisfy the following recursive relations:

$$\begin{aligned} g_0(z) &= \lambda pz - \mu_1(1-\alpha-\beta), \\ g_1(z) &= a(z)g_0(z) + z\mu_2[\vartheta c(z) - \mu_1\alpha(1-\gamma)], \\ g_k(z) &= a(z)g_{k-1}(z) - zc(z)b(z)g_{k-2}(z), \quad k \geq 2. \end{aligned} \quad (3.5)$$

Investigating properties of polynomials $D_k(z)$ and $g_k(z)$ one may easily see that lemma 2.1 of section 2 holds true in this case except property 2°, which takes form

$$D_k(0) = (-1)^{k+1} \mu_1^{k+1} (1-\alpha-\beta)^k (1-\beta).$$

Also holds true for discipline *BS* theorem 2.1 of section 2 from which follows

Corollary 3.1. Necessary and sufficient condition for ergodicity of QN ($M|M|1|\infty; M|1|m$) is

$$\begin{aligned} & \{ \{ [\mu_2(1-\gamma)]^R \mu_2 [(1-\gamma)(\lambda p - \mu_1(1-\beta)) - \vartheta(\lambda(1-p) + \mu_1\alpha)] - \\ & - [\lambda(1-p) + \mu_1\alpha]^R [\lambda(1-p) + \mu_1\alpha] (\lambda p - \mu_1(1-\alpha-\beta) + \mu_2\vartheta) - \\ & - \mu_1\mu_2(1-\gamma)\alpha \} / \{ \mu_2(1-\gamma) - (\lambda(1-p) + \mu_1\alpha) \} \} < 0. \end{aligned} \quad (3.6)$$

Calculation of the stationary state probabilities $p_{i,j}, i \geq 0, j = \overline{0, R}$, is performed in the way similar to the case with discipline *L*, but polynomials $L_k(z), D_k(z)$ and $g_k(z)$ are defined by the other recursive relations.

4. Discipline B

It is easy to see that if $p=1$ and $\alpha=1$ then the analysis of a given QN with discipline *BS* and *B* leads to solution of similar SEE's and reduction from discipline *BS* to *B* may be done by a simple substitution of m for $m+1$. In other words holds

Statement 4.1. Necessary and sufficient ergodicity condition for QN $(M|M|1|\infty; M|1|m)$ with discipline *B* when $p=\alpha=1$ is the following:

$$\left\{ \left\{ [\mu_2(1-\gamma)]^{R+2} (\lambda - \mu_1 - \vartheta \mu_1) - \mu_1^{R+2} [(\lambda + \mu_2 \vartheta) - \mu_1 \mu_2 (1-\gamma)] \right\} / [\mu_2(1-\gamma) - \mu_1] \right\} < 0. \quad (4.1)$$

In a general case when p or α are not equal to unit the analysis of QN with discipline *B* leads to some additional difficulties. But such an analysis may be carried out for arbitrary p and α in case when blocking the first-node server is caused also by the customers from the exogenous source arriving at the second node when its buffer is fully occupied (discipline *B'*). In this case holds

Statement 4.2. Necessary and sufficient ergodicity condition for QN $(M|M|1|\infty; M|1|m)$ with discipline *B'* is the following:

$$\left\{ \left\{ [\mu_2(1-\gamma)]^{R+1} [\mu_2(1-\gamma)(\lambda p - \mu_1(1-\beta)) + \vartheta(\lambda(1-p) + \mu_1 \alpha)] - [\lambda(1-p) + \mu_1 \alpha]^{R+1} [(\lambda(1-p) + \mu_1 \alpha)(\lambda p + \mu_2 \vartheta) + \mu_1 \mu_2 (1-\beta)(1-\gamma)] \right\} / [\mu_2(1-\gamma) - (\lambda(1-p) + \mu_1 \alpha)] \right\} < 0. \quad (4.2)$$

5. Conclusion

In conclusion let us note that if $m = \infty$ then the considered QN forms classical two-stage open exponential network. In this case by Jackson Theorem [7] we get

$$p_{i,j} = (1 - \rho_1)(1 - \rho_2) \rho_1^i \rho_2^j, \quad i, j \geq 0, \quad (5.1)$$

where $\rho_i = \lambda_i / \mu_i$ and $\lambda_i, i = 1, 2$, are found from the system of equations

$$(\lambda_0, \lambda_1, \lambda_2) \mathbf{T} = (\lambda_0, \lambda_1, \lambda_2), \quad (5.2)$$

where $\lambda_0 = \lambda$, and are given by

$$\begin{aligned} \lambda_1 &= \{ [p(1-\gamma) + (1-p)\vartheta] / [(1-\beta)(1-\gamma) - \alpha\vartheta] \} \lambda = \alpha_1 \lambda, \\ \lambda_2 &= \{ [(1-p)(1-\beta) + \alpha p] / [(1-\beta)(1-\gamma) - \alpha\vartheta] \} \lambda = \alpha_2 \lambda. \end{aligned} \quad (5.3)$$

Formula (5.1) may be used for the check of calculation when the parameter m is large for all the node conjugation disciplines.

It is known from [8] that when $m = \infty$ the necessary and sufficient condition for ergodicity of the QN is

$$\lambda < \min_{i=1,2} \frac{\mu_i}{\alpha_i}. \quad (5.4)$$

Allowing in each of the conditions (2.11), (3.6), (4.1) and (4.2) $m \rightarrow \infty$, we get the ergodicity condition (5.4).

References

1. Pujolle, G., File d'attente en serie et application au taux de charge maximale d'un réseau d'ordinateurs. Revue AFJRO—Supplement vol. 10, 5, 1976.
2. Bocharov, P. P., On multi-stage system of finite capacity and blocking with reseriving, "Teoriya telegrafika i seti s upravlyaemymi elementami", Nauka Publ., 1980 (in Russian).
3. Bocharov, P. P., Albores, F. J., On two-stage exponential queueing system with internal losses or blocking. Problems of Control and Information Theory, vol. 9, 5, 1980.
4. Konheim, A. G., Reiser, M., A queueing model with finite waiting room and blocking. J. ACM, vol. 23, 2, 1976.
5. Artamonov, G. T., Brekhov, O. M., Analytical probabilistic models of computer operation, Energiya Publ., 1978 (in Russian).
6. Aven, O. I., Kogan, Ya. A., Control of computations, Energiya Publ., 1978 (in Russian).
7. Jackson, J. R., Networks of waiting-lines. JORSA, vol. 5, 4, 1957.
8. Kaufmann, A., Cruon, R., Les phénomènes d'attente. Théorie et applications. Dunod, Paris, 1961.

О двухузловой экспоненциальной сети массового обслуживания с внутренними потерями или блокировками

П. П. БОЧАРОВ, Ф. Х. АЛЬБОРЕС

(Москва)

Рассмотрена разомкнутая экспоненциальная сеть массового обслуживания, состоящая из двух однолинейных узлов. Первый узел неограниченной емкости, число мест для ожидания на втором узле ограничено. Рассматриваются три дисциплины сопряжения узлов, согласующие их функционирование с учетом накопителя конечной емкости на втором узле: дисциплина с внутренними потерями заявок, блокировка с остановкой первого прибора и блокировка с повторением обслуживания на первом приборе. Получены необходимые и достаточные условия эргодичности сети и алгоритм расчета стационарных вероятностей состояний сети.

П. П. Бочаров, Ф. Х. Альборес

Университет дружбы народов им. П. Лумумбы

СССР, 117923, Москва, ГСП,

ул. Орджоникидзе, 3

DUAL ALGORITHM OF OPTIMIZATION OF A LINEAR DYNAMIC SYSTEM*

R. GABASOV, F. M. KIRILLOVA, O. I. KOSTYUKOVA

(Minsk)

(Received June 26, 1982)

A linear terminal problem of optimization of a dynamic system in the class of bounded continuous functions possessing bounded piece-wise continuous derivatives is considered. A new approach is suggested which is based on the methods worked out by the authors earlier. Dual method is investigated in detail. New notions of support and support control are introduced. The formula of increment is obtained. Optimality criterion in the form of the maximum principle is proved without the use of measures. New ε -maximum principle — suboptimality criterion is grounded. Finite dual algorithm of construction of optimal (suboptimal) control is described in detail. The example of optimization of a vibrating system is considered.

1. Introduction

Modern theory of optimal processes [1] has received its most complete development for the case of admissible controls which are represented by piece-wise continuous functions. The problems of optimization of dynamic systems with the help of controls having bounded piece-wise continuous derivatives are reduced in the majority of cases to mathematical optimal control problems with phase constraints. These are the most difficult problems of optimal control theory. The qualitative results achieved in this field are far from being sufficient for the numerical solution of such problems [2].

In the given paper the dual algorithm of optimization of a linear system with the help of control having bounded piece-wise continuous derivatives has been constructed on the basis of the methods of solution of linear programming problems and optimal control problems without phase constraints worked out by the authors [3–8] earlier. The class of controls with constraints on the speed of the changes of influences has been considered. The used technique of investigation allows us to transfer the received results on other classes of controls.

The algorithm is illustrated by the example of optimization of a vibrating system.

* After the material of the report to Equadiff V (Bratislava, August 1981).

In the paper the questions of existence of optimal controls are not discussed as the suggested algorithm practically solves any problem of the class considered. It is clear that the questions of stability of computation, of computer experiment and so on arise. But these topics are the matter of independent investigations. The experience of numeric solution of other classes of optimal control problems [4-8] which has been accumulated by now proves the advantages of the approach on which the algorithm of the given paper is based.

Let us remark that with the help of the references of the article one can establish independently the ties between some results of the given paper and the known facts of qualitative theory of optimal control.

2. Statement of the problem. Support control

Continuous upper bounded functions $u(t)$, $t \in T = [0, t^*]$, satisfying the conditions

$$u(t) \leq \alpha, \quad t \in T, \quad u(0) = u_0, \quad (1)$$

taking top value $u(t) = \alpha$ on the finite number of segments and having on T bounded piece-wise continuous derivatives

$$|\dot{u}(t)| \leq 1, \quad t \in T, \quad (2)$$

are said to be admissible controls.

In the class of admissible controls consider the problem of maximization of the cost function

$$J(u) = c'x(t^*) \rightarrow \max \quad (3)$$

on the trajectories $x(t)$, $t \in T$, of the controlled according to Kalman system

$$\dot{x} = Ax + bu, \quad x(0) = 0, \quad (4)$$

where $x = x(t)$ is n -vector of state of the system at moment t , $u = u(t)$ is scalar (controlling influence); A is an $n \times n$ -matrix, b , c are n -vectors; α , $u_0 \leq \alpha$ are the given numbers; $\dot{x} = \dot{x}(t) = dx/dt$.

The existence of two constraints (1), (2) on control makes problem (1)-(4) related to the problem with phase constraints: after introducing new phase constraints: after introducing a new phase variable $x_{n+1} = u$ and a new control variable $v = \dot{x}_{n+1}$ we get the optimal control problem with constraint on control $|v| \leq 1$ and on phase variable $x_{n+1} \leq \alpha$.

On the set T let us pick out the segments $T_i = [\tau_i, \tau^i]$, $i = \overline{1, M}$ constructed in such a way that $\tau_i \leq \tau^i < \tau_{i+1}$. The totality $Q = \{T_i, i = \overline{1, M}\}$ is said to be the support of problem (1)-(4).

Functions

$$\xi(t) = \psi'(t)b, \quad t \in T_{\Phi}^0 = \bigcup_{i=1, M} (\tau_i, \tau^i); \quad \xi(t) = 0, \quad t \in T \setminus T_{\Phi}^0; \quad (5)$$

$$\Delta(t) = \int_i^{t^*} (\xi(\tau) - \psi'(\tau)b) d\tau, \quad t \in T, \quad (6)$$

are put in accordance with the support Q , where $\psi(t)$, $t \in T$, is the solution of the conjugate system

$$\dot{\psi}(t) = -A'\psi(t), \quad \psi(t^*) = c. \quad (7)$$

The pair $\{u(\cdot), Q\}$ consisting of the admissible control $u(\cdot) = \{u(t), t \in T\}$ and support of the problem is said to be support control.

The support $Q_u = \{T_i, i = 1, M\}$ is said to accompany the admissible control $u(\cdot)$ if segments T_i consist of such and only such points on which the control takes the value of α .

3. Maximum and ε -maximum principles

Together with the admissible control $u(\cdot)$ consider the admissible control $\bar{u}(\cdot) = u(\cdot) + \omega(\cdot)$ and calculate the increment of the cost function

$$\Delta J(u) = J(\bar{u}) - J(u) = c' \Delta x(t^*), \quad (8)$$

where $\Delta x(t)$, $t \in T$ is the solution of equation

$$\Delta \dot{x}(t) = A \Delta x(t) + b \omega(t), \quad \Delta x(0) = 0, \quad t \in T. \quad (9)$$

As $\psi'(t) = c' F(t^*, t)$, $t \in T$, where $F(t, \tau)$ is a fundamental matrix of solutions;

$$\partial F(t, \tau) / \partial t = A F(t, \tau);$$

$$\partial F(t, \tau) / \partial \tau = -F(t, \tau) A, \quad F(t, t) = E;$$

then denoting $f(t)$, $t \in T$, the solution of equation $df/dt = Af + b$, $f(0) = 0$, and expressing the solution of equation (9) with the help of the Cauchy formula from (8) we get

$$c' \Delta x(t^*) = c' \int_0^{t^*} [f(t^*) - F(t^*, t) f(t)] \dot{\omega}(t) dt = \int_0^{t^*} c(t) \dot{\omega}(t) dt. \quad (10)$$

Here $c(t) = c' f(t^*) - c' F(t^*, t) f(t)$.

For any piece-wise continuous function $\xi(t)$, $t \in T$, the equalities

$$\Delta(t) = \int_t^{t^*} (\xi(\tau) - \psi'(\tau)b)d\tau = \int_t^{t^*} \xi(\tau)d\tau - c(t), \quad t \in T. \quad (11)$$

are true.

Having placed (11) into (10) we get a formula of increment of the cost function

$$\Delta J(\bar{u}) = c' \Delta x(t^*) = \int_0^{t^*} (-\Delta(t)\dot{c}(t) + \xi(t)\omega(t))dt. \quad (12)$$

Suppose that $\bar{u}(\cdot) = u^0(\cdot)$ is optimal control, $\xi(t)$, $\Delta(t)$, $t \in T$ are functions (5), (6) corresponding to the support Q . Then from (12) follows the inequality

$$J(u^0) - J(u) \leq \beta(u(\cdot), Q), \quad (13)$$

where

$$\beta(u(\cdot), Q) = \int_0^{t^*} (\varepsilon_v(t) + \varepsilon_u(t))dt, \quad (14)$$

$$\varepsilon_v(t) = \Delta(t)\dot{u}(t) - \min_{v \in V} \Delta(t)v,$$

$$\varepsilon_u(t) = \max_{u \in U} \xi(t)u - \xi(t)u(t). \quad (15)$$

(4) is said to be a suboptimality estimate of the support control $\{u(\cdot), Q\}$.

From (14), (15) it is given that $\beta(u(\cdot), Q) < \infty$ only if $\xi(t) \geq 0$, $t \in T$.

By traditions of the theory of optimal processes the optimality and suboptimality criteria will be formulated in the form of extremal principles.

Theorem 1. The admissible control $u(t)$, $t \in T$, is optimal if and only if support control $\{u(\cdot), Q_u\}$ satisfies the conditions of extremum

$$\xi(t)u(t) = \max_{u \in U} \xi(t)u, \quad t \in T; \quad (16)$$

$$\Delta(t)\dot{u}(t) = \min_{|\vartheta| \leq 1} \Delta(t)\vartheta, \quad t \in T. \quad (17)$$

The proof is given in the supplement.

Theorem 2. For ε -optimality of the admissible control $u(t)$, $t \in T$ the existence of such support Q is necessary and sufficient that for the functions $\xi(t)$, $\Delta(t)$, $t \in T$, (5), (6) constructed according to it the conditions of ε extremum

$$\xi(t)u(t) = \max_{u \leq \alpha} \xi(t)u - \varepsilon_u(t), \quad t \in T,$$

$$\Delta(t)\dot{u}(t) = \min_{|\vartheta| \leq 1} \Delta(t)\vartheta + \varepsilon_v(t), \quad t \in T,$$

$$\int_0^{t^*} (\varepsilon_v(t) + \varepsilon_u(t)) dt \leq \varepsilon. \quad (18)$$

be fulfilled.

The proof is given in the supplement.

The optimal support Q^0 (i.e. the support at which for $\{u^0(\cdot), Q^0\}$ conditions (18) are fulfilled) possesses the following properties

$$\begin{aligned} \xi(t) \geq 0, \quad t \in T; \quad \Delta(t) = 0, \quad t \in [\tau_i, \tau^i], \\ \Delta(\tau_i - \delta) \leq 0, \quad \Delta(\tau_i + \delta) \geq 0, \quad i = \overline{1, M}, \end{aligned} \quad (19)$$

at sufficiently small $\delta > 0$.

The support with properties (19) is said to be regular. The empty support $Q = \emptyset$ (i.e. $M = 0, T_\emptyset^0 = \emptyset$) is regular.

4. Algorithm

Let $\xi(t)$ be piece-wise continuous, $\psi(t)$ and $\Delta(t)$, $t \in T$ be continuous functions. The problem

$$F(\xi) = (\alpha - u_0) \int_T \xi(t) dt + \int_T \zeta |\Delta(t)| dt \rightarrow \min,$$

$$\begin{aligned} \psi'(t) = -\psi'(t)A, \quad \psi(t^*) = c, \quad \dot{\Delta}(t) = -\psi'(t)b + \xi(t), \\ \Delta(t^*) = 0, \quad \xi(t) \geq 0, \quad t \in T, \end{aligned} \quad (20)$$

is said to be dual to problem (1)–(4). It can be easily verified that the function $\xi(t)$, $t \in T$, (5) corresponding to the regular support Q satisfies the constraints of problem (20).

Let the initial support Q be empty: $M = 0, T_\emptyset^0 = \emptyset$. The functions $\xi(t)$, $\Delta(t)$, $t \in T$, (5), (6) corresponding to Q have the form

$$\xi(t) = 0, \quad \Delta(t) = \int_{t^*}^t \psi'(\tau)b d\tau, \quad t \in T.$$

Let us construct pseudocontrol $\tilde{u}(t)$, $t \in T$, on the support Q in such a way that $\beta(\tilde{u}(\cdot), Q) = 0$:

$$\tilde{u}(0) = u_0, \quad \tilde{u}(t) = \tilde{u}(0) - \int_0^t \text{sign } \Delta(\tau) d\tau. \quad (21)$$

If $\tilde{u}(t) \leq \alpha, t \in T$ then $u^0(t) = \tilde{u}(t), t \in T$, is optimal control in problem (1)–(4). Otherwise we find $t_0 = \min \{t: \max \tilde{u}(t), t \in T\}$ and carry out the procedure of ascent described below using the initial data $Q = \emptyset; \Delta(t), t \in T; t_0; DB(0) = \alpha - \tilde{u}(t_0)$.

The procedure of ascent consists in the following. To begin the work of the procedure of the ascent the initial information is given:

the support $Q = \{[\tau_i, \tau^i], i = \overline{1, M}\}$ and function $\Delta(t), t \in T$, (6) corresponding to it;
the point t_0 ,
the number $DB(0)$.

Denote by $w_1 = 0 + 0, w_2, \dots, w_m = t_0 - 0$ ($w_i < w_{i+1}$) the points in which the function $\psi'(t)b, t \in [0, t_0]$, changes the sign; through $g_i(\Delta) = t$ denote the function reverse to the function $\Delta = \Delta(t), t \in (w_i, w_{i+1}), i = \overline{1, m-1}$. The numbers $\Delta_i = \Delta(w_i), i = \overline{1, m-1}$ satisfying the inequality $\Delta_i < 0$, are put in decreasing order: $0 > \Delta_1 > \Delta_2 > \dots > \Delta_{i_p}$.

Suppose $\Delta_{i_0} = 0, p = 1$ and find $DB(\Delta_{i_p})$ where

$$DB(\Delta) = DB(\Delta_{i_{p-1}}) + 2 \sum_{j \in R_p} |g_j(\Delta_{i_{p-1}}) - g_j(\Delta)| + \\ + |g_{m-1}(\Delta_{i_{p-1}}) - g_{m-1}(\Delta)|; \quad R_p = \{j: \min \{\Delta(w_j), \Delta(w_{j+1})\} < \\ < \Delta_{i_{p-1}} \leq \max \{\Delta(w_j), \Delta(w_{j+1})\}, \quad j = \overline{1, m-2}\}. \quad (22)$$

Calculate $\bar{\Delta}: \bar{\Delta} = \Delta_{i_p}$ if $DB(\Delta_{i_p}) < 0$; $\bar{\Delta}$ equals to the root of the equation $DB(\bar{\Delta}) = 0$ if $DB(\Delta_{i_p}) \geq 0$.

Let the inequality

$$\int_t^{g_{m-1}(\Delta)} \text{sign}(\Delta(\tau) - \Delta) d\tau \leq 0 \quad (23)$$

be true for all $t \in (0, g_{m-1}(\Delta))$ at $\Delta = \bar{\Delta}$.

Consider the cases: a) $DB(\bar{\Delta}) < 0$, b) $DB(\bar{\Delta}) = 0$. In the case of a) change p to $p+1$ and repeat the described operations beginning with the calculation of $DB(\Delta_{i_{p+1}})$ according to formula (22). In a finite number of steps either case (b) is realized or inequality (23) is broken.

In the case of b) we finish the procedure of ascent by constructing the support \bar{Q} which is obtained from Q by the following changes:

$$\text{introduce a new support segment } [g_{m-1}(\bar{\Delta}), t_0]. \quad (24)$$

Consider now the situation when inequality (23) at $\Delta = \bar{\Delta}$ is broken. Let $\hat{\Delta}$ ($\Delta_{i_{p-1}} \geq \hat{\Delta} > \bar{\Delta}$) be a minimal number at which inequality (23) is true; $\hat{t}_0 \in (0, g_{m-1}(\hat{\Delta}))$ is

a minimal moment such that

$$\int_{\hat{t}_0}^{\theta_{m-1}(\hat{A})} \text{sign}(\Delta(\tau) - \hat{A}) d\tau = 0.$$

The support Q is changed into the intermediate support \bar{Q} according to rule (24). Here \hat{A} plays the role of \bar{A} . Once more (with $p=1$) repeat the procedure of ascent proceeding from the initial data

$$\bar{Q}; \bar{A}(t), \quad t \in T; \quad \hat{t}_0; \quad DB(0) = DB(\hat{A}) < 0.$$

In a finite number of operations case (b) is realized and the procedure of ascent is finished by the construction of the regular support \bar{Q} .

Let support \bar{Q} consist of segments $[\tau_i, \tau^i]$, $i = \overline{1, M}$. Construct pseudocontrol $\bar{u}(\cdot)$ according to \bar{Q} in such a way that

$$\begin{aligned} \bar{u}(0) \sim u_0; \quad \bar{u}(t) &= u_0 - \int_0^t \text{sign} \bar{A}(\tau) d\tau, \quad t \in [0, \tau_1]; \\ \bar{u}(t) &= \alpha, \quad t \in [\tau_i, \tau^i]; \quad \bar{u}(t) &= \alpha - \int_{\tau^i}^t \text{sign} \bar{A}(\tau) d\tau, \\ t \in [\tau^i, \tau_{i+1}], \quad i &= \overline{1, M}; \quad \tau_{M+1} &= t^*. \end{aligned} \quad (25)$$

Here $\bar{A}(t)$, $t \in T$, is function (6) constructed in accordance with the support \bar{Q} . It is easily verified that the function $\bar{u}(t)$, $t \in T$, is continuous and $\bar{u}(t) \leq \alpha$, $t \in T$. Hence $\bar{u}(t)$, $t \in T$, is admissible control. By force of the fact that according to the construction $\beta(\bar{u}(\cdot), Q) = 0$ we conclude that $u^0(t) = \bar{u}(t)$, $t \in T$, is optimal control of problem (1)–(4).

Note. The algorithm above has been described for the case when the solution of the problem begins with the special regular support $Q = \emptyset$. A dual algorithm can be described which begins its work with an arbitrary regular support Q (with the initial dual feasible solution $\zeta(t)$, $t \in T$, (5) corresponding to the support Q). In this case the general variant will not differ in principle from that described above.

As the procedure of the ascent consists of a finite number of steps, the described algorithm solves problem (1)–(4) for a finite number of operations.

5. Example

Consider the problem of optimal control of vibrations

$$\begin{aligned} x_2(3\pi) \rightarrow \max; \quad \dot{x}_1 &= x_2, \quad \dot{x}_2 = -x_1 + u, \\ x_1(0) &= x_2(0) = 0, \quad u(t) \leq 2, \quad u(0) = 0, \\ |\dot{u}(t)| &\leq 1, \quad t \in [0, 3\pi]. \end{aligned} \quad (26)$$

In problem (26) function (6) corresponds to the support $Q = \emptyset$:

$$\Delta(t) = -\sin t, \quad t \in [0, 3\pi] \quad \text{and pseudocontrol (21)}$$

$$\tilde{u}(t) = t, \quad t \in [0, \pi); \quad \tilde{u}(t) = -t + 2\pi, \quad t \in [\pi, 2\pi);$$

$$\tilde{u}(t) = t - 2\pi, \quad t \in [2\pi, 3\pi].$$

Calculate $\min \{t \in T: \tilde{u}(t) = \max_{\tau \in T} \tilde{u}(\tau)\} = \pi, u(\pi) = \pi$ and carry out the procedure of ascent using the initial information (Fig. 1):

$$Q = \emptyset; \quad \Delta(t) = -\sin t, \quad t \in [0, 3\pi]; \quad t_0 = \pi, \quad DB(0) = 2 - \pi.$$

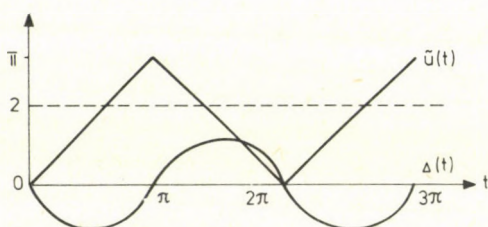


Fig. 1

On the segment $[0, \pi]$ the function $\psi'(t)b = -\cos t$ changes the sign once in the point $\pi/2$. Hence

$$w_1 = 0, \quad w_2 = \pi/2, \quad w_3 = \pi, \quad t = g_1(\Delta) = -\arcsin \Delta,$$

$$t \in [0, \pi/2); \quad t = g_2(\Delta) = -\arcsin \Delta, \quad t \in (\frac{\pi}{2}, \pi].$$

Let us put in decreasing order the numbers $\Delta_1 = 0, \Delta_2 = -1, \Delta_3 = 0$ satisfying the inequality $\Delta_i < 0: \Delta_{i_1} = \Delta_2 = -1$. Let $\Delta_{i_0} = 0, p = 1$ and find $DB(\Delta_{i_1})$ in conformity with (22):

$$DB(\Delta_{i_1}) = DB(0) + 2|g_1(0) - g_1(-1)| + |g_2(0) - g_2(-1)| = 2 + \pi/2 > 0.$$

Calculate $\bar{\Delta} = -\sin(\pi - 2)/3$ as the solution of the equation $DB(\bar{\Delta}) = 0$. Conditions (23) for $\bar{\Delta} = -\sin(\pi - 2)/3$ are fulfilled. Hence the case b) has been realized. We finish the procedure of the ascent on the support $\bar{Q} = \{[\pi - (\pi - 2)/3, \pi]\}$, to which function (5) $\bar{\Delta}(t), t \in T$, corresponds (Fig. 2). In accordance with the algorithm, optimal control

$u^0(\cdot) = \bar{u}(\cdot)$ is constructed in conformity with formulae (25) and has the form (Fig. 2):

$$\begin{aligned} u^0(t) &= -t, & t \in [0, v]; & \quad u^0(t) = t - 2v, & t \in [v, \pi - v]; \\ u^0(t) &= 2, & t \in [\pi - v, \pi]; & \quad u^0(t) = -t + 2 + \pi, & t \in [\pi, 2\pi]; \\ u^0(t) &= t - 3\pi + 2, & t \in [2\pi, 3\pi], & \quad v = (\pi - 2)/3. \end{aligned}$$

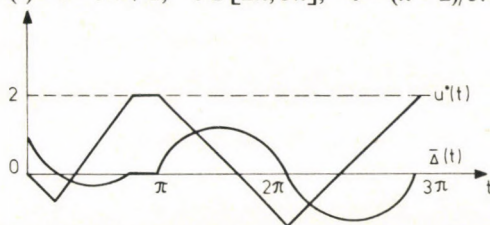


Fig. 2

Supplement

Proof of theorem 1. Sufficiency. If for the support control $\{u(\cdot), Q\}$ extremum conditions (16), (17) are fulfilled then $\beta(u(\cdot), Q) = 0$ and from (13) follows optimality of the control $u(\cdot)$.

Necessity. Let $u(\cdot)$ be optimal control. Let's show that for the support control $\{u(\cdot), Q_u\}$ conditions (16), (17) are true. As Q_u is an accompanying support, condition (16) for $\{u(\cdot), Q_u\}$ can be broken only in the case if there can be found such index $s \in \{1, \dots, M\}$ and number $\bar{t} \in (\tau_s, \tau^s)$, that $\xi(\bar{t}) < 0$.

Suppose that for $\{u(\cdot), Q_u\}$ condition (16) are not fulfilled. From the continuity of the function $\xi(t)$, $t \in (\tau_s, \tau^s)$, follows the existence of such section $T_0 = [t_0, t^0] \subset (\tau_s, \tau^s)$, $t_0 < t^0$, that $\xi(t) < 0$, $t \in T_0$. Construct the variation of control

$$\begin{aligned} \omega(t) &= 0, & t \in T \setminus T_0, & \quad \omega(t) = -t + t_0, & t \in [t_0, (t^0 + t_0)/2]; \\ \omega(t) &= t - t^0, & t \in [(t^0 + t_0)/2, t^0]. \end{aligned}$$

As $\omega(t) \leq 0$, $t \in T$; $\omega(t) = -1$, $t \in [t_0, (t^0 + t_0)/2]$, $\dot{\omega}(t) = 1$, $t \in [(t^0 + t_0)/2, t^0]$, then the control $\bar{u}(\cdot) = u(\cdot) + \omega(\cdot)$ will be admissible, moreover

$$\begin{aligned} J(\bar{u}) - J(u) &= \int_{T_0} (-\Delta(\Delta)\dot{\omega}(t) + \xi(t)\omega(t))dt = \\ &= \text{const} \left[\int_{t_0}^{(t^0+t_0)/2} dt - \int_{(t^0+t_0)/2}^{t^0} dt \right] + \int_{T_0} \xi(t)\omega(t)dt = \int_{T_0} \xi(t)\omega(t)dt > 0. \end{aligned}$$

The inequality obtained contradicts to optimality of the control $u(\cdot)$. Hence for the support control $\{u(\cdot), Q_u\}$ conditions (16) are fulfilled.

Let us pass to the proof of conditions (17). Note that according to the construction the points $\tau_i, \tau^i, i=1, \overline{M}$, are the points of local maxima of the function $u(t), t \in T$. Hence there exists such $\delta > 0$, that $\dot{u}(t) \geq 0, t \in [\tau_i - \delta, \tau_i]$;

$$\dot{u}(t) \leq 0, \quad t \in [\tau^i, \tau^i + \delta]. \quad (\text{S.1})$$

Let for $t \in [\tau_s, t^*], s \in \{1, \dots, M+1\}$ condition (17) be fulfilled and $\Delta(\tau_s) = 0$ (it can always be achieved supposing $s = M+1, \tau_{M+1} = t^*$). Let us show that condition (17) is true for

$$t \in [\tau_{s-1}, t^*] = [\tau_{s-1}, \tau^{s-1}] \cup (\tau^{s-1}, \tau_s) \cup [\tau_s, t^*].$$

Consider the interval (τ^{s-1}, τ_s) . Suppose that on it condition (17) is broken, i.e. there exists such $\bar{t} \in (\tau^{s-1}, \tau_s)$, that either a) $\Delta(\bar{t}) > 0, \dot{u}(\bar{t}) > -1$ or b) $\Delta(\bar{t}) < 0, \dot{u}(\bar{t}) < 1$.

Consider case a). From continuity of the function $\Delta(t), t \in T$, and semi-continuity from the left part of the function $\dot{u}(t), t \in T$, follows the existence of such section $T_0 = [t_0, t^0] \subset (\tau^{s-1}, \tau_s), t_0 < t^0$ that $\Delta(t) > \rho, \dot{u}(t) > -1 + \mu, t \in T_0$, where $\rho > 0, \mu > 0$ are some constants. Let exist support sections of non-zero length (i.e. $\tau_i < \tau^i$) on the right of T_0 and $T_k = [\tau_k, \tau^k]$ is the nearest of them. Find the number η in the following way: let $\eta = \tau^k$ if $\int_{\tau_k}^{\tau^k} \psi'(t) b dt \leq \rho/2$; otherwise η equals the root of equation $\int_{\tau_k}^{\eta} \psi'(t) b dt = \rho/2$. Let us substitute the function $\xi(t), t \in T$, for

$$\bar{\xi}(t), \quad t \in T: \quad \bar{\xi}(t) = \xi(t), \quad t \in T \setminus (\tau_k, \eta), \quad \bar{\xi}(t) = 0, \quad t \in (\tau_k, \eta),$$

and construct the function $\bar{\Delta}(t)$ (6) corresponding to $\bar{\xi}(t), t \in T$:

$$\bar{\Delta}(t) = \Delta(t), \quad t \in [\eta, t^*]; \quad \bar{\Delta}(t) = - \int_t^{\eta} \psi'(\tau) b d\tau < 0, \quad t \in (\tau_k, \eta);$$

$$\bar{\Delta}(t) = \Delta(t) - \int_{\tau_k}^{\eta} \psi'(\tau) b d\tau \geq \Delta(t) - \rho/2, \quad t \in [0, \tau_k].$$

Construct the variation of control:

$$\omega(t) = 0, \quad t \in [0, t_0] \cup [\eta, t^*]; \quad \omega(t) = -t + t_0, \quad t \in (t_0, t_0 + v),$$

$$\omega(t) = -v, \quad t \in (t_0 + v, \eta - v); \quad \omega(t) = t - \eta, \quad t \in [\eta - v, \eta),$$

where $v = \min \{t^0 - t_0, \eta - \tau_k\}$ and calculate the step

$$\vartheta = \min \{1, \mu\} > 0.$$

It can be easily verified that control $\bar{u}(\cdot) = u(\cdot) + \vartheta \omega(\cdot)$ will be admissible and by force of (12) $\bar{J}(\bar{u}) - J(u) = \vartheta \left(\int_{t_0}^{t_0+v} \bar{\Delta}(t) dt - \int_{\eta-v}^{\eta} \bar{\Delta}(t) dt \right) > 0$. The last inequality contradicts to optimality of control $u(\cdot)$.

Consider now the situation when there are no support sections of non-zero length on the right of T_0 . In this case construct the variation:

$$\begin{aligned} \omega(t) &= 0, \quad t \in [0, t_0]; \quad \omega(t) = -t + t_0 < 0, \\ t \in (t_0, t^0); \quad \omega(t) &= t_0 - t^0 < 0, \quad t \in [t^0, t^*], \end{aligned}$$

and calculate the step $\Theta = \min \{1, \Theta_1\}$, $\Theta_1 = \min (1 + \dot{u}(t))$, $t \in T_0$. Control $\bar{u}(\cdot) = u(\cdot) + \Theta \omega(\cdot)$ will be admissible and $J(\bar{u}) - J(u) = \Theta \int_{T_0} \Delta(t) dt > 0$ which contradicts to optimality of $u(\cdot)$. Consider case b). From the continuity of the function $\Delta(t)$, $t \in T$, semi-continuity on the left of the function $\dot{u}(t)$, $t \in T$, conditions (S.1) and the fact that $\Delta(\tau_s) = 0$, $\Delta(\bar{t}) < 0$, $\dot{u}(\bar{t}) < 1$, follows the existence of such segments $T_0 = [t_0, t^0]$, $[\eta, \tau_s]$ ($t_0 < t^0$, $\eta < \tau_s$) that

$$\begin{aligned} \min_{t \in [\eta, \tau_s]} \Delta(t) > \max_{t \in T_0} \Delta(t); \quad \dot{u}(t) < 1 - \mu, \quad t \in T_0; \\ \dot{u}(t) \geq 0, \quad t \in [\eta, \tau_s], \end{aligned} \quad (\text{S.2})$$

where $\mu = \text{const} > 0$. Construct the variation

$$\begin{aligned} \omega(t) &= 0, \quad t \in [0, t_0] \cup [\tau_s, t^*]; \quad \omega(t) = \alpha - u(t), \quad t \in [\tau_s - v, \tau_s]; \\ \omega(t_0 + \tau) &= \omega(\tau_s - \tau) = \alpha - u(\tau_s - \tau), \quad \tau \in [0, v]; \quad \omega(t) = \alpha - u(\tau_s - v), \\ t &\in [t_0 + v, \tau_s - v] \end{aligned}$$

where $v = \min \{t^0 - t_0, \tau_s - \eta\}$, and calculate the step

$$\begin{aligned} \bar{\Theta} &= \min \{1, \Theta_1, \Theta_2\}, \quad \Theta_1 = \min (1 - \dot{u}(t_0 + \tau)) / \dot{u}(\tau_s - \tau), \quad \tau \in [0, v]; \\ \Theta_2 &= \min (\alpha - u(t)) / \omega(t), \quad t \in [t_0, \tau_s - v]. \end{aligned}$$

As $u(t) < \alpha - \mu_1$, $t \in [t_0, \tau_s - v]$, $\mu_1 > 0$; $\dot{u}(t) < 1 - \mu$, $t \in T_0$, $\mu > 0$ then $\bar{\Theta} > 0$. The control $\bar{u}(\cdot) = u(\cdot) + \bar{\Theta} \omega(\cdot)$ is admissible and by force of (S.2)

$$J(\bar{u}) - J(u) > \bar{\Theta} (u(\tau_s) - u(\tau_s - v)) \left[\min_{t \in [t_0, t_0 + v]} \Delta(t) - \max_{t \in [\tau_s - v, \tau_s]} \Delta(t) \right] > 0.$$

The inequality received contradicts to optimality of control $u(\cdot)$.

The contradictions obtained in cases a), b) prove the statement that for moments $t \in (\tau^{s-1}, \tau_s)$ conditions (17) are fulfilled. Using the proved statement, relations (S.1) and continuity $\Delta(t)$, $t \in T$, we get

$$\Delta(t) \geq 0, \quad t \in [\tau^{s-1}, \tau^{s-1} + \delta].$$

Let us show that $\Delta(\tau^{s-1})=0$. Suppose the reverse, i.e. $\Delta(\tau^{s-1})>0$. Substitute the function $\xi(t)$, $t \in T$, for the function $\bar{\xi}(t)=\xi(t)$, $t \in T(\tau_{s-1}, \tau^{s-1})$; $\bar{\xi}(t)=0$, $t \in (\tau_{s-1}, \tau^{s-1})$. For the function $\bar{\Delta}(t)$, $t \in T$, (6) corresponding to $\bar{\xi}(t)$, $t \in T$, the equality

$$\bar{\Delta}(t)=\Delta(t), \quad t \in [\tau^{s-1}, t^*] \quad (\text{S.3})$$

is true.

From (S.1, S.3) and continuity of $\bar{\Delta}(t)$, $t \in T$, follows the existence of such $v>0$ that $\bar{\Delta}(t)>0$, $\dot{u}(t) \geq 0$, $t \in [\tau^{s-1}-v, \tau^{s-1}]$. Reasoning in the way analogous to case a), we get contradiction with optimality of control $u(\cdot)$. So $\Delta(\tau^{s-1})=0$. As $\Delta(t)=\text{const}$, $t \in [\tau_{s-1}, \tau^{s-1}]$, then $\Delta(t) \equiv 0$, $t \in [\tau_{s-1}, \tau^{s-1}]$. Hence, conditions of extremum (17) are also fulfilled for $t \in [\tau_{s-1}, \tau^{s-1}]$.

So it has been proved that condition (17) is fulfilled for all $t \in [\tau_{s-1}, t^*]$. Diminishing s and repeating the given reasoning we prove the correctness of condition (17) on all the on $T=[0, t^*]$. The theorem is proved.

Proof of theorem 2. Necessity. If support Q is such that $\xi(t) \geq 0$, $t \in T$, then suboptimality estimate $\beta(u(\cdot), Q)$ (14) of support control $(u(\cdot), Q)$ can be written in the form

$$\begin{aligned} \beta(u(\cdot), Q) &= \int_{T_+} \Delta(t)(\dot{u}(t)+1)dt + \int_{T_-} \Delta(t)(\dot{u}(t)-1)dt + \\ &+ \int_T \xi(t)(\alpha - u(t))dt = -J(u) + I(Q), \end{aligned} \quad (\text{S.4})$$

where

$$\begin{aligned} I(Q) &= c' f(t^*)u_0 + \int_T |\Delta(t)|dt + (\alpha - u_0) \int_T \xi(t)dt, \\ T_+ &= \{t \in T: \Delta(t) > 0\}, \quad T_- = \{t \in T: \Delta(t) < 0\}. \end{aligned}$$

In the last expression the first item $J(u)$ depends only on control, the second one $I(Q)$ depends only on support Q .

In problems (1)–(4) there exists optimal control $u(\cdot)$. It follows from theorem 1 that such support Q^0 can be found that for $\{u^0(\cdot), Q^0\}$ the conditions of extremum (16), (17) are fulfilled from which it follows that $\xi^0(t) \geq 0$, $t \in T$, and $\beta(u^0(\cdot), Q^0)=0$. Using (S.4) we get

$$J(u^0) = I(Q^0). \quad (\text{S.5})$$

Let $u(\cdot)$ be ε -optimal control, i.e. $J(u^0) - J(u) \leq \varepsilon$. Let us assign support Q^0 to the control $u(\cdot)$. Then in accordance with (S.4), (S.5) for support control $\{u(\cdot), Q^0\}$ we have

$$\beta(u(\cdot), Q^0) = -J(u) + I(Q^0) = -J(u) + J(u^0) \leq \varepsilon.$$

Sufficiency follows from (19) and inequality (13).

References

1. Pontryagin L. S., Boltyansky V. G., Gamkrelidze R. V., Mishchenko E. F., *Mathematical theory of optimal processes*. M., Nauka, 1976.
2. Fedorenko R. P., *Approximate solution of optimal control problems*. M., Nauka, 1978.
3. Gabasov R., Kirillova F. M., Kostyukova O. I. A method of solution of general linear programming problem. — *Doklady AN BSSR*, 1979, vol. 3, No. 3, pp. 197–200.
4. Gabasov R., Kirillova F. M., *Linear programming methods. I–III*, Minsk, BGU publishing House, 1977, 1978, 1980.
5. Gabasov R., Kirillova F. M., *Constructive methods of Parametric and Functional Optimization*. Preprints of VIIIth IFAC Congress, Kyoto, 1981.
6. Gabasov R., Kirillova F. M., Kostyukova O. I., *Adaptive method of solving linear programming problems*. Prepr. IFAC Symposium on Optimization Methods (Varna, Bulgaria, 1979).
7. Gabasov R., Kirillova F. M., *Constructive methods of solving extremal problems*. Prepr. 3rd Polish–English Seminar on realtime processes, Poland, 1980.
8. Gabasov R., Kirillova F. M., *New Linear Programming Methods and there Application to Optimal Control Problems*. Proc. I Workshop on Control Applications of Nonlinear Programming, Denver, USA, 1979. Pergamon Press, 1980.

Двойственный алгоритм оптимизации линейной динамической системы

Р. ГАБАСОВ, Ф. М. КИРИЛЛОВА, О. И. КОСТЮКОВА

(Минск)

Предлагается новый алгоритм решения линейной задачи оптимального управления в классе ограниченных непрерывных функций с ограниченными кусочно-непрерывными производными. Алгоритм является двойственным, т. е. оптимальное управление строится путем специального преобразования информации о двойственной задаче. Начальная информация предполагается известной или просто строится по элементам исходной задачи. В основе алгоритма лежит метод, ранее разработанный авторами для решения задач линейного программирования (Габасов Р., Кириллова Ф. М. *Методы линейного программирования*. ч. II–III, Изд.-во БГУ, Минск, 1977, 1978, 1980. Широкий численный эксперимент показал существенные преимущества метода перед известными методами линейного программирования. Метод был использован для построения алгоритмов решения специальных задач нелинейного программирования (кусочно-линейное, квадратичное, геометрическое программирование), дискретных задач оптимального управления (Gabasov R., Kirillova F. M., *Constructive methods of Parametric and Functional Optimization*. Preprints of VIII IFAC Congress, vol. 4, Kyoto, 1981). Особенность излагаемого алгоритма состоит в использовании специального элемента, названного опорой. Опора позволяет проверить текущую информацию на оптимальность или субоптимальность и в случае необходимости преобразовать ее на лучшую. Попутно доказаны критерии оптимальности в виде принципов максимума и β -максимума, в которых используются не меры (как в большинстве работ по задачам с фазовыми ограничениями), а вполне регулярные функции. Алгоритм конечен и реализуется в виде достаточно простых операций, легко осуществимых на ЭВМ. Работа алгоритма иллюстрируется на примере оптимального управления колебательными движениями. Доказательство теоретических результатов основано на новых формулах приращения.

Ф. М. Кириллова

О. И. Костюкова

Институт математики АН БССР

СССР Минск, Сурганова, 11

A NEW CONCEPTION OF DIGITAL ADAPTIVE PSD CONTROL

J. MARŠÍK

(Prague)

(Received May 5, 1982)

A simple heuristic approach to adaptive control without any need of plant identification, test signals or extremum seeking operations is presented. Having introduced a new performance index the adaptive loops can be treated as ordinary control loops. This index has been defined as a ratio of zero crossing frequencies of the control error and its difference. For any stable process its value always lies within the interval (0, 1), with the optimum being in the middle, if certain conditions are fulfilled. The common controller gain is adjusted so as to maintain the index at a constant value equal to one half. The parameters of the other controller terms are adapted so that each of them may supply a signal with the same effective level into the resulting controller output. Because of its simplicity the algorithm is suitable for microprocessor implementation.

1. Basic idea of the approach

In a previous paper [1], the practical criterion of "maximum stability margin" was formulated which enables us to adjust the values of all controller parameters into the middle of their stability range. For that purpose the stability boundaries of each parameter had to be found by means of generating limit cycle oscillations with an appropriate amplitude (using a relay characteristic), which is often inconvenient or even inadmissible.

To avoid this drawback, we succeeded in finding a new performance index with similar properties which need not use this limit cycle excitation. The basic idea is very simple:

Observing the zero crossing frequencies of the control error f_e and of its first difference f_v , we can find the dependence of their ratio $\kappa = \frac{f_e}{f_v}$, e.g., on the controller gain α . Meanwhile, only the common gain of a PSD controller is considered. Adaptation of the other parameters will be discussed later.

It can be shown that this ratio is very roughly proportional to the gain, reaching its upper limit value equal to one if the system becomes unstable. Thus the possible range of this ratio reaches from zero to one; small values indicate an overdamped control process, high values testify to an oscillating one. As well as in the case of the

previous performance criterion [1], the first intuitive suggestion was to choose just the middle as the probable optimum, i.e. $\kappa = 0.5$.

This has proved to be true, however, some conditions must be met which will be defined in the next chapter. Nevertheless, even if these conditions are not fulfilled, the optimum value can be found by trial as well.

2. Assumptions for a correct design of the adaptation loop

The analyzed control error must correspond to the true control loop response, otherwise it must be deprived of all parasitic signals to which the system cannot react.

For this aim a proper choice of the sampling period is often sufficient, further smoothing being seldom necessary (about 5 samples for a half-wave of the response are recommended). Otherwise the moving averages have proved to be the best.

Having met these conditions, we may hope that the optimal value of the above mentioned ratio of zero-crossing frequencies will be about one half. This is rather a "rule of thumb", because there are some exceptions, such as dead-beat systems, whose corresponding optimal performance index may have different values.

The controlled plant must be stable and not too oscillatory unless the controller would correct this property; further, permanently oscillating low frequency disturbances must be avoided. Since the algorithms cannot discern external oscillations from those caused by the high gain, such oscillations entail gain lowering.

On the contrary, the high frequency noise, if not suppressed, caused increasing the zero-crossing frequency in the control error difference, however, not so much in the control error itself. Thus the ratio of zero-crossings decreases and this results in gain raising. That is the reason why the signal must be smoothed.

These conditions may be sufficient for regulation. In addition, when tracking is also to be considered, step changes of the reference signal are not advisable because of evoking undesirable peak overshoots of the controlled variable. In such cases the reference signal should be prefiltered by a first order filter analogous to that described in Section 4, Eqs. (4.8)–(4.11). Having respected these recommendations, we can use the proposed approach in all cases where a standard PSD controller is applicable. For this purpose only the control error must be accessible and no further process quantities need to be treated.

3. Formulation of two basic algorithms for adaptation of the gain α

Let us denote:

$M \dots$	length of observation interval given as a number of samples
$e \dots$	the control error
$v = \Delta e \dots$	the difference of the control error ("velocity")
$f_e \dots$	the zero-crossing frequency of e
$f_v \dots$	the zero-crossing frequency of Δe
$N_e \dots$	the number of zero-crossings of e at the M -th sampling instant
$N_v \dots$	the number of zero-crossings of Δe at the M -th sampling instant
$\kappa = \frac{f_e}{f_v} \dots$	the ratio of the zero-crossing frequencies
$\alpha \dots$	the controller gain
$\kappa_{\text{ref}} \dots$	the reference (desired) value of κ
$\lambda \dots$	the coefficient of adaptation speed

As mentioned in the introduction, the dimensionless ratio

$$\kappa(\alpha) = \frac{f_e}{f_v} \approx \frac{N_e}{N_v} \quad (3.1)$$

is generally a nonlinear function of the controller gain. However, it is always valid that

$$0 \leq \kappa \leq 1. \quad (3.2)$$

For our purpose it would be desirable if κ were a linear function of α within this interval, because it would enable us to adjust α after one step by means of the following simple algorithm:

$$\alpha_{kM} = \alpha_{(k-1)M} \frac{\kappa_{\text{ref}}}{\kappa_{(k-1)M}} \quad (3.3)$$

Since if $\kappa = \frac{\alpha}{\alpha_{\text{crit}}}$, then $\alpha_{kM} = \kappa_{\text{ref}} \alpha_{\text{crit}}$, where α_{crit} denotes the critical value of α .

Nevertheless, the algorithm (3.3) can be applied iteratively. It uses the stationary values of κ measured during M sampling intervals, M being sufficiently large to ensure the settling of the whole control process including the random disturbances.

When a value of α has been adjusted according to algorithm (3.3), new values of κ are computed using new counted values N_e and N_v , and all the values from the preceding step are replaced by the new ones.

In this way the influence of the dynamic properties of the control loop on the adaptation has been eliminated. Note that algorithm (3.3) resembles a manual control,

where the operator, having evaluated the effect of the previous correction, makes a new one.

Thus algorithm (3.3) can be also used for checking the control process and adjusting α manually, if necessary. (Sometimes it may happen that the iterations cannot converge. According to our experience, the divergence is mainly caused by violating the recommendations of Chapter 2.)

Besides this well comprehensive algorithm there are some other ones [2] with continual adaptation which, however, must reflect the entire system dynamics. It means, roughly speaking, that fast systems may adapt quickly, while the slow ones must adapt slowly. For this purpose it is necessary to know certain characteristic frequency of the system to which the adaptation speed is made proportional. It has shown that the frequency f_v is just what we need.

Let us try to formulate an adaptation algorithm which will meet the above stated demands, e.g. the following one:

$$\Delta\alpha = \alpha\lambda f_v(\kappa_{\text{ref}} - \kappa). \quad (3.4)$$

It can be rewritten as

$$\Delta\alpha = \alpha\lambda(f_v\kappa_{\text{ref}} - f_e). \quad (3.5)$$

The adaptation speed is proportional to that of the control process given by f_v , to the adaptation error ($\kappa_{\text{ref}} - \kappa$) and to the gain adapted so as to ensure the same relative speed for each value of α .

As a matter of fact, however, since κ depends on the critical value of α , the adaptation speed is not the same for each α . If $\alpha < \alpha_{\text{crit}}$, we get an idle adaptation of α . This can be seen, for example, if we assume again that

$$\kappa = \frac{\alpha}{\alpha_{\text{crit}}} \quad (3.6)$$

and substitute this into Eq. (3.4).

We obtain

$$\Delta\alpha = \lambda f_v \frac{\alpha}{\alpha_{\text{crit}}} (\kappa_{\text{ref}} \alpha_{\text{crit}} - \alpha), \quad (3.7)$$

i.e.

$$\Delta\alpha = \lambda f_v \kappa (\kappa_{\text{ref}} \alpha_{\text{crit}} - \alpha). \quad (3.8)$$

Here the dependence on κ is clearly undesirable. The remedy for this drawback is evident: we shall divide the original algorithm (3.4) or (3.5) by κ so that

$$\Delta\alpha = \alpha \frac{\lambda f_v}{\kappa} (\kappa_{\text{ref}} - \kappa) \quad (3.9)$$

or

$$\Delta\alpha = \alpha\lambda \frac{f_v}{f_e} (\kappa_{\text{ref}} \cdot f_v - f_e). \quad (3.10)$$

Having formulated algorithms (3.3) or (3.9), we are obliged to solve the problem of acquiring the frequencies f_e and f_v . Direct measuring is possible but not too advisable due to providing only a scarce information about the system behaviour. We must wait for several sampling intervals before being able to compute the frequencies f_e and f_v and the ratio κ . Between two zero-crossings of e practically one half of the transient response time passes away, having to be regarded as a pure transportation lag.

As follows from the essence of the approach, a single step disturbance is hardly sufficient to bring about a noticeable adaptation. Only frequent random disturbances are able to evoke the adaptive action.

Fortunately, these difficulties can be overcome, if the frequencies f_e, f_v are not obtained by counting the actual zero crossings but, indirectly, by computing from the control error and its first and second differences, as will be shown in the next chapter. This indirect calculation gives useful data at every sampling instant so that the undesirable transportation lag is avoided. This results in a faster adaptation. Notwithstanding certain inexactness of this approach it leads to a considerable improvement, though it has some non-minimum-phase properties as well.

4. Indirect computing of the zero-crossing frequencies

About forty years ago, Rice [3] derived a formula for zero crossings of a continuous Gaussian random signal with zero mean

$$N_e = \frac{T}{\pi} \sqrt{\frac{R_{vv}(0)}{R_{ee}(0)}} = \frac{T}{\pi} \sqrt{\frac{\sigma_v^2}{\sigma_e^2}} \quad (4.1)$$

where the following notation is used:

$R_{ee}(0), R_{vv}(0) \dots$	the values of autocorrelation functions of the signal e and of its derivative
$T \dots$	observation time
$N_e \dots$	number of zero crossings of e
$\sigma_e^2, \sigma_v^2 \dots$	dispersions of e and of its derivative.

The formulas are applicable, if $R_{vv}(0)$ is finite, hence the white noise cannot be examined immediately and must be filtered, at least, by a filter of relative second order (the difference between the denominator and numerator orders must be equal to, or greater than, two).

For the zero-crossing frequency we obtain

$$f_e = \frac{N_v}{T} = \frac{1}{\pi} \sqrt{\frac{R_{vv}(0)}{R_{ee}(0)}} = \frac{1}{\pi} \sqrt{\frac{\sigma_v^2}{\sigma_e^2}}. \quad (4.2)$$

In an analogous manner, we get for v

$$N_v = \frac{T}{\pi} \sqrt{\frac{R_{aa}(0)}{R_{vv}(0)}} = \frac{T}{\pi} \sqrt{\frac{\sigma_a^2}{\sigma_v^2}} \quad (4.3)$$

$$f_v = \frac{1}{\pi} \sqrt{\frac{R_{aa}(0)}{R_{vv}(0)}} = \frac{1}{\pi} \sqrt{\frac{\sigma_a^2}{\sigma_v^2}}. \quad (4.4)$$

The filter which the white noise must pass through is then of at least third order.

It can easily be proved that formulas (4.3) and (4.4) are also valid for any sine wave $A \sin \omega t$, because of $R_{vv}(0) = \omega^2 R_{ee}(0)$. Then

$$N_e = \frac{T}{\pi} \omega = \frac{T}{\pi} \cdot 2\pi f = 2fT. \quad (4.5)$$

Thus the frequency of zero crossings corresponds to the number of half-waves.

For sampled signals analogous formulas may be applied. Instead of derivatives the differences and instead of integrals the sums are used. Since even higher differences of the sampled white noise always exist, the formulas are applicable without the above stated restrictions, although their accuracy is the lower, the less smoothed the signal is.

For algorithm (3.3) working with steady-state values κ we can adapt formulas (4.3) and (4.4) as follows:

$$\kappa = \frac{f_e}{f_v} = \frac{\frac{1}{\pi} \sqrt{\frac{\sum_{i=0}^M (\Delta e_i)^2}{\sum_{i=0}^M e_i^2}}}{\frac{1}{\pi} \sqrt{\frac{\sum_{i=0}^M (\Delta^2 e_i)^2}{\sum_{i=0}^M (\Delta e_i)^2}}} = \frac{\sum_{i=0}^M (\Delta e_i)^2}{\sqrt{\sum_{i=0}^M e_i^2 \cdot \sum_{i=0}^M (\Delta^2 e_i)^2}}. \quad (4.6)$$

For algorithm (3.9) or (3.10) with nonstationary κ the moving averages or other type averages with forgetting old values must be used:

$$\kappa = \frac{f_e}{f_v} = \frac{\sqrt{\frac{(\Delta e)^2}{e^2}}}{\sqrt{\frac{(\Delta^2 e)^2}{e^2}}} = \frac{(\Delta e)^2}{\sqrt{(\Delta^2 e)^2 \cdot e^2}} = \frac{\overline{v^2}}{\sqrt{a^2 \cdot e^2}}. \quad (4.7)$$

The simplest average values $\overline{e^2}$, $\overline{(\Delta e)^2}$, $\overline{(\Delta^2 e)^2}$ may be obtained by letting e^2 , $(\Delta e)^2$, $(\Delta^2 e)^2$ pass through a first-order filter. The corresponding equations of this filter have the form

$$\overline{e_n^2} = \frac{e_n^2 + \tau_{n-1} \overline{e_{n-1}^2}}{1 + \tau_{n-1}} \quad (4.8)$$

$$\overline{v_n^2} = \overline{(\Delta e_n)^2} = \frac{(\Delta e)_n^2 + \tau_{n-1} \overline{v_{n-1}^2}}{1 + \tau_{n-1}} \quad (4.9)$$

$$\overline{a_n^2} = \overline{(\Delta^2 e_n)^2} = \frac{(\Delta^2 e)_n^2 + \tau_{n-1} \overline{a_{n-1}^2}}{1 + \tau_{n-1}}. \quad (4.10)$$

To make τ adaptive as well, we shall define it as

$$\tau_n = 2\pi \sqrt{\frac{\overline{v_n^2}}{a_n^2}} = \frac{2}{f_v}. \quad (4.11)$$

Here τ is the double average time between two zero crossings of Δe . When the optimum is reached, τ equals to the average time between zero crossings of e .

Substituting for f_e and f_v into algorithm (3.10) gives

$$\Delta\alpha = \alpha \frac{\lambda}{\pi} \frac{\sqrt{a^2 \cdot e^2}}{v^2} \left(\kappa_{\text{ref}} \sqrt{\frac{a^2}{v^2}} - \sqrt{\frac{v^2}{e^2}} \right) \quad (4.12)$$

with $0.1 < \lambda < 0.3$.

Similarly, after inserting f_v and κ into the other form (3.9) of the algorithm, we get

$$\Delta\alpha = \alpha \frac{\lambda}{\pi} \sqrt{\frac{a^2}{v^2}} \left(\frac{\kappa_{\text{ref}} \sqrt{a^2 \cdot e^2}}{v^2} - 1 \right). \quad (4.13)$$

As pointed out before, the computation of the zero-crossing frequencies has certain nonminimum-phase properties, even if no pure delay exists. It behaves like a system with zeros outside the unit circle in its transfer-function numerator.

In addition, not only the actual zero crossings are calculated in this way, but some fictive ones as well, corresponding rather to approaching instead of crossing zero. This enables the algorithm to adapt even in the case of a single step disturbance. Thus the algorithm (4.12) or (4.13) is universal, its use being not confined only to cases with random disturbances.

5. Adaptation algorithms for the P and D terms of the PSD controller

We shall consider the following controller structure:

$$u_n = \sum_0^n \alpha e_i + \alpha \beta e_n + \alpha \gamma \Delta e_n \quad (5.1)$$

where u_n is the controller output, α is the gain, β is the parameter of the proportional term and γ is that of the difference term.

To proceed further, let us transform Eq. (5.1) into

$$u_n = \sum_{i=0}^n \alpha [e_i + \beta \Delta e_i + \gamma \Delta^2 e_i] \quad (5.2)$$

as if we had a PDD² controller terminated by a summing term. Now, our idea is to make the effective control effort of all components equal, so that

$$\beta \sqrt{(\Delta e)^2} = \gamma \sqrt{(\Delta^2 e)^2} = \sqrt{e^2}. \quad (5.3)$$

Hence

$$\beta = \sqrt{\frac{e^2}{(\Delta e)^2}} = \sqrt{\frac{e^2}{v^2}} \quad (5.4)$$

$$\beta = \sqrt{\frac{e^2}{(\Delta^2 e)^2}} = \sqrt{\frac{e^2}{a^2}}. \quad (5.5)$$

Thus Eq. (5.4) and Eq. (5.5) are the adaptation algorithms for the P and D terms of the controller, α being adapted by the algorithm (4.12) or (4.13). Exact mathematical justification of this approach can hardly be presented, except for the improvement of the phase angle. Making all components equal, the phase improvement amounts to about $\frac{\pi}{2}$ (for a sine wave). This phase amendment results in better control.

It is a matter of fact that the complete algorithm composed of Eqs (4.12), (5.4) and (5.5) determines a controller which is adaptive to any gain variations as well as to time-scale changes in the control loop including also the characteristics of the disturbances.

We do not venture to assert that the controller parameters are actually optimal from the point of view of some known criterion. Nevertheless, the responses obtained have shown to be very good. Otherwise it is not difficult to find correcting coefficients for each term by trial. Instability cannot occur, even if β or γ are made 2–3 times greater, since stability is watched by adapting the common gain α . A significant advantage consists in the fact that no higher differences of e need to be calculated than those being necessary for adapting the gain α . In addition, all variables for determining the parameters β and γ are available from calculating the gain α only.

6. Experimental results

Many examples were examined both on digital and hybrid computers. The simulated plants were nontrivial, of up to fourth order, with non-minimum phase and with a transportation lag, too. Random disturbances, as well as steps or other deterministic signals were used. The digital adjustment was employed not only for the digital PSD controller but also for an analogue PID controller, without any difficulties.

Having applied step disturbances, we got the "optimum" setting after 2–4 steps, even if we started with an initial value of the gain α very far from the optimum. For illustration, a typical control error, caused by a step disturbance and obtained after four steps, is shown in Fig. 1. The disturbance was applied at the input with the transfer function $1/(z-0.7)^2$.

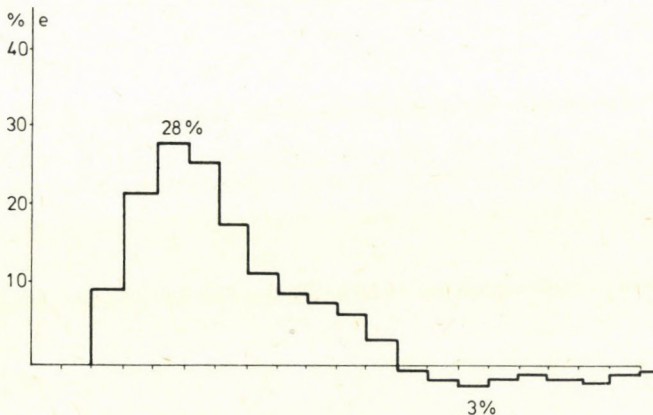


Fig. 1. A typical control error subsequent to a step disturbance at the plant input. The required value of the performance index was $\kappa=0.5$

Regardless of the plant order, all the examples exhibited good results provided that the conditions from Chapter 2 had been met.

It can be said that the shape of such step responses is always very similar, well damped and with two prevalent halfwaves only.

This similarity of closed-loop-control processes is widely known (a higher-order loop behaves like a second-order one due to the predominant pair of complex roots in its characteristic equation).

Having two peaks and one zero crossing, the shape of e corresponds to $\kappa = 0.5$. Since the zero crossings are computed indirectly, using the quadratic values of e , Δe and $\Delta^2 e$, the rapidly decaying tail of e with its further zero crossings is neglected automatically.

If single disturbances with very long pauses occurred so that $e \rightarrow 0$, it should be necessary to switch off the adaption in order to avoid division by zero.

As for initial conditions, it is advisable to take $\bar{e}_0^2 = \bar{v}_0^2 = \bar{a}_0^2 < 0, 1 \sigma_e^2$, where σ_e^2 denotes the variance of e .

7. Conclusion

A new performance index, formulated as a ratio of zero-crossing frequencies of the control error and its difference, enabled us to design a simple adaptive controller. The value of this performance index represents a measure of oscillatory character of the control process, as can easily be demonstrated on simple step responses.

References

1. *Marsík J., Černý P., Bláha S.*, A Simple Algorithm for Automatic Optimum Setting of Controller Parameters. The 2nd IFAC/IFIP Symposium of Software for Computer Control "SOCOCO" 79, paper A—XII, Czechoslovak Academy of Sciences, Prague (1979).
2. *Marsík J.*, Simple Algorithms for Digital Adaptive Control (in Czech). Research Report No. 1106 of the Institute of Information Theory and Automation, Prague (1981).
3. *Rice S. O.*, Mathematical Analysis of Random Noise. Bell Techn. Journal vol. 23, pp. 282–332 (1944); vol. 24, pp. 46–150 (1945).

Новая концепция цифрового адаптивного регулирования ПСД действия

Я. МАРШИК

(Прага)

Предлагается простой алгоритм адаптивного регулирования и управления, который не требует ни идентификации объекта регулирования, ни пробных воздействий, ни поиска экстремумов.

Благодаря введению нового косвенного показателя качества, адаптивные петли могут работать в режиме обычного регулирования. Показатель качества определяется как отношение частот переходов через нуль ошибки регулирования и ее разности. Для устойчивых процессов значение показателя лежит всегда внутри интервала $(0, 1)$, и оптимуму соответствует одна половина, если выполнены некоторые условия.

Общее усиление регулятора настраивается так, чтобы поддерживать значение показателя на одной половине. Параметры остальных каналов регулируются так, чтобы каждый из них на своем выходе отдавал сигнал с тем же самым эффективным уровнем, как и другие. Отличаясь значительной простотой, алгоритм пригоден для эксплуатации микропроцессоров.

J. Maršík

Institute of Information Theory and Automation

Czechoslovak Academy of Sciences

182 08 Praha 8

Pod vodárenskou věží 4

Czechoslovakia

SIMULTANEOUS OPTIMIZATION OF MAINTENANCE AND REPLACEMENT POLICY FOR MACHINES

I. MUNTEAN

(Cluj-Napoca)

(Received June 25, 1982)

Optimal control methods are used for simultaneous optimization of maintenance policy and sale date of a machine. The customary assumption that the maintenance investments are not so effective as to enhance the value of the machine over previous values, is not necessarily required in this paper. For linear and nonlinear variants of the associated optimal control problem we derive explicit expressions for optimal strategies, and analyse some numerical examples.

1. Introduction

A machine is to be bought, used for productive purposes for some period, and then replaced. The machine is supposed to generate a constant rate of revenue over the capital invested in it. However, due to the physical depreciation and obsolescence phenomenon, the revenue rate decreases forcing the management to initiate a maintenance action in order to slow down the degradation of the machine's capability. The maintenance action here means money spent over and above the minimum spent on current repairs, and involves preventive maintenance, reliability equipment, partial modernization, ergonomic improvements etc. The problem is to choose a maintenance schedule and replacement date so as to maximize the net revenue associated with the operation of the machine.

Using the optimal control methods, B. Naslund [6], G. L. Thompson [11] and J.-Y. Helmer [2], pp. 127–138, solved this problem under the assumption that maintenance costs are not so effective as to enhance the value of the machine over previous values. However, for a modification of the Thompson's linear model S. R. Arora and P. T. Lele [1], and D. N. Khandelwall, J. Sharma and L. M. Ray [5] have presented situations in which the salvage value of the machine may be increasing everywhere, or returning periodically to the purchased price, respectively. In the present paper we study the optimal strategies of machine maintenance and replacement without requiring the above restrictive assumption. The equations for determining the switching instant and the replacement date are established in [11] and [2], pp. 130–137, by a less convincing argument of successive optimization. In what follows

we reobtain the same equations from the transversality conditions for an optimal control problem associated with the *simultaneous* optimization of maintenance and replacement date. For linear and nonlinear variants of the last problem we derive explicit expression for optimal strategies, and analyse some numerical examples.

Other machine replacement models in deterministic or stochastic setting are discussed in the papers of M. I. Kamien and N. L. Schwartz [4], S. P. Sethi [8], C. S. Tapiero and I. Venezia [10], S. P. Sethi and S. Chand [9], and D. G. Nguyen and D. N. P. Murthy [7].

2. Mathematical model and necessary conditions of optimality

The machine is purchased at time $t=0$ at price $x_0 > 0$, put into service at the same time, and replaced at time $t_1 > 0$. The residual value $x(t)$ of the machine at time $t \in [0, t_1]$ depends upon the following economic factors:

1° In absence of maintenance action, the value $x(t)$ decreases with a quantity $a(t)\Delta t$ on the time interval $[t, t + \Delta t]$, where $a: [0, +\infty[\rightarrow R$ is the obsolescence function;

2° The assignment of a sum of money $u(t)$ for maintenance action on $[t, t + \Delta t]$ produces an increase of $x(t)$ with a quantity $F(t, u(t))\Delta t$, where $u: [0, t_1] \rightarrow R$ is a nonnegative function and $F: [0, +\infty[\times R \rightarrow R$ is a given function.

Taking into account both economic factors, the variation of the machine's value $\Delta x(t) = x(t + \Delta t) - x(t)$ is given by $\Delta x(t) = -a(t)\Delta t + F(t, u(t))\Delta t$, whence, assuming the differentiability of the function $x: [0, t_1] \rightarrow R$, we obtain

$$\dot{x}(t) = -a(t) + F(t, u(t)).$$

In establishing the net revenue we admit that the earnings $y(t)\Delta t$, due to the operation of the machine on the interval $[t, t + \Delta t]$, are proportional to the residual value $x(t)$ and to the length of this interval, i.e., $y(t)\Delta t = rx(t)\Delta t$, where the constant $r > 0$ is the production rate. The net revenue is the sum of the following three terms:

1° The negative of the initial price, i.e., $-x_0$;

2° The excess of earnings $y(t)\Delta t$ over the maintenance expenditure $u(t)\Delta t$, which is to be discounted by a weight $\exp(-it)$, and summed over all the operation period, i.e.,

$$\int_0^{t_1} [y(t) - u(t)] \exp(-it) dt,$$

where the positive constant i is the rate of interest, $i < r$;

3° The discounted residual value of the machine at the replacement date t_1 , i.e., $x(t_1) \exp(-it_1)$.

The net revenue now is given by

$$-x_0 + \int_0^{t_1} [rx(t) - u(t)] \exp(-it) dt + x(t_1) \exp(-it_1). \quad (1)$$

Finally, let us denote by u_0 the upper bound of the maintenance costs, and suppose that $0 < u_0 \leq +\infty$.

A triple $s = (t_1, u, x)$ is said to be a *strategy* associated with the above data, if t_1 is a positive number; $u: [0, t_1] \rightarrow R$ is a piecewise continuous function (i.e., u is left-continuous and possesses a finite right-limit at each point in the open interval $]0, t_1[$, and it is continuous on the closed interval $[0, t_1]$ except for a finite subset E_u of $]0, t_1[$, such that for each t in $[0, t_1]$ we have $0 \leq u(t) \leq u_0$ when $u_0 < +\infty$, and $0 \leq u(t)$ when $u_0 = +\infty$, respectively; and $x: [0, t_1] \rightarrow R$ is a continuous function on $[0, t_1]$, which is differentiable on $[0, t_1] \setminus E_u$ and satisfies $x(0) = x_0$ and

$$\dot{x}(t) = -a(t) + F(t, u(t)) \quad \text{for} \quad t \in [0, t_1] \setminus E_u. \quad (2)$$

The components u and x of a strategy are usually called *control function* and *state function*, respectively. Let us denote by S the set of all strategies and by $C: S \rightarrow R$ the cost functional defined as follows

$$C(s) = \int_0^{t_1} L(t, u(t), x(t)) dt \quad \text{for} \quad s = (t_1, u, x) \in S,$$

where $L: [0, +\infty[\times R^2 \rightarrow R$ is the function given by

$$L(t, u, x) = -[(r-i)x - u - a(t) + F(t, u)] \exp(-it).$$

Recalling (2), the negative of the net revenue in (1) associated with a strategy $s = (t_1, u, x)$ can be rewritten as

$$\begin{aligned} & \int_0^{t_1} -[rx(t) - u(t)] \exp(-it) dt - [x(t_1) \exp(-it_1) - x(0)] = \\ & = \int_0^{t_1} -[rx(t) - u(t)] \exp(-it) dt + \int_0^{t_1} -\frac{d}{dt} [x(t) \exp(-it)] dt = \\ & = \int_0^{t_1} L(t, u(t), x(t)) dt = C(s). \end{aligned}$$

Now the maximization of the net revenue is equivalent to the problem: find a strategy $s_* = (t_*, u_*, x_*)$ in S such that $C(s_*) \leq C(s)$ for every strategy $s = (t_1, u, x)$ in S .

This is an optimal control problem with integral cost, free final time and free final state. We derive necessary conditions of optimality from the well-known maximum

principle of Pontrjagin (see [3]). To this end suppose that the functions a and F are continuous together with their partial derivatives of first order, and introduce the Hamiltonian of the problem:

$$H(t, u, x, p) = [(r-i)x - u - a(t) + F(t, u)] \exp(-it) + [-a(t) + F(t, u)]p.$$

Let us remark that the constraint functions $h_1, h_2: R \rightarrow R$, defined by $h_1(u) = -u$, and $h_2(u) = u - u_0$ when $u_0 < +\infty$, satisfy the constraint qualification in [3]:

$$\text{rank}(h'_1(u), h_1(u)) = \text{rank}(-1, -u) = 1$$

and

$$\text{rank} \begin{pmatrix} h'_1(u) & h_1(u) & 0 \\ h'_2(u) & 0 & h_2(u) \end{pmatrix} = \text{rank} \begin{pmatrix} -1 & -u & 0 \\ 1 & 0 & u - u_0 \end{pmatrix} = 2$$

when $u_0 < \infty$.

Now suppose that $s_* = (t_*, u_*, x_*) \in S$ is an optimal strategy, i.e., it is a solution of the above optimal control problem. Then, by the maximum principle of Pontrjagin, there exists a continuous function $p: [0, t_*] \rightarrow R$, which is differentiable on $[0, t_*] \setminus E_{u_*}$, such that the differential equation

$$\dot{p}(t) = - \frac{\partial H}{\partial x}(t, u_*(t), x_*(t), p(t)) = -(r-i) \exp(-it) \quad (3)$$

is satisfied for $t \in [0, t_*] \setminus E_{u_*}$, the maximum principle

$$H(t, u, x_*(t), p(t)) \leq H(t, u_*(t), x_*(t), p(t)), \quad t \in [0, t_*] \setminus E_{u_*}, \quad (4)$$

holds for $u \in [0, u_0]$ when $u_0 < +\infty$, and for $u \geq 0$ when $u_0 = +\infty$, and the transversality conditions

$$p(t_*) = 0 \quad \text{and} \quad H(t_*, u_*(t_*), x_*(t_*), p(t_*)) = 0 \quad (5)$$

are fulfilled. Since p is continuous, the integration of (3) with the final condition $p(t_*) = 0$ in (5) leads to

$$p(t) = \frac{r-i}{i} [\exp(-it) - \exp(-it_*)], \quad t \in [0, t_*],$$

so that the value of the Hamiltonian in the left side of (4) can be written as

$$H(t, u, x_*(t), p(t)) = \varphi(t) + h(t, u),$$

where

$$\varphi(t) = (r-i)x_*(t) \exp(-it) - a(t)g(t),$$

$$h(t, u) = -u \exp(-it) + F(t, u)g(t)$$

and

$$g(t) = \exp(-it) \left[\frac{r}{i} + \left(1 - \frac{r}{i} \right) \exp(-i(t_* - t)) \right].$$

Clearly, $g(t) > 0$ for $t \in [0, t_*]$. Now, the maximum principle (4) becomes

$$h(t, u) \leq h(t, u_*(t)), \quad t \in [0, t_*] \setminus E_{u_*}, \quad (6)$$

for $u \in [0, u_0]$ when $u_0 < +\infty$, and for $u \geq 0$ when $u_0 = +\infty$, and the equalities in (5) imply

$$(r-i)x_*(t_*) - u_*(t_*) - a(t_*) + F(t_*, u_*(t_*)) = 0. \quad (7)$$

From (6) we see that, for each t in $[0, t_*] \setminus E_{u_*}$, the value $u_*(t)$ of the optimal control is a maximum point for the function $u \rightarrow h(t, u)$ on the interval $[0, u_0]$ when $u_0 < +\infty$, and on the interval $[0, +\infty[$ when $u_0 = +\infty$. Hence, it will be useful to search the sign of the partial derivative

$$\frac{\partial h}{\partial u}(t, u) = g(t) \left[-G(t) + \frac{\partial F}{\partial u}(t, u) \right], \quad t \in [0, t_*] \setminus E_{u_*}, \quad (8)$$

near the point $u = u_*(t)$. Here the function $G: [0, t_*] \rightarrow R$, given by

$$G(t) = \frac{1}{g(t)} \exp(-it), \quad (9)$$

is positive, continuous and strictly increasing since

$$\frac{dG}{dt}(t) = \frac{r-i}{[g(t)]^2} \exp[-i(t+t_*)] > 0 \quad \text{for} \quad t \in [0, t_*].$$

Relations (6) and (8) facilitate the study of the optimal maintenance policy u_* on the length of the time interval $[0, t_*]$, and equation (7) permits the computation of the replacement date t_* .

The optimal strategies will be investigated for the linear and nonlinear variants of the optimal control problem, which correspond to two particular forms of the partial derivative $\frac{\partial F}{\partial u}$. The last derivative means the rate of the machine's residual value with respect to the maintenance expense, so that it can be regarded as a measure of effectiveness of the maintenance expense.

3. Linear variant

Admit that the function $(t, u) \rightarrow F(t, u)$ is linear with respect to the second variable, i.e., F has the form $F(t, u) = f(t)u$, where the function $f: [0, +\infty[\rightarrow \mathbb{R}$ is continuous together with its first derivative. In this case the differential equation (2) takes on the form

$$\dot{x}(t) = -a(t) + f(t)u. \quad (2')$$

3.1. THEOREM. Suppose $u_0 < +\infty$, f is decreasing on the interval $[0, +\infty[$ and $s_* = (t_*, u_*, x_*)$ is an optimal strategy. Then the optimal control u_* can be determined as follows:

a) If $f(t) < G(t)$ for all $t \in]0, t_*]$, we have

$$u_*(t) = 0 \quad \text{when} \quad t \in [0, t_*];$$

b) If $f(t) > G(t)$ for all $t \in [0, t_*[$, we have

$$u_*(t) = u_0 \quad \text{when} \quad t \in [0, t_*];$$

c) If neither of the conditions in a) and b) are satisfied, we have

$$u_*(t) = \begin{cases} u_0, & \text{when } t \in [0, t_0], \\ 0, & \text{when } t \in]t_0, t_*], \end{cases}$$

where t_0 is the single root in the open interval $]0, t_*[$ of the equation $f(t) = G(t)$.

Proof. If the condition in a) is satisfied, from (8) we get

$$\frac{\partial h}{\partial u}(t, u) = g(t) [-G(t) + f(t)] < 0$$

for all $t \in]0, t_*] \setminus E_{u_*}$ and all $u \in [0, u_0]$. Therefore, the function $u \rightarrow h(t, u)$ is strictly decreasing on the interval $[0, u_0]$, so that it attains its maximum value at the point $u = 0$. Hence $u_*(t) = 0$ when $t \in]0, t_*] \setminus E_{u_*}$. Since E_{u_*} is a finite subset of $]0, t_*[$ and u_* is a continuous function at the point $t = 0$ and left-continuous on the interval $]0, t_*[$, it follows that $u_*(t) = 0$ for every $t \in [0, t_*]$.

If the condition in b) is satisfied, then the function $u \rightarrow h(t, u)$ is strictly increasing on $[0, u_0]$ for each $t \in [0, t_*[\setminus E_{u_*}$, hence this function attains its maximum value at the point $u = u_0$. As above, we have $u_*(t) = u_0$ for each $t \in [0, t_*]$.

If the condition in c) is fulfilled, there exist a point t' in $]0, t_*]$ and a point t'' in $[0, t_*[$ such that $f(t') \geq G(t')$ and $f(t'') \leq G(t'')$. Since the function $t \rightarrow -G(t) + f(t)$ is strictly decreasing and continuous on $[0, t_*]$, the inequality $t' \leq t''$ holds, and there exists a single t_0 in $[t', t'']$ such that $f(t_0) = G(t_0)$. Consequently, $f(t) > G(t)$ for $0 \leq t < t_0$,

and $f(t) < G(t)$ for $t_0 < t \leq t_*$. Now, we take into account (8) to complete the proof of the theorem.

3.2. REMARK. G. L. Thompson [11] established Theorem 3.1 in the supplementary hypothesis that the obsolescence function a is increasing and satisfies the inequality

$$f(t)u_0 \leq a(t) \quad \text{for all } t \geq 0. \quad (10)$$

This inequality means that the maintenance action cannot be so effective as to enhance the value of the machine over previous values. However, as the following example shows, our Theorem 3.1 applies even if inequality (10) is not satisfied.

3.3. EXAMPLE. We keep up the data of the Thompson's Example 1:

$$x_0 = 100; \quad a(t) = 2; \quad r = 0.1; \quad i = 0.05 \quad \text{and} \quad u_0 = 1;$$

but we choose for maintenance effectiveness a greater function, namely the function $f: [0, +\infty[\rightarrow \mathbb{R}$ given by $f(t) = 4/(1+t)^{1/2}$. Suppose that $s_* = (t_*, u_*, x_*)$ is an optimal strategy. We shall prove that condition c) in Theorem 3.1 is satisfied. To this end we first remark that there exists a t' in $]0, t_*]$ such that $f(t') > G(t')$, since $f(0) = 4 > G(0)$ and the functions f and G are continuous on $[0, t_*]$. Also, putting $\bar{t} = t_* + 20 \cdot \ln 2$ and defining the extension $\bar{G}: [0, \bar{t}] \rightarrow \mathbb{R}$ of the function G in (9) by

$$\bar{G}(t) = [2 - \exp(-0.05(t_* - t))]^{-1},$$

there exists a single t'' in $]0, \bar{t}[$ such that $f(t'') = \bar{G}(t'')$, since \bar{G} is continuous and strictly increasing on $[0, \bar{t}[$, and $\bar{G}(t) \rightarrow +\infty$ as $t \rightarrow \bar{t}$.

It remains to prove the inequality $t'' < t_*$. Suppose the contrary: $t'' \geq t_*$. From

$$(1 + t'')^{1/2} = 8 - 4 \cdot \exp(-0.05(t_* - t'')) \leq 4$$

we derive $t_* \leq t'' \leq 15$. Since $f(t) > G(t)$ for all t in $[0, t_*[$, condition b) in Theorem 3.1 is fulfilled, hence $u_*(t) = 1$ when $t \in [0, t_*]$. Integrating (2') with $u = 1$ and $x(0) = 100$, we get

$$x_*(t) = 92 + 8(1+t)^{1/2} - 2t.$$

Now, equality (7) becomes

$$\frac{1}{20}(92 + 8(1+t_*)^{1/2} - 2t_*) - 3 + 4/(1+t_*)^{1/2} = 0,$$

hence the inequalities $0 < t_* \leq 15$ lead to the contradiction

$$120 < 32(1+t_*)^{1/2} + 8(1+t_*) + 80 = 2t_*(1+t_*)^{1/2} \leq 120.$$

Therefore, condition c) in Theorem 3.1 must be satisfied, and so we have

$$u_*(t) = \begin{cases} 1, & \text{when } t \in [0, t_0], \\ 0, & \text{when } t \in]t_0, t_*], \end{cases}$$

and

$$x_*(t) = \begin{cases} 92 + 8(1+t)^{1/2} - 2t, & \text{when } t \in [0, t_0], \\ 92 + 8(1+t_0)^{1/2} - 2t, & \text{when } t \in [t_0, t_*]. \end{cases}$$

Now, (7) and $f(t_0) = G(t_0)$ can be written as the system

$$\begin{cases} t_* = 26 + 4(1+t_0)^{1/2}, \\ (1+t_0)^{1/2} = 8 - 4 \cdot \exp(-0.05(t_* - t_0)), \end{cases}$$

whose solution is $t_0 = 35.75$, $t_* = 50.25$. The value of the machine at the replacement time is $x_*(t_*) = 40.00$ and the maximal net revenue is $-C(s_*) = 53.20$.

For comparison, we mention that the effectiveness f in Thompson's example is given by $f(t) = 2(1+t)^{-1/2}$ and the corresponding numerical results are $t_0 = 10.70$, $t_* = 34.84$, $x_*(t_*) = 40.00$ and $-C(s_*) = 34.69$.

4. Nonlinear variant

In this section we suppose that F does not depend on time, i.e., F has the form $F(t, u) = F_0(u)$, where the function $F_0: R \rightarrow R$ is continuous together with its first derivative. Here the differential equation (2) takes on the form

$$\dot{x}(t) = -a(t) + F_0(u). \quad (2'')$$

4.1. THEOREM. Suppose $u_0 = +\infty$, the derivative F'_0 is strictly decreasing on the interval $[0, +\infty[$, $F'_0(u) \rightarrow 0$ as $u \rightarrow +\infty$, and $s_* = (t_*, u_*, x_*)$ is an optimal strategy. Then the optimal control u_* can be determined as follows:

a) If $F'_0(0) \geq 1$, we have

$$u_*(t) = f_0^{-1}(G(t)) \quad \text{when } t \in [0, t_*];$$

b) If $F'_0(0) \leq G(0)$, we have

$$u_*(t) = 0 \quad \text{when } t \in [0, t_*];$$

c) If $1 > F'_0(0) > G(0)$, we have

$$u_*(t) = \begin{cases} f_0^{-1}(G(t)), & \text{when } t \in [0, t_0], \\ 0, & \text{when } t \in]t_0, t_*], \end{cases}$$

where f_0^{-1} is the inverse of the function $f_0: [0, +\infty[\rightarrow]0, F'_0(0)]$, given by $f_0(u) = F'_0(u)$, and t_0 is the single root in the open interval $]0, t_*[$ of the equation $F'_0(0) = G(t)$, namely

$$t_0 = t_* + \frac{1}{i} \ln [(rF'_0(0) - i)((r - i)F'_0(0))^{-1}]. \quad (11)$$

Proof. If the condition in a) is satisfied, we get

$$F'_0(0) \geq 1 = G(t_*) > G(t) \quad \text{for all } t \in [0, t_*[, \quad (12)$$

since G is strictly increasing. Also, for each t in $[0, t_*[\setminus E_{u_*}$ we have $u_*(t) > 0$. Indeed, supposing the contrary: $u_*(t) = 0$ for a t in $[0, t_*[\setminus E_{u_*}$, from (6) we derive $h(t, 0) = h(t, u_*(t)) = \sup \{h(t, u) : u \geq 0\}$ which yields

$$\frac{\partial h}{\partial u}(t, 0) = \lim_{u \rightarrow 0, u > 0} \frac{h(t, u) - h(t, 0)}{u} \leq 0;$$

and hence, by (8), $-G(t) + F'_0(0) \leq 0$. Combining this with (12), we arrive at the contradiction $F'_0(0) \leq G(t) < F'_0(0)$. Therefore, $u_*(t) > 0$ for each t in $[0, t_*[\setminus E_{u_*}$. Now, by

Fermat's theorem, $\frac{\partial h}{\partial u}(t, u_*(t)) = 0$, hence from (8) we deduce

$$F'_0(u_*(t)) = G(t) \quad \text{when } t \in [0, t_*[\setminus E_{u_*}. \quad (13)$$

Since the function f_0 has an inverse, (12) and (13) lead to

$$u_*(t) = f_0^{-1}(G(t)) \quad \text{when } t \in [0, t_*[\setminus E_{u_*}.$$

In fact, the last equality holds for all t in $[0, t_*]$, because the set E_{u_*} is finite, the function u_* is piecewise continuous on $[0, t_*]$, and the functions G and f_0^{-1} are continuous on $[0, t_*]$ and $]0, F'_0(0)]$, respectively.

If the condition in b) is satisfied, we get $F'_0(u) \leq F'_0(0) \leq G(0) < G(t)$ for $t \in]0, t_*]$ and $u \geq 0$, whence $\frac{\partial h}{\partial u}(t, u) < 0$ for $t \in]0, t_*] \setminus E_{u_*}$ and $u \geq 0$. The same argument as in the proof of Theorem 3.1, condition a), shows that $u_*(t) = 0$ for all $t \in [0, t_*]$.

If the condition in c) is fulfilled, it follows that the open interval $]0, t_*[$ contains a single point t_0 with $f_0(0) = G(t_0)$, since $G(t_*) = 1 > F'_0(0) > G(0)$. The arguments developed in the presence of condition a) yield the formula

$$u_*(t) = f_0^{-1}(G(t)) \quad \text{for } t \in [0, t_0[\setminus E_{u_*}$$

and then the formula

$$u_*(t) = f_0^{-1}(G(t)) \quad \text{for all } t \in [0, t_0].$$

When $t \in]t_0, t_*] \setminus E_{u_*}$, we have $u_*(t) = 0$ since the contrary relation $u_*(t) > 0$ for a t in $]t_0, t_*] \setminus E_{u_*}$ implies $\frac{\partial h}{\partial u}(t, u_*(t)) = 0$, and so we arrive at the contradiction $G(t) = F'_0(u_*(t)) < F'_0(0) = G(t_0) < G(t)$. Hence, $u_*(t) = 0$ for all t in $]t_0, t_*]$, the proof of the theorem is complete.

4.2. REMARKS. (i) Theorem 4.1 remains true if the condition $F'_0(u) \rightarrow 0$ as $u \rightarrow +\infty$ is replaced by the less restrictive condition $\lim_{u \rightarrow +\infty} F'_0(u) < G(0)$. In the last case the range of f_0 will be the interval $] \lim_{u \rightarrow +\infty} F'_0(u), F'_0(0)]$.

(ii) With the supplementary hypothesis $F'_0(u) \leq 1$ for all $u \geq 0$, a variant of Theorem 4.1 has been established by J.-Y. Helmer [2], pp. 128–137. This hypothesis is fulfilled if $a(0) = 0$ and for any maintenance strategy it is impossible to surpass the initial value of the machine.

4.3. EXAMPLES. We shall illustrate by examples the three situations appearing in Theorem 4.1. The common data of these examples are:

$$x_0 = 100; \quad a(t) = 2; \quad r = 0.1; \quad i = 0.05;$$

and the function $F_0: R \rightarrow R$ given by

$$F_0(u) = \begin{cases} ku, & \text{when } u \leq 0, \\ k \cdot \ln(1+u), & \text{when } u > 0, \end{cases}$$

where k is a positive number. Suppose that $s_* = (t_*, u_*, x_*)$ is an optimal strategy.

First admit $k = 1$. Then condition a) in Theorem 4.1 is satisfied, hence

$$u_*(t) = \frac{1}{G(t)} - 1 \quad \text{when } t \in [0, t_*].$$

By (2'') we obtain

$$x_*(t) = 100 - 2t - \int_0^t \ln G(s) ds \quad \text{when } t \in [0, t_*].$$

Equation (7) becomes

$$2t_* - \int_0^{t_*} \ln [2 - \exp(-0.05(t_* - t))] dt = 60$$

and possesses the solution $t_* = 38.16$. The net revenue is $-C(s_*) = 31.72$.

Next, suppose that $0 < k \leq \frac{1}{2}$. Then

$$F'_0(0) = k \leq \frac{1}{2} < \left[\frac{r}{i} + \left(1 - \frac{r}{i} \right) \exp(-it_*) \right]^{-1} = G(0)$$

and so condition b) is fulfilled. In particular for $k = \frac{1}{3}$ we obtain

$$u_*(t) = 0 \quad \text{and} \quad x_*(t) = 100 - 2t \quad \text{when} \quad t \in [0, t_*],$$

and then $t_* = 30.00$ and $-C(s_*) = 28.92$.

Finally, suppose that $G(0) < k < 1$, where

$$G(0) = [2 - \exp(-3/2)]^{-1} = 0.56278742.$$

Then $F'_0(0) > G(0)$, since otherwise condition b) in Theorem 4.1 would imply $t_* = 30$ and then the contradiction

$$k = F'_0(0) \leq G(0) = [2 - \exp(-it_*)]^{-1} < k.$$

Therefore, the condition in c) is satisfied and so we have

$$u_*(t) = \begin{cases} \frac{k}{G(t)} - 1, & \text{when } t \in [0, t_0], \\ 0, & \text{when } t \in]t_0, t_*], \end{cases}$$

and

$$x_*(t) = \begin{cases} x_0 - \int_0^t \left(a + k \cdot \ln \frac{G(s)}{k} \right) ds, & \text{when } t \in [0, t_0], \\ x_*(t_0) - a \cdot (t - t_0), & \text{when } t \in [t_0, t_*], \end{cases}$$

where $t_0 \in]0, t_*[$ is the number in (11) and it is given by

$$t_0 = t_* + \frac{1}{i} \cdot \ln \frac{kr - i}{k(r - i)}. \quad (14)$$

Moreover, by (7) we obtain another equation for the unknowns t_* and t_0 :

$$(r - i) \left[x_0 + t_0 k \cdot \ln k - k \cdot \int_0^{t_0} \ln G(t) dt - at_* \right] = a. \quad (15)$$

From (14) and (15) we derive the equation determining t_0 :

$$\int_0^{t_0} \ln [kr - (kr - i) \exp(-i(t_0 - t))] dt - \left(\frac{a}{k} + \ln i \right) t_0 + \frac{1}{k} \left[x_0 - \frac{a}{r - i} - \frac{a}{i} \cdot \ln \frac{k(r - i)}{kr - i} \right] = 0. \quad (16)$$

The net revenue is given by

$$\begin{aligned}
 -C(s_*) = & -x_0 + \frac{1+rx_0}{i} + \frac{r}{i^2} \left\{ -a - k + \right. \\
 & + k \cdot \ln \left[k \left(\frac{r}{i} + \left(1 - \frac{r}{i} \right) \exp(-it_*) \right) \right] \left. \right\} + \left\{ -\frac{1}{i} (1 + krt_0) + \right. \\
 & \left. + \frac{kr}{i^2} \left[1 - \ln \left[k \left(\frac{r}{i} \exp(-it_0) + \left(1 - \frac{r}{i} \right) \exp(-it_*) \right) \right] \right] \right\} \cdot \\
 & \cdot \exp(-it_0) + \left[-kt_0 + \frac{1}{i} (krt_0 - a) + \frac{ar}{i^2} + \right. \\
 & \left. + \frac{k(r-i)}{i^2} \cdot \ln \frac{r \cdot \exp(-it_0) + (i-r) \exp(-it_*)}{r + (i-r) \exp(-it_*)} \right] \exp(-it_*). \quad (17)
 \end{aligned}$$

In particular, for $k=0.8$ we obtain $t_* = 32.49$; $t_0 = 26.74$ and $-C(s_*) = 29.60$. The approximate evaluation of the definite integral in (16) is performed by the trapezoidal formula.

We remark that the same equalities (14), (16) and (17) can be used for computation of optimal strategies in the first two cases, if we take $t_0 = t_*$, $k=1$, and $t_0=0$, $k \rightarrow 0$, respectively.

References

1. Arora, S. R., Lele, P. T., A note on optimal maintenance policy and sale date of a machine. *Management Sci.*, **17** (1970), 170-173.
2. Helmer, J.-Y., La commande optimale en économie. Dunod, Paris, 1972.
3. Hestenes, M. R., On variational theory and optimal control theory. *J. SIAM on Control*, **3** (1963), 23-48.
4. Kamien, M. I., Schwartz, N. L., Optimal maintenance and sale age for a machine subject to failure. *Management Sci.*, **17** (1971), B495-B504.
5. Khandelwall, D. N., Sharma, J., Ray, L. M., Optimal periodic maintenance of a machine. *IEEE Trans. Automatic Control*, **24** (1979), 513.
6. Naslund, B., Simultaneous determination of optimal repair policy and service life. *Swedish J. of Economics*, **68** (1966), 63-73.
7. Nguyen, D. G., Murthy, D. N. P., Optimal preventive maintenance policies for repairable systems. *Operations Res.*, **29** (1981), 1181-1194.
8. Sethi, S. P., A survey of management science applications of the deterministic maximum principle. *Applied optimal control*, pp. 33-68. *Studies in the Management Sci.*, **9**, North-Holland, Amsterdam, 1978.
9. Sethi, S. P., Chand, S., Planning horizon procedures for machine replacement models. *Management Sci.*, **25** (1979/1980), 140-161. Erratum, *ibid.*, **26** (1980), 342.

10. *Tapiero, C. S., Venezia, I.*, A mean variance approach to the optimal machine maintenance and replacement problem. *J. Oper. Res. Soc.*, **30** (979), 457466.
11. *Thompson, G. L.*, Optimal maintenance policy and sale date of a machine. *Management Sci.*, **14** (1968), 543-550.

Одновременная оптимизация стратегии оделуживания и замены машин

И. МУНТЯН

(Клуж—Напока)

Для одновременной оптимизации стратегии обслуживания и замены машин, а также их коммерческих характеристик применяются методы оптимального управления. Обычное предположение о том, что затраты на обслуживание не столь велики, как увеличение стоимости машины по сравнению с первоначальной, в данной статье не требуются. Для линейного и нелинейного случая задачи совместной оптимизации управления выводятся в явном виде выражения для оптимальной стратегии и анализируются некоторые численные примеры.

I. Muntean

Faculty of Mathematics

University "Babeş-Bolyai"

Str. M. Kogălniceanu 1

3400 Cluj-Napoca, Romania

DIRECT POLYNOMIAL APPROACH TO DISCRETE-TIME STOCHASTIC TRACKING

M. ŠEBEK

(Prague)

(Received July 27, 1982)

A new technique to design optimal linear *discrete-time* multivariable stochastic systems is presented. The technique is based on polynomial matrices. The optimal controller is shown to consist of feedback and feedforward parts and it is designed by solving two polynomial matrix equations whose coefficients are obtained by spectral factorization.

Introduction

The problem of signals tracking in the presence of disturbances, the so-called *stochastic tracking problem (STP)*, is one of the most significant problems of optimal control. It was not satisfactorily solved until the development of state space theory. The prevalent philosophy used in the state space approach (see Kwakernaak and Sivan [4]) is to reformulate the given tracking problem to a *regulator* problem. This can be successfully accomplished by means of the well known trick which consists in augmenting the plant and the reference dynamics into a single composite system (see Kalman and Koepcke [6]). Since the regulator problem is to be solvable for the augmented system, this procedure is limited to stable reference generators. However, this restriction is somewhat artificial, for the composite system does not exist in reality at all.

As an alternative to the state space approach, a polynomial solution to the STP has been reported recently for single-input single-output systems by Kučera [3]. Although his approach is different, the prevalent philosophy of the state space approach is kept: tracking is again reformulated to regulation. As a consequence, the restriction to stable reference generators cannot be circumvented.

The aim of this paper is to show that the polynomial techniques make it possible to abandon the philosophy above and to derive, for multi-input multi-output discrete-time plants, a new *direct* solution to the STP which is general enough to handle unstable and/or nonminimum-phase plants and reference generators with rectangular transfer matrices as well as nonnegative definite noise intensities and matrices in the measure of performance.

For a continuous time version of this approach the reader is referred to [5].

A prominent role throughout this paper will play real polynomial matrices. They are treated in detail in the books by Wolovich [8], Kučera [2] and Kailath [1]. As usual, for any polynomial matrix $H(d)$ define $H_*(d) = H^T(d^{-1})$ and $\langle H \rangle = H(0)$.

Formulation

Consider a multivariable discrete-time stochastic plant modeled by the controlled ARMA process

$$A(d)y = B(d)u + C(d)w \quad (1)$$

where y is the vector output sequence, u is the vector input sequence and w is the background noise. The $A(d)$, $B(d)$ and $C(d)$ is a left coprime triple of polynomial matrices in the delay operator d such that $A(0)$ is invertible while $B(0) = 0$.

Let the measured output of the plant be corrupted by an additive observation noise v .

Further consider a reference vector sequence \bar{y} modeled by the ARMA process

$$\bar{A}(d)\bar{y}(d) = \bar{C}(d)\bar{w} \quad (2)$$

where $\bar{A}(d)$ and $\bar{C}(d)$ are left coprime polynomial matrices in d such that $\bar{A}(0)$ is invertible.

Let the available version of the reference sequence be corrupted by an additive observation noise \bar{v} .

All four vector random sources v , w and \bar{v} , \bar{w} are pairwise uncorrelated zero-mean covariance-stationary white vector random processes with intensities Λ , Ω and $\bar{\Lambda}$, $\bar{\Omega}$, respectively, which all are real nonnegative definite matrices.

For a given plant and reference, the design of the optimal controller

$$P(d)u = -Q(d)(y + v) + \bar{Q}(d)(\bar{y} + \bar{v}) \quad (3)$$

evolves from minimization of the weighted sum of steady-state variances of the plant input and the tracking error, i.e., of the cost

$$J = \text{tr} \langle \Phi W_u \rangle + \text{tr} \langle \Psi W_{\bar{y}-y} \rangle \quad (4)$$

where W_u and $W_{\bar{y}-y}$ are correlation functions of u and $\bar{y} - y$, respectively, in steady state and Φ and Ψ are real nonnegative definite weighting matrices.

Thus the objective of our design is to minimize (4) subject to the constraint that the tracking system defined by (1) through (3) be asymptotically stable. At the same time we shall assume that all spectral factors defined below by (6)–(7) exist.

Design procedure

The purpose of this section is to present the entire design procedure in an easy-to-follow way. Setting aside for the moment all questions of solvability, we start with the primary data $A(d), B(d), C(d), \bar{A}(d), \bar{C}(d)$ and $\Phi, \Psi, \Lambda, \Omega, \bar{\Lambda}, \bar{\Omega}$ and attempt to construct $P(d), Q(d)$ and $\bar{Q}(d)$.

For simplicity of notation, the function arguments are omitted wherever convenient. The optimal design is carried out in the following steps.

- 1) Find left coprime polynomial matrices A_0, B_0 and right coprime polynomial matrices A_1, B_1 such that

$$A_0^{-1}B_0 = B_1 A_1^{-1} = A^{-1}B \tag{5}$$

- 2) Perform the spectral factorization to obtain stable polynomial matrices F, G and \bar{G} satisfying

$$A_{1*} \Phi A_1 + B_{1*} \Psi B_1 = F_* F \tag{6}$$

$$A \Lambda A_* + C \Omega C_* = G G_* \tag{7}$$

$$\bar{A} \bar{\Lambda} \bar{A}_* + \bar{C} \bar{\Omega} \bar{C}_* = \bar{G} \bar{G}_* \tag{8}$$

- 3) Calculate any left coprime matrix fraction

$$\bar{A}_3^{-1} A_3 = A_0 \bar{A}^{-1} \tag{9}$$

and right coprime polynomial matrix fractions

$$\begin{aligned} B_2 G_1^{-1} &= G^{-1} B & A_2 G_2^{-1} &= G^{-1} A \\ DE^{-1} &= (A_3 \bar{G})^{-1} \bar{A}_3 B_0 & \bar{A}_2 \bar{G}_2^{-1} &= \bar{G}^{-1} \bar{A} \end{aligned} \tag{10}$$

- 4) Solve the equations

$$F_* [X, Y] + Z_* [B_2, -A_2] = [A_{1*} \Phi G_1, B_{1*} \Psi G_2] \tag{11}$$

$$F_* [\bar{X}, \bar{Y}] + \bar{Z}_* [D, -\bar{A}_2] = [A_{1*} \Phi E, B_{1*} \Psi \bar{G}_2] \tag{12}$$

for polynomial matrices X, Y, Z and $\bar{X}, \bar{Y}, \bar{Z}$ such that $\langle Z \rangle = 0$ and $\langle \bar{Z} \rangle = 0$.

- 5) The desired P, Q and \bar{Q} are now read from a numerator of a left coprime matrix fraction

$$\bar{G}^{-1} [P, Q, \bar{Q}] = [X G_1^{-1}, Y G_2^{-1}, \bar{Y} \bar{G}_2^{-1}] \tag{13}$$

The optimal controller is to be realized as a single dynamical system of least order.

There are efficient algorithms to implement all steps of the design procedure. They can be found gathered together, e.g., in [2].

Solvability

The preceding design procedure produces optimal controller whenever one exists. It is the purpose of this section to prove this claim.

Theorem. The discrete-time optimal tracking problem is solvable iff

- the greatest common left divisor of A and B is a stable polynomial matrix
- \bar{A}_3 is a stable polynomial matrix (i.e., the unstable part of \bar{A} is a right divisor of A_0)

The solution is unique and the optimal controller is given by (13).

Proof. First we shall justify the design procedure by constructing the optimal controller provided it exists. The proof differs from its continuous-time counterpart (given in [5]) enough to be given here, although its underlying philosophy is similar: to express the cost as a sum of squares with only two depending on the controller matrices.

To this effect, write D for a greatest common left divisor of A and B so that $A = DA_0$, $B = DB_0$ and define rational matrices M, N and M_1, N_1 via relations

$$P^{-1}Q = M^{-1}N \quad (14)$$

and

$$\begin{bmatrix} A_0 & B_0 \\ N & -M \end{bmatrix} \begin{bmatrix} M_1 & B_1 \\ N_1 & -A_1 \end{bmatrix} = I. \quad (15)$$

Furthermore, write H for any greatest common right divisor of A_0 and \bar{A} so that $A_0 = \tilde{A}_0 H$ and $\bar{A} = \tilde{A} H$ and define a right coprime polynomial fraction to $\tilde{A}^{-1} B_0$,

$$B_4 A_4^{-1} = \tilde{A}^{-1} B_0 \quad (16)$$

and denote $H_1 = A_4^{-1} A_1$. It can be shown that H_1 is a polynomial matrix.

Finally define rational matrices \bar{M}, \bar{N} and \bar{M}_1, \bar{N}_1 via the relations

$$P^{-1}\bar{Q} = M^{-1}\bar{N} \quad (17)$$

and

$$\begin{bmatrix} H & B_4 \\ \bar{N} & -\bar{M} \end{bmatrix} \begin{bmatrix} \bar{M}_1 & B_1 \\ \bar{N}_1 & -H_1 \end{bmatrix} = I. \quad (18)$$

Straightforward analysis then yields

$$\begin{aligned} u &= -A_1 N v - A_1 N A^{-1} C w + A_1 \bar{N} \bar{v} + A_1 \bar{N} \bar{A}^{-1} \bar{C} \bar{w} \\ \bar{y} - y &= B_1 N v - (I - B_1 N) A^{-1} C w - B_1 \bar{N} \bar{v} + (I - B_1 \bar{N}) \bar{A}^{-1} \bar{C} \bar{w} \end{aligned} \quad (19)$$

or, equivalently,

$$\begin{aligned} u &= -A_1 N v - N_1 D^{-1} C w + A_1 \bar{N} \bar{v} + A_4 \bar{N}_1 \bar{A}_0^{-1} \bar{C} \bar{w} \\ \bar{y} - y &= B_1 N v - M_1 D^{-1} C w - B_1 \bar{N} \bar{v} + \bar{M}_1 \bar{A}^{-1} - \bar{C} \bar{w}. \end{aligned} \quad (20)$$

Now,

$$\begin{aligned} W_u &= A_1 N A N_* A_{1*} + A_1 N A^{-1} C \Omega C_* A_*^{-1} N_* A_{1*} + \\ &+ A_1 \bar{N} \bar{\Lambda} \bar{N}_* A_{1*} + A_1 \bar{N} \bar{A}^{-1} \bar{C} \bar{\Omega} \bar{C}_* \bar{A}_*^{-1} \bar{N}_* A_{1*} = \\ &= A_1 N A^{-1} G G_* A_*^{-1} N_* A_{1*} + A_1 \bar{N} \bar{A}^{-1} \bar{G} \bar{G}_* \bar{A}_*^{-1} N_* A_{1*} \end{aligned} \quad (21)$$

where use has been made of (7) and (8), while, similarly,

$$\begin{aligned} W_{y-y} &= (I - B_1 N) A^{-1} G G_* A_*^{-1} (I - B_1 N)_* + B_1 N A + \Lambda N_* B_{1*} - A \\ &+ (I - B_1 \bar{N}) \bar{A}^{-1} \bar{G} \bar{G}_* \bar{A}_*^{-1} (I - B_1 \bar{N})_* + B_1 \bar{N} \bar{\Lambda} + \bar{\Lambda} \bar{N}_* B_{1*} - \bar{A}. \end{aligned} \quad (22)$$

A simple calculation based on the result $\text{tr } KL = \text{tr } LK$ and on the assumption $\langle B_1 \rangle = 0$ then gives, after completing the squares,

$$J = \text{tr} \langle U_* U \rangle + \text{tr} \langle V \rangle + \text{tr} \langle \bar{U}_* \bar{U} \rangle + \text{tr} \langle \bar{V} \rangle - \text{tr} \Psi(A + \bar{A}) \quad (23)$$

where

$$\begin{aligned} U &= (FN - F_*^{-1} B_{1*} \Psi) A^{-1} G \\ V &= G_* A^{-1} (\Psi - \Psi B_1 F^{-1} F_*^{-1} B_{1*} \Psi) A^{-1} G \\ \bar{U} &= (F\bar{N} - F_*^{-1} B_{1*} \Psi) \bar{A}^{-1} \bar{G} \\ \bar{V} &= \bar{G}_* \bar{A}^{-1} (\Psi - \Psi B_1 F^{-1} F_*^{-1} B_{1*} \Psi) \bar{A}^{-1} \bar{G}. \end{aligned}$$

With a help of the second part of equation (11) written in the form

$$F_*^{-1} B_{1*} \Psi A^{-1} G = -F_*^{-1} Z_* Y A_2^{-1}$$

we have

$$U = W + T_*$$

where

$$W = F N A^{-1} G - Y A = W_0 + W_1 d + \dots \quad (24)$$

is a formal power series in nonnegative powers of d while

$$T_* = F_*^{-1} Z_* = T_1 d^{-1} + T_2 d^{-2} + \dots \quad (25)$$

is a formal power series in negative powers.

Now simply

$$\text{tr} \langle U_* U \rangle = \sum_{i=0}^{\infty} \text{tr} W_i W_i^T + \sum_{i=1}^{\infty} \text{tr} T_i T_i^T$$

and the best we can do to minimize this term is to put $W=0$.

Along the same lines, on employing the first part of (12) the best we can do to minimize the term

$$\text{tr} \langle \bar{U}_* \bar{U} \rangle$$

is to put $\bar{W}=0$, where

$$\bar{W} = F\bar{N}A^{-1}G - YA. \quad (26)$$

Since the other terms of (23) are independent of the controller we can do nothing more. From a glance at (24) and (26), the optimal controller is given by

$$N = F^{-1}YG_2^{-1} \quad (27)$$

$$\bar{N} = F^{-1}\bar{Y}\bar{G}^{-1} \quad (28)$$

Up to now, merely the second part of (11), i.e.,

$$F_*Y - Z_*A_2 = B_{1*}\Psi G_2 \quad (29)$$

has been derived. However, employing (6), (15) and (29), simple algebraic manipulations yield

$$F_*FMG_1 + Z_*B_2 = A_{1*}\Phi G_1 \quad (30)$$

Since, evidently, FMG_1 must be polynomial matrix, we have

$$M = F^{-1}XG_1^{-1} \quad (31)$$

for a polynomial matrix X and substituting it in (30) the first part of (11) is derived. Moreover, comparing (27), (28) and (31) with (14), (17) and (13) results.

Analogously, the first part of (12) can be derived along with the relation

$$\bar{M} = F^{-1}\bar{X}\bar{G}^{-1} \quad (32)$$

where \bar{G}^{-1} is given by

$$\bar{B}\bar{G}^{-1} = \bar{G}_2^{-1}B_1.$$

In order for the cost to represent a finite weighted sum of steady-state variances, clearly, all terms in (19) are to be stable rational matrices. Now M and N are stable due to (27), (31) and, on employing Theorem 5.9 from [2], M_1 and N_1 are stable as well. Along the same lines, \bar{M} , \bar{N} , \bar{M}_1 and \bar{N}_1 are all stable rational matrices. Thus the stability of u and $y - \bar{y}$ hinges on the polynomial matrices D and \bar{A}_0 . Since \bar{D} and C are left coprime (for A , B and C is a left coprime triple by assumption) as well as \bar{A} and \bar{C} are left coprime (for \bar{A} and \bar{C} are left coprime) and, moreover, $\det \bar{A}_3 = \det \bar{A}$, the conditions a) and b) result.

Finally, the resulting tracking system is asymptotically stable for M , N (M_1 , N_1) are stable rational matrices (see [2]).

Conclusion

A polynomial solution to discrete-time stochastic tracking problems has been presented. The approach is based on external polynomial models and the design procedure is reduced to the solution of linear polynomial matrix equations whose coefficients are obtained by spectral factorization. There are efficient algorithms to implement all steps of the design procedure, see e.g. Kučera [2] and Vostrý [7].

The optimal controller has been shown to consist of an output feedback and a reference feedforward. This configuration is clearly superior to processing just the tracking error $r - y$. Moreover, the feedback part of the controller turns out to be independent of the reference. This means that the controller ensures optimal regulation and optimal tracking at the same time.

The feedback part of the controller is specified by equation (1). This equation can be interpreted as the assignment of desired pole locations, which have been determined by spectral factorizations. The feedforward part of the controller is specified by equation (12). This equation can then be thought of as adjusting the zeros of the optimal system.

It follows from (15) that the characteristic polynomial of the optimal system is the multiple of determinants of the matrices F , G , and \bar{G} . F corresponds to a complete state feedback, G to an observer which estimates the plant's state, and \bar{G} to an observer for the reference generator. These matrices must be chosen in an optimal way employing spectral factorizations (6), (7) and (8).

The polynomial equation approach is thus seen to be an alternative to the state space methods. Rather than constructing the state feedback and the observers separately, the optimal controller is found here as a single dynamical system whose realization is left to the designer.

References

1. Kailath, T., Linear Systems. Prentice Hall, Englewood Cliffs, 1980.
2. Kučera, V., Discrete Linear Control: The Polynomial Equation Approach. Wiley, Cichester, 1979.
3. Kučera, V., Discrete stochastic regulation and tracking, *Kybernetika*, **16**, 3, 263-272, 1980.
4. Kwakernaak, H., Sivan, R., Linear Optimal Control Systems. Wiley, New York, 1972.
5. Šebek, M., Polynomial design of stochastic tracking systems, *IEEE Trans. Automat. Contr.*, **AC27**, 2, April 1982.
6. Kalman, R. E., Koepcke, R. W., Optimal synthesis of linear sampling control systems using generalized performance indexes. *Trans. ASME Ser. D*, Vol. **80**, 1958.
7. Vostrý, Z., New algorithm for polynomial spectral factorization with quadratic convergence, *Kybernetika*, **11**, 415-422, 1975; *Kybernetika*, **12**, 248-259, 1976.
8. Wolovich, W. A., Linear Multivariable Systems. Springer, New York, 1974.

Прямой полиномиальный подход к дискретным стохастическим сервосистемам

М. ШЕБЕК

(Прага)

В статье предлагается новый метод синтеза оптимальных линейных дискретных много-
связных стохастических сервосистем. Оказывается, что оптимальная управляющая система
содержит как канал обратной связи так и канал связи, по задающим воздействиям. Она получается с
помощью алгебры полиномиальных матриц.

Синтез сведен к решению двух уравнений с матрицами полиномов и к спектральной
факторизации матриц.

M. Šebek

Institute of Information Theory and Automation

Czechoslovak Academy of Sciences

182 08 Praha 8

Pod vodárenskou věží 4

Czechoslovakia

ТУДОМАНУОС АКАДЕМИА
KÖNYVTÁRA

PRINTED IN HUNGARY

Akadémiai Nyomda, Budapest

A SURVEY OF SUPERIMPOSED CODE THEORY

A. G. DYACHKOV, V. V. RYKOV

(Moscow)

Generalization of superimposed code concept introduced by Kautz and Singleton is considered in this paper. The survey of new results of superimposed code theory is presented. Some open problems are formulated.

1. Notations, definitions of superimposed codes and their properties

Let $1 \leq s < t$, $1 \leq L \leq t - s$, $N \geq 1$ be integers, and let $\mathbf{u}(j) = (u_1(j), u_2(j), \dots, u_N(j))$, $j = 1, s$ denote the binary columns (of 0 and 1) of length N . The Boolean sum $\mathbf{u} = \mathbf{u}(1) \vee \mathbf{u}(2) \vee \dots \vee \mathbf{u}(s)$ of columns $\mathbf{u}(1), \dots, \mathbf{u}(s)$ is the binary column $\mathbf{u} = (u_1, u_2, \dots, u_N)$ with components

$$u_i = \begin{cases} 0, & \text{if } u_i(1) = u_i(2) = \dots = u_i(s) = 0, \\ 1 & \text{otherwise.} \end{cases}$$

Let us say that column \mathbf{u} covers column \mathbf{v} , iff $\mathbf{u} \vee \mathbf{v} = \mathbf{u}$.

Let $\mathbf{x} = \|\mathbf{x}_i(j)\|$, $i = 1, N$, $j = 1, t$ be a binary $N \times t$ -matrix of 0 and 1, and the symbol $\mathbf{x}(j)$ denote the j -th column of \mathbf{x} . Later on the matrix \mathbf{x} is interpreted as a set of t binary columns $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)$.

Definition 1. An $N \times t$ matrix \mathbf{x} is called a *superimposed* (s, t, L) -code of length N if the Boolean sum of any s -subset of its columns can cover not more than $L - 1$ columns that are not components of the given Boolean sum.

Note, that in the most important particular case $L = 1$, definition 1 is equivalent to the following condition. The Boolean sum of any s -subset of columns \mathbf{x} covers those and only those columns that are the components of given Boolean sum.

Superimposed $(s, t, 1)$ -codes were introduced in reference [1] by Kautz and Singleton. Applied problems leading to the definition of $(s, t, 1)$ -codes and some methods of construction of such codes are described in [1]. New applications of (s, t, L) -codes, $L \geq 1$, for multiple access channel are considered in [2].

Remark 1. Let $[N]$ be the set of integers from 1 to N . Each column of the matrix \mathbf{x} is identified with the subset of $[N]$, which consists of positions where this column

contains 1's. Then using the terminology of sets, the construction of an (s, t, L) -code of length N is equivalent to the following combinatorial problem. A family of t subsets of the set $[N]$ should be constructed in which the union of no s members of the family can include more than $L-1$ members of the family, different from the members of this union.

Let $N(s, t, L)$ be the minimal possible length of an (s, t, L) -code. The following symbols will be used:

$\lceil b \rceil$ — the least integer $\geq b$,

$\lfloor b \rfloor$ — the largest integer $\leq b$,

\log — logarithm of base 2,

\triangleq — equality by definition,

e — base of natural logarithm,

$h(u) = -u \log u - (1-u) \log (1-u)$ — binary entropy.

The following propositions follow easily by definition 1.

Proposition 1. Any $(s+1, t, L)$ -code is an (s, t, L) -code, and any (s, t, L) -code is an $(s, t, L+1)$ -code. Hence,

$$N(s+1, t, L) \geq N(s, t, L) \geq N(s, t, L+1).$$

Proposition 2. For any s -subset of columns of an (s, t, L) -code there are not more than $\binom{s+L-1}{s}$ s -subsets of columns, such that the Boolean sums of columns of these s -subsets coincide with the Boolean sum of columns of the given s -subset. Hence

$$N(s, t, L) \geq \left\lceil \log \binom{t}{s} - \log \binom{s+L-1}{s} \right\rceil \quad (1)$$

Proposition 3. Let $\lfloor t/L \rfloor$ pairwise disjoint L -subsets of the columns of an (s, t, L) -code be fixed. Let \mathbf{x}' be a $(0, 1)$ -matrix of N rows and $\lfloor t/L \rfloor$ columns, which columns are the Boolean sums of the columns of the given L -subsets. Then \mathbf{x} is an $(\lfloor s/L \rfloor, \lfloor t/L \rfloor, 1)$ -code, and hence

$$N(s, t, L) \geq N(\lfloor s/L \rfloor, \lfloor t/L \rfloor, 1). \quad (2)$$

Inequality (2) allows us to obtain lower bounds for $N(s, t, L)$ using the corresponding lower bounds for the particular case $L=1$.

Definition 2. A matrix \mathbf{x} is called a *superimposed* (\tilde{s}, t) -code of length N , if all the Boolean sums composed of s different columns of \mathbf{x} are different.

Let $\tilde{N}(s, t)$ be the minimal possible number of rows of an (\tilde{s}, t) -code. The definition of (\tilde{s}, t) -codes occurred in connection with the design of screening experiments [3]. The following propositions follow from definitions 1 and 2.

Proposition 4.

$$\tilde{N}(s, t) \geq \left\lceil \log \binom{t}{s} \right\rceil. \quad (3)$$

Asymptotically, when

$$t \rightarrow \infty, \quad s - \text{const}, \quad L - \text{const}, \quad (4)$$

the following inequalities take place

$$N(s, t, L) \geq s \log t(1 + o(1)), \quad (5)$$

$$\tilde{N}(s, t) \geq s \log t(1 + o(1)).$$

These inequalities follow from (1) and (3), accordingly. Let \mathbf{x} be an $N \times t$ -matrix, where $t = \binom{N}{\lfloor N/2 \rfloor}$, all the columns are different and contain the same number $\lfloor N/2 \rfloor$ of 1's. Then \mathbf{x} is a (\tilde{N}, t) -code and a $(1, t, L)$ -code simultaneously. Hence, when $s = 1$, there is equality in (5).

Proposition 5. Any $(s, t, 1)$ -code is an (\tilde{s}, t) -code, and any (\tilde{s}, t) -code is an $(s-1, t, 2)$ -code, i.e.

$$N(s, t, 1) \geq \tilde{N}(s, t) \geq N(s-1, t, 2). \quad (6)$$

Proposition 6. The matrix \mathbf{x} simultaneously satisfies the definitions of the $(s-1, t, 1)$ -codes and the (\tilde{s}, t) -codes iff all the Boolean sums composed of not more than s columns of \mathbf{x} are different.

Inequalities (2) and (6) allow us to obtain another lower bound for $\tilde{N}(s, t)$ by means of any lower bound for $N(\lfloor s-1 \rfloor/2, \lfloor t/2 \rfloor, 1)$. Proposition 6 is very important for the design of screening experiments. This proposition allows us to estimate the minimal length of the code satisfying the condition of proposition 6 in terms of minimal lengths of $(s-1, t, 1)$ -codes and (\tilde{s}, t) -codes.

The following questions are considered below. Some results of [4] improving the lower bounds (1), (3), (5) are formulated in section 2. Random coding bounds are given in Section 3. A special class of superimposed codes that was introduced by Kautz and Singleton is investigated in section 4. This class is an important particular case of $(s, t, 1)$ -codes, because almost all of the known regular constructions belong to this class. Bounds for the optimal length of such codes are given in Section 4.

2. Lower bounds of length of superimposed codes

At first, let us consider the lower bounds for $N(s, t, 1)$, and then using (2) and (6) we generalise these bounds for (s, t, L) -codes, and (\tilde{s}, t, L) -codes. In this section we formulate only the theorems. Their proofs can be found in reference [4]. Theorem 1 was proved by L. A. Bassalygo in 1975.

Theorem 1. (L. A. Bassalygo)

$$N(s, t, 1) \geq \min \{(s+1)(s+2)/2; t\}$$

and, therefore, $N(s, t, 1) = t$ if $s \geq \sqrt{2t}$. In other words, for $s \geq \sqrt{2t}$ no $(s, t, 1)$ -code is better than the trivial one of length $N = t$, whose matrix is diagonal.

Theorem 2. Let $2 \leq s < t$ and $d = N(s-1, t-1)$. Then the length N of any $(s, t, 1)$ -code satisfies the inequality

$$t \leq N + s^2 \sum_{k=1}^{\lfloor (N-d)/s \rfloor} \binom{N}{k+1} / \binom{ks}{k}.$$

We introduce the function

$$f_s(v) = h(v/s) - vh(1/s), \quad s \geq 1,$$

of argument v , $0 < v < 1$, where $h(u)$ is the binary entropy. From theorem 2 follows

Theorem 3. If (4) is true, then

$$N(s, t, 1) \geq K(s) \log t(1 + o(1)),$$

where the sequence $K(1) = 1, K(2), K(3), \dots$ is defined recurrently: $K(s), s \geq 2$, is the unique solution of the equation

$$K(s) = \left[\max_{(*)} f_s(v) \right]^{-1},$$

where the maximum is taken over all v , satisfying the condition

$$0 < v \leq \frac{K(s) - K(s-1)}{K(s)}. \quad (*)$$

The properties of the sequence $K(s), s \geq 2$, are described in theorem 4.

Theorem 4.

$$1) \quad K(2) = \left[\max_{0 < v < 1} f_2(v) \right]^{-1},$$

$$K(s) = \left[f_s \left(\frac{K(s) - K(s-1)}{K(s)} \right) \right]^{-1} \quad s \geq 3.$$

$$2) \quad K(s) \geq s^2/2 \log [e(s+1)/2], \quad s \geq 2.$$

$$3) \quad K(s) = s^2/2 \log s(1 + o(1)), \quad s \rightarrow \infty.$$

For comparison with asymptotic bound (5) the numerical values $K(s)$ for $s = \overline{s, 17}$ are given;

s	$K(s)$	s	$K(s)$	s	$K(s)$	s	$K(s)$
2	3.10628	6	12.0482	10	24.5837	14	40.3950
3	5.01802	7	14.8578	11	28.2402	15	44.8306
4	7.11964	8	17.8876	12	32.0966	16	49.4536
5	9.46603	9	21.1313	13	36.1493	17	54.2612

Inequalities (2) and (6) allow us to generalize the results of theorem 3 for (s, t, L) -codes and (\tilde{s}, \tilde{t}) -codes.

Corollary 1. Under conditions (4)

$$N(s, t, L) \geq c(s, L) \log t(1 + o(1)),$$

$$\tilde{N}(s, t) \geq \tilde{c}(s) \log t(1 + o(1)),$$

where

$$c(s, L) = \max \{s, K(\lfloor s/L \rfloor)\},$$

$$\tilde{c}(s) = \max \{s; K(\lfloor (s-1)/2 \rfloor)\}.$$

From theorem 4 and the numerical values of $K(s)$ we can conclude:

$$1) \quad c(s, 1) = K(s) > s.$$

2) for any $L \geq 2$ there is such a number $s(L)$ that for all $s \geq s(L)$ the factor $c(s, L)$ satisfies $c(s, L) > s$. In particular $s(2) = 16$. 3) for $s \geq 19$, $\tilde{c}(s) > s$ holds. In terms of coding theory this means that for the Boolean model of the design of screening experiments (see [3]) the capacity for average error probability differs from the zero error capacity (the capacity for maximal error probability).

3. Random coding bounds

Let $\mathbf{x} = \|x_i(j)\|$, $i = \overline{1, N}$, $j = \overline{1, t}$ — be a random matrix, all entries of which are independent identically distributed random variables with distribution

$$P\{x_i(j) = 0\} = \beta; \quad P\{x_i(j) = 1\} = 1 - \beta, \quad 0 < \beta < 1.$$

Let the event $\mathcal{A}_N(s, t, L)$ mean that the matrix \mathbf{x} does not satisfy definition 1. It is not hard to see that

$$P\{\mathcal{A}_N(s, t, L)\} \leq \binom{t}{s+L} \binom{s+L}{L} q(s, L, \beta)^N, \quad (7)$$

where

$$q(s, L, \beta) = 1 - \beta^s(1 - \beta^L). \quad (8)$$

Since the right-hand side of (8) takes its minimal value for

$$\beta = \beta_s(L) = \left(\frac{s}{s+L}\right)^{1/L}, \quad (9)$$

then from (7), using random coding arguments (see [5]), it is easy to prove the following statement.

Theorem 5

$$N(s, t, L) \leq \left\lceil C(s, L) \log t + \frac{\log(s!L!)}{\log q(s, L, \beta_s(L))} \right\rceil,$$

$$(1 \leq s < t, L \leq s - t)$$

where notations (8), (9) are used, and

$$C(s, L) = \frac{s+L}{-\log q(s, L, \beta_s(L))}. \quad (10)$$

From theorem 5 follows

Corollary 2. Under conditions (4)

$$N(s, t, L) \leq C(s, L) \log t(1 + o(1)), \quad (11)$$

where $C(s, L)$ is defined by (8)–(10).

For $L=1$

$$C(s, 1) = \frac{s+1}{-\log [1 - s^s/(s+1)^{s+1}]}$$

This means that for $s \rightarrow \infty$

$$C(s, 1) = \frac{e}{\log e} s^2(1 + o(1)) = 1.884 17 \cdot s^2(1 + o(1)).$$

Note, that the corresponding asymptotic formula for factor $c(s, 1)$ of the lower bound in Corollary 1 takes the form:

$$c(s, 1) = K(s) = \frac{s^2}{2 \log s} (1 + o(1)).$$

We give some numerical values of $C(s, 1)$ for comparison with table $K(s)$ from Section 2:

$$C(3.1) = 24.8762 \quad C(4.1) = 40.5487 \quad C(5.1) = 59.9883 \quad C(6.1) = 83.1955$$

If $L = s$, where $l \geq s^{-1}$ is a constant not depending on s , then (11) can be rewritten

$$N(s, t, ls) \leq sg(l) \log t (1 + o(1)), \quad t \rightarrow \infty, \quad (12)$$

where the function $g(l)$ of the parameter $l \geq s^{-1}$ is given by the formula

$$g(l) = (1+l) / -\log \left[1 - \left(\frac{1}{1+l} \right)^{1/l} \frac{l}{1+l} \right].$$

Let l_0 be the unique value of the parameter $l > 0$ for which

$$g(l_0) = \min_{l > 0} g(l).$$

Numerical computations show that $l_0 = 2.235$, and

$$g(l_0) = g(2.235) = 4.269315.$$

For comparison we give some values of the function $g(l)$:

$$\begin{aligned} g(1/2) &= 6.48436 & g(2/3) &= 5.61682 & g(3/4) &= 5.33940 \\ g(1) &= 4.81884 & g(2) &= 4.27895 & g(3) &= 4.33521. \end{aligned}$$

It is easy to see that for $L = ls$, where l is near to l_0 and s is large, the factor $C(s, L) = sg(l)$ in the bound of (12) becomes essentially less than the factor $C(s, 1) = sg(1/s)$.

For the case of (\tilde{s}, \tilde{t}) -code we formulate only the random coding bounds, which are similar to those in theorem 5 and corollary 2. The following statement was proved in [6].

Theorem 6.

$$\tilde{N}(s, t) \leq \lceil \tilde{C}(s) \log t + b_s \rceil,$$

where

$$\tilde{C}(s) = \frac{s+1}{-\log [1 - 2s^s / (s+1)^{s+1}]},$$

$$b_s = \begin{cases} \frac{\log(2^2/s!)}{-\log [1 - 2s^s / (s+1)^{s+1}]}, & \text{if } s = \overline{1, 3}, \\ 0, & \text{if } s \geq 4. \end{cases}$$

Corollary 3. Under conditions (4)

$$\tilde{N}(s, t) \leq \tilde{C}(s) \log t(1 + o(1)).$$

For the comparison of the lower and upper bounds we give some numerical values of $\tilde{C}(s)$ and asymptotic ($s \rightarrow \infty$) formulas:

$$\tilde{C}(2) = 5.92, \quad \tilde{C}(3) = 11.70, \quad \tilde{C}(4) = 19.37, \quad \tilde{C}(5) = 28.92,$$

$$\tilde{C}(s) = \frac{e}{2 \log e} s^2(1 + o(1)) = 0.94208 \cdot s^2(1 + o(1)),$$

$$\tilde{c}(s) = \frac{s^2}{8 \log s} (1 + o(1)).$$

4. Kautz–Singleton codes

The theorems of Section 3 are only theorems of existence. They do not give any method for the construction of the “good” codes. The first question, arising when one tries to apply theorem 5, is the following. How many steps Q of computation one must make to verify, that a given matrix \mathbf{x} with sizes corresponding to the bound of theorem 5, satisfies the definition of an $(s, t, 1)$ -code? If one step is the computation of a Boolean sum and testing of covering of the two binary columns of length N , then the number Q evidently has the order of t^{s+1} . For $t = \overline{10^3}, 10^4$, and $s = 5, \dots, 15$, which occur in applications (see [1]), Q becomes astronomically great

$$Q = 10^{18}, \dots, 10^{64}.$$

Is it possible to find any simple sufficient condition for matrix \mathbf{x} to be an $(s, t, 1)$ -code and the verification of this condition takes essentially less computation steps? One of such evident conditions is given in [1] and formulated below as theorem 7.

Theorem 7. Let \mathbf{x} be a binary $N \times t$ matrix, whose columns have the same number of 1's

$$w = \sum_{i=1}^N x_i(j), \quad j = \overline{1, t}, \quad (13)$$

and for $k \neq j$

$$\lambda_{kj} \triangleq \sum_{i=1}^N x_i(k)x_i(j), \quad (14)$$

$$\lambda = \max_{k \neq j} \lambda_{kj}. \quad (15)$$

Then the matrix x is $(s, t, 1)$ -code for any s , satisfying the inequality

$$s \leq \left\lfloor \frac{w-1}{\lambda} \right\rfloor. \quad (16)$$

The verification of this sufficient condition of Kautz and Singleton, i.e. the computation of the number $\lfloor (w-1)/\lambda \rfloor$ for the matrix x , whose columns have the same number W of 1's takes $Q = \binom{t}{2} \sim t^2$ computation steps (one step is the computation of λ_{k_j}). This number for the above considered values of parameters s and t has the order $Q = 10^6, \dots, 10^8$, which is acceptable from the practical point of view.

Let $1 \leq \lambda \leq w \leq N$ be given integers.

Definition 3. The $N \times t$ matrix x satisfying conditions (13)–(15) is called a $\{t, w, \lambda\}$ -matrix. The minimal possible number of rows of a $\{t, w, \lambda\}$ -matrix is denoted by $n\{t, w, \lambda\}$.

Definition 4. We say that a $\{t, w, \lambda\}$ -matrix is an $(\overline{s, t})$ -code (or an $(s, t, 1)$ -Kautz–Singleton code), if inequality (16) holds.

Denote

$$\bar{N}(s, t) = \min_{(16)} n\{t, w, \lambda\},$$

where the minimum is taken over the parameters w and λ , satisfying (16).

The aim of this section is to investigate the lower and upper bounds for $\bar{N}(s, t)$. It follows from theorem 7 that any $(\overline{s, t})$ -code is also an $(s, t, 1)$ -code. Thus, the lower bound of $N(s, t, 1)$, described in Sections 1 and 2 can be considered as lower bounds of $\bar{N}(s, t)$. For their improvement we need.

Lemma. 1) Let N be the number of rows of a $\{t, w, \lambda\}$ -matrix, then

$$N \geq \frac{tw^2}{\lambda(t-1) + w}, \quad (17)$$

$$\binom{N}{\lambda+1} \geq \binom{w}{\lambda+1} \cdot t. \quad (18)$$

2) If a $\{t, w, \lambda\}$ -matrix is an $(\overline{s, t})$ -code of length $N \leq t-1$, then $w \geq s+1$.

Inequality (17) is an evident consequence of Johnson's bound [7]. Inequality (18) and statement 2) were proved in [1]. From (16), (17) and statement 2) follows

Theorem 8.

$$\bar{N}(s, t) \geq \min \left\{ t, \frac{s(s+1)}{1+s/t} \right\} \geq \min \{t, s^2\}.$$

The bound of theorem 8 is roughly twice better than the lower bound for $\bar{N}(s, t)$, given by theorem 1. From theorem 8 and inequality (1) we have

Corollary 4. For any $s < t$

$$\bar{N}(s, t) \geq d(s, t) \triangleq \max \{d_1(s, t); d_2(s, t)\},$$

$$d_1(s, t) = \left\lceil \log \binom{t}{s} \right\rceil, \quad (19)$$

$$d_2(s, t) = \min \{t, s^2\}.$$

From (16)–(17) it follows that for any (s, t) -code the ratio

$$v \triangleq \frac{w}{N} \leq v(s, t) \triangleq s^{-1} + t^{-1} \quad (20)$$

and the ratio

$$u = \frac{\lambda}{N} \leq \frac{w}{sN} = \frac{v}{s}. \quad (21)$$

If we use some known bounds for the binomial coefficients (see [5], problem 5.8) to estimate the left and right hand sides of (18) and take account of (20)–(21), then it is easy to prove (details are left to reader) the following theorem.

Theorem 9. Let s, t, d be given integers such that

$$3 \leq s < t, \quad v(s, t) \leq 1/2, \quad (22)$$

$$\frac{v(s, t)}{s} + \frac{1}{d} < \frac{1}{es + 1}.$$

Then for any (s, t) -code with length $N \geq d$ the following inequality holds

$$N \geq F(s, t, d) \triangleq \left\lceil \frac{\log t + \frac{1}{2} \log(\pi/4)}{h(v(s, t) \cdot s^{-1} + d^{-1}) - v(s, t)h(s^{-1})} \right\rceil,$$

where $h(u)$ is the binary entropy.

We shall show that bound (19) can be improved by theorem 9, if the value $d = d(s, t)$, defined by (19), satisfies (22). Introduce the recurrent sequence $D_0, D_1, D_2, \dots, D_k = D_k(s, t), \dots$ of integers

$$D_k = \begin{cases} D_{k-1}, & \text{if } F(s, t, D_{k-1}) \leq D_{k-1}, \\ F(s, t, D_{k-1}), & \text{if } F(s, t, D_{k-1}) > D_{k-1}, \end{cases}$$

where $D_0 \triangleq d(s, t)$ is defined by (19). Let $k_0, k_0 = 0, 1, 2, \dots$ be the least non-negative integer k satisfying $D_k = D_{k+1}$. Let

$$D(s, t) \triangleq D_{k_0}(s, t).$$

Then it is evident that $D(s, t) \geq d(s, t)$ and

$$\bar{N}(s, t) \geq D(s, t) \quad (23)$$

holds.

Corollary 5. *As $t \rightarrow \infty$, $s = \text{const}$, the asymptotic inequality*

$$\begin{aligned} \bar{N}(s, t) &\geq K_s \log t (1 + o(1)), \\ K_s &= [h(s^{-2}) - s^{-1}h(s^{-1})]^{-1}, \end{aligned} \quad (24)$$

holds similarly to the lower bound of corollary 1.

Applying the inequality $h(u) \leq u \log(e/u)$, one can verify that $K_s \geq s^2/\log(se)$ from which follows

$$\bar{N}(s, t) \geq \frac{s^2 \log t}{\log(se)} (1 + o(1))$$

Using algebraic methods of coding theory [8], Kautz and Singleton in [1] obtained a family of (\bar{s}, t) -codes with the following parameters. Let $k \geq 2$ be an integer, and $q \geq k-1$ be a prime or prime power. Then

$$\begin{aligned} t &= q^k, & s &= \lfloor q/(k-1) \rfloor, \\ N &= q[1 + (k-1)s], & w &= q+1, & \lambda &= k-1. \end{aligned} \quad (25)$$

For acquaintance with other known constructive classes of (\bar{s}, t) -codes the reader is referred to [1]. We shall give for comparison the numerical table of parameters (25) and of bounds (19), (23) and (24).

k	q	t	s	N (25)	$d(s, t)$ (19)	$D(s, t)$ (23)	K_s (24)
3	23	12 167	11	529	125	235	34.363
4	31	923 581	10	961	177	390	29.505
3	32	32 768	16	1056	256	455	63.318
3	16	4 096	8	272	81	128	20.760
5	16	1 048 576	4	272	76	120	7.437

We complete this section formulating the upper asymptotic bound of $\bar{N}(s, t)$, obtainable by the random coding method, as

$$t \rightarrow \infty, \quad s - \text{const.} \quad (26)$$

Theorem 10. Under condition (26)

$$\bar{N}(s, t) \leq E_s \log t(1 + o(1)),$$

where

$$E_s = \left[\max_{0 < \beta < s^{-1}} E(s, \beta) \right]^{-1}$$

$$E(s, \beta) = h(\beta) - \beta h(s^{-1}) - (1 - \beta)h\left(\frac{\beta(s-1)}{(1-\beta)s}\right).$$

For $s=2$ and $s=3$ the following values are obtained

$$E_2 = 10.6213, \quad E_3 = 28.6090.$$

It can be also shown that

$$E_s = as^2(1 + o(1)), \quad s \rightarrow \infty,$$

where $a = 4.28127$.

5. Open problems

Concluding this paper, we formulate three open problems of superimposed codes theory in the form of questions. We find these problems the most interesting.

- 1) Is it possible to improve the lower bound of theorem 3?
- 2) Is it possible to strengthen inequality (2), and then to improve bound (5) for all $L \geq 2, s \geq 2$?
- 3) Does exist a generalization of the Kautz–Singleton sufficient condition of Section 4 for the case of (s, t, L) -codes, $L \geq 2$?

References

1. Kautz W. H., Singleton R. C. Nonrandom Binary Superimposed Codes. IEEE Trans. Inform. Theory, 1964, 10, 363–377.
2. Dyachkov A. G., Rykov V. V. One application of codes for multiple access channel to ALOHA-system. 6-th All-Union Seminar on Computer Networks. Papers, 1981, v. 4, pp. 18–24.

3. *Dyachkov A. G.* Error probability bounds for two models of randomized design of screening experiments. *Problem of information transmission*, 1979, **15**, 4, 17–31. (Russian).
4. *Dyachkov A. G., Rykov V. V.* Length bounds of superimposed codes. *Problems of information transmission* 1982, **18**, 3, 7–13. (Russian).
5. *Gallager R. G.* *Information theory and reliable communication*, J. Wiley, New York 1968.
6. *Malyutov M. B.* On planning screening experiments, *Proceedings of the 1975 IEEE-USSR Joint Workshop on Inform. Theory*. N. Y., IEEE Inc. 1976, 144–147.
7. *Johnson S. M.* A new upper bound for error correcting codes, *IEEE Trans. Inform. Theory*, 1962, **8**, 203–207.
8. *Berlecamp E. R.* *Algebraic coding theory*, McGraw-Hill Book Company, New York 1968.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H - 1111 Budapest XI., Stoczek u. 2, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper with wide margins (4–5 cm), should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice-versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме – реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в трех экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 отисков их статей.

Рукописи непринятых статей возвращаются авторам.

052.812

CONTENTS · СОДЕРЖАНИЕ


<i>Дьячков А. Г., Рыков В. В.</i> Обзор теории дизъюнктивных кодов (<i>Dyachkov, A. G., Rykov, V. V.</i> : A survey of superimposed code theory)	229
<i>Bocharov, P. P., Albores, F. J.</i> : On two-node exponential queueing network with internal losses or blocking (<i>Бочаров П. П., Альборес Ф. Х.</i> О двухузловой экспоненциальной сети массового обслуживания с внутренними потерями или блокировками)	243
<i>Gabasov, R., Kirillova, F. M., Kostyukova, O. I.</i> : Dual algorithm of optimization of a linear dynamic system (<i>Габасов Р., Кириллова Ф. М., Костюкова О. И.</i> Двойственный алгоритм оптимизации линейной динамической системы)	253
<i>Maršik, J.</i> : A new conception of digital adaptive PSD control (<i>Маршик Я.</i> Новая концепция цифрового адаптивного регулирования ПСД действия)	267
<i>Muntean, I.</i> : Simultaneous optimization of maintenance and replacement policy for machines (<i>Мунтян И.</i> Одновременная оптимизация стратегии и замены машин)	279
<i>Šebek, M.</i> : Direct polynomial approach to discrete-time stochastic tracking (<i>Шебек М.</i> Прямой полиномиальный подход к дискретным стохастическим сервосистемам)	293

316.920

VOL. 12 • NUMBER 5
TOM HOMEP

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY



ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1983

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H-1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 0BW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1983 Subscription Rate US Dollars 140.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

N. N. KRASOVSKII (USSR)

G. BOGNÁR (HUNGARY)

COORDINATING EDITORS

USSR

S. V. EMEL'YANOV

Е. П. ПОПОВ

V. S. PUGACHEV

V. I. SIFOROV

E. D. TERYAEV

HUNGARY

T. VÁMOS

L. VARGA

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJČ

РЕДАКТОРЫ ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

Г. БОГНАР (ВНР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

Е. Д. ТЕРЯЕВ

ВНР

Т. ВАМОШ

Л. ВАРГА

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

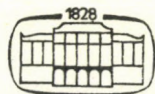
Л. КЕВИЦКИ

Я. ҚОЧИШ

ЧССР

И. БЕНЕШ

В. СТРЕЙЦ



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

PRINTED IN HUNGARY
Akadémiai Nyomda, Budapest

TUDOMÁNYOS KÖNYVTÁRA

AN ABSTRACT SOURCE-CHANNEL TRANSMISSION THEOREM*

I. CSISZÁR
(Budapest)

(Received 1 October, 1982)

An abstract graph-theoretic model of information transmission over a noisy channel is considered and an analogue of Shannon's source-channel transmission theorem is proved.

Introduction

The basic mathematical structure of the Shannon theory of discrete memoryless systems is more combinatorial than probabilistic, at least as far as block coding without feedback is considered. The reason is that sets in product spaces can be decomposed into "exponentially few" subsets according to types, resp. joint types, whereby the relevant distributions will be uniform on these subsets and the problem of bounding probabilities boils down to counting. An approach to exponential error bounds for DMC's using this idea (rather than, e.g., Chernoff bounding techniques) was first developed by Csiszár, Körner and Marton [5]. Their approach was further elaborated and applied to a variety of two-terminal and multi-terminal coding problems in Csiszár and Körner [3] and other recent works of these authors.

As regards coding theorems without exponential error bounds, their inherently combinatorial nature transpires already with the cruder "typical sequences" approach. An elegant treatment of this subject in the framework of hypergraph theory has been put forward by Ahlswede [1]. In order to more deeply understand the mathematical structure of coding theorems, he proposed to consider also "abstract versions" of them; in particular, he proved "channel coding theorems" for general bipartite graphs. The power of elementary graph-theoretic methods for proving coding theorems with exponential error bounds has been demonstrated by Csiszár and Körner [4].

* This work was done while the author was visiting professor at Universität Bielefeld, FRG, sponsored by Deutsche Forschungsgemeinschaft.

The model

Now we sketch an "abstract" source-channel transmission model in the spirit of Ahlswede [1]. An abstract channel is a bipartite graph $(\mathcal{B}, \mathcal{C}, \mathcal{F})$ where \mathcal{B} and \mathcal{C} are finite sets and \mathcal{F} is a subset of $\mathcal{B} \times \mathcal{C}$. The vertices $b \in \mathcal{B}$, resp. $c \in \mathcal{C}$ of this graph are interpreted as the possible channel inputs, resp. outputs, while the edges $(b, c) \in \mathcal{F}$ specify the possible input-output pairs. Up to a difference in the interpretation of vertices and edges, an abstract source is also a bipartite graph, say $(\mathcal{A}, \mathcal{D}, \mathcal{G})$. Here the vertices in \mathcal{A} are the possible messages, those in \mathcal{D} the possible reproductions of messages, and an edge $(a, d) \in \mathcal{G}$ indicates that d is an acceptable reproduction of a .

To define the transmissibility of an abstract source over an abstract channel, for given mappings $f: \mathcal{A} \rightarrow \mathcal{B}$ ("encoder") and $\varphi: \mathcal{C} \rightarrow \mathcal{D}$ ("decoder") we shall call a four-tuple $(a, b, c, d) \in \mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \mathcal{D}$ a path, resp. a circuit if $b = f(a)$, $(b, c) \in \mathcal{F}$, $d = \varphi(c)$, resp., in addition, also $(a, d) \in \mathcal{G}$.

Definition. Designate the ratio of non-circuits among the paths starting at $a \in \mathcal{A}$ by $p(a) = p(a, f, \varphi)$. An abstract source is ε -transmissible resp. average ε -transmissible over an abstract channel if for suitable mappings $f: \mathcal{A} \rightarrow \mathcal{B}$, $\varphi: \mathcal{C} \rightarrow \mathcal{D}$ we have $p(a, f, \varphi) \leq \varepsilon$ for every $a \in \mathcal{A}$, resp. $\sum_{a \in \mathcal{A}} p(a, f, \varphi) \leq \varepsilon |\mathcal{A}|$. In the particular case $\varepsilon = 0$, i.e., when all paths are actually circuits, we speak of strict transmissibility.

For convenience, we shall consider regular bipartite graphs only. Given a vertex a of an arbitrary graph, we designate by $\mathcal{S}(a)$ the set of vertices connected with a by an edge. A bipartite graph $(\mathcal{B}, \mathcal{C}, \mathcal{F})$, say, is called regular if $\mathcal{S}(b)$ has the same cardinality for every $b \in \mathcal{B}$, and the same holds for $\mathcal{S}(c)$, $c \in \mathcal{C}$. Equivalently, this means that there exists a positive number $r(\mathcal{F})$ such that

$$\frac{|\mathcal{B}|}{|\mathcal{S}(c)|} = \frac{|\mathcal{C}|}{|\mathcal{S}(b)|} = r(\mathcal{F}) \quad \text{for each } b \in \mathcal{B}, c \in \mathcal{C}. \quad (1)$$

The typical example motivating our "abstract" considerations is $\mathcal{B} = \mathcal{T}_P \subset \mathcal{X}^n$, $\mathcal{C} = \mathcal{T}_Q \subset \mathcal{Y}^n$ (the sequences in \mathcal{X}^n resp. \mathcal{Y}^n of some fixed type P , resp. Q , cf. [3]), with $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}$ iff \mathbf{x} and \mathbf{y} have a prescribed joint type. Then, visualizing this joint type as the joint distribution of dummy RV's X and Y , we have that $\log r(\mathcal{F})$ equals $nI(X \wedge Y)$, up to an error term of order $\log n$. Thus, for general regular bipartite graphs, $\log r(\mathcal{F})$ may be interpreted as an abstract analogue of mutual information. This interpretation is further supported by the "abstract source resp. channel coding theorems" stated in the next Lemma. More general versions of these appear in Ahlswede [1]. For the reader's convenience, we shall nevertheless give the simple proof of this Lemma.

Lemma. For a regular abstract source $(\mathcal{A}, \mathcal{D}, \mathcal{G})$ there exist elements d_1, \dots, d_N of \mathcal{D} , with $\triangleq \lfloor r(\mathcal{G}) \ln |\mathcal{A}| \rfloor$, such that

$$\mathcal{A} = \bigcup_{i=1}^N \mathcal{S}(d_i). \quad (2)$$

Further, for a regular abstract channel $(\mathcal{B}, \mathcal{C}, \mathcal{F})$ there exist elements b_1, \dots, b_M of \mathcal{B} , with $M \triangleq \left\lceil r(\mathcal{F}) \frac{\varepsilon}{4} \right\rceil$, such that

$$\frac{|\mathcal{S}(b_i) \cap \bigcup_{j \neq i} \mathcal{S}(b_j)|}{|\mathcal{S}(b_i)|} \leq \varepsilon, \quad i = 1, \dots, M. \tag{3}$$

Here $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer, i.e., $\lfloor t \rfloor = k$ iff $k - 1/2 < t \leq k + 1/2$.

Proof. A fixed $a \in \mathcal{A}$ is not contained in the union of the sets $\mathcal{S}(d_i)$ iff $\mathcal{S}(a)$ contains neither d_i . Choosing d_1, \dots, d_N at random, this event has probability $(1 - 1/r(\mathcal{G}))^N$. Hence $|\mathcal{A}|(1 - 1/r(\mathcal{G}))^N$ is an upper bound to the probability that (2) does not hold. Obviously, this bound is less than 1 iff $N \geq r(\mathcal{G}) \ln |\mathcal{A}|$. The same can be checked by an easy calculation also if $N \geq r(\mathcal{G}) \ln |\mathcal{A}| - 1/2$, providing $\ln |\mathcal{A}| > 1$. Since the case $|\mathcal{A}| = 2$ is trivial, this proves the first assertion. To prove the second one, choose $2M$ elements b_1, \dots, b_{2M} from \mathcal{B} at random. Since $c \in \mathcal{S}(b_i) - \mathcal{S}(b_j)$ iff both b_i and b_j are in $\mathcal{S}(c)$, the probability of this event is $r(\mathcal{F})^{-2}$, cf. (1). Fixing i and summing for all $j \neq i$ and $c \in \mathcal{C}$, it follows that the expectation of $|\mathcal{S}(b_i) \cap \bigcup_{j \neq i} \mathcal{S}(b_j)|$ is less than $(2M - 1)|\mathcal{C}|r(\mathcal{F})^{-2}$. Hence, using (1) once more, we get for

$$S_i \triangleq \frac{|\mathcal{S}(b_i) \cap \bigcup_{j \neq i} \mathcal{S}(b_j)|}{|\mathcal{S}(b_i)|}$$

that

$$E\left(\frac{1}{2M} \sum_{i=1}^{2M} S_i\right) \leq \frac{2M - 1}{r(\mathcal{F})} \leq \frac{\varepsilon}{2}$$

if $M = \left\lceil r(\mathcal{F}) \frac{\varepsilon}{4} \right\rceil$. Thus for some particular choice of b_1, \dots, b_{2M} , the average within the expectation is not greater than $\frac{\varepsilon}{2}$. Supposing as we may that $S_1 \leq S_2 \leq \dots \leq S_{2M}$, this implies $S_i \leq \varepsilon$ for $i = 1, \dots, M$, which is (3).

Now we can prove the following "abstract source-channel transmission theorem".

Theorem. A sufficient condition for the ε -transmissibility of a regular abstract source $(\mathcal{A}, \mathcal{D}, \mathcal{G})$ over a regular abstract channel $(\mathcal{B}, \mathcal{C}, \mathcal{F})$ is

$$r(\mathcal{F}) \geq r(\mathcal{G}) \cdot \frac{4}{\varepsilon} \ln |\mathcal{A}|. \tag{4}$$

Further, a necessary condition, even for average ε -transmissibility, is

$$r(\mathcal{F}) \geq r(\mathcal{G}) \cdot (1 - \varepsilon). \quad (5)$$

Proof. Consider d_1, \dots, d_N and b_1, \dots, b_M as in the Lemma. Notice that condition (4) implies $N \leq M$. Choose to each $a \in \mathcal{A}$ a d_i connected with it by an edge; this is possible by (2). Then define $f: \mathcal{A} \rightarrow \mathcal{B}$ by letting $f(a) \triangleq b_i$. Further, for each $c \in \mathcal{C}$ connected with exactly one b_i , $1 \leq i \leq M$, define $\varphi(c) \triangleq d_i$; otherwise φ may be arbitrary. Then, with the notation in the definition of ε -transmissibility, we have by (3)

$$p(a, f, \varphi) \leq \frac{|\mathcal{S}(b_i) \cap \bigcup_{j \neq i} \mathcal{S}(b_j)|}{|\mathcal{S}(b_i)|} \leq \varepsilon,$$

proving the direct part of the Theorem.

For the converse, suppose that some mappings $f: \mathcal{A} \rightarrow \mathcal{B}$, $\varphi: \mathcal{C} \rightarrow \mathcal{D}$ meet the condition

$$\sum_{a \in \mathcal{A}} p(a, f, \varphi) \leq \varepsilon |\mathcal{A}|$$

of average ε -transmissibility. Notice that $1 - p(a, f, \varphi)$ equals, by definition, the number of circuits among the paths starting at a , divided by the number of all such paths. Since by (1) the latter is $|\mathcal{S}(f(b))| = |\mathcal{C}|/r(\mathcal{F})$, the number of all circuits is

$$\frac{|\mathcal{C}|}{r(\mathcal{F})} \sum_{a \in \mathcal{A}} (1 - p(a, f, \varphi)) \geq \frac{|\mathcal{C}|}{r(\mathcal{F})} (1 - \varepsilon) |\mathcal{A}|.$$

Among the paths ending at any fixed $d = \varphi(c)$, at most $|\mathcal{S}(d)| = |\mathcal{A}|/r(\mathcal{G})$ can be circuits. Thus the last bound implies

$$|\mathcal{C}| \cdot \frac{|\mathcal{A}|}{r(\mathcal{G})} \geq \frac{|\mathcal{C}|}{r(\mathcal{F})} (1 - \varepsilon) |\mathcal{A}|,$$

proving (5).

The Theorem is satisfactory when ε is "not very small". On the other hand, having in mind the typical example mentioned after eq. (1), there is a large gap between the sufficient and necessary conditions when ε tends to zero exponentially as $n \rightarrow \infty$. For small ε , an improvement of the direct part of the Theorem via an improvement of the second assertion of the Lemma is sometimes possible, using an expurgation argument in the selection of the b_i 's. Still, asymptotically optimal results need not be attainable by composing good "source codes" and "channel codes". E.g., for $(\mathcal{A}, \mathcal{D}, \mathcal{G}) = (\mathcal{B}, \mathcal{C}, \mathcal{F})$ strict transmissibility is attained "without coding" (compare this with the example in Csizsár [2], Section 4).

References

1. Ahlswede, R., Coloring hypergraphs: A new approach to multi-user source coding I-II. *Journal of Combinatorics, Information and System Sciences*, Vol. 4 (1979) pp. 76-115; Vol. 5 (1980) pp. 220-268.
2. Csiszár, I., On the error exponent of source-channel transmission with a distortion threshold. *IEEE Trans. Vol. IT-28* (1982).
3. Csiszár, I., Körner, J., *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, Budapest — Academic, New York, 1981.
4. Csiszár, I., Körner, J., Graph decomposition: a new key to coding theorems. *IEEE Trans. Vol. IT-27* (1981) pp. 5-12.
5. Csiszár, I., Körner, J., Marton, K., A new look at the error exponent of a discrete memoryless channel. Paper presented at the International Symposium on Information Theory, Cornell Univ., Ithaca, NY, 1977.

Об абстрактной теореме рассеяния в передающих устройствах

И. ЧИСАР
(Будапешт)

Рассматривается абстрактная графовая модель рассеяния информации в линиях с шумом и доказывается теорема аналогичная теореме Шеннона о рассеивании в передающих устройствах.

I. Csiszár
Mathematical Institute of the Hungarian Academy of Sciences
H-1053 Budapest, Reáltanoda u. 13-15.
Hungary

ADAPTIVE TRANSFORM PICTURE CODING FOR ROBOT VISION SYSTEM

F. ŠOLC, J. HALABALA
(Brno)

(Received 15 September, 1982)

A multipurpose digital scanning television system was developed. For effective storage of pictures in digital form an adaptive transform picture coding system was chosen. The picture is divided into subpictures at first. Individual subpictures are transformed by fast Hadamard transformation. Hadamard coefficients are quantized and encoded according to subpicture properties. Results of this coding system are demonstrated by several examples.

A chain for digital processing of TV picture was developed. This chain consists of TV camera, sampling unit, universal minicomputer and special TV monitor which is connected to the computer. This system enables processing of TV picture, which is digitized to an image consisting of 128×128 array of pels (picture elements), each of them is quantized to 16 grey levels. The chain is to be used for robot vision system, so the problem of effective encoding of image is vital. When one uses direct digital code for encoding of this image it takes $128 \times 128 \times 4 = 65\,536$ bits, i.e. 4 096 words of 16 bits from memory of the minicomputer (maximum capacity of the minicomputer's memory is 32k words). It is known from practice that pictorial data contain significant structure, that is why the entropy of typical pictures is much lower than the maximum possible entropy, which is in this case 65 536 bits. According to [1], the structure inherent in pictorial data is such that the entropy of typical pictures is about one bit per pel, this value depends highly on the class "typical pictures" of course. Thus it should be possible to encode our picture into 128×128 bits at least without any loss of information. Many coding strategies and techniques exist. If the class of typical pictures were properly defined it would be possible to choose a coding technique which would match the pictures and would yield the most effective encoding. Unfortunately, the definition of the class of typical pictures is rather difficult and sometimes very vague. In this case it is necessary to choose a technique which is sufficiently universal. So it was decided to use transform picture coding technique and, from several transformations, the Hadamard transformation was chosen for its relative simplicity of computation.

It is known [2] that the Hadamard transformation of picture can be interpreted as two-dimensional transformation of $n \times n$ array of pels of original picture X into $n \times n$ array of coefficients of transformed picture Y . One can write this transformation and its inverse in form [3]

$$\mathbf{Y} = \mathbf{H}\mathbf{X}\mathbf{H} \quad (1)$$

$$\mathbf{X} = \frac{1}{n^2} \mathbf{H}\mathbf{Y}\mathbf{H} \quad (2)$$

where \mathbf{H} is an ordered $n \times n$ Hadamard matrix. Another form of this transformation is [4]

$$y_{kl} = \sum_{i=1}^n \sum_{j=1}^n a_{kl ij} x_{ij} \quad (3)$$

$$\mathbf{X} = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n y_{kl} \mathbf{a}_{kl} \quad (4)$$

where \mathbf{a}_{kl} is an $n \times n$ matrix whose elements are $+1$ or -1 . Matrices \mathbf{a}_{kl} are basis pictures and (4) is interpreted as a series expansion of the $n \times n$ picture \mathbf{X} onto n^2 $n \times n$ basis pictures with the y_{kl} , $l=1, 2, \dots, n$ the coefficients of the expansion. As it was shown in [4], the coefficients of the series have unequal variance and it is possible to truncate the series (4) after the first η terms, whose coefficients have the highest variance. Then one can write

$$\mathbf{X} \approx \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^{\eta} y_{kl} \mathbf{a}_{kl} \quad (5)$$

After such truncation the coefficients y_{kl} are encoded and stored instead of encoding and storing of picture \mathbf{X} . For the sake of computing simplicity and reliability, subpictures of 8×8 pels from picture of 128×128 pels were chosen to be transformed and encoded. For this case the Hadamard matrix is

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix} \quad (6)$$

and $64, 8 \times 8$ basis pictures are in Fig. 1, where black represents $+1$ and white -1 . The structure of the Hadamard matrix (6) allows us to calculate transformation (1) and its

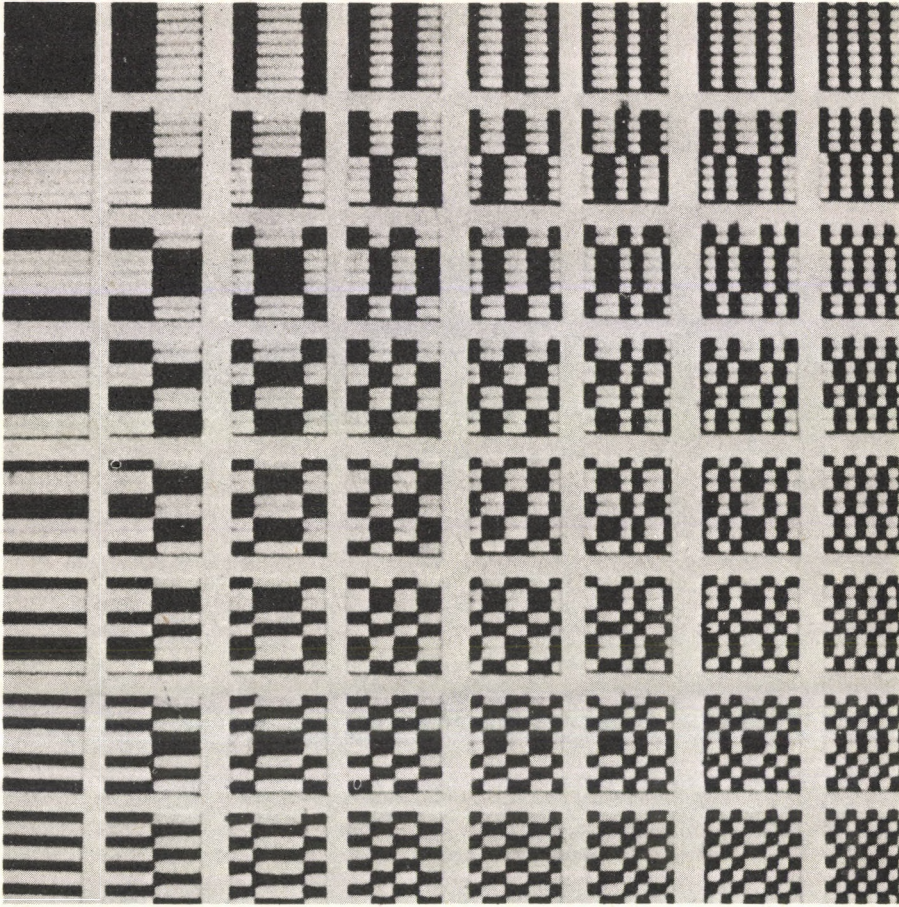


Fig. 1. The 64×8 sequency ordered basis pictures for the Hadamard transformation

inverse (2) by means of a fast algorithm which uses only 384 additions instead of 1024 additions by the classical way. According to [5] variance of the coefficients y_{kl} for a wide class of pictures can be estimated by the following formula for standard deviation

$$\sigma_{kl} = S \exp\left(-\frac{k^2 + l^2}{p}\right) \quad (7)$$

where S and p are coefficients which depend on the given class of pictures. That is why it was decided to truncate series (4) by "zonal filtering". In (4) only the coefficients y_{kl} having ones in the corresponding entries in matrices (8) are retained.

$$\begin{array}{cc}
 \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
 \text{(a)} & \text{(b)} \\
 \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
 \text{(c)} & \text{(d)}
 \end{array} \tag{8}$$

After several tests one can recommend zonal filtering according to matrix (8a). Now each coefficient of the resulting series (5) must be quantized and encoded. From (3) we can see that the values of the coefficients y_{kl} can be within the interval $\langle -480; +480 \rangle$ for $k, l = 2, 3, \dots, n$ and within the interval $\langle 0; 960 \rangle$ for $k = l = 1$ in this case (subpicture 8×8 and 16 grey levels). Since the variances of the coefficients vary widely, it would be inefficient to use the same quantizer for each coefficient. Because each coefficient is a linear combination of original pels, the central limit theorem indicates that the distribution of individual coefficients tends toward Gaussian. That is why each Gaussian distributed y_{kl} is transformed into a uniformly distributed \hat{y}_{kl} according to the relation

$$\begin{aligned}
 \hat{y}_{kl} &= \text{int} \left(n_{kl} \text{erf} \left(\frac{y_{kl}}{\sigma_{kl}} \right) \right); & \text{for } k, l = 1, 2, \dots, n \\
 \hat{y}_{11} &= \text{int} \left(n_{11} \text{erf} \left(\frac{y_{11} - 480}{\sigma_{11}} \right) \right) & (k, l) \neq (1, 1)
 \end{aligned} \tag{9}$$

where n_{kl} is a number of levels available for quantizing the individual coefficient y_{kl} , $\text{int}(\cdot)$ and $\text{erf}(\cdot)$ are integer and error functions respectively. After quantization into n_{kl} levels each coefficient \hat{y}_{kl} is encoded into direct binary code, thus each n_{kl} is chosen to be a natural power of base two. Thus the total number of bits into which the picture is encoded is $256 \sum_{l=1}^8 \sum_{k=1}^8 \log_2 n_{kl}$. Now the problem is to determine the tradeoff among S , p , allocation of quantization levels and the total number of bits into which the picture is encoded, i.e. the compression ratio. Some hints for solving this problem are in [4]. By means of the system mentioned above subjective and objective tests (mean square error between the original and the reconstructed picture was used for objective tests) were performed. From these tests one can make the following conclusions.

As to quantization levels, quantization according to the following matrices of n_{kl} can be recommended.

$$\begin{matrix}
 \begin{bmatrix}
 32 & 16 & 16 & 8 & 4 & 0 & 0 & 0 \\
 16 & 16 & 16 & 8 & 4 & 0 & 0 & 0 \\
 16 & 16 & 8 & 8 & 0 & 0 & 0 & 0 \\
 8 & 8 & 8 & 4 & 0 & 0 & 0 & 0 \\
 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix} & \text{(a)} &
 \begin{bmatrix}
 64 & 32 & 16 & 8 & 4 & 0 & 0 & 0 \\
 32 & 16 & 8 & 8 & 4 & 0 & 0 & 0 \\
 16 & 8 & 8 & 4 & 0 & 0 & 0 & 0 \\
 8 & 8 & 4 & 4 & 0 & 0 & 0 & 0 \\
 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix} & \text{(b)} \\
 \\
 \begin{bmatrix}
 128 & 32 & 16 & 8 & 4 & 0 & 0 & 0 \\
 32 & 32 & 16 & 8 & 4 & 0 & 0 & 0 \\
 16 & 16 & 16 & 8 & 4 & 0 & 0 & 0 \\
 8 & 8 & 4 & 4 & 0 & 0 & 0 & 0 \\
 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix} & \text{(c)} &
 \begin{bmatrix}
 128 & 64 & 32 & 16 & 8 & 0 & 0 & 0 \\
 64 & 64 & 16 & 16 & 8 & 0 & 0 & 0 \\
 32 & 16 & 16 & 8 & 0 & 0 & 0 & 0 \\
 16 & 16 & 8 & 8 & 0 & 0 & 0 & 0 \\
 8 & 8 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix} & \text{(d)} \\
 \end{matrix} \tag{10}$$



Fig. 2. The face of a man—original from 128×128 pels, 16 grey levels

Matrix (10b) can be recommended in general. Maximum recommended compression ratio is 4:1, i.e. the original picture is encoded with 1 bit per pel. Recommended values of S and p are $S \in \langle 120; 200 \rangle$ $p \in \langle 8; 13 \rangle$. For general case $S = 160$, $p = 10$ was chosen.

The compression ratio can be increased by means of an adaptive technique. The adaptive technique makes the transform coder to adapt to the subpicture structure and to choose the mode that is most efficient for that subpicture from the allowed number of modes. One can suppose that the picture will contain larger areas of constant brightness, e.g. picture background, thus corresponding subpictures are coded by means of y_{11} coefficient only, other coefficients are discarded. Subpictures of variable brightness are encoded as described before. As a criterion for decision between these two modes the variance of pel brightness in the subpicture was chosen. After subjective and objective tests with several pictures, the variance within the interval $\langle 1; 13 \rangle$ can



Fig. 3. Reconstructed pictures of the face of a man from fig. 2.

(a) Compression ratio 4:1, $S=160$, $p=10$, zonal filtering (8a), quantization (10b), nonadaptive algorithm.

be recommended as a decision level. Subpictures with lower variance are encoded with the help of y_{11} only.

Decoding of encoded pictures is straightforward. As the mapping (9) is not one-to-one, one can calculate only $\bar{y}_{kl} \approx y_{kl}$ according to the following formula

$$\bar{y}_{kl} = \text{int}(\sigma_{kl}(\text{erf}^{-1}(\hat{y}_{kl} + 0.5)) + 0.5); \quad \text{for } k, l = 1, 2, \dots, n$$

$$(k, l) \neq (1, 1)$$

$$\bar{y}_{11} = \text{int}(\sigma_{11}(\text{erf}^{-1}(\hat{y}_{11} + 0.5)) + 480.5).$$
(11)

Corrective factors 0.5 are used to minimize the error of inverse transformation. The last



(b) Compression ratio 4:1, $S=200$, $p=8$, zonal filtering (8a), quantization (10b), nonadaptive algorithm.

step in calculation of the original picture $\bar{X} \approx X$ is inverse Hadamard transformation

$$X \approx \bar{X} = \frac{1}{n^2} H \bar{Y} H \quad (12)$$

where \bar{Y} is a matrix of coefficients \bar{y}_{kl} . The elements of matrix \bar{X} need not be integers and sometimes can be outside of interval $\langle 0; 15 \rangle$. Thus, finally, each element of matrix \bar{X} is rounded to the closest integer according to the function $\text{int}(x+0.5)$, and all results which are outside of the interval $\langle 0; 15 \rangle$ are taken as 0 or 15, respectively.

Figures 2-5 demonstrate achieved results.



(c) Compression ratio 6.45:1, $S = 160$, $p = 10$, zonal filtering (8a), quantization (10b), adaptive algorithm, decision level of variance 1.3

Conclusions

Some results of current research in the area of Digital Image Processing for Robot Vision System were presented. All mentioned work was done with the help of a universal minicomputer and specially developed chain of digital-analogue TV. The encoding process takes approximately 3 minutes with the used equipment, the same concerns the decoding process. The low operating speed of the minicomputer is the main culprit of such a long time. Thus the coding and decoding process cannot be used in real time robot vision system yet, nevertheless the system can be used for storing of templates in robot's memory in contemporary form. If the class of the pictures being more precisely specified, one can reach better compression ratio and speed.

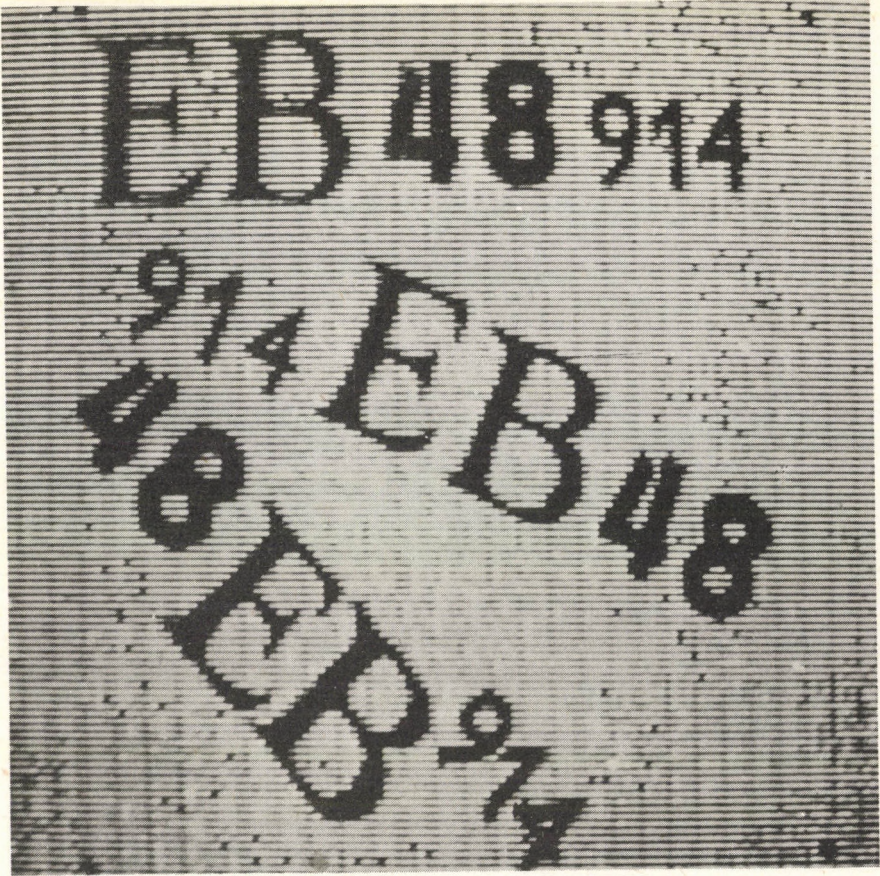


Fig. 4. Text—original from 128×128 pels, 16 grey levels

Acknowledgement

The authors wish to thank Assistant Professor P. Vavřín, Head of research team, and the staff of the Department, who keep things going so that this paper could be written.

References

1. Habibi, A., Robinson, G. S., A Survey of Digital Picture Coding, IEEE Computer, vol. 7, pp. 23–34, May 1974.
2. Pratt, W. K., Digital Image Processing, J. Wiley, N. Y., 1978.
3. Huang, T. S., Picture Processing and Digital Filtering, Springer Verlag, Berlin, 1975.
4. Wintz, P. A., Transform Picture Coding, Proc. IEEE, vol. 60, pp. 809–820, July 1972.
5. Pratt, W., Kane, J., Andrews, H., Hadamard Transform Image Coding, Proc. IEEE, vol. 57, pp. 58–68, Jan. 1969.

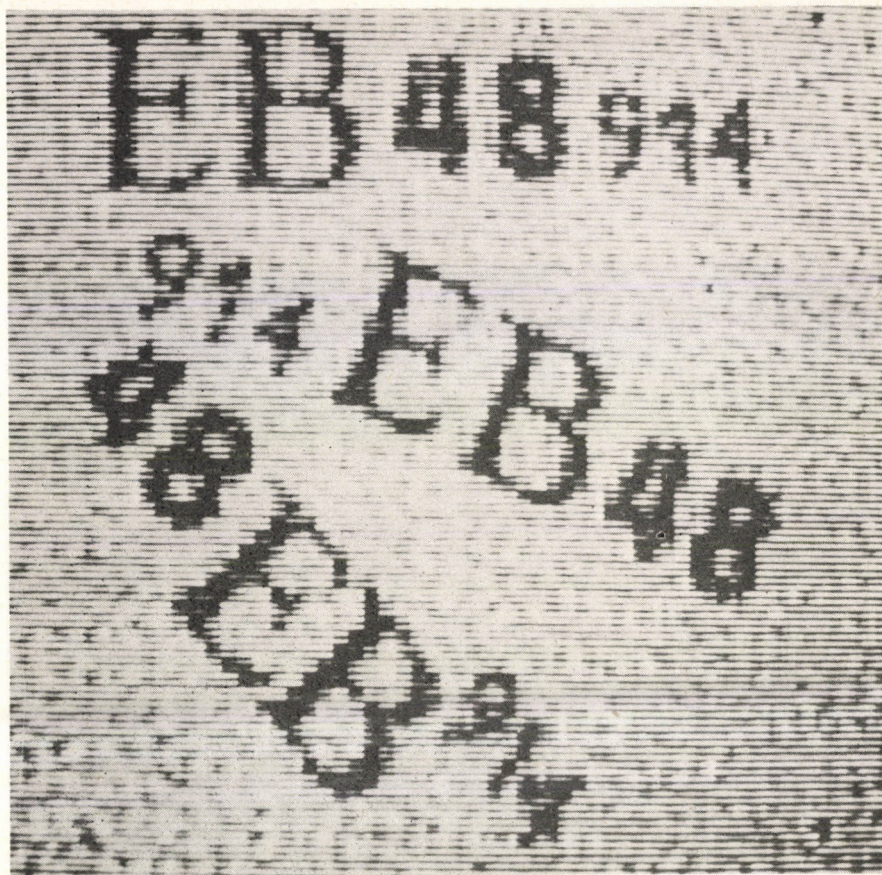


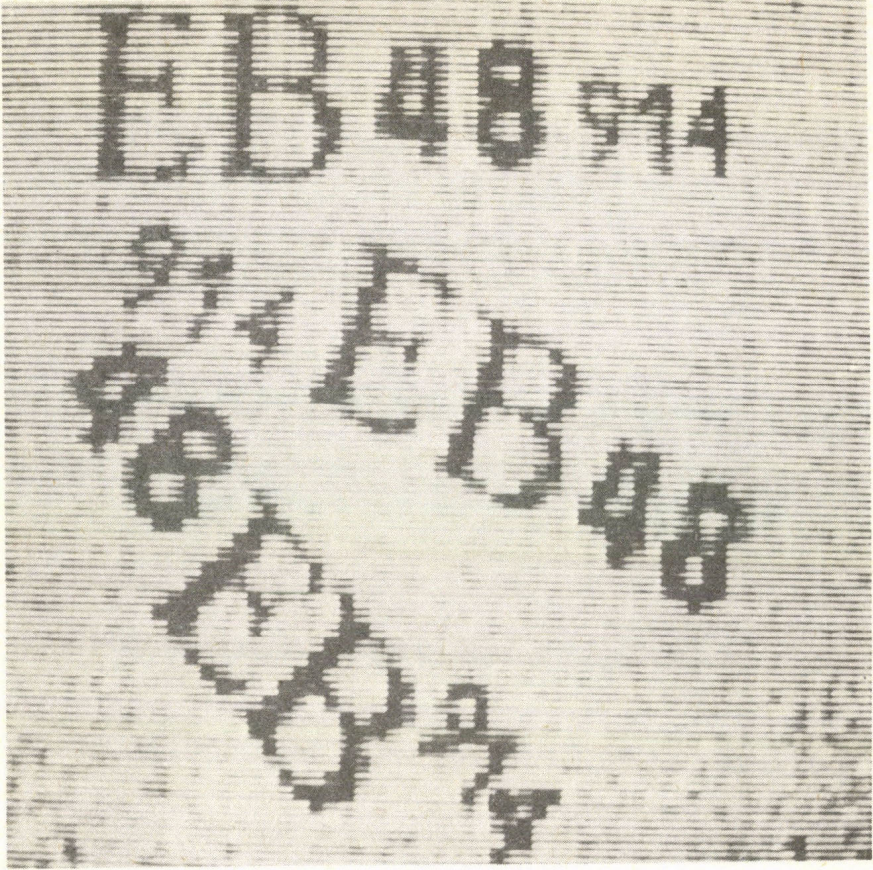
Fig. 5. Reconstructed pictures of the text from fig. 4.

(a) Compression ratio 4:1, $S = 160$, $p = 10$, zonal filtering (8a), quantization (10b), nonadaptive algorithm.

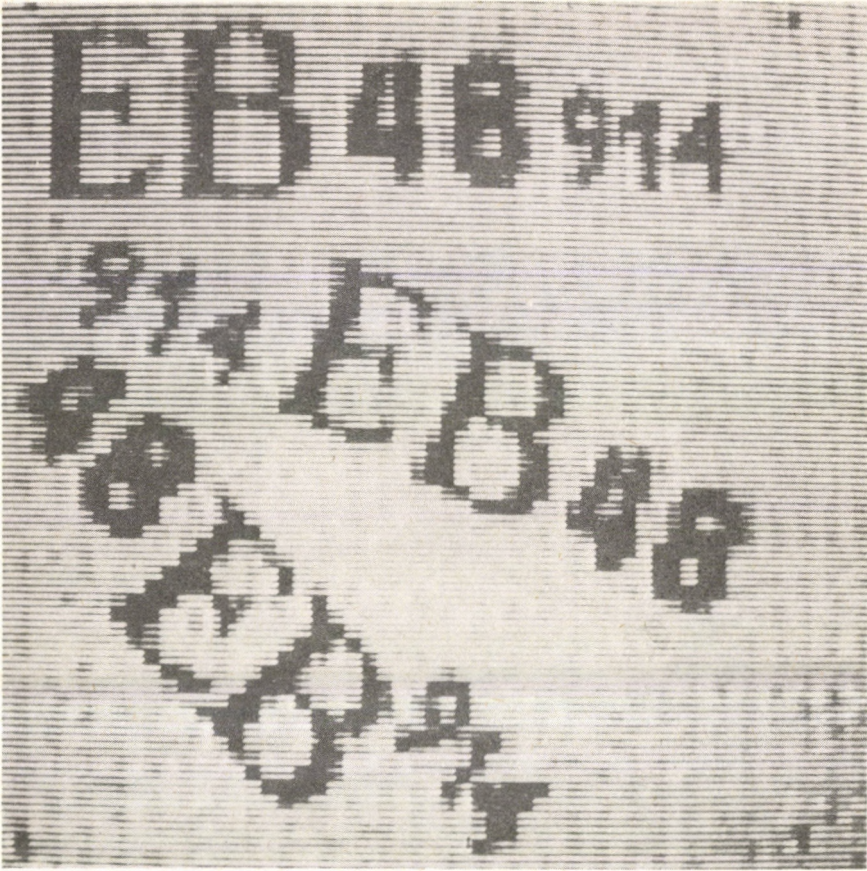
Адаптивная кодировка изображений с помощью двухмерного преобразования для визуальной системы робота

Ф. ШОЛЫЦ И ХАЛАБАЛА
(Брно)

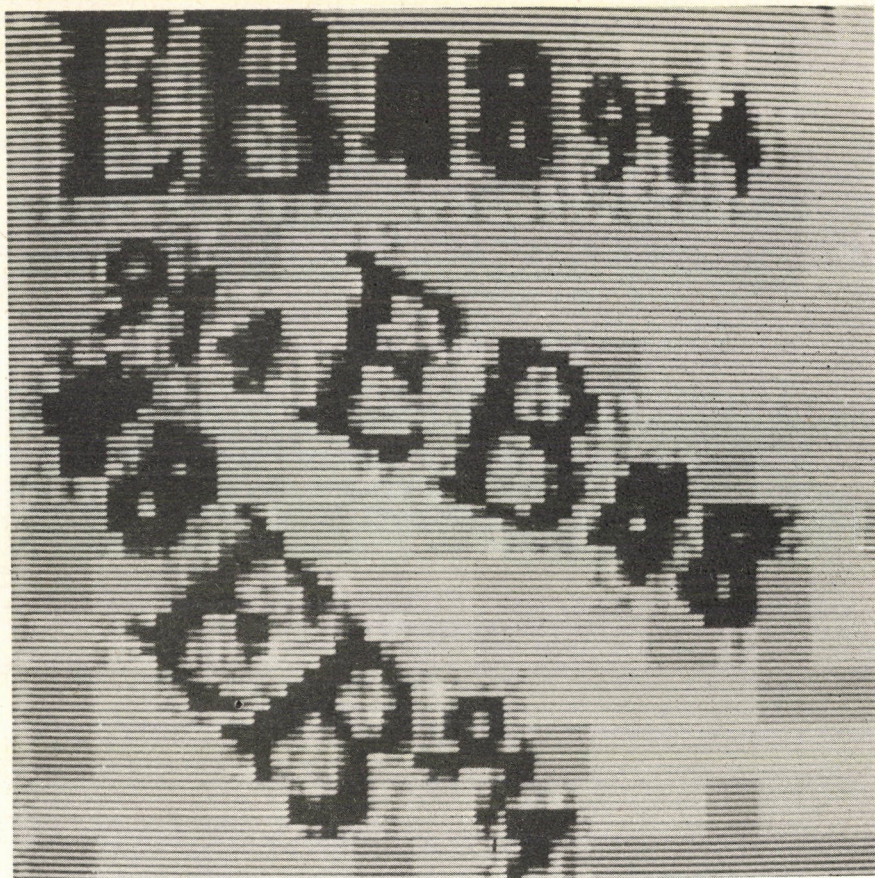
Для обработки оптической информации была построена система цифровой обработки телевизионных изображений. Для эффективного хранения изображений был выбран метод адаптивного кодирования с помощью двухмерного преобразования. В начале изображение разделяется на частные изображения. Отдельные частные изображения преобразуются с помощью скорого преобразования Адамара. Затем коэффициенты преобразования квантуются и кодируются в соответствии со свойствами частного изображения. Результаты демонстрируются на нескольких примерах.



(b) Compression ratio 4:1, $S = 125$, $p = 13$, zonal filtering (8a), quantization (10b), nonadaptive algorithm.



(c) Compression ratio 3:1, $S = 125$, $p = 13$, zonal filtering (8a), quantization (10d), nonadaptive algorithm.



(d) Compression ratio 7.1:1, $S=160$, $p=10$ zonal filtering (8a), quantization (10b), adaptive algorithm, decision level of variance 1.7

F. Šolc, J. Halabala
Technical University Brno
Czechoslovakia

ON NONLINEAR DIFFERENTIAL GAMES OF PURSUIT

S. A. VAHRAMEEV
(Moscow)

(Received 6 July, 1982)

Nonlinear differential games of pursuit on smooth manifolds are considered and sufficient condition of pursuit to be finished are also proved. In this work the technique of chronological calculus is used which was developed in [2, 3]. A brief summary of some notions and formulas of this calculus are presented in first section of the article. The sufficient conditions, presented in the paper are analogous to conditions of Pontryagin's first direct method of pursuit [1].

1. Preliminaries

1.1. *Vector fields and diffeomorphisms.* Let M be a C^∞ -manifold smoothly imbedded in the space R^d , $\Phi = \Phi(M)$ is an algebra of all smooth functions $\varphi : M \rightarrow R$, $\Phi^d = \Phi \times \dots \times \Phi$, E is an abbreviation on M of identical map in R^d , $T_x M$ is a tangent space to M at the point $x \in M$.

Each differentiation \vec{X} of algebra Φ , i.e. the linear map which satisfy the "differentiation of the product" rule

$$\vec{X}(\varphi\psi) = (\vec{X}\varphi)\psi + \varphi(\vec{X}\psi) \quad \forall \varphi, \psi \in \Phi$$

is referred to as vector field on M . The set $\text{Der}(\Phi)$ of all vector fields on manifold M have natural structure of Lie algebra with multiplication according to the formula

$$[\vec{X}\vec{Y}] = \vec{X} \circ \vec{Y} - \vec{Y} \circ \vec{X} \quad \forall \vec{X}, \vec{Y} \in \text{Der}(\Phi).$$

As in any Lie algebra to arbitrary element $\vec{X} \in \text{Der}(\Phi)$ corresponds to the linear map

$$\text{ad } \vec{X} : \text{Der}(\Phi) \rightarrow \text{Der}(\Phi)$$

defined by the formula

$$\text{ad } \vec{X}(\vec{Y}) = [\vec{X}\vec{Y}] \quad \forall \vec{X}, \vec{Y} \in \text{Der}(\Phi).$$

Let $\pi(x)$ be an orthogonal projection $\pi(x) : R^d \rightarrow T_x M$. With arbitrary element $\vec{h} \in R^d$ let us relate the vector field $\vec{h} \in \text{Der}(\Phi)$ such that

$$\vec{h}\varphi(x) = \langle d\varphi(x), \pi(x)h \rangle \quad \forall \varphi \in \Phi.$$

Because $\vec{X}\varphi(x) = \langle d\varphi(x), \vec{X}E(x) \rangle$, we have

$$\vec{h}E(x) = \pi(x)h.$$

Let

$$\|\varphi\|_{s,k} = \sup_{x \in k} \sum_{\alpha=1}^n \sup_{|h_j|=1} |\vec{h}_1 \circ \dots \circ \vec{h}_\alpha \varphi(x)|,$$

where $s \geq 0$ is an integer, and $k \subset M$ is a compact set. Then the set of seminorms $\|\cdot\|_{s,k}$ defines on Φ locally convex metrizable topology. In the sequel the space Φ is considered with this topology.

Let $\mathcal{L}(\Phi)$ be an associative algebra of all linear continuous maps in Φ with multiplication—composition of the maps. Let us relate to each smooth map $P: M \rightarrow M$ the linear transformation $P^*: \Phi \rightarrow \Phi$ by to the formula

$$P^*\varphi = \varphi \circ P \quad \forall \varphi \in \Phi.$$

It is obvious that to each diffeomorphism $P: M \rightarrow M$ there corresponds automorphism $P^*: \Phi \rightarrow \Phi$ which we also call the diffeomorphism. If P^* is a diffeomorphism, one may check that

$$\text{Ad } P^* \vec{X} = P^* \vec{X} P^{*-1} \in \text{Der}(\Phi)$$

$$\forall \vec{X} \in \text{Der}(\Phi).$$

1.2. *One-parameter families of functions, fields and diffeomorphisms.* Let $\varphi_t, t \in R$ be an one-parameter family of elements Φ . The continuity and differentiability of such family with respect to $t \in R$ are defined in a usual way. We call the family $\varphi_t, t \in R$ measurable iff the function $t \rightarrow \varphi_t(x)$ is measurable. We say that the family $\varphi_t, t \in R$ is locally integrable iff it is measurable and

$$\int_{t_1}^{t_2} \|\varphi_\tau\|_{s,k} d\tau < \infty$$

$$\forall t_1, t_2 \in R, s \geq 0, k \subset M.$$

The integral of locally integrable family $\varphi_t, t \in R$ is the function

$$x \mapsto \int_{t_1}^{t_2} \varphi_\tau(x) d\tau \in \Phi.$$

Let us consider now the one-parameter family of operators $L_t, t \in R$ in $\mathcal{L}(\Phi)$. For such operators the correspondent notions are defined in weak sense, i.e. the family $L_t, t \in R$ of elements $\mathcal{L}(\Phi)$ has the Q -property iff such property has the family $L_t \varphi, t \in R$ $\forall \varphi \in \Phi$.

Nonstationary field or merely field we call a locally integrable family $\vec{X}_t, t \in R$ of vector fields on M . Nonstationary fields are bounded iff

$$\int_{t_1}^{t_2} \|\varphi_\tau\|_{s,m} d\tau < \infty \quad \forall s \geq 0, t_1, t_2 \in R.$$

An arbitrary absolute continuous family of diffeomorphisms $P_t^*, t \in R$ which satisfy the condition $P_0^* = Id$ referred to as a flow on M .

Let \vec{X}_t, \vec{Y}_t be arbitrary (nonstationary) vector fields. By the symbols

$$\overline{\exp} \int_0^t \vec{X}_\tau d\tau \quad \text{and} \quad \overline{\exp} \int_0^t -\vec{Y}_\tau d\tau$$

we denote the solutions of operator equations

$$\frac{d}{dt} G_t = G_t \vec{X}_t, G_0 = Id$$

and

$$\frac{d}{dt} F_t = -\vec{Y}_t F_t, F_0 = Id.$$

It turns out [2, 3] that the solutions of such operator equations exist, are unique and are mutually converse flows if $\vec{X}_t = \vec{Y}_t$ and \vec{X}_t is bounded.

Let \vec{X}_t, \vec{Y}_t be arbitrary (bounded) fields. It turns out [2] that there are the following formulas of the "variation of a constant"

$$\begin{aligned} \overline{\exp} \int_0^t (\vec{X}_\tau + \vec{Y}_\tau) d\tau &= \overline{\exp} \int_0^t Ad \overline{\exp} \int_0^\theta \vec{X}_\tau d\tau d\theta \overline{\exp} \int_0^t \vec{X}_\tau d\tau = \\ &= \overline{\exp} \int_0^t \overline{\exp} \int_0^\theta ad \vec{X}_\tau d\tau \vec{Y}_\theta d\theta \Big| \int_0^t \vec{X}_\tau d\tau. \end{aligned} \tag{1.1}$$

By means of this formula it is easy to derive the formula

$$\overline{\exp} \int_0^t \vec{X}_\tau d\tau \overline{\exp} \int_0^t \vec{Y}_\tau d\tau = \overline{\exp} \int_0^t \overline{\exp} \int_0^\tau -ad \vec{Y}_\theta d\theta (\vec{X}_\tau + \vec{Y}_\tau) d\tau \tag{1.2}$$

which will be used below.

Note that if the flow P_t^* is the solution of an equation

$$\frac{d}{dt} P_t^* = P_t^* \circ \vec{X}_t, P_0^* = Id, \tag{1.3}$$

then the absolute continuous family $P_t, t \in R$, of diffeomorphisms of manifold M corresponding to it is defined by means of ordinary differential equation on M

$$\dot{x} = \vec{X}_t E(x), \quad x \in M \tag{1.4}$$

because

$$\frac{d}{dt} P_t = \frac{d}{dt} P_t^* E(x) = P_t^* \circ \vec{X}_t E(x) = (\vec{X}_t E) \circ P(x).$$

Conversely, if $P_t, t \in R, P_0 = Id$ is an absolutely continuous family of diffeomorphisms of manifolds M generated by equation (1.4), then P_t^* is the flow which satisfies equation (1.4).

Thus the solution $x(t), t \in R$ of equation (1.4), which satisfies the initial condition $x(0) = x_0$ may be expressed by the formula

$$x(t) = \overline{\exp} \int_0^t \vec{X}_\tau d\tau E(x_0), \quad t \in R.$$

The more circumstantial exposition of notion and facts mentioned here one can find in articles [2-3].

2. Nonlinear differential game of pursuit

2.1. *The statement of the problem.* Let M be smooth manifold, \vec{f} be a complete vector field on M ,

$$G_t = \{\vec{g}_t(u); u \in P \subset R^p\}, \quad H_t = \{\vec{h}_t(v); v \in Q \subset R^q\}$$

be the families of nonstationary bounded vector fields on M , which depend on parameters $u \in P \subset R^p$ and $v \in Q \subset R^q$. Let the submanifold $N \subset M$ and point $x_0 \in M$ be given. The admissible control $u(t), 0 \leq t \leq T$ of pursuer we call the measurable function of time which have values in the set P , and admissible control of evader is the measurable function which have values in the set Q .

Let us agree that if $u(t), v(t), t \in R$ are the admissible controls, then the families $\vec{g}_t(u(t)), \vec{h}_t(v(t)), t \in R$ are nonstationary fields in the sense of definition given in Section 1.

The differential game of pursuit $\Gamma = \langle M, N, G_t, H_t, \vec{f} \rangle$ which we considered is defined by

$$\begin{aligned} \dot{x} &= (\vec{f} + \vec{g}_t(u) - \vec{h}_t(v)) E(x), \quad x \in M, \\ x(0) &= x_0. \end{aligned} \tag{2.1}$$

We say that in the game $\Gamma = \langle M, N, G_t, H_t, \vec{f} \rangle$ the pursuit can be finished in time $T > 0$ from initial point $x_0 \in M$ iff for arbitrary admissible control of evader $v(t), 0 \leq t \leq T$ there exists the admissible control of pursuer $u(t), 0 \leq t \leq T$ such that solution

$$x(t) = \overline{\exp} \int_0^t (\vec{f} + \vec{g}_\tau(u(\tau)) - \vec{h}_\tau(v(\tau))) d\tau E(x_0), \quad t \in R$$

of equation (2.1) and satisfies the condition $x(T) \in N$.

2.2. *Sufficient condition of pursuit.* Let us denote by $\Xi(t)$ the geometrical difference in the space $\text{Der}(\Phi)$ of the sets $e^{tad\vec{f}}G_t$ and $e^{tad\vec{f}}H_t$:

$$\Xi(t) = e^{tad\vec{f}}G_t * e^{tad\vec{f}}H_t = \{ \vec{\xi} \in \text{Der}(\Phi) \mid \vec{\xi} + e^{tad\vec{f}}H_t \subset e^{tad\vec{f}}G_t \}.$$

Theorem 1. Suppose that there exists a nonstationary field $\vec{\xi}_t \in \text{Der}(\Phi)$ such that $\vec{\xi}_t \in \Xi(t), 0 \leq t \leq T$ and

$$\overline{\text{exp}} \int_0^T \vec{\xi}_t d\tau e^{T\vec{f}} E(x_0) \in N.$$

Then the pursuit may be finished on time T .

Proof. Let $v(t), 0 \leq t \leq T$ be an arbitrary control of evader. Then because $\forall \tau, 0 \leq \tau \leq T$

$$\vec{\xi} + e^{\tau ad\vec{f}} \vec{h}_\tau(v(\tau)) \in e^{\tau ad\vec{f}} G_\tau$$

there exists the admissible control $u(t), 0 \leq t \leq T$ of pursuer such that

$$\vec{\xi} + e^{\tau ad\vec{f}} \vec{h}_\tau(v(\tau)) = e^{\tau ad\vec{f}} \vec{g}(u(\tau))$$

for almost all $\tau, 0 \leq \tau \leq T$. Let us prove that

$$\overline{\text{exp}} \int_0^T (\vec{f} + \vec{g}_\tau(u(\tau)) - \vec{h}_\tau(v(\tau))) d\tau = \overline{\text{exp}} \int_0^T \vec{\xi} d\tau e^{T\vec{f}}. \tag{2.2}$$

By means of the variations formula we have

$$\begin{aligned} & \overline{\text{exp}} \int_0^T (\vec{f} + \vec{g}_\tau(u(\tau)) - \vec{h}_\tau(v(\tau))) d\tau = \\ & = \overline{\text{exp}} \int_0^T (e^{\tau ad\vec{f}} \vec{g}_\tau(u(\tau)) - e^{\tau ad\vec{f}} \vec{h}_\tau(v(\tau))) d\tau e^{T\vec{f}}. \end{aligned}$$

Thus in order to prove equality (2.2) it is sufficient to prove that

$$\overline{\text{exp}} \int_0^T (e^{\tau ad\vec{f}} \vec{g}(u(\tau)) - e^{\tau ad\vec{f}} \vec{h}(v(\tau))) d\tau = \overline{\text{exp}} \int_0^T \vec{\xi}_\tau d\tau. \tag{2.3}$$

Because

$$\begin{aligned} & \overline{\text{exp}} \int_0^T [e^{\tau ad\vec{f}} \vec{g}_\tau(u(\tau)) - e^{\tau ad\vec{f}} \vec{h}_\tau(v(\tau))] d\tau = \\ & = \overline{\text{exp}} \int_0^T \overline{\text{exp}} \int_0^t a d e^{\tau ad\vec{f}} \vec{g}_\tau(u(\tau)) d\tau e^{tad\vec{f}} \vec{h}_t(v(t)) dt \overline{\text{exp}} \int_0^T e^{\tau ad\vec{f}} \vec{g}_\tau(u(\tau)) d\tau \end{aligned}$$

then (2.3) is equivalent to the equality

$$\begin{aligned} & \overline{\exp} \int_0^T \overline{\exp} \int_0^t ad e^{\tau ad f} g_\tau(u(\tau)) d\tau e^{t ad f} \vec{h}_t(v(t)) dt = \\ & = \overline{\exp} \int_0^T \vec{\xi}_\tau d\tau \left(\overline{\exp} \int_0^T e^{t ad f} \vec{g}_t(u(t)) dt \right)^{-1}. \end{aligned} \quad (2.4)$$

By the choice of $u(t)$, $0 \leq t \leq T$ we have

$$\begin{aligned} & \vec{\xi}_t - e^{t ad f} \vec{g}_t(u(t)) = -e^{t ad f} \vec{h}_t(v(t)); \\ & \overline{\exp} \int_0^T \vec{\xi}_\tau d\tau \left(\overline{\exp} \int_0^T e^{t ad f} \vec{g}_t(u(t)) dt \right)^{-1} = \\ & = \overline{\exp} \int_0^T \vec{\xi}_\tau d\tau \left(\overline{\exp} \int_0^T -e^{t ad f} \vec{g}_t(u(t)) dt \right) = \\ & = \overline{\exp} \int_0^T \overline{\exp} \int_0^t ad e^{\tau ad f} \vec{g}_\tau(u(\tau)) d\tau (\vec{\xi}_t - e^{t ad f} \vec{g}_t(u(t))) dt = \\ & = \overline{\exp} \int_0^T -\overline{\exp} \int_0^t ad e^{\tau ad f} \vec{g}_\tau(u(\tau)) d\tau e^{t ad f} \vec{g}_t(u(t)) dt. \end{aligned}$$

Here we use the correspondence, (1.2), and the equality

$$\overline{\exp} \int_0^T \vec{X}_\tau d\tau = \left(\overline{\exp} \int_0^T \vec{X}_\tau d\tau \right)^{-1}.$$

So we show that (2.4) is the identity, thus the identity is (2.2). The theorem is proved.

3. Examples

3.1. *Bilinear differential games.* Let us denote by $Gl(n; R)$ the general linear Lie group and by $gl(n; R)$ its Lie algebra and consider the conflict control right-invariant system

$$\dot{X} = AX + \sum_{i=1}^p u_i B_i X - \sum_{j=1}^q v_j C_j X$$

on Lie group $Gl(n; R)$. Here are $u = (u_1, \dots, u_p) \in P$, $v = (v_1, \dots, v_q) \in Q$ are the controls of pursuer and evader, respectively P and Q are the compacts in the spaces R^p and R^q . Let us denote by

$$\overline{\exp} \int_0^t A_\tau d\tau \quad \text{and} \quad \overline{\exp} \int_0^t B_\tau d\tau$$

the solutions of matrix equations

$$\begin{aligned} \dot{X}_t &= X_t A_t, & X_0 &= I, \\ \dot{Y}_t &= B_t Y_t, & Y_0 &= I, \end{aligned}$$

where I is identity matrix. The differential game of pursuit to matrix system (2.1) is posed in the next way.

There is prescribed a nonempty set $\mathcal{X} \subset Gl(n; R)$. For each admissible control of evader $v(t)$, $0 \leq t \leq T$ one has to find the admissible control $u(t)$, $0 \leq t \leq T$ of pursuer such that the corresponding solution

$$X_t = \overline{\exp} \int_0^t \left(A + \sum_{i=1}^p u_i(\tau) B_i - \sum_{j=1}^q v_j(\tau) C_j \right) d\tau, \quad t \in R$$

of matrix differential equation

$$\begin{aligned} \dot{X}_t &= AX_t + \sum_{i=1}^p u_i(t) B_i X_t - \sum_{j=1}^q v_j(t) C_j X_t, \\ X_0 &= I \end{aligned} \quad (3.1)$$

satisfies the condition

$$X_T \in \mathcal{X}.$$

From theorem 1 follows

Theorem 2. Let for $\forall t, 0 \leq t \leq T$

$$\omega_T(t) = \left\{ e^{(T-t)adA} \sum_{i=1}^p u_i B_i; u \in P \right\} * \left\{ e^{(T-t)adA} \sum_{j=1}^q v_j C_j; v \in Q \right\} \neq \emptyset$$

and there exists the measurable family $\xi_t, t \in R$ of elements $gl(n; R)$ such that

$$\xi_t \in \omega_T(t), \quad 0 \leq t \leq T$$

and

$$\overline{\exp} \int_0^T \xi_\tau d\tau e^{TA} \in \mathcal{X}.$$

Then the pursuit may be finished in time T .

As an illustration of application of theorem 2, let us consider the system

$$\begin{cases} \dot{x}^1 = x^2 + \alpha u x^2 - c x^1 v, \\ \dot{x}^2 = u \beta x^1, |u| \leq r, |v| \leq \tau. \end{cases} \quad (3.2)$$

Because

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in sl(2; R),$$

an easy calculation shows that

$$e^{tA} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad e^{tadA}B = \begin{pmatrix} t\beta & -t^2\beta + \alpha \\ \beta & -t\beta \end{pmatrix}, \quad e^{tadA}C = \begin{pmatrix} 0 & C \\ 0 & 0 \end{pmatrix},$$

$$\omega_T(t) = e^{(T-t)adA}BP^* e^{(T-t)adA}CQ =$$

$$= \left\{ \begin{pmatrix} (T-t)\beta & -(T-t)^2\beta + \alpha \\ \beta & (t-T)\beta \end{pmatrix} u; |u| \leq \sigma \right\}^* \left\{ \begin{pmatrix} 0 & C \\ 0 & 0 \end{pmatrix} v; |v| \leq \tau \right\},$$

thus when $\sigma \min_{t \in [0, T]} |\alpha - (T-t)^2\beta| = \gamma_T^* \geq |c|\tau$ we have

$$\omega_T(t) = \left\{ \begin{pmatrix} (T-t)\beta & \gamma_t \\ \beta & (T-t)\beta \end{pmatrix}; \gamma_t \in \left\{ \hat{t} | \hat{t} | \leq \sigma |\alpha - (T-\hat{t})\beta| - \sigma |c| \right\} \right\}.$$

This shows that for system (3.2) the pursuit can be finished in time T from the initial position $x_0 = (x_0^1, x_0^2)$ if $\sigma \min_{t \in [0, T]} |\alpha - (T-t)^2\beta| \geq |c|\tau$ on an arbitrary subset which includes the attainable set in time T from initial position (x_0^1, x_0^2) for the following control system

$$\dot{x}^1 = \alpha u x^2, \quad \dot{x}^2 = \beta u x^1,$$

$$u \in P_T(t) = \{ |\xi| |\dot{\xi}| \leq \sigma |\alpha - (T-t)^2\beta| - \sigma |c| \}.$$

3.2. *Commutative systems.* Bilinear system (3.1) is called commutative if $\forall t, s \in R$

$$[e^{tadA}A_i e^{sadA}A_j] = 0 \quad (3.3)$$

for all $i, j = 1, \dots, p+q$, where $A_i = B_i, i = 1, \dots, p, A_{p+j} = C_j, j = 1, \dots, q$. There exist the constructive criteria of the verification of commutative condition (3.3) (see [4]). Namely, system (3.1) is commutative iff

$$[ad^k AA_i A_j] = 0, \quad \forall k = 0, \dots, n^2 - 1, \quad i, j = 1, \dots, p+q.$$

Let $u(t), v(t)$ be admissible controls. Then the correspondence solution X_t of equation (3.2) can be represented in the following way [2]:

$$X_t = e_0^t \int_0^t e^{(t-\tau)adA} \left(\sum_{i=1}^p u_i(\tau) B_i - \sum_{j=1}^q v_j(\tau) C_j \right) d\tau e^{tA},$$

so on trajectories of systems (3.2) from I lie only those matrices $X \in Gl(n; R)$ which have the factorisation

$$X = e^H e^{tA} \quad (H \in gl(n; R)).$$

For this case, theorem 2 can be formulated in the next way:

Theorem 3. Suppose that the system is commutative. Let

$$\mathcal{X} = \{e^H e^{TA}, H \in \mathcal{H} \subset gl(n; R)\}.$$

If

$$\omega_T(t) \neq \emptyset, \quad 0 \leq t \leq T$$

then the pursuit may be finished in time T if there exists the measurable family ξ_t , $t \in R$ of elements of $gl(n; R)$ such that $\xi_t \in \omega_T(t)$, $0 \leq t \leq T$ and

$$\int_0^T \xi_\tau d\tau \in \mathcal{H}.$$

Let us show now from theorem 3 that one may derive the conditions of the first direct method of Pontryagin for linear system

$$\begin{aligned} \dot{x} &= Ax + Bu - Cv = Ax + \sum_{i=1}^p u_i b_i - \sum_{j=1}^q v_j c_j; \\ x \in R^n, u &= (u_1, \dots, u_p) \in P, v = (v_1, \dots, v_q) \in Q, \\ x(0) &= x_0 \end{aligned} \tag{3.4}$$

with closed terminal set $N \subset R^n$. For this purpose let us identify the space R^n with affine plane $\Pi = \left\{ z = \begin{pmatrix} x \\ 1 \end{pmatrix}; x \in R^n \right\}$ in R^{n+1} and set

$$\begin{aligned} \tilde{A} &= \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{B}_i = \begin{pmatrix} 0 & b_i \\ 0 & 0 \end{pmatrix}, \quad \tilde{C}_i = \begin{pmatrix} 0 & c_i \\ 0 & 0 \end{pmatrix} \\ \tilde{M} &= \left\{ \begin{pmatrix} 0 & v \\ 0 & 0 \end{pmatrix}; v \in N \right\}, \quad \mathcal{H} = \left\{ \begin{pmatrix} 0 & h \\ 0 & 0 \end{pmatrix}; h \in R^n \right\}. \end{aligned}$$

Then to a linear control system (3.4) corresponds a bilinear control system in R^{n+1}

$$\begin{aligned} \dot{z} &= \tilde{A}z + \sum_{i=1}^p u_i \tilde{B}_i z - \sum_{j=1}^q v_j \tilde{C}_j z, \\ z \in R^{n+1}, \quad z(0) &= \begin{pmatrix} x_0 \\ 1 \end{pmatrix}, \quad u \in P, \quad v \in Q. \end{aligned}$$

To this system corresponds the commutative matrix system

$$\begin{aligned} \dot{Z} &= \tilde{A}Z + \sum_{i=1}^p u_i \tilde{B}_i Z - \sum_{j=1}^q v_j \tilde{C}_j Z, \\ Z(0) &= I \end{aligned} \tag{3.5}$$

with terminal set $\mathcal{X} = \{e^H e^{TA}, H \in \mathcal{H}\}$.

By theorem 3 we have that the pursuit for this system may be finished in time T if

$$\begin{aligned} \omega_T(t) &= \left\{ e^{(T-t)ad\tilde{A}} \sum_{i=1}^q u_i \tilde{B}_i; u \in P \right\}^* \\ &= \left\{ e^{(T-t)ad\tilde{A}} \sum_{j=1}^q v_j \tilde{C}_j; v \in Q \right\}^* = \\ &= \begin{pmatrix} 0 & e^{(T-t)A}BP^* e^{(T-t)A}CQ \\ 0 & 0 \end{pmatrix} \neq \emptyset, \end{aligned}$$

$0 \leq t \leq T$ and there exists a measurable family $\xi_t \in \omega_T(t)$, $0 \leq t \leq T$ such that

$$\int_0^T \xi_\tau d\tau \in \mathcal{H}.$$

Thus we received the conditions

$$(i) \quad e^{(T-t)A}BP^* e^{(T-t)A}CQ \equiv \hat{\omega}_T(t) \neq \emptyset, \quad 0 \leq t \leq T;$$

$$(ii) \quad J\hat{v} \in R^n: \hat{v} = \int_0^T \hat{\xi}_\tau d\tau,$$

where

$$\hat{\xi}_\tau \in \hat{\omega}_T(\tau), \quad 0 \leq \tau \leq T.$$

If we claim the condition $Z(t) \in N$ to be hold, we have

$$(iii) \quad Z(T) \in e^{\begin{pmatrix} 0 & \hat{v} \\ 0 & 0 \end{pmatrix} T} \tilde{M}.$$

Conditions (i)–(iii) show that the linear differential games (3.4) may be finished in time T from initial point $x_0 \in R^n$ if

$$\hat{\omega}_T(t) \neq \emptyset, \quad 0 \leq t \leq T$$

and

$$e^{TA}x_0 \in - \int_0^T \hat{\omega}_T(\tau) d\tau + N.$$

Thus we receive conditions which are analogous to the conditions of Pontryagin's first direct method [1].

The author thanks A. A. Agrachev and R. W. Gamkrelidze for the helpful discussions.

Reference

1. Pontryagin, L. S., The linear differential games of pursuit. Math. Coll., 1980, **112**, No. 3, pp. 309–330 (in Russian).
2. Agrachev, A. A., Gamkrelidze, R. W., Exponential representation of flows and chronologic calculus. H. Math. Coll., 1978, **107**, No. 4, pp. 463–532 (in Russian).
3. Agrachev, A. A., Gamkrelidze, R. W., Chronological algebras and nonstationary vector fields. Geometrical Problems. M., 1979, **11**, 135–176 (in Russian).
4. Brockett, R. W., System theory on group manifolds and coset spaces. Siam J. Contr., 1972, **10**, No. 2, pp. 265–284.

О нелинейных дифференциальных играх преследования

С. А. ВАХРАМЕЕВ
(Москва)

Рассматриваются нелинейные дифференциальные игры преследования на гладких многообразиях. Динамика таких игр описывается обыкновенными дифференциальными уравнениями вида

$$\dot{x} = f(x) + g(u, x) - h(x, v), \quad x \in M,$$

$$v \in Q, u \in P.$$

Здесь \vec{f} — гладкое векторное поле на n -мерном дифференцируемом многообразии M , $\{\vec{g}(u); u \in P\}$ и $\{\vec{h}(v); v \in Q\}$ — семейства гладких векторных на M , непрерывно зависящие от $u \in P$ и $v \in Q$, PQ — непустые компакты в конечномерных евклидовых пространствах. Предлагаются достаточные условия, при которых возможно завершение преследования. Эти условия обобщают условия Л. С. Понтрягина на рассматриваемый класс дифференциальных игр. Метод, который используется в статье, основан на хронологическом исчислении, развитом в работах А. А. Аграчёва и Р. В. Гамкrelидзе. В билинейном случае условия, предлагаемые в данной работе, применимы к достаточно широкому классу дифференциальных игр и допускают эффективную проверку. Приводятся примеры и выводятся условия первого прямого метода Л. С. Понтрягина для линейных дифференциальных игр преследования.

С. А. Вахрамеев
ВИНИТИ АН СССР, ГКНТ
СССР, Москва, А-219
Балгийская ул., 14

THE CONTROL OF ROTATION FOR ASYMMETRIC RIGID BODY

A. A. AGRACHEV, A. V. SARYCHEV
(Moscow)

(Received 6 July, 1982)

The control of angular momentum vector for asymmetric rigid body is under consideration. The control torque is directed along the axis \bar{L} which is fixed in the body. The problems of controllability are studied. The attainable sets of the corresponding control system are constructed in angular momentum space. The results concerning the simultaneous control of angular momentum and attitude are formulated.

1. Introduction

We consider an asymmetric rigid body which rotates about its center of mass. Control torque can be applied along the axis \bar{L} which is fixed in the body. Let the value of angular velocity at the initial time $t = 0$ be $\bar{\Omega}$. We study whether it is possible to come from the rotation with angular velocity $\bar{\Omega}$ to one with angular velocity $\bar{\Omega}$ for the finite time T . Probably the most interesting is to find the possibility for stopping the body for the finite time and the possibility for coming to the stationary rotation around one of the principal axes.

The problem of control of the rigid body rotating about its center of mass was considered in a lot of publications, see for example [1], which contains vast bibliography. A problem in [2] is similar to our one.

In one of the recent publications [3] "complete controllability" of rigid body is established when axis \bar{L} is in a "general position". This result is formulated in i. 9A of our work. Nonetheless we have considered it is worthwhile to offer our proof of statement 9A which differs from the proof given in [3]. Our proof is based on the study of the first variation of the system only (it means that during the calculations only expressions linear on \bar{L} are used (see (6.1), (6.2)). It can be said that not simply the controllability of the control system but the controllability in linear approach is proved. It will give us later a possibility to describe basing upon developed research the attainable sets of the control system (1.2) when control $u(M)$ is a smooth function of phase vector M (smooth feedback control) unlike measurable functions of time $u(t)$ (measurable program controls) which are studied in this paper.

At i.i. 9B-9F possible breaks of the general position of the axis \bar{L} are studied in succession; the attainable sets of the control system (1.2) are described correspondingly.

In i. 10 we study the possibility to control simultaneously the angular velocity and orientation of the body.

Let us give exact definitions. When accounting the theory of motion of the rigid body around the center of mass we shall follow [4]. The configuration of the rigid body is described by the inertia tensor A . In a co-ordinate system connected with the body, A is a symmetric (3×3) -matrix. The eigenvectors of this matrix are called principal axes of the body, its eigenvalues are the principal moments of inertia. The body is called asymmetric if all its principal moments of inertia are different. The angular velocity vector in the coordinate system connected with the body is denoted Ω ($\Omega \in \mathbf{R}^3$). Vector $M = A\Omega$ is an angular momentum vector ($M \in \mathbf{R}^3$). Further we shall describe the motion of angular momentum vector M which linearly depends upon the angular velocity vector.

According to [4] the equation of the motion of the angular momentum vector (Euler equation) is

$$\dot{M} = M \times \Omega = M \times BM. \quad (1.1)$$

Here “ \times ” denotes vector product in \mathbf{R}^3 , B is a matrix reversed to A .

If a torque proportional to the current value of the scalar control function $u(t)$ is applied along the axis \bar{L} which is fixed in the body then the equation of motion is

$$\dot{M} = M \times BM + Lu. \quad (1.2)$$

(Here L is a unit vector, lying on \bar{L} .) The control functions $u(t)$ are measurable, $\forall t |u(t)| \leq \kappa$.

Control system (1.2) is a special case of the autonomous control system

$$\dot{x} = f(x) + g(x)u \quad (1.3)$$

where $x \in \mathbf{R}^n$, $u \in \mathbf{R}^1$, f, g are smooth vector fields; control functions $u(t)$ are measurable, $|u(t)| \leq \kappa$.

Definition 1. Point $\hat{x} \in \mathbf{R}^n$ is attainable from point $\tilde{x} \in \mathbf{R}^n$ for system (1.3) if the admissible control function $u(t)$ exists such that the solution of the equation

$$\dot{x} = f(x) + g(x)u(t),$$

starting at \tilde{x} ($x(0) = \tilde{x}$) equals to \hat{x} in some finite moment of time T ($x(T) = \hat{x}$). The set of points which are attainable from \tilde{x} is called an attainable set for system (1.3) from \tilde{x} . It is denoted $D_{\tilde{x}}$.

The purpose of this article is to describe the attainable set $D_{\tilde{M}}$ for control system (1.2) for arbitrary \tilde{M} .

2. Local controllability

Let us reduce the problem of constructing the attainable set to the study of local controllability of the system.

Let us consider again control system (1.3)

$$\dot{x} = f(x) + g(x)u.$$

Definition 2. System (1.3) is locally controllable at the point $\tilde{x} \in \mathbf{R}^n$, if there exists the vicinity W of point \tilde{x} in \mathbf{R}^n such that: a) every point of W can be attained from \tilde{x} ; b) \tilde{x} can be attained from every point of W .

The next statement follows from the definition of local controllability.

Proposition 1. Let Q be an open connected domain, and every point of Q be a point of local controllability of system (1.3); $\tilde{x} \in Q$. Then every point $x \in Q$ can be attained from \tilde{x} and vice versa \tilde{x} can be attained from every point of Q .

If changing in (1.3) the direction of time we get system (1.3⁻)

$$\dot{x} = -f(x) - g(x)u.$$

The attainable set from the point \tilde{x} of system (1.3⁻) is denoted by $D_{\tilde{x}}^-$. Evidently systems (1.3) and (1.3⁻) are or are not locally controllable simultaneously. Besides that $x \in D_y^-$ if and only if $y \in D_x$.

Proposition 2. Let Q be an open connected subdomain of \mathbf{R}^n . Suppose that system (1.3) is locally controllable at every point of Q except points of a subset $\hat{Q} \subset Q$. Suppose also that $Q \setminus \hat{Q} = \bigcup_{i=1}^m Q_i$, where Q_i are open connected components of $Q \setminus \hat{Q}$. If for every point $\hat{x} \in \hat{Q}$, there exist points $y_i, z_i \in Q_i$ ($i = 1, \dots, m$), such that $y_i \in D_{\hat{x}}$, $z_i \in D_{\hat{x}}^-$, then for every point $\tilde{x} \in Q$ the attainable set $D_{\tilde{x}}$ of system (1.3) contains Q .

Proof is a sequence of simple reasonings: 1) it follows from Proposition 1 that $\forall \tilde{x} \in Q_i D_{\tilde{x}} \supseteq Q_i$; 2) According to the statement $\forall \hat{x} \in \hat{Q} \exists y_i \in Q_i$ such that $y_i \in D_{\hat{x}}$ and thus $D_{\tilde{x}} \supseteq D_{y_i} \supseteq Q_i$. As far as i is arbitrary, so $D_{\tilde{x}} \supseteq \bigcup_{i=1}^m Q_i = Q \setminus \hat{Q}$. Let \hat{x} be an arbitrary point of \hat{Q} , $\hat{x} \neq \tilde{x}$. By assumption $\exists z_i \in Q_i$, $z_i \in D_{\hat{x}}^-$, i.e. $\hat{x} \in D_{z_i}$. Thus $\forall \hat{x} \in \hat{Q} D_{\tilde{x}} \supseteq Q$. 3) Let $\tilde{x} \in Q \setminus \hat{Q}$, for example $\tilde{x} \in Q_i$. Let us take an arbitrary point $\hat{x} \in \hat{Q}$. By assumption $\exists z_i \in Q_i$, $z_i \in D_{\hat{x}}^-$, i.e. $\hat{x} \in D_{z_i}$. Since evidently $z_i \in D_{\tilde{x}}$, $\hat{x} \in D_{z_i}$ and $D_{\tilde{x}} \supseteq Q$, then $D_{\tilde{x}} \supseteq Q$.

3. The sufficient condition of local controllability

The next proposition gives us a sufficient condition of local controllability. In its formulation we use the following terms: $[f, g](x) = \frac{\partial g}{\partial x} f(x) - \frac{\partial f}{\partial x} g(x)$ Lie bracket

(commutator) of two vector fields f, g . Vector field $(adf)^k g$ is defined inductively

$$(adf)g = [f, g], \quad (adf)^k g = [f, (adf)^{k-1} g].$$

Proposition 3. Let $\tilde{x}(t)$ be a periodic trajectory of the equation $\dot{x} = f(x)$, $\tilde{x}(t+T) = \tilde{x}(t)$ and for some t_0 the vectors $g(\tilde{x}(t_0)), ((adf)g)(\tilde{x}(t_0)), \dots, ((adf)^{n-1}g)(\tilde{x}(t_0))$ are linearly independent. Then control system (1.3) is locally controllable at every point of the trajectory $\tilde{x}(t)$ (see for example [5]).

4. Euler equation for the free rotation of the body

It is known that Euler equation (1.1) has two integrals of motion: energy integral $\mathcal{E} = \frac{1}{2} \langle M, \Omega \rangle = \frac{1}{2} \langle M, BM \rangle$ and angular momentum integral $\mathfrak{M}^2 = \langle M, M \rangle$. Here and further $\langle \cdot, \cdot \rangle$ denotes scalar product in \mathbf{R}^3 . Let us consider the orthonormal basis e_1, e_2, e_3 , where e_1, e_2, e_3 lie on the principal axes of the body. The principal axes will be denoted $\bar{e}_i = \{\alpha e_i; -\infty < \alpha < +\infty\}$. The eigenvalues of matrix B are $J_1 > J_2 > J_3$. In coordinate system e_1, e_2, e_3 the integrals of motion look like $\mathcal{E} = \frac{1}{2}(J_1 M_1^2 + J_2 M_2^2 + J_3 M_3^2)$, $\mathfrak{M}^2 = M_1^2 + M_2^2 + M_3^2$. The existence of these integrals implies

4a. Each trajectory of Eq. (1.1) lies in the intersection of the ellipsoid defined by the equation $\mathcal{E} = \text{const.}$, and sphere defined by the equation $\mathfrak{M}^2 = \text{const.}$

4b. Each point of each of the axes \bar{e}_i ($i = 1, 3$) is a stationary point of Eq. (1.1). For stationary trajectories of Eq. (1.1) one of the next three relations holds: $2\mathcal{E} = J_i \mathfrak{M}^2$ ($i = 1, 3$).

4c. If $2\mathcal{E} \neq J_2 \mathfrak{M}^2$ then the corresponding solution of Eq. (1.1) is periodical.

4d. If $2\mathcal{E} = J_2 \mathfrak{M}^2$ then: i) each point of axis \bar{e}_2 is a stationary point of Eq. (1.1); ii) trajectories of Eq. (1.1) lie in this case in one of two specific planes Π_1, Π_2 , which are defined on the basis $\{e_1, e_2, e_3\}$ by the equations

$$\sqrt{J_1 - J_2} M_1 + \sqrt{J_2 - J_3} M_3 = 0,$$

$$\sqrt{J_1 - J_2} M_1 - \sqrt{J_2 - J_3} M_3 = 0.$$

These trajectories are the arcs of a circle whose center lies in the origin; iii) middle principal, axis \bar{e}_2 lies in intersection of the planes Π_1, Π_2 ; iv) each of the full trajectories of Eq. (1.1), lying in Π_1 is a semicircle which begins at the point of positive semi-axis \bar{e}_2 and ends in the symmetrically situated point of negative semi-axis; in Π_2 trajectories begin on the negative semi-axis and end on the positive one.

5. Some formulae of vector algebra

Further we need some formulae of vector algebra. N will denote a self-adjoint operator in \mathbf{R}^3 .

Proposition 4. The following equality holds:

$$N(x \times y) = (\text{Tr } N)(x \times y) - Nx \times y - x \times Ny. \tag{5.1}$$

Proof. It is sufficient to verify this equality for some orthonormal basis in \mathbf{R}^3 . Let us choose basis in \mathbf{R}^3 for which the operator N has diagonal form. In this case the verification of (5.1) reduces to trivial calculation.

In particular, when $y = Nx$ we get

$$N(x \times Nx) = (\text{Tr } N)(x \times Nx) - x \times N^2x. \tag{5.2}$$

Scalar multiplication of both parts of (5.1) on vector $(u \times v)$ gives us

$$\langle u \times v, N(x \times y) \rangle = (\text{Tr } N)\langle u \times v, x \times y \rangle - \langle u \times v, Nx \times y \rangle - \langle u \times v, x \times Ny \rangle. \tag{5.3}$$

Further we also use a well-known formula for the double vector product

$$(x \times y) \times z = \langle z, x \rangle y - \langle z, y \rangle x. \tag{5.4}$$

Let us generalize formula (5.4), i.e. calculate double vector product $(N(x \times y)) \times z$. Using (5.1) for the calculation of $N(x \times y)$ we get

$$(N(x \times y)) \times z = (\text{Tr } N)((x \times y) \times z) - ((Nx \times y) \times z) - ((x \times Ny) \times z). \tag{5.5}$$

Using (5.4) for the calculation of the double vector products in (5.5) we get

$$\begin{aligned} (N(x \times y)) \times z &= \langle y, z \rangle Nx - \langle x, z \rangle Ny + \langle y, Nz \rangle x - \\ &- \langle x, Nz \rangle y + \langle x, z \rangle (\text{Tr } N)y - \langle y, z \rangle (\text{Tr } N)x. \end{aligned} \tag{5.6}$$

6. The study of local controllability of system (1.2)

Let us go over to the research of the control system (1.2). We shall denote $F(M) = M \times BM$; a constant vector field equal to L at every point, will also be denoted by L .

Let us calculate vector fields $V_1 = (adF)L$ and $V_2 = (adF)^2L$. As L does not depend on M , and F depends on M quadratically, then

$$V_1 = (adF)L = [F, L] = [M \times BM, L] = BL \times M - L \times BM. \tag{6.1}$$

$$\begin{aligned} V_2 &= (adF)^2L = [F[F, L]] = [M \times BM, BL \times M - L \times BM] = \\ &= BL \times (M \times BM) - (BL \times M) \times BM + L \times B(M \times BM) - M \times B(BL \times M) + \\ &+ (L \times BM) \times BM + M \times B(L \times BM). \end{aligned} \tag{6.2}$$

We remind that in (6.1) and (6.2) B is a matrix inverse to the inertia tensor. Let us transform double vector products in the right part of (6.2) using (5.4) and (5.6). Putting in order similar terms, we get

$$V_2 = (adF)^2 L = (\langle BM, M \rangle (\text{Tr } B) - 2\langle BM, BM \rangle) L + \\ + (\langle BM, M \rangle - \langle M, M \rangle (\text{Tr } B)) BL + \langle M, M \rangle B^2 L.$$

It follows from proposition 3 that the linear independence of the values of vector fields $V_0 = L$, $V_1 = (adF)L$, $V_2 = (adF)^2 L$ calculated at a point \tilde{M} lying on a periodic trajectory of Eq. (1.1) is sufficient for local controllability of system (1.2) at any point of this trajectory. On the other hand, linear independence of V_0, V_1, V_2 is equivalent to the inequality

$$\langle V_1, V_0 \times V_2 \rangle \neq 0.$$

Vector product $V_0 \times V_2$ is defined by the expression

$$V_0 \times V_2 = (\langle BM, M \rangle - (\text{Tr } B) \langle M, M \rangle) (L \times BL) + \langle M, M \rangle (L \times B^2 L) = \\ = \langle BM, M \rangle (L \times BL) + \langle M, M \rangle ((L \times B^2 L) - (\text{Tr } B) (L \times BL)).$$

According to (5.2), $L \times B^2 L - (\text{Tr } B) (L \times BL) = -B(L \times BL)$.

Thus

$$V_0 \times V_2 = \langle BM, M \rangle (L \times BL) - \langle M, M \rangle B(L \times BL). \quad (6.3)$$

Let us denote $X = L \times BL$. Then $V_0 \times V_2 = \langle BM, M \rangle X - \langle M, M \rangle BX$, and the mixed product $\langle V_1, V_0 \times V_2 \rangle$ is defined by the expression

$$\langle V_1, V_0 \times V_2 \rangle = \langle BM, M \rangle \langle X, V_1 \rangle - \langle M, M \rangle \langle BX, V_1 \rangle.$$

The condition of linear dependence of V_0, V_1, V_2 is

$$\langle BM, M \rangle \langle X, V_1 \rangle - \langle M, M \rangle \langle BX, V_1 \rangle = 0. \quad (6.4)$$

Let us note that as far as V_1 is linear on M , (6.4) defines cubic surface in \mathbf{R}^3 .

7. The study of local controllability of system (1.2) (continuation)

Let us study surface (6.4) and its connection with control system (1.2). As it was shown above, the local controllability of system (1.2) at some points of the periodic trajectory of (1.1) implies local controllability at every point of this trajectory. Thus, in order to find points where system (1.2) is not locally controllable, it is necessary to pick out the trajectories of the Eq. (1.1), which are entirely lying on the surface (6.4).

Let us note that the values $\langle M, M \rangle$, $\langle M, BM \rangle$ are constants along every trajectory of Eq. (1.1) and equal \mathfrak{M}^2 and $2\mathcal{E}$, correspondingly. Hence the points of the trajectory of Eq. (1.1), which is entirely lying on the surface (6.4), satisfy the equation

$$2\mathcal{E}\langle X, V_1 \rangle - \mathfrak{M}^2 \langle BX, V_1 \rangle = 0 \tag{7.1}$$

where V_1 was defined above as $V_1 = BL \times M - L \times BM$.

For an arbitrary vector $Y \in \mathbf{R}^3$ we denote ω_Y the linear operator mapping \mathbf{R}^3 to \mathbf{R}^3 according to the formula $\forall z \in \mathbf{R}^3 \ \omega_Y z = Y \times z$. Then V_1 can be regarded as a result of applying the operator $T = \omega_{BL} - \omega_L \circ B$ to the vector M . The adjoint operator T^* is defined by the formula: $T^* = B \circ \omega_L - \omega_{BL}$. Denoting $2\mathcal{E}/\mathfrak{M}^2 = \alpha > 0$, we represent Eq. (7.1) as follows

$$\langle T^*(\alpha E - B)X, M \rangle = 0 \tag{7.2}$$

where E is an identical operator in \mathbf{R}^3 . Let us elucidate when the equality

$$T^*(\alpha E - B)X = 0 \tag{7.3}$$

holds.

Equation (7.3) has the following solutions

$$X = L \times BL = 0. \tag{7.3_1}$$

It means that \bar{L} coincides with one of the axes $\bar{e}_1, \bar{e}_2, \bar{e}_3$. Evidently in this case (6.4) is satisfied at every point of \mathbf{R}^3 .

$$X = L \times BL \neq 0, \quad (\alpha E - B)(L \times BL) = 0. \tag{7.3_2}$$

It holds if and only if $\alpha = J_i$, $X = L \times BL$ is collinear to e_i ($i = \overline{1, 3}$). Let us note that $L \times BL$ is collinear to e_i if and only if $\bar{L} \perp \bar{e}_i$. When $\alpha = J_1, \alpha = J_3$ the corresponding solutions of Eq. (1.1) are stationary, i.e. they are points lying on the axes \bar{e}_1 or \bar{e}_3 , correspondingly. Thus if $\bar{L} \perp \bar{e}_1, \bar{L} \perp \bar{e}_3$, then (6.4) is satisfied at every point of axis \bar{e}_1 or \bar{e}_3 correspondingly. If $\alpha = J_2$ then the trajectory of Eq. (1.1) is an arc of the circle, lying in one of the planes Π_1, Π_2 that is if $\bar{L} \perp \bar{e}_2$, then (6.4) is satisfied at every point of $\Pi_1 \cup \Pi_2$.

$$(\alpha E - B)(L \times BL) \neq 0, \quad T^*(\alpha E - B)(L \times BL) = 0. \tag{7.3_3}$$

In this case operator T^* must be degenerate. In bases e_1, e_2, e_3 , T^* is represented by the matrix

$$T^* = \begin{pmatrix} 0 & (J_3 - J_1)L_3 & (J_1 - J_2)L_2 \\ (J_2 - J_3)L_3 & 0 & (J_1 - J_2)L_1 \\ (J_2 - J_3)L_2 & (J_3 - J_1)L_1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & L_3 & L_2 \\ L_3 & 0 & L_1 \\ L_2 & L_1 & 0 \end{pmatrix} \begin{pmatrix} J_2 - J_3 & 0 & 0 \\ 0 & J_3 - J_1 & 0 \\ 0 & 0 & J_1 - J_2 \end{pmatrix},$$

$$\det T^* = 2(J_2 - J_3)(J_3 - J_1)(J_1 - J_2)L_1L_2L_3.$$

As the body is asymmetric, that is $J_1 > J_2 > J_3$ then, the degeneration of T^* implies one of the equalities $L_i = 0$, i.e. $\bar{L} \perp \bar{e}_i$. It can be easily shown that it implies $(\alpha E - B)(L \times BL) = 0$ or $T^*(\alpha E - B)(L \times BL) \neq 0$, so (7.3₃) does not give new solutions as compared with (7.3₂).

Let us assume now that

$$T^*(\alpha E - B)X \neq 0. \quad (7.4)$$

Then (7.2) defines a plane in \mathbf{R}^3 , i.e. the trajectory lying on the surface (6.4) must be a plane curve. As it is known, any trajectory of Eq. (1.1) which is a plane curve satisfies one of the equalities $2\mathcal{E} = J_i \mathcal{M}^2$ ($i = \overline{1, 3}$). Each of these trajectories is a stationary point lying on one of the axes \bar{e}_i or an arc of the circle lying in one of the planes Π_k ($k = 1, 2$). Thus if (7.4) holds, then the sufficient condition of local controllability of system (1.2) (see Proposition 3) is fulfilled at every point of \mathbf{R}^3 with the possible exception of the points of the subset $H = \Pi_1 \cup \Pi_2 \cup \bar{e}_1 \cup \bar{e}_2 \cup \bar{e}_3$. It is obvious that if $L \notin H$ then for the set $\mathbf{R}^3 \setminus H$, consisting of four connected components, the requirements of Proposition 2 are met, so for every $\tilde{M} \in \mathbf{R}^3$ the attainable set $D_{\tilde{M}}$ coincides with \mathbf{R}^3 .

Let us note that if L satisfies (7.3), but is not collinear to e_i ($i = \overline{1, 3}$) then in this case too (as it was shown above) the conditions of local controllability may be violated only at the points of H , so if L does not belong to H , then for every \tilde{M} $D_{\tilde{M}} = \mathbf{R}^3$.

So we ascertain that if $L \notin H$, then for every \tilde{M} the attainable set $D_{\tilde{M}}$ of the control system (1.2) coincides with \mathbf{R}^3 .

8. Singular positions of the axis \bar{L} ($L \in H$)

8A. Let axis \bar{L} lie in one of the planes Π_1, Π_2 , for example in the plane Π_1 ($L \notin \bar{e}_2$). Then plane Π_1 is invariant for the control systems (1.2) and (1.2⁻) (the corresponding vector fields are tangent to Π_1). Hence for every $\tilde{M} \in \Pi_1$ the attainable set $D_{\tilde{M}} \subset \Pi_1$. The detailed construction of the attainable set $D_{\tilde{M}}$ for the case $\tilde{M}, L \in \Pi_i$ ($i = 1, 2$) will be given in further publications.

The set $\mathbf{R}^3 \setminus \Pi_1$ consists of two open semispaces Q_1, Q_2 . Let us prove that for every $\tilde{M} \in Q_i$ the attainable set $D_{\tilde{M}}$ coincides with Q_i . In fact the semispace Q_i is invariant for control system (1.2), when $L \in \Pi_1$, i.e. $D_{\tilde{M}} \subseteq Q_i$, and the sufficient conditions of local controllability for system (1.2) are violated only at the points of the set $Q_i \cap \Pi_2$, since these points belong to the aperiodic solutions of Eq. (1.1). The set $Q_i \setminus (Q_i \cap \Pi_2)$ consists of two connected open components. Since $L \in \Pi_1$ and L is not collinear to e_2 , then $L \notin \Pi_2$ and for the system (1.2) restricted onto the set Q_i the conditions of Proposition 2 are fulfilled, so for every $\tilde{M} \in Q_i$ $D_{\tilde{M}} = Q_i$.

8B. Let L be collinear to e_1 , for example $L = e_1$. Then $V_0 = L = e_1$, V_2 is collinear to V_0 ,

$$V_1 = BL \times M - L \times BM = (J_1 e_1 \times M) - e_1 \times BM = ((B - J_1 E)M) \times e_1.$$

Let us consider in \mathbf{R}^3 a two-dimensional distribution generated by the vector fields $V_0 = e_1$, $V_1 = ((B - J_1 E)M) \times e_1$. A moment's consideration shows that this distribution is integrable; corresponding integral manifolds are the elliptic cylinders with an element parallel to \bar{e}_1 . The ellipses lying at the foot of these cylinders are the integral curves of the equation

$$\dot{M} = ((B - J_1 E)M) \times e_1,$$

which is equivalent to the system

$$\dot{M}_1 = 0, \quad \dot{M}_2 = (J_1 - J_3)M_3, \quad \dot{M}_3 = (J_2 - J_1)M_2. \quad (8.1)$$

The trajectories of (8.1) are ellipses defined by the equation

$$(J_1 - J_2)M_2^2 + (J_1 - J_3)M_3^2 = c_1 = \text{const.} \quad (8.2)$$

Equation (8.2) considered in $\mathbf{R}^3 = \{(M_1, M_2, M_3)\}$, defines an elliptic cylinder with an element parallel to \bar{e}_1 .

Let us prove that the elliptic cylinders (8.2) are invariant manifolds for system (1.2). Since $L = e_1$ is tangent to all cylinders (8.2), then it is sufficient to prove that the vector field $F = M \times BM$ is tangent to cylinder (8.2) at every point. In fact the mixed product

$$\begin{aligned} \langle M \times BM, V_0 \times V_1 \rangle &= \langle M \times BM, e_1 \times ((B - J_1 E)M \times e_1) \rangle = \\ &= \langle M \times BM, \langle e_1, e_1 \rangle (B - J_1 E)M - \langle e_1, (B - J_1 E)M \rangle e_1 \rangle = \\ &= \langle M \times BM, (B - J_1 E)M \rangle - \langle M \times BM, e_1 \rangle \langle (B - J_1 E)e_1, M \rangle = \\ &= \langle M \times BM, (B - J_1 E)M \rangle = \langle M \times BM, BM \rangle - J_1 \langle M \times BM, M \rangle = 0. \end{aligned}$$

Let us consider the restriction of system (1.2) on the arbitrary cylinder (8.2). If constant c_1 in (8.2) differs from zero, then the values of the fields V_0 , V_1 are linearly independent at every point of the cylinder. Thus every point of cylinder (8.2) which belongs to the periodic trajectory of Eq. (1.1) is a point of local controllability for the restricted system (1.2). On every cylinder C defined by (8.2), the set of points belonging to the unperiodic trajectories of (1.1), consists of two circles γ_1, γ_2 lying in the planes Π_1, Π_2 correspondingly. The set $C \setminus (\gamma_1 \cup \gamma_2)$ consists of four connected open components. It is easy to show that when $L = e_1$ then the conditions of Proposition 2 are fulfilled for the system (1.2) restricted onto C , so for every $\tilde{M} \in C$, $D_{\tilde{M}} = C$. If \tilde{M} lies on axis \bar{e}_1 , then obviously $D_{\tilde{M}}$ coincides with \bar{e}_1 .

Similarly if $L = e_3$ then the invariant manifolds of control system (1.2) are elliptic cylinders with an element parallel to \bar{e}_3 , defined by the equations

$$(J_1 - J_3)M_1^2 + (J_2 - J_3)M_2^2 = c_2 = \text{const.},$$

and axis \bar{e}_3 . Cylinders (8.3) and axis \bar{e}_3 are the attainable sets $D_{\tilde{M}}$ for every one of their points \tilde{M} .

When $L = e_2$ the invariant manifolds of the system (1.2) are axis \bar{e}_2 and hyperbolic cylinders of two sheets with an element parallel to \bar{e}_2 , defined by the equation

$$(J_1 - J_2)M_1^2 - (J_2 - J_3)M_3^2 = c_3 = \text{const.} \quad (8.4)$$

If constant c_3 in (8.4) equals zero, the hyperbolic cylinder degenerates into the pair of planes Π_1, Π_2 . The control of motion of the angular momentum M in the plane Π_k , when $L = e_2$, will be considered in further publications. Each of the sheets of the hyperbolic cylinders (8.4) and axis \bar{e}_2 is an attainable set $D_{\tilde{M}}$ for its every point \tilde{M} .

9. Attainable sets in the space of angular momenta (results)

The results of the work are as follows.

9A. If the axis \bar{L} neither coincide with one of the principal axes nor lies in one of the planes Π_1, Π_2 , then for every \tilde{M} the attainable set $D_{\tilde{M}}$ of control system (1.2) coincides with the whole \mathbf{R}^3 .

9B. If \bar{L} coincides with the major axis \bar{e}_1 , then \mathbf{R}^3 is stratificated to two-dimensional invariant manifolds of the system (1.2), which are the elliptic cylinders with an element parallel to \bar{e}_1 . These cylinders are defined by Eq. (8.2)

$$(J_1 - J_2)M_2^2 + (J_1 - J_3)M_3^2 = \text{const.}$$

Axis \bar{e}_1 is also an invariant manifold for control system (1.2). The attainable set $D_{\tilde{M}}$ coincides with that of the elliptic cylinders (8.2), which contain \tilde{M} . If \tilde{M} lies on \bar{e}_1 , then the attainable set $D_{\tilde{M}}$ coincides with \bar{e}_1 .

9C. If \bar{L} coincides with \bar{e}_3 , then \mathbf{R}^3 is stratificated to two-dimensional invariant manifolds of control system (1.2), which are elliptic cylinders with an element parallel to \bar{e}_3 . These cylinders are defined by Eq. (8.3)

$$(J_1 - J_3)M_1^2 + (J_2 - J_3)M_2^2 = \text{const.}$$

Each of these cylinders and axis \bar{e}_3 are the attainable sets $D_{\tilde{M}}$ for each point \tilde{M} they contain.

9D. If \bar{L} coincides with the middle axis \bar{e}_2 , then the invariant submanifold of control system (1.2) are the hyperbolic cylinders of two sheets with an element parallel to \bar{e}_2 defined by Eq. (8.4)

$$(J_1 - J_2)M_1^2 - (J_2 - J_3)M_3^2 = c_3 = \text{const.}$$

Axis \bar{e}_2 is also an invariant submanifold of system (1.2). If constant c_3 in (8.4) equals zero then the hyperbolic cylinder degenerates into the pair of planes Π_1, Π_2 .

Each of the sheets of the hyperbolic cylinders (8.4) and axis \bar{e}_2 are the attainable sets $D_{\tilde{M}}$ for every point \tilde{M} they contain. Let us note that the case $c_3=0, \tilde{M} \notin \bar{e}_2$ needs specific study.

9E. If \bar{L} lies in one of the planes $\Pi_k (k=1, 2)$ and does not coincide with \bar{e}_2 , and $\tilde{M} \notin \Pi_k$, then an attainable set $D_{\tilde{M}}$ coincides with one of those open semispaces of \mathbf{R}^3 separated by the plane Π_k which contains \tilde{M} . In particular, it means that in this case the zero value of the vector of angular momentum is not attainable.

9F. If L lies in one of the planes Π_1, Π_2 and \tilde{M} lies in the same plane Π_k , then the attainable set $D_{\tilde{M}} \subset \Pi_k$. The description of the attainable set $D_{\tilde{M}}$ in this case will be given in a separate publication.

Remark. It follows from the formulated results that if at least one of two vectors L, \tilde{M} does not lie in $\Pi_k (k=1, 2)$ then the attainable set $D_{\tilde{M}}$ does not depend upon constant \varkappa , which bounds the control.

Besides that the relation $\hat{M} \in D_{\tilde{M}}$ is in this case symmetrical: if \hat{M} can be attained from \tilde{M} , then \tilde{M} can be attained from \hat{M} . It can be shown that the situation is quite different when $\tilde{M}, L \in \Pi_k$.

10. The control of angular momentum and attitude of the body

Let us consider now the control of the angular momentum and the attitude in parallel. In this case the phase space of the control system is $SO(3) \times \mathbf{R}^3$ (where $SO(3)$ is a group of rotations of \mathbf{R}^3). The equations of motion in notations of our article are

$$\begin{aligned} \dot{Q} &= Q(\widehat{BM}), \\ \dot{M} &= M \times BM + Lu \end{aligned} \tag{10.1}$$

where $Q \in SO(3)$ is a matrix defining the attitude of the body, $(\widehat{BM}) \in SO(3)$ is an antisymmetric (3×3) -matrix

$$(\widehat{BM}) = \begin{pmatrix} 0 & -J_3 M_3 & J_2 M_2 \\ J_3 M_3 & 0 & -J_1 M_1 \\ -J_2 M_2 & J_1 M_2 & 0 \end{pmatrix}.$$

It is shown in [3] that if \bar{L} is in general position (i.e. L satisfies the conditions of 9A) then for every point $(\tilde{Q}, \tilde{M}) \in SO(3) \times \mathbf{R}^3$ the attainable set $D_{\tilde{Q}, \tilde{M}}$ of the control system (10.1) coincides with the whole phase space $SO(3) \times \mathbf{R}^3$.

Let us supplement this result by the description of the attainable sets $D_{\tilde{Q}, \tilde{M}}$ for the singular positions of $\bar{L} (L \in H)$.

10A. If \bar{L} coincides with \bar{e}_i ($i = 1, 3$) then for every $\tilde{M} \notin H$ the attainable set $D_{\tilde{Q}, \tilde{M}}$ of control system (10.1) coincides with $SO(3) \times D_{\tilde{M}}$, where $D_{\tilde{M}}$ is an attainable set of system (1.2) (elliptic or hyperbolic cylinder) described in one of the i.i 9B-9D correspondingly.

10B. If \bar{L} lies in one of the planes Π_k , $L \notin \bar{e}_2$, $\tilde{M} \notin \Pi_k$ then for every \tilde{Q} the attainable set $D_{\tilde{Q}, \tilde{M}}$ coincides with $SO(3) \times D_{\tilde{M}}$, where $D_{\tilde{M}}$ is an attainable set of system (1.2) (semispace) described in 9E.

10C. If $L, \tilde{M} \in \bar{e}_i$ then the attainable set of the control system (10.1) is a Cartesian product of \bar{e}_i lying in the space \mathbf{R}^3 of angular momenta and one-dimensional subgroup of $SO(3)$, consisting of all rotations around $\bar{L} = \bar{e}_i$, i.e. Cartesian product of circumference and straight line.

11. Conclusion

In this work we studied the rotation of the rigid body, which is controlled by the torque directed along the axis \bar{L} which is fixed in the body. It was proved that if \bar{L} is in general position (\bar{L} does not belong to the singular set whose measure equals zero) then the body is completely controllable, i.e. for the finite time it can reach simultaneously any preassigned value of the angular momentum vector and given attitude. All the cases of singular disposed \bar{L} have been studied too, and the attainable sets in the phase space of the corresponding control system were described.

References

1. Chernousko, F. L., Akulenko, L. D., Sokolov, B. N., Control of oscillations. "Nauka", Moscow, 1980.
2. Lobry, C., Controlabilite des systemes non lineaires. "Outils et modeles math. autom. Anal. syst. et trait signale." Vol. 1, Paris, 1981, 187-214
3. Bonnard, B., Controle de latitude d'un satellite rigide. R.A.I.R.O Automatique, Systems Analysis and Control, vol. 16, No. 1, 1982, pp. 85-93.
4. Arnold, V. I., Mathematical methods of classical mechanics. "Nauka", Moscow, 1974.
5. Hermes, N., On local and global controllability. SIAM. J. Control, vol. 12, 1974, 252-261.

Управление вращением асимметричного твердого тела

А. А. АГРАЧЁВ, А. В. САРЫЧЕВ
(Москва)

Рассматривается асимметричное твердое тело, вращающееся вокруг центра масс. В теле фиксирована проходящая через центр масс ось L , вдоль которой прилагается управляющий момент, пропорциональный по величине текущему значению скалярного управления $u(t)$ ($|u(t)| \leq \kappa$). Ставится вопрос о возможности перехода за конечное нефиксированное время T от вращения с мгновенной

угловой скоростью $\tilde{\Omega}$ к вращению с мгновенной угловой скоростью $\tilde{\Omega}$. Ради упрощения выкладок результаты формулируются не для угловой скорости, а для вектора кинетического момента M , линейно связанного с Ω . Для любого начального значения \tilde{M} построено множество достижимости $D_{\tilde{M}}$ — множество значений кинетического момента, достижимых из \tilde{M} за конечное нефиксированное время T .

В формулировке результатов фигурирует пара особых плоскостей Π_1, Π_2 , содержащих неперриодические траектории $M(t)$ кинетического момента при свободном вращении тела. В пересечении плоскостей Π_1, Π_2 лежит ось \tilde{e}_2 среднего момента инерции тела.

Получены следующие результаты.

1. Если ось \tilde{L}_2 не совпадает ни с одной из главных осей инерции и не лежит ни в одной из плоскостей Π_1, Π_2 , то для любого \tilde{M} множество достижимости $D_{\tilde{M}}$ совпадает со всем пространством \mathbb{R}^3 кинетических моментов.

2. Если ось \tilde{L} совпадает с осью \tilde{e}_1 (\tilde{e}_3) наименьшего (наибольшего) момента инерции тела, то множество достижимости $D_{\tilde{M}}$ представляет собой эллиптический цилиндр с образующей, параллельной \tilde{e}_1 (\tilde{e}_3) или ось \tilde{e}_1 (\tilde{e}_3), если $\tilde{M} \in \tilde{e}_1$ ($\tilde{M} \in \tilde{e}_3$).

3. Если ось \tilde{L} совпадает с осью \tilde{e}_2 среднего момента инерции и $\tilde{M} \notin (\Pi_k \setminus \tilde{e}_2)$, то множество достижимости $D_{\tilde{M}}$ есть гиперболический цилиндр с образующей, параллельной \tilde{e}_2 , или сама ось \tilde{e}_2 , если $\tilde{M} \in \tilde{e}_2$.

Если ось \tilde{L} лежит в одной из плоскостей Π_k ($k = 1, 2$) и не совпадает с \tilde{e}_2 , то $\forall \tilde{M} \notin \Pi_k$ множество достижимости $D_{\tilde{M}}$ совпадает с тем из двух открытых полупространств, получающихся при делении \mathbb{R}^3 плоскостью Π_k , в котором лежит \tilde{M} .

5. Если ось \tilde{L} и вектор \tilde{M} лежат одновременно в плоскости Π_k , то $D_{\tilde{M}} \subset \Pi_k$. Подробное описание множества достижимости в этом случае будет дано во второй части работы.

6. Из сказанного выше следует, что если ось L и вектор \tilde{M} не лежат одновременно в одной из плоскостей Π_k , то множество достижимости $D_{\tilde{M}}$ не зависит от константы κ , ограничивающей управление.

В работе рассмотрено также одновременное управление угловой скоростью и ориентацией тела. Фазовое пространство системы в этом случае есть $SO(3) \times \mathbb{R}^3$, где $SO(3)$ — группа вращений пространства \mathbb{R}^3 , то есть группа ортогональных (3×3) -матриц Q , $\det Q = 1$. Построено множество достижимости $D_{\tilde{Q}, \tilde{M}}$ для такой системы. Доказано, что

7. Если \tilde{L} удовлетворяет условиям п. 1, то $\forall \tilde{Q}, \tilde{M}$ — множество достижимости $\tilde{D}_{\tilde{Q}, \tilde{M}} = SO(3) \times \mathbb{R}^3$.

8. Если L, \tilde{M} удовлетворяют одному из условий п.п. 2–4, и \tilde{M} не лежит на главной оси инерции, то $D_{\tilde{Q}, \tilde{M}} = SO(3) \times D_{\tilde{M}}$, где $D_{\tilde{M}}$ — множество достижимости, описанное в соответствующем из п.п. 2–4.

9. Если $L, \tilde{M} \in \tilde{e}_i$, то $\forall \tilde{Q}$ множество достижимости $D_{\tilde{Q}, \tilde{M}}$ есть декартово произведение прямой $\tilde{e}_i \subset \mathbb{R}^3$ и одномерной подгруппы группы $SO(3)$, состоящей из поворотов вокруг оси $L = \tilde{e}_i$.

А. А. Аграчев

ВИНИТИ АН СССР, ГКНТ,

СССР, 125219 Москва А-219,

Балтийская ул., 14

ON OPTIMAL CONTROL PROBLEM WITH INITIAL STATE NOT A PRIORI GIVEN

W. KOTARSKI, A. KOWALEWSKI
(Katowice, Kraków)

(Received 1 August, 1982)

In this paper an optimal control problem for a system described by a linear partial differential equation of the parabolic type with the Dirichlet's boundary condition is considered.

We impose some constraints on the control. The performance functional has the integral form. The control time T is fixed. The initial condition is not given by a known function but it belongs to a certain set (the initial state is not a priori given).

The problem formulated in the paper describes the process of optimal heating in which we have some freedom in choosing the initial temperature of the heated object.

We also present a particular example in which the set of admissible controls and the one of initial conditions are given by means of the norm constraints.

The application of the well-known projective gradient method in the Hilbert space allows us to obtain the numerical solution for our optimization problem.

The problem considered here is the generalization (more general performance index) of similar type as was presented by Prof. J. L. Lions at his lecture in Stefan Banach International Mathematical Center in Warsaw in 1980 during the Semester on Optimal Control Theory.

1. Statement of optimal control problem. Optimization theorem

We consider the parabolic equation describing the dynamic of controlled system

$$\frac{\partial y}{\partial t} + A(t)y = u \quad x \in \Omega, \quad t \in (0, T) \quad (1.1)$$

$$y(x, t) = 0 \quad x \in \Gamma, \quad t \in (0, T) \quad (1.2)$$

$$y(x, 0) \in K \quad x \in \Omega \quad (1.3)$$

where $\Omega \in R^n$ is a bounded, open set with boundary Γ , which is a C^∞ -manifold of dimension $(n-1)$. Locally, Ω is totally on one side of Γ ,

$K \in H_0^1(\Omega)$ is a closed, convex subset with non-empty interior in the space $H_0^1(\Omega)$,
 $u \equiv u(x, t)$, $y \equiv y(x, t)$.

The right-hand side of Eq. (1.1) and the initial condition (1.3) are not continuous functions, but they are measurable ones belonging to L^2 or L^∞ spaces. Therefore we shall look for the solution of Eqs (1.1)–(1.3) in some Sobolev spaces.

Let us denote by $Y = L^2(0, T; H_0^1(\Omega))$ the space of states, and by $U = L^2(Q)$ the space of controls.

The operator $A(t)$ has the form

$$A(t)y = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x, t) \frac{\partial y(x, t)}{\partial x_j} \right)$$

and the functions $a_{ij}(x, t)$ satisfy the conditions

$$\sum_{i,j=1}^n a_{ij}(x, t) \xi_i \xi_j \geq \alpha \sum_{i=1}^n \xi_i^2, \quad \alpha > 0, \quad \forall (x, t) \in Q, \quad \forall \xi_i \in R^1$$

$$a_{ij}(x, t) \in C^1(Q) \quad Q = \Omega \times (0, T).$$

It is known that if initial state $y(x, 0)$ is an arbitrary fixed function, then Eqs (1.1)–(1.3) have the unique solution $y \in W(0, T)$ continuously dependent on the right-hand side and on the initial condition (Theorem 1.2 [6]) with the assumptions mentioned above.

The control time T is fixed in our problem.

The performance functional is given by

$$J(y, u) = \int_Q F(x, t, y, u) dx dt \rightarrow \min \quad (1.4)$$

where

$F: \Omega \times [0, T] \times R^1 \times R^1 \rightarrow R^1$ satisfies the following conditions:

A1) $F(x, t, y, u)$ is continuous with respect to (x, t, y, u)

A2) there exist $F_y(x, t, y, u)$, $F_u(x, t, y, u)$ which are continuous with respect to (x, t, y, u)

A3) $F(x, t, y, \cdot)$ is strictly convex with respect to u , i.e.

$$F(x, t, y, \lambda u_1 + (1 - \lambda)u_2) < \lambda F(x, t, y, u_1) + (1 - \lambda)F(x, t, y, u_2)$$

$$\forall u_1, u_2 \in R^1, u_1 \neq u_2, (x, t, y) \in \Omega \times [0, T] \times R^1, \lambda \in (0, 1)$$

A4) $F(x, t, \cdot, u)$ is strictly convex with respect to y , i.e.

$$F(x, t, \lambda y_1 + (1 - \lambda)y_2, u) < \lambda F(x, t, y_1, u) + (1 - \lambda)F(x, t, y_2, u)$$

$$\forall y_1, y_2 \in R^1, y_1 \neq y_2, (x, t, u) \in \Omega \times [0, T] \times R^1, \lambda \in (0, 1)$$

or instead of (A3)–(A4) we can assume the following condition:

A5) $F(x, t, \cdot, \cdot)$ is strictly convex with respect to the pair (y, u) , i.e.

$$F(x, t, \lambda y_1 + (1 - \lambda)y_2, \lambda u_1 + (1 - \lambda)u_2) < \lambda F(x, t, y_1, u_1) + (1 - \lambda)F(x, t, y_2, u_2)$$

$$\forall y_1, y_2, u_1, u_2 \in R^1, (y_1, u_1) \neq (y_2, u_2), \lambda \in (0, 1).$$

We assume the following constraints on controls: $u \in U_{ad}$ is a closed, convex subset of U with non-empty interior in the space $L^2(Q)$. (1.5)

Remark 1. In the following, to abbreviate the formulas, we shall write F, F_y, F_u, p without the arguments.

We shall also denote by

$$y(0) \equiv y(x, 0), \quad p(0) \equiv p(x, 0), \quad y(T) \equiv y(x, T), \quad p(T) \equiv p(x, T).$$

The optimal control problem (1.1)–(1.5) will be solved as the optimization one in which the function u and the initial condition $y(0)$ are the unknown functions. It is easy to show that optimal control u^0 which suits to optimal condition $y^0(0)$ gives the smallest value for the performance functional. Every other optimal control $u^{0'}$ which suits to initial condition $y(0) \in K$ gives the greater value for the performance functional. So proceeding in this way with the lack of additional information about the set K and the performance functional, we reach a certain compromise.

Making use of Milutin–Dubovicki's Theorem we shall derive the necessary and sufficient conditions of optimality for the optimization problem (1.1)–(1.5).

The idea of Milutin–Dubovicki's method was particularly described in [4, 5], therefore we shall not present this method here.

The solution of the stated optimal control problem is equivalent to seeking a triplet $(y^0(0), y^0, u^0) \in E = H_0^1(\Omega) \times Y \times U$ which satisfies Eqs (1.1)–(1.3) and minimizing the performance functional (1.4) with the constraints on controls u (1.5).

We formulate the necessary and sufficient conditions of optimality in the form of Theorem 1.

Theorem 1. The solution of the optimization problem (1.1)–(1.5) exists and is unique with the assumptions mentioned above; the necessary and sufficient conditions of the optimality are characterized by the following system of partial differential equation and inequalities:

State equation

$$\frac{\partial y^0}{\partial t} + A(t)y^0 = u^0, \quad x \in \Omega, \quad t \in (0, T), \quad (1.6)$$

$$y^0(x, t) = 0, \quad x \in \Gamma, \quad t \in (0, T). \quad (1.7)$$

Adjoint equation

$$-\frac{\partial p}{\partial t} + A^*(t)p = F_y, \quad x \in \Omega, \quad t \in (0, T), \quad (1.8)$$

$$p(x, t) = 0, \quad x \in \Gamma, \quad t \in (0, T), \quad (1.9)$$

$$p(x, T) = 0, \quad x \in \Omega. \quad (1.10)$$

Maximum condition

$$\int_{\Omega} (p + F_u)(u - u^0) dx dt \geq 0, \quad \forall u \in U_{ad} \quad (1.11)$$

$$\int_{\Omega} p(0)[k - y^0(0)] dx \geq 0, \forall k \in K \quad (1.12)$$

where " 0 " denotes the optimal element,

$A^*(t)$ is the adjoint operator to $A(t)$,

F_y, F_u are the Fréchet derivatives of F with respect to y, u , respectively in the point (y^0, u^0) .

Outline of the proof

According to Milutin-Dubovicki's Theorem [4] we approximate the set representing the inequality constraints by the regular admissible cone, the equality constraint by the regular tangent cone and the performance functional by the regular improvement cone.

a) Analysis of the equality constraint

The set Q_1 representing the equality constraint has the form

$$Q_1 = \left\{ \begin{array}{ll} \frac{\partial y}{\partial t} + A(t)y = u & x \in \Omega, \quad t \in (0, T) \\ y(0), y, u \in E; & \\ y(x, t) = 0 & x \in \Gamma, \quad t \in (0, T) \\ y(x, 0) = y(0) & y \in \Omega \end{array} \right\}. \quad (1.13)$$

We construct the regular tangent cone of the set Q_1 using the Lusternik Theorem (Theorem 9.1 [4]).

For this purpose we define the operator $P(y(0), y, u)$ in the form

$$P(y(0), y, u) = \left(\frac{\partial y}{\partial t} + Ay - u, y(x, 0) - y(0), y(x, t)|_{x \in \Gamma} \right). \quad (1.14)$$

The operator $P(y(0), y, u)$ is the mapping from the space $H_0^1(\Omega) \times L^2(0, T; H_0^1(\Omega)) \times L^2(Q)$ into the space $L^2(0, T; H^{-1}(\Omega)) \times H_0^1(\Omega) \times L^2(0, T; H_0^1(\Omega))$.

The Fréchet differential of the operator $P(y(0), y, u)$ can be written in the following form

$$P(y^0(0), y^0, u^0)(\bar{y}(0), \bar{y}, \bar{u}) = \left(\frac{\partial \bar{y}}{\partial t} + A\bar{y} - \bar{u}, \bar{y}(x, 0) - \bar{y}(0), \bar{y}(x, t)|_{x \in \Gamma} \right). \quad (1.15)$$

Really, $\frac{\partial}{\partial t}$ (Theorem 2.8 [7]) and $A(t)$ (Theorem 2.1 [6]) are linear and bounded mappings.

Using Theorem 1.2 [6] we can prove that $P'(y^0(0), y^0, u^0)$ is the operator "one to one" from the space $H_0^1(\Omega) \times L^2(0, T; H_0^1(\Omega)) \times L^2(Q)$ onto $L^2(0, T; H^{-1}(\Omega)) \times H_0^1(\Omega) \times L^2(0, T; H_0^1(\Omega))$.

Considering that there are fulfilled the assumptions of Lusternik's Theorem we can write down the regular tangent cone for the set Q_1 in the point $(y^0(0), y^0, u^0)$ in the form

$$RTC(Q_1, (y^0(0), y^0, u^0)) = \{(\bar{y}(0), \bar{y}, \bar{u}) \in E; P'(y^0(0), y^0, u^0)(\bar{y}(0), \bar{y}, \bar{u}) = 0\}. \quad (1.16)$$

It is easy to notice that it is a subspace. Therefore using Theorem 10.1 [4] we know the form of the functional belonging to the adjoint cone

$$f_1(\bar{y}(0), \bar{y}, \bar{u}) = 0 \quad \forall (\bar{y}(0), \bar{y}, \bar{u}) \in RTC(Q_1, (y^0(0), y^0, u^0)). \quad (1.17)$$

b) Analysis of the constraint on controls

The set $Q_2 = K \times Y \times U_{ad}$ representing the inequality constraint is a closed and a convex one with non-empty interior in the space E .

Using Theorem 10.5 [4] we find the functional belonging to the adjoint regular admissible cone, i.e.

$$f_2(\bar{y}(0), \bar{y}, \bar{u}) \in [RAC(Q_2, (y^0(0), y^0, u^0))]^*.$$

We can note if E_1, E_2, E_3 are three linear topological spaces, then the adjoint space to $E = E_1 \times E_2 \times E_3$ has the form

$$E^* = \{f = (f_1, f_2, f_3), \quad f_1 \in E_1^*, f_2 \in E_2^*, f_3 \in E_3^*\}$$

and

$$f(x) = f_1(x_1) + f_2(x_2) + f_3(x_3).$$

So we note the functional $f_2(\bar{y}(0), \bar{y}, \bar{u})$ as follows

$$f_2(\bar{y}(0), \bar{y}, \bar{u}) = f_1'''(\bar{y}) + f_2'(\bar{y}(0)) + f_2''(\bar{u}) \quad (1.18)$$

where $f_1'''(\bar{y}) = 0, \forall y \in Y$ (Theorem 10.1 [4])

$f_2'(\bar{y}(0)), f_2''(\bar{u})$ are the support functionals to the sets K and U_{ad} in the points $y^0(0)$ and u^0 , respectively (Theorem 10.5 [4]).

c) Analysis of the performance functional

Using Theorem 7.5 [4] we find the regular improvement cone

$$RFC(J, (y^0(0), y^0, u^0)) = \{(\bar{y}(0), \bar{y}, \bar{u}) \in E; J'(y^0(0), y^0, u^0)(\bar{y}(0), \bar{y}, \bar{u}) < 0\}, \quad (1.19)$$

where $J'(y^0(0), y^0, u^0)(\bar{y}(0), \bar{y}, \bar{u})$ is the Fréchet differential of the performance functional. By the assumptions (A1), (A2) this differential exists (it can be proved in the same way as in example 7.2 [4]) and can be written as

$$\int_Q (F_y \bar{y} + F_u \bar{u}) dx dt.$$

On the basis of Theorem 10.2 [4] we find the functional belonging to the adjoint regular improvement cone, which has the form

$$f_3(\bar{y}(0), \bar{y}, \bar{u}) = \lambda_0 \int_Q (F_y \bar{y} + F_u \bar{u}) dx dt \quad (1.20)$$

where $\lambda_0 > 0$.

d) Analysis of Euler–Lagrange's equation

The Euler–Lagrange equation for our optimization problem has the form

$$\sum_{i=1}^3 f_i = 0. \quad (1.21)$$

Let $p(x, t)$ be the solution of (1.8)–(1.10) for $u^0, y^0(0), y^0$ and denote by \bar{y} the solution of $P'(\bar{y}(0), \bar{y}, \bar{u}) = 0$ for any fixed $\bar{y}(0)$ and \bar{u} . Next taking into account (1.17), (1.18) and (1.20) we can express (1.21) in the form

$$\begin{aligned} f'_2(\bar{y}(0)) + f'_2(\bar{u}) &= \lambda_0 \int_Q F_y \bar{y} dx dt + \lambda_0 \int_Q F_u \bar{u} dx dt \\ \forall (\bar{y}(0), \bar{y}, \bar{u}) &\in RTC(Q_1, (\bar{y}(0), \bar{y}, \bar{u})). \end{aligned} \quad (1.22)$$

We transform the first component of the right-hand side of (1.22) introducing the adjoint variable by Eq. (1.8) and using formulas (1.10), (1.15) and (1.16).

In turn, we get

$$\begin{aligned} \lambda_0 \int_Q F_y \bar{y} dx dt &= \lambda_0 \int_Q \left(-\frac{\partial p}{\partial t} + A^*(t)p \right) \bar{y} dx dt = \\ &= -\lambda_0 \int_Q \frac{\partial p}{\partial t} \bar{y} dx dt + \lambda_0 \int_Q p A(t) \bar{y} dx dt = \\ &= \lambda_0 \int_{\Omega} [-p(T) \bar{y}(T) + p(0) \bar{y}(0)] dx - \lambda_0 \int_Q p \frac{\partial \bar{y}}{\partial t} dx dt + \\ &+ \lambda_0 \int_Q p A(t) \bar{y} dx dt = \lambda_0 \int_{\Omega} p(0) \bar{y}(0) dx + \lambda_0 \int_Q p \bar{u} dx dt. \end{aligned} \quad (1.23)$$

Substituting (1.23) into (1.22), we obtain

$$f'_2(\bar{y}(0)) + f''_2(\bar{u}) = \lambda_0 \int_{\Omega} p(0) \bar{y}(0) dx + \lambda_0 \int_Q (p + F_u) \bar{u} dx dt. \quad (1.24)$$

Using the definition of the support functional [4] and dividing both members of the obtained inequality by λ_0 , we finally get

$$\int_{\Omega} p(0) (k - y^0(0)) dx + \int_Q (p + F_u) (u - u^0) dx dt \geq 0 \\ \forall k \in K, \quad \forall u \in U_{ad}. \quad (1.25)$$

This last inequality is equivalent to the maximum conditions (1.11), (1.12).

In order to prove the sufficiency of the derived conditions of optimality, we use the fact that the constraints and the performance functional are convex and Slater's condition is satisfied (Theorem 15.2 [4]).

Really, there exists a point $(\tilde{y}(0), \tilde{y}, \tilde{u}) \in \text{int } Q_2$ such that $(\tilde{y}(0), \tilde{y}, \tilde{u}) \in Q_1$. This fact follows immediately from the existence of non-empty interior in the set Q_2 and Theorem 1.2 [6] about solution of a parabolic equation.

The uniqueness of the triplet $(y^0(0), y^0, u^0)$ follows from the strict convexity of the performance functional (1.4) (assumptions (A3) and (A4) or (A5)).

This last remark completes the proof of Theorem 1.

Table 1

Maximum condition and state equation which suit to different sets K and U_{ad}

Case	The set of initial conditions K	The set of admissible controls u U_{ad}	Maximum condition	State equation
1	$H_0^1(\Omega)$	$L^2(Q)$	$p = -F_u$ $p(0) = 0$	(1.6), (1.7)
2	$H_0^1(\Omega)$	U_{ad}	(1.11) $p(0) = 0$	(1.6), (1.7)
3	K	$L^2(Q)$	$p = -F_u$ (1.12)	(1.6), (1.7)
4	$\{0\}$	U_{ad}	(1.11)	(1.6), (1.7) $y^0(0) = 0$
5	$\{0\}$	$L^2(Q)$	$p = -F_u$	(1.6), (1.7) $y^0(0) = 0$

Remark 2. Theorem 1 can be obtained as the generalization of similar ones given in [6] for quadratic performance index and can be also proved using the methods of [6].

Remark 3. On the basis of Theorem 1 we can derive various modifications in particular cases of the sets K and U_{ad} . In Table 1 we indicate maximum condition and state equation which suit to different sets K and U_{ad} .

For instance, case 1 suits to the situation when there are no constraints on the control u and on the initial condition $y(0)$, while case 4 suits to the one when the initial condition is exactly determining.

Remark 4. We must notice that the conditions of optimality given in Theorem 1 do not provide any analytical formulas for optimal control. Thus we must resign from the exact determining of the optimal control and must use approximation methods. For instance, the ones used in particular case of functional in [5].

Further, let us consider a particular problem (1.1)–(1.5) in which

$$U_{ad} = \{u \in L^2(Q); \int_Q u^2(x, t) dx dt \leq M^2, \quad M \text{ is const.}\} \quad (1.26)$$

$$K = \{y(0) \in H_0^1(\Omega); \int_{\Omega} y^2(0) dx \leq N^2, \quad N \text{ is const.}\}. \quad (1.27)$$

Because of, it is possible to obtain the evident form of projective operator for K and U_{ad} given above, to get numerical solution of the problem (1.1)–(1.5) one can use the well-known projective gradient method.

2. Projective gradient method

Let us assume that

V is a closed, convex and bounded subset in a Hilbert space,

$J: V \rightarrow R^1$ is a functional belonging to $C^1(V)$.

We must find the point $v^0 \in V$, so that $J(v^0) = \inf_{v \in V} J(v)$.

For this purpose we construct the sequence $\{v_n\}$ ($n=0, 1, \dots$) according to the formula

$$v_{n+1} = P_V(v_n - \alpha_n J'(v_n)) \quad (2.1)$$

where P_V denotes the projective operator on the set V

$\alpha_n > 0$ can be chosen using one of the method given in [1].

The proof of the weak convergence of the projective gradient method is given in [1], for example. To obtain the strong convergence one can use the method of regularization [1].

3. Application of projective gradient method for solving a particular problem (1.1)–(1.5)

Let us consider problem (1.1)–(1.5) in which U_{ad} and K have the form (1.26) and (1.27), respectively.

Next let us denote

$$V = \{(y(0), u) \in H_0^1(\Omega) \times L^2(Q) : \int_Q u^2(x, t) dx dt \leq M^2, \int_{\Omega} y^2(0) dx \leq N^2\}. \quad (3.1)$$

It is known that the space $H_0^1(\Omega) \times L^2(Q)$ is a Hilbert one. Performing the same calculations as in (d) we can see that

$$J'(y(0), u) = (p(0), p + F_u). \quad (3.2)$$

Admitting that in the n -th iteration the control u is equal u_n , we get y_n as the solution of the following equation

$$\frac{\partial y_n}{\partial t} + A(t)y_n = u_n \quad x \in \Omega, \quad t \in (0, T) \quad (3.3)$$

$$y_n(x, t) = 0 \quad x \in \Gamma, \quad t \in (0, T) \quad (3.4)$$

with the initial condition $y_n(0)$.

The adjoint equation has the form

$$-\frac{\partial p_n}{\partial t} + A^*(t)p_n = F_y \quad x \in \Omega, \quad t \in (0, T) \quad (3.5)$$

$$p_n(x, t) = 0 \quad x \in \Gamma, \quad t \in (0, T) \quad (3.6)$$

$$p_n(x, T) = 0 \quad x \in \Omega. \quad (3.7)$$

Knowing the n -th approximation $v_n = (y_n(0), u_n)$ we can find $v_{n+1} = (y_{n+1}(0), u_{n+1})$ using the projective gradient method. Taking into account the form of projective operator P_V on the set V [2], we get

$$u_{n+1} = \begin{cases} u_n - \alpha_n(p_n + F_u) & \text{if } \int_Q [u_n - \alpha_n(p_n + F_u)]^2 dx dt \leq M^2 \\ \frac{M[u_n - \alpha_n(p_n + F_u)]}{\left[\int_Q [u_n - \alpha_n(p_n + F_u)]^2 dx dt \right]^{1/2}} & \text{otherwise,} \end{cases} \quad (3.8)$$

$$y_{n+1}(0) = \begin{cases} y_n(0) - \alpha_n p_n(0) & \text{if } \int_{\Omega} [y_n(0) - \alpha_n p_n(0)]^2 dx \leq N^2 \\ \frac{N[y_n(0) - \alpha_n p_n(0)]}{\left[\int_{\Omega} [y_n(0) - \alpha_n p_n(0)]^2 dx \right]^{1/2}} & \text{otherwise} \end{cases} \quad (3.9)$$

where $\alpha_n > 0$ can be chosen on the basis of [1], for instance

1) α_n may be calculated from the condition

$$f_n(\alpha_n) = \inf_{\alpha > 0} f_n(\alpha)$$

$$f_n(\alpha) = J[P_V((y_n(0), u_n) - \alpha J'(y_n(0), u_n))].$$

2) α_n can be arbitrarily given as the sequence such that

$$\sum_{n=0}^{\infty} \alpha_n = \infty, \quad \sum_{n=0}^{\infty} \alpha_n^2 < \infty.$$

To solve Eqs (3.3)–(3.7) one can use the convergent and stable difference method.

4. Conclusions

The derived conditions of optimality (Theorem 1) are original with point of view of applications Milutin–Dubovicki's Theorem in solving optimal control problems for distributed parameter system. Making use of Milutin–Dubovicki's method the similar conditions of the optimality may be derived for the boundary conditions of the second and the third types. In this case the theorem about the existence and uniqueness of the solution of parabolic equation (Theorem 1.2 [6]) is also valid (for the suitable choice of the spaces).

References

1. Vasiljev, F. P., Methods of Solving of Extremal Problems, "Nauka", Moscow 1981 (in Russian)
2. Vasiljev, F. P., On Gradient Methods for Solving Optimal Control Problems for Systems Described by Parabolic Equations, "Znanija", Moscow 1978, 118–142 (in Russian), Sbornik
3. Balakrishnan, A. V., Introduction to Optimization Theory in a Hilbert Space, Springer–Verlag, Berlin, New York 1971 (this monograph is in Russian, too)
4. Girsanov, I. V., Lectures on Mathematical Theory of Extremal Problems, Publication of University of Moscow, 1970 (in Russian)
5. Kowalewski, A., Kotarski, W., Application of Milutin–Dubovicki's Method to Solving an Optimal Control Problem for Hyperbolic Systems, Problems of Control and Information Theory, Vol. 9 (3) 1980, 183–193
6. Lions, J. L., Optimal Control of Systems Governed by Partial Differential Equations, Springer–Verlag, Berlin, New York, 1971 (this monograph is in French and in Russian, too)
7. Maslov V. P., Operators Methods, "Nauka", Moscow 1973 (in Russian)

Об одной задаче оптимального управления с неопределенным начальным состоянием

В. КОТАРСКИ, А. КОВАЛЕВСКИ
(Катовице) (Краков)

В работе рассматривается задача оптимального управления для объекта, описываемого дифференциальным уравнением в частных производных параболического типа с граничным условием Дирихле.

В работе накладываются некоторые ограничения на управление. Функция стоимости имеет интегральный вид. В нашей задаче время управления фиксировано. Начальное условие не определяется известной функцией, но является элементом определенного множества.

Проблема, сформулированная в работе, описывает процессы оптимального нагрева, в которых есть возможность выбора начальной температуры нагреваемого объекта.

Представлен также частный пример, в котором множество допустимых управлений, а также одно из начальных условий даны в виде ограничений на их норму.

Применение хорошо известного метода проекции градиента в гильбертовом пространстве делает возможным получение численного решения данной задачи оптимизации.

Проблема, рассматриваемая в работе, представляет собой обобщение задачи, которая была поставлена в докладе проф. Ж. Л. Лионса в международном Математическом Центре в Варшаве в 1980 году во время семестра по оптимальному управлению.

W. Kotarski
Institute of Mathematics, Silesian University,
ul. Bankowa 14, 40-007 Katowice, Poland

A. Kowalewski
Institute of Automatic Control, Systems Engineering and Telecommunication,
Stanisław Staszic University of Mining and Metallurgy,
Al. Mickiewicza 30, 30-059 Kraków, Poland

АСИМПТОТИЧЕСКИЕ ФОРМУЛЫ И ОЦЕНКИ ДЛЯ ВЕРОЯТНОСТНЫХ ХАРАКТЕРИСТИК ПОЛНОДОСТУПНОГО ПУЧКА С АБСОЛЮТНО НАСТОЙЧИВЫМ АБОНЕНТОМ

С. Н. СТЕПАНОВ

(Москва)

(Поступила в редакцию 7 июля 1982г.)

Для модели полнодоступного пучка, в которой абонент с вероятностью равной единице повторяет требование к соединению после получения отказа в обслуживании, получены оценки и асимптотические формулы для вероятностных характеристик при стремлении интенсивности поступления первичных вызовов к предельному значению, обусловленному максимальной пропускной способностью системы.

1. Введение

Для более адекватного описания реальных систем связи используются модели, в которых абонент после получения отказа в обслуживании повторяет попытку соединения через некоторое случайное время [1–4]. Интенсивность потока повторных вызовов существенно увеличивается в случае абсолютно настойчивого абонента, особенно когда интенсивность потока первичных вызовов стремится к предельному значению, обусловленному максимальной пропускной способностью системы. Изучение этой экстремальной ситуации и будет целью статьи. Отметим, что, несмотря на важность проблемы, решение этой задачи ранее в литературе по повторным вызовам не рассматривалось.

2. Схема функционирования модели

На полнодоступный пучок из v линий поступает пуассоновский поток первичных вызовов интенсивности λ . Поступивший вызов занимает произвольную свободную линию на случайное время обслуживания, распределенное экспоненциально с параметром, равным единице. Если свободных линий нет, абонент с вероятностью H_1 после первого отказа и вероятностью 1 после повторного отказа посылает следующий вызов через случайное время, имеющее экспоненциальное распределение с параметром, равным $\mu > 0$.

Обозначим через $P(j, i)$ вероятность того, что в стационарном состоянии имеется j повторяющихся абонентов и i линий занято в пучке. Система уравнений статистического равновесия выглядит следующим образом:

$$P(j, i)(\lambda + j\mu + i) = P(j, i-1)\lambda + P(j+i, i-1)(j+1)\mu + P(j, i+1)(i+1),$$

$$j=0, 1, \dots; i=0, 1, \dots, v-1. \quad (2.1)$$

$$P(j, v)(\lambda H_1 + v) = P(j, v-1)\lambda + P(j-1, v)\lambda H_1 + P(j+1, v-1)(j+1)\mu,$$

$$j=0, 1, \dots$$

$$\sum_{j=0}^{\infty} \sum_{i=0}^v P(j, i) = 1.$$

Введем обозначения:

$$P(i) = \sum_{j=0}^{\infty} P(j, i); \quad J(i) = \sum_{j=0}^{\infty} P(j, i)j.$$

Суммируя (2.1) по j , получаем соотношения:

$$P(i)\lambda + J(i)\mu = P(i+1)(i+1), \quad i=0, 1, \dots, v-1. \quad (2.2)$$

Основными вероятностными характеристиками модели являются:

— вероятность потерь первичного вызова $P(v)$,

— среднее число занятых линий $I = \sum_{i=1}^v P(i)i$,

— среднее число повторяющихся абонентов $J = \sum_{i=0}^v J(i)$.

Нетрудно показать [5], что значения $P(v)$, I , J , $J(v)$ связаны соотношениями:

$$\lambda + J\mu = P(v)\lambda + J(v)\mu + I, \quad (2.3)$$

$$J\mu = P(v)\lambda H_1 + J(v)\mu. \quad (2.4)$$

3. Оценки и асимптотические формулы

Вначале укажем ограничения на изменение интенсивности входного потока первичных вызовов. Из (2.3), (2.4) следует

$$\sum_{i=0}^{v-1} P(i)(v-i+\lambda(1-H_1)) = v - \lambda H_1. \quad (3.1)$$

Откуда получаем $\lambda H_1 \leq v$. Непосредственной подстановкой нетрудно убедиться в том, что выполнение равенства $\lambda H_1 = v$ противоречит неразложимости

марковского процесса, описывающего функционирование модели. Таким образом, окончательно получаем необходимое условие $\lambda H_1 < v$. Просуммируем систему уравнений (2.1) по i от 0 до v при фиксированном j . Получаем соотношение

$$(P(j+1, 0) + \dots + P(j+1, v-1))(j+1)\mu = P(j, v)\lambda H_1. \quad (3.2)$$

Подставим (3.2) в уравнение равновесия (2.1), содержащее $P(j, v)$ в левой части, и после умножения на j просуммируем по j от 0 до ∞ . Имеем

$$J(v)\mu(v - \lambda H_1) = P(v)\lambda H_1(\lambda + \mu) - \sum_{j=0}^{\infty} \sum_{i=0}^{v-2} \Theta(j, i), \quad (3.3)$$

где

$$\Theta(j, i) = P(j, i)j\mu(\lambda + (j-1)\mu).$$

Поскольку значения $\Theta(j, i) \geq 0$, то соотношение (3.3) можно применять для оценки характеристик $J\mu$ и $J(v)\mu$. Используя (2.3), (2.4), оценим $P(v)$, $P(v-1)$. Далее из (3.3), (2.4) получаем

$$J\mu \leq \frac{\lambda(\lambda H_1(\lambda(1-H_1) + v + \mu) - v(v-1))}{(v - \lambda H_1)(\lambda(1-H_1) + v)} + \quad (3.4)$$

$$+ \frac{v-1}{\lambda(1-H_1)+1} \left(\lambda(1-H_1) + v + \mu - 2 - \frac{\lambda H_1 \mu - v(v-1)}{\lambda(1-H_1) + v} \right),$$

при

$$\lambda_1 \leq \lambda < v/H_1,$$

где

$$\lambda_1 = -\frac{v+\mu}{2(1-H_1)} + \sqrt{\frac{(v+\mu)^2}{4(1-H_1)^2} + \frac{v(v-1)}{H_1(1-H_1)}}.$$

Аналогично находится верхняя граница для $J(v)\mu$.

Более точные результаты при приближенном расчете вероятностных характеристик модели в случае $\lambda H_1 \rightarrow v$ дает использование асимптотических формул. Найдем два первых члена разложения характеристик по степеням $v - \lambda H_1$ при $\lambda H_1 \rightarrow v$. Из (3.1) и $\sum_{i=0}^v P(i) = 1$ следует:

$$\lim_{\lambda H_1 \rightarrow v} P(v) = 1, \quad \lim_{\lambda H_1 \rightarrow v} I = v.$$

Воспользовавшись этими соотношениями, получаем из (2.1), (2.2), что при $\lambda H_1 \rightarrow v$

$$\sum_{i=0}^{v-2} J(i) = o(1), \quad \sum_{j=0}^{\infty} \sum_{i=0}^{v-3} P(j, i) j^2 = o(1). \quad (3.5)$$

Из (3.3), (3.5) находим первый член разложения характеристик $J\mu$ и $J(v)\mu$ по степеням $v - \lambda H_1$ при $\lambda H_1 \rightarrow v$:

$$J\mu = \frac{v(v(1-H_1) + H_1(1+\mu))}{(v - \lambda H_1)H_1} + o(1),$$

$$J(v)\mu = \frac{v(v(1-H_1) + H_1(1+\mu))}{(v - \lambda H_1)H_1} + o(1).$$

Для определения следующего члена разложения рассмотрим поведение вероятностей $P(j, v-1)$ при $\lambda H_1 \rightarrow v$. Подставляя (3.2) в уравнение системы (2.1), содержащее $P(j, v-1)$ в левой части и, проделав необходимые алгебраические преобразования, получаем

$$P(v-1) \geq P(0, v-1)A(\lambda),$$

где $\lim_{\lambda H_1 \rightarrow v} A(\lambda) = +\infty$. Отсюда и (2.1) следует, что при $\lambda H_1 \rightarrow v$

$$P(0) + \dots + P(v-2) = o(v - \lambda H_1). \quad (3.6)$$

Из (2.1)–(2.4), (3.3), (3.6) находим формулы, содержащие два члена разложения основных вероятностных характеристик модели по степеням $(v - \lambda H_1)$ при $\lambda H_1 \rightarrow v$

$$P(v) = 1 - \frac{(v - \lambda H_1)H_1}{v(1 - H_1) + H_1} + o(v - \lambda H_1),$$

$$I = v - \frac{(v - \lambda H_1)H_1}{v(1 - H_1) + H_1} + o(v - \lambda H_1),$$

$$J\mu = \frac{v(v(1 - H_1) + H_1(1 + \mu))}{(v - \lambda H_1)H_1} -$$

$$- (1 - v + \mu) \left(1 + \frac{H_1}{v(1 - H_1) + H_1} \right) - \frac{2v}{H_1} + o(1),$$

$$J(v)\mu = \frac{v(v(1 - H_1) + H_1(1 + \mu))}{(v - \lambda H_1)H_1} -$$

$$- (1 - v + \mu) \left(1 + \frac{H_1}{v(1 - H_1) + H_1} \right) - \frac{2v}{H_1} - v + o(1). \quad (3.7)$$

4. Применение асимптотических формул

Найденные асимптотические формулы и оценки можно использовать для приближенного расчета вероятностных характеристик модели при $\lambda H_1 \rightarrow v$. Для иллюстрации точности вычислений рассмотрим пример. В таблице 1 приведены результаты точного расчета характеристики J (столбец под номером 1), а также ее оценки, полученные по формулам (3.7) (столбец под номером 2) и (3.4) — (столбец под номером 3). Входные параметры принимают следующие значения: $\mu = 5$, $H_1 = 0,5$. Из анализа численных данных можно сделать вывод о том, что асимптотические формулы имеют приемлемую для практических нужд точность. Отметим, что найденные оценки могут достаточно эффективно использоваться не только в области больших потерь, когда λH_1 близко к v , но и при значениях λH_1 существенно меньших v (см. первые три строки таблицы 1).

Таблица 1

λ	$v=2$			λ	$v=10$		
	1	2	3		1	2	3
3	3,7991	3,7333	4,08	15	6,4476	5,6727	10,377
3,5	10,111	10,133	10,3	18	25,179	24,873	28,260
3,7	18,622	18,667	18,765	19	57,026	56,873	59,867
3,9	61,286	61,333	61,366	19,5	120,95	120,87	123,68
3,95	125,30	125,33	125,35	19,9	632,89	632,87	635,53
3,99	637,31	637,33	637,34	19,95	1272,88	1272,87	1275,5
3,995	1277,3	1277,3	1277,3	19,99	6392,87	6392,87	6395,5

Другой важной областью применения асимптотических формул является теоретическое исследование точности приближенных методов расчета характеристик систем с повторными вызовами. Наиболее естественным способом построения приближенного метода является упрощение исходной модели так, чтобы в дальнейшем использовать более простые системы, для которых вероятностные характеристики находятся либо в виде явных формул, либо имеют менее трудоемкий алгоритм вычисления. Рассмотрим два подхода: а) сравнение с однолинейными системами; б) изменение порядка обслуживания повторных вызовов. Исследуем каждый из этих случаев. Будем предполагать, что $H_1 = 1$.

а) Сравнение с однолинейными системами

Использование той или иной однолинейной системы зависит от того, какая характеристика оценивается. Для оценки $P(v)$ заменим исходную модель

на v однолинейных систем с интенсивностями входного потока $\frac{\lambda}{v}$ и абсолютно настойчивым абонентом. Каждая из моделей функционирует независимо. Значение $P(v)$ оценим вероятностью $P'(v)$ занятости всех однолинейных систем. Очевидно

$$P'(v) = \left(\frac{\lambda}{v}\right)^v. \quad (4.1)$$

Интуитивно понятно, что это будет оценка снизу для $P(v)$. Для оценки J возьмем соответствующую характеристику J' однолинейной системы с интенсивностью входного потока $\frac{\lambda}{v}$ и абсолютно настойчивым абонентом. Из (3.3), (2.4) имеем:

$$J' = \frac{\lambda^2(1+\mu)}{\mu v(v-\lambda)}. \quad (4.2)$$

Раскладывая в ряд по $(v-\lambda)$ при $\lambda \rightarrow v$ $P'(v)$ и J' , нетрудно убедиться в том, что два первых члена разложения совпадают с соответствующим разложением вероятностных характеристик $P(v)$ и J исходной модели (3.7), т.е. имеем при $\lambda \rightarrow v$

$$P(v) - P'(v) = o(v-\lambda), \quad J - J' = o(1). \quad (4.3)$$

Таким образом, доказано, что введенные оценки (4.1), (4.2) являются асимптотически точными при $\lambda \rightarrow v$.

Сравнение различных оценок вероятностных характеристик $P(v)$, J для пучка с абсолютно настойчивым абонентом ($H_1 = 1$) проведено в таблице 2 для модели с параметрами $v = 5$, $\mu = 5$. Столбец под номером 1 — точное значение характеристики; 2 — результаты расчетов по формулам (3.7); 3 — оценки сверху (3.4) для J и $\lambda/(\lambda(1-H_1)+v)$ для $P(v)$ (для $P(v)$ оценка следует из (2.3), (2.4)); 4 — использование приближенных формул (4.1) для $P(v)$ и (4.2) для J . Как и следовало ожидать из (4.3) при $\lambda \rightarrow v$ погрешность, которую дает использование (4.1), (4.2) и (3.7), примерно одинакова.

Таблица 2

λ	$P(v)$				J			
	1	2	3	4	1	2	3	4
4	0,50348	0	0,8	0,32768	2,8888	3,6	9,6	3,84
4,5	0,71482	0,5	0,9	0,59049	8,6381	9,6	15	9,72
4,9	0,93081	0,9	0,98	0,90392	56,639	57,6	62,52	57,624
4,95	0,96342	0,95	0,99	0,95099	116,72	117,6	122,46	117,61
4,99	0,99195	0,99	0,998	0,99004	596,88	597,6	602,41	597,60

б) Изменение порядка обслуживания повторных вызовов

Используя соотношения (2.2), нетрудно показать, что $P(v) \leq P_0(v)$, где $P_0(v)$ вероятность потерь вызова для пучка с ожиданием с той же интенсивностью входного потока. Раскладывая в ряд по степеням $v - \lambda$, вероятность $P_0(v)$ получаем из (3.7)

$$\lim_{\lambda \rightarrow v} \frac{P_0(v) - P(v)}{v - \lambda} > 0, \quad v > 1.$$

Таким образом, оценка $P(v)$ вероятностью потерь для пучка с ожиданием при $\lambda \rightarrow v$ менее точная, чем оценка (4.1). Построим модель с ожиданием, вероятность потерь которой также дает оценку сверху для $P(v)$, но является более точной при $\lambda \rightarrow v$.

Рассмотрим модель, функционирующую как исходный пучок с повторными вызовами, если число занятых линий v , $v - 1$ или в системе не имеется источников повторных вызовов. Если в результате освобождения линии процесс переходит в состояние с $(v - 2)$ -мя занятыми линиями, то один из имеющихся повторяющих абонентов занимает свободную линию. Оценим характеристики исходной модели $P(v)$, J соответствующими характеристиками $P''(v)$, J'' построенной модели. Значения $P''(v)$, J'' могут быть найдены из системы уравнений статистического равновесия. Она имеет вид:

$$\begin{aligned} P_{j,v}(\lambda + v) &= P_{j,v-1}\lambda + P_{j-1,v}\lambda + P_{j+1,v-1}(j+1)\mu, \\ P_{j,v-1}(\lambda + j\mu + v - 1) &= P_{j,v}v + P_{j+1,v-1}(v-1) + P_{0,v-2}\lambda\delta(j, 0), \\ P_{0,k}(\lambda + k) &= P_{0,k-1}\lambda + P_{0,k+1}(k+1), \quad k = v-2, v-3, \dots, 0. \end{aligned} \quad (4.4)$$

$$\delta(j, 0) = \begin{cases} 1, & j=0 \\ 0, & j>0, \end{cases}$$

где $P_{j,i}$ — стационарные вероятности модели: j — число источников повторных вызовов, i — число занятых линий в пучке. Характеристики $P''(v)$, J'' определяются как

$$P''(v) = \sum_{j=0}^{\infty} P_{j,v}; \quad J'' = \sum_{j=1}^{\infty} (P_{j,v} + P_{j,v-1})j. \quad (4.5)$$

Легко показать, что при $\lambda \rightarrow v$

$$P''(v) - P(v) = o(v - \lambda), \quad \lim_{\lambda \rightarrow v} (J'' - J) = -2(v - 1)/\mu.$$

Следовательно, показано, что оценки (4.5) дают асимптотически точные результаты, когда $\lambda \rightarrow v$ (правда точность J'' ниже, чем точность J').

Аналогичным образом, асимптотические формулы (3.7) можно использовать для теоретического изучения точности других приближенных методов при $\lambda H_1 \rightarrow v$. Отметим, что такое исследование полезно при построении эффективных приближенных алгоритмов расчета моделей с повторными вызовами. Если $H_2 < 1\lambda$ может стремиться к ∞ . В этом случае асимптотические формулы для более сложных моделей с повторными вызовами получены в [6].

Литература

1. Gosztony, G., Comparison of calculated and simulated results for trunk groups with repeated attempts. 8 ITC, Melbourne, 1976, paper 321, 1-11.
2. Le Gall, P., Sur l'influence des répétitions d'appels dans l'écoulement du trafic téléphonique. Annales télécomm., 1970, 25, 9-10, pp. 339-348.
3. Корпышев Ю. Н., Повторные вызовы при междугородной связи. Электросвязь, 1974, № 1, с. 35-41.
4. Ionin, G. L., Sedol, I. I., Full availability groups with repeated calls and time of advanced service. 7 ITC, Stockholm, 1973, paper 137, 1-4.
5. Шнепс-Шнеппе М. А., Степанов С. Н., Некоторые соотношения для систем с повторными попытками. В кн.: Информационные сети и их анализ. М.: Наука, 1978, с. 26-31.
6. Степанов С. Н., Численные методы расчета систем с повторными вызовами. М., Наука, 1983.

Asymptotic formulae and estimations for probabilistic characteristics of full-available group with absolutely persistent subscriber

S. N. STEPANOV

(Moscow)

For a model of full-available group with repeated calls and absolutely persistent subscriber estimations and asymptotic formulae are derived as intensity of primary calls tends to its limit value.

S. N. STEPANOV

Institute for Problem of Information Transmission

USSR, Moscow, GSP-4,

Ermolovoy 19

ASYMPTOTIC FORMULAE AND ESTIMATIONS FOR PROBABILISTIC CHARACTERISTICS OF FULL-AVAILABLE GROUP WITH ABSOLUTELY PERSISTENT SUBSCRIBER

S. N. STEPANOV

Estimations and asymptotic formulae of probabilistic characteristics with intensity of incoming primary calls tends to its limit value that is conditioned by the maximum capacity of the system are obtained for a model of full-available group in which a subscriber with a probability equal to 1 repeats the call after receiving a refusal.

1. Introduction

For a more adequate description of actual service systems the models in which a subscriber after receiving a refusal repeats the call in a random distributed time are used [1-4]. The intensity of the repeated call flow essentially decreases in a case of absolutely persistent subscriber particularly when the intensity of primary calls tends to its limit value stipulated by the maximum capacity of the system. The study of such an extreme situation is the subject of the paper. In spite of the urgency of the problem small attention was paid to it in literature devoted to the repeated calls.

2. Model functioning

Primary call intents arrive according to a Poisson process with intensity λ to a group of v lines. Having found an arbitrary free line the call seizes it for a random holding time exponentially distributed with a parameter equal to one. If the call arrives at the time when all the lines are busy it gets refusal. Then with a probability of H_1 (if it was the first refusal) and with a probability of 1 (if the refusal comes in a retry) the subscriber makes another attempt in an exponentially distributed time with a parameter equal to $\mu > 0$. The repeated call is served as a primary call.

Let $P(j, i)$ denote the probability of stationary model states, where j is the number of subscribers, repeating the call and i is the number of busy lines. The system of state equations can be written as

$$P(j, i) (\lambda + j\mu + i) = P(j, i-1)\lambda + P(j+1, i-1)(j+1)\mu + P(j, i+1) \times (i+1),$$

$$j=0,1,\dots; \quad i=0,1,\dots, v-1, \quad (2.1)$$

$$P(j, v)(\lambda H_1 + v) = P(j, v-1)\lambda + P(j-1, v)\lambda H_1 + P(j+1, v-1)(j+1)\mu$$

$$\sum_{j=0}^{\infty} \sum_{i=0}^v P(j, i) = 1. \quad j=0,1,\dots$$

Let us introduce the notation

$$P(i) = \sum_{j=0}^{\infty} P(j, i), \quad J(i) = \sum_{j=0}^{\infty} P(j, i)j.$$

Summing (2.1) with respect to j we obtain the relations

$$P(i)\lambda + J(i)\mu = P(i+1)(i+1), \quad i=0,1,\dots, v-1. \quad (2.2)$$

The basic probabilistic characteristics of the model are:

- the loss probability of the primary call - $P(v)$;
- the mean number of busy lines $I = \sum_{i=1}^v P(i)i$;
- the mean number of subscribers repeating their calls $J = \sum_{i=0}^v J(i)$.

It is easy to demonstrate [5] that the values of $P(v)$, I , J , $J(v)$ are connected by equations:

$$\lambda + J\mu = P(v)\lambda + J(v)\mu + I, \quad (2.3)$$

$$J\mu = P(v)\lambda H_1 + J(v)\mu. \quad (2.4)$$

3. Estimations and asymptotic formulae

At first we find the restriction on the value of the intensity of primary calls. From (2.3), (2.4) follows

$$\sum_{i=0}^{v-1} P(i)(v-i+\lambda(1-H_1)) = v - \lambda H_1, \quad (3.1)$$

from which we obtain $\lambda H_1 \leq v$. After direct substitution it is easy to make sure that the realization of the equality $\lambda H_1 = v$ contradicts to the irreducibility of the Markov process, describing the functioning of the model. Thus we have finally the necessary condition: $\lambda H_1 < v$. Summing (2.1) with respect to i from 0 to v at j fixed we obtain

$$(P(j+1,0) + \dots + P(j+1, v-1))(j+1)\mu = P(j, v)\lambda H_1. \quad (3.2)$$

We substitute (3.2) into (2.1) containing $P(j, v)$ in the left part and after multiplication by j and summing up with respect to j from 0 to ∞ we have

$$J(v)\mu(v - \lambda H_1) = P(v)\lambda H_1(\lambda + \mu) - \sum_{j=0}^{\infty} \sum_{i=0}^{v-2} \Theta(j, i), \quad (3.3)$$

where

$$\Theta(j, i) = P(j, i)j\mu(\lambda + (j-1)\mu).$$

It is clear that $\Theta(j, i) \geq 0$. We can therefore use (3.3) for estimation of the characteristics $J\mu$ and $J(v)\mu$. Using (2.3), (2.4) we evaluate $P(v)$, $P(v-1)$ and finally from (3.3), (2.4) we have

$$J\mu \leq \frac{\lambda(\lambda H_1(\lambda(1-H_1)+v+\mu)-v(v-1))}{(v-\lambda H_1)(\lambda(1-H_1)+v)} + \frac{v-1}{\lambda(1-H_1)+1} \times \\ \times \left(\lambda(1-H_1)+v+\mu-2 - \frac{\lambda H_1 \mu - v(v-1)}{\lambda(1-H_1)+v} \right), \quad (3.4)$$

where

$$\lambda_1 \leq \lambda < \frac{v}{H_1}$$

and

$$\lambda_1 = -\frac{v+\mu}{2(1-H_1)} + \sqrt{\frac{(v+\mu)^2}{4(1-H_1)^2} + \frac{v(v-1)}{H_1(1-H_1)}}.$$

In a similar way can find the upper bound for $J(v)\mu$.

More accurate results at approximate computation of probabilistic characteristics of the model in the case $\lambda H_1 \rightarrow v$ are obtained by using asymptotic formulae. Let us find the first two terms of expansion of the characteristics into series of powers $(v-\lambda H_1)$ as $\lambda H_1 \rightarrow v$. From (3.1) and $\sum_{i=1}^v (P(i)=1)$ we have

$$\lim_{\lambda H_1 \rightarrow v} P(v) = 1, \quad \lim_{\lambda H_1 \rightarrow v} I = V.$$

Using these relations from (2.1) and (2.2), we obtain

$$\sum_{i=1}^{v-2} J(i) = o(1), \quad \sum_{j=1}^{\infty} \sum_{i=0}^{v-3} P(j, i)j^2 = o(1), \quad (3.5)$$

as $\lambda H_1 \rightarrow v$.

From (3.3) and (3.5) we find the first member of the expansion of characteristics $J\mu$ and $J(v)\mu$ into series of powers $(v-\lambda H_1)$ as $\lambda H_1 \rightarrow v$

$$* J\mu = \frac{v(v(1-H_1)+H_1(1+\mu))}{(v-\lambda H_1)H_1} + o(1); \\ J(v)\mu = \frac{v(v(1-H_1)+H_1(1+\mu))}{(v-\lambda H_1)H_1} + o(1).$$

To find the next member of the expansion let us consider the behaviour of probability $P(j, v-1)$ as $\lambda H_1 \rightarrow v$. Substituting (3.2) into equation (2.1) containing $P(j, v-1)$ in the left part and making the necessary algebraic transforms we have

$$P(v-1) \geq P(0, v-1)A(\lambda),$$

where $\lim_{\lambda H_1 \rightarrow v} A(\lambda) = +\infty$. From this and (2.1) follows that as $\lambda H_1 \rightarrow v$

$$P(0) + \dots + P(v-2) = o(v - \lambda H_1). \quad (3.6)$$

From (2.1)–(2.4), (3.3) and (3.6) we find formulae containing two members of the decomposition of the main probabilistic characteristics of the model into series of powers $(v - \lambda H_1)$ as $\lambda H_1 \rightarrow v$

$$\begin{aligned} P(v) &= 1 - \frac{(v - \lambda H_1)H_1}{v(1 - H_1) + H_1} + o(v - \lambda H_1), \\ I &= v - \frac{(v - \lambda H_1)H_1}{v(1 - H_1) + H_1} + o(v - \lambda H_1), \\ J\mu &= \frac{v(v(1 - H_1) + H_1(1 + \mu))}{(v - \lambda H_1)H_1} - \\ &\quad - (1 - v + \mu) \left(1 + \frac{H_1}{v(1 - H_1) + H_1} \right) - \frac{2v}{H_1} + o(1), \\ J(v)\mu &= \frac{v(v(1 - H_1) + H_1(1 + \mu))}{(v - \lambda H_1)H_1} - \\ &\quad - (1 - v + \mu) \left(1 + \frac{H_1}{v(1 - H_1) + H_1} \right) - \frac{2v}{H_1} - v + o(1). \end{aligned} \quad (3.7)$$

4. Application of asymptotic formulae

The estimations and asymptotic formulae obtained in the previous chapter may be used for approximate calculation of the probabilistic characteristics in the case of λH_1 tends to v . To illustrate the accuracy of the calculations we consider an example. In Table 1 the results of accurate calculation of J are given (column 1) so as its estimations obtained from (3.7) (column 2) and from (3.4) (column 3). The input parameters take the following numerical values: $\mu = 5$, $H_1 = 0.5$. From the analysis of the numerical data one can conclude that the asymptotic formulae have the accuracy acceptable for practical necessities. Note that the found estimations one can use with sufficient efficiency not only in the area of large losses, when λH_1 is close to v , but for the values of λH_1 considerably smaller than v also (see the first three lines of Table 1).

Table 1

λ	$v=2$			λ	$v=10$		
	1	2	3		1	2	3
3	3.7991	3.7333	4.08	15	6.4476	5.6727	10.377
3.5	10.111	10.133	10.3	18	25.179	24.873	28.260
3.7	18.622	18.667	18.765	19	57.026	56.873	59.867
3.9	61.286	61.333	61.366	19.5	120.95	120.87	123.68
3.95	125.30	125.33	125.35	19.9	632.89	632.87	635.53
3.99	637.31	637.33	637.34	19.95	1272.88	1272.87	1275.5
3.995	1277.3	1277.3	1277.3	19.99	6392.87	6392.87	6395.5

The other important area for the application of asymptotic formulae is the theoretical study of the accuracy of approximate methods for calculation of the characteristics of systems with repeated calls. The most natural way for constructing the approximate method is the simplification of initial model so that later on it will be possible to use more simple systems for which the probabilistic characteristics are represented in explicit formulae or have less labour-consuming calculation algorithm. Let us consider two approaches: a) the comparison with one-line systems; b) changing of the repeated call service order. We consider both cases. Suppose $H_1 = 1$.

a) Comparison with one-line systems

The choice of the one-line system depends on estimated characteristic. For the estimation of $P(v)$ we replace the initial model by v one-line systems with intensity of the incoming flow λ/v and with absolutely persistent subscriber. Each of the models is independent.

Let $P'(v)$ denote the probability of all one line systems being busy. We shall use $P'(v)$ as an estimate for $P(V)$. Obviously

$$P'(v) = \left(\frac{\lambda}{v}\right)^v. \quad (4.1)$$

It is clear that it will be a lower bound for $P(v)$. To estimate J we take the corresponding characteristic J' of one line system with the incoming flow intensity λ/v and absolutely persistent subscriber. From (3.3), (2.4) we have

$$J' = \frac{\lambda^2(1+\mu)}{\mu v(v-\lambda)}. \quad (4.2)$$

Let us consider an expansion of $P'(v)$ and J' into the series of powers of $(v-\lambda)$ where $\lambda \rightarrow v$. It is clear that the first two members of the expansion coincide with the

corresponding expansion of the probabilistic characteristics $P(v)$ and J of the initial model (3.7). As λ tends to v we have

$$P(v) - P'(v) = o(v - \lambda), \quad J - J' = o(1). \quad (4.3)$$

Thus it is proved that estimations (4.1), (4.2) give asymptotically exact results as λ tends to v .

Some comparisons of different estimations of probabilistic characteristics $P(v)$, J for a group with absolutely persistent subscriber ($H_1 = 1$) are given in Table 2 for a model with parameters $v = 5$, $\mu = 5$. In column 1 the accurate value of characteristic is given; in column 2 the results of calculations of formulae (3.7) are given; in column 3 upper bounds (3.4) for J and $\lambda/(\lambda(1 - H_1) + v)$ for $P(v)$ are given (for $P(v)$ that comes from (2.3), (2.4); in column 4 the use of approximate formulae (4.1) for $P(v)$ and (4.2) for J is presented. Just as we expected from (4.3) when $\lambda \rightarrow v$ the accuracy of (4.1), (4.2) and (3.7) is approximately the same.

Table 2

λ	$P(v)$				J			
	1	2	3	4	1	2	3	4
4	0.50348	0	0.8	0.32768	2.8888	3.6	9.6	3.84
4.5	0.71482	0.5	0.9	0.59049	8.6381	9.6	15	9.72
4.9	0.93081	0.9	0.98	0.90392	56.639	57.6	62.52	57.624
4.95	0.96342	0.95	0.99	0.95099	116.72	117.6	122.46	117.61
4.99	0.99195	0.99	0.998	0.99004	596.88	597.6	602.41	597.60

b) Changing the order of service for repeated calls

From (2.2) it is easy to show that $P(v) \leq P_0(v)$, where $P_0(v)$ denotes the probability of all the lines in full available group $M/M/v$ being busy. Both models with repeated calls and delay have the same intensity of primary calls.

Let us consider an expansion of $P_0(v)$ into the series of powers of $(v - \lambda)$ as $\lambda \rightarrow v$. From this and (3.7) we have

$$\lim_{\lambda \rightarrow v} \frac{P_0(v) - P(v)}{v - \lambda} > 0, \quad v > 1.$$

Therefore the accuracy of the estimation $P_0(v)$ for $P(v)$ is less than the accuracy of (4.1). Now we introduce the model with delay which also gives the upper bound for $P(v)$ but this bound is more accurate as $\lambda \rightarrow v$.

Let us consider the model functioning as initial model with repeated calls if the number of busy lines is v or $v - 1$ for if the number of repeating subscribers is equal to 0.

If the process from the state with $(v-1)$ busy lines transforms to the state with $(v-2)$ busy lines (after completing the service of one call), then one of the repeating subscribers immediately seizes an arbitrary free line. We evaluate the characteristics of initial model $P(v), J$ by corresponding characteristics $P''(v), J''$ of constructed model. The value of $P''(v), J''$ may be found from the system of state equations. It can be written as

$$\begin{aligned} P_{j,v}(\lambda+v) &= P_{j,v-1}\lambda + P_{j-1,v}\lambda + P_{j+1,v-1}(j+1)\mu, \\ P_{j,v-1}(\lambda+j\mu+v-1) &= P_{j,v} + P_{j+1,v-1}(v-1) + P_{0,v-2}\lambda\delta(j,0), \\ P_{0,k}(\lambda+k) &= P_{0,k-1}\lambda + P_{0,k+1}(k+1); \quad k=v-2, v-3, \dots, 0; \\ \delta(j,0) &= \begin{cases} 1, j=0 \\ 0, j>0, \end{cases} \end{aligned} \quad (4.4)$$

where $P_{j,i}$ are the stationary probabilities, j is the number of sources of repeated calls, i is the number of busy lines in the group. Characteristics $P''(v), J''$ are defined as

$$P''(v) = \sum_{j=0}^{\infty} P_{j,v}; \quad J'' = \sum_{j=1}^{\infty} (P_{j,v} + P_{j,v-1})j. \quad (4.5)$$

It is easy to show that as $\lambda \rightarrow v$

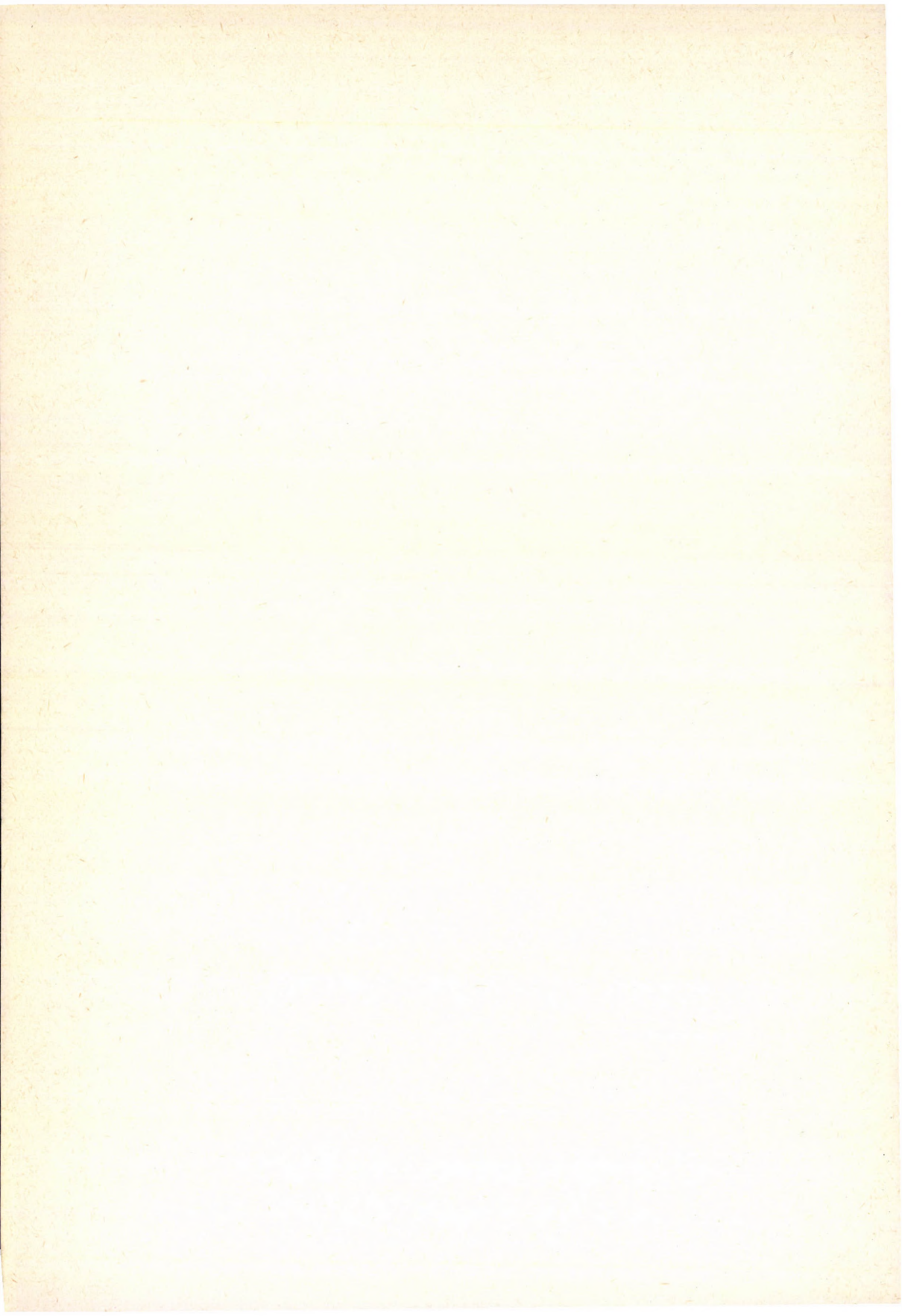
$$P''(v) - P(v) = o(v - \lambda), \quad \lim_{\lambda \rightarrow v} (J'' - J) = -2(v-1)/\mu.$$

Therefore we proved that estimations (4.5) give asymptotically exact results as λ tends to v (but the accuracy of J'' is less than the accuracy of J').

In a similar way we can use asymptotic formulae (3.7) for theoretical investigation of the accuracy of the other approximate methods as $\lambda H_1 \rightarrow v$. Note that such study is useful for constructing efficient approximate algorithms for models with repeated calls. If $H_2 < 1$ λ can tend to ∞ . In this case asymptotic formulae for more complicated model with repeated calls are obtained in [6].

References

1. Gosztony, G., Comparison of calculated and simulated results for trunk groups with repeated attempts. 8 ITC, Melbourne, 1976, paper 321, 1-11.
2. Le Gall, P., Sur l'influence des répétitions d'appels dans l'écoulement du trafic téléphonique. Annales télécomm., 1970, 25, 9-10, pp. 339-348.
3. Kornishev, J. N., Repeated attempts in toll traffic. Elektrosviaz, 1974, 1, pp. 35-41 (in Russian).
4. Ionin, G. L., Sedol, I. I., Full availability groups with repeated calls and time of advanced service. 7 ITC, Stockholm, 1973, paper 137, 1-4.
5. Shneps-Shneppe, M. A., Stepanov, S. N., Some relations for systems with repeated calls, in the book "Information networks and their analysis", M., Nauka, 1978, pp. 26-31 (in Russian).
6. Stepanov, S. N., Numerical methods for calculation of systems with repeated calls. M., Nauka, 1938 (in Russian).



NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Lehinsky Prospect 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H-1111 Budapest XI., Stoczek u. 2, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper with wide margins (4–5 cm), should carry the title of the contribution, the author(s) name, and the name of the country. At the end of the typescript the name and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary — possibly in Russian if the paper is in English and *vice-versa* — should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме — реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в трех экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициях. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Csiszár, I.</i> : An abstract source-channel transmission theorem (<i>Чисар И.</i> Об абстрактной теореме рассеяния в передающих устройствах)	303
<i>Šolc, F., Halabala, J.</i> : Adaptive transform picture coding for robot vision system (<i>Шольц Ф., Халабала И.</i> Адаптивная кодировка изображений с помощью двухмерного преобразования для визуальной системы робота)	309
<i>Vahrameev, S. A.</i> : On nonlinear differential games of pursuit (<i>Вахрамеев С. А.</i> О нелинейных дифференциальных играх преследования)	323
<i>Agrachev, A. A., Sarychev, A. V.</i> : The control of rotation for asymmetric rigid body (<i>АгрACHEV А. А., Сарычев А. В.</i> Управление вращением асимметричного твердого тела)	335
<i>Kotarski, W., Kowalewski, A.</i> : On optimal control problem with initial state not a priori given (<i>Котарски В. Ковалевски А.</i> Об одной задаче оптимального управления с неопределенным начальным состоянием)	349
<i>Степанов С. Н.</i> Асимптотические формулы и оценки для вероятностных характеристик полнодоступного пучка с абсолютно настойчивым абонентом (<i>Stepanov, S. N.</i> : Asymptotic formulae and estimations for probabilistic characteristics of full-available group with absolutely persistent subscriber)	361

316.920

VOL. 12 • NUMBER 6
TOM HOMEP

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1983

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H-1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 0BW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1983 Subscription Rate US Dollars 140.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

N. N. KRASOVSKII (USSR)
G. BOGNÁR (HUNGARY)

COORDINATING EDITORS

USSR

S. V. EMELYANOV
E. P. POPOV
V. S. PUGACHEV
V. I. SIFOROV
E. D. TERYAEV

HUNGARY

T. VÁMOS
L. VARGA
A. PRÉKOPA
S. CSIBI
I. CSISZÁR
L. KEVICZKY
J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ
V. STREJČ

РЕДАКТОРЫ ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)
Г. БОГНАР (ВНР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

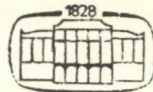
С. В. ЕМЕЛЬЯНОВ
Е. П. ПОПОВ
В. С. ПУГАЧЕВ
В. И. СИФОРОВ
Е. Д. ТЕРЯЕВ

ВНР

Т. ВАМОШ
Л. ВАРГА
А. ПРЕКОПА
Ш. ЧИБИ
И. ЧИСАР
Л. КЕВИЦКИ
Я. КОЧИШ

ЧССР

И. БЕНЕШ
В. СТРЕЙЦ



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

OPTIMAL STRATEGIES IN A HIERARCHICAL DIFFERENTIAL GAME

A. F. KLEIMENOV
(Sverdlovsk)

(Received February 6, 1983)

A definition of optimal strategies in a hierarchical nonantagonistic differential game is given on the basis of the formalization of an antagonistic differential game introduced in [1, 2]. For certain class of games the existence of optimal strategies is proved, their explicit form is given. The proof is essentially based on the fact of the existence of an universal saddle point in an antagonistic differential game [2]. An illustrating example is considered.

1. Introduction

Mathematical models of control systems with hierarchical decision making are applied to investigations of large control systems, in particular, of those with objects having economical or engineering nature. Such models are characterized by the presence of many decision makers (players) whose interests are not coinciding and not contradicting strictly. The study of such models is a subject of the theory of hierarchical games. In many works (see e.g. [3-9] and bibliography in these papers) hierarchical games are considered under the following assumption. The upper level player (ULP) (the center [4, 6], the leader [3, 5, 8, 9]) announces his strategy before the beginning of the play. Having received this information the lower level player (LLP) (the follower [3, 5, 8, 9]) chooses his strategy rationally, that is, he takes into account his own interests only. The purpose of the ULP is to choose the best strategy. A series of works, in particular [5, 7, 8, 9] is devoted to the investigation of hierarchical differential games under the assumption mentioned above. The idea of using penalty strategies for the construction of best strategies of the ULP was applied in [7]; this idea was suggested for static games in [4]. The method of finding the best strategies of the ULP based on different representations of solution of team problem was proposed in [9]. This method was applied to linear-quadratic games under some additional assumptions.

This paper studies a two-person hierarchical differential game with dynamics described by nonlinear equation in the sufficiently general form. The basic result presented in the paper is the definition of such class of games for which explicit expressions of optimal strategies of the ULP and the LLP are given.

2. Problem formulation

We shall consider a control system described by the differential equation

$$\dot{x} = f(t, x, u, v), \quad u \in P, \quad v \in Q \quad (1)$$

where x is an n -dimensional phase vector; u and v are vectors of control actions governed by players 1, 2 respectively; P and Q are compacts in the sets R^p and R^q . The function $f: G \times P \times Q \rightarrow R^n$ is continuous and satisfies the Lipschitz condition in x . Here G is a compact set in $R^1 \times R^n$ whose projection onto the time axis is equal to the given interval $[t_0, \vartheta]$. We assume that all trajectories of system (1) which began at any position $(t_*, x_*) \in G$ remain in the set G by all $t \in [t_*, \vartheta]$.

Player 1 chooses his control u to minimize the payoff $\sigma_1(x[\vartheta])$ and player 2 chooses his control v to minimize the payoff $\sigma_2(x[\vartheta])$. Here $\sigma_i: R^n \rightarrow R^1$ ($i=1, 2$) are given continuous functions.

We assume in general that the function f does not satisfy the condition of saddle point in the minor game [1, p. 56]. As it follows from the theory of antagonistic differential games [1] it is necessary to distinguish three cases of informational assumptions for the players. In all cases both players are informed of $x[t]$ at time t . In case 1° player 2 is informed of the control u for player 1. In case 2° both players are not informed of the control for the opponent. In case 3° player 1 is informed of the control v for player 2. Then the players' actions are formalized in the following classes: {pure strategies of player 1 — counterstrategies of player 2} in case 1°, {mixed strategies of player 1 — mixed strategies of player 2} in case 2° and {counterstrategies of player 1 — pure strategies of player 2} in case 3°.

In this paper players' actions in nonantagonistic differential game are assumed to be in the classes mentioned above depending on what case 1°, 2° or 3° is valid. For the sake of simplicity we shall assume further that the function f satisfies the condition of saddle point in the minor game. Therefore we suppose that both players have at their disposal pure strategies as their admissible actions. However the results obtained below are held in the general case also.

We shall identify a pure strategy of player 1 with a pair $U = \{u(t, x, \varepsilon), \beta_1(\varepsilon)\}$ where $u(t, x, \varepsilon)$, $(t, x) \in G$, $\varepsilon > 0$ is any function with values in P and $\beta_1(\cdot) \in A(0, \infty)$. By symbol $A(0, \infty)$ we shall denote the class of continuous monotonous functions $\beta: (0, \infty) \rightarrow (0, \infty)$ which satisfy a condition $\beta(\varepsilon) \rightarrow 0$ if $\varepsilon \rightarrow 0$. The definition of strategy as a function of the position (t, x) and of the precision parameter ε was proposed in [2]. Here the function $\beta_1(\cdot)$ is included in the expression of U for technical purposes, it shall be used below for definition of motions. Similarly, the pure strategy of player 2 is identified with pair $V = \{v(t, x, \varepsilon), \beta_2(\varepsilon)\}$ where function $v(t, x, \varepsilon)$, $(t, x) \in G$, $\varepsilon > 0$ has values in Q and $\beta_2(\cdot) \in A(0, \infty)$.

We shall assume further that an initial position (t_*, x_*) is fixed. Let be given: pure strategies U and V , ε_1 and ε_2 , which are values of the precision parameter ε chosen by players 1 and 2 respectively. Let $\Delta_1 = \{\tau_i^{(1)}\}$ and $\Delta_2 = \{\tau_j^{(2)}\}$ be subdivisions of the interval $[t_*, \vartheta]$ with families of non-overlapping semi-intervals $[\tau_i^{(1)}, \tau_{i+1}^{(1)})$ and $[\tau_j^{(2)}, \tau_{j+1}^{(2)})$ chosen by players 1 and 2 respectively provided $\delta(\Delta_1) \leq \beta_1(\varepsilon_1)$ and $\delta(\Delta_2) \leq \beta_2(\varepsilon_2)$ where the step of the subdivision Δ_1 is denoted by $\delta(\Delta_1) = \max_i (\tau_{i+1}^{(1)} - \tau_i^{(1)})$ and $\delta(\Delta_2)$ is the step of the subdivision Δ_2 .

An Euler spline generated by the pair of strategies (U, V) , by fixed $\varepsilon_1, \varepsilon_2$, by the subdivisions Δ_1 and Δ_2 from the initial position (t_*, x_*) is a continuous function $x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[t] = x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[t, t_*, x_*, U, V]$ satisfying the multistage differential equation

$$\begin{aligned} \dot{x}_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[t] &= f(t, x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[t], u(\tau_i^{(1)}, x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[\tau_i^{(1)}], \varepsilon_1), \\ &\quad v(\tau_{j-1}^{(2)}, x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[\tau_{j-1}^{(2)}], \varepsilon_2)), \\ t \in [\tau_i^{(1)}, \tau_j^{(2)}), \quad \text{where} \quad \tau_{j-1}^{(2)} &\leq \tau_i^{(1)} < \tau_j^{(2)} \leq \tau_{i+1}^{(1)}. \end{aligned}$$

A motion generated by the pair of strategies (U, V) from initial position (t_*, x_*) is a continuous function $x[t] = x[t, t_*, x_*, U, V]$ for which there exists a sequence of Euler splines $x_{\Delta_1^k, \Delta_2^k}^{\varepsilon_1^k, \varepsilon_2^k}[t, t^k, x^k, U, V]$ uniformly converging to $x[t]$ on $[t_*, \vartheta]$ if $k \rightarrow \infty, \varepsilon_1^k \rightarrow 0, \varepsilon_2^k \rightarrow 0, t^k \rightarrow t_*, x^k \rightarrow x_*, \delta(\Delta_1^k) \leq \beta_1(\varepsilon_1^k), \delta(\Delta_2^k) \leq \beta_2(\varepsilon_2^k)$. The pair (U, V) generates generally a set of motions $x[t, t_*, x_*, U, V]$. We shall denote this set compact in $C[t_*, \vartheta]$ by the symbol $X(t_*, x_*, U, V)$.

In the set $X(t_*, x_*, U, V)$ we single out such motions that are limits of sequences of Euler splines satisfying an additional condition $\varepsilon_2^k \leq \varepsilon_1^k$. The set of such motions will be denoted by the symbol $X_1(t_*, x_*, U, V)$. This set is compact in $C[t_*, \vartheta]$.

We assume the following informational hierarchy in the system. Player 1 has a complete information about the system, that is, he knows equation (1), the sets P and Q , the payoffs σ_1 and σ_2 . On the other hand, player 2 has the same information except the knowledge of the payoff σ_1 . Then we say that player 1 is the ULP and player 2 is the LLP.

Assumption A. Player 1 is the first to choose his strategy $U^* = \{u^*(t, x, \varepsilon), \beta_1^*(\varepsilon)\}$ and announces it to player 2 before the beginning of the play.

Assumption B. When the strategy U^* is known to player 2 his action is assumed to be rational. This means that player 2 chooses his strategy V^* satisfying the following condition

$$\begin{aligned} \rho^*(t_*, x_*, U^*) &= \max_{x[\cdot] \in X_1(t_*, x_*, U^*, V^*)} \sigma_2(x[\vartheta, t_*, x_*, U^*, V^*]) = \\ &= \min_V \max_{x[\cdot] \in X_1(t_*, x_*, U^*, V)} \sigma_2(x[\vartheta, t_*, x_*, U^*, V]). \end{aligned} \tag{2}$$

Here we shall not discuss the question about the existence of strategies V^* in the general case. We only note that rational strategy V^* exists in the class of games considered below. A set of rational strategies V^* of player 2 corresponding to announced strategy U^* of player 1 will be denoted by the symbol $K(t_*, x_*, U^*)$.

To formulate condition (2) in the approximate form we make one more assumption.

Assumption C. Player 1 chooses a value ε_1 of his precision parameter and informs this value to player 2 simultaneously with the beginning of the play.

Under Assumption C the condition (2) means that for any $\zeta > 0$ there exists $\varkappa(\zeta) > 0$ such that for any $\varepsilon_1 \leq \varkappa(\zeta)$, $\varepsilon_2 \leq \varepsilon_1$ and any subdivisions Δ_1, Δ_2 , $\delta(\Delta_1) \leq \beta_1^*(\varepsilon_1)$, $\delta(\Delta_2) \leq \beta_2^*(\varepsilon_2)$ the following inequality is true

$$\sigma_2(x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[\vartheta, t^*, x^*, U^*, V^*]) \leq \rho^*(t_*, x_*, U^*) + \zeta \quad (3)$$

where $(|t^* - t_*|^2 + \|x^* - x_*\|^2)^{1/2} \leq \min(\beta_1^*(\varepsilon_1), \beta_2^*(\varepsilon_2))$.

Assumption C makes the fulfilment of the inequality $\varepsilon_2 \leq \varepsilon_1$ for Euler splines feasible.

When player 1 announces the strategy U , his guaranteed result is

$$\rho(U) = \sup_{V \in K(t_*, x_*, U)} \max_{x[\cdot] \in X_1(t_*, x_*, U, V)} \sigma_1(x[\vartheta, t_*, x_*, U, V]).$$

Problem 1. One is to find a strategy of player 1 $U^0 = \{u^0(t, x, \varepsilon), \beta_1^0(\varepsilon)\}$ such that

$$\rho(U^0) = \min_U \rho(U). \quad (4)$$

The strategy U^0 will be called an optimal strategy of player 1 in a hierarchical differential game. Any strategy V^0 belonging to set $K(t_*, x_*, U^0)$ will be called an optimal strategy of player 2.

Some results of the theory of antagonistic differential games [1, 2] will be given now. Consider the following games Γ_1 and Γ_2 . The dynamics of both games are described by equation (1). In the game Γ_1 player 1 that governs the control u minimizes $\sigma_1(x[\vartheta])$, player 2 is opposed to him. In the game Γ_2 player 2 that governs the control v minimizes $\sigma_2(x[\vartheta])$, player 1 is opposed to him. As it was earlier assumed for the sake of simplicity, the function f satisfies the condition of saddle point in the minor game. Therefore, both players use pure strategies in the games Γ_1 and Γ_2 .

It is a known fact [2] that such differential games Γ_1 and Γ_2 have values given by continuous functions $\gamma_1(t, x)$ and $\gamma_2(t, x)$, $(t, x) \in G$ and universal saddle points

$$\begin{aligned} U_1^0 &= \{u^{(1)0}(t, x, \varepsilon), \beta_{11}^0(\varepsilon)\}, & V_1^0 &= \{v^{(1)0}(t, x, \varepsilon), \beta_{21}^0(\varepsilon)\}, \\ U_2^0 &= \{u^{(2)0}(t, x, \varepsilon), \beta_{12}^0(\varepsilon)\}, & V_2^0 &= \{v^{(2)0}(t, x, \varepsilon), \beta_{22}^0(\varepsilon)\}, \end{aligned} \quad (5)$$

respectively.

Now we formulate the following auxiliary problem.

Problem 2 (team problem for player 1). If the dynamics of the control system is described by equation (1), one is to find a pair of measurable functions $(u(t), v(t), t_* \leq t \leq \vartheta)$ minimizing the payoff $\sigma_1(x(\vartheta))$ where $x(t), t_* \leq t \leq \vartheta$ is a trajectory of system (1) generated by controls $u(\cdot)$ and $v(\cdot)$ from the initial position (t_*, x_*) .

It is a well-known fact that a solution of Problem 2 exists if the set $S(t, x) = \{s \in R^n: s = f(t, x, u, v), u \in P, v \in Q\}$ is convex.

3. Basic result

Let a pair $(u^*(t), v^*(t), t_* \leq t \leq \vartheta)$ be a solution of Problem 2 and $x^*(\cdot)$ a corresponding trajectory. It is natural to assume that the following inequality is true

$$\gamma_1(t_*, x_*) > \sigma_1(x^*(\vartheta)) \quad (6)$$

where $\gamma_1(t, x)$ is a value of the game Γ_1 . Inequality (6) means that both players when acting together obtain a result for player 1 which is better than the one in the antagonistic game Γ_1 . In other words, player 2 essentially affects the result of the game Γ_1 . Assume that the following condition is fulfilled

$$\gamma_2(t, x^*(t)) > \sigma_2(x^*(\vartheta)), \quad \forall t \in [t_*, \vartheta]. \quad (7)$$

It means that player 2 prefers to follow the trajectory $x^*(\cdot)$ right to the end of the play (as compared to the result of the antagonistic game Γ_2).

In general, Problem 2 may have more than one solution satisfying conditions (6), (7). In this case player 1 chooses such a solution for which the payoff $\sigma_2(x(\vartheta))$ has the least value, that is, player 1 shows goodwill to player 2.

So let the functions $u^*(\cdot), v^*(\cdot)$ be given. By using Lusin's theorem, one can find families of piecewise continuous functions $\{u^\varepsilon(\cdot)\}, \{v^\varepsilon(\cdot)\}$ such that $\|x^\varepsilon(t) - x^*(t)\| \leq \varepsilon$ is valid for all $t \in [t_*, \vartheta]$. Here $x^\varepsilon(\cdot)$ is a trajectory of system (1) generated by the controls $u^\varepsilon(\cdot)$ and $v^\varepsilon(\cdot)$.

Consider the following strategies $U^* = \{u^*(t, x, \varepsilon), \beta_1^*(\varepsilon)\}$ and $V^* = \{v^*(t, x, \varepsilon), \beta_1^*(\varepsilon)\}$ where

$$u^*(t, x, \varepsilon) = \begin{cases} u^\varepsilon(t), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^\varepsilon(t)\| \leq \varepsilon, \\ u^{(2)0}(t, x, \varepsilon), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^\varepsilon(t)\| > \varepsilon, \end{cases} \quad (8)$$

$$v^*(t, x, \varepsilon) = \begin{cases} v^\varepsilon(t), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^\varepsilon(t)\| \leq \varepsilon, \\ v^{(2)0}(t, x, \varepsilon), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^\varepsilon(t)\| > \varepsilon, \end{cases} \quad (9)$$

and the majorant function $\beta_1^*(\varepsilon)$, being the same for both strategies, is chosen so as to ensure the inequality

$$\|x_{\Delta_1, \Delta_2}^{\varepsilon_1, \varepsilon_2}[t, t_*, x_*, U^*, V^*] - x^*(t)\| < \varepsilon \quad (10)$$

$$\forall t \in (t_*, \vartheta]$$

for Euler splines provided $\varepsilon_2 \leq \varepsilon$, $\delta(\Delta_1) \leq \beta_1^*(\varepsilon)$, $\delta(\Delta_2) \leq \beta_1^*(\varepsilon_2)$. The functions $u^{(2)0}(t, x, \varepsilon)$, $v^{(2)0}(t, x, \varepsilon)$ were determined in (5).

Theorem. Let the dynamics of a control system be described by equation (1), an initial position (t_*, x_*) be fixed and Assumptions A, B be fulfilled. Assume that a pair of measurable functions $u^*(\cdot)$, $v^*(\cdot)$ is a solution of Problem 2 (team problem for player 1) and the corresponding trajectory $x^*(\cdot)$ satisfies conditions (6), (7). Then the strategies U^* , V^* (8)–(10) are optimal strategies in the considered hierarchical differential game.

Proof. Assume that player 1 chooses the strategy U^* (8), (10) and announced it to player 2. Then $V^* \in K(t_*, x_*, U^*)$. In fact, player 2 keeping the strategy V^* (9), (10) guarantees to himself the result $\sigma_2(x^*(\vartheta))$ on ideal limit motions which is not worse than the one which can be gained by using any other strategy. On the other hand, the strategy U^* is the best one for player 1 since it ensures him a minimal possible result $\sigma_1(x^*(\vartheta))$.

Remark 1. On constructing the strategy U^* (8) it is essential that a "penalty strategy" defined by the function $u^{(2)0}(t, x, \varepsilon)$ be universal, that is, suitable for all positions which may occur during the game.

Remark 2. Let Assumption C be fulfilled, that is, player 1 announced a chosen value ε_1^* of the precision parameter simultaneously with the beginning of the play. Then Euler Splines

$$x_{\Delta_1, \Delta_2}^{\varepsilon_1^*, \varepsilon_2}[t, t_*, x_*, U^*, V^*], \quad \varepsilon_2 \leq \varepsilon_1^*, \delta(\Delta_1) \leq \beta_1^*(\varepsilon_1^*), \delta(\Delta_2) \leq \beta_1^*(\varepsilon_2)$$

do not fall outside the limits of the " $2\varepsilon_1^*$ -tube" constructed along the trajectory $x^*(\cdot)$.

Remark 3. A value ε on the right-hand side of (10) may be replaced by a smaller value, e.g. $k\varepsilon$ where $0 < k < 1$. Then the layer of "thickness" $(1-k)\varepsilon_1^*$ may be used to mitigate after-effects caused by unpremeditated deviation of player 2 from the rational strategy.

Remark 4. The family of functions $\{v^*(\cdot)\}$ may be announced by player 1 before the beginning of the play.

Remark 5. If the functions $u^*(t)$ and $v^*(t)$ are piecewise continuous, then the strategies U^* , V^* may be defined more simply. Namely, instead of (8), (9) one can write

$$u^*(t, x, \varepsilon) = \begin{cases} u^*(t), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^*(t)\| \leq \varepsilon, \\ u^{(2)0}(t, x, \varepsilon), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^*(t)\| > \varepsilon, \end{cases} \quad (8')$$

$$v^*(t, x, \varepsilon) = \begin{cases} v^*(t), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^*(t)\| \leq \varepsilon, \\ v^{(2)0}(t, x, \varepsilon), & t_* \leq t \leq \vartheta, \quad \varepsilon > 0, \quad \|x - x^*(t)\| > \varepsilon. \end{cases} \quad (9')$$

Remark 6. As it was noted in Section 2, if the function f does not satisfy the condition of saddle point in the minor game then the players' actions are formalized by three classes of strategies depending on the informational assumptions 1°, 2° or 3°. It is natural to assume that the ULP may choose these informational assumptions so as to obtain the fulfilment of condition (7). E.g., if inequality (7) is fulfilled only in case 1° then it will be preferable for player 1 to give information about the control u to player 2.

4. An example

The following equations

$$\begin{aligned} \dot{\xi}_1 &= u_1 + v_1, & \dot{\xi}_2 &= u_2 + v_2, \\ \|u\| &= (u_1^2 + u_2^2)^{1/2} \leq 0.5, & \|v\| &= (v_1^2 + v_2^2)^{1/2} \leq 1.0 \end{aligned} \quad (11)$$

describe a motion of a material point of unit mass on the plane (ξ_1, ξ_2) under an action of a force $F = u + v$. Player 1 (player 2) that governs the control $u(v)$ minimizes $\sigma_1(\xi[\vartheta])$, $(\sigma_2(\xi[\vartheta]))$ where

$$\begin{aligned} \sigma_i(\xi[\vartheta]) &= \|\xi[\vartheta] - a^{(i)}\|, \\ \xi &= (\xi_1, \xi_2), & a^{(i)} &= (a_1^{(i)}, a_2^{(i)}), & i &= 1, 2. \end{aligned} \quad (12)$$

Here $a^{(i)}$ ($i = 1, 2$) are given points on the plane (ξ_1, ξ_2) , ϑ is final time.

The right-hand side of (11) is such that the condition of saddle point in the minor game is fulfilled. Therefore the actions of both players are formalized in a class of pure strategies. Let Assumptions A, B be fulfilled. It is required to find optimal strategies in this hierarchical game.

By setting $y_1 = \xi_1$, $y_2 = \xi_2$, $y_3 = \dot{\xi}_1$, $y_4 = \dot{\xi}_2$ and making the following change of variables $x_1 = y_1 + (\vartheta - t)y_3$, $x_2 = y_2 + (\vartheta - t)y_4$, $x_3 = y_3$, $x_4 = y_4$ we shall get a system whose first and second equations are

$$\begin{aligned} \dot{x}_1 &= (\vartheta - t)(u_1 + v_1), \\ \dot{x}_2 &= (\vartheta - t)(u_2 + v_2). \end{aligned} \quad (13)$$

Further, (12) may be written as

$$\sigma_i(x[\vartheta]) = \|x[\vartheta] - a^{(i)}\|, \quad x = (x_1, x_2), \quad i = 1, 2. \quad (14)$$

We consider now the shortened system (13) with the payoffs (14). It can easily be shown that the values in antagonistic differential games Γ_1 and Γ_2 are given by the formulae

$$\begin{aligned} \gamma_1(t, x) &= \|x - a^{(1)}\| + 0.25(\vartheta - t)^2, \\ \gamma_2(t, x) &= \|x - a^{(2)}\| - 0.25(\vartheta - t)^2 \end{aligned}$$

and the functions $u^{(2)0}(t, x, \varepsilon)$ and $v^{(2)0}(t, x, \varepsilon)$ in (5) are given by

$$u^{(2)0}(t, x, \varepsilon) = 0.5 \frac{x - a^{(2)}}{\|x - a^{(2)}\|}, \quad v^{(2)0}(t, x, \varepsilon) = - \frac{x - a^{(2)}}{\|x - a^{(2)}\|}. \quad (15)$$

Let the following initial conditions be given: $t_0 = 0$, $\xi_1(0) = 2.1$, $\xi_1(0) = -1.5$, $\xi_2(0) = -0.1$, $\xi_2(0) = 1.5$. The following numerical values of other parameters were chosen: $\vartheta = 1.4$, $a_1^{(1)} = 0$, $a_2^{(1)} = -1.0$, $a_1^{(2)} = 1.0$, $a_2^{(2)} = 0$. We have then $x_1(0) = 0$, $x_2(0) = 2.0$.

It can be easily shown that a solution of the corresponding team problem for player 1 is given by

$$\begin{aligned} u_1^*(t) = 0, \quad u_2^*(t) = -0.5, \quad v_1^*(t) = 0, \quad v_2^*(t) = -1.0 \\ 0 \leq t \leq 1.4. \end{aligned} \quad (16)$$

The corresponding trajectory $x^*(\cdot)$ is given by

$$x_1^*(t) = 0, \quad x_2^*(t) = 0.75t^2 - 2.1t + 2.0, \quad 0 \leq t \leq 1.4.$$

The function $\gamma_2(t, x)$ calculated along the trajectory $x^*(\cdot)$ is given by

$$\gamma_2^*(t) = \gamma_2(t, x^*(t)) = [1 + (0.75t^2 - 2.1t + 2.0)^2]^{1/2} - 0.25(1.4 - t)^2.$$

We are convinced immediately that $\gamma_2^*(t)$ decreases strongly when t changes from 0 to 1.4. Hence condition (7) is fulfilled. According to the theorem of Section 3, optimal strategies are given by formulae (8), (9) with the substitution of functions (15) and (16).

A projection of a phase trajectory corresponding to an ideal limit motion generated by a pair of the optimal strategies (U^*, V^*) is represented in Fig. 1.

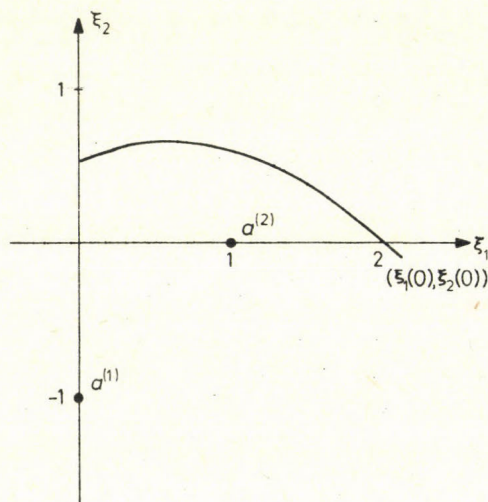


Fig. 1

References

1. Krasovskii, N. N., Subbotin, A. I., Positional differential games. Moscow, "Nauka", 1974 (Russian).
2. Krasovskii, N. N., Differential games. Approximation and formal models. *Mathem. sbornik*, **107**, 4 (1978).
3. Von Stackelberg, H., The theory of the market economy. Oxford, Oxford Univ. Press, 1952.
4. Germeier, Iu. B., On two-person games with fixed sequence of moves, *Dokl. Acad. Nauk SSSR*, **198**, 5 (1971).
5. Chen, C. I., Cruz, J. B., Jr., Stackelberg solution for two-person games with biased information patterns. *IEEE Trans. Automat. Contr.* **AC-17**, 6 (1972).
6. Germeier, Iu. B., Nonantagonistic games. Moscow, "Nauka", 1976 (Russian).
7. Kononenko, A. F., On multi-step conflict with information exchange. *Ž. Vyčisl. Mat. i Mat. Fiz.*, **17**, 4 (1977).
8. Cruz, J. B., Jr., Leader-follower strategies for multilevel systems. *IEEE Trans. Automat. Contr.*, **AC-23**, 2 (1978).
9. Bařar, T., Selbuz, H., Closed-loop Stackelberg strategies with applications in the optimal control of multilevel systems. *IEEE Trans. Automat. Contr.*, **AC-24**, 2 (1979).

Оптимальные стратегии в одной иерархической дифференциальной игре

А. Ф. КЛЕЙМЕНОВ

(Свердловск)

Рассматривается неантагонистическая иерархическая дифференциальная игра двух лиц при предположении, что первый игрок (игрок верхнего уровня) объявляет свое действие до начала игры, а второй игрок (игрок нижнего уровня), получив эту информацию, действует рационально, т. е. исходя только из собственных интересов. Динамика игры описывается нелинейным дифференциальным уравнением достаточно общего вида. Действия игроков формализуются в тех же классах позиционных стратегий, что и в теории антагонистических дифференциальных игр [1, 2] в зависимости от предположений, которые принимаются относительно информированности игроков о реализующихся управлениях партнера.

В работе вводятся понятия оптимальных стратегий игроков верхнего и нижнего уровней. Выделен класс игр, для которого оптимальные стратегии существуют и допускают явное представление. Результат существенно опирается на факт существования универсальной седловой точки в антагонистической дифференциальной игре [2]. Рассматривается иллюстрирующий пример.

А. Ф. Клейменов

Институт математики и механики

Уральского научного центра АН СССР

620219 Свердловск, ГСП-384,

ул. С. Ковалевской, 16

BLOCK DECOUPLING BY DYNAMIC COMPENSATION WITH INTERNAL PROPERNESS AND STABILITY

V. KUČERA

(Prague)

(Received October 12, 1982)

The problem of decoupling a linear system by dynamic compensation into multi-input multi-output subsystems is solved by transfer matrix methods. A simple necessary and sufficient condition is given for the existence of an admissible decoupling compensator. A design procedure is then proposed which guarantees the internal properness and stability of the decoupled system. These two properties are effectively studied by means of stable proper factorizations of transfer matrices.

1. Introduction

Decoupling is a way of decomposing a complex system into non-interacting subsystems. In fact, certain practical applications necessitate to control independently different parts of the system. Even if this is not required, the absence of interaction can significantly simplify the synthesis of the desired control laws.

Despite several decades of research efforts, the decoupling problem has not yet been completely solved. In its classical setup (one input affects one and only one output) the problem was studied by means of transfer function matrices. Among others, see Voznesenkij [26], Kavanagh [15], Strejc [24], Mejerov [19], Wolovich [27] and Kučera [17]. The problem of internal structure and stability of the decoupled system, however, received just little attention.

A deeper insight was provided by the state space approach. The pioneering work is due to Morgan [20]. Falb and Wolovich [7] were the first to establish a solvability condition for Morgan's problem. Gilbert [10] related this criterion to state feedback invariants of the system.

The block decoupling problem, i.e. that of producing multi-input multi-output non-interacting subsystems, was introduced by Wonham and Morse [29] and Basile and Marro [1]. Using a geometric approach, they determined the solvability of the problem by static state feedback in several special cases. Another approach, based on the structure algorithm, was presented by Silverman and Payne [23]. The decoupling by dynamic state feedback and precompensation was studied by Morse and Wonham [21], who finally obtained a deep insight into the internal structure of the decoupled

system. The reader is referred to Wonham [28] and Morse and Wonham [22] for details. Recently we have witnessed a comeback of the transfer matrix methods, see e.g. Koussiouris [16], Hautus and Heymann [11], Dion [6] and Kučera [18].

Some recent extensions concern the decoupling by static output feedback, either with or without stability. For reference, see e.g. Howze and Pearson [14], Howze [13], Denham [2], Hazlerigg and Sinha [12], Filev [8], [9], Descusse and Malabre [3] and Descusse, Lafay and Kučera [5]. Finally let us mention the works of Vardulakis [25] and Descusse and Dion [4] on the implications of the system structure at infinity for the decoupling.

This paper is inspired by the work of Hautus and Heymann [11] and it aims to solve, in a systematic way, the problem of decoupling a linear system into multi-input multi-output blocks. Such a general formulation suits best the practical needs. The decoupling is to be achieved by a dynamic compensator driven by measured outputs, which need not coincide with the outputs to be controlled.

The transfer matrix approach is adopted to solve the problem. A simple condition for decoupling is established in terms of ranks of certain rational matrices. Similar results were obtained by Hautus and Heymann [11] in the special case of proper systems with measurable internal state. The attention is then focused on the design of a decoupling compensator which ensures the internal properness and stability of the decoupled system. These properties, indispensable in practice, are achieved here through a systematic use of stable proper factorizations of the system transfer matrix.

2. Problem formulation

Let F be the field of reals and denote by $F^{p \times q}(s)$ and $F^{p \times q}\{s\}$ respectively the set of $p \times q$ real *rational* and *stable proper rational* matrices in the Laplace variable s . Recall that a rational matrix is stable if it has no pole in the region $0 \leq \operatorname{Re} s < \infty$ and is proper if it has no pole at $s = \infty$.

The set $F^{n \times n}(s)$ is a ring and its units (invertible elements) are non-singular matrices. The set $F^{n \times n}\{s\}$ is also a ring; here the units are the matrices having neither pole nor zero in $0 \leq \operatorname{Re} s \leq \infty$. When $n = 1$ we write $F(s)$ and $F\{s\}$ for brevity.

Two stable proper rational matrices A and B are said to be relatively left (or right) prime if there exist stable proper rational matrices X_1 and Y_1 (or X_2 and Y_2) such that $AX_1 + BY_1 = I$ (or $X_2A + Y_2B = I$), where I stands for the identity matrix.

Consider a linear system governed by the input-output equation

$$y = Tu \tag{1}$$

where u is the input q -vector, y is the output p -vector and $T \in F^{p \times q}(s)$ is the transfer matrix of the system. Note that T is just rational, not necessarily proper or stable. Let

p_1, \dots, p_k be a given set of positive integers satisfying

$$\sum_{i=1}^k p_i = p$$

and denote

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \quad (2)$$

where y_i is a p_i -dimensional subvector.

System (1) is said to be *decoupled*, or more specifically (p_1, \dots, p_k) -decoupled, if there exist positive integers q_1, \dots, q_k satisfying

$$\sum_{i=1}^k q_i = q$$

such that T has the block diagonal form

$$T = \begin{bmatrix} T_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & T_{kk} \end{bmatrix}$$

where T_{ii} is $p_i \times q_i$.

This is not a generic property of the system, but it can be achieved by suitable compensation. To this effect let w denote the m -vector output of the system which is available for measurement. Now, w may coincide with the entire state of the system or, in particular, with the output y to be controlled. In general, it is related with the input by the equation

$$w = T_w u \quad (3)$$

where $T_w \in F^{m \times q}(s)$ is another transfer matrix, again not necessarily proper or stable.

The most general linear dynamic *compensator* can then be described by the equation

$$u = Kv + K_w w \quad (4)$$

where v is an external input of suitable dimension, say r . As it is seen in Fig. 1, the matrices $K \in F^{q \times r}(s)$ and $K_w \in F^{q \times m}(s)$ represent the feedforward and the feedback parts of the compensator, respectively.

The *decoupling problem* is then to find matrices K and K_w , if they exist, such that the transfer matrix

$$\bar{T} = T(I - K_w T_w)^{-1} K \quad (5)$$

from v to y be suitably block diagonal.

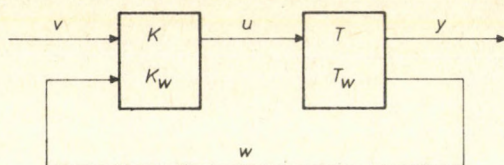


Fig. 1. System and compensator

Obviously, unless additional provisions are made, the decoupling problem is trivial since it could be solved by setting $K=0$. Thus it is necessary to impose certain *admissibility conditions* on the decoupling compensators to make the problem meaningful, for example

$$\text{rank } \bar{T} = \text{rank } T. \quad (6)$$

This condition is equivalent to the preservation of the class of controlled output trajectories. We thus require that no essential loss of control occurs through the decoupling process.

Another requirement, frequently imposed on the decoupled system in practice, is that of internal properness and stability. A linear system is said to be *internally proper* if a step disturbance of its internal state does not excite Dirac pulses anywhere in the system, and it is said to be *internally stable* if the transient excited by such a step disturbance asymptotically dies out anywhere in the system. These properties are to be distinguished from the external properness and stability, which are fully characterized by the properness and stability of the system transfer matrix.

3. Preliminaries

In order to study the internal properties of the decoupled system in a unified way, it is convenient to express the transfer matrices of the system and those defining the compensator in the following factorized form

$$\begin{bmatrix} T_w \\ T \end{bmatrix} = \begin{bmatrix} B \\ C \end{bmatrix} A^{-1}$$

$$[-K_w \quad K] = P^{-1} [Q \quad R]$$

where the matrices

$$A \in F^{q \times q}\{s\}, \quad \begin{bmatrix} B \\ C \end{bmatrix} \in F^{(m+p) \times q}\{s\}$$

are relatively right prime and the matrices

$$P \in F^{q \times q}\{s\}, \quad [Q \ R] \in F^{q \times (m+r)}\{s\}$$

are relatively left prime.

These *stable proper factorizations* exist and are unique up to units of $F^{q \times q}\{s\}$. They make it possible to write system equations (1) and (3) as

$$\begin{aligned} u &= Ax \\ w &= Bx \\ y &= Cx \end{aligned} \quad (7)$$

for some x , and compensator equation (4) as

$$Pu = -Qw + Rv. \quad (8)$$

The overall system then can be given the form shown in Fig. 2 and its transfer matrix (5) reads

$$\bar{T} = C(PA + QB)^{-1}R. \quad (9)$$

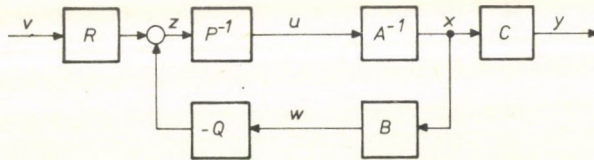


Fig. 2. Overall system in factorized form

The *fundamental assumption* we make here is that the internal states of the given system which are not reachable from its input u or which are not observable at its outputs w, y form an internally proper and stable subsystem. Similarly it is assumed that the compensator is realized in such a way that the internal states which are not reachable from its inputs w, v or which are not observable at its output u form an internally proper and stable subsystem.

The issue of internal properness and stability of the overall system is then solved by the following

Lemma. The overall system (7)–(8) is internally proper and stable if and only if the matrix $PA + QB$ is a unit of $F^{q \times q}\{s\}$.

Proof. Let the compensated system be internally proper and stable. Then, e.g. the transfer matrix $(PA + QB)^{-1}$ from the signal z to x in Fig. 2 is proper and stable. Since $PA + QB$ itself is proper and stable by definition, it is a unit of $F^{q \times q}\{s\}$.

Conversely let $PA + QB$ be a unit of $F^{q \times q}\{s\}$. Then the transfer matrices relating any two signals within the overall system (7)–(8) are all proper and stable since $PA + QB$ is their common denominator. Thus the internal improperness or instability of this system could originate only in the unreachable and/or unobservable parts of system (7) and compensator (8). However, this is excluded by our fundamental assumption.

Q.E.D.

4. Problem solvability

We are going to establish a simple necessary and sufficient condition for the system to be decoupled as well as internally proper and stable.

In view of (2) write

$$C = \begin{bmatrix} C_1 \\ \vdots \\ C_k \end{bmatrix} \quad (10)$$

where C_i is a $p_i \times q$ submatrix. Then we have the following

Theorem. Given system (7) and partition (10), there exists an admissible compensator (8) such that the overall system is

(i) internally proper and stable if and only if

$$A \text{ and } B \text{ are relatively right prime} \quad (11)$$

(ii) block decoupled if and only if

$$\sum_{i=1}^k \text{rank } C_i = \text{rank } C. \quad (12)$$

Proof. (i) Let the overall system be internally proper and stable. By Lemma, the matrix $PA + QB$ is a unit of $F^{q \times q}\{s\}$ whence A and B must be relatively right prime.

Conversely, let the matrices A and B of system (7) be relatively right prime. Then there exist matrices $P \in F^{q \times q}\{s\}$ and $Q \in F^{q \times m}\{s\}$ such that

$$PA + QB = I. \quad (13)$$

Moreover, P can be chosen non-singular. Indeed, let $\bar{P} \in F^{q \times q}\{s\}$ and $\bar{Q} \in F^{q \times m}\{s\}$ satisfy (13) with \bar{P} singular. Then the matrices

$$P = \bar{P} + W\bar{B}$$

$$Q = \bar{Q} - W\bar{A}$$

satisfy (13) as well provided $\bar{A} \in F^{q \times q}\{s\}$ and $\bar{B} \in F^{m \times q}\{s\}$ are relatively left prime matrices such that $\bar{A}^{-1}\bar{B} = B\bar{A}^{-1}$ and $W \in F^{q \times m}\{s\}$ (see Kučera [17]). Choose any point s_0 at which $A(s_0)$ —and hence also $\bar{A}(s_0)$ —is non-singular. Then any W satisfying $W(s_0) = \bar{Q}(s_0)\bar{A}^{-1}(s_0)$ yields $Q(s_0) = 0$. Substituting into (13) the corresponding $P(s_0)$ is seen to be non-singular, which means that P itself is non-singular.

Then compensator (8) defined by the matrices P, Q from (13) and by an arbitrary proper stable rational matrix R satisfying $\text{rank } CR = \text{rank } C$ is admissible since, by (9),

$$\text{rank } \bar{T} = \text{rank } CR = \text{rank } C = \text{rank } T.$$

The resulting system (7)–(8) is internally proper and stable in view of Lemma and identity (13).

(ii) Let (8) be an admissible decoupling compensator for system (7). Denote

$$\bar{K} = (PA + QB)^{-1}R.$$

The block diagonality of the transfer matrix \bar{T} then implies

$$\text{rank } C\bar{K} = \sum_{i=1}^k \text{rank } C_i\bar{K}$$

and the admissibility of the compensator gives

$$\text{rank } C_i\bar{K} = \text{rank } C_i, \quad i = 1, \dots, k.$$

Therefore (12) holds.

The sufficiency will again be proved by construction. Denote

$$r_i = \text{rank } C_i, \quad i = 1, \dots, k.$$

Then there exists a non-singular matrix $U_i \in F^{p_i \times p_i}\{s\}$ such that

$$C_i = U_i \begin{bmatrix} C'_i \\ 0 \end{bmatrix}$$

where the rows of C'_i are linearly independent over $F(s)$ and where the zero matrix has $p_i - r_i$ rows and may be empty. If (12) holds, then

$$C' = \begin{bmatrix} C'_1 \\ \vdots \\ C'_k \end{bmatrix}$$

has linearly independent rows over $F(s)$ and hence has a right inverse G .

Define a compensator (8) by the matrices P and Q from (13) and by the matrix $R = Gt$, where the scalar function $t \in F\{s\}$ is chosen so as to make R a stable proper rational matrix. This compensator is a special case of that constructed earlier in (i);

hence it is admissible. The transfer matrix (9)

$$\bar{T} = CGt = \begin{bmatrix} U_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & U_k \end{bmatrix} \begin{bmatrix} I_1 & & & \\ 0 & \ddots & & \\ & & \ddots & \\ & & & I_k \\ & & & & 0 \end{bmatrix}^t$$

is block diagonal, where I_j stands for the $r_j \times r_j$ identity matrix. The resulting system (7)–(8) is therefore block decoupled. Q.E.D.

It is worthwhile to note that the problem of decoupling and that of internal properness and stability are two *separate* and independent issues. However, this is no longer true for less general compensators, e.g. for the error-actuated compensator described by

$$Pu = Q(v - w).$$

See Kučera [17] for details.

As to the interpretation of the solvability conditions, condition (11) corresponds to the internal properness and stability of the subsystem defined by the internal states of system (7) which are not observable at the measured output w . Condition (12) calls for the linear independence of any two outputs of (7) belonging to different blocks. The solvability of the decoupling problem is thus strongly dependent on the integers p_1, \dots, p_k , that is to say, upon the allocation of the outputs into the blocks.

5. Compensator construction

Let us summarize, in a concise and comprehensive way, the construction of a desired compensator as it follows from the proof of the Theorem.

We are given a system by its transfer matrices T and T_w along with the integers p_1, \dots, p_k . The transfer matrices K and K_w defining an admissible decoupling compensator which ensures the internal properness and stability of the resulting system can be found, if they exist, in the following steps.

- 1) Determine a relatively left prime, stable proper factorization

$$\begin{bmatrix} T_w \\ T \end{bmatrix} = \begin{bmatrix} B \\ C \end{bmatrix} A^{-1}$$

and the partition C_1, \dots, C_k of C .

PROBLEMS OF CONTROL AND INFORMATION THEORY VOL. 12 (1983)

SUBJECT INDEX

- Agrachev, A. A., Sarychev, A. V.*: The control of rotation for asymmetric rigid body. Vol. 12, No. 5, pp. 335-347
- Azimov, A. Ya.*: Duality in nonconvex problems of vector optimization. Vol. 12, No. 3, pp. 209-219
- Bocharov, P. P., Albores, F. J.*: On two-node exponential queueing network with internal losses of blocking. Vol. 12, No. 4, pp. 243-252
- Chaudhuri, A. K., Mukherjee, R. N.*: On minimum time control. Vol. 12, No. 3, pp. 167-178
- Chernousko, F. L.*: On equations of ellipsoids approximating reachable sets. Vol. 12, No. 2, pp. 97-110
- Csiszár, I.*: An abstract source-channel transmission theorem. Vol. 12, No. 5, pp. 303-307
- Dinh The Luc*: Duality in programming under probabilistic constraints with a random technology matrix. Vol. 12, No. 6, pp. 429-437
- Dyachkov, A. G., Rykov, V. V.*: A survey of superimposed code theory. Vol. 12, No. 4, pp. 229-242
- Gabasov, R., Kirillova, F. M., Kostyukova, O. I.*: Dual algorithm of optimization of a linear dynamic system. Vol. 12, No. 4, pp. 253-265
- Gerasimov, A. I.*: Analysis of queueing networks by polynomial approximation. Vol. 12, No. 3, pp. 219-227
- Györfi, L., Vajda, I.*: Block coding and correlation decoding for an M -user weighted adder channel. Vol. 12, No. 6, pp. 405-417
- Kleimenov, A. F.*: Optimal strategies in a hierarchical differential game. Vol. 12, No. 6, pp. 369-377
- Kotarski, W., Kowalewski, A.*: On optimal control problem with initial state not a priori given. Vol. 12, No. 5, pp. 349-359
- Krasovskii, N. N., Tret'yakov, V. E.*: A stochastic program synthesis of a guaranteeing control. Vol. 12, No. 2, pp. 79-95
- Krzyżak, A., Pawlak, M.*: Universal consistency results for Wolwerton-Wagner regression function estimate with application in discrimination. Vol. 12, No. 1, pp. 33-42
- Kučera, V.*: Block decoupling by dynamic compensation with internal properness and stability. Vol. 12, No. 6, pp. 379-389
- Lipcey, Zs.*: N -person nonlinear qualitative differential games with incomplete information, survey of results. Vol. 12, No. 2, pp. 111-122
- Maršik, J.*: A new conception of digital adaptive PSD control. Vol. 12, No. 4, pp. 267-277
- Muntean, I.*: Simultaneous optimization of maintenance and replacement policy for machines. Vol. 12, No. 4, pp. 279-291
- Novovicová, J.*: Robustness of Bayes estimation of dynamic system parameters. Vol. 12, No. 2, pp. 123-132
- Ovseevich, A. I.*: Extremal properties of ellipsoids approximating attainability sets. Vol. 12, No. 1, pp. 43-54
- Pedrycz, W.*: Towards set-theoretic representation of nondeterministic systems. Vol. 12, No. 3, pp. 179-193
- Rykov, A. S.*: Simplex algorithms for unconstrained minimization. Vol. 12, No. 3, pp. 195-207
- Šebek, M.*: Direct polynomial approach to discrete-time stochastic tracking. Vol. 12, No. 4, pp. 293-300
- Šolc, F., Halabala, J.*: Adaptive transform picture coding for robot vision system. Vol. 12, No. 5, pp. 309-322
- Stepanov, S. N.*: Asymptotic formulae and estimations for probabilistic characteristics of full-available group with absolutely persistent subscriber. Vol. 12, No. 5, pp. 361-369
- Subbotina, N. N., Subbotin, A. I.*: On sensitivity of the value function of the differential game with the integral-terminal payoff. Vol. 12, No. 3, pp. 153-166
- Šujan, Š.*: Sinai's theorem and entropy compression. Vol. 12, No. 6, pp. 419-428
- Vahrameev, S. A.*: On nonlinear differential games of pursuit. Vol. 12, No. 5, pp. 323-333
- Vishnevskii, V. M., Gerasimov, A. I.*: Analysis of flows in closed exponential queueing networks. Vol. 12, No. 6, pp. 391-404
- Vostrý, Z.*: Contribution to simulation of distributed parameter systems. Vol. 12, No. 1, pp. 55-62
- Yashkov, S. F.*: A derivation of response time distribution for a $M|H|1$ processor-sharing queue. Vol. 12, No. 2, pp. 133-148
- Yemelyanov, S. V., Korovin, S. K., Ulanov, B. V.*: Control of nonstationary dynamic systems with quasicontinuous generation of the control signal. Vol. 12, No. 1, pp. 11-32
- Yemelyanov, S. V., Soloviev, A. A.*: Application of new feedback types in the problem of signal differentiation. Vol. 12, No. 2, pp. 63-77
- Zinoviev, V. A.*: Cascade equal-weight codes and maximal packings. Vol. 12, No. 1, pp. 3-10

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 12 (1983)

AUTHOR INDEX

- Agrachev, A. A. Vol. 12, No. 5, pp. 335-347
 Albores, F. J. Vol. 12, No. 4, pp. 243-252
 Azimov, A. Ya. Vol. 12, No. 3, pp. 209-219
 Bocharov, P. P. Vol. 12, No. 4, pp. 243-252
 Chaudhuri, A. K. Vol. 12, No. 3, pp. 167-178
 Chernousko, F. L. Vol. 12, No. 2, pp. 97-110
 Csiszár, I. Vol. 12, No. 5, pp. 303-307
 Dinh The Luc Vol. 12, No. 6, pp. 429-437
 Dyachkov, A. G. Vol. 12, No. 4, pp. 229-242
 Gabasov, R. Vol. 12, No. 4, pp. 253-265
 Gerasimov, A. I. Vol. 12, No. 3, pp. 219-227;
 Vol. 12, No. 6, pp. 391-404
 Györfi, L. Vol. 12, No. 6, pp. 405-417
 Halabala, I. Vol. 12, No. 5, pp. 309-322
 Kirillova, F. M. Vol. 12, No. 4, pp. 253-265
 Kleimenov, A. F. Vol. 12, No. 6, pp. 369-377
 Korovin, S. K. Vol. 12, No. 1, pp. 11-32
 Kostyukova, O. I. Vol. 12, No. 4, pp. 253-265
 Kotarski, W. Vol. 12, No. 5, pp. 349-359
 Kowalewski, A. Vol. 12, No. 5, pp. 349-359
 Krasovskii, N. N. Vol. 12, No. 2, pp. 79-95
 Krzyżak, A. Vol. 12, No. 1, pp. 33-42
 Kučera, V. Vol. 12, No. 6, pp. 379-389
 Lipcey, Zs. Vol. 12, No. 2, pp. 111-122
 Maršik, J. Vol. 12, No. 4, pp. 267-277
 Mukherjee, R. N. Vol. 12, No. 3, pp. 167-178
 Muntean, I. Vol. 12, No. 4, pp. 279-291
 Novovičová, J. Vol. 12, No. 2, pp. 123-132
 Ovseevich, A. I. Vol. 12, No. 1, pp. 43-54
 Pawlak, M. Vol. 12, No. 1, pp. 33-42
 Pedrycz, W. Vol. 12, No. 3, pp. 179-193
 Rykov, A. S. Vol. 12, No. 3, pp. 195-207
 Rykov, V. V. Vol. 12, No. 4, pp. 229-242
 Sarychev, A. V. Vol. 12, No. 5, pp. 335-347
 Šebek, M. Vol. 12, No. 4, pp. 293-300
 Šolc, F. Vol. 12, No. 5, pp. 309-322
 Soloviev, A. A. Vol. 12, No. 2, pp. 63-77
 Stepanov, S. N. Vol. 12, No. 5, pp. 361-369
 Subbotin, A. I. Vol. 12, No. 3, pp. 153-166
 Subbotina, N. N. Vol. 12, No. 3, pp. 153-166
 Šujan, Š. Vol. 12, No. 6, pp. 419-428
 Tret'yakov, V. E. Vol. 12, No. 2, pp. 79-95
 Ulanov, B. V. Vol. 12, No. 1, pp. 11-32
 Vahrameev, S. A. Vol. 12, No. 5, pp. 323-333
 Vajda, I. Vol. 12, No. 6, pp. 405-417
 Vishnevskii, V. M. Vol. 12, No. 6, pp. 391-404
 Vostrý, Z. Vol. 12, No. 1, pp. 55-62
 Yashkov, S. F. Vol. 12, No. 2, pp. 133-148
 Yemelyanov, S. V. Vol. 12, No. 1, pp. 11-32;
 Vol. 12, No. 2, pp. 63-77
 Zinoviev, V. A. Vol. 12, No. 1, pp. 3-10

2) Apply a non-singular stable proper rational matrix U_i to bring the matrix C_i , $i = 1, \dots, k$ to the form

$$C_i = U_i \begin{bmatrix} C'_i \\ 0 \end{bmatrix}$$

where C'_i has linearly independent rows over $F(s)$.

3) Calculate a right inverse G of

$$C' = \begin{bmatrix} C'_1 \\ \vdots \\ C'_k \end{bmatrix}.$$

4) Choose a function $t \in F\{s\}$ such that Gt is a stable proper rational matrix.

5) Solve the equation

$$PA + QB = I$$

for stable proper rational matrices P and Q , with P non-singular.

6) Set $R = Gt$.

7) Then

$$[-K_w \ K] = P^{-1}[QR].$$

If Step 3 fails, the decoupling with respect to p_1, \dots, p_k is impossible and if Step 5 fails, the decoupled system cannot be made internally proper and stable.

The above construction yields a decomposition of the overall system into $p_i \times r_i$ non-interacting blocks so that the external input vector has the dimension

$$r = \sum_{i=1}^k r_i = \text{rank } C.$$

When realizing the compensator we must bear in mind the fundamental assumption concerning its unreachable and unobservable internal states.

6. Conclusions

A necessary and sufficient condition has been given for the block decoupling of linear systems by dynamic compensation while ensuring the internal properness and stability. The analysis has been based on transfer matrix concepts and has resulted in a general yet relatively simple design procedure.

The main contribution of the paper is in ensuring the *internal properness and stability* of the decoupled system. Using the notion of stable proper factorizations, these two properties are handled in a unified way to give a complete result under the most

general circumstances. In particular, the system is described by a rational transfer matrix, not necessarily proper or stable, and an output of the system, possibly different from the output to be controlled, is assumed to be available for measurement. No restrictive assumptions are imposed on the compensator but the linearity. The admissibility condition is chosen in a natural way to guarantee that no essential loss of control occurs through the decoupling process rather than insisting on the complete output controllability.

References

1. Basile, G., Marro, G., A state space approach to non interacting controls. Ric. Autom. **1**, 68–77, 1970.
2. Denham, M. J., A necessary and sufficient condition for decoupling by output feedback. IEEE Trans. Automatic Control **AC-18**, 537, 1973.
3. Descusse, J., Malabre, M., Solvability of the decoupling problem for linear constant (A, B, C, D) -quadruples with regular output feedback. IEEE Trans. Automatic Control **AC-27**, 456–458, 1982.
4. Descusse, J., Dion, J. M., On the structure at infinity of linear square decoupled systems. IEEE Trans. Automatic Control **AC-27**, 971–974, 1982.
5. Descusse, J., Lafay, J. F., Kučera, V., Decoupling by restricted static state feedback — The general case. IEEE Trans. Automatic Control, to appear.
6. Dion, J. M., Feedback block decoupling and infinite structure of linear systems. Report L.A.G. 82–24, E.N.S. d'Ingénieurs Electriciens de Grenoble, France, 1982.
7. Falb, P. L., Wolowich, W. A., Decoupling in the design and synthesis of multivariable control systems. IEEE Trans. Automatic Control **AC-12**, 651–659, 1967.
8. Filev, D. P., Some new results in state space decoupling of multivariable systems I — A link between geometric and matrix methods. Kybernetika **18**, 215–233, 1982.
9. Filev, D. P., Some new results in state space decoupling of multivariable systems II — Extensions to decoupling of systems with $D \neq 0$ and output feedback decoupling. Kybernetika **18**, 330–344, 1982.
10. Gilbert, E. G., The decoupling of multivariable systems by state feedback. SIAM J. Control **7**, 50–63, 1969.
11. Hautus, M. L. J., Heymann, M., Linear feedback decoupling — Transfer function analysis. IEEE Trans. Automatic Control, to appear.
12. Hazlerigg, A. D. G., Sinha, P. K., A non-interacting control by output feedback and dynamic compensation. IEEE Trans. Automatic Control **AC-23**, 76–79, 1978.
13. Howze, J. W., Necessary and sufficient conditions for decoupling using output feedback. IEEE Trans. Automatic Control **AC-18**, 44–46, 1973.
14. Howze, J. W., Pearson, J. B., Decoupling and arbitrary pole placement in linear systems using output feedback. IEEE Trans. Automatic Control **AC-15**, 660–663, 1970.
15. Kavanagh, R. J., Noninteracting controls in linear multivariable systems. AIEE Trans. Applications and Industry **76**, 95–100, 1957.
16. Koussiouris, T. G., A frequency domain approach to the block decoupling problem. Internat. J. Control **29**, 991–1010, 1979.
17. Kučera, V., Algebraic Theory of Discrete Linear Control (in Czech). Academia, Prague, 1978.
18. Kučera, V., Block decoupled systems (in Czech). Proc. Internat. Conf. COMEMOP, Bratislava, Czechoslovakia, 1983.
19. Mejerov, M. V., Multivariable Control Systems (in Russian). Nauka, Moscow, 1965.
20. Morgan, B. S., The synthesis of linear multivariable systems by state feedback. Proc. JACC, 468–472, 1964.
21. Morse, A. S., Wonham, W. M., Decoupling and pole placement by dynamic compensation. SIAM J. Control **8**, 317–337, 1970.

22. *Morse, A. S., Wonham, W. M.*, Status of non interacting control. IEEE Trans. Automatic Control **AC-16**, 568–581, 1971.
23. *Silverman, L. M., Payne, H. J.*, Input-output structure of linear systems with application to the decoupling problem. SIAM J. Control **9**, 199–233, 1971.
24. *Strejc, V.*, The general theory of autonomy and invariance of linear systems of control. Acta Technica **5**, 235–258, 1960.
25. *Vardulakis, A. I. G.*, On infinite zeros. Internat. J. Control **32**, 849–866, 1980.
26. *Voznesenskij, I. N.*, A control of systems with many outputs (in Russian). Automat. i Telemekh. **4**, 7–38, 1936.
27. *Wolovich, W. A.*, Linear Multivariable Systems. Springer, New York, 1974.
28. *Wonham, W. M.*, Linear Multivariable Control. Springer, New York, 1974.
29. *Wonham, W. M., Morse, A. S.*, Découpling and pole assignment in linear systems — A geometric approach. SIAM J. Control **8**, 1–18, 1970.

Обеспечение автономности вместе с внутренней правильностью и устойчивостью методом динамической компенсации

В. КУЧЕРА

(Прага)

Задача разбиения линейной многосвязной системы на автономные блоки с помощью динамической компенсации решается методом передаточных матриц. Выведено простое необходимое и достаточное условие разрешимости задачи допустимым компенсатором. Основное внимание обращается на процедуру решения, которая обеспечивает внутреннюю правильность и устойчивость автономной системы. Эффективное исследование этих свойств опирается на понятие устойчивых правильных факторизаций передаточных матриц.

V. Kučera

Institute of Information Theory and Automation

Czechoslovak Academy of Sciences

Pod vodárenskou věží 4

182 08 Praha 8, Czechoslovakia

ИССЛЕДОВАНИЕ ПОТОКОВ В ЗАМКНУТЫХ ЭКСПОНЕНЦИАЛЬНЫХ СЕТЯХ МАССОВОГО ОБСЛУЖИВАНИЯ

В. М. ВИШНЕВСКИЙ, А. И. ГЕРАСИМОВ

(Москва)

(Поступила в редакцию 3 сентября 1982 г.)

Рассматривается замкнутая экспоненциальная сеть произвольной структуры с одним классом сообщений.

Предложен алгоритм отыскания функции распределения интервалов между сообщениями, выходящими из узлов сети.

Для замкнутой циклической экспоненциальной сети массового обслуживания с произвольным числом узлов получены в явном виде распределение и первые два момента интервалов между выходящими из узлов сообщениями.

Введение

При проектировании современных вычислительных сетей и систем широкое применение находит математический аппарат теории стохастических сетей, использование которого для аналитического моделирования вычислительных сетей позволяет получать такие важные характеристики как пропускная способность и задержка сообщений в сети; исследовать эффективность различных методов управления потоками и проводить сравнительный анализ протоколов; осуществлять выбор объема буферной памяти узлов коммутации сообщений (пакетов); осуществлять расчет мультипрограммных вычислительных систем и т. д. В [1, 2] предложен метод полиномиальной аппроксимации для анализа замкнутых, открытых и смешанных сетей массового обслуживания произвольной структуры с несколькими классами сообщений, несколькими подцепями, произвольными функциями распределения времени обслуживания в узлах, приоритетами и блокировками.

Обобщение метода, предложенного в [1, 2], предполагает использование первых двух моментов интервалов времени между последовательными поступлениями сообщений в узлы сети для исследования замкнутых сетей. В связи с этим возникает необходимость анализа функции распределения интервалов времени между последовательными поступлениями (или выходами) сообщений из узлов в замкнутых сетях.

Свойства выходных потоков в однолинейных системах массового обслуживания исследовались в ряде работ [3–10], где, в частности, описаны условия, при которых выходящий поток будет пуассоновским. Эти условия определяют многофазные системы, на фазах которых входящие и выходящие потоки будут пуассоновскими. В [11–15] исследовались потоки в открытых экспоненциальных сетях с обратными связями и пуассоновскими внешними потоками. В [11] было показано, что между узлами i и j поток будет пуассоновским, если вероятность перехода из i в j больше нуля, но узел i не достижим из узла j , причем потоки между другими узлами не будут пуассоновскими и даже не будут процессами восстановления. Для двухузловой циклической экспоненциальной сети в [16] получена функция распределения времени цикла. Потоки в замкнутой циклической сети рассматривались также в [17], где показано, что поток в сети не является рекуррентным.

В настоящей работе анализируются потоки в замкнутой экспоненциальной сети произвольной структуры с одним классом сообщений (каждый узел достижим из любого узла).

Предложен алгоритм отыскания функции распределения интервалов между сообщениями, выходящими из узлов в экспоненциальной сети массового обслуживания произвольной структуры.

Для важного частного случая замкнутой циклической экспоненциальной сети массового обслуживания с произвольным числом узлов получены в явном виде распределение и первые два момента интервалов времени между выходящими из узлов сообщениями.

Замкнутая сеть произвольной структуры

Рассматривается замкнутая сеть массового обслуживания, состоящая из M узлов, между которыми циркулирует N однотипных сообщений в соответствии с матрицей переходных вероятностей $\|P_{ij}\|$.

Обозначим:

$A_i(t) = 1 - e^{-\mu_i t}$ — функция распределения времени обслуживания в i -ом ($i = 1, M$) узле;

μ_i — интенсивность обслуживания в узле i ($i = \overline{1, M}$);

$(\bar{n}, n_M) = (n_1, \dots, n_{M-1}, n_M)$ — состояние сети в стационарном режиме, где n_i — число сообщений в i -ом узле ($i = 1, M$);

$$\varepsilon(n) = \begin{cases} 0 & (n=0) \\ 1 & \text{в противном случае;} \end{cases}$$

$$\begin{aligned} \bar{n} &= (n_1, \dots, n_{M-1}); \\ \bar{n}^{kj} &= \begin{cases} n_1, \dots, n_k - 1, \dots, n_j + 1, \dots, n_{M-1} & (i \neq j) \\ \bar{n} & (i = j) \end{cases}; \\ \bar{n}^j &= (n_1, \dots, n_j + 1, \dots, n_{M-1}); \\ \bar{n}_j &= (n_1, \dots, n_j, 0, \dots, 0), \quad \text{причем } n_j \geq 1, n_{j+1} = \dots = n_{M-1} = 0. \end{aligned}$$

Событие, состоящее в том, что за время t , отсчитываемое от момента окончания обслуживания в узле M , закончится обслуживание очередного сообщения, может осуществиться следующими несовместными способами:

1. В момент окончания обслуживания в M -ом узле сеть находилась в состоянии $(\bar{n}, 0)$, вероятность чего $P_{N-1}(\bar{n}, 0)$; обслуженное в M -ом узле сообщение перешло в узел j ($j = \overline{1, M-1}$), вероятность чего P_{Mj} ; за время t какое-либо сообщение поступило в узел M и обслужилось в нем, вероятность чего есть

$$R(\bar{n}^j, 0, t) * A_M(t)$$

(где $*$ — символ свертки).

$R(\bar{n}, 0, t)$ — функция распределения интервала времени от момента прихода сообщения в какой-либо узел сети (сеть переходит в состояние $(\bar{n}, 0)$) до момента достижения каким-нибудь из сообщений узла M при условии, что в момент прихода сообщения узел M был пуст.

2. В момент окончания обслуживания в M -ом узле сеть находилась в состоянии $(\bar{n}, n_M \geq 1)$, вероятность чего $P_{N-1}(\bar{n}, n_M)$; за время t сообщение обслужилось в M -ом узле, вероятность чего $A_M(t)$.

3. В момент окончания обслуживания в M -ом узле сеть находилась в состоянии $(\bar{n}, 0)$, вероятность чего $P_{N-1}(\bar{n}, 0)$; обслуженное в M -ом узле сообщение вернулось в M -ый узел с вероятностью P_{MM} ; за время t сообщение обслужилось в узле M , вероятность чего $A_M(t)$.

Исходя из этого, преобразование Лапласа–Стилтьеса (пр. Л–С) $\varphi_M(s)$ функции распределения $\Phi_M(t)$ времени между последовательными моментами выхода сообщений из узла M может быть представлена в виде:

$$\begin{aligned} \varphi_M(s) &= \frac{\mu_M}{\mu_M + s} \sum_{j=1}^{M-1} \sum_{|\bar{n}|=N-1} P_{N-1}(\bar{n}, 0) \pi(\bar{n}^j, 0, s) P_{Mj} + \\ &+ \frac{\mu_M}{\mu_M + s} \sum_{\substack{|\bar{n}, n_M|=N-1 \\ n_M \geq 1}} P_{N-1}(\bar{n}, n_M) + \frac{\mu_M}{\mu_M + s} P_{MM} \sum_{\substack{|\bar{n}|=N-1 \\ n_M=0}} P_{N-1}(\bar{n}, 0), \end{aligned} \quad (1)$$

где: $\pi(\bar{n}^j, 0, s)$ — пр. Л-С от функции распределения $R(\bar{n}^j, 0, t)$;

$$|\bar{n}| = \sum_{i=1}^{M-1} n_i; \quad |\bar{n}, n_M| = \sum_{i=1}^M n_i;$$

N — число сообщений, циркулирующих в сети.

Здесь стационарные вероятности $P_{N-1}(\bar{n}, n_M)$ вычисляются в соответствии с [18, 19] как стационарные вероятности состояний замкнутой сети с $N-1$ сообщением в произвольный момент времени.

Легко показать,* что $\pi(\bar{n}, 0, s)$ определяются из следующей системы линейных уравнений:

$$\begin{aligned} \pi(\bar{n}, 0, s) \left(\sum_{k=1}^{M-1} \varepsilon(n_k) \mu_k + s \right) = \\ = \sum_{k=1}^{M-1} \left[\mu_k \varepsilon(n_k) \sum_{j=1}^{M-1} \pi(\bar{n}^{kj}, 0, s) P_{kj} \right] + \sum_{k=1}^{M-1} \varepsilon(n_k) \mu_k P_{kM}. \end{aligned} \quad (2)$$

Решая (2) и подставляя $\pi(\bar{n}, 0, s)$ в (1), получаем $\varphi_M(s)$. Выпишем, например, систему (2) для сети рис. 1, являющейся моделью мультипрограммных

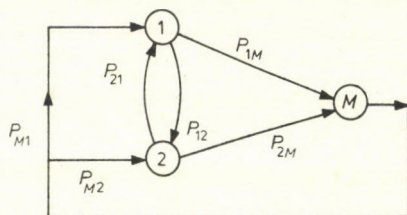


Рис. 1

вычислительных систем [20] при уровне мультипрограммирования $N=2$ (число узлов $M=3$).

$$\begin{aligned} n(1, 1, 0, s) = (\pi(0, 2, 0, s) P_{12} + P_{1M}) \frac{\mu_1}{\mu_1 + \mu_2 + s} + \\ + (\pi(2, 0, 0, s) P_{21} + P_{2M}) \frac{\mu_2}{\mu_1 + \mu_2 + s}, \end{aligned}$$

* Действительно, если рассмотреть узлы, занятые в момент прихода сообщения в какой-нибудь узел сети, то пр. Л-С функции распределения времени до выхода сообщения из (занятого) узла k , при условии, что из остальных узлов сообщение не вышло, представляется в виде

$$\int_0^{\infty} e^{-\sum_{i \neq k} \varepsilon(n_i) \mu_i v - sv} d(1 - e^{-\mu_k v}) = \mu_k \left(\sum_{i=1}^{M-1} \varepsilon(n_i) \mu_i + s \right).$$

Отсюда с учетом того, что переход в состояние $(\bar{n}^{kj}, 0)$ происходит с вероятностью P_{kj} , а вероятность прихода сообщения в узел M равна P_{kM} , получаем систему (2).

$$\pi(0, 2, 0, s) = (\pi(1, 1, 0, s)P_{21} + P_{2M}) \frac{\mu_2}{\mu_2 + s},$$

$$\pi(2, 0, 0, s) = (\pi(1, 1, 0, s)P_{12} + P_{1M}) \frac{\mu_1}{\mu_1 + s}.$$

Решение указанной системы и отыскание $\varphi_M(s)$ позволяет в данном случае определить производительность и другие характеристики рассматриваемой мультипрограммой вычислительной системы.

Циклическая сеть

Рассмотрим теперь замкнутую циклическую экспоненциальную сеть, содержащую M узлов, в которой циркулирует N сообщений (Рис. 2).

Обозначим через $\tau_M^{(1)}$ и $\tau_M^{(2)}$ первые два момента интервалов между выходящими из узла M сообщениями.

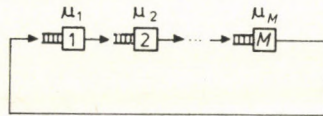


Рис. 2

Введем функции $g(n, m)$ и $G(n)$:

$$g(n, 1) = (a_1)^n \quad \text{для } n = \overline{0, N};$$

$$g(0, m) = 1 \quad \text{для } m = \overline{1, M};$$

$$g(n, m) = g(n, m-1) + a_m g(n-1, m) \quad \text{при } m > 1 \text{ и } n > 0;$$

$$G(n) = g(n, M) \quad \text{для } n = \overline{0, N};$$

где $a_i = 1/\mu_i$ ($i = 1, M$).

Обозначим $\underset{k=j}{*}^M A_k(t) = A_j(t) * A_{j+1}(t) * \dots * A_M(t)$.

Утверждение 1. При указанных выше предположениях

$$\Phi_M(t) = \sum_{j=1}^M \left(Q_j \left(\underset{k=j}{*}^M A_k(t) \right) \right),$$

$$\varphi_M(s) = \sum_{j=1}^M \left(Q_j \prod_{k=j}^M \left(\frac{\mu_k}{\mu_k + s} \right) \right),$$

$$\tau_M^{(1)} = \sum_{j=1}^M \left(Q_j \sum_{k=j}^M \frac{1}{\mu_k} \right) = \frac{G(N)}{G(N-1)},$$

$$\tau_M^{(2)} = \sum_{j=1}^M \left\{ Q_j \left[\left(\sum_{k=1}^M \frac{1}{\mu_k} \right)^2 + \sum_{k=j}^M \frac{1}{\mu_k^2} \right] \right\},$$

где $Q_j = a_j g(N-2, j) / G(N-1)$.

Доказательство. В данном случае матрица переходных вероятностей $\|P_{ij}\|$ преобразуется к виду: $P_{i, i+1} = 1$ ($i = 1, M-1$), $P_{M1} = 1$, остальные

$$P_{ij} = 0. \quad (3)$$

Легко проверить непосредственной подстановкой, что решение системы уравнений (2) имеет вид:

$$\pi(\bar{n}_j, 0, s) = \prod_{k=j}^{M-1} \left(\frac{\mu_k}{\mu_k + s} \right). \quad (4)$$

Из (1) с учетом (3), (4) получаем функцию распределения $\Phi_M(t)$ в виде:

$$\Phi_M(t) = \sum_{j=1}^M \left(\sum_{k=j}^M A_k(t) \right) \sum_{\substack{(n_1, \dots, n_j) \\ n_j \geq 1}} P_{N-1}(n_1, \dots, n_j, 0, \dots, 0).$$

В соответствии с [21]

$$P_{N-1}(n_1, \dots, n_M) = \frac{1}{G(N-1)} \prod_{i=1}^M (a_i)^{n_i}.$$

Примем:

$$\begin{aligned} \sum_{\substack{(n_1, \dots, n_j) \\ n_j \geq 1}} P_{N-1}(n_1, \dots, n_j, 0, \dots, 0) &= \sum_{\substack{(n_1, \dots, n_j) \\ n_j \geq 1}} \frac{1}{G(N-1)} \prod_{i=1}^j (a_i)^{n_i} = \\ &= a_j \frac{1}{G(N-1)} \sum_{\substack{(n_1, \dots, n_j) \\ \sum_{k=1}^j n_k = N-2}} \prod_{i=1}^j (a_i)^{n_i} = a_j \frac{1}{G(N-1)} g(N-2, j) = Q_j. \end{aligned}$$

В частности,

$$Q_1 = a_1 \frac{a_1^{(N-2)}}{G(N-1)}, \quad Q_M = a_M \frac{G(N-2)}{G(N-1)}.$$

Перепишем $\Phi_M(t)$ в виде:

$$\Phi_M(t) = \sum_{j=1}^M \left(Q_j \left(\sum_{k=j}^M A_k(t) \right) \right),$$

откуда

$$\varphi_M(s) = \sum_{j=1}^M Q_j \prod_{k=j}^M \left(\frac{\mu_k}{\mu_k + s} \right). \quad (5)$$

Из (5) находим моменты $\tau_M^{(r)}$ интервалов между сообщениями, выходящими из узла M :

$$\tau_M^{(r)} = (-1)^r \frac{d^r}{(ds)^r} \varphi_M(s) |_{s=0}.$$

В частности, первые два момента $q_M^{(1)}$ и $\tau_M^{(2)}$ имеют вид:

$$\tau_M^{(1)} = \sum_{j=1}^M \left(\sum_{k=1}^M Q_j \frac{1}{\mu_k} \right);$$

$$\tau_M^{(2)} = \sum_{j=1}^M \left\{ Q_j \left[\left(\sum_{k=1}^M \frac{1}{\mu_k} \right)^2 + \sum_{k=1}^M \frac{1}{\mu_k^2} \right] \right\}.$$

В приложении 1 доказано, что

$$\tau_M^{(1)} = \sum_{j=1}^M \left(\sum_{k=j}^M Q_j \frac{1}{\mu_k} \right) = \frac{G(N)}{G(N-1)}.$$

Полученное выражение для $\tau_M^{(1)}$ совпадает с выражением, найденным в [21] (см. приложение 1).

Коэффициент вариации $C_M^{(r)}$ интервалов между выходящими из узла M сообщениями равен:

$$C_M^{(r)} = (\tau_M^{(2)} - (\tau_M^{(1)})^2) / (\tau_M^{(1)})^2 = \tau_M^{(2)} / (\tau_M^{(1)})^2 - 1.$$

В моменты выходов сообщений из любых узлов стационарное распределение вероятностей состояний замкнутой экспоненциальной сети с N сообщениями совпадает со стационарным распределением вероятностей состояний сети с $N-1$ сообщением в произвольный момент [18, 19] независимо от того, из какого узла вышло сообщение. Другими словами, любое сообщение, приходя в любой узел i , каждый раз с одинаковой вероятностью $P_{N-1}(n_i)$ встречает в нем n_i сообщений. Для того, чтобы сообщение, выйдя из любого узла, вернулось в этот же узел, оно должно пройти через все узлы по одному разу. Учитывая экспоненциальный характер обслуживания в узлах сети, получаем, что функция распределения времени цикла одна и та же для всех узлов.

Случайная величина времени цикла относительно любого узла представляет собой сумму N интервалов времени между последовательными выходами сообщений из этого узла, так как за время цикла какого-нибудь сообщения относительно любого узла в циклической сети через этот узел пройдет ровно N сообщений. Из сказанного следует, что функции распределения интервалов времени между последовательными выходами сообщений из узла одинаковы для всех узлов циклической экспоненциальной сети.

Таким образом,

$$\tau_i^{(r)} = \tau_M^{(r)} = \tau^{(r)}; \quad r = 1, 2, \dots; \quad i = \overline{1, M}.$$

Рассмотрим асимптотическое поведение циклической сети при больших $N \rightarrow \infty$ нагрузках.

Обозначим

$$\lambda_1 = 1/\tau^{(1)}, \quad \rho_i = \lambda_i a_i = \lambda_1 a_i, \quad b_i = a_i/a_1 \quad (i = \overline{1, M}),$$

$$b_* = \max_{2 \leq i \leq M} b_i, \quad \lambda_{*1} = 1/a_1, \quad \rho_{*i} = \lambda_{*1} a_i.$$

Пусть, кроме того

$$S(N, M) = \left\{ (n_1, \dots, n_M) \left| \sum_{i=1}^M n_i = N \quad \text{и} \quad n_i \geq 0 \forall i \right. \right\},$$

$$S_0(N, M) = \left\{ (n_1, \dots, n_M) \left| \sum_{i=2}^M n_i = N \quad \text{и} \quad n_i \geq 0 \forall i \right. \right\},$$

$$S_1(N, M) = \left\{ (n_1, \dots, n_M) \left| \sum_{i=2}^M n_i \leq N-1 \quad \text{и} \quad n_i \geq 0 \forall i \right. \right\},$$

$$S_{1k}(N, M) = \left\{ (n_1, \dots, n_M) \left| \sum_{i=2}^M n_i = k \quad \text{и} \quad n_i \geq 0 \forall i \right. \right\}.$$

Предположим (наиболее интересный случай), что $\exists i_*: a_{i_*} = \max_i a_i$. Без ограничения общности пусть $i_* = 1$.

Утверждение 2. В сделанных выше предположениях существует

$$\lim_{N \rightarrow \infty} P\{n_1 \geq 1\} = 1 \quad \text{и} \quad \lim_{N \rightarrow \infty} P(n_1, \dots, n_M) = \prod_{i=2}^M (1 - \rho_{*i}) \rho_{*i}^{n_i},$$

причем скорость сходимости при $k \rightarrow \infty$ не менее

$$C((Mb_* - b_*^2)/(Mb_* - 2b_* + 1))^k,$$

где C — некоторая постоянная. (Доказательство приведено в Приложении 2.)

Из утверждения 2 следует, что асимптотически, при больших нагрузках ($N \rightarrow \infty$), характеристики рассматриваемой циклической сети будут совпадать с характеристиками многофазной экспоненциальной системы, состоящей из $M-1$ узла, на вход которой поступает пуассоновский поток с параметром $1/a_1$.

Для оценки того, насколько потоки в замкнутой циклической экспоненциальной сети отличаются от пуассоновских при различных нагрузках сети (различных N) можно использовать в качестве критерия* величину $\chi = [2 - \tau^{(2)}/(\tau^{(1)})^2] \cdot 100\%$, равную разнице (в %) между коэффициентом вариации экспоненциального распределения и коэффициентом вариации распределения интервалов между выходящими из узлов сообщениями.

* Заметим, что могут быть использованы и другие критерии [22].

Выпишем χ в аналитическом виде:

$$\chi = \left[2 - G(N-1)/(G(N))^2 \sum_{j=1}^M \left\{ a_j g(N-2, j) \left[\left(\sum_{k=j}^M a_k \right)^2 + \sum_{k=j}^M a_k^2 \right] \right\} \right] \cdot 100\%.$$

На рис. 3 представлена зависимость χ от числа сообщений в циклической экспоненциальной сети с параметрами:

$$M=5, \quad a_1=10, \quad a_2=8, \quad a_3=2, \quad a_4=4, \quad a_5=6.$$

Видно, что при малых нагрузках поток существенно отличается от пуассоновского, но с увеличением числа сообщений в сети поток асимптотически приближается к пуассоновскому.

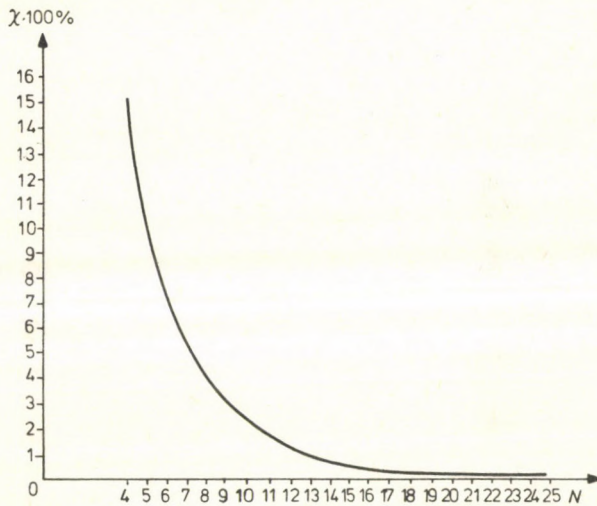


Рис. 3

Приложение 1

$\tau_M^{(1)}$ может быть представлено в следующем виде:

$$\begin{aligned} \tau_M^{(1)} &= \sum_{j=1}^M \left(\sum_{k=j}^M Q_j \frac{1}{\mu_k} \right) = \frac{1}{G(N-1)} \sum_{j=1}^M \sum_{k=j}^M a_j g(N-2, j) a_k = \\ &= \frac{1}{G(N-1)} \sum_{k=1}^M \sum_{j=1}^k a_j g(N-2, j) a_k = \frac{1}{G(N-1)} \sum_{k=1}^M a_k r_k, \end{aligned}$$

где

$$r_k = \sum_{j=1}^k a_j g(N-2, j).$$

Докажем индукцией по k , что $r_k = g(N-1, k)$, используя свойство $g(n, k) = g(n, k-1) + a_k g(n-1, k)$ (при $k > 1$ и $n > 0$),

$$r_1 = a_1 g(N-2, 1) = a_1 a_1^{N-2} = a_1^{N-1} = g(N-1, 1),$$

$$r_2 = r_1 + a_2 g(N-2, 2) = g(N-1, 1) + a_2 g(N-2, 2) = g(N-1, 2).$$

Предполагая, что $r_{k-1} = g(N-1, k-1)$, получим

$$r_k = r_{k-1} + a_k g(N-2, k) = g(N-1, k-1) + a_k g(N-2, k) = g(N-1, k).$$

Обозначим

$$s_i = \sum_{k=1}^i a_k r_k = \sum_{k=1}^i a_k g(N-1, k).$$

Докажем индукцией по i , что $s_i = g(N, i)$,

$$s_1 = a_1 g(N-1, 1) = a_1 a_1^{N-1} = a_1^N = g(N, 1),$$

$$s_2 = s_1 + a_2 g(N-1, 2) = g(N, 1) + a_2 g(N-1, 2) = g(N, 2).$$

Предполагая, что $s_{i-1} = g(N, i-1)$, получаем

$$s_i = s_{i-1} + a_i g(N-1, i) = g(N, i-1) + a_i g(N-1, i) = g(N, i).$$

При $i = M$

$$s_i = \sum_{k=1}^M a_k r_k = g(N, M) = G(N).$$

Таким образом,

$$\tau_M^{(1)} = G(N)/G(N-1).$$

Полученное выражение для $\tau_M^{(1)}$ совпадает с аналогичным выражением в [21], которое можно вычислить непосредственно исходя из вероятности занятости узла M

$$\tau_M^{(1)} = 1/(\mu_M P(n_M \geq 1)) = 1/(G(N-1)/G(N)) = G(N)/G(N-1). \quad (6)$$

Приложение 2

Из (6) следует, что $P\{n_1 \geq 1\} = a \frac{G(N-1)}{G(N)}$,

откуда

$$P\{n_1 \geq 1\} = \left(a_1^N \sum_{S(N-1, M)} \prod_{i=1}^M a_i^{n_i} \right) / \left(a_1^{N-1} \sum_{S(N, M)} \prod_{i=1}^M a_i^{n_i} \right) = \\ = \left(\sum_{S(N-1, M)} \prod_{i=2}^M b_i^{n_i} \right) / \left(\sum_{S(N, M)} \prod_{i=2}^M b_i^{n_i} \right).$$

Но

$$\sum_{S(N, M)} \prod_{i=2}^M b_i^{n_i} = \sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N, M)} \prod_{i=2}^M b_i^{n_i} = \\ = \sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_0(N-1, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N-1, M)} \prod_{i=2}^M b_i^{n_i}. \quad (7)$$

Так как общее количество состояний в сети с k сообщениями равно C_{M+k-1}^k ,

то

$$\sum_{S_0(k, M)} \prod_{i=2}^M b_i^{n_i} < C_{M+k-1}^k b_*^k = d_k.$$

Пусть

$$\varepsilon_k = b_*(M-1)/(k+1), \quad k_0 = b_*(M-1)/(1-b_*), \quad D_0 = d_{k_0} (b_* + \varepsilon_{k_0})^{k_0},$$

тогда при

$$k \geq k_0 \quad b_*(M+k)/(k+1) = b_* + \varepsilon_k \leq b_* + \varepsilon_{k_0} < 1.$$

Таким образом, при

$$k \geq k_0 \quad \varepsilon_{k_0} < d_{k+1} < d_k (b_* + \varepsilon_{k_0}),$$

отсюда при

$$k > k_0 \quad d_k < D_0 (b_* + \varepsilon_{k_0})^k. \quad (8)$$

Так как $b_* + \varepsilon_{k_0} < 1$, то $d_k \rightarrow 0$ при $k \rightarrow \infty$, следовательно, $\sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} \rightarrow 0$ при

$N \rightarrow \infty$. Представим $\sum_{S_1(N, M)} \prod_{i=2}^M b_i^{n_i}$ в виде $\sum_{S_1(N, M)} \prod_{i=2}^M b_i \sum_{k=0}^{N-1} \sum_{S_{1k}(N, M)} \prod_{i=2}^M b_i^{n_i}$.

Так как $\sum_{k=0}^{\infty} \sum_{S_{1k}(N, M)} \prod_{i=2}^M b_i^{n_i}$ есть произведение рядов $\sum_{n_i=0}^{\infty} b_i^{n_i}$ ($i = 2, \overline{M}$) в форме

Коши, то

$$\lim_{N \rightarrow \infty} \sum_{S_1(N, M)} \prod_{i=2}^M b_i^{n_i} = \prod_{i=2}^M \sum_{n_i=0}^{\infty} b_i^{n_i} = \prod_{i=2}^M (1/(1-b_i)).$$

Используя (7), получаем

$$\lim_{N \rightarrow \infty} P\{n_1 \geq 1\} = \lim_{N \rightarrow \infty} \left[\left(\sum_{S_0(N-1, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N-1, M)} \prod_{i=2}^M b_i^{n_i} \right) / \left(\sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_0(N-1, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N-1, M)} \prod_{i=2}^M b_i^{n_i} \right) \right] = 1.$$

Причем из (8) следует, что скорость сходимости при $k \rightarrow \infty$ не менее $C((Mb_* - b_*^2)/(Mb_* - 2b_* + 1))^k$, где C — некоторая постоянная.

Так как $\lim_{N \rightarrow \infty} P\{n_1 \geq 1\} = 1$, то

$$\lim_{N \rightarrow \infty} \lambda_1 = \lambda_{*1} = 1/a_1 \quad \text{и} \quad \lim_{N \rightarrow \infty} \rho_i = \rho_{*i} = b_i.$$

Поэтому

$$\begin{aligned} \lim_{N \rightarrow \infty} P(n_1, \dots, n_M) &= \left(\prod_{i=2}^M \rho_{*i}^{n_i} \right) / \left(\prod_{i=2}^M (1/(1-\rho_{*i})) \right) = \\ &= \prod_{i=2}^M (1-\rho_{*i}) \rho_{*i}^{n_i}. \end{aligned}$$

Литература

1. Вишнеvский В. М., Герасимов А. И. Приближенный метод исследования сетей массового обслуживания с несколькими классами сообщений. Третья школа по автоматизированным системам массового обслуживания. Винница, Винницкий Дом техники, с. 24–25, 1981.
2. Герасимов А. И. Приближенный метод исследования иерархических сетей массового обслуживания с блокировкой узлов сообщениями. Теория и техника автоматизированных систем массового обслуживания, М. Изд. Московского Дома научно-технической пропаганды, с. 82–85, 1982.
3. Александров А. М. О выходящих потоках одного класса систем массового обслуживания. Известия АН СССР. Техническая кибернетика, № 4, с. 3–11, 1968.
4. Яшков С. Ф. Об одном классе дисциплин обслуживания для моделей вычислительных сетей. Информационно-вычислительные сети ЭВМ, М. Изд. Московского Дома научно-технической пропаганды, с. 112–117, 1980.
5. Толмачев А. Л. Об обслуженном потоке системы $GI(M/1)r \leq \infty$. В сб. Системы массового обслуживания и коммутации, М., Наука, с. 8–17, 1974.
6. Вишнеvский В. М., Тимохова Т. А. О выходящем потоке системы массового обслуживания с ненадежным обслуживающим прибором. Актуальные вопросы теории и практики управления, М. Наука, с. 23–32, 1977.

7. Фалин Г. И. Влияние повторных вызовов на выходящий поток одноканальной системы массового обслуживания. Известия АН СССР. Техническая кибернетика № 4, с. 114–118, 1979.
8. Клейнрок Л. Теория массового обслуживания. М. Машиностроение, 1979.
9. Noetzel, A. S., A generalized discipline for product from network solutions, Journal of the ACM, v. 26, No. 4, pp. 779–793, 1979.
10. Наумов В. А. О независимой работе подсистем сложной системы. В кн. Теория массового обслуживания. Труды III Всесоюзной школы — совещания по теории массового обслуживания, т. 2, Изд. Московского Университета, 1976, с. 169–177.
11. Beutler, F. J., Melamed, B., Decomposition and customer streams of feedback networks of queues in equilibrium, Operations research, v. 26, No. 6, pp. 1059–1072, 1978.
12. Pujolle, G., Soula, C. A study of flows in queueing networks and approximate method for solution, 4-th International Symp. Modell. and Performance Eval. Comput. Systems, Vienna. Conf. Prepr., v. 2, pp. 189–203, 1979.
13. Disney, R. L. Random flow in queueing networks: A review and critique, AIEE Transactions, v. 7, No. 3, pp. 268–288, 1975.
14. Daley, D. J., Queueing output processes, Advances in Applied Probability, v. 8, No. 2, pp. 395–415, 1976.
15. Labetoulle, J., Pujolle, G., Soula, C., Stationary distributions of flows in Jacson networks, Math. Oper. Res., 1981, vol. 6, No. 2, pp. 173–185.
16. Chow, W.-M., The cycle time distribution of exponential cyclic queues, Journal of the ACM, v. 27, No. 2, pp. 281–286, 1980.
17. Вишнеvский В. М., Сибирская Т. К. Потоки в замкнутых сетях массового обслуживания. Третья школа по автоматизированным системам массового обслуживания. Винница, Винницкий Дом техники, с. 56, 1981.
18. Толмачев А. Л. Некоторые характеристики замкнутых экспоненциальных сетей. Теория телетрафика и информационные сети. М., Наука, с. 3–6, 1977.
19. Reiser, M., Lavenberg, S. S., Mean-value analysis of closed multichain queueing networks, Journal of the ACM, v. 27, No. 2, pp. 313–322, 1980.
20. Вишнеvский В. М., Твердохлебов А. С. Модели замкнутых сетей с блокировками для анализа мультипрограммных вычислительных систем. Автоматика и телемеханика № 5, с. 172–179, 1980.
21. Buzen, J. P., Computational algorithms for closed queueing networks with exponential servers, Communications of the ACM, v. 16, No. 9, pp. 527–531, 1973.
22. Башарин Г. П., Кокотушкин В. А., Наумов В. А. Метод эквивалентных замен в теории телетрафика. Итоги науки и техники, Электросвязь, т. 11, М., ВИНТИ, 1980, с. 82–122.

Analysis of flows in closed exponential queueing networks

V. M. VISHNEVSKII, A. I. GERASIMOV

(Moscow)

The results of flow structure investigations in closed queueing networks which are of great importance for development of the methods of real computing system and network analysis are presented.

A closed queueing network consisting of M nodes with N identical customers circulating between them with a transfer probability matrix $\|P_{ij}\|$ is considered. For this exponential closed network with the FCFS queueing discipline in the nodes the algorithm of finding distribution functions of intervals between successive customer arrivals or leavings from the network nodes is proposed. This algorithm solves a set of linear equations for Laplace–Stieltjes transforms of distribution functions of the time interval between the customer service termination in the chosen node and the arrival of any customer into this node, provided that the customer which has left the node, saw it empty and goes to one of the remaining $M-1$ nodes.

The distribution, mean and variance of intervals between customers leaving the nodes are explicitly obtained for an important particular case of closed cyclic exponential queueing network with an arbitrary number of nodes. Asymptotic behaviour of the cyclic network under large loads is investigated and the evaluation of closeness of the flows of customers circulating there to the Poisson one under different network loads (different N) is obtained.

В. М. Вишневский
Институт проблем управления
СССР Москва В-485,
Профсоюзная, 65

BLOCK CODING AND CORRELATION DECODING FOR AN M -USER WEIGHTED ADDER CHANNEL

L. GYÖRFI, I. VAJDA
(Budapest)

(Received February 6, 1983)

In this paper we investigate the performance of a block code for an M -user channel, the output of which is the weighted sum of delayed inputs with additive noise. Separated decoders of correlation type are supposed, not knowing the codebooks of the other users, and they are universal in the sense that they do not use the actual channel parameters.

1. Introduction

In the theory of the multiple access channel (MAC, [1]–[5]) the information transmission between separated encoders and common decoder is investigated and the capacity region is characterized. The common decoder obviously implies the knowledge of the codebooks of the encoders.

The results on interference channel (IFC, [10]–[14]) are based on the theory of MAC, therefore the separated decoders know the codebooks of all encoders.

The situation often occurs, where several sender-receiver pairs share a common IFC and each decoder knows only the codebook of the corresponding encoder ([15], Fig. 1). In the sequel the decoder of this restriction is called autonomic decoder. The autonomic decoders are important if, for example, there is no base station for

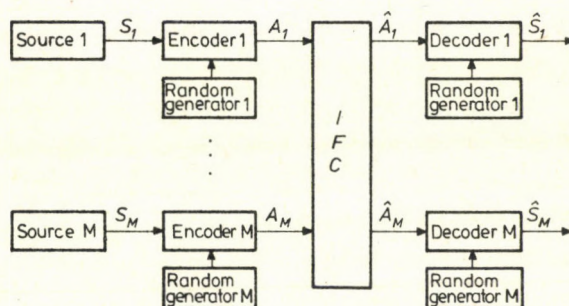


Fig. 1. IFC with autonomic decoders

common decoding or the privacy of communication is important or the population of users is changing from time to time.

On the one hand the information theoretic analysis of IFC is missing for asynchronous encoders and autonomic decoders, on the other hand the performance of particular coding procedures has been analysed if the IFC is generated by spread spectrum communication, and the outputs of the IFC are the weighted sum of the inputs ([6]–[9]). Each transmitter is assigned a unique pseudorandom sequence, which generates the codebook, and autonomic correlation decoders are applied. However, these codes are essentially one bit to block codes.

Our aim is to give a rate region for IFC if the code rates are identical and the IFC produces the weighted sum of the inputs with additive stationary and ergodic noise. To do this, a random block coding and correlation decoding is used.

For code word synchronization between the corresponding encoders and decoders an algorithm is presented, which is a generalization of the simple "sliding" correlator technique. It is shown that in the same rate region the decoding error probability and the synchronization error probability tend to zero as the block length tends to infinity.

2. The model of channel

The i -th encoder produces a binary (+1 and -1 valued) sequence $A_i = \{A_{i,n}, -\infty < n < \infty\}$ ($i = 1, 2, \dots, M$). We consider an interference channel defined as follows:

$$\hat{A}_{i,n} = \sum_{j=1}^M c_{ij}(d_{ij}A_{j,n-k_{ij}-1} + e_{ij}A_{j,n-k_{ij}}) + B_{i,n}, \quad (1)$$

where $C = \{c_{ij}, i, j = 1, 2, \dots, M\}$, $D = \{d_{ij}, i, j = 1, 2, \dots, M\}$, $E = \{e_{ij}, i, j = 1, 2, \dots, M\}$ are unknown matrices and $d_{ij} \geq 0$, $e_{ij} \geq 0$, $d_{ij}^2 + e_{ij}^2 \leq 1$, $d_{ii} = 0$, $e_{ii} = 1$ ($i, j = 1, 2, \dots, M$). The sequences $\{B_{i,n}, -\infty < n < \infty\}$ are strictly stationary and ergodic with zero mean and variance σ_i^2 ($i = 1, 2, \dots, M$).

Example 1. Spread spectrum communication

Consider a popular spread spectrum communication scheme ([6]–[9]). The system includes amplitude modulators, adder channel with delays and attenuations, and coherent demodulators (Fig. 2). The output of the i -th encoder modulates in amplitude the carrier $\sqrt{2W_i} \cos(\Omega t - v_i)\varphi(t)$, where W_i and v_i stand for the power and phase-ambiguity parameter, respectively. Here

$$\varphi(t) = \sum_{n=-\infty}^{\infty} \psi(t - nT_c),$$

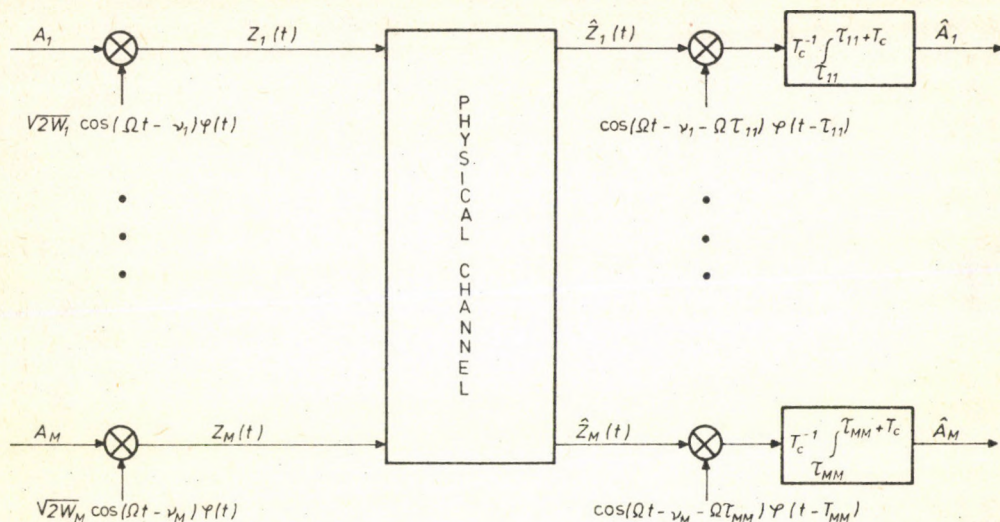


Fig. 2. The spread spectrum scheme

where $\psi(t) \geq 0$ is a time-limited signal, limited to $[0, T_c]$ and $T_c^{-1} \int_0^{T_c} \psi(t)^2 dt = 1$. ψ is called chip waveform and T_c^{-1} stands for the chip rate.

The output of the i -th modulator:

$$Z_i(t) = \sqrt{2W_i} \cos(\Omega t - \nu_i) \sum_{n=-\infty}^{\infty} A_{i,n} \psi(t - nT_c).$$

The input of the i -th demodulator is a weighted sum of Z_j 's ($j = 1, 2, \dots, M$) with delay and a zero mean, second order, strictly stationary and ergodic process $N_i(t)$ with covariance function $K_i(t)$:

$$\hat{Z}_i(t) = \sum_{j=1}^M \sqrt{V_{ij}} Z_j(t - \tau_{ij}) + N_i(t), \quad (2)$$

where $V = \{V_{ij}\}$ and $\tau = \{\tau_{ij}\}$ denote the power attenuation and delay matrix, respectively.

The correlator detectors contain multipliers and integrators, averaging on length T_c . One of the two inputs of the multiplier is the signal $\hat{Z}_i(t)$ and the other is the synchronized replica of the spreading signal of modulator i : $\cos(\Omega t - \nu_i - \Omega \tau_{ii}) \psi(t - \tau_{ii})$.

Because we can ignore the double frequency component of the input signal of each integrator, the output of the i -th demodulator has the form (1), where

$$c_{ij} = \sqrt{\frac{V_{ij} W_j}{2}} \cos(\Omega \tau_{ij} - \nu_j - (\Omega \tau_{ii} + \nu_i)),$$

$$d_{ij} = \frac{1}{T_c} \int_0^{f_{ij}} \psi(t)\psi(t + T_c - f_{ij})dt,$$

$$e_{ij} = \frac{1}{T_c} \int_0^{T_c - f_{ij}} \psi(t)\psi(t + f_{ij})dt,$$

$$f_{ij} = (\tau_{ij} - \tau_{ii}) \bmod T_c,$$

$$k_{ij} = \text{integer part of } \left(\frac{\tau_{ij} - \tau_{ii}}{T_c} \right).$$

Let us define the sequence $\{B_{i,n}\}$ by

$$B_{i,n} \triangleq T_c^{-1} \int_0^{T_c} N_i(t + \tau_{ii} + nT_c) \cos(\Omega(t + nT_c) - v_i) \psi(t) dt.$$

Because of the properties of N_i , the sequence $\{B_{i,n}\}$ is stationary and ergodic with zero mean and variance $\sigma_i^2 \triangleq E(B_{i,n}^2)$,

$$\sigma_i^2 = T_c^{-2} \int_0^{T_c} \int_0^{T_c} K_i(t-s) \cos(\Omega(t + nT_c) - v_i) \cos(\Omega(s + nT_c) - v_i) \psi(t)\psi(s) dt ds$$

($i = 1, 2, \dots, M$).

Example 2. M-user binary adder channel

If $c_{i,j} = 1$, $d_{i,j} = 0$, $e_{i,j} = 1$, $k_{i,j} = 0$ ($i, j = 1, 2, \dots, M$) (no attenuation and chip synchronism between senders), then according to (1) we get:

$$\hat{A}_{i,n} = \sum_{j=1}^M A_{j,n} + B_{i,n} \quad (3)$$

which is the well-known adder channel model of the multiple access information theory ([1]–[5]), assuming $B_{i,n} = B_n$. In these results the capacity regions are characterized if there is a common decoder knowing the codebook of each encoder. In the noiseless case ($B_{i,n} \equiv 0$) it is shown that the maximum achievable sum rate of the encoders, $C_{\text{sum}}(M)$, is asymptotically equal to $1/2 \log_2(2\pi eM)$ as M increases ([4]–[5]).

The capacity region is still open for asynchronous encoders and autonomic decoders.

3. One bit to block coding

Considering the case of one bit to block coding, where the encoded information bits of the source i is a_i taking the values $+1, -1$ ($i=1, \dots, M$). Using the random coding argument let ζ_i be the randomly chosen codeword of the i -th encoder ($i=1, \dots, M$) being independent, N -dimensional random vectors with independent, zero mean, $+1, -1$ valued coordinates. The i -th encoder transmits $\zeta_i = (\zeta_{i,1}, \dots, \zeta_{i,N})$ if the a_i source bit is 1 and $-\zeta_i$ otherwise. If there is code word synchronism between the encoders then the output of the channel is $\sum_{j=1}^M \zeta_j a_j$. The i -th decoder decodes $+1$ if the inner product $(\zeta_i, \sum_{j=1}^M \zeta_j a_j)$ is nonnegative and decodes -1 otherwise.

Then the error probability

$$\begin{aligned} P_e &= P((\zeta_i, \zeta_i + \sum_{j \neq i} \zeta_j a_j) < 0) = \\ &= E(P(N < - \sum_{j \neq i} (\zeta_i, \zeta_j) a_j | \zeta_i, a_1, \dots, a_M)) \\ &= E(P(N < - \sum_{j \neq i} \sum_{l=1}^N \zeta_{il} \cdot \zeta_{jl} a_j | \zeta_i, a_1, \dots, a_M)). \end{aligned}$$

Applying the Bernstein-Chernoff bound, let $\lambda > 0$ be arbitrary, then by independence

$$\begin{aligned} P_e &\leq \frac{E(E(\exp(-\lambda \sum_{j \neq i} \sum_{l=1}^N \zeta_{il} \cdot \zeta_{jl} a_j) | \zeta_i, a_1, \dots, a_M))}{\exp(\lambda N)} \\ &= \frac{E\left(\prod_{j \neq i} \prod_{l=1}^N E(\exp(-\lambda \zeta_{il} \cdot \zeta_{jl} a_j) | \zeta_i, a_1, \dots, a_M)\right)}{\exp(\lambda N)} \\ &= E \prod_{j \neq i} \prod_{l=1}^N \cosh(-\lambda \zeta_{il} a_j) \exp(-\lambda N) \\ &\leq E \prod_{j \neq i} \prod_{l=1}^N \exp(\lambda^2 \zeta_{il}^2 a_j^2 / 2) \exp(-\lambda N) \\ &= \exp(\lambda^2 N(M-1)/2 - \lambda N). \end{aligned}$$

For $\lambda = 1/(M-1)$ we get that

$$P_e \leq \exp(-N/(2(M-1))) \leq \exp(-1/(2R_{\text{sum}})), \quad (4)$$

where $R_{\text{sum}} = M/N$ stands for the sum of the equal rates $1/N$. In the sequel we show a block code which can work at all rates under $1/(2 \ln 2)$. Inequality (4) guarantees an error probability less than 10^{-6} if $R \leq 0.036$, which is far from the limit $1/(2 \ln 2)$.

4. Block to block coding

In this section block coding is considered for the general channel model (1). Each encoder maps the binary L vectors into binary blocks of length N . To use the random coding method the code books of users are supposed to be independent and contain 2^L independent, uniformly distributed code words. One code word of codebook i is denoted by the sequence $A_{i,n}$ which is an i.i.d. sequence of $+1$, -1 valued random variables.

Without loss of generality the decoding error probability of the communication from the 1-st encoder to the 1-st decoder is examined. The random code book of the 1-st encoder is denoted by ξ_1, \dots, ξ_{2^L} . The message j ($j \in \{1, 2, \dots, 2^L\}$) is sent by the block ξ_j , and the received block is denoted by $\alpha + \xi'_j$, where $\xi'_j = c_{11}\xi_j$, $\alpha = (\alpha_1, \dots, \alpha_N)$, $\alpha_n = \hat{A}_{1,n} - c_{11}A_{1,n}$ and α, ξ'_j are independent.

Correlation decoding is used, i.e. decide j^* if j^* maximizes the inner product $(\alpha + \xi'_j, \xi_l)$ over l . Therefore the block error probability, given a message j , can be calculated as follows:

$$\begin{aligned} P_{e,j} &\triangleq P(j^* \neq j | j) \leq P\left(\bigcup_{l \neq j} \{(\alpha + \xi'_j, \xi_l) \geq (\alpha + \xi'_j, \xi_j)\} | j\right) \leq \\ &\leq P((\alpha + \xi'_j, \xi_j) < 0 | j) + \\ &+ P((\alpha + \xi'_j, \xi_j) \geq 0, \bigcup_{l \neq j} \{(\alpha + \xi'_j, \xi_l) \geq (\alpha + \xi'_j, \xi_j)\} | j). \end{aligned} \quad (5)$$

Let $k \neq j$ be fixed and introduce the notation

$$[x]_0^1 = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

Then by (5) we get

$$\begin{aligned} P_{e,j} &\leq P((\alpha + \xi'_j, \xi_j) < 0 | j) + \\ &+ E\left(P\left(\bigcup_{l \neq j} \{(\alpha + \xi'_j, \xi_l) \geq |(\alpha + \xi'_j, \xi_j)|\} \mid \xi_j, \alpha, j\right) | j\right) \\ &\leq P((\alpha + \xi'_j, \xi_j) < 0 | j) + \\ &+ E\left([2^L P((\alpha + \xi'_j, \xi_k) \geq |(\alpha + \xi'_j, \xi_j)|) | \xi_j, \alpha, j]_0^1 | j\right). \end{aligned} \quad (6)$$

Let us apply the Bernstein–Chernoff bound as in Section 3, then

$$\begin{aligned} P((\alpha + \xi'_j, \xi_j) < 0 | j) &= E(P(-(\alpha, \xi_j) > Nc_{11} | \alpha, j) | j) \leq \\ &\leq E \exp(1/2\lambda^2 \|\alpha\|^2 - \lambda Nc_{11}) |_{\lambda = \frac{Nc_{11}}{\|\alpha\|^2}} = E \exp\left(-\frac{N^2 c_{11}^2}{2\|\alpha\|^2}\right), \end{aligned} \quad (7)$$

where $\|\cdot\|$ stands for the Euclidean norm.

In the same way

$$\begin{aligned} P((\alpha + \xi'_j, \xi_k) \geq |(\alpha + \xi'_j, \xi_j)| / j, \xi_j, \alpha) &\leq \\ &\leq \exp\left(-\frac{|(\alpha + \xi'_j, \xi_j)|^2}{2\|\alpha + \xi'_j\|^2}\right). \end{aligned} \quad (8)$$

In our block code the sum rate is

$$R_{\text{sum}} = \frac{ML}{N},$$

therefore (6), (7), and (8) imply that

$$P_{e,j} \leq E \exp\left(-L \frac{1}{2R_{\text{sum}}\eta_L}\right) + E \left[\exp\left(-L \left(\frac{1}{2R_{\text{sum}}\tilde{\eta}_L} - \ln 2\right)\right) \right]_0^1, \quad (9)$$

where

$$\eta_L = \frac{\|\alpha\|^2}{NM c_{11}^2} \quad (10)$$

and

$$\tilde{\eta}_L = \frac{N\|\alpha + \xi'_j\|^2}{M|(\alpha + \xi'_j, \xi_j)|^2}. \quad (11)$$

Let us introduce the notation

$$Q_j = \frac{1}{M c_{jj}^2} \left(\sum_{i=1}^M c_{ji}^2 (d_{ji}^2 + e_{ji}^2) + \sigma_j^2 \right) \quad j=1, 2, \dots, M \quad (12)$$

then by the strong law of large numbers relation (1) implies that

$$\lim_{L \rightarrow \infty} \eta_L = Q_1 - 1/M \quad \text{a.s.} \quad (13)$$

and

$$\lim_{L \rightarrow \infty} \tilde{\eta}_L = Q_1 \quad \text{a.s.} \quad (14)$$

Therefore (9), (13), and (14) give our

Theorem. If M and R_{sum} is fixed and

$$R_{\text{sum}} < \frac{1}{2 \ln 2} \cdot \frac{1}{\max_{1 \leq j \leq M} Q_j}, \quad (15)$$

then $P_{e,j}$ tends to zero as the block length L tends to infinity.

Let us examine the consequences of the Theorem. According to our Example 1 in Section 2,

$$e_{ij}^2 + d_{ij}^2 \leq 1$$

and it is equal to 1 in the case of chip synchronism, therefore if the encoder chips are asynchronous, then by (12) and (15) we get a larger rate region than in the synchronous case.

Put $V_{ij}W_j = W$, $\delta_i = \delta$ and $\tau_{ij} = \tau$ for each i and j . Then

$$\max_j Q_j = \max_j \left\{ \frac{1}{M} \sum_{l=1}^M \cos^2(v_j - v_l) + \frac{2\sigma^2}{WM} \right\}.$$

If there is no noise and the modulators are synchronous, as in Example 2,

$$\max_j Q_j = 1.$$

therefore the error probability tends to zero if

$$R_{\text{sum}} < \frac{1}{2 \ln 2}$$

as we said in Section 3.

If the modulators are asynchronous, then it is easy to show that

$$\frac{1}{2} \leq \max_j \frac{1}{M} \sum_{l=1}^M \cos^2(v_j - v_l) \leq 1$$

for arbitrary M , and if v_1, \dots, v_M are independent and uniformly distributed on $[0, 2\pi]$ then

$$\max_j \frac{1}{M} \sum_{l=1}^M \cos^2(v_j - v_l) \xrightarrow{M} \frac{1}{2} \quad \text{a.s.},$$

therefore if

$$R_{\text{sum}} < \frac{1}{\ln 2},$$

then $P_{e,j} \xrightarrow{L} 0$ if M is sufficiently large.

It is not surprising that from inequality (9) an exponential error bound can be derived if there is no noise or the noise is white Gaussian. One has to mention, however, that the decoding rule is not maximum likelihood. This bound is approximately equal to

$$\exp\left(-L\left(\frac{1}{2R_{\text{sum}}Q_i} - \ln 2\right)\right).$$

5. The code word synchronization

Throughout the paper we have assumed the code word synchronism between the encoder i and decoder i . In what follows the simple "sliding" correlator technique ([8]) is generalized to 2^L code words for searching synchronism. As it was done in Section 4, without loss of generality the synchronization between encoder 1 and decoder 1 is considered.

At decoder 1 we have 2^L "sliding" correlators, whose j -th correlates the instant segment of N bit length of the input of decoder 1 ($\hat{A}_{1,n}$) with the code word ξ_j . The maximal value from the 2^L outputs of correlators is denoted by ρ_L , which is compared to a threshold d . If the threshold is reached, the recognition of synchronism is decided, otherwise a new step is made "sliding" one bit further on the sequence $\hat{A}_{1,n}$.

Let H_τ denote the hypotheses that at a fixed step the shift between the code word frames of the 1-st encoder and the 1-st decoder is τ , where $\tau \in \{1, 2, \dots, N-1\}$. Let H_0 denote the hypothesis that these code words are in synchronism ($\tau=0$).

The false alarm and false dismissal error probabilities can be defined at a fixed step as follows

$$\begin{aligned} P_{f_a} &\triangleq P(\rho_L \geq d | H_\tau) \\ P_{f_d} &\triangleq P(\rho_L < d | H_0), \end{aligned} \quad \tau \neq 0, \quad (16)$$

where we choose $d = (1 - \varepsilon)c_{11}N$ for threshold value, $0 < \varepsilon < 1$. The reason for choosing this is that the maximal value of ρ_L equals to $c_{11}N$ in the one user case ($M=1$), and this value is taken when there is synchronism, i.e. $\tau=0$. The suitable value of ε follows from the required error probabilities.

In the sequel we give upper bounds on the error probabilities P_{f_a} and P_{f_d} .

Under the hypothesis H_0 , ρ_L has the form $\rho_L = \max_j (\alpha + c_{11}\xi_j, \xi_j)$, if message I was transmitted. Thus

$$\begin{aligned} P\{\max_j (\alpha + c_{11}\xi_j, \xi_j) < (1 - \varepsilon)c_{11}N | I = i\} &\leq \\ &\leq P\{(\alpha + c_{11}\xi_i, \xi_i) < (1 - \varepsilon)c_{11}N\} = \\ &= P\{-(\alpha, \xi_i) > \varepsilon c_{11}N\} \leq E \exp\left(-\frac{\varepsilon^2 c_{11}^2 N^2}{2\|\alpha\|^2}\right), \end{aligned} \quad (17)$$

so $P_{f_a} \xrightarrow{L \rightarrow \infty} 0$.

If $0 < \tau < N$, then $\xi^{(\tau)}$ consists of parts of two code words as follows. If the consecutive messages I and J are encoded by code words

$$\xi_I = \{\xi_{I,1}, \xi_{I,2}, \dots, \xi_{I,N}\}$$

and

$$\xi_J = \{\xi_{J,1}, \xi_{J,2}, \dots, \xi_{J,N}\}$$

then

$$\xi^{(\tau)} = \{\xi_{I,N-\tau+1}, \xi_{I,N-\tau+2}, \dots, \xi_{I,N}, \xi_{J,1}, \dots, \xi_{J,N-\tau}\}.$$

Given $I = i_1, J = i_2, i_1 \neq i_2$, the coordinates of $\xi^{(\tau)}$ are independent. If $i_1 = i_2$ then $\xi^{(\tau)}$ is a rotation of the code word ξ_{i_1} . Thus under H_τ

$$\begin{aligned} P(\max_j (\alpha + c_{11} \xi^{(\tau)}, \xi_j) \geq (1-\varepsilon)c_{11}N | I = i_1, J = i_2) & \quad (18) \\ & \leq P((\alpha + c_{11} \xi^{(\tau)}, \xi_{i_1}) \geq (1-\varepsilon)c_{11}N | I = i_1, J = i_2) + \\ & \quad + P((\alpha + c_{11} \xi^{(\tau)}, \xi_{i_2}) \geq (1-\varepsilon)c_{11}N | I = i_1, J = i_2) + \\ & \quad + E([2^L P((\alpha + c_{11} \xi^{(\tau)}, \xi_k) \geq (1-\varepsilon)c_{11}N | I = i_1, J = i_2, \alpha)]_0^1 | I = i_1, J = i_2), \end{aligned}$$

where $k \neq i_1, k \neq i_2$. For the third term of the right hand side of (18), $\alpha + c_{11} \xi^{(\tau)}$ and ξ_k are independent and applying the same technique as in Section 4, we get that

$$\begin{aligned} E([2^L P((\alpha + c_{11} \xi^{(\tau)}, \xi_k) \geq (1-\varepsilon)c_{11}N | I = i_1, J = i_2, \alpha)]_0^1 | I = i_1, J = i_2) & \leq \\ & \leq E \left[2^L \exp \left(-L \frac{(1-\varepsilon)^2 c_{11}^2 N M}{2R_{\text{sum}}(\|\alpha\|^2 + Nc_{11}^2)} \right) \right]_0^1 \quad (19) \end{aligned}$$

which tends to zero if

$$R_{\text{sum}} < \frac{1}{2 \ln 2} \frac{(1-\varepsilon)^2}{Q_1}. \quad (20)$$

For the evaluation of the first and second terms of the right hand side of (18) we apply the following elementary facts on random sequences: let $\eta_1, \eta_2, \dots, \eta_n, \dots$ be i.i.d. ± 1 valued random variables with $E\eta_i = 0$, then vectors $(\eta_1, \eta_2, \dots, \eta_n)$ and $(\eta_1 \cdot \eta_{1+\tau}, \eta_2 \cdot \eta_{2+\tau}, \dots, \eta_n \cdot \eta_{n+\tau})$ have the same distribution, therefore for each $\delta > 0$

$$P\left(\sum_{i=1}^N \eta_i \cdot \eta_{i+\tau} \geq \delta\right) = P\left(\sum_{i=1}^N \eta_i \geq \delta\right) \quad (21)$$

and

$$\begin{aligned}
 & P\left(\sum_{i=1}^N \eta_i \eta_{i+\tau \bmod N} \geq \delta\right) = \\
 & = P\left(\sum_{i=1}^{\tau} \eta_i \eta_{i+N-\tau} + \sum_{i=1}^{N-\tau} \eta_i \eta_{i+\tau} \geq \delta\right) \leq \\
 & \leq P\left(\sum_{i=1}^{\tau} \eta_i \eta_{i+N-\tau} \geq \frac{\delta}{2}\right) + \\
 & + P\left(\sum_{i=1}^{N-\tau} \eta_i \eta_{i+\tau} \geq \frac{\delta}{2}\right) = \\
 & = P\left(\sum_{i=1}^{\tau} \eta_i \geq \frac{\delta}{2}\right) + P\left(\sum_{i=1}^{N-\tau} \eta_i \geq \frac{\delta}{2}\right).
 \end{aligned} \tag{22}$$

If $i_1 \neq i_2$, then by (21) we get that

$$\begin{aligned}
 & P((\alpha + c_{11}\xi^{(\tau)}, \xi_{i_1}) \geq (1-\varepsilon)c_{11}N | I=i_1, J=i_2) \leq \\
 & \leq P((\alpha, \xi_{i_1}) \geq (1-\varepsilon)c_{11}N/2) + \\
 & + P(c_{11}\xi^{(\tau)}, \xi_{i_1}) \geq (1-\varepsilon)c_{11}N/2 | I=i_1, J=i_2) \leq \\
 & \leq E \exp\left(-\frac{(1-\varepsilon)^2 c_{11} N^2}{8\|\alpha\|^2}\right) + \exp\left(-\frac{(1-\varepsilon)^2 N}{8}\right),
 \end{aligned} \tag{24}$$

which tend to zero if $L \rightarrow \infty$. If $i_1 = i_2$ then by (22)

$$\begin{aligned}
 & P((\alpha + c_{11}\xi^{(\tau)}, \xi_{i_1}) \geq (1-\varepsilon)c_{11}N | I=i_1, J=i_2) \leq \\
 & \leq E \exp\left(-\frac{(1-\varepsilon)^2 c_{11}^2 N^2}{8\|\alpha\|^2}\right) + \\
 & + \exp\left(-\frac{(1-\varepsilon)^2 N^2}{32\tau}\right) + \exp\left(-\frac{(1-\varepsilon)^2 c_{11}^2 N^2}{32(N-\tau)}\right) + \\
 & \leq E \exp\left(-\frac{(1-\varepsilon)^2 c_{11}^2 N^2}{8\|\alpha\|^2}\right) + 2 \exp\left(-\frac{(1-\varepsilon)^2 N}{32}\right).
 \end{aligned} \tag{25}$$

Summarizing the above results, the error probabilities P_{f_a} , P_{f_d} tend to zero as the code block length tends to infinity if $\varepsilon > 0$ and

$$R_{\text{sum}} < \frac{1}{2 \ln 2} \max_j Q_j \frac{(1-\varepsilon)^2}{j}. \tag{26}$$

The parameter ε was arbitrary, therefore the synchronization procedure may work with arbitrarily small error probabilities in the same rate region as that of the decoding rule for synchronized case.

Conclusion

A rate region of equal code rates is given for an interference channel, when the asynchronous encoders use block coding and the autonomous decoders working on the correlation decoding rule know the codebook of the corresponding encoder only.

This rate region depends on the channel parameters (attenuations, delays, etc.), however, the decoders do not know these parameters.

If there is code word synchronism between encoders and decoders communicating with each other, then the block decoding error probability is evaluated. If there is no code word synchronism, then a rule is introduced for synchronization and in the same rate region, as before, the synchronization error probability tends to zero as the block length tends to infinity.

The result can easily be applied analysing direct sequence spread spectrum systems from the point of view of achievable sum rates.

References

1. Ahlswede, R., "Multi-way communication channels." in Proc. 2nd Int. Symp. Inform. Theory, Tsahkadsor. Armenian S. S. R., 1971, pp. 23-52. Publishing House of the Hungarian Academy of Sciences, 1973.
2. Ahlswede, R., "The capacity region of a channel with two senders and two receivers", Ann. Prob., vol. 2, pp. 805-814, Oct. 1974.
3. van der Meulen, E. C., "A survey of multi-way channels in information theory: 1961-1976", IEEE Trans. Inform. Theory, vol. IT-23, pp. 1-37, Jan. 1977.
4. Wolf, J. K., "Multi-user communication networks", in Communication Systems and Random Process Theory, J. K. Skwirzynsky, Ed., Alphen aan den Rijn: The Netherlands, 1979, pp. 37-53.
5. Chang, S., Weldon, E. J. Jr., "Coding for T -user multiple access channels", IEEE Trans. Inform. Theory, vol. IT-25, pp. 684-691, Nov. 1979.
6. Pursley, M. B., "Performance evaluation for phase-coded spread-spectrum multiple-access communication-Part I: System analysis", IEEE Trans. Comm., vol. COM-25, pp. 795-799, 1977.
7. Yao, K., "Error probability of asynchronous spread-spectrum multiple-access communications systems", IEEE Trans. Comm., vol. COM-25, pp. 803-809, 1977.
8. Dixon, R. C., "Spread Spectrum Systems", J. Wiley and Sons, Ch. 6, 1976.
9. Cooper, G. R., Nettleton, R. W., "A spread-spectrum technique for high-capacity mobile communications", IEEE Trans. Vehic. Techn., vol. VT-27, pp. 264-275, 1977.
10. Sato, H., "Two-user communication channels", IEEE Trans. Inform. Theory, vol. IT-23, no. 3, p. 295, May, 1977.
11. Carleial, A. B., "Interference channels", IEEE Trans. Inform. Theory, vol. IT-24, no. 1, p. 60, Jan., 1978.
12. Carleial, A. B., "A case where interference does not reduce capacity", IEEE Trans. Inform. Theory, vol. IT-21, no. 5, p. 569, Sept., 1975.

13. Sato, H., "On the capacity region of a discrete two-user channel for strong interference", IEEE Trans. Inform. Theory, vol. IT-24, no. 3, 377, May, 1978.
14. Sato, H., Tanabe, M., "A discrete two user channel with strong interference", Trans. IECE Japan E61, no. 11, p. 880, Nov., 1978.
15. Györfi, L., Kerekes, I., "A Block Code for Noiseless Asynchronous Multiple Access OR Channel", IEEE Trans. Inform. Theory, vol. IT-27, no. 6, p. 788, Nov., 1981.

**Блочное кодирование и корреляционное декодирование
для взвешенного суммирующего канала с M пользователями**

Л. ДЕРФИ, И. ВАЙДА

(Будапешт)

В работе рассматриваются характеристики блочных кодов для канала с M пользователями, в котором выходной сигнал является взвешенной суммой сдвинутых по времени входных сигналов и аддитивного шума. Предполагается, что в системе используются отдельные корреляционные декодеры, каждый из которых не знает кодовые книги других пользователей, и эти декодеры универсальны в том смысле, что они не используют фактические параметры канала.

L. Györfi

I. Vajda

Technical University of Budapest

H-1111 Budapest, Stoczek u. 2.

SINAI'S THEOREM AND ENTROPY COMPRESSION

Š. ŠUJAN
(Bratislava)

(Received October 20, 1982)

A strengthening of Sinai's theorem of ergodic theory is established and used to derive a source coding theorem in the spirit of Gray-Neuhoff-Ornstein alternate approach.

1. Introduction

Ornstein's isomorphism theory [6] is based on approximation techniques involving a strong concept of closeness of two stationary processes (formalized through Ornstein's \bar{d} -distance) and its relations to the usual weak concept of closeness describing closeness in distribution and in entropy. The strong closeness implies the weak one but not vice versa. The converse is known as the property of being finitely determined. The important fact is that finitely determined processes are precisely Bernoulli processes; that is, stationary codings of i.i.d. processes (by an i.i.d. process we mean a sequence of independent and identically distributed random variables).

A central role in Ornstein's theory is played by a strong version of Sinai's theorem [8], which establishes the possibility of passing from weak approximations to an i.i.d. process to strong approximations.

Our aim is to prove a new type of strong Sinai's theorem involving also an average distortion constraint. Two observations suggest its formulation and the way to its proof. First of all, the topological approach to source coding due to Gray, Neuhoff, and Ornstein [2] can be used to find processes close in the weak sense to a given i.i.d. process. Secondly, for reasonable fidelity criteria, the average distortion is continuous in the strong sense. Hence, Ornstein's strong form of Sinai's theorem should do the rest.

Unfortunately, a good deal of technical notions and results is needed which do not seem to be common to information theorists. Thus we include a sketchy survey of basic concepts in the next section. A systematic account may be found in [10].

2. Ornstein's Fundamental Lemma

Let (Ω, \mathcal{F}, m) be a "nice" probability space (e.g., a Lebesgue space [6] or, a standard Borel space). Let $T: \Omega \rightarrow \Omega$ be an automorphism (=invertible measure-

preserving transformation). We shall work with finite ordered measurable partitions of Ω . If $P=(P^1, \dots, P^K)$, we put

$$d(P)=(m(P^1), \dots, m(P^K));$$

$$d(P|F)=(m(P^k \cap F)/m(F); \quad 1 \leq k \leq K)$$

if $F \in \mathcal{F}$ and $m(F) > 0$. The weak partition distance between two ordered partitions $P=(P^1, \dots, P^K), Q=(Q^1, \dots, Q^K)$ is defined to be the number

$$|d(P)-d(Q)| = \sum_{k=1}^K |m(P^k)-m(Q^k)|$$

(of course, this definition extends easily to partitions on different probability spaces), and the strong partition distance is

$$|P-Q| = \sum_{k=1}^K m(P^k \Delta Q^k),$$

where $E \Delta F$ stands for the symmetric difference. If $P=(P^1, \dots, P^K)$ and $Q=(Q^1, \dots, Q^J)$ are two ordered finite partitions, we denote by $P \wedge Q$ the partition whose atoms are the intersections $P^k \cap Q^j$ ordered lexicographically, say. We put

$$P(T, n) = \bigvee_{i=0}^{n-1} T^{-i}P, \quad P(T, 0) = P, \quad (1)$$

and

$$(P)_T = \bigvee_{i \in Z} T^i P = \sigma \left(\bigcup_{i \in Z} T^i P \right), \quad (2)$$

where $Z = \{ \dots, -1, 0, 1, \dots \}$. A measurable partition P of Ω is said to be a generator if $(P)_T = \mathcal{F} \bmod m$ (i.e., if $(P)_T$ and \mathcal{F} give rise to isomorphic measure algebras). A partition P is said to be independent if for each $n \geq 1$, $T^n P$ and $P(T, n)$ are independent. Since $(P)_T$ is a sub- σ -field of \mathcal{F} invariant under each T and T^{-1} , it is but a factorfield as defined in [7]. The action of T on the probability space $(\Omega, (P)_T, m)$ is said to be a factor of T .

Let $\mathcal{I}(T) = \{ F \in \mathcal{F} : T^{-1}F = F \}$. A factor of T determined by a partition P is said to be ergodic if the σ -field $(P)_T \cap \mathcal{I}(T)$ is trivial mod m . If this is true for a generator P then T itself is called ergodic.

A pair (T, P) , where T is an automorphism and P is a finite measurable partition is said to be a process. The process (T, P) is called ergodic if the corresponding factor is.

An automorphism T is called Bernoulli if it admits an independent generator P . The corresponding process (T, P) is said to be Bernoulli, too. Note that this means that by redefining the state space so that the atoms of P will serve as new states we shall get an i.i.d. process. If T is Bernoulli and P is any finite measurable partition then the corresponding factor is Bernoulli [6].

Finally, let us introduce the notion of entropy. If $P=(P^1, \dots, P^K)$, we put $H(P) = -\sum m(P^k) \log m(P^k)$. The entropy of the process (T, P) is defined as the limit

$$h(T, P) = \lim_{n \rightarrow \infty} n^{-1} H(P(T, n)).$$

The entropy of T is the supremum

$$h(T) = \sup h(T, P)$$

taken over all finite measurable partitions P of Ω . Recall from [1] that $h(T) = h(T, P)$ when P is a generator, and that $h(T, P) = H(P)$ when P is independent. The strong form of Sinai's theorem called Ornstein's Fundamental Lemma by Shields [7] reads as follows:

Lemma (Ornstein's Fundamental Lemma). Let \bar{T} be a Bernoulli automorphism with an independent generator \bar{P} . For each $\varepsilon > 0$ there is a $\delta > 0$ such that for any ergodic automorphism T (possibly defined on another probability space) with $h(T) \geq h(\bar{T})$ and for any partition P with $\text{card}(P) = \text{card}(\bar{P})$, the conditions

- (a) $|d(P) - d(\bar{P})| < \delta$, and
- (b) $|h(T, P) - h(\bar{T}, \bar{P})| < \delta$

imply the existence of a partition Q such that

- (c) Q is independent,
- (d) $d(Q) = d(\bar{P})$, and
- (e) $|Q - P| \leq \varepsilon$.

3. Strong Sinai's Theorem

The notions and results of the preceding section easily transform to the setup of discrete stationary information sources [10]. We shall illustrate the "transformation rules" several times below.

Let $A = \{a_1, \dots, a_K\}$ and $\hat{A} = \{b_1, \dots, b_J\}$ be two finite sets (called the source alphabet and the reproduction alphabet, respectively). As usual in information theory, a process $X = (X_i; i \in Z)$ is understood to be the sequence of one-dimensional projections $X_i: A^Z \rightarrow A$:

$$X_i(x) = x_i \quad \text{for } x \in A^Z \quad \text{and } i \in Z.$$

This corresponds to taking $\Omega = A^Z$ and $P = P(A)$, the natural "zero-time" partition of A^Z :

$$P(A) = (\{x \in A^Z: x_0 = a_k\}; 1 \leq k \leq K) = (P^k; 1 \leq k \leq K). \tag{3}$$

Further put

$$P(\hat{A}) = (\{y \in \hat{A}^Z : y_0 = b_j\}; 1 \leq j \leq J) = (\hat{P}^j; 1 \leq j \leq J), \tag{4}$$

$$R(A) = (\{(x, y) : x_0 = a_k\}; 1 \leq k \leq K) = (C^k; 1 \leq k \leq K), \tag{5}$$

$$R(\hat{A}) = (\{(x, y) : y_0 = b_j\}; 1 \leq j \leq J) = (D^j; 1 \leq j \leq J). \tag{6}$$

The latter two partitions partition the joint source-reproduction space $A^Z \times \hat{A}^Z$ (identified as usually with $(A \times \hat{A})^Z$). Let $T_A, T_{\hat{A}}$, and $T_{A \times \hat{A}}$ denote the corresponding shifts, and use μ and ν as generic symbols for distributions of processes X on A^Z and Y on \hat{A}^Z . We suppose that X and Y are stationary (i.e., $\mu = \mu T_A^{-1}, \nu = \nu T_{\hat{A}}^{-1}$), ergodic (i.e., $\mathcal{I}(T_A) = \{\emptyset, A^Z\} \text{ mod } \mu, \mathcal{I}(T_{\hat{A}}) = \{\emptyset, \hat{A}^Z\} \text{ mod } \nu$), and aperiodic (i.e., $\mu\{x\} = \nu\{y\} = 0$ for all $x \in A^Z$ and $y \in \hat{A}^Z$). Since X can be identified with $(T_A, P(A))$, the entropy rate of X is well defined by $h(X) = h(T_A, P(A))$. In what follows we shall equivalently use notations $h(X)$ and $h(\mu)$, where $\mu = \text{dist } X$. This convention will concern also other quantities (like average distortion and $\bar{\rho}$ -distance defined below).

We assume that $\rho : A \times \hat{A} \rightarrow [0, \infty)$ has the property that for any $a \in A$ there is a unique $b \in \hat{A}$ with $\rho(a, b) = 0$. Let $\rho_{\max} = \max \{\rho(a, b) : a \in A, b \in \hat{A}\}$. We use ρ to define a single-letter fidelity criterion. If $x^n = (x_0, \dots, x_{n-1}) \in A^n$ and $y^n = (y_0, \dots, y_{n-1}) \in \hat{A}^n$, we put

$$\rho_n(x^n, y^n) = n^{-1} \sum_{i=0}^{n-1} \rho(x_i, y_i). \tag{7}$$

Any measurable map $\bar{f} : A^Z \rightarrow \hat{A}^Z$ such that $\bar{f} \circ T_A = T_{\hat{A}} \circ \bar{f}$ is called a stationary code. If \bar{f} is used to code the process X (in which case it suffices that \bar{f} be defined and stationary only for μ -almost all $x \in A^Z; \mu = \text{dist } X$) then its rate is defined to be the entropy $h(\bar{f} X)$ of the encoded process, and the average distortion is the number

$$\rho(\bar{f}) = \int \rho(x_0, (\bar{f}x)_0) \mu(dx). \tag{8}$$

Factors and stationary codings can be identified. Indeed, any stationary code \bar{f} gives rise to a measurable map $f : A^Z \rightarrow \hat{A}, f(x) = (\bar{f}x)_0$, and f in turn gives rise to a finite measurable partition

$$Q_f = (f^{-1}\{b_j\}; 1 \leq j \leq J). \tag{9}$$

Then $(Q_f)_{T_A}$ (see (2)) is the corresponding factor-field. Conversely, let $Q = (Q^j; 1 \leq j \leq J)$ be any measurable partition of A^Z into J atoms. Then the formulae

$$f_Q(x) = b_j \quad \text{if } x \in Q^j, \quad 1 \leq j \leq J \tag{10}$$

define a stationary code $\bar{f}_Q : A^Z \rightarrow \hat{A}^Z$, viz.

$$(\bar{f}_Q(x))_i = f_Q(T_A^i x), \quad x \in A^Z, \quad i \in Z. \tag{11}$$

Given X, Y , and $n \geq 1$ let $X^n \vee Y^n$ (or, $\mu^n \vee \nu^n$, where $\mu = \text{dist } X, \nu = \text{dist } Y$) denote the set of all joint probability vectors p on $A^n \times \hat{A}^n$ with marginals $\text{dist } X^n$ and $\text{dist } Y^n$. Observe that $\text{dist } X^n = d(P(A)(T_A, n))$; see (1). Let

$$\begin{aligned} \bar{\rho}(X, Y) &= \limsup_{n \rightarrow \infty} \bar{\rho}_n(X, Y) = \\ &= \limsup_{n \rightarrow \infty} \inf_{p \in X^n \vee Y^n} \int \rho_n(x^n, y^n) dp. \end{aligned} \tag{12}$$

$\bar{\rho}$ was introduced in [3] as a generalization of \bar{d} -distance. If X and Y are jointly ergodic (which is the case, for example, if Y is a stationary coding of X and X is ergodic) then

$$\bar{\rho}(X, Y) = \inf_{p \in X \vee Y} \int \rho(x_0, y_0) dp. \tag{13}$$

Here is our main result:

Theorem 1. Suppose X, Y , and ρ are given as above. Let Y be an i.i.d. process (i.e., $(Y_i; i \in \mathbb{Z})$ is a sequence of independent and identically distributed \hat{A} -valued random variables). Let $h(X) \geq h(Y)$. For any $\gamma > 0$ there exists a stationary code $\bar{f}: A^{\mathbb{Z}} \rightarrow \hat{A}^{\mathbb{Z}}$ such that

- (a) $\bar{f}X = Y$, and
- (b) $\int \rho(x_0, (fx)_0) d\mu \leq \bar{\rho}(X, Y) + \gamma; \quad \mu = \text{dist } X$.

4. The Proof

The idea of the proof is as follows. First of all we use the construction from [2, proof of Lemma 1] in order to get a code \bar{f} such that the conditions (a) and (b) of Lemma above are satisfied. Then using that lemma we get a coding such that the encoded process is exactly Y (thereby proving part (a) of our theorem). Finally, assertion (e) of the Lemma is used to bound the average distortion as in part (b) of the theorem.

In order to make the paper reasonably selfcontained we reproduce a part of the quoted proof from [2]. But first some necessary notions: If T is an automorphism of (Ω, \mathcal{F}, m) and if R is a finite measurable partition of Ω , then the quadruple (T, n, F, R) is said to be an α -gadget if the sets $F, TF, \dots, T^{n-1}F$ are pairwise disjoint, $m(\cup T^i F) \geq 1 - \alpha$, and $d(R(T, n)|F) = d(R(T, n))$; α -gadgets exist by the strong Rokhlin's lemma [7, p. 22]. Two α -gadgets (T, n, F, R) and $(\bar{T}, n, \bar{F}, \bar{R})$ are said to be isomorphic; in symbols, $(T, n, F, R) \sim (\bar{T}, n, \bar{F}, \bar{R})$, if

$$d(R(T, n)|F) = d(\bar{R}(\bar{T}, n)|\bar{F}).$$

Let $v = \text{dist } Y$. The first two steps show that for any $\delta > 0$ we can find a code $\bar{f}: A^Z \rightarrow \hat{A}^Z$ such that

$$h(\bar{f}X) < h(v) + \delta; \quad (14)$$

$$\rho(\bar{f}) < \bar{\rho}(\mu, v) + \delta. \quad (15)$$

By (10), it suffices to find an appropriate partition Q of A^Z into J atoms.

Step 1. Let A, \hat{A}, ρ, μ, v , and $\delta > 0$ be given. Let $p \in \mu \vee v$ approximately yield $\bar{\rho}(\mu, v)$, i.e., let

$$\int \rho(x_0, y_0) dp \leq \bar{\rho}(\mu, v) + \delta/3.$$

We can imagine p as the distribution of a pair process. Then the average distortion between the marginal processes can be expressed as

$$\rho(R(A), R(\hat{A})) = \int \rho(x_0, y_0) dp = \sum_{k=1}^K \sum_{j=1}^J \rho(a_k, b_j) p(C^k \cap D^j) \quad (16)$$

(see (5), (6)). Furthermore,

$$h(v) = h(T_{A \times \hat{A}}, R(\hat{A})) = \lim_{n \rightarrow \infty} n^{-1} H(R(\hat{A})(T_{A \times \hat{A}}, n))$$

(see (1)). Given $0 < \delta < e^{-1}$ choose $l \geq 1$ so large that

$$|l^{-1} H(R(\hat{A})(T_{A \times \hat{A}}, l)) - h(v)| \leq \delta/3,$$

and pick $\alpha > 0$ so small and n so large that

$$KJ \alpha \rho_{\max} \leq \delta/3, \quad J^l (\alpha + (l-1)/n)^{1/2} \leq \delta/3. \quad (17)$$

Step 2. Since the pair process corresponding to p must be aperiodic, we can use strong Rokhlin's lemma and find an α -gadget $(T_{A \times \hat{A}}, n, \tilde{F}, R(A) \vee R(\hat{A}))$; in particular

$$d(R(A) \vee R(\hat{A})(T_{A \times \hat{A}}, n) | \tilde{F}) = d(R(A) \vee R(\hat{A})(T_{A \times \hat{A}}, n)).$$

This shows also that

$$d(R(A)(T_{A \times \hat{A}}, n) | \tilde{F}) = d(R(A)(T_{A \times \hat{A}}, n)). \quad (19)$$

Further, let $(T_A, n, F, P(A))$ denote an α -gadget for the process X . By the definition of $\bar{\rho}$, p induces $\mu = \text{dist } X$ on A^Z so that

$$d(P(A)(T_A, n)) = d(R(A)(T_{A \times \hat{A}}, n)), \quad (20)$$

$$d(R(A)(T_{A \times \hat{A}}, n) | \tilde{F}) = d(P(A)(T_A, n) | F). \quad (21)$$

It follows from (18) through (21) that

$$(T_{A \times \hat{A}}, n, \tilde{F}, R(A)) \sim (T_A, n, F, P(A)). \quad (22)$$

A non-atomic probability space can be partitioned so that an arbitrary given finite probability vector obtains as the distribution of that partition [7, Lemma 4.3]. Combining this with (22) we find a finite partition $\tilde{Q}=(\tilde{Q}^1, \dots, \tilde{Q}^J)$ of the union

$$\bigcup_0^{n-1} T_A^i F \text{ such that}$$

$$(T_A, n, F, \tilde{Q}) \sim (T_{A \times \hat{A}}, n, \tilde{F}, R(\hat{A})). \tag{23}$$

Let $Q=(Q^1, \dots, Q^J)$ be any extension of \tilde{Q} to a partition of A^Z (i.e., we add parts of the rest of A^Z to the atoms of \tilde{Q} in an arbitrary manner). Define \bar{f} by means of Q as in (10) and (11). A technical result of [2] shows that $h(\bar{f} X) \leq h(v) + 2\delta/3$ and $\rho(\bar{f}) \leq \bar{\rho}(\mu, \nu) + 2\delta/3$ so that the claimed inequalities (14) and (15) are valid.

Step 3. Now we shall show that Q can be chosen so that (14), (15), and the inequality

$$|d(Q) - d(P(\hat{A}))| \leq \delta \tag{24}$$

are valid. Given a $\delta > 0$ observe that if the reasoning in steps 1 and 2 is valid for some n and α then it is valid also for that n and any $\alpha' < \alpha$. Hence fix n and pick α so small that the inequalities (17) take place together with

$$J^{2n}\alpha \leq \delta. \tag{25}$$

Any extension of \tilde{Q} to A^Z must satisfy

$$|d(Q(T_A, n)|F) - d(Q(T_A, n))| \leq J^n\alpha, \tag{26}$$

for we can add at most a set of measure α to the atoms of \tilde{Q} , and there are at most J^n atoms of $\tilde{Q}(T_A, n)$. Let $(T_A, n, G, P(\hat{A}))$ be an α -gadget for the reproduction process. Then

$$\begin{aligned} & |d(Q(T_A, n)) - d(P(\hat{A})(T_{\hat{A}}, n))| \leq \\ & \leq |d(Q(T_A, n)) - d(Q(T_A, n)|F)| + \\ & + |d(Q(T_A, n)|F) - d(R(\hat{A})(T_{A \times \hat{A}}, n)|\tilde{F})| + \\ & + |d(R(\hat{A})(T_{A \times \hat{A}}, n)|\tilde{F}) - d(P(\hat{A})(T_{\hat{A}}, n)|G)| + \\ & + |d(P(\hat{A})(T_{\hat{A}}, n)|G) - d(P(\hat{A})(T_{\hat{A}}, n))| \leq \\ & \leq J^n\alpha = J^{2n}\alpha/J^n \leq \delta/J^n. \end{aligned}$$

But then

$$\begin{aligned} |d(Q) - d(P(\hat{A}))| & \leq J^{n-1} |d(Q(T_A, n) - d(P(\hat{A})(T_{\hat{A}}, n))| \leq \\ & \leq J^{n-1}\delta/J^n < \delta, \end{aligned}$$

so that (23) is valid, too.

Step 4. Since δ has been chosen arbitrarily, we use Ornstein's Fundamental Lemma to conclude that for any $\varepsilon > 0$, if Q is chosen as in Step 3 for δ corresponding to that ε , then there exists a partition $\bar{Q} = (\bar{Q}^j; 1 \leq j \leq J)$ of A^Z such that

$$\bar{Q} \text{ is independent,} \quad (27)$$

$$d(\bar{Q}) = d(P(\hat{A})), \quad (28)$$

$$|\bar{Q} - Q| \leq \varepsilon. \quad (29)$$

If $\bar{f} = \bar{f}_{\bar{Q}}$ (cf. (10) and (11)) then $\bar{f}X$ is an i.i.d. process by (27). Since $\text{dist}((\bar{f}X)_0) = \text{dist} Y_0$ (this follows from (28) and the fact that $P(\hat{A})$ partitions \hat{A}^Z according to the zero-time output of the process Y), $\bar{f}X = Y$, for Y is also i.i.d. This proves assertion (a) of Theorem 1.

Step 5. We shall prove that the function

$$Q \mapsto \rho(P(A), Q) = \sum_{k,j} \rho(a_k, b_j) \mu(P^k \cap Q^j)$$

is continuous in the strong partition distance. We claim that

$$|\rho(P(A), Q) - \rho(P(A), \bar{Q})| \leq |Q - \bar{Q}| KJ \rho_{\max}. \quad (30)$$

Clearly

$$\begin{aligned} |\rho(P(A), Q) - \rho(P(A), \bar{Q})| &\leq \\ &\leq KJ \rho_{\max} \max_{k,j} |\mu(P^k \cap \bar{Q}^j) - \mu(P^k \cap Q^j)|. \end{aligned}$$

Since $|\mu(E) - \mu(F)| \leq \mu(E \Delta F)$, we see that

$$\begin{aligned} |\mu(P^k \cap \bar{Q}^j) - \mu(P^k \cap Q^j)| &\leq \\ &\leq \mu[(P^k \cap \bar{Q}^j) \Delta (P^k \cap Q^j)] \leq \mu(\bar{Q}^j \Delta Q^j). \end{aligned}$$

Since this is true for all k and all j , (30) follows from the definition of the strong partition distance. Now

$$\begin{aligned} \rho(\bar{f}) &= \sum_{k,j} \rho(a_k, b_j) \mu(P^k \cap \bar{Q}^j) \leq \\ &\leq \sum_{k,j} \rho(a_k, b_j) \mu(P^k \cap Q^j) + \\ &+ \sum_{k,j} \rho(a_k, b_j) |\mu(P^k \cap \bar{Q}^j) - \mu(P^k \cap Q^j)| \leq \\ &\leq \bar{\rho}(X, Y) + \delta + |Q - \bar{Q}| KJ \rho_{\max} \leq \bar{\rho}(X, Y) + \delta + \varepsilon KJ \rho_{\max} \end{aligned}$$

(the latter inequality follows from (29)). Now, if $\epsilon \rightarrow 0$ then also $\delta \rightarrow 0$ so that the up-right expression can be made smaller than $\bar{\rho}(X, Y) + \gamma$ using ϵ , and hence δ , sufficiently small. This completes the proof of Theorem 1.

5. A Source Coding Problem

We assume that A, \hat{A}, ρ, X , and Y are above. For any rate constraint $R \in [0, h(X)]$ define the optimum performance theoretically attainable (OPTA) using stationary codes to be the number

$$\delta(R, \mu) = \inf \{ \rho(\bar{f}) : h(\bar{f}X) \leq R \} . \tag{31}$$

A combination of results from [2] and [5] shows that the OPTA using block or sliding-block codes of rates at most R coincides with $\delta(R, \mu)$ for all ergodic processes. Gray et al. [2] developed a topological approach to source coding problems, and their coding theorem relates the OPTA with the process form $\bar{\rho}$ -distance (cf. (13)):

$$\delta(R, \mu) = \rho(R, \mu) = \inf \{ \bar{\rho}(X, Y) : h(Y) \leq R \} . \tag{32}$$

Let

$$\delta_0(R, \mu) = \inf \{ \rho(\bar{f}) : h(\bar{f}X) \leq R, \bar{f}X \text{ i.i.d.} \} , \tag{33}$$

$$\rho_0(R, \mu) = \inf \{ \bar{\rho}(X, Y) : h(Y) \leq R, Y \text{ i.i.d.} \} . \tag{34}$$

Theorem 2. Let A, \hat{A}, ρ , and X be as above, and let $\mu = \text{dist } X$. For any $0 \leq R \leq h(X)$ it holds that

$$\delta_0(R, \mu) = \rho_0(R, \mu) .$$

Proof. (1) $\rho_0(R, \mu) \leq \delta_0(R, \mu)$. Let \bar{f} be a stationary code such that $\rho(\bar{f}) \leq \delta_0(R, \mu) + \epsilon$. Then

$$\bar{\rho}(X, \bar{f}X) \leq \int \rho(x_0, (\bar{f}x)_0) d\mu = \rho(\bar{f}) \leq \delta_0(R, \mu) + \epsilon .$$

Hence, for each $\epsilon > 0$ there exists an i.i.d. process Y over alphabet \hat{A} such that $h(Y) \leq R$ and $\bar{\rho}(X, Y) \leq \delta_0(R, \mu) + \epsilon$ (simply put $Y = \bar{f}X$). The claimed inequality follows from (34) using the fact ϵ has been chosen arbitrarily.

(2) $\delta_0(R, \mu) \leq \rho_0(R, \mu)$. Choose an i.i.d. process Y over alphabet \hat{A} such that $h(Y) \leq R$ and $\bar{\rho}(X, Y) \leq \rho_0(R, \mu) + \epsilon/2$. Since $h(Y) \leq R \leq h(X)$, Theorem 1 applies. We use it with $\gamma = \epsilon/2$. This yields a stationary code $\bar{f} : A^Z \rightarrow \hat{A}^Z$ such that $\bar{f}X = Y$ and

$$\rho(\bar{f}) = \int \rho(x_0, (\bar{f}x)_0) d\mu \leq \bar{\rho}(X, Y) + \epsilon/2 \leq \rho_0(R, \mu) + \epsilon .$$

Since Y and ϵ have been chosen arbitrarily, the proof is complete.

The optimum codes from Theorem 2 have the advantage of performing simultaneously entropy compression and redundancy removal, the encoded process

being already i.i.d. (cf. [10] for a detailed discussion on overall source coding operation). On the other hand, the solution is only partial, for we do not know whether $\delta_0(R, \mu) = \delta(R, \mu)$. We conjecture that this is so when X is also an i.i.d. process.

Another open problem is to generalize the results to other classes of processes. Formally, this amounts to replacing the i.i.d. condition in (33) and (34) by the requirement that $\bar{f}X$ and Y are members of some specified class of processes.

If X is also an i.i.d. process and $h(X) > h(Y) > 0$, then the coding $\bar{f}X = Y$ can be carried over by means of a finitary code \bar{f} with finite expected code length [4]. In this special case it is possible by a slight modification of the coding technique from [4] to show that Theorem 1 is valid also for such codes [9].

References

1. Billingsley, P., Ergodic Theory and Information. J. Wiley, New York, 1965.
2. Gray, R. M., Neuhoff, D. L., Ornstein, D. S., Non-block source coding with a fidelity criterion. Ann. Prob., **3** (1975), 478–491.
3. Gray, R. M., Neuhoff, D. L., Shields, P. C., A generalization of Ornstein's \bar{I} -distance with applications to information theory. Ann. Prob., **3** (1975), 315–328.
4. Keane, M., Smorodinsky, M., A class of finitary codes. Israel J. Math., **26** (1977), 352–371.
5. Kieffer, J. C., Extensions of source coding theorems for block codes to sliding-block codes. IEEE Trans. Inform. Theory, **IT-26** (1980), 679–692.
6. Ornstein, D. S., Ergodic Theory, Randomness, and Dynamical Systems. Yale Univ. Press, New Haven–London, 1974.
7. Shields, P. C., The Theory of Bernoulli Shifts. Univ. of Chicago Press, Chicago, 1973.
8. Sinai, Ya. G., On weak isomorphism of transformations with an invariant measure (in Russian). Mat. Sborn., **63** (1964), 23–42.
9. Šujan, Š., Codes in ergodic theory and information: Some examples. In: Ergodic Theory and Related Topics (H. Michel, ed.), Mathematical Research, Vol. **12**, Akademie-Verlag, Berlin 1982, 181–184.
10. Šujan, Š., Ergodic theory, entropy, and coding problems of information theory (to appear as supplement to Kybernetika, **19** (1983)).

Теорема Синяя и энтропийное сжатие

Ш. ШУЙЯН

(Братислава)

Устанавливается усиление теоремы Синяя в эргодической теории. Оно используется для доказательства теоремы кодирования источников в духе подхода Грея–Нейхофи–Орнштейна.

Š. Šujan

Institute of Measurement and Measuring Technique,
Electro-Physical Research Center,
Slovak Academy of Sciences
842 19 Bratislava
Czechoslovakia

DUALITY IN PROGRAMMING UNDER PROBABILISTIC CONSTRAINTS WITH A RANDOM TECHNOLOGY MATRIX

DINH THE LUC
(Budapest)

(Received January 12, 1983)

The dual program of a nonlinear programming problem under probabilistic constraints with a random technology matrix is studied and it is proved that under some adequate conditions the primal program is normal, in this way the primal and dual programs are equivalent.

1. Introduction

Consider the following programming problem

$$\begin{aligned} \min \langle c, x \rangle \\ Dx \geq b \\ P(Ax \leq \beta) \geq p \end{aligned} \tag{P}$$

where c is a constant n -component vector, b is a constant m -vector, D is an $m \times n$ -matrix, x is an n -component variable, A is an $m' \times m$ -matrix and β is an m' -component vector. It is supposed that the entries of A as well as the components of β are random variables.

This is a stochastic programming model of A. Prékopa which has several practical applications (see [5]). In general this program is not convex because of the probabilistic constraint $P(Ax \leq \beta) \geq p$. However, under some additional conditions on A , β and p the set of points satisfying the probabilistic constraint will be convex. The crucial results used in proving the convexity are the theorems on logarithmic concave measures due to A. Prékopa. We refer the reader to [3] and [4] about these topics. In this note we shall study the dual program of Program (P) using the general duality theory developed in [6] and [7], and prove its normality.

2. The dual program

Let w be a vector of R^n , we define the function $\varphi(x, w)$ by the following:

$$\varphi(x, w) = \begin{cases} \langle c, x \rangle & \text{if } Dx - b \geq w \\ & \text{and } P(Ax \leq \beta) \geq p \\ +\infty & \text{otherwise.} \end{cases}$$

Hypothesis 1. The set of all vectors x satisfying $P(Ax \leq \beta) \geq p$ is convex.

Lemma 1. Under hypothesis 1 if program (P) has its feasible set nonempty, i.e. there exists a vector x such that $Dx \geq b$ and $P(Ax \leq \beta) \geq p$, then φ is a closed proper convex function on $R^n \times R^m$.

Proof. Since the feasible set of (P) is nonempty, function φ is not identically equal to $+\infty$ and by definition φ has nowhere the value $-\infty$. This shows that φ is proper. In order to establish the convexity of φ , it suffices to prove that if $\varphi(x_i, w_i) = \langle c, x_i \rangle$, where x_i and w_i satisfy the following:

$$Dx_i - b \geq w_i \quad \text{and} \quad P(Ax_i \leq \beta) \geq p, \quad i = 1, 2,$$

then $\varphi(\lambda x_1 + (1-\lambda)x_2, \lambda w_1 + (1-\lambda)w_2) \leq \lambda \varphi(x_1, w_1) + (1-\lambda)\varphi(x_2, w_2)$ for every $\lambda \in [0, 1]$. Indeed in virtue of hypothesis 1 we have

$$P(A(\lambda x_1 + (1-\lambda)x_2) \leq \beta) \geq p$$

and of course

$$D(\lambda x_1 + (1-\lambda)x_2) - b \geq \lambda w_1 + (1-\lambda)w_2.$$

According to the definition of φ we obtain

$$\begin{aligned} \varphi(\lambda x_1 + (1-\lambda)x_2, \lambda w_1 + (1-\lambda)w_2) &= \langle c, \lambda x_1 + (1-\lambda)x_2 \rangle = \\ &= \lambda \varphi(x_1, w_1) + (1-\lambda)\varphi(x_2, w_2). \end{aligned}$$

For the closedness of φ we note that if sequences $\{x_i\}$ and $\{w_i\}$ converge to x_0 and w_0 respectively, then

$$\lim_{i \rightarrow \infty} \varphi(x_i, w_i) = \begin{cases} \langle c, x_0 \rangle & \text{if } Dx_0 - b \geq w_0 \text{ and } P(Ax_0 \leq \beta) \geq p \\ \geq \infty & \text{otherwise.} \end{cases}$$

This means that $\lim_{i \rightarrow \infty} \varphi(x_i, w_i) \geq \varphi(x_0, w_0)$ and the lemma is proved.

Now we can define the conjugate function of φ as follows

$$\begin{aligned}\varphi^*(\xi, \lambda) &= \sup_{x, w} [\langle \xi, x \rangle + \langle \lambda, w \rangle - \varphi(x, w)] \\ &= \sup_{\substack{x, w \\ Dx - b \geq w \\ P(Ax \leq \beta) \geq p}} [\langle \xi, x \rangle + \langle \lambda, w \rangle - \langle c, x \rangle] \\ &= \begin{cases} \sup_{(x: P(Ax \leq \beta) \geq p)} [\langle \xi, x \rangle + \langle \lambda, Dx - b \rangle - \langle c, x \rangle] & \text{if } \lambda \geq 0 \\ +\infty & \text{otherwise.} \end{cases}\end{aligned}$$

Setting in particular $\xi=0$ we obtain

$$\varphi^*(0, \lambda) = \begin{cases} -\langle \lambda, b \rangle + \sup_{(x: P(Ax \leq \beta) \geq p)} \langle D'\lambda - c, x \rangle & \text{if } \lambda \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Finally we get the dual program:

$$\max \left\{ \langle \lambda, b \rangle - \sup_{x: P(Ax \leq \beta) \geq p} \langle D'\lambda - c, x \rangle \right\}, \quad \lambda \geq 0. \quad (\mathbf{P}^*)$$

3. The normality

First we recall some definitions (see [6]).

The perturbation function of program (P) is defined as follows

$$\Phi(w) = \inf_x \varphi(x, w) = \inf_{\substack{(x: Dx - b \geq w \\ P(Ax \leq \beta) \geq p}} \langle c, x \rangle.$$

Program (P) is said to be normal if the perturbation function Φ is closed at $w=0$.

Before proving the normality of program (P) we need some topological property of the set of vector x satisfying $Dx - b \geq w$ and $P(Ax \leq \beta) \geq p$. Let A_1, A_2, \dots be a sequence of nonempty subsets in R^n . Set A_0 is called the limit of this sequence (write $A_0 = \lim A_i$) if for any positive ε there exists an integer N such that for $i \geq N$ one has

$$A_i \subseteq A_0 + \varepsilon B(0, 1) \quad \text{and} \quad A_0 \subseteq A_i + \varepsilon B(0, 1)$$

where $B(0, 1)$ is the unit ball in R^n .

Let w_i be a sequence of vectors in R^n . Denote

$$D_i = \{x \in R^n: Dx - b \geq w_i\}$$

$$D_0 = \{x \in R^n: Dx - b \geq 0\}.$$

Lemma 2. If the sequence w_i converges to 0, then $\lim_i D_i = D_0$.

Proof. Since the inclusion $D_0 \subseteq D_i + \varepsilon B(0, 1)$ is obvious for any fixed positive ε and $i \geq N$, N is large enough, it is sufficient to prove that $D_i \subseteq D_0 + \varepsilon B(0, 1)$ for $i \geq N$. Suppose

to the contrary that the above inclusion is not true, that is there exists a positive ε_0 and a sequence n_i such that D_{n_i} does not belong to $D_0 + \varepsilon_0 B(0, 1)$. Hence for every n_i an element x_{n_i} may be found in D_{n_i} so that

$$[x_{n_i} + \varepsilon_0 B(0, 1)] \cap D_0 = \emptyset.$$

or in other words, the following inequality holds for no $y \in B(0, 1)$:

$$D(x_{n_i} + \varepsilon_0 y) - b \geq 0. \quad (1)$$

Without loss of generality we can suppose that for every n_i the first inequality in (1) does not hold, i.e.

$$\langle d_1, x_{n_i} + \varepsilon_0 y \rangle - b_1 < 0 \quad (2)$$

where d_1 is the first row of matrix D and b_1 is the first component of vector b . Since x_{n_i} is in D_{n_i} we have $\langle d_1, x_{n_i} \rangle - b_1 \geq w_{n_i}^1$ for every n_i , where $w_{n_i}^1$ is the first component of w_{n_i} . Set $y=0$, then the following relation will hold:

$$0 > \langle d_1, x_{n_i} \rangle - b_1 > w_{n_i}^1.$$

As w_i converges to 0 we have

$$\lim \langle d_1, x_{n_i} \rangle = b_1. \quad (3)$$

Suppose that x_{n_i} has a subsequence which we denote by the same x_{n_i} for convenience, such that $\langle d_1, x_{n_i} \rangle \neq 0$ for every i . In this case the restriction of function $\langle d_1, \cdot \rangle$ on $B(0, 1)$ attains its maximum at some point $y \in B(0, 1)$. Let $\delta = \langle d_1, y \rangle$, then $\delta > 0$. From (3) it follows that

$$|\langle d_1, x_{n_i} \rangle - b_1| < \varepsilon_0 \cdot \delta/3$$

for n_i sufficiently large. Consequently

$$\begin{aligned} \langle d_1, x_{n_i} + \varepsilon_0 y \rangle - b_1 &= \langle d_1, x_{n_i} \rangle - b_1 + \varepsilon_0 \langle d_1, y \rangle = \\ &= \langle d_1, x_{n_i} \rangle - b_1 + \varepsilon_0 \delta > \varepsilon_0 \delta/3 > 0. \end{aligned} \quad (4)$$

It is obvious that (4) contradicts (2) when n_i is sufficiently large. In this way we have only to consider the case when $\langle d_1, x_{n_i} \rangle = 0$ for all i , say more exactly for all i greater than some integer. But in this case from (3) it follows that $b_1 = 0$ and (2) is impossible. The impossibility finishes the proof of the lemma.

Lemma 3. Let M be a closed convex subset of R^n . Assume that $D_0 \cap M$ is nonempty and compact. Then for every $\varepsilon > 0$ there exists a number N such that

$$D_i \cap M \subseteq D_0 \cap M + \varepsilon B(0, 1) \quad \text{for all } i \geq N.$$

Proof. Let ε be an arbitrary positive. In virtue of Lemma 2 there is a number N_1 such that

$$D_i \cap M \subseteq [D_0 + \varepsilon_0 B(0, 1)] \cap M \quad \text{for all } i \geq N_1.$$

In order to prove Lemma 3 it suffices to show that there exists a number k_0 which depends on ε such that the following relation holds

$$[D_0 + \varepsilon B(0, 1)] \cap M \subseteq D_0 \cap M + k_0 \varepsilon B(0, 1). \tag{5}$$

It is obvious that (5) is equivalent to the following

$$D_{-\varepsilon} \cap M \subseteq D_0 \cap M + k_0 \varepsilon B(0, 1) \tag{6}$$

where $D_{-\varepsilon} = \{x \in R^n: Dx - b \geq -(\varepsilon, \dots, \varepsilon)\}$. Suppose the opposite that (6) does not hold, i.e. there is a positive ε_0 such that $D_{-\varepsilon_0} \cap M$ does not belong to $D_0 \cap M + k\varepsilon_0 B(0, 1)$ with $k = 1, 2, \dots$. In other words we can find a sequence $\{x_k\}$ of vectors in $D_{-\varepsilon_0} \cap M$ so that the distance $\rho(x_k, D_0 \cap M)$ from x_k to $D_0 \cap M$ is greater than $k\varepsilon_0$. We first note that the norm $\|x_k\|$ converges to ∞ as well as $\rho(x_k, D_0 \cap M)$ converges to ∞ when k runs to ∞ . Without loss of generality we may assume that $\{x_k/\|x_k\|\}$ converges to some \bar{x} . Taking $x^* \in D_0 \cap M$ we see that $x^* + \lambda \bar{x} \in M$ for every $\lambda \geq 0$ (by the convexity of M). In virtue of the Representation Theorem of polyhedres (see [1]), we conclude that $x^* + \lambda \bar{x}$ belongs to D_0 for all $\lambda \geq 0$. Consequently, $x^* + \lambda \bar{x}, \lambda \geq 0$, is a ray in $D_0 \cap M$. This contradicts the compactness of $D_0 \cap M$. The proof is completed.

Hypothesis 2. The feasible set of program (P) is nonempty and compact.

Theorem 1. Under hypotheses 1 and 2 program (P) is normal.

Proof. To prove the normality of program (P) we have to show that $\Phi(w)$ is closed at $w=0$, or equivalently for any sequence $\{w_i\}$ of vectors in R^n converging to 0 the following inequality holds

$$\lim \Phi(w_i) \geq \Phi(0). \tag{7}$$

First we observe that under hypothesis 2, program (P) has an optimal solution, say x_0 , and $\Phi(0) = \langle c, x_0 \rangle$. If D_i has an empty intersection with $M = \{x \in R^n: P(Ax \leq \beta) \geq p\}$ then by definition $\Phi(w_i) = +\infty$. Hence (7) holds trivially. By this we may assume that $D_i \cap M$ is nonempty for all i . And suppose to the contrary that (7) does not hold. Then for some positive ε_0 we could choose a sequence $\{w_{n_i}\}$ converging to 0 and an integer N so that

$$\Phi(w_{n_i}) < \Phi(0) - \varepsilon_0 \quad \text{for } i \geq N. \tag{8}$$

Let x_i be an element of $D_i \cap M$ satisfying

$$\langle c, x_i \rangle - \Phi(w_{n_i}) < \varepsilon_0 / 3(\|c\| + 1) \quad \text{if } \Phi(w_{n_i}) \text{ is finite} \tag{9}$$

$$\text{and } \langle c, x_i \rangle < \Phi(0) - \varepsilon_0 \quad \text{if } \Phi(w_{n_i}) \text{ is } -\infty. \tag{10}$$

Using Lemma 3 we obtain

$$D_i \cap M \subseteq D_0 \cap M + (\varepsilon_0/3(\|c\| + 1)) \cdot B(0, 1)$$

for every i greater than some integer N' .

Consequently there are elements \bar{x}_i in $D_0 \cap M$ so that

$$\|x_i - \bar{x}_i\| \leq \varepsilon_0/3(\|c\| + 1). \quad (11)$$

Take $i \geq \max(N, N')$ to have

$$\begin{aligned} \langle \bar{x}_i, c \rangle &= \langle c, \bar{x}_i - x_i \rangle + \langle c, x_i \rangle \\ \langle \|c\| \cdot \|\bar{x}_i - x_i\| + \langle c, x_i \rangle &< \varepsilon_0/3 + \langle c, x_i \rangle. \end{aligned} \quad (12)$$

Combine (8), (9), (10) and (12) to arrive at the contradiction: $\langle \bar{x}_i, c \rangle < \Phi(0) - \varepsilon_0/3$ while $\bar{x}_i \in D_0 \cap M$ and $\Phi(0) = \inf \langle c, x \rangle$.

Corollary 1. Under hypotheses 1 and 2 the values of programs (P) and (P*) are equal.

Proof. It is known that for a closed proper convex function $\varphi(x, w)$ the following statements are equivalent [6]

- i) $\Phi(0) = \inf_x \varphi(x, 0) = \sup_\lambda -\varphi^*(0, \lambda)$
- ii) The perturbation function Φ is closed at $w=0$.

The corollary is now derived directly from these statements and from Theorem 1.

Theorem 2. Suppose that $M \cap D_0$ is not bounded. Under hypothesis 1 if the functional $\langle c, \cdot \rangle$ is not constant on every direction of recession of $D_0 \cap M$, then program (P) is normal.

Proof. In order to prove this theorem we have to verify inequality (7) in the proof of the previous theorem. If $\Phi(0)$ is equal to $-\infty$, (7) holds trivially. Now suppose $\Phi(0)$ is finite and the opposite that

$$\Phi(w_{n_i}) < \Phi(0) - \varepsilon_0 \quad \text{for } i \geq N$$

where w_{n_i} is some sequence converging to 0, ε_0 is some positive. For every $i \geq N$ we may find x_i in $D_i \cap M$ so that

$$\langle c, x_i \rangle < \Phi(0) - \varepsilon_0/3. \quad (13)$$

Without loss of generality we may assume that $\{x_i/\|x_i\|\}$ converges to \bar{x} . If $\{\|x_i\|\}$ is bounded then it has a limit point, say x_0 . Clearly x_0 is in $D_0 \cap M$ hence (13) gives the following contradiction

$$\langle c, x_0 \rangle \leq \Phi(0) - \varepsilon_0/3.$$

We have now to consider only the case when $\{\|x_i\|\}$ converges to ∞ . In this case \bar{x} is a direction of recession of M . By lemma 1, \bar{x} is a direction of recession of D_0 and consequently of D_i also. From inequality (13) it follows that $\langle c, \bar{x} \rangle \leq 0$ and the assumption on functional $\langle c, \cdot \rangle$ we have $\langle c, \bar{x} \rangle < 0$. This contradicts the finiteness of $\Phi(0)$ by taking values of $\langle c, \cdot \rangle$ on the ray which starts at some point in $D_0 \cap M$ and runs in direction \bar{x} . This completes the proof.

Corollary 2. Let $D_0 = P + Q$ where P is a polytope and Q is a cone. If $\langle c, q \rangle \neq 0$ for every nonzero q in Q , then program (P) is normal.

This corollary is derived immediately from Theorem 2.

We finish this paper by some remarks on Lemma 3. The assumptions on M and its intersection with D_0 are important. Without the convexity of M the lemma is obviously not true. It is more difficult to see that this problem occurs if the intersection of M with D_0 is not compact. For this case, consider the following example.

Let

$$M = \{(x, y, z) \in R^3: y \geq 1 \text{ and } z \geq x^2/y\}$$

$$D_0 = \{(x, y, z) \in R^3: y \geq 1, x \geq -1 \text{ and } z \leq 0\}.$$

It is obvious that M is convex and

$$M \cap D_0 = \{(x, y, z) \in R^3: x = 0, y \geq 1 \text{ and } z = 0\}$$

is not compact. Lemma 3 is not valid. Consider the following program

$$\begin{aligned} & \min \langle c, (x, y, z) \rangle \\ & (x, y, z) \in D_0 \cap M \\ & c = (1, 0, 0). \end{aligned} \tag{P'}$$

This program has optimal solutions and the optimal value is equal to 0. Its perturbation program will be

$$\begin{aligned} & \min \langle c, (x, y, z) \rangle \\ & (x, y, z) \in D_\varepsilon \cap M \end{aligned}$$

where $D_\varepsilon = \{(x, y, z) \in R^3: y \geq 1, x \geq -1 - \varepsilon \text{ and } z \leq \varepsilon\}$ with $\varepsilon > 0$. Obviously the optimal value of this program is $-1 - \varepsilon$. Thus program (P') is not normal.

Further, if in addition to the assumptions made in Lemma 3 we require D_0 and the relative interior of M meet each other then $D_0 \cap M = \lim_i D_i \cap M$. This is an immediate corollary of Lemma 3 and Robinson's results about stability for systems of inequalities (see [8]).

Finally, it is also worth noting that we have studied only the normality of problem (P). For its stability we refer the reader to [8] and [9], where this topic has

been successfully developed, in particular, under the constraint qualification, i.e., there is $x \in M$ with $Dx - b > 0$, one has

$$\inf_x \Phi(x, 0) = \max_{\lambda} -\Phi^*(0, \lambda)$$

ensuring the solvability of the dual program.

Acknowledgements

The author is indebted to professor A. Prékopa who first suggested studying of such a problem, and to an anonymous referee for helpful comments.

References

1. Prékopa, A., *Lineáris programozás (Linear programming)*, Budapest, 1968.
2. Prékopa, A., Programming under probabilistic constraints with a random technology matrix, *Math. Operationsforsch. u. Statist.* **5** (1974), Heft 2, S. 109–116.
3. Prékopa, A., Logarithmic concave measures with application to stochastic programming, *Acta Sci. Math., Szeged* **32** (1971), 301–316.
4. Prékopa, A., Logarithmic concave measures and related topics, in *Stochastic programming* (edited by M. A. H. Dempster), London, 1974.
5. Prékopa, A., *Studies in Applied stochastic programming*, Tanulmányok 80/1978, MTA SZTAKI, Budapest.
6. Avriel, M., *Nonlinear programming*, Prentice-Hall, INC.
7. Rockafellar, R. T., *Convex analysis*, Princeton, New York, 1970.
8. Robinson, S. M., Stability theory for systems of inequalities. Part 1: Linear systems, *SIAM J. Numer. Anal.*, vol. **12**, pp. 754–769, 1975.
9. Bazaraa, M. S., A theorem of the alternative with application to convex programming: Optimality, duality and stability, *J. Math. Anal. Appl.*, vol. **41**, pp. 701–715, 1973.

Двойственность в задаче программирования при вероятностных ограничениях со случайной матрицей технологии

ДИНЬ ТХЕ ЛУК
(Будапешт)

В статье рассмотрена двойственная задача для задачи типа

$$\min \langle c, x \rangle$$

$$Dx \geq b$$

$$P(Ax \leq \beta) \geq p,$$

где A — случайная матрица и β — случайный вектор. Доказано, что под некоторыми гипотезами исходная задача нормальна и таким образом установлена эквивалентность между двойственной и исходной задачами. Наконец, дается один пример, показывающий, что нормальность не будет иметь место, если множество точек, удовлетворяющих ограничениям, не компактно.

Dinh The Luc
Computer and Automation Institute
Hungarian Academy of Sciences
Kende u. 13-17
H-1502 Budapest
Hungary

PRINTED IN HUNGARY

Akadémiai Nyomda, Budapest

ELŐKÖZMŐ
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

ANALYSIS OF FLOWS IN CLOSED EXPONENTIAL QUEUEING NETWORKS

V. M. VISHNEVSKII, A. I. GERASIMOV

(Moscow)

An arbitrarily structured closed exponential network with one customer class is considered. An algorithm of determining the distribution function of intervals between customers leaving the nodes is proposed. For a closed cyclic exponential queueing network with an arbitrary number of nodes, the distribution, mean and variance of times between customers leaving the nodes are obtained.

Introduction

In analytical modeling of computing networks by mathematical methods of stochastic (queueing) network theory permits calculating such important characteristics as throughput and customer delay in the network; analyzing the effect of different flow control methods and to make comparing protocol analysis; choosing buffer storage size of a message (packet) switching node; calculating multiprogrammed computing systems, etc. In [1, 2] a method of polynomial approximation for analysis of closed, open and mixed networks of queues of an arbitrary structure with different classes of customers, subchains, arbitrary distribution functions of time-service at nodes priorities and blockings was proposed. Generalization of the method suggested in [1, 2] uses the mean and variance of time-intervals between two successive customer arrivals at nodes for studies of closed queueing networks. In this context it is necessary to analyse the distribution function of time-intervals between successive arrivals into (leavings from) nodes of closed network.

The properties of output flows in single-service queueing systems have been considered in a series of papers [3–10], where the conditions under which output flow is Poisson were discussed. These conditions characterize multiphase systems on the phases of which input and output flows are Poisson ones. In [11–15] open exponential networks with feedbacks and external Poisson flows were analysed. It was shown in [11] that a flow between the nodes i and j is Poisson if the probability of transition from i to j is greater than zero, and the node i is not achievable from the node j with the flows between other nodes being non Poisson and even not renewal processes. For a two node cyclic exponential network a cycle time distribution function was obtained in [16]. Flows of closed cyclic networks were also discussed in [17] where it was shown that a network flow is not a recurrent one.

In this paper flows in arbitrarily structured closed exponential network with a single class of customers (every node is achievable from any node) will be analysed. An algorithm of determining a distribution function of intervals between customers (messages) from nodes in an exponential queueing system of arbitrary structure will be proposed.

For an important particular case of a closed cyclic exponential queueing network with an arbitrary number of nodes, the distribution, the mean and variance of time intervals between two successive messages leaving the nodes are found explicitly.

An arbitrarily structured closed network

Let us take up a queueing network of M nodes and N uniform messages circulating between them in accordance with a transition probability matrix $\|P_{ij}\|$.

Denote:

- $A_i(t) = 1 - e^{-\mu_i t}$ — is a service time distribution function at node i ($i = \overline{1, M}$).
- μ_i — the service intensity at node i ($i = \overline{1, M}$).
- $(\bar{n}, n_M) = (n_1, \dots, n_{M-1}, n_M)$ — the network state for stationary mode, where n_i is the number of messages at the node i ($i = \overline{1, M}$);

$$\varepsilon(n) = \begin{cases} 0 & (n=0) \\ 1 & \text{otherwise} \end{cases}$$

$$n = (n_1, \dots, n_{M-1})$$

$$n^{kj} = \begin{cases} (n_1, \dots, n_{k-1}, \dots, n_{j+1}, \dots, n_{M-1}) & (k \neq j) \\ \bar{n} & (k = j) \end{cases}$$

$$\bar{n}^j = (n_1, \dots, n_{j+1}, \dots, n_{M-1});$$

$$\bar{n}_j = (n_1, \dots, n_j, 0, \dots, 0),$$

with $n_j \geq 1, n_{j+1} = \dots = n_{M-1} = 0$.

An event which is the end of servicing in the node M of a message within time t from the end of service of a previous message there (in the node M) can occur in one of the following ways:

1. At the end of service time in the node M , the network state is $(\bar{n}, 0)$ with a probability of $P_{N-1}(\bar{n}, 0)$; the message served in the node M is transferred into the node j ($j = \overline{1, M-1}$) with a probability of P_{Mj} ; some message enters the node and is served there during t with a probability of $R(\bar{n}^j, 0, t) * A_M(t)$ (where $*$ is a convolution symbol).

$R(\bar{n}, 0, t)$ is the distribution function of the time interval from the arrival of a customer in some node of the network (the network changing its state to state $(\bar{n}, 0)$) to arrival of some of the N customers in node M , provided that node M is empty at the customer arrival time.

2. At the end of service time in the node M , the network is in the state $(\bar{n}, n_M \geq 1)$ with a probability of $P_{N-1}(\bar{n}, n_M)$; the message is served in the node M during t , with a probability of $A_M(t)$.

3. At the end of servicing in the node M the network is in the state $(\bar{n}, 0)$, with a probability of $P_{N-1}(\bar{n}, 0)$; the message served in the node M returns to the node M with a probability of P_{MM} ; the message is served in the node M during t , with a probability of $A_M(t)$.

Hence a Laplace-Stieltjes transform (L-S tr.) $\varphi_M(s)$ of the distribution function $\Phi_M(t)$ of times between sequence message leavings from the node M can be represented as:

$$\begin{aligned} \varphi_M(s) = & \frac{\mu_M}{\mu_M + s} \sum_{j=1}^{M-1} \sum_{|\bar{n}|=N-1} P_{N-1}(\bar{n}, 0) \pi(\bar{n}^j, 0, s) P_{Mj} + \\ & + \frac{\mu_M}{\mu_M + s} \sum_{\substack{|\bar{n}, n_M|=N-1 \\ n_M \geq 1}} P_{N-1}(\bar{n}, n_M) + \frac{\mu_M}{\mu_M + s} P_{MM} \sum_{\substack{|\bar{n}|=N-1 \\ n_M=1}} P_{N-1}(\bar{n}, 0), \end{aligned} \quad (1)$$

where: $\pi(\bar{n}^j, 0, s)$ — L-S tr. of the distribution function $R(\bar{n}^j, 0, t)$;

$$|\bar{n}| = \sum_{i=1}^{M-1} n_i; \quad |\bar{n}, n_M| = \sum_{i=1}^M n_i,$$

N is the number of customers (messages) circulated in the network. Here the stationary probabilities $P_{N-1}(\bar{n}, n_M)$ are calculated, following [18, 19] as stationary probabilities of closed network states with $N-1$ messages at an arbitrary time.

It is easy to show¹ that $\pi(\bar{n}, 0, s)$ are found from the following set of linear equations:

$$\pi(\bar{n}, 0, s) \left(\sum_{k=1}^{M-1} \varepsilon(n_k) \mu_k + s \right) = \quad (2)$$

¹ Indeed, for nodes busy at the time of customer arrival in some node of the network, the Laplace-Stieltjes transform of the distribution function of time to customer leaving the (busy) k -th node, provided that none of customers did not leave the other nodes is given in the form

$$\int_0^{\infty} e^{-\sum_{i \neq k} \varepsilon(n_i) \mu_i v - sv} d(1 - e^{-\mu_k v}) = \mu_k \left(\sum_{i=1}^{M-1} \varepsilon(n_i) \mu_i + s \right).$$

Hence, since transition to state $(\bar{n}^k, 0)$ occurs with probability P_{kj} and the probability of customer arrival at node M is equal to P_{kM} , we have the system (2).

$$= \sum_{k=1}^{M-1} \left[\mu_k \varepsilon(n_k) \sum_{j=1}^{M-1} \pi(\bar{n}^{kj}, 0, s) P_{kj} \right] + \sum_{k=1}^{M-1} \varepsilon(n_k) \mu_k P_{kM}$$

Solving it and substituting $\pi(\bar{n}, 0, s)$ in (1) we have $\varphi_M(s)$. Let us write out, for example, a system (2) for the network of Fig. 1 which is a model of a multiprogramming computer system [20] with a multiprogramming level $N = 2$ (the number of nodes $M = 3$).

$$\begin{aligned} \pi(1, 1, 0, s) &= (\pi(0, 2, 0, s)P_{12} + P_{1M}) \frac{\mu_1}{\mu_1 + \mu_2 + s} + \\ &+ (\pi(2, 0, 0, s)P_{21} + P_{2M}) \frac{\mu_2}{\mu_1 + \mu_2 + s} \\ \pi(0, 2, 0, s) &= (\pi(1, 1, 0, s)P_{21} + P_{2M}) \frac{\mu_2}{\mu_2 + s} \\ \pi(2, 0, 0, s) &= (\pi(1, 1, 0, s)P_{12} + P_{1M}) \frac{\mu_1}{\mu_1 + s}. \end{aligned}$$

Solving the above system and finding $\varphi_M(s)$ leads to the productivity and other characteristics of the multiprogramming system.

A cyclic network

Consider a closed cyclic exponential M -node network with N messages cycling in it (Fig. 2).

Denote by $\tau_M^{(1)}$ and $\tau_M^{(2)}$ the mean and variance of time intervals between messages leaving the node M .

Let us introduce the following functions $g(n, m)$ and $G(n)$:

$$\begin{aligned} g(n, 1) &= (a_1)^n && \text{for } n = \overline{0, N}; \\ g(0, m) &= 1 && \text{for } m = \overline{1, M} \\ g(n, m) &= g(n, m-1) + a_m g(n-1, m) && \text{for } m > 1 \text{ and } n > 0 \\ G(n) &= g(n, M) && \text{for } n = \overline{0, N}, \end{aligned}$$

where $a_i = 1/\mu_i$ ($i = \overline{1, M}$).

Denote

$$\underset{k=j}{*}^M A_k(t) = A_j(t) * A_{j+1}(t) * \dots * A_M(t).$$

Statement 1. Under the given above assumptions

$$\begin{aligned}\Phi_M(t) &= \sum_{j=1}^M (Q_j * A_k(t)) \\ \varphi_M(s) &= \sum_{j=1}^M Q_j \prod_{k=j}^M \left(\frac{\mu_k}{\mu_k + s} \right) \\ \tau_M^{(1)} &= \sum_{j=1}^M \left(Q_j \sum_{k=j}^M \frac{1}{\mu_k} \right) = \frac{G(N)}{G(N-1)} \\ \tau_M^{(2)} &= \sum_{j=1}^M \left\{ Q_j \left[\left(\sum_{k=j}^M \frac{1}{\mu_k} \right)^2 + \sum_{k=j}^M \frac{1}{\mu_k^2} \right] \right\},\end{aligned}$$

where $Q_j = a_j g(N-2, j) / G(N-1)$.

Proof. In this case the transition probability matrix is $\|P_{ij}\|$ transformed as follows: $P_{i, i+1} = 1$, $(i = 1, M-1)$, $P_{M1} = 1$, and the remaining $P_{ij} = 0$ (3)

It is easy to check by direct substitution that the solution of equations (2) is:

$$\pi(\bar{n}_j, 0, s) = \prod_{k=j}^{M-1} \left(\frac{\mu_k}{\mu_k + s} \right). \quad (4)$$

From (1) with (3) and (4) we have distribution function $\Phi_M(t)$:

$$\Phi_M(t) = \sum_{j=1}^M \left(\sum_{k=j}^M A_k(t) \right) \sum_{\substack{(n_1, \dots, n_j) \\ n_j \geq 1}} P_{N-1}(n_1, \dots, n_j, 0, \dots, 0)$$

In accordance with [21]

$$P_{N-1}(n_1, \dots, n_M) = \frac{1}{G(N-1)} \prod_{i=1}^M (a_i)^{n_i}$$

and

$$\begin{aligned}\sum_{\substack{(n_1, \dots, n_j) \\ n_j \geq 1}} P_{N-1}(n_1, \dots, n_j, 0, \dots, 0) &= \sum_{\substack{(n_1, \dots, n_j) \\ n_j \geq 1}} \frac{1}{G(N-1)} \prod_{i=1}^j (a_i)^{n_i} = \\ &= a_j \frac{1}{G(N-1)} \sum_{\substack{(n_1, \dots, n_j) \\ \sum_{k=1}^j n_k = N-2}} \prod_{i=1}^j (a_i)^{n_i} = a_j \frac{1}{G(N-1)} g(N-2, j) = Q_j.\end{aligned}$$

In particular

$$Q_1 = a_1 \frac{a_1^{(N-2)}}{G(N-1)}, \quad Q_M = a_M \frac{G(N-2)}{G(N-1)}.$$

Rewrite $\Phi_M(t)$ in the following way

$$\Phi_M(t) = \sum_{j=1}^M \left(Q_j \left(\underset{k=j}{*} A_k(t) \right) \right),$$

hence

$$\varphi_M(s) = \sum_{j=1}^M Q_j \prod_{k=j}^M \left(\frac{\mu_k}{\mu_k + s} \right) \quad (5)$$

From (5) we find the moments $\tau_M^{(r)}$ of intervals between messages leaving the node M :

$$\tau_M^{(r)} = (-1)^r \frac{d^r}{(ds)^r} \varphi_M(s) |_{s=0}$$

In particular, the mean and variance $\tau_M^{(1)}$ and $\tau_M^{(2)}$ are:

$$\tau_M^{(1)} = \sum_{j=1}^M \left(\sum_{k=j}^M Q_j \frac{1}{\mu_k} \right);$$

$$\tau_M^{(2)} = \sum_{j=1}^M \left\{ Q_j \left[\left(\sum_{k=j}^M \frac{1}{\mu_k} \right)^2 + \sum_{k=j}^M \frac{1}{\mu_k^2} \right] \right\}.$$

In Appendix 1 it is proved that

$$\tau_M^{(1)} = \sum_{j=1}^M \left(\sum_{k=j}^M Q_j \frac{1}{\mu_k} \right) = \frac{G(N)}{G(N-1)}$$

The resultant expression for $\tau_M^{(1)}$ coincides with the expressions determined in [21] (see Appendix 1).

The variation coefficient $C_M^{(v)}$ of intervals between messages leaving the node M is equal to:

$$C_M^{(v)} = (\tau_M^{(2)} - (\tau_M^{(1)})^2) / (\tau_M^{(1)})^2 = \tau_M^{(2)} / (\tau_M^{(1)})^2 - 1.$$

At the node leaving times a stationary state distribution of closed exponential message network N coincides with the stationary state network distribution with $N - 1$ messages at arbitrary moment [18, 19] whatever is the node left by the message. In other words, any message entering any node i every time meets n_i messages there with probability of $P_{N-1}(n_i)$. For the message leaving any node to return to the starting point it must pass through every node once. Because of the exponential service at the nodes the cycle time distribution function is the same for all the nodes.

The random variable of cycle time with respect to any node is a sum of N intervals between messages leaving this node, because during a message cycle with respect to any node of the cyclic network, exactly N messages pass through this node. From the above it follows that the distribution functions of intervals between messages leaving the nodes are identical for all nodes of a cyclic exponential network.

Consequently:

$$\tau_i^{(r)} = \tau_M^{(r)} = \tau^{(r)}, \quad r = 1, 2, \dots; \quad i = \overline{1, M}.$$

Consider the asymptotic behaviour of a cyclic network under large $N \rightarrow \infty$ workloads.

Denote $\lambda_1 = 1/\tau^{(1)}$, $\rho_i = \lambda_i a_i = \lambda_1 a_i$, $b_i = a_i/a_1$ ($i = \overline{1, M}$), $b_* = \max_{2 \leq i \leq M} b_i$, $\lambda_{*1} = 1/a_1$,

$$\rho_{*i} = \lambda_{*1} a_i.$$

In addition, let $S(N, M) = \left\{ (n_1, \dots, n_M) \mid \sum_{i=1}^M n_i = N \text{ and } n_i \geq 0 \forall i \right\}$

$$S_0(N, M) = \left\{ (n_1, \dots, n_M) \mid \sum_{i=2}^M n_i = N \quad \text{and} \quad n_i \geq 0 \forall i \right\}$$

$$S_1(N, M) = \left\{ (n_1, \dots, n_M) \mid \sum_{i=2}^M n_i \leq N-1 \quad \text{and} \quad n_i \geq 0 \forall i \right\}$$

$$S_{1k}(N, M) = \left\{ (n_1, \dots, n_M) \mid \sum_{i=2}^M n_i = k \quad \text{and} \quad n_i \geq 0 \forall i \right\}.$$

Suppose (the most interesting case), that $\exists! i_* : a_{i_*} = \max_i a_i$. Without restricting the generality, let $i_* = 1$.

Statement 2. In the above assumptions there exists

$$\lim_{N \rightarrow \infty} P\{n_1 \geq 1\} = 1 \quad \text{and} \quad \lim_{N \rightarrow \infty} P(n_1, \dots, n_M) = \prod_{i=2}^M (1 - \rho_{*i}) \rho_{*i}^{n_i}$$

and the rate of convergence when $k \rightarrow \infty$ is at least

$$C((Mb_* - b_*^2)/(Mb_* - 2b_* + 1))^k$$

where C is a some constant.

The proof is given in Appendix 2.

From Statement 2 it follows that asymptotically, under large workloads $N \rightarrow \infty$ the characteristics of the cyclic network coincide with characteristics of a multiphase exponential system with $M - 1$ nodes, into which a Poisson flow is fed with a parameter $1/a_1$.

To estimate how much closed cyclic exponential network flows differ from Poisson ones under different network workloads (different N), we may use as a criterion¹ the value $\chi = [2 - \tau^{(2)}/(\tau^{(1)})^2] \cdot 100\%$, which is equal to the difference (in percentage) between the exponential distribution variation coefficient and the variation coefficient of the distribution of the intervals between messages leaving the nodes.

¹ Note that other criteria can also work [22].

Write out χ in the analytical form:

$$\chi = \left[2 - G(N-1)/(G(N))^2 \sum_{j=1}^M \left\{ a_j g(N-2, j) \left[\left(\sum_{k=j}^M a_k \right)^2 + \sum_{k=j}^M a_k^2 \right] \right\} \right] \cdot 100\%.$$

The dependence of χ on the number of messages in a cyclic exponential network with parameters $M=5$, $a_1=10$, $a_2=8$, $a_3=2$, $a_4=4$, $a_5=6$ is shown in Fig. 3. Under the small loads the flow is clearly much different from a Poisson one, but with an increasing number of network messages the flow asymptotically approaches the Poisson one.

Appendix 1

$\tau_M^{(1)}$ can be presented as

$$\begin{aligned} \tau_M^{(1)} &= \sum_{j=1}^M \left(\sum_{k=j}^M Q_j \frac{1}{\mu_k} \right) = \frac{1}{G(N-1)} \sum_{k=1}^M \sum_{j=k}^M a_j g(N-2, j) a_k = \\ &= \frac{1}{G(N-1)} \sum_{k=1}^M \sum_{j=1}^k a_j g(N-2, j) a_k = \frac{1}{G(N-1)} \sum_{k=1}^M a_k r_k, \end{aligned}$$

where

$$r_k = \sum_{j=1}^k a_j g(N-2, j).$$

Let us prove by induction with respect to k that $r_k = g(N-1, k)$, using the property $g(n, k) = g(n, k-1) + a_k g(n-1, k)$ for $k > 1$ and $n > 0$

$$r_1 = a_1 g(N-2, 1) = a_1 a_1^{N-2} = a_1^{N-1} = g(N-1, 1)$$

$$r_2 = r_1 + a_2 g(N-2, 2) = g(N-1, 1) + a_2 g(N-2, 2) = g(N-1, 2).$$

Assuming that $r_{k-1} = g(N-1, k-1)$, we have

$$\begin{aligned} r_k &= r_{k-1} + a_k g(N-2, k) = \\ &= g(N-1, k-1) + a_k g(N-2, k) = g(N-1, k) \end{aligned}$$

Denote

$$s_i = \sum_{k=1}^i a_k r_k = \sum_{k=1}^i a_k g(N-1, k).$$

Let us prove by induction with respect to i that $s_i = g(N, i)$

$$s_1 = a_1 g(N-1, 1) = a_1 a_1^{N-1} = a_1^N = g(N, 1)$$

$$s_2 = s_1 + a_2 g(N-1, 2) = g(N, 1) + a_2 g(N-1, 2) = g(N, 2).$$

Assuming that $s_{i-1} = g(N, i-1)$ we have

$$s_i = s_{i-1} + a_i g(N-1, i) = g(N, i-1) + a_i g(N-1, i) = g(N, i).$$

For $i = M$ $s_i = \sum_{k=1}^M a_k r_k = g(N, M) = G(N).$

Consequently $\tau_M^{(1)} = G(N)/G(N-1).$

The resultant expression for $\tau_M^{(1)}$ coincides with a similar one in [21], which can be solved directly, from the expression for the probability that the node M is busy:

$$\tau_M^{(1)} = 1/(\mu_M P(n_M \geq 1)) = 1/(G(N-1)/G(N)) = G(N)/G(N-1) \quad (6)$$

Appendix 2

From (6) it follows that $P\{n_1 \geq 1\} = a \frac{G(N-1)}{G(N)}$

hence
$$P\{n_1 \geq 1\} = \left(a_1^N \sum_{S(N-1, M)} \prod_{i=1}^M a_i^{n_i} \right) / \left(a_1^{N-1} \sum_{S(N, M)} \prod_{i=1}^M a_i^{n_i} \right) =$$

$$= \left(\sum_{S(N-1, M)} \prod_{i=2}^M b_i^{n_i} \right) / \left(\sum_{S(N, M)} \prod_{i=2}^M b_i^{n_i} \right).$$

But

$$\sum_{S(N, M)} \prod_{i=2}^M b_i^{n_i} = \sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N, M)} \prod_{i=2}^M b_i^{n_i} =$$

$$= \sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_0(N-1, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N-1, M)} \prod_{i=2}^M b_i^{n_i}. \quad (7)$$

Because the total number of states in the network with k messages is equal to C_{M+k-1}^k

then
$$\sum_{S_0(k, M)} \prod_{i=2}^M b_i^{n_i} < C_{M+k-1}^k b_*^k = d_k.$$

Let

$$\varepsilon_k = b_*(M-1)/(k+1), \quad k_0 = b_*(M-1)/(1-b_*), \quad D_0 = d_{k_0}/(b_* + \varepsilon_{k_0})^{k_0}.$$

Then with

$$k \geq k_0 \quad b_*(M+k)/(k+1) = b_* + \varepsilon_k \leq b_* + \varepsilon_{k_0} \leq 1.$$

Consequently when

$$k \geq k_0 \quad d_{k+1} < d_k(b_* + \varepsilon_{k_0}).$$

Hence when

$$k > k_0 \quad d_k < D_0(b_* + \varepsilon_{k_0})^k \quad (8)$$

Since

$$b_0 + \varepsilon_{k_0} < 1 \quad \text{then} \quad d_k \rightarrow 0 \quad \text{with} \quad k \rightarrow \infty.$$

Hence
$$\sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} \rightarrow 0 \quad \text{with } N \rightarrow \infty.$$

Represent
$$\sum_{S_1(N, M)} \prod_{i=2}^M b_i$$
 as
$$\sum_{S_1(N, M)} \prod_{i=2}^M b_i = \sum_{k=0}^{N-1} \sum_{S_{1k}(N, M)} \prod_{i=2}^M b_i^{n_i}.$$

As
$$\sum_{k=0}^{\infty} \sum_{S_{1k}(N, M)} \prod_{i=2}^M b_i^{n_i}$$

is the product of $\sum_{n_i=0}^{\infty} b_i^{n_i}$ ($i=2, M$) in the Cauchy form then

$$\lim_{N \rightarrow \infty} \sum_{S_1(N, M)} \prod_{i=2}^M b_i^{n_i} = \prod_{i=2}^M \sum_{n_i=0}^{\infty} b_i^{n_i} = \prod_{i=2}^M (1/(1-b_i)).$$

Using (7), we have

$$\begin{aligned} \lim_{N \rightarrow \infty} P\{n_1 \geq 1\} &= \lim_{N \rightarrow \infty} \left[\left(\sum_{S_0(N-1, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N-1, M)} \prod_{i=2}^M b_i^{n_i} \right) / \right. \\ &\quad \left. \left(\sum_{S_0(N, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_0(N-1, M)} \prod_{i=2}^M b_i^{n_i} + \sum_{S_1(N-1, M)} \prod_{i=2}^M b_i^{n_i} \right) \right] = 1. \end{aligned}$$

From (8) it follows that the rate of convergence when $k \rightarrow \infty$ is not less than

$$C((Mb_* - b_*^2)/(Mb_* - 2b_* + 1))^k,$$

where C is a constant.

Since $\lim_{N \rightarrow \infty} P\{n_1 \geq 1\} = 1$ then

$$\lim_{N \rightarrow \infty} \lambda_1 = \lambda_{*1} = 1/a_1 \quad \text{and} \quad \lim_{N \rightarrow \infty} \rho_i = \rho_{*i} = b_i.$$

Hence

$$\begin{aligned} \lim_{N \rightarrow \infty} P(n_1, \dots, n_M) &= \left(\prod_{i=2}^M \rho_{*i}^{n_i} \right) / \left(\prod_{i=2}^M (1/(1-\rho_{*i})) \right) = \\ &= \prod_{i=2}^M (1-\rho_{*i}) \rho_{*i}^{n_i}. \end{aligned}$$

References

1. Vishnevskii, V. M., Gerasimov, A. I., "An approximate method for the study of queueing networks with several classes of customers". In: Tretya shkola po avtomatizirovannym sistemam massovogo obsluzhivania. Vinnitca; Vinnitskiy Dom Tekhniki; pp. 24-25, 1981.

2. *Gerasimov, A. I.*, "An approximate method for the study of hierarchical queueing networks with node blocking by customers". In: *Teoria i tekhnika avtomatizirovannikh sistem massovogo obsluzhivaniya*, Moscow: Izd. Moskovskogo Doma Nauchno-Tekhnicheskoy Propagandy, pp. 82-85, 1982.
3. *Aleksandrov, A. M.*, "On outcoming flows of one queueing system class". *Izvestia AN SSSR. Tekhnicheskaya kibernetika*, No 4, pp. 3-11, 1968.
4. *Yashkov, S. F.*, "On one service discipline class for computing networks models". In: *Informazionno-Vichislitelnye seti EVM*. Moscow: Izd. Moskovskogo Doma Nauchno-Tekhnicheskoy Propagandy, pp. 112-117, 1980.
5. *Tolmachev, A. L.*, "On the serviced flow of a $GI/M/1r \leq \infty$ system". In: *Sistemy massovogo obsluzhivaniya i kommutatsii*, Moscow: Nauka, pp. 8-17, 1974.
6. *Vishnevskii, V. M., Timokhova, T. A.*, "On the outcoming queueing system flow with unreliable service device". In: *Aktual'nye voprosy teorii i praktiki upravleniya*, Moscow: Nauka, pp. 23-32, 1977.
7. *Falin, G. I.*, "The effect of repeated calls on the outcoming flow of a one-channel queueing system". *Izvestia AN SSSR. Tekhnicheskaya Kibernetika*, No. 4, pp. 114-118.
8. *Kleinrock, L.*, *Queueing systems, volume I: Theory*, New York, 1975.
9. *Noetzel, A. S.*, A generalized discipline for product form network solutions, *Journal of the ACM*, vol. 26, No. 4, pp. 779-793, 1979.
10. *Naumov, V. A.*, On Autonomous Operation of Subsystems in a Complex System. — In: *The Queueing Theory. Proceedings of the Third All-Union Workshop on the Queueing Theory*, vol. 2, Moscow University Press, 1976, pp. 169-177 (in Russian).
11. *Beutler, F. J., Melamed, B.*, Decomposition and customer streams of feedback networks of queues in equilibrium, *Operations research*, vol. 26, No. 6, pp. 1059-1072, 1978.
12. *Pujolle, G., Soula, C.*, A study of flows in queueing networks and approximate method for solution, 4-th International Symp. Modell. and Performance Eval. Comput. Systems, Vienna. Conf. Prepr., vol. 2, pp. 189-203, 1979.
13. *Disney, R. L.*, Random flow in queueing networks: A review and critique, *AIIE Transactions*, vol. 7, No. 3, pp. 268-288, 1975.
14. *Daley, D. J.*, Queueing output processes, *Advances in Applied Probability*, vol. 8, No. 2, pp. 395-415, 1976.
15. *Labetoulle, J., Pujolle, G., Soula, C.*, Stationary distributions of flows in Jacson networks, *Math. Oper. Res.*, 1981, vol. 6, No. 2, pp. 173-185.
16. *Chow, W.-M.*, The cycle time distribution of exponential cyclic queues, *Journal of the ACM*, vol. 27, No. 2, pp. 281-286, 1980.
17. *Vishnevskii, V. M., Sibirskaya, T. K.*, "Flows in closed queueing networks". *Tret'ya shkola po avtomatizirovannym sistemam massovogo obsluzhivaniya*, Vinnitsa, Vinnitskiy Dom Tekhniki, pp. 56, 1981.
18. *Tolmachev, A. L.*, Some characteristics of closed exponential network. In: *Teoria teletrafika i informatsionnye seti* Moscow: Nauka, pp. 3-6, 1977.
19. *Reiser, M., Lavenberg, S. S.*, Mean-value analysis of closed multichain queueing networks, *Journal of the ACM*, vol. 27, No. 2, pp. 313-328, 1980.
20. *Vishnevskii, V. M., Tverdohlebov, A. S.*, "Models of closed networks with blockings for the analysis for multiprogrammed computer systems". *Avtomatika i telemekhanika*, No. 5, pp. 172-179, 1980.
21. *Buzen, J. P.*, Computational algorithms for closed queueing networks with exponential servers, *Communications of the ACM*. vol. 16, No. 9, pp. 527-531, 1973.
22. *Basharin, G. P., Kokotushkin, V. A., Naumov, V. A.*, The Method of Equivalent Substitutions in the Teletraffic Theory. *Itogi nauki i tekhniki. Elektrosvyaz*, vol. 11, Moscow: VINITI, 1980, pp. 82-122. (in Russian)

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospekt 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H-1111 Budapest XI., Stoczek u. 2, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper with wide margins (4–5 cm), should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice-versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50–100 слов и приложено резюме – реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в трех экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициях. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Kleimenov, A. F.</i> Optimal strategies in a hierarchical differential game (Клейменов А. Ф. Оптимальные стратегии в одной иерархической дифференциальной игре)	369
<i>Kučera, V.</i> Block decoupling by dynamic compensation with internal properness and stability (Кучера В. Обеспечение автономности вместе с внутренней правильностью и устойчивостью методом динамической компенсации)	379
<i>Вишневецкий В. М., Герасимов А. И.</i> Исследование потоков в замкнутых экспоненциальных сетях массового обслуживания (Vishnevskii, V. M., Gerasimov, A. I. Analysis of flows in closed exponential queueing networks)	391
<i>Györfi, L., Vajda, I.</i> Block coding and correlation decoding for an M -user weighted adder channel (Дёрфи, Л., Вайда, И. Блочное кодирование и корреляционное декодирование для взвешенного суммирующего канала с M пользователями)	405
<i>Šujan, Š.</i> Sinai's theorem and entropy compression (Шуйян Ш. Теорема Синая и энтропийное сжатие)	419
<i>Dinh The Luc</i> Duality in programming under probabilistic constraints with a random technology matrix (Дин Тхе Лук Двойственность в задаче программирования при вероятностных ограничениях со случайной технологической матрицей)	429