VOL. 15 · NUMBER 1
ТОМ 15 · НОМЕР 1

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 СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1986

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
- — theory of adaptive systems
- — theory of estimation and identification
- — theory of controlling robot-technologic and flexible manufacturing systems
- — information theory
- — information-theoretic aspects of multiple access networks.

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;
- — теория адаптивных систем;
- — теория оценивания и идентификации;
- — теория управления робототехническими и гибкими производственными системами;
- — теория ипформации;
- — теория ипформации в области сетей с множественным доступом.

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 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1985 Subscription Rate US Dollars 160.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY VOLUME **15** (1986)

SUBJECT INDEX

Rafajłowicz, E.: *L*-optimal input signals for distributed-parameter systems identification. **15,** *1,* pp. 79–89.

Serkov, D. A.: Synthesis of feedback control and finite dimensional models. **15,** *3,* pp. 239–251.

Shaikhet, L. E.: On a necessary condition of optimality for stochastic systems with noise by control. **15,** *1,* pp. 35–46.

Sinitsyn, I. N.: Stochastic hereditary control systems. **15,** *4,* pp. 287–298.

Subbotin, A. I., Tarasyev, A. M.: Stability properties of the value function of a differential game and viscosity solutions of Hamilton–Jacobi equations. **15,** *6,* pp. 451–463.

Ustyuzanin, A. M.: On the problem of matrix parameter identification. **15,** *4,* pp. 265–273.

Vajda, I.: Robust estimation in discrete and continuous families by means of a minimum chi-square method. **15,** *2,* pp. 111–127.

Venetsanopoulos, A. N., Singh, I.: Topological optimization of communication network subject to reliability constraints. **15,** *1,* pp. 63–78.

Víšek, J. Á.: A note on numerical aspects of rob ıt testing. **15,** *4,* pp. 299–307.

Zenkevich, S. L.: Logical control of adaptive robot. Principles of control system organization. **15,** *4,* pp. 319–331.

Zítek, P.: Anisochronic modelling and stability criterion of hereditary systems. **15,** *6,* pp. 413–423.

PROBLEMS OF CONTROL AND INFORMATION THEORY VOLUME **15** (1986)

AUTHOR INDEX

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

1828

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

Problems of Control and Information Theory, Vol. 15 (1), pp. 3–17 (1986)

THE OPTIMAL CONTROL OF STOCHASTIC SEQUENCES WITH DELAY

O. V. ALISEENKO, V. M. KHAMETOV, A. B. PIUNOVSKI

(*Moscow*)

(Received December 11, 1984)

The controlled Markov chain with control's delay is investigated. The definition of the τ-models' class is given. For such models the methods similar to those of dynamic programming are developed. The characteristic properties of the τ-models are presented; the concrete example is considered.

1. Introduction

Very often the usual dynamic systems do not describe the behaviour of the real processes completely enough. Sometimes the control on the output of regulator has a delay. For example: the problem of control of remote moving objects; the problem of service of technical systems in which latent refusals can appear and so on. The main feature of such problems is the following: the system is broken on the interval $[0, \tau)$, where τ is the value of delay. Because of this, the difficulty in the control's choice appears. The purpose of the present work is to study the controlled Markov chain with delay.

In the works [1–7] devoted to the optimal control of different Markov processes without delay the Bellman's equation is obtained and the necessary and sufficient conditions of the simple Markov policies' optimality are presented. Contrary to the works noted above we consider the problem of optimal synthesis in the models with control's delay. The consideration of such models results usually in essential difficulties associated with the following fact: the process is not Markovian even for simplest control policies.

In the present paper sufficient conditions of the optimal Markov policies' existence are formulated and a theorem giving the necessary and sufficient conditions of the policy's optimality is proved. The main results concern the case when the model Z is the τ-model. The characteristic property of τ-models is presented in Theorem 3.

Theorem 4 gives the sufficient conditions of the τ-models' existence. The optimization task for the linear system with the square criterion is considered to be an

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

example. Note that our task differs from those investigated earlier [1, 2, 8] because of the varying control's delay.

All the statements can be obviously formulated and proved for the models with varying delay $\tau=\tau(t)$. Note that in the case $\tau=0$ all the results reduce into the well-known facts of the dynamic programming theory [2–4].

2. Notations and definitions

Description of the model. Let $(\Omega, F, P, (\mathscr{F}_t)_{t\in N_T})$ be a complete stochastic basis $(N_T=\{0, 1, \ldots, T\})$ on which the Markov chain ξ_t^x with values in the Borel space $(X, \mathscr{B}(X))$, is given; $\xi_0^x=x\in X$; $\mathscr{F}_t=\sigma\{\xi_s^x, s\leqq t\}$. Let the Borel space A be given (the control space). Let the transition probability of $\{\xi_t^x\}$ for one step (by a fixed value of the control parameter $a_t=a\in A$) be defined by $P_a(t, y, \Gamma)\triangleq P\{\xi_t^x\in\Gamma\,|\,\xi_{t-1}^x=y, a_t=a\}$. Let $R(t, y, a)$ be the profit value on the step t. The control a_t is assumed to be $\mathscr{F}_{(t-\tau-1)\vee 0}$-measurable. Here and everywhere further $\theta_1\vee\theta_2\triangleq\max\{\theta_1,\theta_2\}$; $\theta_1\vee\theta_2\triangleq\min\{\theta_1,\theta_2\}$. Define the final profit by $S(\xi_T^x)$. The object $Z=\{N_T, X, A, P, R, S\}$ is called the model as usually [2]. The functions P, R, S are assumed to be measurable.

Def. The control policy $\pi(x)$ is any $\mathscr{F}_{(t-\tau-1)\vee 0}$-measurable function $\pi[(y)_0^{(t-\tau-1)\vee 0}]$ with values in A. (The symbol $(y)_0^\theta$ denotes a trajectory of the process ξ_t^x on the interval $[0, \theta]$.) Similarly to the notation $\pi(x)$, we shall use the notation a_t^x.

Def. Denote $W(\pi(x))\triangleq M^{\pi(x)}\left[\sum_{t=1}^{T} R(t, \xi_{t-1}^x, a_t^x)+S(\xi_T^x)\right]$. The function $W(\pi(x))$ will be called the policy's $\pi(x)$ value.

Here and everywhere further the symbol $M^{\pi(x)}$ means the averaging (the integral) by the measure $P^{\pi(x)}$ on $\mathscr{F}=\sigma\left\{\bigcup_{t\in N_T}\mathscr{F}_t\right\}$ corresponding to the policy $\pi(x)$.

Everywhere further we assume that the available policies' class Π includes only those policies for which the conditional means $M^{\pi(x)}\left[\sum_{\theta=t+1}^{T} R(\theta, \xi_{\theta-1}^x, a_\theta^x)+\right.$ $\left.+S(\xi_T^x)\,|\,\mathscr{F}_{(t-\tau)\vee 0}\right]$ exist for every $t\in N_T$.

Def. The control policy $\hat{\pi}(x)$ is called optimal if $W(\hat{\pi}(x))=\sup\limits_{\pi(x)} W(\pi(x))$.

Def. The control policy $\pi(x)$ is called Markov policy if $\pi[(y)_0^{(t-\tau-1)\vee 0}]=\pi_t(P_{t-1}^{\pi(x)})$. Here and everywhere further $P_t^{\pi(x)}(\Gamma, (y)_0^{(t-\tau)\vee 0})\triangleq P^{\pi(x)}\{\xi_t^x\in\Gamma\,|\,\mathscr{F}_{(t-\tau)\vee 0}\}$ is the conditional stochastic measure with respect to σ-algebra $\mathscr{F}_{(t-\tau)\vee 0}$ (if $t\leqq\tau$ then $P_t^{\pi(x)}$ is an unconditional probability). We shall denote the set $P_t^{\pi(x)}(\,\cdot\,, (y)_0^{(t-\tau)\vee 0})$ by the symbol B_t. The set of all Markov policies will be denoted by Π^m.

Remark. According to the last definition all the policies

$$a_t^x = \pi[(y)_0^{(t-\tau-1)\vee 0}] \triangleq \pi(M_{t-1}^1, \ldots, M_{t-1}^N)$$

will be Markov policies

$$(M_{t-1}^k \triangleq M^{\pi(x)}[(\xi_{t-1}^x)^k \mid \mathscr{F}_{(t-\tau-1)\vee 0}], \quad k=1,2,\ldots,N;\ N<\infty;\ X=\mathscr{R}^1).$$

Remark. In the case $\tau=0$ all the measures $P_t^{\pi(x)}(\cdot, (y)_0^t)$ are indicators of one-point sets; $B_t = B$ (B does not depend on t) is exactly the set of values of the reflection

$$x \to \chi_x(\cdot, (y)_0^t);$$

here

$$\chi_x(\Gamma, (y)_0^t) \triangleq \begin{cases} 1, & x \in \Gamma, \\ 0, & x \notin \Gamma. \end{cases}$$

So the model Z is bijectively connected with the model $\tilde{Z}$, in which the phase space is B. It is obvious that the values of corresponding policies in the models Z and $\tilde{Z}$ are equal, i.e. the investigation of the model Z is equivalent to the investigation of the model $\tilde{Z}$. Moreover the Markov policies in the model Z correspond to the classic Markov policies [2–6] in the model $\tilde{Z}$.

Def. A function $f(y, a): X \times A \to R^1$ will be called uniformly upper semi-continuous (with respect to a) if

$$\forall\, a \in A\ \forall\, \varepsilon > 0\ \exists\, \delta > 0\ \forall\, y \in X\ \forall\, \tilde{a} \in A \rho_A(\tilde{a}, a) < \delta \to f(y, \tilde{a}) < f(y, a) + \varepsilon.$$

(Here ρ_A is the metric in the Borel space A.)

Let us assume that the following conditions (c) are satisfied:

C1. The function $S(R)$ is uniformly continuous (upper semi-continuous) with respect to a.

C2. The function $g(y, a) \triangleq \int_X f(z) P_a(t, y, dz)$ is uniformly upper semi-continuous with respect to a for every bounded continuous function $f(y)$.

C3. The space A is a compact.

Together with the condition C2 we shall need the following stronger condition:

C2′. The function $g(y, a) \triangleq \int_X f(z) P_a(t, y, dz)$ is continuous with respect to the pair (y, a) for every bounded continuous function $f(y)$.

Remark. As for the conditions C2, C2′ note that if ξ_t^x is a solution of the equation $\xi_t^x = \Phi_t(\xi_{t-1}^x, a_t, \eta_t)$ ($\{\eta_t\}$ is a sequence of independent random variables), then C2, C2′ are satisfied if the function Φ_t is measurable with respect to all its arguments and uniformly continuous with respect to (ξ_{t-1}^x, a_t).

3. Main results

Theorem 1. Let the set Π coincide with the set of all policies and conditions C1–C3 be satisfied. Assume that $F_t(y)$ is an arbitrary bounded continuous function. Then

1) with $P^{\pi(x)}$-probability 1 there exists

$$\sup_{a \in A} \{M^{\pi(x)}[R(t, \xi_{t-1}^x, a) - F_{t-1}(\xi_{t-1}^x) + \int_X F_t(y) \times$$

$$\times P_a(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}]\} < \infty$$

which is achieved for all $t \in N_T$ and for every policy $\pi(x)$.

2) for every $\varepsilon > 0$ there exists a Markov policy $\hat{\pi}(x) = \hat{a}_t^x$ such that the following inequalities are satisfied:

$$M^{\hat{\pi}(x)}[R(t, \xi_{t-1}^x, \hat{a}_t^x) - F_{t \mp 1}(\xi_{t-1}^x) + \int_X F_t(y) \times P_{\hat{a}_t^x}(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}] >$$

$$> \sup_{a \in A} \{M^{\hat{\pi}(x)}[R(t, \xi_{t-1}^x, a) - F_{t-1}(\xi_{t-1}^x) +$$

$$+ \int_X F_t(y) P_a(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}]\} - \varepsilon (t \in N_T)$$

with $P^{\hat{\pi}(x)}$-probability 1.

3) there exists a Markov policy $\hat{\pi}(x) = \hat{a}_t^x$ such that the following equality is satisfied:

$$\sup_{a \in A} \{M^{\hat{\pi}(x)}[R(t, \xi_{t-1}^x, a) - F_{t-1}(\xi_{t-1}^x) + \int_X F_t(y) \times$$

$$\times P_a(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}]\} = M^{\hat{\pi}(x)}[R(t, \xi_{t-1}^x, \hat{a}_t^x) - F_{t-1}(\xi_{t-1}^x) +$$

$$+ \int_X F_t(y) P_{\hat{a}_t^x}(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}]\} \tag{1}$$

$(t \in N_T)$ with $P^{\hat{\pi}(x)}$-probability 1. ("Equality (1) has a measurable choice".)

Remark. Point 2) of Theorem 1 is true if there exists a finite supremum in (1) with $P^{\hat{\pi}(x)}$-probability 1 for every policy $\hat{\pi}(x)$.

Hypothesis H. Assume that there exists a measurable function $F_t(y)$ such that $F_T(y) = S(y)$; the integral $\int_X F_t(y) P_a(t, z, dy)$ is defined for all $a \in A$, $z \in X$, $t \in N_T$; the function

$$f(t, x) \triangleq \sup_{a \in A} \{M^{\pi(x)}[R(t, \xi_{t-1}^x, a) - F_{t-1}(\xi_{t-1}^x) +$$

$$+ \int_X F_t(y) P_a(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}]\}$$

is $\mathscr{B}(X)$-measurable, independent of $\mathscr{F}_{(t-\tau-1)v_0}$ for every policy $\pi(x) \in \Pi$.

Def. The models in which Hypothesis H is true will be called τ-models.

Remark. The examples of nontrivial τ-models are given in Corollary 2 (Theorem 2) and in paragraph 4. The complete description of τ-models is presented in Theorem 3.

Theorem 2. For τ-models the following statements are equivalent:

1) the policy $\hat{\pi}(x)$ is optimal;

2) the $F_{(t-\tau\mp 1)v_0}$-measurable function $\hat{a}_t^x = \hat{\pi}(x) \in \Pi$ satisfies equality (1) with $P^{\hat{\pi}(x)}$-probability 1.

Corollary 1. If equality (1) in some τ-model has a measurable choice then there exists an optimal Markov policy.

Corollary 2. Let us assume that $\tau = 0$, X is a compact, the functions R, S are bounded from above and conditions C1, C2′, C3 are satisfied. Then Z is the τ-model and the controls $\hat{a}_t^x (= \hat{\pi}(x))$ are optimal if and only if they satisfy the Bellman's equation:

$$\begin{cases} 0 = \sup\limits_{a \in A} \{R(t, \xi_{t-1}^x, a) - F_{t-1}(\xi_{t-1}^x) + \int\limits_X F_t(y) P_a(t, \xi_{t-1}^x, dy)\} = \\[2ex] = R(t, \xi_{t-1}^x, \hat{a}_t^x) - F_{t-1}(\xi_{t-1}^x) + \int\limits_X F_t(y) P_{\hat{a}_t^x}(t, \xi_{t-1}^x, dy); \\[2ex] F_T(y) = S(y) \end{cases} \tag{2}$$

with $P^{\hat{\pi}(x)}$-probability 1.

Let $\pi(x)$ be an arbitrary policy on $[1, t]$ and $\pi^m(x)\,(= a_\theta^m)$ be a Markov policy on $[t+1, T]$. The symbol $W_t\,(\pi(x), \pi^m(x), (y)_0^{(t-\tau)v_0})$ will denote the random variable:

$$W_t(\pi(x), \pi^m(x), (y)_0^{(t-\tau)v_0}) \triangleq M^{\langle \pi(x), \pi^m(x) \rangle}$$

$$\left[\sum_{\theta = t+1}^{T} R(\theta, \xi_{\theta-1}^x, a_\theta^m) + S(\xi_T^x) | \mathscr{F}_{(t-\tau)v_0} \right].$$

Here

$$\langle \pi(x), \pi^m(x) \rangle \;\; [(y)_0^{(\tilde{t}-\tau-1)v_0}] \triangleq \begin{cases} \pi[(y)_0^{(\tilde{t}-\tau-1)v_0}], & \tilde{t} \leqq t; \\ \pi^m[(y)_0^{(\tilde{t}-\tau-1)v_0}], & \tilde{t} > t \end{cases}$$

is the natural composition of the policies $\pi(x)$ and $\pi^m(x)$.

Let

$$V_t(\pi(x), (y)_0^{(t-\tau)v_0}) \triangleq \sup_{\pi^m(x) \in \Pi_t^m} W_t(\pi(x), \pi^m(x), (y)_0^{(t-\tau)v_0})$$

where the supremum is to be taken (for every trajectory $(y)_0^{(t-\tau)v_0}$) on the class of all Markov policies on $[t+1, T]$.

Assume that the following condition is satisfied:

C4. The random variable $V_t(\pi(x), (y)_0^{(t-\tau)v_0})$ is the functional on B_t depending on the conditional measure $P_t^{\pi(x)}\,(\cdot\,, (y)_0^{(t\;\;\tau)v_0})$: $G_t(P_t^{\pi(x)}) \triangleq V_t(\pi(x), (y)_0^{(t-\tau)v_0})$ only.

Remark. It is obvious that the conditional measure $P_t^{\pi(x)}(\cdot, (y)_0^{(t-\tau)\vee 0})$ depends only on $\xi^x_{(t-\tau)\vee 0} \in X$ and on the control parameters' set

$$\{a^x_{(t-\tau)\vee 0}, a^x_{(t-\tau+1)\vee 0}, \dots, a^x_t\} \in \underbrace{A \times A \times \dots \times A}_{(\tau-1)\wedge t} \triangleq A^{(\tau-1)\wedge t}.$$

So condition C4 is satisfied if the mapping

$$\{\xi^x_{(t-\tau)\vee 0}, a^x_{(t-\tau)\vee 0}, a^x_{(t-\tau+1)\vee 0}, \dots, a^x_t\} \to P_t^{\pi(x)}$$

is reversible for all $\pi(x)$. For example, let the controlled process ξ^x_t be a martingale, defined by the equality $\xi^x_t = \xi^x_{t-1} + \eta_t + a_t z^t \zeta_t$ where $\{\eta_t\}$ and $\{\zeta_t\}$ are some independent sequences of independent random variables with the means being equal to zero and the dispersions being equal to one; $A = \{0, 1\}$. If $z^2 \leqq \frac{1}{2}$ then condition C4 is satisfied.

Furthermore, C4 is satisfied if $\tau = 0$; besides that the example of nontrivial τ-model for which condition C4 is satisfied, is presented in paragraph 4.

Theorem 3. Let the functions R, S be bounded from above. Then the following statements are equivalent:

1) condition C4 is satisfied and there exists a measurable function $g_t(z)$ such that $G_t(P_t^{\pi(x)}) = \int_X g_t(z) P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)\vee 0})$;

2) the model Z is the τ-model.

Remark. The sufficient conditions for the existence of the function $g_t(z)$ are given in Theorem 4.

Corollary. Let D be the set of functions satisfying condition 1) of Theorem 3. Then the model Z is the τ-model for the function $F_t(y)$ (see Hypothesis H) if and only if $F_t(y) = g_t(y) + \varphi_t(x)$, where $g_t(y) \in \varphi_t(x)$ is an arbitrary measurable function depending on the initial state and satisfying the condition $\varphi_T(x) = 0$.

The symbol $\mathscr{R}$ will denote the space of Radon-measures on $(X, \mathscr{B}(X))$ [9].

If X is a compact, then the symbol $C^*(X)$ will denote the space of bounded Radon-measures (the linear space which is conjugate to the Banach-space $C(X)$ of continuous functions with uniform convergement topology). Remember that a sequence $\{r_n\} \in C^*(X)$ is called weakly* converging to the element $r \in C^*(X)$ if $\lim_{n\to\infty} \int_X f(y) r_n(dy) = \int_X f(y) r(dy)$ for each function $f \in C(X)$.

In order to formulate Theorem 4 we shall need the following condition:

C5. a) The elements of B_t are measures which are absolutely continuous relatively to some σ-finite measure $\mu(\cdot)$ on the space $(X, \mathscr{B}(X))$: $P_t(\Gamma) = \int_\Gamma p_t(y)\mu(dy)$; in this case $\Psi(P_t) \triangleq p_t(y)$.

b) There exists an ideal space $E(X, \mu)$ [12] such that $\forall\ P_t \in B_t\, \Psi(P_t) \in E(X, \mu)$. (For example the spaces $L_q(X, \mu)$, $1 \leqq q \leqq \infty$ are ideal.)

Theorem 4. Let condition C4 be satisfied for all $t \in N_T$ and the functions R, S be bounded from above. Assume that $B_t \subset \mathscr{R}$ and every finite measure's subset of B_t is linearly independent. Then Z is a τ-model if one of the following conditions is satisfied:

a) The set B_t is a finite set of bounded Radon-measures.

b) The space X is a compact; $B_t \subset C^*(X)$; the functional $G_t(P_t^{\pi(x)})$ can be extended into some linear functional $\bar{G}_t(r)$ on $C^*(X)$ and the set $N(\bar{G}_t) \triangleq \{r \in C^*(X) : \bar{G}_t(r) = 0\}$ is weakly* closed.

c) Condition C5 is satisfied; the functional $\tilde{G}_t(\Psi(P_t^{\pi(x)})) \triangleq G_t(P_t^{\pi(x)})$ on the subset $\Psi(B_t) \subset E(X, \mu)$ can be extended into some linear functional $\bar{G}_t(r)$ on the ideal space $E(X, \mu)$ which has the following property: if the sequence $\{r_n\} \in E(X, \mu)$ converges to zero by the measure μ on every finite-measure-set and $\mu\{y : |r_n(y)| > f(y)\} = 0$ $(n = 1, 2, \ldots)$ for some function $f(y) \in E(X, \mu)$ ---- then $\bar{G}_t(r_n) \xrightarrow[n \to \infty]{} 0$.

Remark. Theorem 4 needs the condition of linear independence of all finite measure's subsets of B_t. In the case when this condition is not satisfied the following reasoning can be useful. Let us assume that condition C4 is satisfied for all $t \in N_T$, the functions R, S are bounded from above; $B_t \subset C^*(X)$. Let

$$G_\theta(P_\theta^{\pi(x)}) = \int_X g_\theta(z) P_\theta^{\pi(x)}(dz, (y)_0^{(\theta - \tau)v_0})$$

for all $\theta > t$. (This assumption is true for $t = T - 1$.) Let us take an arbitrary linearly dependent measure's set $B_t^L = \{P_t^{(1)}, P_t^{(2)}, \ldots, P_t^{(L)}\} \subseteq B_t$ such that $\forall\ l \leqq L$ the set $B_t^L \setminus P_t^{(l)}$ is linearly independent. According to [9] the functional $G_t(P_t^{\pi(x)})$ will be linear on the space $\bar{B}_t^L$ if the set of continuous functions

$$f(P_t^{\pi(x)}, a) \triangleq \int_X [R(t+1, \tilde{z}, a) + \int_X g_{t+1}(z) P_a(t+1, \tilde{z}, dz)] P_t^{\pi(x)}(d\tilde{z}, (y)_0^{(t - \tau)v_0})$$

is right-filtered. (The control parameter $a \in A$ is the filter parameter.) Note that in the case $L = 3$ the condition having been formulated is also necessary for the linearity of the functional $G_t(P_t^{\pi(x)})$ on the space $\bar{B}_t^L$.

It follows from the proof of Theorem 4 that if the previous reasoning is correct for all linearly dependent sets $B_t^L \subseteq B_t$ and one of conditions a), b), c) of Theorem 4 is satisfied then there exists a function $g_t(z)$ such that

$$G_t(P_t^{\pi(x)}) = \int_X g_t(z) P_t^{\pi(x)}(dz, (y)_0^{(t - \tau)v_0}).$$

According to Theorem 3 if these reasonings are correct for the moments $t = T - 2$, $T - 3, \ldots, 0$ then Z is a τ-model.

4. Example. The linear system with square criterion

Let

$$X=A=R^1; \quad R(t,y,a)=K(t)a^2+M(t)y^2; \quad S(y)=Gy^2,$$

where $G, K(t), M(t)<0$. The process ξ_t^x is defined by the equality

$$\xi_t^x=A(t)\xi_{t-1}^x+B(t)a_t+C(t)\eta_t,$$

where $\{\eta_t\}$ is a sequence of independent standard Gauss random variables.

As it was mentioned all the results presented above are true for the case of varying delay $\tau=\tau(t)$. Contrary to [8] (the case $\tau=\text{const}$ was investigated) we assume that $\tau(t)$ is an arbitrary function with values in the natural numbers' set.

Let

$$D(t)\triangleq\begin{bmatrix} M(t)B^2(t)+A^2(t)K(t) & M(t)K(t) \\ B^2(t) & K(t) \end{bmatrix}$$

be 2×2-matrix;

$$g(t)\triangleq\frac{(1,0)\prod\limits_{\theta=t+1}^{T} D(\theta)\,(G,1)^T}{(0,1)\prod\limits_{\theta=t+1}^{T} D(\theta)\,(G,1)^T};$$

$$q(t)\triangleq\sum_{\theta=t+1}^{T}\{g(\theta)\times C^2(\theta)+$$

$$+\frac{g(\theta)A(\theta)B(\theta)}{k(\theta)+g(\theta)B^2(\theta)}\sum_{s=(\theta-\tau(\theta))v_1}^{\theta-1} C^2(s)\prod_{r=s+1}^{\theta-1} A^2(r)\}; \quad F_t(y)\triangleq g(t)y^2+q(t).$$

One can easily prove that in this model Hypothesis H is true for $F_t(y)$ and $f(t,x)=0$. The unique $\mathscr{F}_{(t-\tau(t)-)v_0}$-measurable control policy satisfying equality (1) is the following

$$\hat{a}_t^x=\hat{\pi}[(y)_0^{(t-\tau(t)-1)v_0}]=-\frac{g(t)A(t)B(t)\bar{y}_{t-1}}{K(t)+g(t)B^2(t)}, \tag{3}$$

where $\bar{y}_{t-1}=M^{\hat{\pi}(x)}[\xi_{t-1}^x\,|\,\mathscr{F}_{(t-\tau(t)-1)v_0}]^{\cdot}$

According to Theorem 2, policy (3) is optimal. One can calculate $\bar{y}_{t-1}$ with the help of the following formula:

$$\bar{y}_{t-1}=\xi^x_{(t-\tau(t)-1)v_0}\cdot\prod_{s=(t-\tau(t))v_1}^{t-1} A(s)+\sum_{s=(t-\tau(t))v_1}^{t-1} B(s)\hat{a}_s^x\times\prod_{\theta=s+1}^{t-1} A(\theta).$$

5. Proof of Theorem 1

1) Let $\varphi_t(y, a) \triangleq R(t, y, a) - F_{t-1}(y) + \int_X F_t(z) P_a(t, y, dz)$. According to conditions C1, C2 $\varphi_t(y, a)$ is uniformly upper semi-continuous (with respect to a). Let us show that the function $\Phi((y)_0^{(t-\tau-1)v_0}, a) \triangleq M^{\pi(x)}[\varphi_t(\xi_{t-1}^x, a) | \mathscr{F}_{(-\tau-1)v_0}]$ is upper semi-continuous (with respect to a). Really, $\forall\, a \in A\; \forall\, \varepsilon > 0\; \exists\, \delta > 0$

$$\forall\, \tilde{a} \in A \rho_A(\tilde{a}, a) < \delta \rightarrow \varphi_t(y, a) - \varphi_t(y, \tilde{a}) < \varepsilon \rightarrow \Phi((y)_0^{(t-\tau-1)v_0}, a) -$$

$$- \Phi((y)_0^{(t-\tau-1)v_0}, \tilde{a}) = \int_X [\varphi_t(z, a) - \varphi_t(z, \tilde{a})] P_{t-1}^{\pi(x)}(dz, (y)_0^{(t-\tau-1)v_0}) < \varepsilon.$$

So, for each trajectory $(y)_0^{(t-\tau-1)v_0}$ there exists the finite $\sup\limits_{a \in A} \Phi((y)_0^{(t-\tau-1)v_0}, a)$ which is achieved on the compact A.

2) Let us prove this point with the help of the mathematical induction method. For $T=1$ the statement is true according to the definition of supremum. Let statement 2) be true for $T-1$ and $\hat{\pi}(x) = \hat{a}_t^x$ be the corresponding Markov policy ($t = 1, 2, \ldots, T-1$). It is sufficient to build a Markov control $\hat{a}_T^x$ depending only upon the conditional measure $P_{T-1}^{\hat{\pi}(x)}(\cdot, (y)_0^{(T-\tau-1)v_0})$ and satisfying the inequality

$$M^{\hat{\pi}(x)}[\varphi_T(\xi_{T-1}^x, \hat{a}_T^x) | \mathscr{F}_{(T-\tau-1)v_0}] >$$

$$> \sup_{a \in A} \{M^{\hat{\pi}(x)}[\varphi_T(\xi_{T-1}^x, a) | \mathscr{F}_{(T-\tau-1)v_0}]\} - \varepsilon.$$

Let

$$A^\varepsilon \triangleq \{(a, P_{T-1}^{\hat{\pi}(x)}) \in A \times B_{T-1} : M^{\hat{\pi}(x)}[\varphi_T(\xi_{T-1}^x, a) | \mathscr{F}_{(T-\tau-1)v_0}] >$$

$$> \sup_{a \in A} M^{\hat{\pi}(x)}[\varphi_T(\xi_{T-1}^x, a) | \mathscr{F}_{(T-\tau-1)v_0}] - \varepsilon\}.$$

It is obvious that A^ε is a Borel space [2]. The control policy $\hat{a}_t^x$, $t < T$ corresponds to the stochastic measure $\bar{P}$ on the set of all measures $P_{T-1}^{\hat{\pi}(x)}(\cdot, (y)_0^{(T-\tau-1)v_0}) \in B_{T-1}$. Using the Yankov–Neiman's lemma [2] for the projection $A^\varepsilon \rightarrow B_{T-1}$ we obtain that there exists such a measurable reflection $\hat{\pi}(P_{T-1}^{\hat{\pi}(x)}) = \hat{a}_T^x$ that $(\hat{a}_T^x, P_{T-1}^{\hat{\pi}(x)}) \in A^\varepsilon$ with $\bar{P}$-probability 1. So $(\hat{a}_T^x, P_{T-1}^{\hat{\pi}(x)}) \in A^\varepsilon$ with $P^{\hat{\pi}(x)}$-probability 1 and the Markov control $\hat{a}_T^x$ is built.

The proof of point 3) coincides with the proof presented above by $\varepsilon = 0$.

6. Proof of Theorem 2

Lemma. Let $F_t(x)$ be an arbitrary measurable function for which the integral $\int_X F_\theta(y)P_a(\theta, z, dy)$ exists for all $a \in A$, $z \in X$, $\theta \in N_T$. Then for each policy $\pi(x) \in \Pi$ the following equality is true:

$$M^{\pi(x)}\Bigg[\sum_{\theta=t+1}^{T} R(\theta, \xi^x_{\theta-1}, a^x_\theta) + S(\xi^x_T) | \mathscr{F}_{(t-\tau)v_0}\Bigg] =$$

$$= \int_X F_t(z)P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v_0}) + M^{\pi(x)}\Bigg[\sum_{\theta=t+1}^{T} M^{\pi(x)}[R(\theta, \xi^x_{\theta-1}, a^x_\theta) - F_{\theta-1}(\xi^x_{\theta-1}) +$$

$$+ \int_X F_\theta(y) \times P_{a^x_\theta}(\theta, \xi^x_{\theta-1}, dy) | \mathscr{F}_{(\theta-\tau-1)v_0}] + S(\xi^x_T) - F_T(\xi^x_T) | \mathscr{F}_{(t-\tau)v_0}\Bigg].$$

Furthermore if the right-hand expectations exist for each $t \in N_T$ then $\pi(x) \in \Pi$.

Proof. Let all the right-hand expectations exist. Then

$$\int_X F_t(z)P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v_0}) + M^{\pi(x)}\Bigg[\sum_{\theta=t+1}^{T} M^{\pi(x)}[R(\theta, \xi^x_{\theta-1}, a^x_\theta) -$$

$$- F_{\theta-1}(\xi^x_{\theta-1}) + \int_X F_\theta(y)P_{a^x_\theta}(\theta, \xi^x_{\theta-1}, dy) | \mathscr{F}_{(\theta-\tau-1)v_0}] + S(\xi^x_T) -$$

$$- F_T(\xi^x_T) | \mathscr{F}_{(t-\tau)v_0}\Bigg] = M^{\pi(x)}\Bigg[\sum_{\theta=t+1}^{T} R(\theta, \xi^x_{\theta-1}, a^x_\theta) + S(\xi^x_T) +$$

$$+ \int_X F_t(z)P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v_0}) + \sum_{\theta=t+1}^{T-1} \{ \int_X F_\theta(y)P_{a^x_\theta}(\theta, \xi^x_{\theta-1}, dy) -$$

$$- F_{\theta-1}(\xi^x_{\theta-1})\} - F_{T-1}(\xi^x_{T-1}) + \int_X F_T(y)P_{a^x_T}(T, \xi^x_{T-1}, dy) - F_T(\xi^x_T) | \mathscr{F}_{(t-\tau)v_0}\Bigg] =$$

$$= M^{\pi(x)}\Bigg[\sum_{\theta=t+1}^{T} R(\theta, \xi^x_{\theta-1}, a^x_\theta) + S(\xi^x_T) + \int_X F_t(z)P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v_0}) +$$

$$+ \sum_{\theta=t+1}^{T-1} \{ \int_X F_\theta(y)P_{a^x_\theta}(\theta, \xi^x_{\theta-1}, dy) - F_{\theta-1}(\xi^x_{\theta-1})\} - F_{T-1}(\xi^x_{T-1}) + M^{\pi(x)}[\int_X F_T(y) \times$$

$$\times P_{a^x_T}(T, \xi^x_{T-1}, dy) - F_T(\xi^x_T) | \mathscr{F}_{T-1}] | \mathscr{F}_{(t-\tau)v_0}\Bigg] = M^{\pi(x)}\Bigg[\sum_{\theta=t+1}^{T} R(\theta, \xi^x_{\theta-1}, a^x_\theta) +$$

$$+ S(\xi^x_T) + \int_X F_t(z)P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v_0}) + \sum_{\theta=t+1}^{T-1} \{ \int_X F_\theta(y) \times$$

$$P_{a_\theta^x}(\theta, \xi_{\theta-1}^x, dy) - F_{\theta-1}(\xi_{\theta-1}^x)\} - F_{T-1}(\xi_{T-1}^x) | \mathscr{F}_{(t-\tau)v_0}\Bigg] = \ldots =$$

$$= M^{\pi(x)}\left[\sum_{\theta=t+1}^{T} R(\theta, \xi_{\theta-1}^x, a_\theta^x) + S(\xi_T^x) | \mathscr{F}_{(t-\tau)v_0}\right]. \qquad \text{q.e.d.}$$

Proof of Theorem 2. Let statement 2) be true. Assume that the $\mathscr{F}_{(t-\tau-1)v_0}$-measurable function $\hat{a}_t^x = \hat{\pi}(x)$ satisfies equality (1) with $P^{\hat{\pi}(x)}$-probability 1. According to the Lemma's statement for $t=0$ and $P_t^{\pi(x)} = \chi_x$ the following inequality is true for each policy $\pi(x) \in \Pi$:

$$W(\pi(x)) = F_0(x) + M^{\pi(x)}\Bigg[\sum_{t=1}^{T} M^{\pi(x)}[R(t, \xi_{t-1}^x, a_t^x) - F_{t-1}(\xi_{t-1}^x) +$$

$$+ \int_X F_t(y) P_{a_t^x}(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}] + S(\xi_T^x) -$$

$$- F_T(\xi_T^x) | \mathscr{F}_0\Bigg] \leqq F_0(x) + \sum_{t-1}^{T} f(t, x).$$

On the other hand $W(\hat{\pi}(x)) = F_0(x) + \sum_{t=1}^{T} f(t, x)$. Hence the policy $\hat{\pi}(x)$ is optimal.

Let statement 2) be not true. So the inequality

$$M^{\hat{\pi}(x)}[R(t, \xi_{t-1}^x, \hat{a}_t^x) - F_{t-1}(\xi_{t-1}^x) + \int_X F_t(y) P_{\hat{a}_t^x}(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}] < f(t, x)$$

is true for some $t \in N_T$ on a nonzero $P^{\hat{\pi}(x)}$-measure set. Hence, according to the Lemma $W(\hat{\pi}(x)) < F_0(x) + \sum_{t=1}^{T} f(t, x)$. Let show that for each $\varepsilon > 0$ there exists such Markov policy $\tilde{\pi}(x) = \tilde{a}_t^x \in \Pi$ that $W(\tilde{\pi}(x)) > F_0(x) + \sum_{t=1}^{T} f(t, x) - \varepsilon$. Really, according to 2) of Theorem 1, there exists a Markov policy $\tilde{\pi}(x) = \tilde{a}_t^x$ such that the inequalities

$$M^{\tilde{\pi}(x)}[R(t, \xi_{t-1}^x, \tilde{a}_t^x) - F_{t-1}(\xi_{t-1}^x) +$$

$$+ \int_X F_t(y) P_{\tilde{a}_t^x}(t, \xi_{t-1}^x, dy) | \mathscr{F}_{(t-\tau-1)v_0}] > f(t, x) - \frac{\varepsilon}{T} \quad (t \in N_T)$$

are true and the left-hand conditional expectations are bounded functions with $P^{\tilde{\pi}(x)}$-probability 1. According to Lemma $\tilde{\pi}(x) \in \Pi$ and $W(\tilde{\pi}(x)) > F_0(x) + \sum_{t=1}^{T} f(t, x) - \varepsilon$. So $W(\hat{\pi}(x)) < \sup_{\pi(x) \in \Pi} W(\pi(x))$, i.e. the policy $\hat{\pi}(x)$ is not optimal. The theorem is proved.

Proof of Corollary 1 follows from the definition of measurable choice (point 3) of Theorem 1).

Proof of Corollary 2. Let us consider the equation

$$F_{t-1}(y)=\sup_{a\in A}\{R(t,y,a)+\int_X F_t(z)P_a(t,y,dz)\}$$

under the condition $F_T(y)=S(y)$. According to the assumptions it has a unique solution. Equality (1) transfers to equality (2) and has a measurable choice. One can find the detailed proofs in [2]. So the model Z is the τ-model. As all the conditions of Theorem 2 are satisfied the statement of Corollary 2 is obviously true.

7. Proof of Theorem 3

Let statement 1) be true. Then the following equalities are true for each policy $\pi(x)$.

$$\int_X g_{t-1}(z)P^{\pi(x)}_{t-1}(dz,(y)_0^{(t-\tau-1)v_0})=G_{t-1}(P^{\pi(x)}_{t-1})=$$

$$=\sup_{\pi^m(x)\in\Pi^m_{t-1}}\Big\{M^{\langle\pi(x),\pi^m(x)\rangle}[R(t,\xi^x_{t-1},a^x_t)|\mathscr{F}_{(t-\tau-1)v_0}]+$$

$$+M^{\langle\pi(x),\pi^m(x)\rangle}\Big[\sum_{\theta=t+1}^{T}R(\theta,\xi^x_{\theta-1},a^x_\theta)+S(\xi^x_T)|\mathscr{F}_{(t-\tau-1)v_0}\Big]\Big\}=$$

$$=\sup_{\pi^m(x)\in\Pi^m_{t-1}}\Big\{M^{\pi(x)}[R(t,\xi^x_{t-1},a^x_t)|\mathscr{F}_{(t-\tau-1)v_0}]+$$

$$+M^{\langle\pi(x),\pi^m(x)\rangle}\Big[M^{\langle\pi(x),\pi^m(x)\rangle}\Big[\sum_{\theta=t+1}^{T}R(\theta,\xi^x_{\theta-1},a^x_\theta)+$$

$$+S(\xi^x_T)|\mathscr{F}_{(t-\tau)v_0}\Big]\Big|\mathscr{F}_{(t-\tau-1)v_0}\Big]\Big\}=\sup_{a^x_t\in A}\{M^{\pi(x)}[R(t,\xi^x_{t-1},a^x_t)|$$

$$|\mathscr{F}_{(t-\tau-1)v_0}]+M^{\pi(x)}[G_t(P_t^{\langle\pi(x),a^x_t\rangle})|\mathscr{F}_{(t-\tau-1)v_0}]\}=$$

$$=\sup_{a\in A}\{M^{\pi(x)}[R(t,\xi^x_{t-1},a)|\mathscr{F}_{(t-\tau-1)v_0}]+M^{\pi(x)}[\int_X g_t(\tilde{z})\times$$

$$\times\{\int_X P^{\pi(x)}\{\xi^x_{t-1}\in dz|\mathscr{F}_{(t-\tau)v_0}\}P_a(t,z,d\tilde{z})\}|\mathscr{F}_{(t-\tau-1)v_0}]\}=$$

$$=\sup_{a\in A}\{M^{\pi(x)}[R(t,\xi^x_{t-1},a)|\mathscr{F}_{(t-\tau-1)v_0}]+M^{\pi(x)}[\int_X g_t(\tilde{z})\times$$

$$\times P_a(t,\xi^x_{t-1},d\tilde{z})|\mathscr{F}_{(t-\tau-1)v_0}]\}\,.$$

So Hypothesis H is true (for the function $g_t(y)$), q.e.d.

Let statement 2) be true. If Hypothesis H is true for some functions $\tilde{F}_t(y)$ and $\tilde{f}(t, x)$ then it is also true for the functions

$$F_t(y) \triangleq \tilde{F}_t(y) + \sum_{\theta=t+1}^{T} \tilde{f}(\theta, x); \quad f(t, x) \triangleq 0.$$

Let us show that

$$\sup_{\pi^m(x) \in \Pi_t^m} W_t(\pi(x), \pi^m(x), (y)_0^{(t-\tau)v0}) = \int_X F_t(z) \times P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v0}).$$

Really, according to the Lemma

$$W_t(\pi(x), \pi^m(x), (y)_0^{(t-\tau)v0}) \leqq \int_X F_t(z) P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v0}).$$

On the other hand, according to point 2) of Theorem 1, there exists a Markov policy $\hat{\pi}^m(x) = \hat{a}_t^x \in \Pi$ such that

$$M^{\langle \pi(x), \hat{\pi}^m(x) \rangle} \left[\sum_{\theta=t+1}^{T} R(\theta, \xi_{\theta-1}^x, \hat{a}_\theta^x) + S(\xi_T^x) | \mathscr{F}_{(t-\tau)v0} \right] >$$

$$> \int_X F_t(z) P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v0}) - \varepsilon.$$

(The proof is analogous to that of the second part of Theorem 2.) Hence

$$\sup_{\pi^m(x) \in \Pi_t^m} W_t(\pi(x), \pi^m(x), (y)_0^{(t-\tau)v0}) =$$

$$= \int_X F_t(z) \times P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v0}) = G_t(P_t^{\pi(x)}).$$

The theorem is proved.

Proof of the Corollary. It has been shown during the proof of Theorem 3 that if $D \neq \emptyset$ then Hypothesis H is true for the function $F_t(y)$ under $f(t, x) = 0$ if and only if $F_t(y) \in D$. Let Hypothesis H be true for a function $F_t(y)$ under some function $f(t, x)$. Then Hypothesis H is true for $\tilde{F}_t(y) = F_t(y) + \sum_{\theta=t+1}^{T} f(\theta, x)$ under $\tilde{f}(t, x) = 0$. Hence

$$F_t(y) = \tilde{F}_t(y) - \sum_{\theta=t+1}^{T} f(\theta, x) = \tilde{F}_t(y) + \varphi_t(x)$$

where

$$\varphi_T(x) = 0; \ \tilde{F}_t(y) \in D.$$

Let $F_t(y)=g_t(y)+\varphi_t(x)$ where $\varphi_T(x)=0$; $g_t(y)\in D$. One can reduce that Hypothesis H is true for $F_t(y)$ under $f(t, x)=\varphi_t(x)-\varphi_{t-1}(x)$ with the help of a direct substitution. The proof is completed.

8. Proof of Theorem 4

a) In this case there exists a biorthogonal system (see Lemma [10]). So the functional $G_t(P_t^{\pi(x)})$ has the form:

$$\int_X g_t(z)P_t^{\pi(x)}(dz, (y)_0^{(t-\tau)v0})$$

where $g_t(z)$ is some continuous function. Note that if X is a compact then $g_t(z)$ is bounded. (According to Theorem 1 the boundedness of the function $g_t(z)$ is one of the conditions involving the existence of optimal Markov policy.)

b) The integral presentation of the functional $G_t(P_t^{\pi(x)})$ follows from Eberlein–Shmuljan's Theorem [11].

c) According to Theorem 1.1 in [12], the linear functional $G_t(P_t^{\pi(x)})$ has the form:

$$\int_X g_t(z)p_t^{\pi(x)}(z, (y)_0^{(t-\tau)v0})\mu(dz)$$

where

$$p_t^{\pi(x)}(\,\cdot\,, (y)_0^{(t-\tau)v0})=\Psi(P_t^{\pi(x)}(\,\cdot\,, (y)_0^{(t-\tau)v0}).$$

In order to complete the proof one must use Theorem 3.

References

1. *Lipčer, R. Sh., Shiryaev, A. N.*, The stochastic processes' statistics. Moscow, Nauka, 1974.
2. *Dynkin, E. B., Juškevič, A. A.*, The controlled Markov processes and their applications. Moscow, Nauka, 1975; English transl.: Springer-Verlag, 1979.
3. *Arkin, V. I., Evstigneev, I. V.*, The stochastic methods of economic's dynamics. Moscow, Nauka, 1979.
4. *Piunovski, A. B., Khametov, V. M.*, On the optimal control of continuously discrete jump processes. Isv. AN SSSR, Techn. Kib., 1983, *3*.
5. *Juškevič, A. A.*, On one policies' class in common controlled Markov models. Teor. Verojatnost. i ejo primen., 1973, *4*.
6. *Feinberg, E. A.*, ε-optimal control of a finite Markov chain by the mean criterion. Teor. Verojatnost. i ejo primen., 1980, *1*.
7. *Pressman, E. L., Sonin, I. M.*, The successive control by imperfect data. Moscow, Nauka, 1982.
8. *Vlasuck, B. A., Shternberg, A. A.*, The structural properties of optimal linear stochastic systems with control's delay. Avtomatika i telemekhanika, 1983, *3*.
9. *Meyer, P. A.*, Probability and potentials. Toronto–London, 1966.
10. *Trenogin, V. A.*, Functional analysis. Moscow, Nauka, 1980.
11. *Yosida, K.*, Functional analysis. Springer-Verlag, Berlin, 1965.
12. *Korotkov, V. B.*, Integral operators. Moscow, Nauka, 1983.

Оптимальное управление с запаздыванием случайными последовательностями

О. В. АЛИСЕЕНКО, В. М. ХАМЕТОВ, А. Б. ПИУНОВСКИЙ

(Москва)

В статье исследуется управляемая цепь Маркова с запаздыванием по управлению, заданная на конечном интервале времени. В подобных моделях марковское свойство может нарушаться даже для простейших стратегий управления. В работе приведены достаточные условия существования оптимальных марковских стратегий. Доказана теорема, дающая необходимые и достаточные условия оптимальности управлений. Метод исследования аналогичен методу динамического программирования.

Основные результаты относятся к случаю, когда изучаемая модель является τ-моделью. В статье приводятся характеристические свойства и условия существования τ-моделей.

В качестве примера рассмотрена задача оптимального управления линейной системой с квадратичным критерием, отличающаяся от исследованных ранее наличием переменного запаздывания в управлении.

О. В. Алисеенко, А. Б. Пиуновский, В. М. Хаметов
Московский институт электронного машиностроения,
каф. исследования операций
СССР, Москва 109028, Б. Вузовский пер., 3/12

DISTRIBUTION PROPERTIES OF FEEDBACK SHIFT REGISTER SEQUENCES*

H. NIEDERREITER

(*Vienna*)

(Received November 10, 1984)

We consider the distribution of s-tuples in output sequences of feedback shift registers with finite field arithmetic. We characterize the s-tuples that can appear, and we prove results on the number of occurrences of a given s-tuple. We also establish bounds for the autocorrelation coefficients of feedback shift register sequences.

1. Introduction and statement of results

A feedback shift register is a switching circuit that consists of a simple loop in which adders, constant multipliers, and delay elements (or flip-flops) are connected. We suppose that the arithmetic of the adders and constant multipliers is that of a finite field F_q with q elements. In many practical applications the feedback shift register will operate with binary arithmetic, that is $q=2$. In this case no constant multipliers are needed in the feedback shift register, since the effect of multiplying by an element of F_2 (i.e. by 1 or 0) can be simulated by a wire connection or a disconnection. For detailed information on feedback shift registers we refer to Lidl and Niederreiter [4, Ch. 8].

The output $y(n)$ of a given feedback shift register after n time units ($n=0, 1, \ldots$) is an element of F_q that can be described algebraically by the recursion

$$y(n+k)=a_{k-1}y(n+k-1)+\ldots+a_0y(n) \qquad \text{for} \quad n=0, 1, \ldots, \tag{1}$$

where k is the number of delay elements in the register and $a_0, \ldots, a_{k-1}$ are fixed elements of F_q that correspond to the constant multipliers in the register. The output sequence $(y(n))$, $n=0, 1, \ldots$, is called a feedback shift register sequence. Such sequences arise in various branches of applied algebra. For instance, the encoding of cyclic codes is equivalent to the generation of feedback shift register sequences (see Lidl and Niederreiter [4, Ch. 9], Peterson and Weldon [8, Ch. 8]), and in cryptography one uses feedback shift register sequences to produce pseudorandom stream ciphers.

* This paper was presented at the International Conference on Applied Discrete Mathematics in Debrecen, September 10–13, 1984.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

In connection with these applications it is of importance to study distribution properties of feedback shift register sequences. Results on such distribution properties yield information on Hamming weights of code words in cyclic codes (see Niederreiter [6]). In cryptography such results are useful for the study of patterns occurring in pseudorandom stream ciphers generated by feedback shift register sequences. We will discuss distribution properties in the general setting of the distribution of s-tuples of output elements.

Let $\mathbf{j}=(j_1, \ldots, j_s)$ be an s-tuple of integers with $0 \leqq j_1 < j_2 < \ldots < j_s$. The s-tuple $\mathbf{j}$ will be fixed throughout the paper. Let $\mathbf{b}=(b_1, \ldots, b_s) \in F_q^s$ and let N be a positive integer. Then for a given feedback shift register sequence $(y(n))$ we let $A(\mathbf{b}; \mathbf{j}; N)$ be the number of $n, 0 \leqq n < N$, with $y(n+j_i)=b_i$ for $1 \leqq i \leqq s$. Our results will provide information on the numbers $A(\mathbf{b}; \mathbf{j}; N)$. The values of these numbers depend very much on how the monomials $x^{j_1}, \ldots, x^{j_s}$ are connected with the characteristic polynomial

$$f(x)=x^k-a_{k-1}x^{k-1}-\ldots-a_0 \in F_q[x] \tag{2}$$

of the sequence $(y(n))$. We can assume w.l.o.g. that $a_0 \neq 0$ in (1), so that $f(0) \neq 0$. The sequence $(y(n))$ is then purely periodic; let d denote the length of the least period of $(y(n))$. It suffices to consider $A(\mathbf{b}; \mathbf{j}; N)$ only for $N \leqq d$. The case that has been studied in greater detail so far is $s=1$. Here Hall [1] treated the case where $N=d$ and the characteristic polynomial $f(x)$ is irreducible over F_q, and Selmer [9] treated the case where $N=d$, $q=2$, and $f(x)$ is a product of two distinct irreducible polynomials over F_2. The general situation in the case $s=1$, namely when q is arbitrary, $f(x)$ is arbitrary, and $N \leqq d$ is arbitrary, was dealt with by Niederreiter [5]. For general $s \geqq 1$ there are some results of Laksov [3] and Selmer [9] on the number of $\mathbf{b} \in F_q^s$ for which $A(\mathbf{b}; \mathbf{j}; d) > 0$ for some $(y(n))$ with characteristic polynomial $f(x)$.

We will assume w.l.o.g. that $(y(n))$ is not the zero sequence and that (1) is the recursion with the least value of k satisfied by $(y(n))$, or equivalently that $f(x)$ is the minimal polynomial of $(y(n))$. Then by [4, Theorem 8.44] the period length d of the sequence is equal to $\operatorname{ord}(f(x))$, where $\operatorname{ord}(f(x))$ is defined as the least positive integer e for which $f(x)$ divides x^e-1. We note that always $d \leqq q^k-1$. If (f) denotes the principal ideal generated by $f=f(x)$ in $F_q[x]$, then the residue class ring $F_q[x]/(f)$ can be considered as a vector space over F_q of dimension k. Let $V(f)$ be the linear subspace of $F_q[x]/(f)$ spanned by the residue classes of $x^{j_1}, \ldots, x^{j_s} \bmod f(x)$. We set

$$m(f)=\dim(V(f)). \tag{3}$$

The significance of the number $m(f)$ is the following. Let $B=B(f)$ be the set of those $\mathbf{b}=(b_1, \ldots, b_s) \in F_q^s$ for which there exists a feedback shift register sequence $(w(n))$ with characteristic polynomial $f(x)$ such that $w(j_i)=b_i$ for $1 \leqq i \leqq s$. Then we will see in Lemma 2 that $\operatorname{card}(B)=q^{m(f)}$. We note that if $\mathbf{b} \notin B$, then for the given sequence $(y(n))$ we have $A(\mathbf{b}; \mathbf{j}; N)=0$ for all N, for if for some n_0 we had $y(n_0+j_i)=b_i$

for $1 \leqq i \leqq s$, then $w(n)=y(n_0+n)$ defines a feedback shift register sequence with characteristic polynomial $f(x)$ and $w(j_i)=b_i$ for $1 \leqq i \leqq s$, a contradiction. Therefore it suffices to consider only the case $\mathbf{b} \in B$ in the sequel. An equivalent description of the set B will be given in Lemma 2.

To state our main results, we need some further notation. For the polynomial $f=f(x)$ we let $M(f)$ be the number of $(c_1, \ldots, c_s) \in F_q^s$ with

$$\gcd(c_1 x^{j_1} + \ldots + c_s x^{j_s}, f(x)) = 1.$$

An explicit formula for $M(f)$ will be established in (10). We set

$$C_N = \begin{cases} q^{k/2} & \text{for} \quad N=d, \\ q^{k/2}\left(\dfrac{2}{\pi}\log d + \dfrac{7}{5}\right) & \text{for} \quad 1 \leqq N < d. \end{cases} \tag{4}$$

We recall the assumption made about the sequence $(y(n))$, namely that $f(x)$ is its minimal polynomial. If we now assume that $f(x)$ has degree $k \geqq 1$, then this implies that $(y(n))$ is not the zero sequence, since the minimal polynomial of the zero sequence is defined to be the constant polynomial $f(x)=1$.

Theorem 1. For $\mathbf{b} \in B$ we have

$$|A(\mathbf{b}; \mathbf{j}; N) - Nq^{-m(f)}| \leqq M(f) C_N q^{-s} + (1 - q^{-m(f)} - M(f) q^{-s}) N \qquad \text{for} \quad 1 \leqq N \leqq d,$$

where $m(f)$, $M(f)$, and C_N are defined as above.

In the special case where $f(x)$ is irreducible over F_q, this result can be improved. We set

$$B_N = \begin{cases} q^{k/2} - \dfrac{d}{1+q^{k/2}} & \text{for} \quad N=d, \\ q^{k/2}\left(\dfrac{2}{\pi}\log d + \dfrac{7}{5}\right) & \text{for} \quad 1 \leqq N < d. \end{cases} \tag{5}$$

Theorem 2. Let $f(x)$ be irreducible over F_q. Then for $\mathbf{b} \in B$ we have

$$|A(\mathbf{b}; \mathbf{j}; N) - Nq^{-m(f)}| \leqq (1 - q^{-m(f)}) B_N$$

for $1 \leqq N \leqq d$, where $m(f)$ and B_N are as defined above.

Some further special cases will be listed in Section 3. In connection with Theorem 2 the following lower bound can be established.

Theorem 3. Let $f(x)$ be irreducible over F_q. Then for every N with $1 \leqq N \leqq d = \operatorname{ord}(f(x))$ there exists a feedback shift register sequence $(y(n))$ with minimal polynomial $f(x)$ and a $\mathbf{b} \in B$ such that

$$|A(\mathbf{b}; \mathbf{j}; N) - Nq^{-m(f)}| \geqq N^{1/2} \left(\frac{(1 - q^{-m(f)})(q^k - N)}{q^{m(f)}(q^k - 1)} \right)^{1/2}.$$

2. Auxiliary results

Let $(y(n))$ be a feedback shift register sequence with minimal polynomial $f(x)$ and let $\mathbf{c}=(c_1, \ldots, c_s)\in F_q^s$. Then it is immediate that

$$z(n; \mathbf{c})=c_1 y(n+j_1)+\ldots+c_s y(n+j_s) \qquad \text{for} \quad n=0, 1, \ldots \tag{6}$$

defines a feedback shift register sequence with characteristic polynomial $f(x)$. The minimal polynomial of this sequence is given by the following result.

Lemma 1. The minimal polynomial of the sequence $(z(n; \mathbf{c}))$, $n=0, 1, \ldots$, in (6) is

$$\frac{f(x)}{\gcd(c_1 x^{j_1}+\ldots+c_s x^{j_s}, f(x))}.$$

Proof. Let

$$F(x)=\sum_{n=0}^{\infty} y(n)x^n, \qquad G(x)=\sum_{n=0}^{\infty} z(n; \mathbf{c})x^n$$

be the generating functions of $(y(n))$ and $(z(n; \mathbf{c}))$, respectively. Then

$$\begin{aligned} G(x) &= \sum_{n=0}^{\infty}\left(\sum_{i=1}^{s} c_i y(n+j_i)\right)x^n=\sum_{i=1}^{s} c_i \sum_{n=0}^{\infty} y(n+j_i)x^n \\ &= \sum_{i=1}^{s} c_i x^{-j_i} \sum_{n=0}^{\infty} y(n+j_i)x^{n+j_i}=\sum_{i=1}^{s} c_i x^{-j_i} \sum_{n=j_i}^{\infty} y(n)x^n \\ &= \sum_{i=1}^{s} c_i x^{-j_i}\left(F(x)-\sum_{n=0}^{j_i-1} y(n)x^n\right). \end{aligned}$$

Since $f(x)$ is the minimal polynomial of $(y(n))$, we have by [4, pp. 418–419] the representation $F(x)=r(x)/f^*(x)$, where $r\in F_q[x]$ of degree $<k$, $f^*(x)=x^k f(1/x)$, and $r^*(x)=x^{k-1}r(1/x)$ is relatively prime to $f(x)$. Thus we obtain

$$f^*(x)G(x)=\sum_{i=1}^{s} c_i p_i(x) \tag{7}$$

with

$$p_i(x)=x^{-j_i}(r(x)-f^*(x)\sum_{n=0}^{j_i-1} y(n)x^n) \qquad \text{for} \quad 1\leqq i\leqq s.$$

Since

$$r(x)-f^*(x)\sum_{n=0}^{j_i-1} y(n)x^n=f^*(x)(F(x)-\sum_{n=0}^{j_i-1} y(n)x^n)\equiv 0 \bmod x^{j_i},$$

it follows that $p_i(x)$ is a polynomial over F_q of degree $<k$. Now

$$p_i^*(x)=x^{k-1}p_i\left(\frac{1}{x}\right)=x^{k-1+j_i}\left(r\left(\frac{1}{x}\right)-f^*\left(\frac{1}{x}\right)\sum_{n=0}^{j_i-1}y(n)x^{-n}\right)$$
$$=x^{j_i}\left(r^*(x)-f(x)\sum_{n=0}^{j_i-1}x^{-n-1}\right)\equiv x^{j_i}r^*(x)\bmod f(x),$$

and so

$$\gcd(c_1p_1^*(x)+\ldots+c_sp_s^*(x),f(x))$$
$$=\gcd((c_1x^{j_1}+\ldots+c_sx^{j_s})r^*(x),\,f(x))$$
$$=\gcd(c_1x^{j_1}+\ldots+c_sx^{j_s},\,f(x))$$

since $\gcd(r^*(x),f(x))=1$. The result of the lemma follows then from (7) and [4, pp. 418–419]. □

For any $g\in F_q[x]$ define

$$K(g)=\{(c_1,\ldots,c_s)\in F_q^s\text{:}c_1x^{j_1}+\ldots+c_sx^{j_s}\equiv 0\bmod g(x)\}. \tag{8}$$

It is clear that $K(g)$ is a linear subspace of F_q^s. Thus

$$\operatorname{card}(K(g))=q^{\dim(K(g))}. \tag{9}$$

We can now establish a formula for the number $M(f)$ defined in Section 1. Let $f_1,\ldots,f_t$ be the distinct monic irreducible factors of f, and for $1\leqq i_1<\ldots<i_u\leqq t$ put

$$d(i_1,\ldots,i_u)=\dim(K(f_{i_1}\ldots f_{i_u}))$$
$$=\dim(K(f_{i_1})\cap\ldots\cap K(f_{i_u})).$$

Then it follows from (9) and the inclusion–exclusion principle of combinatorics that

$$M(f)=q^s+\sum_{u=1}^{t}(-1)^u\sum_{1\leq i_1<\ldots<i_u\leq t}q^{d(i_1,\ldots,i_u)}. \tag{10}$$

For $\mathbf{b}=(b_1,\ldots,b_s)$, $\mathbf{c}=(c_1,\ldots,c_s)\in F_q^s$ we define the inner product

$$\mathbf{b}\cdot\mathbf{c}=b_1c_1+\ldots+b_sc_s.$$

If $K(f)$ is given by (8), then the set B defined in Section 1 can also be described as follows.

Lemma 2. We have

$$B=K(f)^{\perp}=\{\mathbf{b}\in F_q^s\text{:}\,\mathbf{b}\cdot\mathbf{c}=0\qquad\text{for all}\quad\mathbf{c}\in K(f)\}\,.$$

In particular, we have $\operatorname{card}(B)=q^{m(f)}$ with $m(f)$ given by (3).

Proof. If $\mathbf{b}=(b_1, \ldots, b_s) \in B$, then there exists a feedback shift register sequence $(w(n))$ with characteristic polynomial $f(x)$ such that $w(j_i)=b_i$ for $1 \leqq i \leqq s$. For $\mathbf{c}= =(c_1, \ldots, c_s) \in K(f)$ put

$$v(n; \mathbf{c})=c_1 w(n+j_1)+\ldots+c_s w(n+j_s) \qquad \text{for} \quad n=0, 1, \ldots.$$

Let $g(x)$ be the minimal polynomial of $(w(n))$. Then $g(x)$ divides $f(x)$ by [4, Theorem 8.42], and so

$$c_1 x^{j_1}+\ldots+c_s x^{j_s} \equiv 0 \bmod g(x)$$

since $\mathbf{c} \in K(f)$. It follows from Lemma 1 that the minimal polynomial of the sequence $(v(n; \mathbf{c}))$ is the constant polynomial 1. In other words, $(v(n; \mathbf{c}))$ is the zero sequence. Consequently,

$$\mathbf{b} \cdot \mathbf{c}=b_1 c_1+\ldots+b_s c_s=c_1 w(j_1)+\ldots+c_s w(j_s)=v(0; \mathbf{c})=0.$$

Thus we have shown $B \subseteq K(f)^{\perp}$.

Since B and $K(f)^{\perp}$ are linear subspaces of F_q^s, it suffices now to prove that $\dim(B) \geqq \dim(K(f)^{\perp})$. Let φ: $F_q[x] \rightarrow F_q[x]/(f)$ be the canonical homomorphism and define λ: $F_q^s \rightarrow F_q[x]/(f)$ by

$$\lambda(c_1, \ldots, c_s)=\varphi(c_1 x^{j_1}+\ldots+c_s x^{j_s}).$$

Then λ is an F_q-linear mapping whose range is the vector space $V(f)$ in (3) and whose kernel is $K(f)$. Therefore

$$\dim(K(f))=s-m(f). \tag{11}$$

It follows that

$$\dim(K(f)^{\perp})=m(f). \tag{12}$$

Now let $\mathbf{b}=(b_1, \ldots, b_s) \in B$ and let $(w(n))$ be a feedback shift register sequence with characteristic polynomial $f(x)$ such that $w(j_i)=b_i$ for $1 \leqq i \leqq s$. Let

$$\mathbf{w}(n)=(w(n), \quad w(n+1), \ldots, w(n+k-1)) \in F_q^k \qquad \text{for} \quad n=0, 1, \ldots$$

be the state vectors of $(w(n))$. If

$$A=\begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & a_0 \\ 1 & 0 & 0 & \ldots & 0 & a_1 \\ 0 & 1 & 0 & \ldots & 0 & a_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & a_{k-1} \end{bmatrix}$$

is the companion matrix of $f(x)$, then by [4, Lemma 8.12] we have

$$\mathbf{w}(j_i)=\mathbf{w}(0)A^{j_i} \qquad \text{for} \quad 1 \leqq i \leqq s.$$

Multiplying by the k-dimensional vector

$$\mathbf{e}=\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

from the right, we get

$$b_i=\mathbf{w}(0)A^{j_i}\mathbf{e} \qquad \text{for} \quad 1 \leqq i \leqq s.$$

If we let $\mathbf{w}(0)$ run through F_q^k, then we get all feedback shift register sequences with characteristic polynomial $f(x)$. Therefore

$$\dim(B)=\dim(\langle A^{j_1}\mathbf{e}, \ldots, A^{j_s}\mathbf{e}\rangle),$$

where $\langle A^{j_1}\mathbf{e}, \ldots, A^{j_s}\mathbf{e}\rangle$ is the linear subspace of F_q^k spanned by the indicated vectors. Consider the F_q-linear mapping

$$\psi\colon \mathbf{c}=(c_1, \ldots, c_s) \in F_q^s \mapsto c_1 A^{j_1}\mathbf{e}+ \ldots +c_s A^{j_s}\mathbf{e} \in F_q^k.$$

The range of ψ is $\langle A^{j_1}\mathbf{e}, \ldots, A^{j_s}\mathbf{e}\rangle$, hence

$$\dim(B)=\dim(\langle A^{j_1}\mathbf{e}, \ldots, A^{j_s}\mathbf{e}\rangle)=s-\dim(\ker\psi). \tag{13}$$

If $\mathbf{c} \in \ker\psi$, then with

$$C=c_1 A^{j_1}+ \ldots +c_s A^{j_s}$$

we have $C\mathbf{e}=\mathbf{0}$. Thus the first column of the $k \times k$ matrix C is $\mathbf{0}$. Now for $1 \leqq h \leqq k-1$ we have $\mathbf{0}=A^h C\mathbf{e}=C(A^h\mathbf{e})$, and $A^h\mathbf{e}$ is the vector with entry 1 in the $(h+1)$st coordinate and entry 0 elsewhere. It follows that the $(h+1)$st column of C is $\mathbf{0}$. Altogether, C is the zero matrix, and so

$$c_1 x^{j_1}+ \ldots +c_s x^{j_s} \equiv 0 \bmod f(x)$$

since $f(x)$ is the minimal polynomial of A (see [2, Ch. 7]). Thus we have shown $\ker\psi \subseteq K(f)$. From (11), (12), and (13) we get then $\dim(B) \geqq \dim(K(f)^\perp)$. The second part of the lemma follows from (12). □

We need some results about character sums over the finite field F_q. Let χ be a nontrivial additive character of F_q; see Lidl and Niederreiter [4, Ch. 5] for an explicit description of such a character. The quantities C_N and B_N are defined in (4) and (5), respectively.

Lemma 3. If $(w(n))$ is a feedback shift register sequence with minimal polynomial $f(x)$ and least period d, then

$$\left|\sum_{n=0}^{N-1} \chi(w(n))\right| \leqq C_N \qquad \text{for} \quad 1 \leqq N \leqq d.$$

Proof. For $N=d$ this follows from [5, Corollary 3.2] and for $1 \leqq N < d$ from [5, Corollary 3.7]. □

Lemma 4. If $(w(n))$ is a feedback shift register sequence with irreducible minimal polynomial $f(x)$ and least period d, then

$$\left|\sum_{n=0}^{N-1} \chi(w(n))\right| \leqq B_N \qquad \text{for} \quad 1 \leqq N \leqq d.$$

Proof. For $N=d$ this follows from [7, Corollary] and for $1 \leqq N < d$ from [5, Corollary 3.7]. □

3. Proof of Theorems 1 and 2

Let $(y(n))$ be the given feedback shift register sequence with minimal polynomial $f(x)$, let $1 \leqq N \leqq d$, and let $\mathbf{b}=(b_1, \ldots, b_s) \in B$. For $b \in F_q$ let δ_b be the characteristic function of $\{b\}$, i.e. $\delta_b(b)=1$ and $\delta_b(a)=0$ for $a \in F_q$, $a \neq b$. Writing $A=A(\mathbf{b}; \mathbf{j}; N)$ for the sake of brevity, we have then

$$A=\sum_{n=0}^{N-1} \delta_{b_1}(y(n+j_1)) \ldots \delta_{b_s}(y(n+j_s)).$$

Now

$$\delta_b(a)=\frac{1}{q} \sum_{c \in F_q} \chi(c(a-b)) \qquad \text{for all} \quad a \in F_q$$

by [4, p. 192], and so with $\mathbf{c}=(c_1, \ldots, c_s) \in F_q^s$:

$$\begin{aligned} A &= q^{-s} \sum_{n=0}^{N-1} \sum_{c_1 \in F_q} \chi(c_1(y(n+j_1)-b_1)) \ldots \sum_{c_s \in F_q} \chi(c_s(y(n+j_s)-b_s)) \\ &= q^{-s} \sum_{\mathbf{c} \in F_q^s} \bar{\chi}(\mathbf{b} \cdot \mathbf{c}) \sum_{n=0}^{N-1} \chi(z(n; \mathbf{c})), \end{aligned}$$

where $\bar{\chi}$ is the conjugate character of χ and $z(n; \mathbf{c})$ is defined by (6). We split up the outer sum according as

$$\begin{aligned} \mathbf{c} \in K(f) &= \{(c_1, \ldots, c_s) \in F_q^s : c_1 x^{j_1} + \ldots + c_s x^{j_s} \equiv 0 \bmod f(x)\}, \\ \mathbf{c} \in J(f) &= \{(c_1, \ldots, c_s) \in F_q^s : \gcd(c_1 x^{j_1} + \ldots + c_s x^{j_s}, f(x))=1\}, \\ \mathbf{c} \in L(f) &= F_q^s \backslash (K(f) \cup J(f)). \end{aligned}$$

Since for $\mathbf{c} \in K(f)$ we have $z(n; \mathbf{c}) = 0$ for all n by Lemma 1, we obtain

$$\begin{aligned} A = Nq^{-s} \sum_{\mathbf{c} \in K(f)} \bar{\chi}(\mathbf{b} \cdot \mathbf{c}) + q^{-s} \sum_{\mathbf{c} \in J(f)} \bar{\chi}(\mathbf{b} \cdot \mathbf{c}) \sum_{n=0}^{N-1} \chi(z(n; \mathbf{c})) \\ + q^{-s} \sum_{\mathbf{c} \in L(f)} \bar{\chi}(\mathbf{b} \cdot \mathbf{c}) \sum_{n=0}^{N-1} \chi(z(n; \mathbf{c})) . \end{aligned} \tag{14}$$

Furthermore, for $\mathbf{b} \in B$ we have $\mathbf{b} \cdot \mathbf{c} = 0$ for all $\mathbf{c} \in K(f)$ by Lemma 2, and also card $(K(f)) = q^{s-m(f)}$ by (11), thus

$$A = Nq^{-m(f)} + T_2 + T_3 , \tag{15}$$

where T_2 (resp. T_3) is the second (resp. third) term on the right-hand side of (14). For $\mathbf{c} \in J(f)$ the minimal polynomial of the sequence $(z(n; \mathbf{c}))$ is equal to $f(x)$ by Lemma 1, hence

$$| T_2 | \leqq q^{-s} \sum_{\mathbf{c} \in J(f)} \left| \sum_{n=0}^{N-1} \chi(z(n; \mathbf{c})) \right| \leqq M(f) C_N q^{-s}$$

by Lemma 3 and the fact that card $(J(f)) = M(f)$. For $\mathbf{c} \in L(f)$ we use the trivial bound

$$\left| \sum_{n=0}^{N-1} \chi(z(n; \mathbf{c})) \right| \leqq N$$

and the fact that

$$\text{card}\,(L(f)) = q^s - q^{s-m(f)} - M(f) .$$

This yields

$$| T_3 | \leqq (1 - q^{-m(f)} - M(f) q^{-s}) N .$$

By combining these bounds for T_2 and T_3 with (15), we obtain Theorem 1. □

To prove Theorem 2, we go back to (15) and note that for irreducible $f(x)$ the set $L(f)$ is empty, so that $T_3 = 0$. For $\mathbf{c} \in J(f)$ the minimal polynomial of the sequence $(z(n; \mathbf{c}))$ is equal to $f(x)$ by Lemma 1, hence

$$| T_2 | \leqq q^{-s} \sum_{\mathbf{c} \in J(f)} \left| \sum_{n=0}^{N-1} \chi(z(n; \mathbf{c})) \right| \leqq q^{-s} (q^s - q^{s-m(f)}) B_N$$

by Lemma 4 and the fact that card $(J(f)) = q^s -$ card $(K(f)) = q^s - q^{s-m(f)}$. The result of Theorem 2 follows immediately. □

We consider now some interesting special cases of Theorems 1 and 2. Suppose we have

$$\gcd(c_1 x^{j_1} + \ldots + c_s x^{j_s},\ \ f(x)) = 1 \qquad \text{for all} \quad (c_1, \ldots, c_s) \neq \mathbf{0} . \tag{16}$$

This condition is, for instance, satisfied if j_s is less than the degree of any irreducible factor of $f(x)$.

Corollary 1. Under condition (16) we have

$$|A(\mathbf{b};\mathbf{j};N)-Nq^{-s}|\leqq(1-q^{-s})C_N$$

for $1\leqq N\leqq d$ and all $\mathbf{b}\in F_q^s$. If $f(x)$ is irreducible over F_q, we can replace C_N by B_N.

Proof. Condition (16) implies that in (3) we have $m(f)=s$. Furthermore, (16) implies that $M(f)=q^s-1$. The result follows now from Theorems 1 and 2. □

Next we consider the case where $f(x)$ is a power of an irreducible polynomial, say $f(x)=g(x)^e$ with a monic irreducible polynomial $g(x)$ over F_q and $e\geqq 2$.

Corollary 2. If $f(x)$ is a power of the irreducible polynomial $g(x)$, then for $\mathbf{b}\in B$ we have

$$|A(\mathbf{b};\mathbf{j};N)-Nq^{-m(f)}|\leqq(1-q^{-m(g)})C_N+(q^{-m(g)}-q^{-m(f)})N$$

for $1\leqq N\leqq d$.

Proof. By formula (10) we have

$$M(f)=q^s-q^{\dim(K(g))},$$

and by (11) we have $\dim(K(g))=s-m(g)$. The result follows now from Theorem 1. □

4. Proof of Theorem 3

First let $(y(n))$ be a fixed feedback shift register sequence with irreducible minimal polynomial $f(x)$. Using the notation from Section 3 and writing $m=m(f)$, $J=J(f)$, and $K=K(f)$, we get from (15) for $1\leqq N\leqq d=\operatorname{ord}(f(x))$,

$$\begin{aligned}
S:&=\sum_{\mathbf{b}\in B}|A(\mathbf{b};\mathbf{j};N)-Nq^{-m}|^2\\
&=q^{-2s}\sum_{\mathbf{b}\in B}\left|\sum_{\mathbf{c}\in J}\sum_{n=0}^{N-1}\chi(z(n;\mathbf{c})-\mathbf{b}\cdot\mathbf{c})\right|^2\\
&=q^{-2s}\sum_{\mathbf{b}\in B}\sum_{\mathbf{c},\mathbf{c}'\in J}\sum_{n,p=0}^{N-1}\chi(z(n;\mathbf{c})-z(p;\mathbf{c}')-\mathbf{b}\cdot(\mathbf{c}-\mathbf{c}'))\\
&=q^{-2s}\sum_{\mathbf{c},\mathbf{c}'\in J}\sum_{n,p=0}^{N-1}\chi(z(n;\mathbf{c})-z(p;\mathbf{c}'))\sum_{\mathbf{b}\in B}\chi(\mathbf{b}\cdot(\mathbf{c}'-\mathbf{c})).
\end{aligned}$$

If $\mathbf{c}'-\mathbf{c}\in K$, then it follows from Lemma 2 that

$$\sum_{\mathbf{b}\in B}\chi(\mathbf{b}\cdot(\mathbf{c}'-\mathbf{c}))=\operatorname{card}(B)=q^m.$$

If $\mathbf{c}'-\mathbf{c}\notin K$, then there exists a $\mathbf{b}\in B$ with $\mathbf{b}\cdot(\mathbf{c}'-\mathbf{c})\neq 0$. In fact, since B is a linear subspace of F_q^s and χ is a nontrivial character, there exists $\mathbf{b}\in B$ with $\chi(\mathbf{b}\cdot(\mathbf{c}'-\mathbf{c}))\neq 1$.

Thus $\chi(\mathbf{b}\cdot(\mathbf{c}'-\mathbf{c}))$, considered as a function of $\mathbf{b}$, is a nontrivial character of the additive group B, and so

$$\sum_{\mathbf{b}\in B}\chi(\mathbf{b}\cdot(\mathbf{c}'-\mathbf{c}))=0$$

by [4, Theorem 5.4]. It follows that

$$S=q^{m-2s}\sum_{\substack{\mathbf{c},\mathbf{c}'\in J\\ \mathbf{c}'-\mathbf{c}\in K}}\sum_{n,p=0}^{N-1}\chi(z(n;\mathbf{c})-z(p;\mathbf{c}')).$$

For $\mathbf{c}'-\mathbf{c}\in K$, say $\mathbf{c}'=(c_1',\ldots,c_s')$ and $\mathbf{c}=(c_1,\ldots,c_s)$, we have

$$z(p;\mathbf{c}')-z(p;\mathbf{c})=\sum_{i=1}^{s}(c_i'-c_i)y(p+j_i)=0$$

for all p by Lemma 1, hence

$$S=q^{m-2s}\sum_{\substack{\mathbf{c},\mathbf{c}'\in J\\ \mathbf{c}'-\mathbf{c}\in K}}\sum_{n,p=0}^{N-1}\chi(z(n;\mathbf{c})-z(p;\mathbf{c}))$$

$$=q^{m-2s}\sum_{\substack{\mathbf{c},\mathbf{c}'\in J\\ \mathbf{c}'-\mathbf{c}\in K}}\left|\sum_{n=0}^{N-1}\chi(z(n;\mathbf{c}))\right|^2.$$

Since the terms of this sum do not depend on $\mathbf{c}'$, and since for each $\mathbf{c}\in J$ there are $\operatorname{card}(K)=q^{s-m}$ choices of $\mathbf{c}'$ by (11), we obtain

$$S=q^{-s}\sum_{\mathbf{c}\in J}\left|\sum_{n=0}^{N-1}\chi(z(n;\mathbf{c}))\right|^2. \tag{17}$$

We use now an explicit formula for the terms of the sequence $(y(n))$. Let F be the extension field of F_q of degree k, i.e. the finite field with q^k elements, let F^* be the group of nonzero elements of F, let $\alpha\in F^*$ be a fixed root of $f(x)$, and let $\operatorname{Tr}: F\rightarrow F_q$ be the trace function from F onto F_q (see [4, Ch. 2]). Then by [4, Theorem 8.24] there exists a uniquely determined $\theta\in F^*$ such that

$$y(n)=\operatorname{Tr}(\theta\alpha^n)\qquad\text{for all}\quad n. \tag{18}$$

If we let θ run through F^* in formula (18), then $(y(n))$ runs through all the q^k-1 feedback shift register sequences with minimal polynomial $f(x)$. We write now $S(\theta)$ for the expression S in (17) to indicate its dependence on θ. For $\mathbf{c}=(c_1,\ldots,c_s)\in J$ we have

$$z(n;\mathbf{c})=\sum_{i=1}^{s}c_iy(n+j_i)=\sum_{i=1}^{s}c_i\operatorname{Tr}(\theta\alpha^{n+j_i})=\operatorname{Tr}(\theta\beta(\mathbf{c})\alpha^n)$$

with

$$\beta(\mathbf{c})=c_1\alpha^{j_1}+\ldots+c_s\alpha^{j_s}\neq 0.$$

Therefore from (17),

$$S(\theta)=q^{-s}\sum_{\mathbf{c}\in J}\left|\sum_{n=0}^{N-1}\chi(\mathrm{Tr}\,(\theta\beta(\mathbf{c})\alpha^n))\right|^2$$
$$=q^{-s}\sum_{\mathbf{c}\in J}\left|\sum_{n=0}^{N-1}\mu(\theta\beta(\mathbf{c})\alpha^n)\right|^2,$$

where $\mu(\gamma)=\chi(\mathrm{Tr}(\gamma))$ for $\gamma\in F$ is a nontrivial additive character of F. Thus we have

$$T:=\sum_{\theta\in F^*}S(\theta)=q^{-s}\sum_{\theta\in F^*}\sum_{\mathbf{c}\in J}\left|\sum_{n=0}^{N-1}\mu(\theta\beta(\mathbf{c})\alpha^n)\right|^2$$
$$=q^{-s}\sum_{\mathbf{c}\in J}\sum_{\theta\in F^*}\left|\sum_{n=0}^{N-1}\mu(\theta\beta(\mathbf{c})\alpha^n)\right|^2$$
$$=q^{-s}\sum_{\mathbf{c}\in J}\sum_{\theta\in F^*}\left|\sum_{n=0}^{N-1}\mu(\theta\alpha^n)\right|^2,$$

since $\beta(\mathbf{c})\neq 0$ implies that $\theta\beta(\mathbf{c})$ runs through F^* with θ. The inner sum is independent of $\mathbf{c}$ and we have card $(J)=q^s-q^{s-m}$ (see the proof of Theorem 2), hence

$$T=(1-q^{-m})\sum_{\theta\in F^*}\left|\sum_{n=0}^{N-1}\mu(\theta\alpha^n)\right|^2.$$

Now

$$\sum_{\theta\in F^*}\left|\sum_{n=0}^{N-1}\mu(\theta\alpha^n)\right|^2=\sum_{\theta\in F^*}\sum_{n,p=0}^{N-1}\mu(\theta(\alpha^n-\alpha^p))$$
$$=\sum_{n,p=0}^{N-1}\sum_{\theta\in F^*}\mu(\theta(\alpha^n-\alpha^p)).$$

If $n=p$, then the inner sum has the value q^k-1. If $n\neq p$, then from $0\leqq n,\ p<N\leqq d=$ $=\mathrm{ord}\,(f(x))$ and the fact that the element α has order d in the group F^* by [4, Theorem 3.3] we obtain $\alpha^n\neq\alpha^p$, and so

$$\sum_{\theta\in F^*}\mu(\theta(\alpha^n-\alpha^p))=\sum_{\theta\in F^*}\mu(\theta)=\sum_{\theta\in F}\mu(\theta)-\mu(0)=-1.$$

Therefore

$$\sum_{\theta\in F^*}\left|\sum_{n=0}^{N-1}\mu(\theta\alpha^n)\right|^2=N(q^k-1)-(N^2-N)=N(q^k-N),$$

hence

$$T=(1-q^{-m})N(q^k-N).$$

Going back to the definition of T, it follows that there exists a $\theta \in F^*$, or equivalently a feedback shift register sequence $(y(n))$ with minimal polynomial $f(x)$, such that

$$S \geqq N \frac{(1-q^{-m})(q^k-N)}{q^k-1}.$$

Going back to the definition of S at the beginning of this section and using the fact that card $(B)=q^m$ by Lemma 2, we deduce the existence of a $\mathbf{b} \in B$ such that

$$|A(\mathbf{b}; \mathbf{j}; N)-Nq^{-m}|^2 \geqq N \frac{(1-q^{-m})(q^k-N)}{q^m(q^k-1)}.$$

Taking square roots, we obtain the result of Theorem 3. □

To illustrate Theorem 3, we consider a special case appearing frequently in the applications, namely when $f(x)$ is a primitive polynomial over F_q of degree k, i.e. a monic irreducible polynomial over F_q with the largest possible value q^k-1 of ord $(f(x))$; compare with [4, p. 89]. For odd q we put $N=(q^k+1)/2$ and for even q we put $N=\frac{1}{2}q^k$. Then the lower bound in Theorem 3 is greater than

$$\frac{1}{2} q^{(k-m)/2}(1-q^{-m})^{1/2}.$$

For such $f(x)$ the upper bound in Theorem 2 is equal to

$$(1-q^{-m})q^{k/2}\left(\frac{2}{\pi}\log(q^k-1)+\frac{7}{5}\right).$$

A comparison of these bounds shows that if the upper bound in Theorem 2 is considered as a function of q^k, then it is best possible apart from a logarithmic factor.

5. Autocorrelation

We mention that the auxiliary results in Section 2 lead to bounds for the autocorrelation coefficients of feedback shift register sequences. Let $(y(n))$ again be a feedback shift register sequence with minimal polynomial $f(x)$ of degree $k \geqq 1$ and $f(0) \neq 0$. For positive integers j and N and a nontrivial additive character χ of F_q, the corresponding autocorrelation coefficient is defined by

$$C(j; N)=\sum_{n=0}^{N-1} \chi(y(n))\bar{\chi}(y(n+j)),$$

where $\bar{\chi}$ is the conjugate character of χ. With

$$z(n)=y(n)-y(n+j) \qquad \text{for} \quad n=0,1,\ldots$$

we get then

$$C(j;N)=\sum_{n=0}^{N-1}\chi(y(n)-y(n+j))=\sum_{n=0}^{N-1}\chi(z(n)). \tag{19}$$

It follows from Lemma 1 that the minimal polynomial of $(z(n))$ is given by

$$g(x)=\frac{f(x)}{\gcd(x^j-1,f(x))}.$$

Corollary 3. If $g(x)$ is as above, $D=\operatorname{ord}(g(x))$, and $N=QD+R$ with integers Q and R and $0\leqq R<D$, then

$$|C(j;N)|\leqq Q\min(D,q^{\deg(g)/2})+\min\left(R,q^{\deg(g)/2}\left(\frac{2}{\pi}\log D+\frac{7}{5}\right)\right).$$

Proof. By (19) we can write

$$C(j;N)=\sum_{n=0}^{QD-1}\chi(z(n))+\sum_{n=QD}^{QD+R-1}\chi(z(n))=Q\sum_{n=0}^{D-1}\chi(z(n))+\sum_{n=0}^{R-1}\chi(z(n)),$$

since $(z(n))$ has period D. Thus

$$|C(j;N)|\leqq Q\left|\sum_{n=0}^{D-1}\chi(z(n))\right|+\left|\sum_{n=0}^{R-1}\chi(z(n))\right|,$$

and the result follows from Lemma 3 and the definition of C_N in (4). □

If $f(x)$ divides x^j-1, then $(z(n))$ is the zero sequence, and so (19) yields $C(j;N)=N$ for all N. If $f(x)$ is now irreducible over F_q, then the only other alternative is $\gcd(x^j-1,f(x))=1$. In this case we can improve Corollary 3. We write again $d=\operatorname{ord}(f(x))$.

Corollary 4. If $f(x)$ is irreducible over F_q, $\gcd(x^j-1,f(x))=1$, and $N=Qd+R$ with integers Q and R and $0\leqq R<d$, then

$$|C(j;N)|\leqq Q\min\left(d,q^{k/2}-\frac{d}{1+q^{k/2}}\right)+\min\left(R,q^{k/2}\left(\frac{2}{\pi}\log d+\frac{7}{5}\right)\right).$$

Proof. We proceed as in the proof of Corollary 3, but use Lemma 4 instead of Lemma 3 and the fact that $g(x)=f(x)$. □

If we consider the autocorrelation over the full period, i.e. with $N=d$, and if we take a maximal period sequence $(y(n))$, i.e. a sequence with a primitive minimal polynomial $f(x)$, then Corollary 4 yields $|C(j;d)|\leqq 1$ whenever $\gcd(x^j-1,f(x))=1$. However, in this case it is well known that $C(j;d)=-1$. Therefore Corollary 4 can be viewed as a generalization of the latter result.

By [4, Lemma 3.6], $f(x)$ divides x^j-1 if and only if $d=\operatorname{ord}(f(x))$ divides j. If $f(x)$ is irreducible over F_q, it follows therefore that the condition $\gcd(x^j-1, f(x))=1$ in Corollary 4 is satisfied if and only if j is not a multiple of the period length d of the sequence $(y(n))$.

In the binary case $q=2$, there is only one nontrivial additive character of $F_2=\{0,1\}$, and it is given by $\chi(0)=1$, $\chi(1)=-1$. The formula for the autocorrelation coefficient reduces then to

$$C(j;N)=\sum_{n=0}^{N-1}(-1)^{y(n)-y(n+j)}.$$

This number can be interpreted as follows: write the shifted sequence $(y(n+j))$ underneath the sequence $(y(n))$ and count the agreements and disagreements among the first N corresponding terms; then $C(j;N)$ is equal to the number of agreements minus the number of disagreements.

References

1. *Hall, M.*, Equidistribution of residues in sequences, Duke Math. J. **4,** 691–695 (1938).
2. *Hoffman, K., Kunze, R.*, Linear Algebra, 2nd ed., Prentice-Hall, Englewood Cliffs, N. J., 1971.
3. *Laksov, D.*, Linear recurring sequences over finite fields, Math. Scand. **16,** 181–196 (1965).
4. *Lidl, R., Niederreiter, H.*, Finite Fields, Encyclopedia of Math. and Its Appl., vol. **20,** Addison-Wesley, Reading, Mass., 1983.
5. *Niederreiter, H.*, On the cycle structure of linear recurring sequences, Math. Scand. **38,** 53–77 (1976).
6. *Niederreiter, H.*, Weights of cyclic codes, Information and Control **34,** 130–140 (1977).
7. *Niederreiter, H.*, Exponential sums over finite fields, Math. J. Okayama Univ., to appear.
8. *Peterson, W. W., Weldon, E. J.*, Error-Correcting Codes, 2nd ed., M. I. T. Press, Cambridge, Mass., 1972.
9. *Selmer, E. S.*, Linear Recurrence Relations over Finite Fields, Lecture Notes, Univ. of Bergen, 1966.

Свойства распределения последовательностей регистров сдвига с обратной связью

Х. НИДЕРРЕЙТЕР

(Вена)

Рассматривается распределение s-групп в выходной последовательности регистров сдвига с обратной связью с арифметикой в конечном пространстве. Характеризуются s-группы, которые могут появляться и доказываются результаты по количеству появления заданной s-группы. Устанавливаются границы для коэффициентов автокорреляции рассматриваемых последовательностей.

H. Niederreiter
Mathematical Institute
Austrian Academy of Sciences
Dr. Ignaz-Seipel-Platz 2
A-1010 Vienna
Austria

Problems of Control and Information Theory, Vol. 15 (1), pp. 35–46 (1986)

ON A NECESSARY CONDITION OF OPTIMALITY FOR STOCHASTIC SYSTEMS WITH NOISE BY CONTROL

L. E. SHAIKHET

(*Donetsk*)

(Received January 18, 1985)

The optimal control problem of linear stochastic differential equation with noise by control and quadratic cost functional is regarded. Necessary condition of control optimality for that sort of problems is obtained. There are examples which demonstrate a possibility of construction of the optimal control synthesis by virtue the necessary optimality condition.

Introduction

In this paper the optimal control problem of linear stochastic differential equation with noise by control and quadratic cost functional is regarded. The stochastic perturbations are described with Wiener process and centred Poisson measure, which are independent one of another. Admissible control is defined as a stochastic process which is adapted to certain given family of σ-fields and which has a uniformly bounded second moment.

For admissible controls so-called McShane's variations are introduced and the limit of corresponding cost functional variations is calculated in final form.

Particularity of this method is the supplementary integration of cost functional variations with respect to some parameter. Without this integration the above-mentioned limit for systems with noise by control does not exist.

If the control varied is optimal, then the limit of cost functional variations is nonnegative. This fact gives some necessary condition of control optimality. By virtue of this condition we can in some cases, construct the optimal control synthesis in the final form.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

1. Problem statement

Consider the optimal control problem of linear stochastic differential equation [1–4]

$$d\xi(t)=(A_0(t)+B_0(t)u(t)+G_0(t)\xi(t))\,dt+$$

$$+\sum_{r=1}^{m}(A_r(t)+B_r(t)u(t)+G_r(t)\xi(t))\,dw_r(t)+$$

$$+\int_0^T(A(z,t)+B(z,t)u(t)+G(z,t)\xi(t))\tilde{v}(dt,dz)\,, \tag{1.1}$$

$$\xi(0)=\xi_0, \qquad t\in[0,T]$$

with quadratic cost functional

$$I(u)=M[\xi^*(T)D_0\xi(T)+$$

$$+\int_0^T(\xi^*(s)D_1(s)\xi(s)+u^*(s)D_2(s)u(s))\,ds] \tag{1.2}$$

and admissible control set U.

Here the m-dimensional Wiener process $w(t)=(w_1(t),\ldots,w_m(t))$ and the centred Poisson measure $\tilde{v}(t,A)$ with parameter $t\Pi(A)$ are independent and f_t-adapted, $\{f_t\}$ is a family of σ-fields on probability space $\{\Omega,\sigma,P\}$; $M(\xi_0|^2<\infty$, $\xi(t)\in R^n$, $u(t)\in R^l$. Equation and cost functional coefficients have the corresponding dimensionals, nonrandom and uniformly bounded, D_i $(i=0,1,2)$ are nonnegatively definite matrix.

An arbitrary f_t-adapted l-dimensional process $u(t)$, for which $\|u\|^2=$ $=\sup\limits_{0\leq t\leq T}M|u(t)|^2<\infty$ will be called admissible control.

Let $u_0(t)$ be optimal control of problems (1.1), (1.2), i.e. $I(u_0)=\inf\limits_{u\in U}I(u)$, the arbitrary stochastic variable v is $f_{\tau-\varepsilon}$-adapted, $0<\varepsilon<\tau<T$, $M|v|^2<\infty$,

$$u_\varepsilon^\tau=u_\varepsilon^\tau(t)=\begin{cases}v, & t\in[\tau-\varepsilon,\tau),\\ u_0(t), & t\in[0,T]\backslash[\tau-\varepsilon,\tau),\end{cases} \tag{1.3}$$

$$I'_\varepsilon(u_0)=\frac{1}{\varepsilon}\int_\varepsilon^T(I(u_\varepsilon^\tau)-I(u_0))\,d\tau\,,$$

$$I'_0(u_0)=\lim_{\varepsilon\to 0}I'_\varepsilon(u_0)\,. \tag{1.4}$$

It is evident that the value $I'_0(u_0)$ (if the limit exists) must be nonnegative. In this way, the condition $I'_\varepsilon(u_0)\geqq 0$ is necessary for the optimality of control u_0.

The aim of the paper is to find limit (1.4) for control problem (1.1), (1.2).

Theorem. For arbitrary admissible control u_0, limit (1.4) exists and is equal to

$$I'_0(u_0)=M\Big[\int_0^T (v^* D_2(s)v-u_0^*(s)D_2(s)u_0(s))\,ds+$$

$$+2\xi_0^*(T)D_0 p_0(T)+2\int_0^T \xi_0^*(s)D_1(s)p_0(s)\,ds\Big]+$$

$$+\mathrm{Sp}\,[D_0 P_0(T)]+\int_0^T \mathrm{Sp}\,[D_1(s)P_0(s)]\,ds\,.$$

Here $\xi_0(t)$ is a solution of equation (1.1) for control u_0, $p_0(t)$ is a solution of equation ($t\in[0,T]$)

$$p_0(t)=p_0^0(t)+\int_0^t G_0(s)p_0(s)\,ds$$

$$+\sum_{r=1}^m\int_0^t G_r(s)p_0(s)\,dw_r(s)+\int_0^t\int G(z,s)p_0(s)\tilde{v}(ds,dz)\,, \tag{1.5}$$

$$p_0^0(t)=\int_0^t B_0(s)V(s)\,ds+\sum_{r=1}^m\int_0^t B_r(s)V(s)\,dw_r(s)+$$

$$+\int_0^t\int B(z,s)V(s)\tilde{v}(ds,dz),\qquad V(s)=v-u_0(s)\,, \tag{1.6}$$

$P_0(t)$ is a solution of the equation ($t\in[0,T]$)

$$P_0(t)=P_0^0(t)+\int_0^t (G_0(s)P_0(s)+P_0(s)G_0^*(s))\,ds+$$

$$+\sum_{r=1}^m\int_0^t G_r(s)P_0(s)G_r^*(s)\,ds+$$

$$+\int_0^t\int G(z,s)P_0(s)G^*(z,s)\Pi(dz)\,ds\,, \tag{1.7}$$

$$P_0^0(t)=\sum_{r=1}^m\int_0^t B_r(s)W(s)B_r^*(s)\,ds+$$

$$+\int_0^t\int B(z,s)W(s)B^*(z,s)\Pi(dz)\,ds\,, \tag{1.8}$$

$$W(s)=MV(s)V^*(s)\,.$$

2. Proof of the theorem

To prove the theorem we need the next assertions. C will denote arbitrary positive constants.

Lemma 1. Let $\xi_\varepsilon^\tau(t)$ be the solution of equation (1.1) for control (1.3), $q_\varepsilon^\tau(\tau)= = \frac{1}{\varepsilon}(\xi_\varepsilon^\tau(t)-\xi_0(t))$. Then uniformly on $\tau\in[\varepsilon, T]$

$$\|q_\varepsilon^\tau\|^2 \leqq \frac{C}{\varepsilon}. \tag{2.1}$$

Proof. Evidently, $q_\varepsilon^\tau(t)=0$ for $t<\tau-\varepsilon$. Let $t\in[\tau-\varepsilon, \tau)$. Then (1.1)

$$q_\varepsilon^\tau(t)=\frac{1}{\varepsilon}\int_{\tau-\varepsilon}^{t} B_0(s)V(s)\,ds+\sum_{r=1}^{m}\frac{1}{\varepsilon}\int_{\tau-\varepsilon}^{t} B_r(s)V(s)\,dw_r(s)+$$

$$+\frac{1}{\varepsilon}\int_{\tau-\varepsilon}^{t}\int B(z,s)V(s)\tilde{v}(ds,dz)+\int_{\tau-\varepsilon}^{t} G_0(s)q_\varepsilon^\tau(s)\,ds+$$

$$+\sum_{r=1}^{m}\int_{\tau-\varepsilon}^{t} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)+\int_{\tau-\varepsilon}^{t}\int G(z,s)q_\varepsilon^\tau(s)\tilde{v}(ds,dz). \tag{2.2}$$

Hence

$$M|q_\varepsilon^\tau(t)|^2\leqq C\left[\frac{1}{\varepsilon}+\int_{\tau-\varepsilon}^{t} M|q_\varepsilon^\tau(s)|^2\,ds\right]$$

and on the basis of Gronwoll's–Bellman's lemma we have

$$\sup_{\tau-\varepsilon\leqq t\leqq\tau} M|q_\varepsilon^\tau(t)|^2\leqq\frac{C}{\varepsilon}. \tag{2.3}$$

Analogous with it for $t\in[\tau, T]$

$$q_\varepsilon^\tau(t)=q_\varepsilon^\tau(\tau)+\int_{\tau}^{t} G_0(s)q_\varepsilon^\tau(s)\,ds+$$

$$+\sum_{r=1}^{m}\int_{\tau}^{t} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)+\int_{\tau}^{t}\int G(z,s)q_\varepsilon^\tau(s)\tilde{v}(ds,dz), \tag{2.4}$$

$$M|q_\varepsilon^\tau(t)|^2\leqq C[M|q_\varepsilon^\tau(\tau)|^2+\int_{\tau}^{t} M|q_\varepsilon^\tau(s)|^2\,ds].$$

Using (2.3) and Gronwoll's–Bellman's lemma we obtain (2.1). Lemma is proved.

Lemma 2. Let

$$p_\varepsilon^0(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \int\limits_\varepsilon^t q_\varepsilon^\tau(\tau)\,d\tau, & t\in[\varepsilon,T]\,,\end{cases}$$

$$p_\varepsilon(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \int\limits_\varepsilon^t q_\varepsilon^\tau(t)\,d\tau, & t\in[\varepsilon,T]\,.\end{cases}$$

Then uniformly on $t\in[0,T]$

$$\lim_{\varepsilon\to 0} M\,|p_\varepsilon^0(t)-p_0^0(t)|^2=0\,, \tag{2.5}$$

$$\lim_{\varepsilon\to 0} M\,|p_\varepsilon(t)-p_0(t)|^2=0\,. \tag{2.6}$$

Proof. From (2.2) follows

$$p_\varepsilon^0(t)=\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau B_0(s)V(s)\,ds\,d\tau+\frac{1}{\varepsilon}\sum_{r=1}^m\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau B_r(s)V(s)\,dw_r(s)\,d\tau+$$

$$+\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau\int B(z,s)V(s)\tilde{\nu}(ds,dz)+\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau G_0(s)q_\varepsilon^\tau(s)\,ds\,d\tau+$$

$$+\sum_{r=1}^m\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)\,d\tau+\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau\int G(z,s)q_\varepsilon^\tau(s)\tilde{\nu}(ds,dz)\,d\tau\,.$$

Let $z(s)=B_0(s)V(s)$. Since $\|z\|<\infty$ and

$$\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau z(s)\,ds\,d\tau=\frac{1}{\varepsilon}\int\limits_0^\varepsilon sz(s)\,ds+\int\limits_\varepsilon^{t-\varepsilon}z(s)\,ds+$$

$$+\frac{1}{\varepsilon}\int\limits_{t-\varepsilon}^t(t-s)z(s)\,ds=\int\limits_0^t z(s)\,ds+\frac{1}{\varepsilon}\int\limits_0^\varepsilon(s-\varepsilon)z(s)\,ds+$$

$$+\frac{1}{\varepsilon}\int\limits_{t-\varepsilon}^t(t-s-\varepsilon)z(s)\,ds\,, \tag{2.7}$$

then

$$M\left|\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^\tau B_0(s)V(s)\,ds\,d\tau-\int\limits_0^t B_0(s)V(s)\,ds\right|^2\leqq C\varepsilon^2\,.$$

Analogously,

$$M\left|\frac{1}{\varepsilon}\int_{\varepsilon}^{t}\int_{\tau-\varepsilon}^{\tau} B_r(s)V(s)\,dw_r(s)\,d\tau-\int_{0}^{t} B_r(s)V(s)\,dw_r(s)\right|^2\leqq C\varepsilon\,,$$

$$M\left|\frac{1}{\varepsilon}\int_{\varepsilon}^{t}\int_{\tau-\varepsilon}^{t}\int B(z,s)V(s)\tilde{\nu}(ds,dz)\,d\tau-\int_{0}^{t}\int B(z,s)V(s)\tilde{\nu}(ds,dz)\right|^2\leqq C\varepsilon\,.$$

Let $z_\varepsilon^\tau(s)=G_0(s)q_\varepsilon^\tau(s)$. Since $\|z_\varepsilon^\tau\|^2\leqq\dfrac{C}{\varepsilon}$ and

$$\int_{\varepsilon}^{t}\int_{\tau-\varepsilon}^{\tau} z_\varepsilon^\tau(s)\,ds\,d\tau=\int_{0}^{\varepsilon}\int_{\varepsilon}^{s+\varepsilon} z_\varepsilon^\tau(s)\,d\tau\,ds+$$

$$+\int_{\varepsilon}^{t-\varepsilon}\int_{s}^{s+\varepsilon} z_\varepsilon^\tau(s)d\tau\,ds+\int_{t-\varepsilon}^{t}\int_{s}^{t} z_\varepsilon^\tau(s)\,d\tau\,ds\,, \tag{2.8}$$

then

$$M\left|\int_{\varepsilon}^{t}\int_{\tau-\varepsilon}^{\tau} G_0(s)q_\varepsilon^\tau(s)\,ds\,d\tau\right|^2\leqq C\varepsilon\,.$$

Analogously

$$M\left|\int_{\varepsilon}^{t}\int_{\tau-\varepsilon}^{\tau} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)\,d\tau\right|^2\leqq C\varepsilon\,,$$

$$M\left|\int_{\varepsilon}^{t}\int_{\tau-\varepsilon}^{\tau}\int G(z,s)q_\varepsilon^\tau(s)\tilde{\nu}(ds,dz)\,d\tau\right|^2\leqq C\varepsilon\,.$$

Hence

$$M|p_\varepsilon^0(t)-p_0^0(t)|^2\leqq C\varepsilon\,. \tag{2.9}$$

Condition (2.5) is proved. From (2.4) follows

$$p_\varepsilon(t)=p_\varepsilon^0(t)+\int_{\varepsilon}^{t} G_0(s)p_\varepsilon(s)\,ds+$$

$$+\sum_{r=1}^{m}\int_{\varepsilon}^{t} G_r(s)p_\varepsilon(s)\,dw_r(s)+\int_{\varepsilon}^{t}\int G(z,s)p_\varepsilon(s)\tilde{\nu}(ds,dz)\,. \tag{2.10}$$

Subtracting (1.5) from (2.10) and using (2.9) and Gronwoll's–Bellman's lemma we obtain (2.6). Lemma 2 is proved.

Lemma 3. Let $Q_\varepsilon^\tau(t)=\varepsilon M q_\varepsilon^\tau(t)\,(q_\varepsilon^\tau(t))^*$,

$$P_\varepsilon^0(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \displaystyle\int_{\varepsilon}^{t} Q_\varepsilon^\tau(\tau)\,d\tau\,, & t\in[\varepsilon,T]\,,\end{cases}$$

$$P_\varepsilon(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \displaystyle\int_{\varepsilon}^{t} Q_\varepsilon^\tau(t)\,d\tau\,, & t\in[\varepsilon,T]\,.\end{cases}$$

Then uniformly on $t \in [0, T]$

$$\lim_{\varepsilon \to 0} |P^0_\varepsilon(t) - P^0_0(t)| = 0\,, \tag{2.11}$$

$$\lim_{\varepsilon \to 0} |P_\varepsilon(t) - P_0(t)| = 0\,. \tag{2.12}$$

Proof. Let $t \in [\tau - \varepsilon, \tau)$, $X^\tau_\varepsilon(t) = Mq^\tau_\varepsilon(t) V^*(t)$. Using Ito's formula from (2.2) it is easy to obtain

$$\begin{aligned}
P^0_\varepsilon(t) = &\int_\varepsilon^t \int_{\tau-\varepsilon}^\tau [G_0(s) Q^\tau_\varepsilon(s) + Q^\tau_\varepsilon(s) G^*_0(s)]\, ds\, d\tau + \\
&+ \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau [B_0(s)\,(X^\tau_\varepsilon(s))^* + X^\tau_\varepsilon(s) B^*_0(s)]\, ds\, d\tau + \\
&+ \sum_{r=1}^m \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau [G_r(s) X^\tau_\varepsilon(s) B^*_r(s) + B_r(s)\,(X^\tau_\varepsilon(s))^* G^*_r(s)]\, ds\, d\tau + \\
&+ \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau \int [G(z,s) X^\tau_\varepsilon(s) B^*(z,s) + B(z,s)\,(X^\tau_\varepsilon(s))^* G^*(z,s)]\, \Pi(dz)\, ds\, d\tau + \\
&+ \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau \left[\sum_{r=1}^m G_r(s) Q^\tau_\varepsilon(s) G^*_r(s) + \int G(z,s) Q^\tau_\varepsilon(s) G^*(z,s) \Pi(dz) \right] ds\, d\tau + \\
&+ \frac{1}{\varepsilon} \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau \left[\sum_{r=1}^m B_r(s) W(s) B^*_r(s) + \int B(z,s) W(s) B^*(z,s) \Pi(dz) \right] ds\, d\tau\,.
\end{aligned} \tag{2.13}$$

Since $|Q^\tau_\varepsilon(s)| \leqq C$, $|W(s)| \leqq C$, $|X^\tau_\varepsilon(s)| \leqq \dfrac{C}{\sqrt{\varepsilon}}$, then, analogously with (2.7), (2.8), it can be shown that

$$|P^0_\varepsilon(t) - P^0_0(t)| \leqq C\sqrt{\varepsilon}\,. \tag{2.14}$$

Thus, condition (2.11) is proved. Analogously with (2.13) from (2.4) follows

$$\begin{aligned}
P_\varepsilon(t) = P^0_\varepsilon(t) + &\int_\varepsilon^t (G_0(s) P_\varepsilon(s) + P_\varepsilon(s) G^*_0(s))\, ds + \\
&+ \sum_{r=1}^m \int_\varepsilon^t G_r(s) P_\varepsilon(s) G^*_r(s)\, ds + \\
&+ \int_\varepsilon^t \int G(z,s) P_\varepsilon(s) G^*(z,s) \Pi(dz)\, ds\,.
\end{aligned} \tag{2.15}$$

Subtracting (1.7) from (2.15) and using (2.14) and Gronwoll's–Bellman's lemma we obtain (2.12). Lemma 3 is proved.

Proof of the theorem. Let $f(x, A, y) = x^*Ax - y^*Ay$. Then $I'_\varepsilon(u_0) = \sum_{i=1}^{4} \alpha_i(\varepsilon)$,

$$\alpha_1(\varepsilon) = \frac{1}{\varepsilon} \int_\varepsilon^T Mf(\xi_\varepsilon^\tau(T), D_0, \xi_0(T)\, d\tau\,,$$

$$\alpha_2(\varepsilon) = \frac{1}{\varepsilon} \int_\varepsilon^T \int_\tau^T Mf(\xi_\varepsilon^\tau(s), D_1(s), \xi_0(s))\, ds\, d\tau\,,$$

$$\alpha_3(\varepsilon) = \frac{1}{\varepsilon} \int_\varepsilon^T \int_{\tau-\varepsilon}^{\tau} Mf(\xi_\varepsilon^\tau(s), D_1(s), \xi_0(s))\, ds\, d\tau\,,$$

$$\alpha_4(\varepsilon) = \frac{1}{\varepsilon} \int_\varepsilon^T \int_{\tau-\varepsilon}^{\tau} Mf(v, D_2(s), u_0(s))\, ds\, d\tau\,.$$

Since

$$\alpha_1(\varepsilon) = M \int_\varepsilon^T (\xi_\varepsilon^\tau(T) + \xi_0(T))^* D_0 q_\varepsilon^\tau(T)\, d\tau =$$

$$= \mathrm{Sp}\,[D_0 P_\varepsilon(T)] + 2M\xi_0^*(T) D_0 p_\varepsilon(T)\,,$$

then (Lemmas 2, 3)

$$\lim_{\varepsilon\to 0} \alpha_1(\varepsilon) = \mathrm{Sp}\,[D_0 P_0(T)] + 2M\xi_0^*(T) D_0 p_0(T)\,.$$

Analogously with it and (2.7), (2.8), it is easily shown that

$$\lim_{\varepsilon\to 0} \alpha_2(\varepsilon) = \int_0^T (\mathrm{Sp}\,[D_1(s) P_0(s)] + 2M\xi_0^*(s) D_1(s) p_0(s))\, ds\,,$$

$$\lim_{\varepsilon\to 0} \alpha_3(\varepsilon) = 0\,,$$

$$\lim_{\varepsilon\to 0} \alpha_4(\varepsilon) = \int_0^T (v^* D_2(s) v - u_0^*(s) D_2(s) u_0(s))\, ds\,.$$

The theorem is proved.

3. Examples

In some cases, by virtue of the necessary optimality condition, the optimal control can be obtained in final form. Consider two examples for illustration.

Example 1. For the control problem

$$\dot{\xi}(t) = \alpha + (\beta + \gamma u(t))\dot{w}(t)\,, \tag{3.1}$$

$$I(u) = M[\mu\xi^2(T) + \lambda \int_0^T u^2(s)\, ds]\,, \tag{3.2}$$

the variable $I_0'(u_0)$ has the form

$$I_0'(u_0)=M[2\mu\xi_0(T)p_0(T)+\mu\gamma^2\int_0^T(v-u_0(s)]^2\,dt+$$

$$+\lambda\int_0^T(v^2-u_0^2(t))\,ds, \qquad \dot p_0(t)=\gamma(v-u_0(t))\dot w(t).$$

Since

$$M\xi_0(T)p_0(T)=\gamma\int_0^T M(\beta+\gamma u_0(t))(v-u_0(t))\,dt,$$

then

$$I_0'(u_0)=(\lambda+\mu\gamma^2)M\int_0^T[(v-u_0(t))^2+$$

$$+2(v-u_0(t))(u_0(t)+\mu\gamma\beta/(\lambda+\mu\gamma^2))]\,dt.$$

For nonnegative $I_0'(u_0)$ it is necessary and sufficient that the optimal control of problem (3.1), (3.2) might be of the form

$$u_0(t)=-\mu\gamma\beta/(\lambda+\mu\gamma^2).$$

This result can also be obtained by virtue of Bellman's equation.

Example 2. For the control problem of equation

$$\dot\xi(t)=(\alpha+\beta\dot w(t))u(t) \tag{3.3}$$

with cost functional (3.2) the variable $I_0'(u_0)$ has the form

$$I_0'(u_0)=M[2\mu\xi_0(T)p_0(T)+$$

$$+(\lambda+\mu\beta^2)\int_0^T(v-u_0(t))^2\,dt+2\lambda\int_0^T(v-u_0(t))u_0(t)\,dt],$$

$$\dot p_0(t)=(\alpha+\beta\dot w(t))(v-u_0(t)).$$

From Ito's formula

$$M\xi_0(t)p_0(t)=\alpha\int_0^t Mu_0(s)p_0(s)\,ds+$$

$$+\int_0^t M(\alpha\xi_0(s)+\beta^2u_0(s))(v-u_0(s))\,ds.$$

Let us suppose that u_0 has the form

$$u_0(t)=-q(t)\xi_0(t) \tag{3.4}$$

where $q(t)$ is a nonrandom function. Then

$$M\xi_0(t)p_0(t)=\int_0^t(\alpha-\beta^2q(s))M\xi_0(s)(v-u_0(s))\exp\left[-\alpha\int_s^t q(\tau)\,d\tau\right]ds\,,$$

$$I_0'(u_0)=M\{(\lambda+\mu\beta^2)\int_0^T(v-u_0(t))^2\,dt+$$

$$+2\int_0^T\xi_0(t)(v-u_0(t))\left[\mu(\alpha-\beta^2q(t))\exp\left[-\alpha\int_t^T q(s)\,ds\right]-\lambda q(t)\right]dt\}.$$

For nonnegative $I_0'(u_0)$ it is necessary and sufficient that

$$\alpha=q(t)\left[\beta^2+\frac{\lambda}{\mu}\exp\left(\alpha\int_t^T q(s)\,ds\right)\right]. \tag{3.5}$$

Hence the optimal control of problem (3.3), (3.2) have the form (3.4), (3.5).

Let us show that this solution coincides with the solution which is obtained by virtue of Bellman's equation. It is known [1] that the Bellman's equation gives the optimal control in the form

$$u_0(t)=-\alpha p(t)\xi_0(t)/(\lambda+\beta^2p(t))\,,$$

where $p(t)$ is a solution of differential equation

$$\dot p(t)=\frac{\alpha^2p^2(t)}{\lambda+\beta^2p(t)}\,,\qquad p(T)=\mu\,. \tag{3.6}$$

Hence the following condition must hold

$$q(t)=\alpha p(t)/(\lambda+\beta^2p(t))\,,$$

or $\dot p(t)=\alpha p(t)q(t)$. In fact, substituting $q(t)=\dot p(t)/\alpha p(t)$ into (3.5), we obtain (3.6).

4. Conclusion

It should be noted that instead of limit (1.4) we can generally speaking find the limit

$$I^\tau(u_0)=\lim_{\varepsilon\to0}\frac{1}{\varepsilon}\left[I(u_\varepsilon^\tau)-I(u_0)\right] \tag{4.1}$$

for arbitrary fixed $\tau\in(0,T)$. The inequality $I^\tau(u_0)\geqq0$, too, is a necessary condition for the optimality of control u_0. This was made in [5–9] for different (differential, integral, partial differential) nonlinear stochastic equations without noise by control. But for systems with noise by control, limit (4.1) in contrast to limit (1.4) does not exist.

Besides that, founding limit (4.1) in [5–9] for some parameters of control problem (also for optimal control) imposed rather strict Hölderian's conditions on time. Transition from limit (4.1) to its integral form (1.4) allows us to manage without these conditions.

At the same time, contrary to [5–9], the linear system is considered here. If there is a noise by control the nonlinearity of system considerably complicates the founding of limit (1.4).

References

1. *Gikhman, I. I., Skorokhod, A. V.*, Control stochastic processes. Kiev: Naukova dumka, 1977, 252 pp. (in Russian).
2. *Khasminsky, R. Z.* Stability of systems of differential equations for stochastic perturbations of their parameters. Moscow: Nauka, 1969, 365 pp. (in Russian).
3. *Chernousko, F. L., Kolmanovsky, V. B.*, Optimal control for stochastic perturbations. Moscow: Nauka, 1978, 352 pp. (in Russian).
4. *Kolmanovsky, V. B., Nosov, V. R.*, Stability and periodical regimes of regulation systems with after-effect. Moscow: Nauka, 1981, 448 pp. (in Russian).
5. *Shaikhet, L. E.*, Necessary conditions of control optimality for some stochastic systems. Resp. Konf. on stoch. differen. equat. theory. Thes. dokl., Donetsk, 1982, pp. 104–105 (in Russian).
6. *Shaikhet, L. E.*, On a necessary condition of control optimality for stochastic differential equations of hyperbolic type. Theory of stochastic processes. Kiev: Naukova dumka, 1984, **12**, pp. 96—101 (in Russian).
7. *Warfield, V. M.*, A stochastic maximum principle. SIAM J. Control and Optimization. August, 1976, **14**, *5*, pp. 803–826.
8. *Shaikhet, L. E.*, On optimal control of Volterra's equations. Problems of Control and Information Theory, Budapest, 1984, **13**, *3*, pp. 141–152.
9. *Shaikhet, L. E.*, Optimal control of stochastic integral-functional equations. Stochastic Optimization. (Intern. Conf. Kiev, USSR, 9–16 Sept. 1984): Abstracts of Papers, Kiev: V. M. Glushkov Institute of Cybernetics Ac. Sci. Ukr. SSR, 1984, pp. 212–213.

Об одном необходимом условии оптимальности для стохастических систем с шумом при управлении

Л. Е. ШАЙХЕТ

(Донецк)

Рассматривается задача оптимального управления линейным стохастическим дифференциальным уравнением с шумом при управлении и квадратичным критерием качества $I(u)$.

Цель работы — вычисление предела

$$I'_0(u_0)=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{\varepsilon}^{T}(I(u_\varepsilon^\tau)-I(u_0))d\tau,$$

где

$$u_\varepsilon^\tau = u_\varepsilon^\tau(t) = \begin{cases} v, & t \in [\tau-\varepsilon, \tau), \quad 0<\varepsilon<\tau<T, \\ u_0(t), & t \in [0, T]\backslash[\tau-\varepsilon, \tau) \end{cases}$$

для произвольного допустимого управления u_0.

Неравенство $I_0'(u_0) \geqq 0$ является необходимым условием оптимальности управления u_0.

Приведены примеры, демонстрирующие возможность построения синтеза оптимального управления с помощью этого необходимого условия оптимальности.

Л. Е. Шайхет
Всесоюзный научно-исследовательский
институт горной механики им. М. М. Фёдорова
СССР, 340055, Донецк-55,
пр. Театральный, 7.

Problems of Control and Information Theory, Vol. 15 (1), pp. 47–61 (1986)

THE RELATIONAL APPROACH TO THE COMPUTER-COMMUNICATION ARCHITECTURES

J. PUŽMAN

(*Prague*)

(Received October 1, 1984)

In the paper the formalized notion of architectures is introduced. In the basis of the algebra of binary relations, it enables us to classify architectures, to find out the properties of the classes of architectures, to derive their topology, and to construct their modular structure. The main accent is laid on communication network architectures which are arranged in layers (layered architectures), each consisting of modules providing communication functions and communication services.

1. Introduction

The notion of architecture occurs and is dealt with in many branches, however, in computer science and communication theory it possesses a specific meaning and plays sometimes a central role. In spite of the fact that the architecture of computer-communication systems has been of interest of many contributions we have not yet met its sufficiently universal and precise explication. The architecture is considered either statically and its structure (topology) is dealt with, or dynamically, for which the well-known tools have been worked out (finite automata, Petri nets, interlocutors, formal grammars [8]). However, the general notion of architectures remains still at the intuitive stage.

The architecture is generally considered as a space, time, and functional arrangement of any system. The system is, of course, understood to be modular and the arrangement refers to system modules. There is, however, a principal difference between a system proper and its architecture. While the system involves i.a. the influence of its environment and the impact on the environment, the notion of architecture is narrower and covers neither the outer services and their performance, nor the common goal, although sometimes there are endeavours to extend this notion to the whole system. We, however, shall deal with the architecture in the above narrower sense and no equality sign will be put between a system and its architecture.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Before we define an architecture more precisely we briefly sketch the requirements which the definition should satisfy. It would help to solve such problems as:

— to study superiorities and subordinations in the system,
— to define some special system elements or classes of system elements (if any),
— to classify the architectures according to some criteria,
— to derive the topology of systems,
— to enable the modular construction of the system,
— to define interfaces and corresponding protocols between system elements relating to the functions and services to be provided,
— to formalize the system description into the form suitable for computer processing, etc.

The above survey relates to the architecture of any system and from that point of view we shall lead further explanations. Nevertheless, the main attention will be paid to computer-communication systems and networks.

2. Basic preliminaries

As our approach is completely based upon the algebra of binary relations, it needs some definitions and implications some of which are not commonly found in textbooks and monographs [2, 3, 5]. Moreover, the following survey will help those who are not acquainted with the applied mathematical tool.

Throughout the paper we shall deal with a binary relation R on a finite nonempty set M. The binary relation (briefly relation) is a subset of $M \times M$ and is supposed to be nonempty as well.

The following properties of relation will be of use:

— reflexivity: for each $m \in M$, mRm
— irreflexivity: for each $m \in M$, $m\bar{R}m$ ($\bar{R}$ denotes the complement to R)
— symmetry: if for $m_i, m_j \in M$ m_iRm_j, then m_jRm_i
— asymmetry: if for $m_i, m_j \in M$ m_iRm_j, then $m_j\bar{R}m_i$
— transitivity: if for $m_i, m_j, m_k \in M$ m_iRm_j and m_jRm_k, then m_iRm_k
— atransitivity: if for $m_i, m_j, m_k \in M$ m_iRm_j and m_jRm_k, then $m_i\bar{R}m_k$
— negative transitivity: if for $m_i, m_j, m_k \in M$ $m_i\bar{R}m_j$ and $m_j\bar{R}m_k$, then $m_i\bar{R}m_k$
— cyclicity: for each $m_j \in M$ there exist unique $m_i, m_k \in M$, $m_i \neq m_k$, different from m_j, such that m_iRm_j and m_jRm_k
— acyclicity: if for each chain $m_{i_j}R\, m_{i_{j+1}}$, $m_{i_j} \in M$, $j=1, 2, \ldots, k-1$, $k>1$, $m_{i_k}\bar{R}m_{i_1}$ holds

— weak connectivity: for each m_i, $m_j \in M$, $m_i \neq m_j$, either $m_i R m_j$, or $m_j R m_i$, or both relations hold

— quasi-connectivity: for each m_{i_1}, $m_{i_k} \in M$ there exists a chain of $m_{i_j} \in M$ such that either $m_{i_j} R m_{i_{j+1}}$, or $m_{i_{j+1}} R m_{i_j}$, or both relations hold for all $j = 1, 2, \ldots, k-1$, $k > 1$

— centralization: there exists a unique $m_c \in M$ such that $m_c R m$ for all $m \in M$, $m \neq m_c$.

An asymmetric relation is obviously irreflexive. If R is weakly connected it is centralized, if it is centralized it is also quasi-connected (it follows directly from the definitions). The definition of cyclicity as well as that of acyclicity gives the asymmetry. If R is not acyclic but is required to be irreflexive then it must not be transitive.

While the above assertions are treated elsewhere [2, 5] the following ones are more special and deserve to be proved.

Lemma 1. Negative transitivity implies quasi-connectivity.

Proof. The quasi-connectivity of R on M is violated if there exist at least two different elements $m_{i_1}, m_{i_k} \in M$ such that it is impossible to find a connected sequence of $m_{i_j} \in M, j = 1, 2, \ldots, k$, adjacent of which being pairwise in any relation. Suppose such two elements and the sequence broken between m_{i_j} and $m_{i_{j+1}}$ but connected between $m_{i_{j+1}}$ and m_{i_j}. Hence, $m_{i_{j-1}} \bar{R} m_{i_{j+1}}$ (if the opposite held the sequence would continue to be connected due to the direct relation between $m_{i_{j-1}}$ and $m_{i_{j+1}}$), $m_{i_{j+1}} \bar{R} m_{i_j}$ but, at the same time, $m_{i_{j-1}} R m_i$ which contradicts the negative transitivity of R.

Lemma 2. If R is irreflexive and symmetric it cannot be transitive.

Proof. Suppose R to be transitive. Due to the symmetry $m_i R m_j$ implies $m_j R m_i$ and, using the transitivity, $m_i R m_i$ which violates the irreflexivity.

Lemma 3. Irreflexivity together with transitivity implies asymmetry.

Proof. Suppose R to be symmetric, i.e. each $m_i R m_j$ implies $m_j R m_i$. The transitivity of R, however, gives $m_i R m_i$ which again violates the irreflexivity.

Lemma 4. Negative transitivity together with asymmetry implies transitivity.

Proof. Let $m_i R m_j$ and $m_j R m_k$. Due to the asymmetry $m_j \bar{R} m_i$ and $m_k \bar{R} m_j$ which imply $m_k \bar{R} m_i$ (negative transitivity). Then either $m_i R m_k$ and R is transitive, or $m_i \bar{R} m_k$. If the latter held, it together with $m_k \bar{R} m_j$ would imply $m_i \bar{R} m_j$ which violates the initial supposition.

Note that the asymmetric and negatively transitive relation is usually called quasi-ordering (Fishburn in [5] uses the notion of weak ordering).

3. General definition of an architecture

Due to the requirement of modular construction from Section 1 we shall consider only modular systems which are composed of further indivisible elements — modules. Hence, the object we shall study is a finite nonempty set $M = \{m_i, i = 1, 2, \ldots, N\}$ where m (with or without indices) denotes a module.

The modules provide certain functions and well defined services. If $m_i, m_j \in M$ are two system modules, there exist three possibilities:

(a) m_i asks m_j for a service,
(b) m_i provides a service to m_j,
(c) between m_i and m_j there exists no correspondence, i.e. m_i neither asks m_j for a service nor m_i can provide it to m_j.

The relations "to ask for a service" and "to provide a service" can be considered as couplings between modules in a system which support the common global service of the system. We do not deal with the specifications of services (functions) because they depend upon the behaviour of the system under consideration. For communication services and functions see e.g. [8].

Instead of to say m_i asks m_j for a service we may say m_i is in a relation with m_j. Thus the system of couplings between modules can be expressed as a relation R on M. The denotation $m_i R m_j$ means that m_i may ask m_j for a service, while $m_i R' m_j$ (where R' is a reverse relation to R) means that m_i may provide a service to m_j. If we need not distinguish between R and R' we say shortly m_i (m_j) corresponds to m_j (m_i).

The third case (c) when m_i cannot ask m_j for a service and cannot provide it to m_j will be put down as $m_i \bar{R} m_j$ and $m_i \bar{R}' m_j = m_j \bar{R} m_i$.

The above relations between m_i and m_j from M can be rewritten as follows:

(a) $m_i R m_j$,
(b) $m_i R' m_j = m_j R m_i$,
(c) $m_i \bar{R} m_j$ and, at the same time, $m_j \bar{R} m_i$.

The first two cases, however, do not express exactly the relation between m_i and m_j. We must distinguish between

(aa) $m_i R m_j$ and, at the same time, $m_j R m_i$ (two modules provide services to each other), and
(ab) $m_i R m_j$ and, at the same time, $m_j \bar{R} m_i$ (m_i can only ask m_j for a service but cannot provide it to m_j).

Sometimes it is convenient to assign to R two additional relations, in particular,

— the strict preference P for which $m_i P m_j$ iff* $m_i R m_j$ and, at the same time, $m_j \bar{R} m_i$, and
— the indifference I for which $m_i I m_j = m_j I m_i$ iff $m_i R m_j$ and, at the same time, $m_j R m_i$.

* if and only if

By means of the above defined relations P and I the subcases (aa) and (ab) look as follows:

(aa) $m_i I m_j$,
(ab) $m_i P m_j$.

Even if we have not yet precised the relation R on M we shall generally state that a pair $A=(M, R)$ is an architecture. Thus, we involve among architectures such degenerated cases as so-called independent architecture for which $R=\emptyset$ (there are no relations between modules at all). In order to avoid it we require first of all R to be nonempty.

For practical reasons the notion of architecture should not permit the direct influence of any module to itself, in other words, the relation R should comply with

(i) irreflexivity.

The nonemptiness of R is necessary but not sufficient to keep all modules in some correspondence: no module or no subset of modules may be isolated (from the sense of relation) from the other modules. Such a property of the relation R will be called

(ii) quasi-connectivity,

in order to distinguish it from the weak connectivity (see Section 1).

Definition 1. The architecture A is a pair (M, R) where M is a finite nonempty set of modules and R is an irreflexive and quasi-connected binary relation on M (R satisfies (i) and (ii)).

From this place only architectures of the sense of Definition 1 will be dealt with.

4. Classification of architectures

The wide class of architectures is formed from architectures whose modules are joined with a strict preference ($I=\emptyset$) so that between modules there is only a relation of superiority (and subordination). We shall call such architectures asymmetric due to the fact that their relations fulfil instead of (i)

(iii) asymmetry

(the irreflexivity of R follows directly from its asymmetry).

Up to now we omitted the transitivity of R which expresses that a module can ask another module for a service through the third module which corresponds in appropriate direction to both modules. Some architectures occurring in practice are not in fact transitive. Moreover, as we shall see later, some architectures require even the atransitivity of R.

Let the asymmetry of R be replaced by a stronger condition, viz.

(iv) transitivity

which, with the irreflexivity, implies not only asymmetry but also acyclicity. Thus the transitive architectures cannot contain cycles (circles in terms of graph theory [1]) and hence their name — acyclic.

On the other hand, the architectures whose relations do not fulfil the asymmetry or are even symmetric (in the latter case in order that the irreflexivity may hold the relation has to be atransitive) are called distributed. In distributed architectures the modules can correspond between each other and within the pair of such modules there does not exist the superiority (subordination).

The special acyclic architecture containing a unique module m_c, called control, to which all other modules are subordinated (m_c can ask any module for a service either directly or via another modules), is called centralized. The relation of a centralized architecture must satisfy (i) and (iv) but (ii) is replaced by a stronger property

(v) centralization.

Acyclic architectures which are not centralized are called decentralized, i.e. they contain more than one control module which do not correspond between themselves and each of the other modules is subordinated at least to one control module. Special centralized architectures are star ones (see the next section).

The very interesting class of architectures are rings. All modules in a ring architecture form a circle in which each module is placed before the unique module as well as is subordinated to the unique module. More precisely, the relation of ring architectures must satisfy besides the irreflexivity (i) and the quasi-connectivity (ii) also

(vi) cyclicity.

As the cyclicity entirely excludes the transitivity the ring architecture has to be atransitive. In other words, we must not allow the influence of a module to another module unless it is directly subordinated in the ring. For the same reasons, ring architectures cannot be controlled from one place (cannot be centralized).

In order to illustrate the main classes of architectures the oriented graph representation is of use. Let modules correspond to vertices of a graph and the arc (oriented branch) leads from m_i to m_j iff $m_i R m_j$. Figure 1 shows some examples of such graphs of irreflexive and quasi-connected relations and hence of architectures.

The relation sub a) is transitive (and of course asymmetric) and is therefore acyclic, while the relation sub b) is atransitive but asymmetric and is composed of two cycles: (m_1, m_2, m_4) and (m_1, m_3, m_4). The case c) shows the relation which is neither asymmetric nor transitive and is an example of a distributed architecture. The remaining relation (d) fulfils the cyclicity and represents a ring.

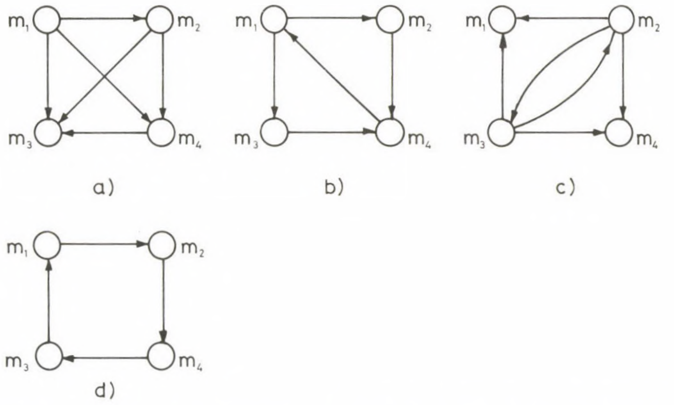


Fig. 1. Graphical examples of some relations: a) acyclic; b) asymmetric; c) distributed; d) cyclic

Table 1. Classes of architectures with properties of their relations

Property of relation	Class of architectures						
	general	asymmetric	acyclic	distributed	ring	centralized	decentralized
Irreflexivity	×	(×)	×	×	×	×	×
Quasi-connectivity	×	×	×	×	×	(×)	×
Asymmetry	–	×	(×)	○	(×)	(×)	(×)
Transitivity	–	–	×	(○)	(○)	×	×
Centralization	–	–	–	○	(○)	×	○
Cyclicity	–	–	(○)	–	×	(○)	(○)

Table 1 surveys the properties of just defined classes of architectures. The mark "x" used in the table stands for the property required in the definition, "o" indicates that the property is prohibited, "–" means that the relation need not possess the property, any mark in parenthesis is a consequence of the properties in the definition.

5. Layered architectures

In computer-communication systems (computer networks, data networks) we meet most often the notion of layered architecture. Sometimes, network architectures are treated only as layered ones [4, 6, 8, 9]. In this section we shall pay our attention to layered architectures and we shall analyze them more deeply.

The layered architecture is understood to be a set of modules partitioned in subsets—layers—within which the modules must not correspond. The only correspondence is allowed between layers and the direction of correspondence (the hierarchy) leads from upper layers to lower ones. The layers are supposed to be in a certain sense ordered. Layered architectures belong therefore to asymmetric ones.

From this follows that the layered architecture A_L is a pair (M, R_L) where R_L satisfies the basic properties, i.e. quasi-connectivity and asymmetry which are necessary, though, but not sufficient. The condition of unidirectional correspondence between layers can be expressed as follows:

$$\text{for all} \quad m_i, m_j, m_k \in M, m_i R_L m_j \quad \text{implies either} \quad m_i R_L m_k \quad \text{or} \quad m_k R_L m_j. \qquad (1)$$

Lemma 5 (Fishburn). Implication (1) is equivalent to the negative transitivity of R_L.

For the proof, see [5].

The modules in each layer are not comparable (related to R_L) and the layers are ordered according to the direction of R_L from the highest one (whose modules only ask for a service and control all modules in lower layers) to the lowest one (whose modules only provide services and are controlled by modules staying at higher layers).

The following definition is a direct consequence of Lemma 1 from Section 1 and of Lemma 5.

Definition 2. A layered architecture A_L is a pair (M, R_L) where M is a finite nonempty set and R_L is a nonempty asymmetric and negatively transitive relation, or shortly, R_L is a quasi-ordering.

Theorem 1. A_L belongs to acyclic architectures.

The proof follows from Lemma 4 and from the definition of acyclic architecture.

Now let us derive from R_L the relation ρ of uncomparability: $m_i \rho m_j$ iff $m_i \bar{R}_L m_j$ and $m_j \bar{R}_L m_i$. It is easily seen that ρ is reflexive, symmetric, and transitive, and hence ρ is an equivalence. The equivalence allows us to partition M into disjoint nonempty equivalence classes. The set $V(m) = \{m_i \in M: m \rho m_i\}$ is an equivalence class generated by m and is, in fact, a layer containing m. Returning to (1) m_i is in a layer while m_j and m_k lay in lower layers, or m_j is in a layer and m_i and m_k belong to upper layers.

Let us examine some types of layered architectures according to the criterion of the number of layers. There exists no 1-layered architecture because the definition excludes nonempty set of relations and modules within a layer cannot be in any correspondence.

A special case is, however, a 2-layered architecture. In such architectures if $m_i R_L m_j$ for any modules m_i and m_j from M, there exists no module $m_k \in M$, different from m_i and m_j, such that $m_i R_L m_k$ and $m_k R_L m_j$. The services in the 2-layered architecture are demanded by modules in the upper layer directly from modules in the adjacent lower layer without any transfer via a third module belonging to an intermediate layer. Two-layered architectures are either centralized and then the upper layer consists exactly from one module (control one) while the other modules form the lower layer, or decentralized. The former architecture is sometimes called a star.

The property of centralization can be spread out to any layered architecture. The common feature of centralized layered architectures is the one-module highest layer. The decentralization means that several modules in the highest layer play the role of control and may request services from all modules in lower layers.

Another extreme case (as for the number of layers) is an N-layered architecture where N is, as usually, the number of modules. Since the N-layered architecture contains only one module in each layer, we call it linear or sequential and denote it by $A_S=(M, R_S)$.

Theorem 2. R_S is the strict ordering.

Proof. Remind that a relation is the strict ordering iff it is asymmetric, transitive, and weakly connected [2]. The asymmetry and transitivity follow directly from the definition of R_S (layering). There is a direct correspondence between modules in adjacent layers. The correspondence keeps the validity also between modules in nonadjacent layers due to the layer hierarchy and transitivity of R_S. Hence, R_S is weakly connected.

Since R_S is the strict ordering there must exist a unique "minimal" module m_c (the minimality refers to R_S: m_i is "less" than m_j iff $m_i R_S m_j$) which is control. Due to the weak connectivity of R_S the architecture A_S is centralized at m_c.

The sequential architecture is often met in communication systems. A chain "periphery unit — control unit (which together form the DTE — data terminal equipment) — DCE (data circuit — terminating equipment) — line (circuit)" is one example.

Layered architectures with more than two layers form the basis for centralized and decentralized computer-communication systems and networks (see e.g. the model of OSI [6, 9]).

The above detailed discussion allows us to state the necessary and optional properties of relations R_L defining the types of layered architectures. The classification similar to Table 1 is in Table 2.

Table 2. Some layered architectures with properties of their relations

Property of relation	2-layered	More than 2-layered		
	star	sequential (linear)	centralized	decentralized
Irreflexivity	(×)	(×)	(×)	(×)
Quasi-connectivity	(×)	(×)	(×)	(×)
Weak connectivity	○	×	○	○
Transitivity	(×)	(×)	(×)	(×)
Negative transitivity	×	×	×	×
Centralization	×	(×)	×	○

6. The topology of architectures

Under the topology of architecture we mean the layout of modules and their interconnections. Figure 1 has shown a certain representation of relations by means of directed graphs, however, it involves all interconnections between modules and is redundant. For example, the arc (m_1, m_3) in Fig. 1a can be removed because the coupling between modules m_1 and m_3 is guaranteed by arcs (m_1, m_2) and (m_2, m_3) and by the transitivity of R.

For the representation of the architecture topology we use again a directed graph based, however, on the dominating relation with respect to R rather than on R proper [7].

Let R be a relation on M. A set of all pairs $(m_i, m_j) \in M \times M$ for which $m_i R m_j$ and there exists no $m \in M$ such that $m_i R m$ and $m R m_j$, is called a dominating relation with respect to R, or briefly dominance (also Hasse's relation) to R. The dominance to R will be denoted by R^x (with or without indices).

It is obvious that R^x is always atransitive and a subset of R. Only if R is atransitive, $R^x = R$.

The dominance keeps many properties of the original relation R, e.g. irreflexivity, asymmetry, cyclicity, acyclicity, quasi-connectivity. On the other hand, R^x need not be weakly connected and centralized even if R complies with these properties.

The dominance R^x seems to be an appropriate tool to express the topology of any architecture because it keeps only necessary interconnections between modules and does not violate the main properties of the original relation. We define the topology of architectures as follows:

Definition 3. The topology of architecture $A = (M, R)$ is an oriented graph $G = (M, B)$ where M is a set of vertices and B is a set of arcs such that $b_{ij} \in B$ iff $m_i R^x m_j$.

Such a graph is also called the graph of dominance relation [2].

On the basis of the properties of relations defining classes and types of architectures we can study their topologies. It is obvious that G is connected (in terms of graph theory [1, 7]) and loopless. G of acyclic architecture does not contain oriented circles while that of asymmetric architecture can do. The control vertex (if any) is characterized by only outgoing arcs.

Theorem 3. The topology of layered architectures is an oriented connected graph without loops and circles. The graph consists of layers, i.e. disjoint (with respect to vertices) subgraphs of isolated vertices. The layers can be ordered in such a way that each vertex in a layer is incident with ingoing arcs from all vertices in the adjacent higher layer (if any) and with outgoing arcs to all vertices in the adjacent lower layer (if any).

Proof. The first part of the theorem is a direct consequence of the properties of R_L. The layering and the absence of arcs between vertices within layers is ensured by ρ_L assigned to R_L. Suppose m_i to be in a layer and m_j, m_k to be in another layer. If $m_i R_L m_j$ then $m_i R_L m_k$ (according to Lemma 5) and m_j, m_k belong to a lower layer. If there exist no m and m' such that $m_i R_L m$, $m R_L m_j$, and $m_i R_L m'$, $m' R_L m_k$, then m_j, m_k stay in the adjacent layer and the arcs lead from m_i both to m_j and m_k. In the opposite case it is sufficient to replace m_j by m and m_k by m' and iteratively to find out the adjacent lower layer. The same considerations hold if $m_j R_L m_i$ and m_j, m_k belong to higher layers. The interconnection of each vertex in a layer with all vertices in adjacent layers is ensured again by the negative transitivity of R_L. Finally, if an arc led from a vertex in a layer to a vertex in the nonadjacent layer the property of R_L^x would be violated (remind R_L^x is atransitive and hence not negatively transitive).

As a consequence of Theorem 3 the topology of star architecture is a star, and that of sequential (linear) architecture is a serpent.

Figure 2 shows several examples of topologies of layered architectures and needs no commentary.

This representation of architecture structures is very graphic but of less use in practical computations. The graph, however, can be expressed in a matrix form which enables to solve practically all problems by means of the matrix algebra [3]. One appropriate matrix is defined as follows:

Definition 4. The relational matrix to a relation R on M is an $N \times N$ $(0-1)$ matrix where N is the cardinality of M and the elements r_{ij} equal to 1 iff $m_i R m_j$, and to 0 otherwise.

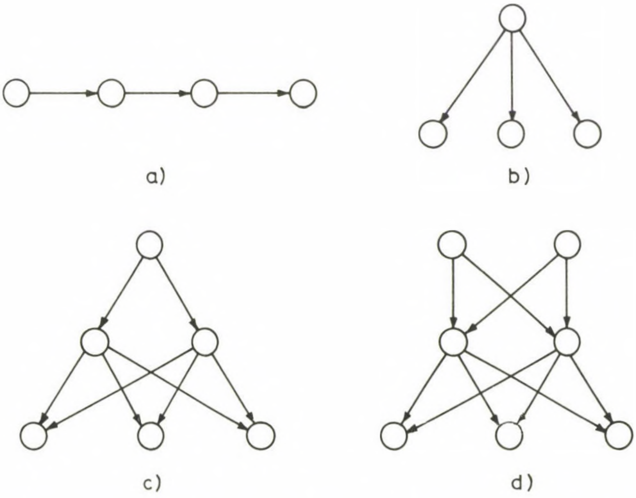


Fig. 2. Topologies of layered architectures: a) linear; b) star; c) 3-layered centralized; d) 3-layered decentralized

The topology of architecture (M, R) is of course put down by the relational matrix derived from R^x rather than from R. From the properties of dominances corresponding to some types of architectures the exact forms of relational matrices can be stated. For example, only matrices with all zero elements at the main diagonal may represent the topology of an architecture (due to the irreflexivity of R^x) while if we add the condition $r_{ij}+r_{ji}\leqq 1$ for all $i, j=1, 2, \dots, N$, we get the topology of an asymmetric architecture.

As layered architectures occupy the dominating place the properties of their relational matrices are useful to be examined. From Definition 4 and Theorem 3 it is obvious that the relational matrix of A_L is of (or can be modified by permutation of rows and columns in) the form of partitioned matrix. Each submatrix (field) is either zero or consisting of all ones and the dimensions of each of them are equal to the cardinality of adjacent layers V_i and V_{i+1} (the number of modules in layers V_i and V_{i+1}).

For our example in Fig. 2d we get the 7×7 relational matrix partitioned into 9 submatrices two of which (2×2 and 2×3) consist of all ones:

$$\left[\begin{array}{cc|cc|ccc} 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

From this follows that the relational matrix of sequential architecture contains exactly $N-1$ ones and can be modified in a triangle matrix with nonzero elements staying closely above (below) the main diagonal, that of star architecture consists of $N-1$ ones in one row corresponding to the control module, etc.

7. Interfaces, protocols, and other applications

Let $A=(M, R)$ be a (network) architecture. Any element of the dominating relation R^x with respect to R is the interface between corresponding modules. More precisely, $m_i R^x m_j$ states the existence of an access means by which the pair m_i and m_j uses or provides services. Physically it represents a system of circuits over which the two modules cooperate. The direction $m_i R^x m_j$ need not be comprehended that the circuits are unidirectional. Most of the control circuits follow the direction of R^x, however, some of them (and other than control ones) of the opposite direction are not excluded (and are even necessary).

Remind that the interface is generally treated as mechanical, electrical, functional, and procedural characteristics of circuits between network modules [4]. The latter two characteristics are commonly called the interface protocol (control procedure). Thus the dominance R^x involves also interface protocols.

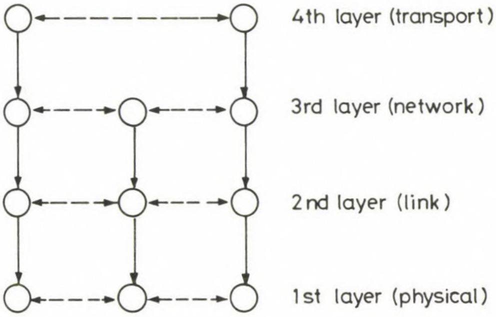


Fig. 3. The 4-layered model of OSI

In layered architectures $A_L=(M, R_L)$ we distinguish between two kinds of protocols: interlayer and layer (peer-to-peer). The interlayer protocol governs the access (communication) between modules in adjacent layers while the layer protocol effects between modules within one layer (peer modules).

Since modules within the layer are in the relation ρ assigned to R_L (see Section 5) we may consider elements of ρ as the existence of layer protocols or simply layer protocols. In contradistinction to R_L which is unidirectional, ρ is always bidirectional (hence the name peer-to-peer). Protocols as well as interfaces are, however, always bidirectional regardless of the type of relation: $m_i P m_j$ means that m_i is superior to m_j and controls it while $m_i I m_j$ points out the same control level of both modules.

Usually layer protocols are directly given by ρ, however, we do not exclude the case when the set of layer protocols is a proper subset of ρ or even is identical with the dominance ρ^x to ρ.

Attentive reader can object that the topology of layered architecture does not completely correspond to that of OSI [6, 9]. In the model of OSI the only interfaces are admitted between modules staying in subgraphs into which the whole graph is vertically partitioned (systems in the OSI terminology) and the layer protocols operate only between neighbouring modules within layers (Fig. 3). The topology in Fig. 3 shows the model of OSI up to the transport layer and differs from the general layered topology from the previous section. Nevertheless, the relational approach manages to cope with this problem.

All relations of layered architecture form a set $R_0 = R_L \cup \rho$ so that the protocols must belong to the subset of R_0. Such a subset is the dominance R_0^x which includes both interlayer protocols and layer protocols. Layer protocols proper form the atransitive dominance ρ^x (dashed lines in Fig. 3) and interlayer protocols are expressed by the difference $R_0^x \backslash \rho^x$ (solid lines).

The relational approach may help to solve the modular construction of systems and networks. Let $A = (M, R)$ be an architecture and $m \notin M$ be some new module. By the introduction of m into A the new architecture $A' = (M', R')$ results, where $M' = M \cup \{m\}$, $R' = R \cup R_m$ and $R_m \subseteq M \times \{m\} \cup \{m\} \times M$, $R_m \neq \emptyset$. Note that A' is indeed an architecture because R_m is irreflexive and the nonemptiness of R_m guarantees the quasi-connectivity of R'.

This construction does not change the original architecture and only adds interfaces between the new module and some original modules in A.

By means of the procedure having been just described one can centralize any decentralized acyclic architecture. If $A = (M, R)$ is such an architecture and $m_c \notin M$ should become control, the centralized architecture is defined on $M \cup \{m_c\}$ with the relation $R \cup (m_c \times M)$. The interfaces between m_c and A connect the control module with all initial modules in A (modules for which no other superior modules exist).

The union of a module and an architecture may be simply spread out to the union of two architectures $A_1 = (M_1, R_1)$ and $A_2 = (M_2, R_2)$ in the following way: $A = A_1 \cup A_2 = (M, R)$ where $M = M_1 \cup M_2$ and $R \subseteq (M_1 \times M_2) \cup (M_2 \times M_1)$.

8. Conclusion

The relational approach to define the general architecture has enabled to answer all questions stated at the beginning of the paper. The author believes, however, that the approach is applicable to a wider spectrum of problems concerning the architectures, in particular the computer-communication architectures. All propositions and complements as well as critical notes in this area will be welcome.

References

1. *Berge, C.*, The Theory of Graphs. New York, J. Wiley, 1962.
2. *Birkhoff, T. C.*, *Bartee, T. C.*, Modern Applied Algebra. New York, McGraw-Hill, 1970.
3. *Birkhoff, T. C.*, *McLane, S.*, A Survey of Modern Algebra. New York, MacMillan, 1965.
4. CCITT Recommendation X. 25. In: Yellow Book, Vol. VIII, Fasc. VIII.2, Data Communication Networks, Services and Facilities, Terminal Equipment and Interfaces. Geneva, UIT, 1982.

5. *Fishburn, P. C.*; Utility Theory for Decision-Making. New York, J. Wiley, 1970.
6. Information Processing Systems — Open Systems Interconnection — Basic Reference Model. Int. Standard IS 7498, ISO, 1983.
7. *Ore, O.*, Graphs and Their Uses. New York, Random House, 1963.
8. *Pužman, J., Pořízek, R.*, Communication Control in Computer Networks. Chichester, J. Wiley, 1981.
9. Reference Model of Open Systems Interconnection. Recommendation CCITT X. 200. Geneva, UIT, 1984.

Реляционный подход к архитектурам вычислительных сетей

Й. ПУЖМАН

(Прага)

В работе введено формальное понятие архитектуры, основанное на алгебре бинарных отношений. Этот подход позволяет классифицировать архитектуры, найти общие свойства классов архитектур, вывести их топологию и построить модульную структуру. Основное внимание уделяется архитектурам коммуникационных сетей, которые упорядочены в уровни (многоуровневые архитектуры) и где каждый уровень состоит из модулей, выполняющих коммуникационные функции и предоставляющих коммуникационные службы.

J. Pužman
State Commission for the Development of Science,
Technology, and Investments
Slezská 9, 120 29 Prague 2
Czechoslovakia

Problems of Control and Information Theory, Vol. 15 (1), pp. 63–78 (1986)

TOPOLOGICAL OPTIMIZATION OF COMMUNICATION NETWORKS SUBJECT TO RELIABILITY CONSTRAINTS*

A. N. Venetsanopoulos, I. Singh

(*Toronto*)

(Received January 10, 1985)

There is an increasing demand for computer networks. As these networks are put to use, there is a growing concern about their reliability and cost. One factor which influences the reliability and cost of computer networks is their topology. This paper examines the problem of optimizing a network's topology subject to specified reliability constraints. In doing so, a new reliability measure is defined and a new heuristic optimization procedure is proposed. Some examples are also presented.

1. Introduction

The use of computer networks has been rapidly increasing. There are two main reasons for this: (1) Computer networks provide an economical means of sharing expensive computer resources, such as specialized hardware, specialized software and databases. (2) They provide access to all users, irrespective of geographical locations. However, as these networks are increasingly put to use, there is a growing concern about their reliability and cost [1]. One of the chief factors affecting both the reliability and cost of a computer network is its topology. The topology influences the reliability in that it determines the extent of redundancy incorporated in the network. It influences the cost since it determines the link requirements, hence the communication cost. Minimizing the communication cost is particularly important, since it appears to be declining relatively slower than either the computing cost or the storage cost [2].

One of the popular topologies is the star topology. In this topology all the nodes of the network are directly connected to a central node. This configuration poses serious reliability problems since the failure of a single link would disconnect the network. In addition, the failure of the central node would cripple the network. On the otheı extreme the completely connected topology maximizcs reliability by providing a

* The research reported in this paper was supported by the Natural Sciences and Engineering Research Council of Canada under Grant No. A7397.

direct link between each node pair. If this link should fail, communication is still possible, through any other node connected to both end points of the failed link. The obvious disadvantage of this topology is its high cost. The mesh topology however shown in Fig. 1 is a more practical topology, as it offers a compromise between cost and reliability. The topology shown has an edge connectivity of two, that is at least two links must fail before the network is disconnected. This connectivity can be included as a constraint in the topological design problem, and is frequently used by designers as a measure of the network's reliability [3]. However, this measure is only adequate for networks with relatively small number of nodes and small component failure probability. For large networks and higher failure rates more rigid constraints must be applied.

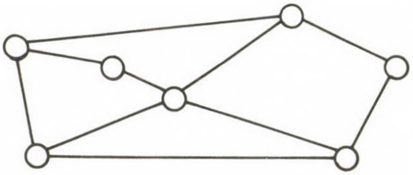


Fig. 1

A better and widely used parameter for characterizing network reliability is the "terminal reliability of the network", defined as the probability of having at least one path existing between the two nodes in the worst condition [1]. There are numerous techniques for determining this reliability [4–8], most of which are based on the enumeration of cut-sets and tie-sets. Its main drawback lies in the fact that it requires knowledge of the nodes or users in the worst condition, and this condition is very sensitive to changes in the network's topology.

To get around this problem the "probability of connectedness" was suggested [9]. This is defined as the probability that the network is simply connected. However, computation of this probability is very complex. Networks using connectedness as the reliability criterion were investigated in [9] and was concluded that the topology of maximally reliable networks depends on the magnitude of the link failure probability. This complicates the problem of designing optimum networks and makes it difficult if not impossible to derive analytical solutions. Most of the available techniques for the solution of such problems are heuristic.

Among the more widely used heuristic methods was the Branch Exchange Method [3, 10]. This method was used in the design of the ARPA network. The method starts with an arbitary topological configuration and reaches a local optimum by means of local transformations or branch exchanges, which consist in each case of the elimination of one or more old links and the insertion of one or more new links. Once a local optimum network is found, the entire procedure is repeated again with a different starting topology. By finding local optimum solutions for different starting networks, a variety of solutions can be generated. The practicality of the approach is based on the assumption that with a high probability some of the local optima found are close to the global optimum. It was claimed that the discrepancy between the optimal and heuristically generated optimal solution is in the order of 5% to 20% [2].

2. Statement of the problem

The general network design problem could be defined as follows [3]:

Given	*Node Locations*
	Peak-hour traffic requirements between node pairs
Minimize	*Total cost*
Over the design variables	*Topology*
	Channel capacities
	Routing policy
Subject to	Link capacity constraint
	Average delay constraints
	Reliability constraints

In this work we are concerned with the subproblem specified by the items in italic characters of the previous problem. As the cost of the network we shall consider only the communications cost. A restricted class of probabilistic networks are considered in this paper. The following assumptions, in terms of a linear graph model and similar to those used by Leggett [11], are made about these networks:

(1) The nodes of the network are perfect: i.e., each node exists with probabililty one.
(2) The branches are uncorrelated: i.e., the state of one branch does not affect the state of any other branch.
(3) Branch i is entirely ON with probability p_i and OFF with probability $q_i = 1 - p_i$.
(4) Branch capacities are not considered. They are assumed to be adequate.
(5) The system is stationary: i.e., it has reached a steady state.

Practical networks satisfying these assumptions exist. For example, node availability approaching one is easily achieved by duplicating the processing unit. Although this is an expensive practice it is frequently done, where high security is required. Also, in many applications, where the capacities are small, capacity allocation problems are usually neglected.

In this paper the reliability of networks will be studied, with respect to their topological properties. Our objective is to design minimum cost computer networks, satisfying specified reliability constraints. In doing so, a new reliability measure is proposed and a new heuristic procedure using optimization techniques is established. In section 3, network reliability and network cost are discussed. In section 4 the optimization algorithm is outlined. Simulation results are presented in section 5. The conclusions follow in section 6.

3. Network reliability and network cost

In this section expressions are given, based on the proposed reliability criterion, for the overall reliability of centralized and distributed networks. Finally, the network cost is considered.

3.1 *Overall Network Reliabililty*

The "terminal reliability" for a number of configurations was presented in [12]. This is a good measure if the pair of nodes selected is the one most likely to be disconnected. The problem with this method lies in the selection of the pair of nodes in the worst condition. For analysis this may not be a problem, but for synthesis no tractable analytical expression exists, since the pair of nodes in the worst condition may change by slightly changing the network's topology or the unavailability of the network components. In this paper, a new reliability measure is proposed. This measure is a simple extension of the terminal reliability measure and is now introduced for the two general categories of networks, namely, centralized and distributed.

Centralized networks, also referred to as terminal oriented networks, are characterized by having one central processing facility (central node) servicing a number of remote and local terminals. The communication problem here is that of providing access to the central facility from the remote terminals. A simple solution to the problem is to connect the terminals directly to the central facility in a star configuration. Such a connection avoids conflict for communication capacity, but is very expensive. At other extreme, the least expensive solution is the minimal spanning tree. This consists of connecting the terminals among themselves and the computer using the shortest total line length. The disadvantage of this solution is that all the terminals, in effect, share the same communication facilities, thus a fair amount of control is required for allocating this resource among the terminals in what is known as polled or multidrop systems.

Between these two configurations, there exist other configurations, in which clusters of terminals are connected to remote concentrators, that are in turn connected to the central computer. At the concentrator, data from the terminals are multiplexed and then transmitted to the computer. Such configurations are very popular and will be considered here. We now define for centralized networks the "overall network reliabililty", as the average probability of the concentrators being connected to the central facility. In a practical network, the importance of the concentrators may differ. For example, the importance may be associated with the number of terminals connected to the concentrator. In view of this, the overall network reliability R, is given by:

$$R=\frac{\sum_{i=1}^{N} W_i P_i}{\sum_{i=1}^{N} W_i} \tag{1}$$

where N is the number of concentrators, W_i is the weight associated with the ith concentrator and P_i is the probabililty of the central computer being connected to the ith concentrator.

If each of the concentrators are replaced by a computer facility itself, each with its own set of concentrators and local terminals, what result is a distributed computer network. Interconnecting computers in this fashion provide a way for sharing computer facilities, thus providing more computer power than might otherwise be possible.

In distributed networks, since the nodes represent computer centres, it is necessary that each node be able to communicate with all other nodes in the network. As a result, the "overall network reliability" is defined as the average probability of any two nodes in the network being connected. Often, in a practical network, the resources and computational load of each computer vary, thus having certain computer centres connected may be more important than others. In view of this, the "overall network reliability" is given by

$$R=\frac{\sum_{ij} W_{ij} P_{ij}}{\sum_{ij} W_{ij}} \qquad (i=1,\ \ldots,\ N-1, j=i+1,\ \ldots,\ N) \tag{2}$$

where N is the number of computer centers or nodes in the network, W_{ij} is the weight associated with the connection between the ith and jth nodes and P_{ij} is the probability of having a connection between nodes i and j. Note that with the appropriate choice of the weights, (1) can be considered a special case of (2).

3.2 Communication Link Reliabililty

A factor which affects the overall network reliability is the reliability of the individual communication links. In general, a link consists of a series arrangement of terminal equipment, such as transmitters, receivers, equalizers, etc., and a communication channel, wire or radio (microwave). Associated with each component is a failure probability, which contributes to the overall link reliability. For this study the simplified model, shown in Fig. 2, is used. In this model, the terminal equipment, local loops, interface equipment, etc., are lumped together with an equivalent failure probability. Assume that A_i and B_i of the link i have unavailabilities q_{A_i} and q_{B_i} respectively and the channel C_i has an unavailability that is a function of its length, l, $q_{C_i}(l)$, then the availability of link i, P_i, is given by:

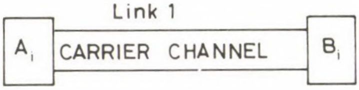


Fig. 2

$$P_i=(1-q_{A_i})(1-q_{B_i})(1-q_{C_i}(l))=P_{T_i}(1-q_{C_i}(l)) \tag{3}$$

where

$$P_{T_i}=(1-q_{A_i})(1-q_{B_i}) \tag{4}$$

is the availability of all equipment associated with link i. In general, the unavailability of the equipment associated with each link is assumed to be constant. This is a reasonable assumption if it is assumed the equipment has already passed through its "burn-in" period and has not yet reached its "wear-out" stage. That is, the equipment availability is affected only by random failures. The channel availability however is a function of its length. Montanari [1], presented some models of the channel. In one of his models he assumed the channel is built as a series of elementary hops (microwave connection) of equal length λ and the failure of each hop is statistically independent. Assuming the experimental function relating the hop length to its unavailability is $h(\lambda)$, then the unavailability of channel i of length L_i is given by:

$$q_{C_i}=1-(1-h(\lambda_i))^{L_i/\lambda_i}\simeq(L_i/\lambda_i)(h(\lambda_i)). \tag{5}$$

Experimental results given in [1] show that for the hop lengths of 7.5 miles and 15 miles unavailabilities of 5.4×10^{-5}/mile and 2.0×10^{-4}/mile were found respectively.

3.3 Overall Network Cost

The overall network cost is primarily dependent on the associated hardware cost (storage facilities, concentrators, etc.), the software cost (control algorithms) and the communications cost. Since the communications cost constitutes the major part of the network cost, it is reasonable to consider this cost only in the optimization process. Thus, the network cost $C(N)$ is given by

$$C(N)=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} s_{ij}c_{ij} \tag{6}$$

where s_{ij} represents the i, j entry in the connection matrix $S(N_k)$, where

$s_{ij}=1$ indicates a link exists between nodes i and j.
$s_{ij}=0$ indicates no link exists between nodes i and j.

c_{ij} denotes the i, j entry in the cost matrix $[C]$ and represents the cost of establishing a link i, j between nodes i and j. The cost matrix is symmetric, signifying the cost of establishing link i, j and j, i is the same. Physically, for full or half duplex communication i, j and j, i designate the same link.

In a practical situation the link cost would be determined by the applicable tariffs and would depend on the line capacity, type of connection and line leased, distance between nodes, etc. An example of Bell Canada's rates for interexchange service on voice grade circuits or data channels is given in [14]. These channels are two-wire, half duplex. In this rate structure, the termination cost is built in and rates are given strictly as a function of distance. Some typical rates for Telepak service are presented in [3]. In this rate structure, there is a fixed termination cost and a mileage cost.

4. The optimization algorithm

The network optimization problem considered here requires the minimization of the communication cost, subject to reliability constraint, with the network topology being the design variable. The constraint imposed is that the network reliability, $R(N)$, be greater than or equal to some value $\in$, less than 1, $(R(N)\geqq \in, 0< \in <1)$. In solving this problem there are many considerations, which make it extremely difficult to derive computationally feasible algorithmic solutions. In this section, a simple heuristic algorithm has been constructed to solve the optimization problem.

The algorithm proposed is based on two optimization techniques, namely: The marginal analysis and branch-and-bound technique. Basically, given a starting network configuration, the marginal analysis technique is used to arrive at a feasible solution. Then the branch-and-bound technique takes over in an attempt to find the local optimum solution. Repeating this process for different starting network configurations and selecting the best local solution, gives an approximation of the true or global optimum solution.

4.1 The Marginal Analysis Technique

In linear programming marginal analysis or the gradient search technique [15] is generally used in seeking out the optimum solution by improving the trial solution in the direction of greatest improvement. In this work, this technique is used to derive a feasible solution, not necessarily the optimum, from any given network configuration. A block diagram of the procedure is shown in Fig. 3.

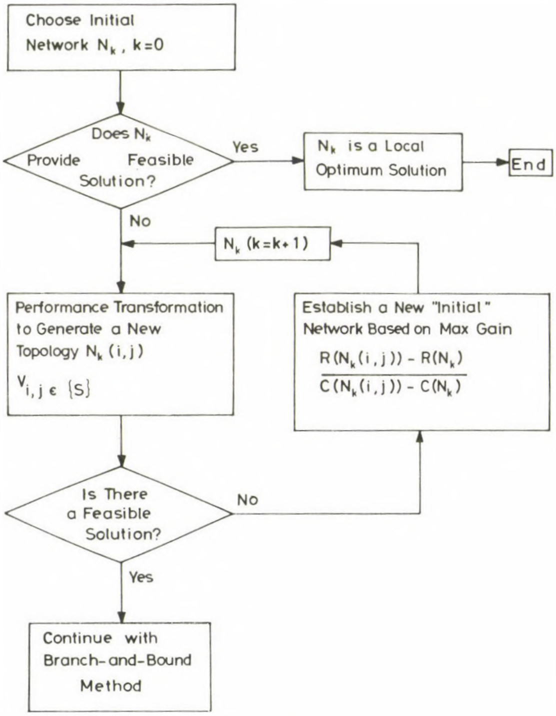

Choose Initial Network N_k, k=0
Does N_k Provide Feasible Solution?
Yes
N_k is a Local Optimum Solution
End
No
N_k (k=k+1)
Performance Transformation to Generate a New Topology $N_k(i,j)$ $\forall_{i,j} \epsilon \{S\}$
Establish a New "Initial" Network Based on Max Gain $\frac{R(N_k(i,j)) - R(N_k)}{C(N_k(i,j)) - C(N_k)}$
Is There a Feasible Solution?
No
Yes
Continue with Branch-and-Bound Method

Fig. 3

4.2 Branch-and-Bound Technique

The branch-and-bound technique [15], [16] is an optimization procedure, with principal advantages that it makes minimal assumptions about the structure of the problem and is fairly easy to implement. Consider $F^{(1)}$ and $\overline{F}^{(1)}$, which denote the subsets of feasible and infeasible solutions at level 1 of the branch-and-bound process. Level 0 of this process could be viewed as containing the network N_k (max k), established in the marginal analysis procedure. The following steps describe the branch-and-bound procedure:

Step 1: Establish a lower bound, $C(N_k(l, m))$ to the problem by selecting from the level 1 subset of feasible solution, $F^{(1)}$, the network $N_k(l, m)$ with the lowest cost. This rules out all the remaining elements in subset $F^{(1)}$ from further consideration. From hereon $N_k(l, m)$ shall be designated as N_{LB}, the network providing the lower bound on cost. Also set $i=1$.

Step 2: From the subset of infeasible solutions $\overline{F}^{(i)}$ generate subset $\overline{FF}^{(i)}$, consisting of level i networks with cost less than $C(N_{LB})$. Networks with cost greater than $C(N_{LB})$ are no longer considered. If subset $\overline{FF}^{(i)}$ has less than two elements, then there are no new topologies which could be generated with a cost less than $C(N_{LB})$ and the algorithm ends with N_{LB} being the local optimum solution. Otherwise, sort the elements in $\overline{FF}^{(i)}$ in ascending order of cost

$$\overline{FF}^{(i)}=\{N_{\overline{FF}_1}, N_{\overline{FF}_2}, \ldots, N_{\overline{FF}_s}\}$$

where

$$C(N_{\overline{FF}_1})<C(N_{\overline{FF}_2})<\ldots<C(N_{\overline{FF}_s})$$

this would minimize the remaining search.

Step 3: Generate a level $(i+1)$ topology, $N^{(i+1)}_{\overline{FF^{(i)}}}(p, q)$, by taking the union of the element $N_{\overline{FF}_p}$ and $N_{\overline{FF}_q}$, $(N_{\overline{FF}_p} \cup N_{\overline{FF}_q})$ where $((q=p+1,\ p+2, \ldots, s)$, $p=1, 2, \ldots, s-1)$.

If $C(N^{(i+1)}_{\overline{FF^{(i)}}}(p, q))>C(N_{LB})$ this step ends, as no further level $(i+1)$ network will be cheaper than $C(N_{LB})$. Otherwise, if $R(N^{(i+1)}_{\overline{FF^{(i)}}}(p, q)) \geqq \in$, then $N^{(i+1)}_{\overline{FF^{(i)}}}(p, q)$ provides as even lower bound on the cost.

Set $N_{LB}=N^{(i+1)}_{\overline{FF^{(i)}}}(p, q)$ and go to step 4. However, if $R(N^{(i+1)}_{\overline{FF^{(i)}}}(p, q)) < \in$ include $N^{(i+1)}_{\overline{FF^{(i)}}}(p, q)$ in the subset of level $(i+1)$ infeasible solution, $\overline{FF}^{(i+1)}$, and repeat this step.

Step 4: If subset $\overline{FF}^{(i+1)}$ has less than two elements the algorithm ends as no new network generated will cost less than $C(N_{LB})$. Otherwise, set $i=i+1$ and go to step 3.

4.3 Evaluation of the Algorithm

Among the other heuristic procedures used for solving the network topological design problem are the branch exchange and the exhaustive search methods. Both these methods require an exhaustive exploration of all the alternative solutions and tend to be very time consuming even when applied to moderately sized networks. The algorithm proposed here has the advantage over these methods in that it substantially reduces the search process, consequently reducing the computing time.

To obtain an appreciation for the saving in computing time, the algorithm is compared with a "restricted" form of the exhaustive search method. In the full exhaustive search method, all the topologies generated by the addition of one or more links to the starting network are examined. However, in the "restricted" form, only the topologies generated by the addition of a specified number of links to the starting network are investigated. To some extent, this is similar to the branch exchange method, where given an arbitrary starting configuration, it searches through all the local solutions by eliminating one or more old links and inserting one or more new links.

Table 1 shows the computing time for solving some small networks using both methods, the algorithm proposed in this work and the "restricted" form of the exhaustive search method. The starting configuration, the reliability constraint and the node-pair information were the same for both methods. It is worthwhile to point out the solutions obtained were identical for both cases. It is clearly seen that the algorithm proposed in this network is more efficient with the saving in computer time increasing considerably with the size of the network.

Table 1. Computing time for different sized network

Network Size (No. of Nodes)	Computing Time (Sec)	
	Proposed Algorithm	Exhaustive Search (Restricted Form)
5	2	2
6	8	28
7	28	283
8	130	>1200

An important question concerning heuristic generated solutions is how close are the characteristics of these solutions to the optimum solutions. To determine this, knowledge of the optimum solution is required. However, obtaining the optimum solution is a tedious if not impossible, as all possible starting configurations must be considered and for each one of these configurations an exhaustive search of all the local solutions must be carried out.

Table 2. Node-pair information for 5-node network

Links Source-Dest Node Pairs	Weights	Cost ($)	Link Availability
1 – 2	7.0	920	0.90
1 – 3	10.0	2420	0.90
1 – 4	7.0	2000	0.90
1 – 5	7.0	1010	0.90
2 – 3	6.0	2150	0.90
2 – 4	5.0	1850	0.90
2 – 5	5.0	1250	0.90
3 – 4	6.0	1400	0.90
3 – 5	6.0	2300	0.90
4 – 5	5.0	1850	0.90

In an attempt to evaluate the accuracy of heuristic generated solutions, a 5-node network was solved using the ring as the starting configuration. All the twelve possible ring configurations were considered. The weights were chosen arbitrarily and the cost was based on Telepak rate structure. The availability of all the links were assumed to be 0.9. The results obtained indicate that the global and local optimum solutions do not differ by more than 11%. This result confirms the claim made in [2] according to which this difference is in the order of 5% to 20%.

5. Simulation results

In this section, a few of the examples generated by the optimization algorithm, are presented. The two general categories of networks, centralized and distributed, are considered. The solutions obtained are compared with the star and loop solutions, for the centralized and distributed networks respectively.

5.1 Centralized Networks

In studying the design of centralized networks, subject to specified reliability constraints, comparisons are made with the star network. Furthermore, since in a star network, the remote nodes are connected directly to the CPU, it is assumed that its overall network reliability can only be increased, by introducing redundancy in the connections between the remote nodes and the CPU.

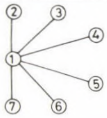
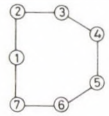
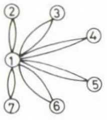
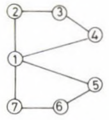
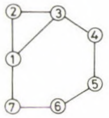
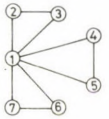
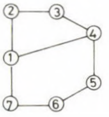
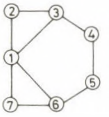
In the first example, the cost and reliability of different network topologies are compared. A seven node network is investigated, whose node pair information is given in Table 3 for link availability of 0.8. An identical network with link availability of 0.9 is also investigated. In this example, node 1 is assumed to be the central processing center or central node and it is assumed to be equidistant from the remote concentrators (Node 2—Node 7). The weights associated with the links are assumed to be equal. The cost of the links are not related to any particular cost structure and were chosen such that the cost of providing a communication link between the central node and any of the concentrators is twice the cost of providing a communication link between adjacent concentators. The cost of the other links vary with their relative position.

Table 3. Node-pair information for 7-node centralized network
$p = 0.8$

Links Source-Dest Node Pairs	Weights	Cost ($)	Link Availability
1 – 2	10.0	1500	0.80
1 – 3	10.0	1500	0.80
1 – 4	10.0	1500	0.80
1 – 5	10.0	1500	0.80
1 – 6	10.0	1500	0.80
1 – 7	10.0	1500	0.80
2 – 3	—	750	0.80
2 – 4	—	1000	0.80
2 – 5	—	1500	0.80
2 – 6	—	2000	0.80
2 – 7	—	2500	0.80
3 – 4	—	750	0.80
3 – 5	—	1000	0.80
3 – 6	—	2500	0.80
3 – 7	—	2000	0.80
4 – 5	—	750	0.80
4 – 6	—	1000	0.80
4 – 7	—	1500	0.80
5 – 6	—	750	0.80
5 – 7	—	1000	0.80
6 – 7	—	750	0.80

Table 4 shows examples of different network topologies. The last column represents the normalized cost with respect to the single line star configuration, topology *A*. The results indicate, the double loop configuration, topology *F*, although having the same cost as the single line star configuration, provides an overall network reliability of 0.892 and 0.970, an increase in reliability of 11.50% and 7.8%, for link availabilities of 0.8 and 0.9 respectively. Also, for $p = 0.9$, the reliability of topology *H* is approximately equal to the reliability of the double line star configuration, topology *B*, yet it could be achieved for about half its cost.

Table 4. Examples of centralized network topologies

Topology	Configuration	Network Reliability		Network Cost	Normalized Cost	Topology	Configuration	Network Reliability		Network Cost	Normalized Cost
		p=0.8	p=0.9	(3)				p=0.8	p=0.9	(3)	
A		0.800	0.900	9000	1.000	E		0.774	0.927	6750	0.750
B		0.960	0.990	18000	2.000	F		0.892	0.970	9000	1.000
C		0.871	0.966	8250	0.917	G		0.928	0.981	11250	1.250
D		0.892	0.974	8250	0.917	H		0.937	0.986	9750	1.083

5.2 *Distributed Networks*

For a distributed network, the overall reliability is defined as the average probability of any two nodes in the network being connected. In studying the optimum design of these networks, the loop or ring configuration will be used as a basis for comparison. The reason for this selection is that the loop is simple, flexible and practical for computer communication networks [17]. The disadvantage, however, is for larger networks this configuration has poor reliability. This is shown in Table 5(a) and in Fig. 4, where the relationship between reliability and network size (number of nodes), for link availability of 0.9, are given. In [17] a number of schemes were studied, for increasing the reliability of this configuration and found the bi-directional, double loop or full redundant scheme provides the best overall reliability for systems with less than 100 nodes. Table 5(b) and the corresponding plot, Fig. 4 demonstrate the high reliability the redundant scheme provides. The obvious disadvantage of this scheme is its cost. However, less reliable, yet acceptable networks at lower cost can be achieved by using partial redundancy.

Table 5. Relationship between the reliability of loop communication networks and networks size for $p=0.9$
(a) Single loop
(b) Double loop

Network Size (No. of Nodes)	Reliability of Loop Configuration
4	0.0699
5	0.9571
6	0.9428
7	0.9274
8	0.9111

(a)

Network Size (No. of Nodes)	Reliability of Loop Configuration
4	0.9997
5	0.9995
6	0.9993
7	0.9991
8	0.9988

(b)

As a last example discussed here a practical network is considered. The nodes of this network are assumed to be located at the main light cities across Canada (Toronto,

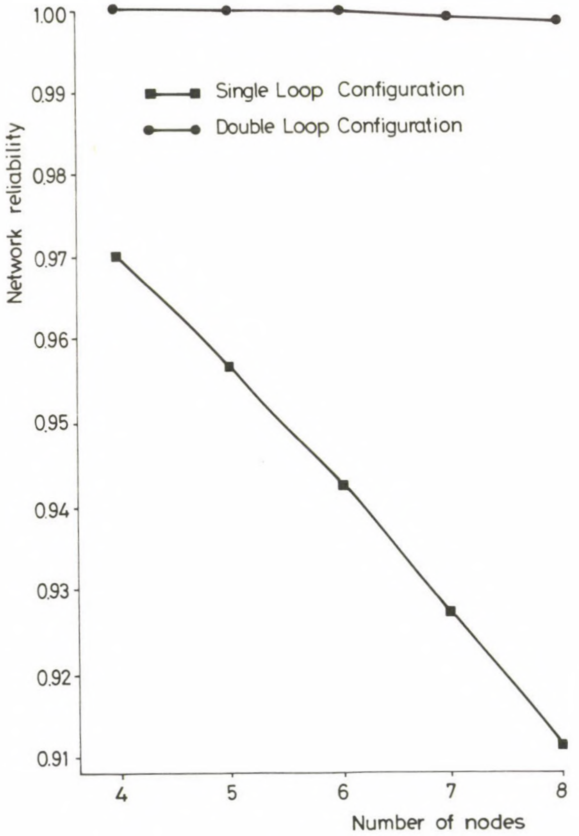


Fig. 4

ɔttawa, Montreal, Halifax, Winnipeg, Regina, Edmonton and Vancouver). Note that in this example, the weights were chosen on the assumption that it is very important for Toronto to communicate with other cities. The cost of the links were calculated on the basis of Telepak rates for a channel capacity of 19.2 K bits/sec. The total link availability was based on a terminal reliability of 0.9 and channel unavailability of 5.4×10^{-5}/mile (based on 7 mile hops). The details of the results will not be presented here due to lack of space but are available from the first author. The results for this example confirm earlier observations, that for a desired reliability requirement, the network generated by the optimization process, provides higher reliability with lower cost than the corresponding ring network.

6. Conclusions

The reliability measure proposed here is a simple extension of the terminal reliability method and takes into consideration the effects of all the terminal pair reliabilities in the network. Consequently, this measure is more representative of the overall network reliability and in addition, it is quite simple to compute. Like other methods used to compute network reliability, the disadvantage with this method is that for large distributed networks, the computing time is long. However, by considering only the important (higher weighted) node-pairs, approximate solutions could be obtained with considerable reduction in computing time. For centralized networks, where the prime concern is the communication between the remote nodes and the central processor, this is usually not a problem. The results confirm that the topology of maximally reliable networks depends on the availability of the individual network links. As this makes it difficult to derive computationally feasible algorithms, heuristic techniques are used. In general, the heuristic generated optimal solution is only an approximation of the true or global optimum. The heuristic procedure presented employs the marginal analysis and the branch-and-bound optimization techniques. This procedure minimizes the search process and consequently is efficient. The restricted class of networks considered here assumed the network nodes were perfect and the branch capacities were always adequate with uncorrelated failure probabilities. However, with little work, the failure probability of the nodes and the correlation between branch failures could be included. The effect of inadequate branch capacities on the network reliability is a more difficult problem and work in this area, together with the effects of software failure on the overall network reliability, is worthwhile. Finally, since most networks tend to expand soon after they are established, another problem needs investigation; namely how should the new node(s) be connected to the existing network, so as to have minimum effect on the optimal design.

References

1. *Fratta, L., Montanari, U. G.*, Synthesis of Available Networks, IEEE Trans. on Reliability, Vol. **R-25,** No. *2*, June 1976.
2. *Kimbleton, S. R., Schneider, G. M.*, Computer Communication Networks: Approaches, Objectives, and Performance Considerations, Computing Surveys, Vol. **7,** No. *3*, Sept. 1975.
3. *Gerla, M., Kleinrock, L.*, On the Topological Design of Distributed Computer Networks, IEEE Trans. on Communications, Vol. **COM-25,** No. *1*, Jan. 1977.
4. *Rai, S., Aggarwal, K. K.*, An Efficient Method for Reliability Evaluation of a General Network, IEEE Trans. on Reliability, Vol. **R-27,** No. *3*, August 1978.
5. *Aggarwal, K. K., Misra, K. B.*, A Fast Algorithm for Reliability Evaluation, IEEE Trans. on Reliability, Vol. **R-24,** No. *1*, April 1975.
6. *Kim, Y. H., Case, K. E., Ghare, P. M.*, A Method for Computing Complex System Reliability, IEEE Trans. on Reliability, Vol. **R-21,** No. *4*, Nov. 1972.
7. *DeMercado, J., Spyratos, N., Bowen, B. A.*, A Method for Calculation of Network Reliability, IEEE Trans. on Reliability, Vol. **R-25,** No. *2*, June 1976.
8. *Nelson, A. C., Batts, J. R., Beadles, R. L.*, A Computer Program for Approximating System Reliability, IEEE Trans. on Reliability, Vol. **R-19,** No. *2*, 1970.
9. *Kellmans, A. K.*, Connectivity of Probabilistic Network, Automat. Remote Contr., No. *3*, 1967.
10. *Frank, H., Frisch, I. T., Chow, W.*, Topological Considerations in the Design of the ARPA Computer Network, Spring Joint Computer Conference, 1970.
11. *Leggett, J. D., Bedrosian, S. D.*, Synthesis of Reliable Networks, IEEE Trans. on Circuit Theory, August 1969.
12. *Shooman, M. L.*, Probabilistic Reliability: An Engineering Approach, McGraw-Hill, 1968.
13. *Schwartz, M.*, Computer-Communication Network Design and Analysis, Prentice-Hall, 1977.
14. Bell Canada Ontario Region — Communications Handbook.
15. *De Neufille, R., Stafford, J. H.*, Systems Analysis for Engineers and Managers, McGraw-Hill, 1971.
16. *Aronofsky, J. S.*, Progress in Operations Research, Wiley, 1969.
17. *Akiyama, M., Sakaue, K.*, Reliability of Loop Communication Networks, Electronics and Communications in Japan, Vol. **60-A,** No. *3*, 1977.

Топологическая оптимизация сетей связи с ограничением по надёжности

А. Н. ВЕНЕЦАНОПОУЛОС, И. СИНГХ

(Торонто)

В статье рассматривается проблема оптимизации топологии сети в соответствии с определенными ограничениями по надежности. Дается новая мера надежности и предлагается новая евристическая процедура оптимизации. Приводятся примеры.

A. N. Venetsanopoulos
I. Singh
Department of Electrical Engineering
University of Toronto
Toronto, Ontario M5S 1A4
Canada

Problems of Control and Information Theory, Vol. 15 (1), pp. 79–89 (1986)

L-OPTIMAL INPUT SIGNALS FOR DISTRIBUTED-PARAMETER SYSTEMS IDENTIFICATION

E. RAFAJŁOWICZ

(*Wrocław*)

(Received December 15, 1984)

In the paper the problem of optimization of an input signal from the view-point of identification accuracy is considered for a class of linear time-invariant distributed-parameter systems. A linear functional of the inverse of the information matrix is chosen as the optimality criterion. Necessary and sufficient optimality conditions are derived. In contrast to earlier author's papers on *D*-optimal input signal choice, these conditions when applied to a certain subclass of considered problems lead to a closed-form solution. A straightforward application of these condition allows us to construct a computational algorithm working in an infinite dimensional function space. Such an algorithm is not presented here, since in a next paper efficient algorithm working in a finite dimensional space will be proposed.

1. Introduction and problem formulation

Developments in the area of distributed-parameter system (DPS) identification algorithms (see [3], [8] for survey papers) as well as a fruitful frequency-domain approach to the experiment design problem for lumped parameter system (LPS) identification (see e.g. [6], [2]) inspired a number of papers on experimental design for DPS identification [5], [9]–[13]. In these papers, excluding [11], [12], the distributed nature of inputs was not considered and the optimization has been confined to the temporal characteristics of the pointwise and boundary excitations. The determinant of the information matrix, i.e. *D*-optimality criterion have been used as measure of an estimation accuracy — in all these papers. On the other hand, the so-called *L*-optimality criterion, i.e. a linear functional of the inverse of the information matrix is known to be statistically meaningful and it is widely used in experimental design for LPS identification [1], [7].

The aim of this paper is to find necessary and sufficient optimality conditions for *L*-optimal input signals for identification of linear, time-invariant DPS. Extension of a frequency-domain approach from the LPS case to the present one seems not to be straightforward due to presence of spatial variables, varying in a bounded domain.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

As a basis for our study system description we choose the following

$$q(\kappa, t; a) = \int_0^t \int_{\Omega_u} G(\kappa, x, t-\tau; a) u(x, \tau)\, dx\, d\tau, \qquad \kappa \in \Omega, \quad t \in (0, T_0). \tag{1.1}$$

Here, $q(\kappa, t; a)$ denotes the system state at a spatial point κ of an open, bounded spatial domain Ω, resulting at time t on applying an input signal u, which is defined on $\Omega_u \times (0, T_0)$, where $\Omega_u \subset \Omega$ is a closed spatial domain. In (1.1), $G(\kappa, x, t; a)$ is the system Green's function (impulse response at κ, t on applying the Dirac delta excitation $\delta(t)$ at $x \in \Omega_u$), which is known to within a vector of unknown, constant parameters $a \in R^r$. These parameters influence the system state $q(\kappa_1, t; a)$ at a measurement point $\kappa_1 \in \Omega$, and they are estimated from the measurements $z(t)$, $t \in (0, T_0)$ of the form

$$z(t) = q(\kappa_1, t; a) + \varepsilon(t), \quad t \in (0, T_0), \tag{1.2}$$

where $\varepsilon(t)$ is the Gaussian white noise with zero mean and $E[\varepsilon(t)\varepsilon(t+\tau)] = \delta(t-\tau)$.

As is known (see e.g. [2]), the estimation accuracy of any asymptotically efficient estimator for a attains its lower bound as $T_0 \to \infty$. Under mild regularity condition this lower bound is equal to the inverse of Fisher's information matrix M_{T^0}, which is of the form (see [9]):

$$M_{T_0} = \int_0^{T_0} w(\kappa_1, t; a) w^T(\kappa_1, t; a)\, dt, \tag{1.3}$$

where the column vector $w(\kappa_1, t; a) = \operatorname{grad}_a q(\kappa_1, t; a)$, while Tdenotes transposition. As in earlier works on experimental design we choose the averaged information gathered per unit of experiment time as a basis for constructing an experimental optimality criterion, i.e.

$$M = \lim_{T_0 \to \infty} [M_{T_0}/T_0], \tag{1.4}$$

In order to ensure consistent estimation of a, we restrict the admissible input signals to those, for which M^{-1} exists. An input signal u is admitted to be a function or realization of a stochastic process, for which the following spectral representation exists:

$$S_u(x, y, j\omega) = \int_{-\infty}^{\infty} \left[\lim_{T_0 \to \infty} \frac{1}{T_0} \int_0^{T_0} [u(x, t+\tau) u(y, t)]\, dt \right] \cdot \exp(-j\omega\tau)\, d\tau \tag{1.5}$$

for $y, y \in \Omega_u$. Due to practical limitations we further restrict the admissible input signals to those with bounded spatially averaged mean power and having the spectrum contained in a bounded interval $[-\omega_0, \omega_0]$. In terms of spectral representation these constraints can be expressed as follows:

$$S_u(x, y, j\omega) = 0 \quad \text{for} \quad |\omega| > \omega_0, \quad x, y \in \Omega_u \tag{1.6}$$

$$\frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\int_{\Omega_U} S_u(x, x, j\omega)\,dx\,d\omega \leqq 1\,. \tag{1.7}$$

The above defined class of spectral representations will be denoted by $\mathscr{S}$. For ease of future references we remark that

$$S_u(x, y, j\omega) = \lim_{T_0\to\infty}\frac{1}{T_0}[\tilde{u}_{T_0}(x, j\omega)\tilde{u}_{T_0}(y, -j\omega)]\,, \tag{1.8}$$

where $\tilde{u}_{T_0}(x, j\omega)$ denotes the Fourier transform of the function

$$u_{T_0}(x, t) = \begin{cases} u(x, t) & \text{for } \ t\in(0, T_0) \\ 0 & \text{for } \ t \geqq T_0, \end{cases} \tag{1.9}$$

while constraint (1.7) can be expressed as follows

$$\lim_{T_0\to\infty}\frac{1}{2\pi T_0}\int_{-\omega_0}^{\omega_0}\int_{\Omega_U} |\tilde{u}_{T_0}(x, j\omega)|^2\,dx\,d\omega \leqq 1\,. \tag{1.10}$$

Using (1.1), (1.3), (1.4) and the Parseval equality, one can express M as follows

$$M(S_u) = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\int_{\Omega_U}\int_{\Omega_U} K(x, j\omega)K^T(y, -j\omega)S_u(x, y, j\omega)\,dx\,dy\,d\omega\,, \tag{1.11}$$

where

$$K(x, j\omega) \triangleq \int_{-\infty}^{\infty}[\operatorname{grad}_a G(\kappa_1, x, t; a)\cdot\exp(-j\omega t)]\,dt\,. \tag{1.12}$$

In order to stress the dependence of M on S_u the notation $M(S_u)$ is used in (1.11), while the dependence of $K(x, j\omega)$ and $M(S_u)$ on κ_1 and a is not made explicit since these variables remain constant. Define $\mathscr{M} = \{M(S_u): S_u \in \mathscr{S}\}$ and let $L[M^{-1}]$ be a linear positive functional of a matrix in brackets, i.e. for every hermitian, positive definite matrices A, B we require:

$$L[A] \geqq 0 \tag{1.13}$$

$$L[\alpha A] = L[A]\cdot\alpha\,, \tag{1.14}$$

$$L[A+B] = L[A] + L[B]\,. \tag{1.15}$$

We assume that L is a strictly convex and differentiable functional.

Remark 1. These requirements hold for $L[M^{-1}] = \operatorname{tr}[M^{-1}\cdot C]$, where C is a given positive definite matrix. Conversely, these assumptions imply L to be of the above trace form (see [7]).

For large T_0, $L_0[M^{-1}] \triangleq \operatorname{tr}[M^{-1}]$ is proportional to the mean variance of parameter estimates and it is known as *A*-optimality criterion. Our main problem is to find $\hat{S}_u \in \mathscr{S}$ such that

$$\forall S_u \in \mathscr{S} \qquad L[M^{-1}(\hat{S}_u)] \leqq L[M^{-1}(S_u)]\,. \tag{1.16}$$

Only analytical aspects of this problem are considered here. An efficient numerical algorithm will be proposed in a next paper. We remark that in [11] the following problem has been considered: find $S_u^* \in \mathscr{S}$ such that

$$\inf_{S_u \in \mathscr{S}} \det[M^{-1}(S_u)] = \det[M^{-1}(S_u^*)]\,, \tag{1.17}$$

where det [.] means the determinant of a matrix in brackets. In both cases, (1.16) and (1.17), necessary optimality conditions are relatively easy to obtain. The difficulty is in finding such forms of these conditions so as to assure sufficiency in each case (1.16) and (1.17). This will be done in Section 2 for the problem (1.16). Furthermore, an explicit solution for the problem (1.16) with L_0 criterion will be given in Section 3 for a wide class of systems (1.1).

2. Optimality conditions

It is easy to verify that the set $\mathscr{S}$ is convex. Let us assume that elements of the vector $K(x, j\omega)$ are linearly independent and continuous functions defined on $\Omega_u \times [-\omega_0, \omega_0]$. Then, reasoning exactly in the same way as in LPS case (see [1]), we obtain that the set $\mathscr{M}$ is convex, closed and bounded, which together with the continuity of $L[.]$, yields the existence of a solution of (1.16).

In order to derive optimality conditions for $\hat{S}_u$ we need the following function:

$$\hat{Q}(x, y, j\omega) \triangleq L[M^{-1}(\hat{S}_u) \cdot K(x, j\omega) K^T(y, -j\omega) M^{-1}(\hat{S}_u)]\,. \tag{2.1}$$

Assuming that for every w, $v \in R^r$ $L[v \cdot w^T] = L[w \cdot v^T]$, we get from (2.1)

$$\hat{Q}(x, y, j\omega) = \hat{Q}(y, x, -j\omega)\,. \tag{2.2}$$

Furthermore, for every $\omega \in [-\omega_0, \omega_0]$ the function $\hat{Q}(x, y, j\omega)$ is nonnegative definite in the following sense: for every complex valued, square integrable function f

$$\int_{\Omega_u} \int_{\Omega_u} \hat{Q}(x, y, j\omega) f(x) f^*(y)\, dx\, dy \geqq 0\,, \tag{2.3}$$

where $f^*(x)$ means the complex conjugate of $f(x)$. We also remark that $\hat{Q}(x, y, j\omega)$ can be expressed as a linear combination of the functions $K_i(x, j\omega) K_m(y, -j\omega)$; $i, m = 1, 2, \ldots, r$, where $K_i(x, j\omega)$ $i = 1, 2, \ldots, r$ are elements of the vector $K(x, j\omega)$. The above facts allow us to use known results of the theory of integral equations with symmetric, nonnegative definite, degenerated kernels (see. e.g. [14]) to study the equation:

$$\mu(\omega)\varphi(x,j\omega)=\int_{\Omega_u}\hat{Q}(x,y,j\omega)\varphi(y,j\omega)\,dy\,. \tag{2.4}$$

In this equation $\omega\in[-\omega_0,\omega_0]$ is treated as a parameter. From the above mentioned theory we have that (2.4) possesses a finite number of real, nonnegative eigenvalues $\hat{\mu}_1(\omega),\ldots,\hat{\mu}_p(\omega)$, while the corresponding eigenfunctions $\varphi_1(x,j\omega),\ldots,\varphi_p(x,j\omega)$ are orthogonal and they can be normalized in such a way that:

$$\int_{\Omega_u}\varphi_k(x,j\omega)\varphi_m(x,-j\omega)\,dx=\begin{cases}1 & \text{for}\quad k=m\\ 0 & \text{for}\quad k\neq m\,.\end{cases} \tag{2.5}$$

Lemma 1. Denote $\hat{M}=M(\hat{S}_u)$. Then,

$$\sup_{|\omega|\leqq\omega_0}\max_{1\leqq k\leqq p}[\hat{\mu}_k(\omega)]\geqq L[\hat{M}^{-1}]\,. \tag{2.6}$$

Proof.

$$L[\hat{M}^{-1}]=L[\hat{M}^{-1}\hat{M}\hat{M}^{-1}]=$$

$$=\frac{1}{2\pi}\int_{-\omega_0}^{\omega_0}\int_{\Omega_u}\int_{\Omega_u}\hat{Q}(x,y,j\omega)\,\hat{S}_u(x,y,j\omega)\,dx\,dy\,d\omega \tag{2.7}$$

where the last equality follows from the linearity and continuity of L. Using extremal properties of eigenvalues of integral equations with symmetric kernels [14] and (1.8) we obtain from (2.7) and (1.10):

$$L[\hat{M}^{-1}]=\lim_{T_0\to\infty}\frac{1}{2\pi T_0}\int_{-\omega_0}^{\omega_0}\int_{\Omega_u}\int_{\Omega_u}\hat{Q}(x,y,j\omega)\cdot\tilde{u}_{T_0}(x,j\omega)\cdot\tilde{u}_{T_0}(y,-j\omega)\,dx\,dy\,d\omega\leqq$$

$$\leqq\sup_{|\omega|\leqq\omega_0}\max_{1\leqq k\leqq p}\hat{\mu}_k(\omega)\cdot\lim_{T_0\to\infty}\frac{1}{2\pi T_0}\int_{-\omega_0}^{\omega_0}\int_{\Omega_u}|\tilde{u}_{T_0}(x,j\omega)|^2\,dx\,d\omega \tag{2.8}$$

for every $\omega\in[-\omega_0,\omega_0]$. □

We also need the following property of the eigenvalues:

$$\hat{\mu}_k(\omega)=\hat{\mu}_k(-\omega),\qquad k=1,2,\ldots,p\,, \tag{2.9}$$

which immediately follows from the equality:

$$\hat{\mu}_k(\omega)=\int_{\Omega_u}\int_{\Omega_u}\hat{Q}(x,y,j\omega)\varphi_k(y,j\omega)\varphi_k(x,-j\omega)\,dx\,dy \tag{2.10}$$

and from (2.2).

Theorem 1. The input spectral density $\hat{S}_u \in \mathscr{S}$ is L-optimal if and only if

$$\sup_{|\omega| \leqq \omega_0} \max_{1 \leqq k \leqq p} \hat{\mu}_k(\omega) = L[M^{-1}(\hat{S}_u)]\,. \tag{2.11}$$

Proof. In order to prove the necessity of (2.11) let us consider the spectral density of the form:

$$S_u^\alpha(x, y, j\omega) = (1-\alpha)\hat{S}_u(x, y, j\omega) + \alpha \cdot s(x, y, j\omega)\,, \tag{2.12}$$

where $s \in \mathscr{S}$ is as follows

$$s(x, y, j\omega) = \pi[\delta(\omega+\omega') + \delta(\omega-\omega')]\,\hat{\varphi}(x, j\omega')\hat{\varphi}(y, -j\omega')\,. \tag{2.13}$$

In (2.13), $\hat{\varphi}(x, j\omega')$ is the eigenfunction of (2.4) corresponding to the largest eigenvalue between $\hat{\mu}_k(\omega')$, $k=1, 2, \ldots, p$, while $|\omega'| \leqq \omega_0$. Convexity of $\mathscr{S}$ implies that for every $\alpha \in [0, 1]$, $S_u^\alpha \in \mathscr{S}$. From the optimality of $\hat{S}_u$ it follows that:

$$\frac{d}{d\alpha} L[M^{-1}(S_u^\alpha)]_{\alpha=0} < 0 \tag{2.14}$$

and on differentiation we obtain (see [1] for the rule of differentiation):

$$L[\hat{M}^{-1}M(s)\hat{M}^{-1}] \leqq L[\hat{M}^{-1}]\,. \tag{2.15}$$

Inequalities (2.9) and (2.13) yield

$$L[\hat{M}^{-1}] \geqq \max_{1 \leqq k \leqq p} \hat{\mu}_m(\omega') \text{ for every } |\omega'| \leqq \omega_0, \tag{2.16}$$

which together with Lemma 1 implies the necessity of (2.11).

Suppose that (2.11) holds but $\hat{S}_u$ is not L-optimal, i.e. there exists $S_u^0 \in \mathscr{S}$ such that $L[M^{-1}(S_u^0)] < L[M^{-1}(\hat{S}_u)]$. This fact and the convexity of L imply that

$$\frac{\partial}{\partial\alpha} L[M^{-1}(\tilde{S}_u^\alpha)]|_{\alpha=0} < 0\,, \tag{2.17}$$

where for $\alpha \in [0, 1]$ $\tilde{S}_u^\alpha \triangleq (1-\alpha)\hat{S}_u + \alpha S_u^0 \in \mathscr{S}$. Reasoning similarly as in (2.8), we obtain from (2.17) and (1.10)

$$L[\hat{M}^{-1}] < L[\hat{M}^{-1}M(S_u^0)\hat{M}^{-1}] = \tag{2.18}$$

$$= \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} \int_{\Omega_u} \int_{\Omega_u} \hat{Q}(x, y, j\omega) S_u^0(x, y, j\omega)\, dx\, dy\, d\omega \leqq \sup_{|\omega| \leqq \omega_0} \max_{1 \leqq k \leqq p} \hat{\mu}_k(\omega)\,,$$

which contradicts (2.11) and proves its sufficiency.

Collorary 1. Let $L_0[M^{-1}]=\mathrm{tr}\,[M^{-1}]$. If for every $\omega\in[-\omega_0,\omega_0]$ the elements of the vector $K(x,j\omega)$ are linearly independent functions over Ω_u, then $\hat{S}_u\in\mathscr{S}$ is L_0-optimal if and only if

$$\sup_{|\omega|\leqq\omega_0}\lambda_{\max}[M^{-1}(\hat{S}_u)\cdot\mathscr{K}(j\omega)M^{-1}(\hat{S}_u)]=\mathrm{tr}\,[M^{-1}(\hat{S}_u)]\,, \tag{2.19}$$

where $\lambda_{\max}[\,.\,]$ denotes the largest eigenvalue of a matrix in brackets, while

$$\mathscr{K}(j\omega)=\int_{\Omega_u}K(x,j\omega)K^T(x,-j\omega)\,dx\,. \tag{2.20}$$

Proof. Let us look for the eigenfunctions of (2.4) in the form $\varphi(x,j\omega)=h^T(j\omega)\hat{M}^{-1}K(x,j\omega)$, where $h(j\omega)$ is $r\times 1$ vector to be chosen. Substituting this function in (2.4) and using the linear independence of elements of $K(x,j\omega)$ as functions of x, we obtain

$$\mu(\omega)h^T(j\omega)=h^T(j\omega)M^{-1}\mathscr{K}(j\omega)\hat{M}^{-1}\,. \tag{2.21}$$

Thus, if $h_1(j\omega),\ \ldots,\ h_r(j\omega)$ are eigenvectors of the matrix $\hat{M}^{-1}\mathscr{K}(j\omega)\hat{M}^{-1}$ then $\hat{\mu}_1(\omega),\ \ldots,\ \hat{u}_r(\omega)$ are simultaneously eigenvalues of (2.4) and (2.21). This finishes the proof by invoking Theorem 1 since (2.4) with the kernel $\hat{Q}(x,y,j\omega)=K^T(y,\ -j\omega)\hat{M}^{-2}K(x,j\omega)$ possesses at most r eigenfunctions. □

3. L_0-optimal experimental design

In this section Collorary 1 is used to find a closed form solution of the L_0-optimal experimental design problem. To this aim we have to restrict a class of considered DPS to those, for which $\mathscr{K}(j\omega)$ is a real matrix, i.e.

$$\mathscr{K}(j\omega)=\mathscr{K}(-j\omega),\ \omega\in[-\omega_0,\omega_0]\,, \tag{3.1}$$

where $\mathscr{K}(j\omega)$ is defined by (2.20).

Collorary 2. Let all the assumptions of Collorary 1 and (3.1) hold. Suppose that there exists $\hat{\omega}\in[-\omega_0,\omega_0]$ such that

$$\sup_{|\omega|\leqq\omega_0}\lambda_{\max}\,[\mathscr{K}^{-1/2}(j\hat{\omega})\mathscr{K}(j\omega)\mathscr{K}^{-1/2}(j\hat{\omega})]=1\,. \tag{3.2}$$

Then the L_0-optimal experimental design is of the form:

$$\begin{aligned}\hat{S}_u(x,y,j\omega)=K^T(x,-j\omega)\mathscr{K}^{-3/2}(j\omega)K(y,j\omega)\cdot[\delta(\omega+\hat{\omega})+\\ +\delta(\omega-\hat{\omega})]\cdot\pi/\mathrm{tr}\,[\mathscr{K}^{-1/2}(j\omega)]\end{aligned} \tag{3.3}$$

or

$$S_u(x, y, j\omega) = K^T(x, -j\hat{\omega})\mathscr{K}^{-3/2}(j\hat{\omega})K(y, j\hat{\omega})\,[\delta(\omega+\hat{\omega})+ \\ +\delta(\omega-\hat{\omega})]\pi/\mathrm{tr}\,[\mathscr{K}^{-1/2}(j\hat{\omega})]\,. \tag{3.4}$$

In both cases $\hat{M} = \mathscr{K}(j\hat{\omega})/\mathrm{tr}\,[\mathscr{K}^{-1/2}(j\hat{\omega})]$.

The proof follows from (2.19) by straightforward inspection. We remark that (3.2) is equivalent to the following condition:

$$\sup_{|\omega|\leq\omega_0} \lambda_{\max}\,[\mathscr{K}^{-1}(j\hat{\omega})\mathscr{K}(j\omega)] = 1\,. \tag{3.5}$$

In order to indicate that (3.1) holds for a large class of DPS, let us dispense ourselves from the precision of formulations and consider a system described by

$$P\left(\frac{\partial}{\partial t}\right)q(x,t) = A_x(a)q(x,t)+u(x,t), \quad x\in\Omega, \quad t>0\,, \tag{3.6}$$

where $P\left(\frac{\partial}{\partial t}\right)$ is formal differential operator obtained by replacing a variable z in the given polynomial $P(z) = b_m z^m + b_{m-1}z^{m-1} + \ldots + b_1 z$ by the operator $\frac{\partial}{\partial t}$.

In (3.6), $A_x(a) = \sum_{i=1}^{s} a_i Q_i$ is a spatial differential operator, which depends on the vector a of unknown, constant parameters a_i, $i=1, 2, \ldots, s$, while Q_i, $i=1, 2, \ldots, s$, are given spatial differential operators. Let us suppose that $A_x(a)$ possesses a complete set of orthonormal eigenfunctions $v_1(x)$, $v_2(x)$, ..., which simultaneously are eigenfunctions of the operators Q_i, $i=1, 2, \ldots, s$, i.e.

$$Q_i v_k(x) = \alpha_{ki}v_k(x), \qquad i=1,2,\ldots,s; \quad k=1,2,\ldots \tag{3.7}$$

where α_{ki} is the corresponding eigenvalue. Under (3.7), the eigenfunctions of $A_x(a)$ do not depend on the vector a. Furthermore, the eigenvalue $\lambda_k(a)$, $k=1, 2, \ldots$ of $A_x(a)$, associated with $v_k(x)$, $k=1, 2, \ldots$, are of the form

$$\lambda_k(a) = \bar{\alpha}_k^T a, \qquad k=1,2,\ldots \tag{3.8}$$

where $\bar{\alpha}_k = [\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{ks}]^T$. Assuming the existence and uniqueness of solution of (3.6) and its differentiability with respect to a, we obtain from (1.12):

$$K(x, j\omega) = -\sum_{k=1}^{\infty} v_k(x)v_k(\kappa_1)\alpha^k/(P(j\omega)+\lambda_k(a))^2\,. \tag{3.9}$$

Thus, (2.20) yields

$$\mathscr{K}(j\omega) = \sum_{k=1}^{\infty} \bar{\alpha}_k\bar{\alpha}_k^T v_k^2(\kappa_1)/|P(j\omega)+\lambda_k(a)|^4\,. \tag{3.10}$$

The above considerations, although informal, shows that condition (3.1) holds for a large class of system described by hyperbolic and parabolic partial differential equations. In order to illustrate an application of Collorary 2, let us consider the following example.

Example. Let the transfer function of a system be of the form:

$$\tilde{G}(\kappa, x, j\omega; a) = k_0 \exp\left[-j\omega \cdot a^T \varphi(\kappa - x)\right], \tag{3.11}$$

where $a \in R^r$ is the vector of unknown parameters, $\varphi(x) \triangleq [\varphi_1(x), \varphi_2(x), \ldots, \varphi_r(x)]^T$ is the vector of given, linearly independent functions on $\Omega \triangleq (0, L)$, $L > 0$, while $\tilde{G}$ is defined as the Fourier transform (with respect to time variable t) of the system pulse response $G(\kappa, x, t; a)$. Such systems are sometimes called systems with distributed delays.

According to (1.12) we have

$$K(x, j\omega) = -j\omega \cdot k_0' \exp\left[-j\omega \cdot a^T \varphi(\kappa - x)\right] \cdot \varphi(\kappa - x)\,. \tag{3.12}$$

Thus, from (2.20) we obtain

$$\mathscr{K}(j\omega) = \omega^2 \Phi, \qquad \Phi \triangleq k_0^2 \cdot \int_{\Omega} \varphi(\kappa - x)\varphi^T(\kappa - x)\,dx \tag{3.13}$$

and the matrix Φ is nonsingular, due to the linear independence of $\varphi_i(\,.\,)$, $i = 1, 2, \ldots, r$. We do not indicate the dependence of Φ on κ, since the sensor location remains constant. Let us consider problem (1.16) with L_0 criterion and fixed $\omega_0 > 0$. The matrix Φ is real and we can apply Collorary 2. It is easy to see that for $\hat{\omega} = \omega_0$ condition (3.5) holds, hence the L_0-optimal experimental design is of the form:

$$\hat{S}_u(x, y, j\omega) = k_1 \varphi^T(\kappa - x)\Phi^{-3/2}\varphi(\kappa - y)\pi\left[\delta(\omega + \omega_0) + \delta(\omega - \omega_0)\right],$$

where $k_1 \triangleq k_0^2 \cdot \omega_0^{(r/2-1)}/\mathrm{tr}\,(\Phi^{-1/2})$.

Concluding remarks

In the paper the problem of optimal experimental design with *L*-optimality criterion has been considered. A simplifying assumption concerning infinite observation time allowed us to use the inverse of information matrix instead of the exact covariance matrix of estimates. Such an approach has two advantages, namely: closed-form formula for the averaged information matrix can be obtained and the resulting experimental design is independent of the identification method used, provided that this method is asymptotically efficient. The necessary and sufficient optimality condition of *L*-optimality obtained in section 2, occurred to be easily applicable, leading to the closed-form formulas for experimental design minimizing averaged estimates errors. Inspection of formula (3.4) leads to the conclusion that the optimal experimental design can be realized by a harmonic input signal whose spatial intensity depends on the appropriately weighted vector of sensitivities to parameter changes. We also remark that the optimal experimental design may depend on unknown parameters to be estimated. This difficulty is inherent to almost all optimal experimental design problems in dynamic systems identification, see e.g. [2], [6], [12], where also ways of overcoming this problem are discussed. The most common of them is to replace the vector a in the optimal design by its a priori estimate.

References

1. *Fedorov, U. V.*, Theory of Optimal Experiments. Academic Press, New York and London, 1972.
2. *Goodwin, G. C., Paÿne, L.*, Dynamic System Identification: Experiment Design and Data Analysis. Academic Press, New York and London, 1977.
3. *Kubrusly, C. S.*, Distributed parameter system identification. A survey. Int. J. Control, vol. **26,** No. *4*, pp. 509–535, 1977.
4. *Kubrusly, C. S., Malebranche, H.*, A survey on optimal sensors and controllers location in DPS. Proc. 3-rd IFAC Symposium on Control of Distributed Parameter Systems, S. P. 59–73, Toulouse, June 1982.
5. *Kuszta, B., Sinha, N. K.*, Design of optimal input signals for the identification of distributed-parameter systems. Int. J. Syst. Sci., vol. **9,** pp. 1–7, 1978.
6. *Mehra, R. K.*, Optimal input signals for parameter estimation in dynamic systems — survey and new results. IEEE Trans. Aut. Contr., vol. **AC-21,** pp. 55–64, 1976.
7. *Pazman, A.*, Some features of the optimal design theory. A survey. Math. Operationsforsch. Statist. ser. Statistics, vol. **11,** pp. 415–446, 1980.
8. *Polis, M. P., Goodson, R. E.*, Parameter identification in distributed systems. A synthesing overwiev. Proc. IEEE, vol. **64,** pp. 45–61, 1976.
9. *Qureshi, Z. H., Ng, T. S., Goodwin, G. C.*, Optimum experimental design for identification of distributed systems. Int. J. Contr., vol. **31,** pp. 21–29, 1980.
10. *Rafajłowicz, E.*, Design of experiments for eigenvalue identification in distributed-parameter systems. Int. J. Contr., vol. **34,** pp. 1079–1094, 1981.
11. *Rafajłowicz, E.*, Optimal input signals for parameter estimation in linear distributed-parameter systems. Int. J. Syst. Sci., vol. **13,** pp. 798–809, 1982.
12. *Rafajłowicz, E.*, Optimum experiment design for parameter identification in distributed systems: Brief survey and new results. Proceedings of 9th World Congress of IFAC'84 (in the press).
13. *Rafajłowicz, E.*, Optimal experiment design for identification of linear distributed-parameter systems. Frequency domain approach. IEEE Trans. Aut. Contr., vol. **AC-28,** pp. 806–808, 1983.
14. *Yosida, K.*, Lectures on Differential and Integral Equations. Interscience Publishers, New York, 1960.

Линейно-оптимальные управляющие сигналы для идентификации систем с распределёнными параметрами

Э. РАФАЙЛОВИЧ

(Вроцлав)

В работе рассматривается проблема выбора управления оптимального с точки зрения качества идентификации параметров линейной системы с распределёнными параметрами. Качество оценивания измеряется линейными функционалами обратной информационной матрицы. Выведены необходимые и достаточные условия оптимальности управления в классе ограниченных по средней мощности сигналов. Используя эти условия, получено аналитическое решение задачи для некоторого класса систем. Во второй части этой работы полученные условия будут служить отправной точкой для конструкции эффективного численного алгоритма.

E. Rafajłowicz
Institute of Engineering Cybernetics,
Technical University of Wrocław,
Wybrzeże Wyspiańskiego 27, 50 370 Wrocław,
Poland.

Problems of Control and Information Theory, Vol. 15 (1), pp. 91–100 (1986)

KEY DISTRIBUTION SYSTEMS BASED ON POLYNOMIAL FUNCTIONS AND ON RÉDEI-FUNCTIONS

R. NÖBAUER

(*Klagenfurt*)

(Received November 20, 1984)

This paper gives a new cryptographic application of the Dickson-polynomials and of a class of rational functions which firstly have been studied by L. Rédei: It is shown how these functions can be used to construct key distribution systems. Further, all key distribution systems of a certain type are determined.

1. Introduction

A cryptosystem is a family of invertible functions $\{f_k:\mathcal{M}\to\mathcal{C}\}_{k\in\mathcal{K}}$ from the plaintext alphabet $\mathcal{M}$ onto the code alphabet $\mathcal{C}$. Having chosen a key $k\in\mathcal{K}$, the message M can be encrypted by representing it as a sequence of elements of the plaintext alphabet

$$M=M_1,\ldots,M_n,\qquad M_i\in\mathcal{M}\quad\text{for}\quad i=1,\ldots,n$$

and by encrypting the message blocks M_i with the encryption function f_k:

$$C_i=f_k(M_i),\qquad i=1,\ldots,n.$$

If A and B want to communicate in a secret way, they first have to exchange a key k_{AB}. In classic cryptography this is done by using a key-channel which has to be absolutely secure against interception—in contrast to the message-channel, which need not be secure. Modern key distribution systems allow the exchange of a key in a public way, without using a secure channel. One realization of such a public key distribution system was proposed by Diffie and Hellman [1]. Their scheme is based on power functions in Galois fields and will be described in the following section. A functional equation of the Diffie–Hellman scheme has been discussed by Henze [2], the application in a hybrid cryptographic system has been proposed by Horak [3], and a generalization to a conference key distribution system has been given by Ingemarsson, Tang and Wong [4].

Akadémiai Kiadó, Budapest
Pergamonn Press, Oxford

2. The Diffie–Hellman scheme

Let p be a prime number for which $p-1$ has a large prime factor p'. Let GF(p) be the Galois field of order p, and let x_0 be a primitive element of GF(p). Let GF(p) and x_0 be known to everyone.

Every potential communication participant C chooses a natural number $k(C)$ which he keeps secret. Further he calculates the value $x_0^{k(C)}$ and stores it in a public file.

If A and B want to exchange a key k_{AB}, the following process is going on:

(1) By inspecting the public file, A gets the information $x_0^{k(B)}$.
(2) A calculates $(x_0^{k(B)})^{k(A)}$.
(3) By inspecting the public file, B gets the information $x_0^{k(A)}$.
(4) B calculates $(x_0^{k(A)})^{k(B)}$.

The value $x_0^{k(A)k(B)}$ is taken as common key k_{AB}.

For the computation of the powers the "square-and-multiply-technique" can be used: For calculating y^l first do successive squaring

$$y, y^2, (y^2)^2, \ldots,$$

and then multiply together the appropriate factors. Using this method, only $0(ld(l))$ field multiplications are required.

Now say a spy wants to get the key k_{AB}. The information available to him is x_0, $x_0^{k(A)}$ and $x_0^{k(B)}$. For computing the key $k_{AB}=x_0^{k(A)k(B)}$ he might try to compute either $k(A)$ or $k(B)$ and thus faces the following problem: Given x_0 and y, calculate a k, w.l.g. $0 \leqq k \leqq q-2$, such that $x_0^k = y$. The number k is called the discrete logarithm of y to the base x_0. Until now no fast algorithm is known for calculating discrete logarithms in fields GF(p) for which $p-1$ contains a large prime factor (cf. [8], [9]). So a spy has no chance to compute the values $k(A)$ and $k(B)$. Of course we cannot exclude the possibility that there might be some way to generate $x_0^{k(A)k(B)}$ from knowledge of $x_0^{k(A)}$ and $x_0^{k(B)}$ only, without computing either $k(A)$ or $k(B)$, although it seems unlikely that such a method exists.

3. Polynomial generalizations

Let us give the following

Definition 1. A PDHS (= polynomial Diffie–Hellman scheme) is a triple $\mathcal{R}=(\mathrm{GF}(q), x_0, \{p_k\}_{k\in N})$, where GF($q$) is a Galois field of order $q=p^n$, $x_0 \in \mathrm{GF}(q)$, N denotes the natural numbers, p_k is a polynomial of degree k over GF(q) for all $k \in N$, and the p_k satisfy the following two conditions:

(i) $p_k \circ p_l = p_l \circ p_k$ for all $k, l \in N$, where $\circ$ denotes composition of polynomials
(ii) the complexity of calculating function values $p_k(y)$, $y \in \mathrm{GF}(q)$, is $0(ld(k))$.

If additionally to (i) and (ii) condition (iii) holds,

(iii) the complexity of the following problem is large: Knowing x_0 and $p_k(x_0)$, calculate an l such that $p_l(x_0)=p_k(x_0)$,

then $\mathscr{R}$ can be used exactly like the original Diffie–Hellman scheme, the polynomials p_k corresponding to the monomials x^k.

We are now going to determine all PDHS with q odd. For this purpose we use the concept of P-chains (= permutable chains) (cf. [5]): A family C of polynomials over the arbitrary field Q is called a P-chain over Q, if every polynomial in C is of degree >0, if for $k>0$ there exists a polynomial in C of degree k, and if $f(x)$, $g(x)$ commute, that is $f \circ g = g \circ f$ for all $f, g \in C$. Since every P-chain over Q contains exactly one polynomial of degree k for all $k \in N$ (cf. [5]), we can always denote the index set by N. We mention that P-chains have an important application in the theory of public key cryptosystems (cf. [6]).

If $C=\{f_k\}_{k\in N}$ is a P-chain over Q and if $l(x)=rx+s$, $r\neq 0$, is a linear polynomial over Q, then evidently $C_1=\{l^{-1}\circ f_k \circ l\}_{k\in N}$ is a P-chain, too, and C_1 is called conjugate (over Q) of C. In our context the following theorem is of central importance: Every P-chain over Q is some conjugate of either the chain of powers $S_1=\{x, x^2, x^3, \ldots\}$ or the chain of Chebyshev-polynomials $S_2=\{t_k | t_k$ the k^{th} Chebyshev polynomial of the first kind$\}_{k\in N}$ (cf. [5]).

Since the family $\{p_k\}_{k\in N}$ of a PDHS is per definition a P-chain, up to conjugacy there exist at most two PDHS with a given GF(q) and x_0. To show that there really exist two PDHS we have to give an evaluation algorithm for the $t_k(x)$ which has complexity $0(ld(k))$.

Since we restrict ourselves to the case q odd, we can replace S_2 by the conjugate chain $S_2'=\{g_k(x)\}_{k\in N}$ where $g_k(x)=2t_k(x/2)$. The polynomials $g_k(x)$ are called Dickson-polynomials; an explicit formula is given by

$$g_k(x)=\sum_{i=0}^{[k/2]}\frac{k}{k-i}\binom{k-i}{i}(-1)^i x^{k-2i} \tag{1}$$

(cf. [7]). We remind that in GF(q)(x), the field of rational functions over GF(q), the following equation holds (cf. [7]):

$$g_k\left(u+\frac{1}{u}\right)=u^k+\frac{1}{u^k}. \tag{2}$$

Let $y=\mathrm{GF}(q)$, we want to calculate $g_k(y)$. For doing this we have to solve

$$u+\frac{1}{u}=y \tag{3}$$

or equivalently

$$u^2-yu+1=0 \tag{4}$$

in some extension ring of GF(q).

It can be easily seen, that $R_y=\mathrm{GF}(q)[u]/(u^2-yu+1)$ is an extension ring of GF(q) and that every element $r\in R_y$ can be uniquely represented in the form

$$r=au+b, \qquad a, b\in \mathrm{GF}(q).$$

If u^2-yu+1 is irreducible over GF(q), then clearly R_y is the quadratic extension field of GF(q). Multiplication in R_y can be easily implemented by using the formula

$$(a_1u+b_1)(a_2u+b_2)=(a_1b_2+a_2b_1+a_1a_2y)u+b_1b_2-a_1a_2. \tag{5}$$

Obviously the element u is a solution of (4). Since $u(y-u)=1$, u is always invertible.

For evaluating $g_k(y)$ just calculate the power u^k in the ring R_y by using the square-and-multiply-technique, that is find elements $a, b\in \mathrm{GF}(q)$ with

$$u^k=au+b.$$

Since u^{-1} also satisfies (4), the equation

$$\frac{1}{u^k}=a\frac{1}{u}+b$$

holds, and therefore

$$g_k(y)=g_k\left(u+\frac{1}{u}\right)=u^k+\frac{1}{u^k}=a\left(u+\frac{1}{u}\right)+2b=ay+2b.$$

The number of required steps is $O(ld(k))$.

Altogether we have proved the following

Theorem 1. All PDHS with q odd are given by $\mathscr{R}=(\mathrm{GF}(q), x_0, \{l^{-1}\circ f_k\circ l\}_{k\in N})$, where $x_0\in\mathrm{GF}(q)$, the f_k are the power polynomials x^k or the Dickson polynomials g_k and l is a linear polynomial $rx+s$ with $r\neq 0$.

We next discuss how to choose the element x_0. We want to choose x_0 in such a way that the complexity of the following problem P is large: Knowing x_0 and $g_k(x_0)$ calculate an $l\in N$ with $g_l(x_0)=g_k(x_0)$. It seems to be quite impossible to give a sufficient condition on x_0; however, we will state a necessary one.

A special algorithm for solving the problem P is given by successive testing of randomly chosen numbers l. Since for naturals k, l with $k\equiv l \bmod q^2-1$ we have $g_k(x_0)=g_l(x_0)$ (cf. [6]), we might restrict ourselves to numbers l with $0\leqq l\leqq q^2-2$, or, equivalently, to residue classes $\bar{l}$ of $Z/(q^2-1)$, the ring of residue classes of the integers modulo q^2-1. Let L be a uniformly distributed random variable on $Z/(q^2-1)$ and let for every residue class $\bar{k}\in Z/(q^2-1)$ the probability $p_0(\bar{k})$ be given by

$$p_0(\bar{k})=P(\omega\,|\,g_{L(\omega)}(x_0)=g_k(x_0)). \tag{6}$$

Then the condition

$$p_0(\bar{k}) < \varepsilon \qquad \text{for all} \quad k \in N \tag{7}$$

with ε depending on the computer power available (e.g. $\varepsilon = 10^{-50}$) is necessary for a high complexity of the problem P. Indeed, if (7) were insulted, that is if numbers k existed with $p_0(\bar{k}) \geqq \varepsilon$, then for these k the successive testing algorithm would be computationally feasible. If we define subsets $D_k(x_0)$ of $Z/(q^2-1)$ by $D_k(x_0) = \{\bar{l} \in Z/(q^2-1) \mid g_l(x_0) = g_k(x_0)\}$ we obtain $p_0(\bar{k}) = \dfrac{|D_k(x_0)|}{q^2-1}$, and another formulation of (7) is given by

$$\frac{|D_k(x_0)|}{q^2-1} < \varepsilon \qquad \text{for all} \quad k \in N. \tag{8}$$

A formula for $|D_k(x_0)|$ is contained in the following

Theorem 2. Let $x_0 \in \mathrm{GF}(q)$, let $u \in \mathrm{GF}(q^2)$ be a solution of $x_0 = u + \dfrac{1}{u}$, and let $\mathrm{ord}\,(u)$ be the multiplicative order of u. Then we have:

(i) $D_l(x_0) = D_k(x_0) \Leftrightarrow l \equiv k \bmod \mathrm{ord}\,(u)$ or
$l \equiv -k \bmod \mathrm{ord}\,(u)$.

(ii) If $\mathrm{ord}\,(u)$ is even then

$$\left.\begin{array}{l} k \equiv 0 \bmod \mathrm{ord}\,(u) \\ k \equiv \dfrac{\mathrm{ord}\,(u)}{2} \bmod \mathrm{ord}\,(u) \end{array}\right\} \Rightarrow |D_k(x_0)| = \frac{q^2-1}{\mathrm{ord}\,(u)}$$

$$k \text{ else} \qquad \Rightarrow |D_k(x_0)| = \frac{2(q^2-1)}{\mathrm{ord}\,(u)}$$

If $\mathrm{ord}\,(u)$ is odd then

$$k \equiv 0 \bmod \mathrm{ord}\,(u) \qquad \Rightarrow |D_k(x_0)| = \frac{q^2-1}{\mathrm{ord}\,(u)}$$

$$k \text{ else} \qquad \Rightarrow |D_k(x_0)| = \frac{2(q^2-1)}{\mathrm{ord}\,(u)}.$$

Proof. (i) We have the following equivalences: $D_l(x_0) = D_k(x_0) \Leftrightarrow g_l(x_0) = g_k(x_0) \Leftrightarrow u^l + \dfrac{1}{u^l} = u^k + \dfrac{1}{u^k} \Leftrightarrow u^l = u^k$ or $u^l = u^{-k} \Leftrightarrow u^{l-k} = 1$ or $u^{l+k} = 1 \Leftrightarrow l \equiv k \bmod \mathrm{ord}\,(u)$ or $l \equiv -k \bmod \mathrm{ord}\,(u)$.

(ii) Let η be the canonical epimorphism from $G_1 = (Z/(q^2-1), +)$ onto $G_2 = (Z/(\mathrm{ord}\,(u)), +)$ and let $\{E_0, E_1, \ldots, E_{\mathrm{ord}\,(u)-1}\}$ be the coset decomposition of G_1 with respect to G_2. Because of (i) $\bar{l} \in D_k(x_0)$ holds iff $l \equiv k \bmod \mathrm{ord}\,(u)$ or

$l \equiv -k \bmod \operatorname{ord}(u)$, and this is equivalent to $l \in E_k$ or $l \in E_{-k}$. The equation $E_k = E_{-k}$ holds iff $k \equiv -k \bmod \operatorname{ord}(u)$, hence iff

$$2k \equiv 0 \bmod \operatorname{ord}(u). \tag{9}$$

If $\operatorname{ord}(u)$ is even, then (9) has the solutions $0 \bmod \operatorname{ord}(u)$ and $\frac{\operatorname{ord}(u)}{2} \bmod \operatorname{ord}(u)$. Thus all classes $D_k(x_0)$, $k \in N$, are given by E_0, $E_{\frac{\operatorname{ord}(u)}{2}}$, $E_1 \cup E_{\operatorname{ord}(u)-1}, \ldots,$ $E_{\frac{\operatorname{ord}(u)}{2}-1} \cup E_{\frac{\operatorname{ord}(u)}{2}+1}$, and by using $|E_i| = \frac{q^2-1}{\operatorname{ord}(u)}$, $i=0, \ldots, \operatorname{ord}(u)-1$, the result follows.

If $\operatorname{ord}(u)$ is odd, then the only solution of (9) is $0 \bmod \operatorname{ord}(u)$. Thus all classes $D_k(x_0)$, $k \in N$, are given by E_0, $E_1 \cup E_{\operatorname{ord}(u)-1}, \ldots, E_{\frac{\operatorname{ord}(u)-1}{2}} \cup E_{\frac{\operatorname{ord}(u)+1}{2}}$, and again the result follows.

Thus x_0 has to be chosen in such a way that for the corresponding u we have $\frac{2}{\varepsilon} < \operatorname{ord}(u)$. Additionally we obtain: The larger $\operatorname{ord}(u)$ is, the better is the choice of x_0. We might ask which is the best x_0. According to [7], all elements $u \in \mathrm{GF}(q^2)$ satisfying an equation

$$x = u + \frac{1}{u}, \qquad x \in \mathrm{GF}(q),$$

are given by $\omega^{(q-1)r}$, $r=0, 1, \ldots, q$, and $\omega^{(q+1)s}$, $s=0, 1, \ldots, q-2$, where ω is a primitive element of $\mathrm{GF}(q^2)$. Thus the maximum possible value of $\operatorname{ord}(u)$ is given by $q+1$, and this value is obtained by choosing $x_0 = \omega^{q-1} + \frac{1}{\omega^{q-1}}$.

4. Generalizations based on rational functions

A wider generalization of the original Diffie–Hellman scheme is given by the following

Definition 2. A MDHS (= modified Diffie–Hellman scheme) is a triple $\mathscr{R} = (\mathrm{GF}(q), x_0, \{f_k\}_{k \in M})$, where $x_0 \in \mathrm{GF}(q)$, M is an infinite subset of N, and the f_k are rational functions $\frac{g_k(x)}{h_k(x)}$ over $\mathrm{GF}(q)$ with the properties

(i) $h_k(v) \not\equiv 0$ for all $v \in \mathrm{GF}(q)$ and $k \in M$

(ii) $f_k \circ f_l = f_l \circ f_k$ for all $k, l \in M$ and

(iii) the complexity of calculating function values $f_k(v)$, $v \in \mathrm{GF}(q)$, is $O(ld(k))$.

An important realization of a MDHS with q odd can be obtained by using a class of rational functions which firstly has been investigated by L. Rédei [10]. We call these functions Rédei-functions. Further, by (a, b) we denote the greatest common divisor of the numbers a and b.

Let $\mathrm{GF}(q)$ be a Galois field of odd order q, let α be a nonsquare of $\mathrm{GF}(q)$, let k be a natural with $(k, q+1)=1$ and let the polynomials $g_k(x)$, $h_k(x)$ be defined by the functional equation

$$(x+\sqrt{\alpha})^k=g_k(x)+h_k(x)\sqrt{\alpha} \tag{10}$$

in $(\mathrm{GF}(q)[y]/(y^2-\alpha))[x]$, where we write $\sqrt{\alpha}$ instead of y. Then the Rédei-function $f_k(x)$ is defined as

$$f_k(x)=\frac{g_k(x)}{h_k(x)}. \tag{11}$$

In [10] it is proved that $h_k(x)$ has no zeros in $\mathrm{GF}(q)$, hence (i) holds. Further it is proved that for naturals k, l with $(k, q+1)=(l, q+1)=1$ we have

$$f_k \circ f_l = f_{kl}, \tag{12}$$

and therefore also (ii) holds. An obvious evaluation algorithm for calculating the function values $f_k(v)$, $v \in \mathrm{GF}(q)$, is given by applying the square-and-multiply-technique on the element $(v+\sqrt{\alpha})^k \in \mathrm{GF}(q)[y]/(y^2-\alpha) \cong \mathrm{GF}(q^2)$. This requires only $O(ld(k))$ operations in $\mathrm{GF}(q)$ and yields $g_k(v)$ and $h_k(v)$, as can be seen from equation (10). Having calculated $g_k(v)$ and $h_k(v)$, only one more field operation is needed to compute $f_k(v)$.

Thus we have proved

Theorem 3. Let $\mathscr{R}=(\mathrm{GF}(q), x_0, \{f_k\}_{k\in M})$, where q odd, $x_0 \in \mathrm{GF}(q)$, $M=\{l \in N \mid (l, q+1)=1\}$ and the $f_{..}$ are the Rédei-functions with a fixed nonsquare element $\alpha \in \mathrm{GF}(q)$. Then $\mathscr{R}$ is a MDHS.

The structure of the mappings induced by the Rédei-functions can be easily revealed (cf. [10]). Let us denote the multiplicative group of a field K by K^*. Since $\mathrm{GF}(q)$ is a subfield of $\mathrm{GF}(q^2)$, we can form the factor group $G=\mathrm{GF}(q^2)^*/\mathrm{GF}(q)^*$. As a factor group of a cyclic group, the group G is cyclic and the order of G is given by $q+1$. We define the following mapping:

$$\psi\colon \begin{cases} \mathrm{GF}(q) \;+\; G\backslash\{0\} \\ \quad v \;\to (v+\sqrt{\alpha})\,\mathrm{GF}(q)^*, \qquad v \in \mathrm{GF}(q). \end{cases}$$

Obviously, ψ is bijective. For k with $(k, q+1)=1$ we have $(\psi(v))^k=\psi(f_k(v))$ for all $v \in \mathrm{GF}(q)$, hence

$$f_k(v)=\psi^{-1}(\psi(v))^k \qquad \text{for all} \quad v \in \mathrm{GF}(q). \tag{13}$$

Now let us discuss how to choose x_0 in order that the complexity of the following problem P is large: Knowing x_0 and $f_k(x_0)$, calculate an $l \in M$ with $f_l(x_0)=f_k(x_0)$.

A special algorithm for solving this problem is given by successive testing of randomly chosen numbers l. Since for naturals k, $l \in M$ with $k \equiv l \bmod q+1$ we have $\psi(x_0)^k=\psi(x_0)^l$, hence $f_k(x_0)=f_l(x_0)$, we can restrict ourselves to numbers $l \in M$ with $0 \leqq l \leqq q$, or equivalently to elements $l \in Z_{q+1}$, the group of prime residue classes modulo $q+1$. Let L be a uniformly distributed random variable on Z_{q+1}, and let for every residue class $\bar{k} \in Z_{q+1}$ the probability $p_0(\bar{k})$ be defined by (6). Then a necessary condition on x_0 is given by

$$p_0(\bar{k})<\varepsilon \qquad \text{for all} \quad k \in M,$$

where ε is as in (7). Defining subsets $|H_k(x_0)|$ of Z_{q+1} by $|H_k(x_0)|= =\{\bar{l} \in Z_{q+1} \mid f_l(x_0)=f_k(x_0)\}$ we obtain $p_0(\bar{k})=\dfrac{|H_k(x_0)|}{\varphi(q+1)}$ for all $k \in M$, where φ denotes Euler's function. Therefore, another formulation of (6) is given by

$$\frac{|H_k(x_0)|}{\varphi(q+1)}<\varepsilon \qquad \text{for all} \quad k \in M.$$

A formula for $|H_k(x_0)|$ is contained in the following

Theorem 4. (i) $H_l(x_0)=H_k(x_0) \Leftrightarrow l \equiv k \bmod \operatorname{ord} \psi(x_0)$.

(ii) $|H_l(x_0)|=\dfrac{\varphi(q+1)}{\operatorname{ord} \psi(x_0)}$ for all $l \in Z_{q+1}$.

Proof. (i) We have the following equivalences: $H_k(x_0)=H_l(x_0) \Leftrightarrow f_k(x_0)= =f_l(x_0) \Leftrightarrow (\psi(x_0))^k=(\psi(x_0))^l \Leftrightarrow k \equiv l \bmod \operatorname{ord} \psi(x_0)$.

(ii) Let η be the canonical epimorphism from $(Z_{q+1}, .)$ onto $(Z_{\operatorname{ord}(x_0)}, .)$, and let $\bar{l}, \bar{k} \in Z_{q+1}$. Following (i) we have $\bar{l} \in H_k(x_0) \Leftrightarrow l \equiv k \bmod \operatorname{ord} \psi(x_0) \Leftrightarrow \eta(\bar{l})=\eta(\bar{k})$. Therefore we obtain

$$|H_k(x_0)|=|\operatorname{kern} \eta|=\frac{|Z_{q+1}|}{|Z_{\operatorname{ord} \psi(x_0)}|}=\frac{\varphi(q+1)}{\varphi(\operatorname{ord} \psi(x_0))},$$

and this completes the proof.

Thus x_0 has to be chosen in such a way that $\dfrac{1}{\varepsilon}<\varphi(\operatorname{ord} \psi(x_0))$. Additionally we obtain: The larger $\varphi(\operatorname{ord} \psi(x_0))$ is, the better is the choice of x_0. Now we note that φ is an order homomorphism from $(N, |)$ to $(N, \leqq)$, where φ denotes the division relation. Therefore the best choice is an x_0 with $\operatorname{ord} \psi(x_0)=q+1$, or in other terms an x_0 such that $\psi(x_0)$ generates G.

It is important to note that the complexity of the problem P does not only depends on the choice of x_0, but also on the choice of q. Obviously, instead of P we could also consider the following problem: Knowing $\psi(x_0)$ and $\psi(x_0)^k$, calculate an

$l \in M$ with $\psi(x_0)^l = \psi(x_0)^k$. Thus we have to choose q in such a way that computing discrete logarithms over the cyclic group G becomes computationally infeasible. Now we observe that the algorithm designed by Pohlig and Hellman [9] for Galois fields can be directly transferred to arbitrary cyclic groups. This algorithm is fast if and only if $|G|$ has only small prime factors. Thus we have to choose q in such a way that $q+1$ contains a large prime factor q'.

The practical computation of an x_0 with $\psi(x_0)$ generates G can be done by trial and error. There exist exactly $\varphi(q+1)$ generators of G, thus on the average

$$\rho = \frac{q+1}{\varphi(q+1)}$$

elements have to be checked until a generator is found. If $q+1 = q'a$, where q' is prime and a is small, we have

$$\rho = \frac{q'a}{(q'-1)\varphi(a)} \approx \frac{a}{\varphi(a)},$$

and this fraction is reasonably small.

References

1. *Diffie, W., Hellman, M. E.*, New directions in cryptography. IEEE Trans. Inform. Theory, **IT–22,** 644–654, 1976.
2. *Henze, E.*, The solution of the general equation for public key distribution systems. IEEE Trans. Inform. Theory, **IT–28,** 933, 1982.
3. *Horak, O.*, Über die Verwaltung kryptographischer Schlüssel in komplexen Kommunikationsnetzen. Elektrotechnik und Maschinenbau **100,** 402–409, 1983.
4. *Ingemarsson, I., Tang, D. T., Wong, C. K.*, A conference key distribution system. IEEE Trans. Inform. Theory, **IT–28,** 714–720, 1982.
5. *Lausch, H., Nöbauer, W.*, Algebra of Polynomials. North Holland Publishing Comp., Amsterdam, 1973.
6. *Lidl, R., Müller, W. B.*, Permutation polynomials in RSA-cryptosystems. Proc. Crypto 83, Univ. of California, Santa Barbara, 1983.
7. *Nöbauer, W.*, Über eine Klasse von Permutationspolynomen und die dadurch dargestellten Gruppen. J. reine angew. Math. **231,** 215–219, 1968.
8. *Odlyzko, A., M.*, Discrete logarithms in finite fields and their cryptographic significance. Preprint, AT&T Bell Lab., Murray Hill, New Jersey 07974, 1984.
9. *Pohlig, S. C., Hellman, M. E.*, An improved algorithm for computing logarithms over GF(p) and its cryptographic significance. IEEE Trans. Inform. Theory, **IT–24,** 106–110, 1978.
10. *Rédei, L.*, Über eindeutig umkehrbare Polynome in endlichen Körpern. Acta Sci. Math. Szeged **11,** 85–92, 1946.

Системы распределения шифров на основании многочленов и функции Редеи

Р. НЬОБАУЭР

(Клагенфурт)

Дается новое применение многочленов Диксона и класса рациональных функций в криптографии, которое впервые было изучено Л. Редеи. Показано, каким образом можно применять эти функции для конструирования систем распределения шифров. Определяются все системы распределения шифров определенного типа.

R. Nöbauer
Institut für Mathematik
UBW Klagenfurt
Universitätsstraße 67
A–9010 Klagenfurt
Austria

РУССКИЙ ПЕРЕВОД

Проблемы управления и теории информации, том **15**, *номер 1 (1986)*

ОБ ОДНОМ НЕОБХОДИМОМ УСЛОВИИ ОПТИМАЛЬНОСТИ ДЛЯ СТОХАСТИЧЕСКИХ СИСТЕМ С ШУМОМ ПРИ УПРАВЛЕНИИ

Л. Е. ШАЙХЕТ

(*Донецк*)

Рассматривается задача оптимального управления линейным стохастическим дифференциальным уравнением с шумом при управлении и квадратичным функционалом качества. Получено необходимое условие оптимальности управления для задач такого типа. Приведены примеры, демонстрирующие возможность построения синтеза оптимального управления с помощью полученного необходимого условия оптимальности.

Введение

В статье рассматривается задача оптимального управления линейным стохастическим дифференциальным уравнением с шумом при управлении и квадратичным функционалом качества. Случайные возмущения описываются независимыми между собой винеровским процессом и центрированной пуассоновской мерой. Допустимое управление определяется как случайный процесс, измеримый относительно заданного потока δ-алгебр и имеющий равномерно ограниченный второй момент.

Для допустимых управлений вводятся так называемые вариации Мак-Шейна и вычисляется в явном виде предел соответствующих вариаций функционала качества.

Особенностью предлагаемого метода является дополнительное интегрирование вариаций функционала качества по некоторому параметру. Без этого интегрирования упомянутый предел для систем с шумом при управлении не существует.

Если варируемое управление оптимально, то предел вариаций функционала качества является неотрицательным. Этот факт определяет некоторое необходимое условие оптимальности управления, с помощью которого в ряде случаев удается в явном виде построить синтез оптимального управления.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

1. Постановка задачи

Рассмотрим задачу оптимального управления линейным стохастическим дифференциальным уравнением [1–4]

$$\begin{gathered} d\xi(t)=(A_0(t)+B_0(t)u(t)+\mathrm{G}_0(t)\xi(t))dt+ \\ +\sum_{r=1}^{m}(A_r(t)+B_r(t)u(t)+G_r(t)\xi(t))dw_r(t)+ \\ +\int_0^T(A(z,t)+B(z,t)u(t)+G(z,t)\xi(t))\tilde{\nu}(dt,dz)\,, \qquad (1.1) \\ \xi(0)=\xi_0\,, \qquad t\in[0,T]\,, \end{gathered}$$

с квадратичным функционалом качества

$$\begin{gathered} I(u)=M[\xi^*(T)D_0\xi(T)+ \\ +\int_0^T(\xi^*(s)D_1(s)\xi(s)+u^*(s)D_2(s)u(s))ds] \qquad (1.2) \end{gathered}$$

и множеством допустимых управлений U.

Здесь m-мерный винеровский процесс $w(t)=(w_1(t),\ \dots,\ w_m(t))$ и центрированная пуассоновская мера $\tilde{\nu}(t,A)$ с параметром $t\Pi(A)$ независимы между собой и f_t-измеримы, $\{f_t\}$-поток σ-алгебр на вероятностном пространстве $\{\Omega,\ \sigma,\ P\}$, $M|\xi_0|^2<\infty$; $\xi(t)\in R^n$, $u(t)\in R^l$, коэффициенты уравнения и функционала качества имеют соответствующие размерности, неслучайны и равномерно ограничены, $D_i(i=0,\ 1,\ 2)$ — неотрицательно определённые матрицы.

Допустимым управлением называется произвольный f_t-измеримый l-мерный процесс $u(t)$, для которого $\|u\|^2=\sup\limits_{0\leqq t\leqq T}M|u(t)|^2<\infty$.

Пусть $u_0(t)$ — оптимальное управление задачи (1.1), (1.2), то есть $I(u_0)=\inf\limits_{u\in U}I(u)$, v — произвольная $f_{\tau-\varepsilon}$-измеримая случайная величина, $0<\varepsilon<\tau<T$, $M|v|^2<\infty$,

$$u_\varepsilon^\tau=u_\varepsilon^\tau(t)=\begin{cases} v, & t\in[\tau-\varepsilon,\tau), \\ u_0(t)\,, & t\in[0,T]\backslash[\tau-\varepsilon,\tau)\,, \end{cases} \qquad (1.3)$$

$$I'_\varepsilon(u_0)=\frac{1}{\varepsilon}\int_\varepsilon^T(I(u_\varepsilon^\tau)-I(u_0))\,d\tau\,,$$

$$I'_0(u_0)=\lim_{\varepsilon\to0}I'_\varepsilon(u_0)\,. \qquad (1.4)$$

Очевидно, что величина $I'_0(u_0)$ (если она существует) должна быть неотрицательной. Таким образом, неравенство $I'_0(u_0) \geqq 0$ является необходимым условием оптимальности управления u_0.

Цель работы — вычисление предела (1.4) для задачи управления (1.1), (1.2).

Теорема. Для любого допустимого управления u_0 величина $I'_0(u_0)$ существует и равна

$$I'_0(u_0) = M\Big[\int_0^T (v^* D_2(s) v - u_0^*(s) D_2(s) u_0(s)) ds +$$

$$+ 2\xi_0^*(T) D_0 p_0(T) + 2\int_0^T \xi_0^*(s) D_1(s) p_0(s)\, ds\Big] +$$

$$+ \mathrm{Sp}\,[D_0 P_0(T)] + \int_0^T \mathrm{Sp}\,[D_1(s) P_0(s)]\, ds\,.$$

Здесь $\xi_0(t)$ — решение уравнения (1.1) при управлении u_0, $p_0(t)$ — решение уравнения ($t \in [0, T]$)

$$p_0(t) = p_0^0(t) + \int_0^t G_0(s) p_0(s)\, ds +$$

$$+ \sum_{r=1}^m \int_0^t G_r(s) p_0(s)\, dw_r(s) + \int_0^t \int G(z, s) p_0(s) \tilde{\nu}(ds, dz)\,, \tag{1.5}$$

$$p_0^0(t) = \int_0^t B_0(s) V(s)\, ds + \sum_{r=1}^m \int_0^t B_r(s) V(s)\, dw_r(s) +$$

$$+ \int_0^t \int B(z, s) V(s) \tilde{\nu}(ds, dz), \qquad V(s) = v - u_0(s)\,, \tag{1.6}$$

$P_0(t)$ — решение уравнения ($t \in [0, T]$)

$$P_0(t) = P_0^0(t) + \int_0^t (G_0(s) P_0(s) + P_0(s) G_0^*(s))\, ds +$$

$$+ \sum_{r=1}^m \int_0^t G_r(s) P_0(s) G_r^*(s)\, ds +$$

$$+ \int_0^t \int G(z, s) P_0(s) G^*(z, s) \Pi(dz)\, ds\,, \tag{1.7}$$

$$P_0^0(t) = \sum_{r=1}^m \int_0^t B_r(s) W(s) B_r^*(s)\, ds +$$

$$+ \int_0^t \int B(z, s) W(s) B^*(z, s) \Pi(dz)\, ds\,, \tag{1.8}$$

$$W(s) = M V(s) V^*(s)\,.$$

2. Доказательство теоремы

Для доказательства теоремы понадобятся некоторые вспомогательные утверждения. Буквой C обозначаются произвольные положительные постоянные.

Лемма 1. Пусть $\xi_\varepsilon^\tau(t)$ — решение уравнения (1.1) при управлении (1.3), $q_\varepsilon^\tau(t)=\frac{1}{\varepsilon}(\xi_\varepsilon^\tau(t)-\xi_0(t))$. Тогда равномерно по $\tau\in[\varepsilon, T]$

$$\|q_\varepsilon^\tau\|^2 \leqq \frac{C}{\varepsilon}. \tag{2.1}$$

Доказательство. Очевидно, что $q_\varepsilon^\tau(t)=0$ при $t<\tau-\varepsilon$. Пусть $t\in[\tau-\varepsilon, \tau]$. Тогда из (1.1)

$$\begin{aligned} q_\varepsilon^\tau(t)=&\frac{1}{\varepsilon}\int\limits_{\tau-\varepsilon}^{t} B_0(s)V(s)\,ds+\sum_{r=1}^{m}\frac{1}{\varepsilon}\int\limits_{\tau-\varepsilon}^{t} B_r(s)V(s)\,dw_r(s)+\\ &+\frac{1}{\varepsilon}\int\limits_{\tau-\varepsilon}^{t}\int B(z,s)V(s)\tilde{v}(ds,dz)+\int\limits_{\tau-\varepsilon}^{t} G_0(s)q_\varepsilon^\tau(s)\,ds+\\ &+\sum_{r=1}^{m}\int\limits_{\tau-\varepsilon}^{t} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)+\int\limits_{\tau-\varepsilon}^{t}\int G(z,s)q_\varepsilon^\tau(s)\tilde{v}(ds,dz). \end{aligned} \tag{2.2}$$

Отсюда

$$M|q_\varepsilon^\tau(t)|^2\leqq C\left[\frac{1}{\varepsilon}+\int\limits_{\tau-\varepsilon}^{t} M|q_\varepsilon^\tau(s)|^2\,ds\right]$$

и на основании леммы Гронуолла–Беллмана

$$\sup_{\tau-\varepsilon\leqq t\leqq\tau} M|q_\varepsilon^\tau(t)|^2\leqq\frac{C}{\varepsilon}. \tag{2.3}$$

Аналогично при $t\in[\tau, T]$

$$\begin{aligned} q_\varepsilon^\tau(t)=&q_\varepsilon^\tau(\tau)+\int\limits_{\tau}^{t} G_0(s)q_\varepsilon^\tau(s)ds+\\ &+\sum_{r=1}^{m}\int\limits_{\tau}^{t} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)+\int\limits_{\tau}^{t}\int G(z,s)q_\varepsilon^\tau(s)\tilde{v}(ds,dz), \end{aligned} \tag{2.4}$$

$$M|q_\varepsilon^\tau(t)|^2\leqq C[M|q_\varepsilon^\tau(\tau)|^2+\int\limits_{\tau}^{t} M|q_\varepsilon^\tau(s)|^2\,ds].$$

Используя (2.3) и лемму Гронуолла–Беллмана, получим (2.1). Лемма 1 доказана.

Лемма 2. Пусть

$$p_\varepsilon^0(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \int\limits_\varepsilon^t q_\varepsilon^\tau(\tau)d\tau, & t\in[\varepsilon,T]\,,\end{cases}$$

$$p_\varepsilon(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \int\limits_\varepsilon^t q_\varepsilon^\tau(t)d\tau, & t\in[\varepsilon,T]\,.\end{cases}$$

Тогда равномерно по $t\in[0,T]$

$$\lim_{\varepsilon\to 0} M\,|p_\varepsilon^0(t)-p_0^0(t)|^2=0\,, \tag{2.5}$$

$$\lim_{\varepsilon\to 0} M\,|p_\varepsilon(t)-p_0(t)|^2=0\,. \tag{2.6}$$

Доказательство. Из (2.2) вытекает

$$p_\varepsilon^0(t)=\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau} B_0(s)V(s)\,ds\,d\tau+\frac{1}{\varepsilon}\sum_{r=1}^m\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau} B_r(s)V(s)\,dw_r(s)\,d\tau+$$

$$+\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau}\int B(z,s)V(s)\tilde{v}(ds,dz)\,d\tau+\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau} G_0(s)q_\varepsilon^\tau(s)\,ds\,d\tau+$$

$$+\sum_{r=1}^m\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)\,d\tau+\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau}\int G(z,s)q_\varepsilon^\tau(s)\tilde{v}(ds,dz)\,d\tau\,.$$

Пусть $z(s)=B_0(s)V(s)$. Так как $\|z\|<\infty$ и

$$\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau} z(s)\,ds\,d\tau=\frac{1}{\varepsilon}\int\limits_0^\varepsilon sz(s)\,ds+\int\limits_\varepsilon^{t-\varepsilon} z(s)\,ds+$$

$$+\frac{1}{\varepsilon}\int\limits_{t-\varepsilon}^t (t-s)z(s)\,ds=\int\limits_0^t z(s)\,ds+\frac{1}{\varepsilon}\int\limits_0^\varepsilon (s-\varepsilon)z(s)\,ds+$$

$$+\frac{1}{\varepsilon}\int\limits_{t-\varepsilon}^t (t-s-\varepsilon)z(s)\,ds\,, \tag{2.7}$$

то

$$M\left|\frac{1}{\varepsilon}\int\limits_\varepsilon^t\int\limits_{\tau-\varepsilon}^{\tau} B_0(s)V(s)\,ds\,d\tau-\int\limits_0^t B_0(s)V(s)\,ds\right|^2\leqq C\varepsilon^2\,.$$

Аналогично

$$M\left|\frac{1}{\varepsilon}\int\limits_{\varepsilon}^{t}\int\limits_{\tau-\varepsilon}^{\tau} B_r(s)V(s)\,dw_r(s)\,d\tau-\int\limits_0^t B_r(s)V(s)\,dw_r(s)\right|^2\leqq C\varepsilon\,,$$

$$M\left|\frac{1}{\varepsilon}\int\limits_{\varepsilon}^{t}\int\limits_{\tau-\varepsilon}^{\tau}\int B(z,s)V(s)\tilde{\nu}(ds,dz)\,d\tau-\int\limits_0^t\int B(z,s)V(s)\tilde{\nu}(ds,dz)\right|^2\leqq C\varepsilon\,.$$

Пусть $z_\varepsilon^\tau(s)=G_0(s)q_\varepsilon^\tau(s)$. Так как $\|z_\varepsilon^\tau\|^2\leqq\dfrac{C}{\varepsilon}$ и

$$\int\limits_{\varepsilon}^{t}\int\limits_{\tau-\varepsilon}^{\tau} z_\varepsilon^\tau(s)\,ds\,d\tau=\int\limits_{0}^{\varepsilon}\int\limits_{\varepsilon}^{s+\varepsilon} z_\varepsilon^\tau(s)\,d\tau\,ds+$$

$$+\int\limits_{\varepsilon}^{t-\varepsilon}\int\limits_{s}^{s+\varepsilon} z_\varepsilon^\tau(s)\,d\tau\,ds+\int\limits_{t-\varepsilon}^{t}\int\limits_{s}^{t} z_\varepsilon^\tau(s)\,d\tau\,ds\,, \tag{2.8}$$

то

$$M|\int\limits_{\varepsilon}^{t}\int\limits_{\tau-\varepsilon}^{\tau} G_0(s)q_\varepsilon^\tau(s)\,ds\,d\tau|^2\leqq C\varepsilon\,.$$

Аналогично

$$M|\int\limits_{\varepsilon}^{t}\int\limits_{\tau-\varepsilon}^{\tau} G_r(s)q_\varepsilon^\tau(s)\,dw_r(s)\,d\tau|^2\leqq C\varepsilon\,,$$

$$M|\int\limits_{\varepsilon}^{t}\int\limits_{\tau-\varepsilon}^{\tau}\int G(z,s)q_\varepsilon^\tau(s)\tilde{\nu}(ds,dz)\,d\tau|^2=C\varepsilon\,.$$

Следовательно,

$$M|p_\varepsilon^0(t)-p_0^0(t)|^2\leqq C\varepsilon\,. \tag{2.9}$$

Соотношение (2.5) доказано. Из (2.4) вытекает

$$p_\varepsilon(t)=p_\varepsilon^0(t)+\int\limits_{\varepsilon}^{t} G_0(s)p_\varepsilon(s)\,ds+$$

$$+\sum_{r=1}^{m}\int\limits_{\varepsilon}^{t} G_r(s)p_\varepsilon(s)\,dw_r(s)+\int\limits_{\varepsilon}^{t}\int G(z,s)p_\varepsilon(s)\tilde{\nu}(ds,dz)\,. \tag{2.10}$$

Вычитая (1.5) из (2.10) и используя (2.9) и лемму Гронуолла–Беллмана, получим (2.6). Лемма 2 доказана.

Лемма 3. Пусть $Q_\varepsilon^\tau(t)=\varepsilon Mq_\varepsilon^\tau(t)\,(q_\varepsilon^\tau(t))^*$,

$$P_\varepsilon^0(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \int\limits_{\varepsilon}^{t} Q_\varepsilon^\tau(\tau)\,d\tau\,, & t\in[\varepsilon,T]\,,\end{cases}$$

$$P_\varepsilon(t)=\begin{cases}0, & t\in[0,\varepsilon),\\ \int\limits_{\varepsilon}^{t} Q_\varepsilon^\tau(t)\,d\tau\,, & t\in[\varepsilon,T]\,.\end{cases}$$

Тогда равномерно по $t \in [0, T]$

$$\lim_{\varepsilon \to 0} |P_\varepsilon^0(t) - P_0^0(t)| = 0\,, \tag{2.11}$$

$$\lim_{\varepsilon \to 0} |P_\varepsilon(t) - P_0(t)| = 0\,. \tag{2.12}$$

Доказательство. Пусть $t \in [\tau - \varepsilon, \tau)$, $X_\varepsilon^\tau(t) = Mq_\varepsilon^\tau(t) V^*(t)$. Используя формулу Ито, из (2.2) легко получить

$$\begin{aligned}
P_\varepsilon^0(t) = {} & \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau [G_0(s) Q_\varepsilon^\tau(s) + Q_\varepsilon^\tau(s) G_0^*(s)]\, ds\, d\tau + \\
& + \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau [B_0(s)\,(X_\varepsilon^\tau(s))^* + X_\varepsilon^\tau(s) B_0^*(s)]\, ds\, d\tau + \\
& + \sum_{r=1}^m \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau [G_r(s) X_\varepsilon^\tau(s) B_r^*(s) + B_r(s)\,(X_\varepsilon^\tau(s))^* G_r^*(s)]\, ds\, d\tau + \\
& + \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau \int [G(z, s) X_\varepsilon^\tau(s) B^*(z, s) + B(z, s)\,(X_\varepsilon^\tau(s))^* G^*(z, s)]\, \Pi(dz)\, ds\, d\tau + \\
& + \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau \left[\sum_{r=1}^m G_r(s) Q_\varepsilon^\tau(s) G_r^*(s) + \int G(z, s) Q_\varepsilon^\tau(s) G^*(z, s)\, \Pi(dz)\right] ds\, d\tau + \\
& + \frac{1}{\varepsilon} \int_\varepsilon^t \int_{\tau-\varepsilon}^\tau \left[\sum_{r=1}^m B_r(s) W(s) B_r^*(s) + \int B(z, s) W(s) B^*(z, s)\, \Pi(dz)\right] ds\, d\tau\,.
\end{aligned} \tag{2.13}$$

Так как $|Q_\varepsilon^\tau(s)| \leqq C$, $|W(s)| \leqq C$, $|X_\varepsilon^\tau(s)| \leqq \dfrac{C}{\sqrt{\varepsilon}}$, то аналогично (2.7), (2.8) можно показать, что

$$|P_\varepsilon^0(t) - P_0^0(t)| \leqq C\sqrt{\varepsilon}. \tag{2.14}$$

Таким образом, (2.11) доказано. Аналогично (2.13) из (2.4) вытекает

$$\begin{aligned}
P_\varepsilon(t) = P_\varepsilon^0(t) + {} & \int_\varepsilon^t (G_0(s) P_\varepsilon(s) + P_\varepsilon(s) G_0^*(s))\, ds + \\
& + \sum_{r=1}^m \int_\varepsilon^t G_r(s) P_\varepsilon(s) G_r^*(s)\, ds + \\
& + \int_\varepsilon^t \int G(z, s) P_\varepsilon(s) G^*(z, s)\, \Pi(dz)\, ds\,.
\end{aligned} \tag{2.15}$$

Вычитая (1.7) из (2.15), а затем используя (2.14) и лемму Гронуолла–Беллмана, получим (2.12). Лемма 3 доказана.

Доказательство теоремы. Пусть $f(x, A, y) = x^*Ax - y^*Ay$. Тогда $I'_\varepsilon(u_0) = \sum_{i=1}^{4} \alpha_i(\varepsilon)$

$$\alpha_1(\varepsilon) = \frac{1}{\varepsilon}\int_\varepsilon^T Mf(\xi_\varepsilon^\tau(T), D_0, \xi_0(T))d\tau,$$

$$\alpha_2(\varepsilon) = \frac{1}{\varepsilon}\int_\varepsilon^T\int_\tau^T Mf(\xi_\varepsilon^\tau(s), D_1(s), \xi_0(s))\,ds\,d\tau\,,$$

$$\alpha_3(\varepsilon) = \frac{1}{\varepsilon}\int_\varepsilon^T\int_{\tau-\varepsilon}^{\tau} Mf(\xi_\varepsilon^\tau(s), D_1(s), \xi_0(s))\,ds\,d\tau\,,$$

$$\alpha_4(\varepsilon) = \frac{1}{\varepsilon}\int_\varepsilon^T\int_{\tau-\varepsilon}^{\tau} Mf(v, D_2(s), u_0(s))\,ds\,d\tau\,.$$

Так как

$$\alpha_1(\varepsilon) = M\int_\varepsilon^T (\xi_\varepsilon^\tau(T) + \xi_0(T))^* D_0 q_\varepsilon^\tau(T)\,d\tau =$$

$$= \mathrm{Sp}\,[D_0 P_\varepsilon(T)] + 2M\xi_0^*(T) D_0 p_\varepsilon(T)\,,$$

то (леммы 2,3)

$$\lim_{\varepsilon\to 0}\alpha_1(\varepsilon) = \mathrm{Sp}\,[D_0 P_0(T)] + 2M\xi_0^*(T) D_0 p_0(T)\,.$$

Аналогично этому и (2.7), (2.8) легко показать, что

$$\lim_{\varepsilon\to 0}\alpha_2(\varepsilon) = \int_0^T (\mathrm{Sp}\,[D_1(s)P_0(s)] + 2M\xi_0^*(s)D_1(s)p_0(s))\,ds\,,$$

$$\lim_{\varepsilon\to 0}\alpha_3(\varepsilon) = 0\,,$$

$$\lim_{\varepsilon\to 0}\alpha_4(\varepsilon) = \int_0^T (v^*D_2(s)v - u_0^*(s)D_2(s)u_0(s))\,ds\,.$$

Теорема доказана.

3. Примеры

В некоторых случаях необходимое условие оптимальности позволяет получить явный вид оптимального управления. В качестве иллюстрации рассмотрим два примера.

Пример 1. Для задачи управления

$$\dot{\xi}(t)=\alpha+(\beta+\gamma u(t))\dot{w}(t)), \tag{3.1}$$

$$I(u)=M[\mu\xi^2(T)+\lambda\int_0^T u^2(s)\,ds] \tag{3.2}$$

величина $I_0'(u_0)$ имеет вид

$$I_0'(u_0)=M[2\mu\xi_0(T)p_0(T)+\mu\gamma^2\int_0^T(v-u_0(t))^2\,dt+$$

$$+\lambda\int_0^T(v^2-u_0^2(t))\,dt, \qquad \dot{p}_0(t)=\gamma(v-u_0(t))\dot{w}(t)\,.$$

Так как

$$M\xi_0(T)p_0(T)=\gamma\int_0^T M(\beta+\gamma u_0(t))(v-u_0(t))\,dt\,,$$

то

$$I_0'(u_0)=(\lambda+\mu\gamma^2)M\int_0^T[(v-u_0(t))^2+$$

$$+2(v-u_0(t))(u_0(t)+\mu\gamma\beta/(\lambda+\mu\gamma^2))]\,dt\,.$$

Для неотрицательности $I_0'(u_0)$ необходимо и достаточно, чтобы оптимальное управление задачи (3.1), (3.2) имело вид

$$u_0(t)=-\mu\gamma\beta/(\lambda+\mu\gamma^2)\,.$$

Этот же результат можно получить с помощью уравнения Беллмана.

Пример 2. Для задачи управления уравнением

$$\dot{\xi}(t)=(\alpha+\beta\dot{w}(t))u(t) \tag{3.3}$$

с функционалом качества (3.2) величина $I_0'(u_0)$ имеет вид

$$I_0'(u_0)=M[2\mu\xi_0(T)p_0(T)+$$

$$+(\lambda+\mu\beta^2)\int_0^T(v-u_0(t))^2\,dt+2\lambda\int_0^T(v-u_0(t))u_0(t)\,dt]\,,$$

$$\dot{p}_0(t)=(\alpha+\beta\dot{w}(t))(v-u_0(t))\,.$$

Из формулы Ито

$$M\xi_0(t)p_0(t)=\alpha\int_0^t Mu_0(s)p_0(s)\,ds+$$

$$+\int_0^t M(\alpha\xi_0(s)+\beta^2u_0(s))(v-u_0(s))\,ds\,.$$

Предположим, что u_0 имеет вид

$$u_0(t) = -q(t)\xi_0(t)\,, \tag{3.4}$$

где $q(t)$ — неслучайная функция. Тогда

$$M\xi_0(t)p_0(t) = \int_0^t (\alpha - \beta^2 q(s))M\xi_0(s)(v - u_0(s))\exp\left[-\alpha\int_s^t q(\tau)\,d\tau\right]ds\,,$$

$$I'_0(u_0) = M\{(\lambda + \mu\beta^2)\int_0^T (v - u_0(t))^2\,dt +$$

$$+ 2\int_0^T \xi_0(t)(v - u_0(t))\left[\mu(\alpha - \beta^2 q(t))\exp\left[-\alpha\int_t^T q(s)\,ds\right] - \lambda q(t)\right]dt\}\,.$$

Для неотрицательности $I'_0(u_0)$ необходимо и достаточно, чтобы

$$\alpha = q(t)\left[\beta^2 + \frac{\lambda}{\mu}\exp\left(\alpha\int_t^T q(s)\,ds\right)\right]. \tag{3.5}$$

Таким образом, оптимальное управление задачи (3.3), (3.2) имеет вид (3.4), (3.5).

Покажем, что это решение совпадает с решением, полученным с помощью уравнения Беллмана. Как известно [1], уравнение Беллмана даёт оптимальное управление в виде

$$u_0(t) = -\alpha p(t)\xi_0(t)/(\lambda + \beta^2 p(t))\,,$$

где $p(t)$ — решение уравнения

$$\dot p(t) = \frac{\alpha^2 p^2(t)}{\lambda + \beta^2 p(t)}\,, \qquad p(T) = \mu\,. \tag{3.6}$$

Следовательно, должно выполняться соотношение

$$q(t) = \alpha p(t)/(\lambda + \beta^2 p(t))\,,$$

или $\dot p(t) = \alpha p(t)q(t)$. Действительно, подставляя $q(t) = \dot p(t)/\alpha p(t)$ в (3.5), получим (3.6).

4. Заключение

Следует отметить, что вместо предела (1.4) можно, вообще говоря, вычислять предел

$$I^{\tau}(u_0) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[I(u_\varepsilon^{\tau}) - I(u_0)\right]. \tag{4.1}$$

при произвольном фиксированном $\tau \in (0, T)$. Неравенство $I^{\tau}(u_0) \geqq 0$ также является необходимым условием оптимальности управления u_0. Так было сделано в [5–9] для различных (дифференциальных, интегральных, с частными производными) нелинейных стохастических уравнений без шума при управлении. Но для систем с шумом при управлении предел (4.1), в отличие от (1.4), не существует.

Кроме того, при вычислении предела (1.4) в [5–9] на некоторые параметры задачи управления (в том числе и на оптимальное управление) накладывались довольно жёсткие условия гёльдеровости по времени. Переход от предела (4.1) к его интегральной форме (1.4) позволяет отказаться от этих условий.

Вместе с тем, в отличие от [5–9] здесь рассматривается линейная система. При наличии шума в управлении нелинейность системы существенно усложняет вычисление предела (1.4).

Литература

1. *Гихман, И. И., Скороход А. В.* Управляемые случайные процессы. Киев, Наук. думка, 1977, 252 с.
2. *Хасьминский Р. З.* Устойчивость систем дифференциальных уравнений при случайных возмущениях их параметров. М., Наука, 1967. 368 с.
3. *Черноусько Ф. Л., Колмановский В. Б.* Оптимальное управление при случайных возмущениях. М., Наука, 1978, 352 с.
4. *Колмановский В. Б., Носов В. Р.* Устойчивость и периодические режимы регулируемых систем с последействием. М., Наука, 1981, 448 с.
5. *Шайхет Л. Е.* Необходимые условия оптимальности управления для некоторых стохастических систем. В кн.: Республиканская конференция по теории стохаст. дифференц. уравнений. (Донецк, 15–17 сент. 1982 г.). Тез. докл., Донецк: ИПММ АН УССР, 1982, с. 104–105.
6. *Шайхет Л. Е.* Об одном необходимом условии оптимальности управления для стохастических дифференциальных уравнений гиперболического типа. В кн.: Теория случайных процессов, 1984, вып. 12, с. 96–101.
7. *Warfield, V. M.*, A stochastic maximum principle. SIAM J. Control and Optimization. August, 1976, **14,** *5*, pp. 803–826.
8. *Shaikhet, L. E.*, On optimal control of Volterra's equations. Problems of Control and Information Theory, Budapest, 1984, **13,** *3*, pp. 141–152.
9. *Shaikhet, L. E.*, Optimal control of stochastic integral-functional equations. Stochastic Optimization. (Intern. Conf. Kiev, USSR, 9–16 Sept. 1984): Abstracts of Papers, Kiev: V. M. Glushkov Institute of Cybernetics Ac. Sci. Ukr. SSR, pp. 212–213.

PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda, Budapest

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

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). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A=||a_{ij}||$. Write diagonal matrices as diag $(d_1, d_2, \ldots d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 129090 Москва И-90, ул. Щепкина, 8. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 208-95-01).

Объём 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 языке в трех экземплярах, 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 и 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 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

HU ISSN 0370-2529

Index: 26 660

VOL. / ТОМ 15 · NUMBER / НОМЕР 2

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 СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1986

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
— theory of adaptive systems
— theory of estimation and identification
— theory of controlling robot-technologic and flexible manufacturing systems
— information theory
— information-theoretic aspects of multiple access networks.

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;
— теория адаптивных систем;
— теория оценивания и идентификации;
— теория управления робототехническими и гибкими производственными системами;
— теория информации;
— теория информации в области сетей с множественным доступом.

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 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1986 Subscription Rate US Dollars 175.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

1828

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

Problems of Control and Information Theory, Vol. 15 (2), pp. 101—109 (1986)

APPROXIMATE METHODS FOR FINDING FINITE-DIMENSIONAL DISTRIBUTIONS OF RANDOM SEQUENCES DETERMINED BY DIFFERENCE EQUATIONS

V. S. PUGACHEV

(*Moscow*)

(Received May 20, 1985)

Approximate methods for finding finite-dimensional distributions of random sequences determined by difference equations are briefly outlined. These methods, proposed here for the first time, are similar to those which are applied to random processes determined by stochastic differential equations [1].

1. General formulae determining finite-dimensional distributions

Let us consider a random sequence $\{X_t\}$ with values in any measurable space $(\mathscr{X}, \mathscr{A})$ generated by the difference equation

$$X_{t+1} = \varphi_t(X_t, V_t), \tag{1}$$

where $\{V_t\}$ is a sequence of independent random variables with values in another measurable space $(\mathscr{V}, \mathscr{E})$, $\varphi_t(x, v)$ $(\mathscr{A} \times \mathscr{E}, \mathscr{A})$-measurable functions mapping $\mathscr{X} \times \mathscr{V}$ into $\mathscr{X}$. Let the initial value X_1 of the sequence $\{X_t\}$ be a random variable independent of the sequence $\{V_t\}$, $\mu_1(A)$, $A \in \mathscr{A}$, the distribution of X_1, $v_t(E)$, $E \in \mathscr{E}$, the distribution of the random variable V_t $(t = 1, 2, \dots)$.

As is well known, $\{X_t\}$ is a Markov sequence under such conditions. So all its finite-dimensional distributions are completely determined by its initial distribution $\mu_1(A)$, $A \in \mathscr{A}$ and the transition distribution $\mu_{t+1}(A \mid X_t)$, $A \in \mathscr{A}$, given by

$$\mu_{t+1}(A \mid x) = v_t(\varphi_t^{-1}(x, A)) \tag{2}$$

where $\varphi_t^{-1}(x, A) \subset \mathscr{V}$ is the inverse image of the set $A \subset \mathscr{X}$ given by the function $\varphi_t(x, v)$ at any fixed x. Formula (2) at $t = 1, 2, \dots$ with given $\mu_1(A)$ solves the problem completely and exactly. This general formula is valid under quite general conditions, for random variables X_t, V_t in any spaces $\mathscr{X}$, $\mathscr{V}$ respectively. But it is of no value for computations. So one must look for approximate methods providing computational solution of the problem.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

First of all we restrict ourselves to finite-dimensional random variables X_t, V_t, finite-dimensional spaces $\mathscr{X}$, $\mathscr{V}$ with Borel σ-algebras $\mathscr{A}$, $\mathscr{E}$. Secondly we shall characterize the distributions of random variables by respective characteristic functions.

2. Equations determining the finite-dimensional characteristic functions

Let $\mathscr{X}$ and $\mathscr{V}$ be finite-dimensional spaces R^p and R^q respectively,

$$g_t(\lambda) = M \exp\{i\lambda^T X_t\} \qquad (t = 1, 2, \dots) \tag{3}$$

the one-dimensional characteristic function of $\{X_t\}$, i.e. the characteristic function of the random variable X_t. From (1) and (3) immediately follows the difference equation for g_t:

$$g_{t+1}(\lambda) = M \exp\{i\lambda^T \varphi_t(X_t, V_t)\}. \tag{4}$$

This is a difference equation for g_t because the right member is completely determined by the distribution of V_t and the distribution of X_t. The latter is in its turn completely determined by $g_t(\lambda)$. Equation (4) with given $g_1(\lambda)$ completely determines $g_t(\lambda)$ for all $t > 1$.

Now let

$$g_{t_1,\dots,t_n}(\lambda_1 \;\dots, \lambda_n) = M \exp\left\{i \sum_{k=1}^{n} \lambda_k^T X_{t_k}\right\} \tag{5}$$

be the n-dimensional characteristic function of $\{X_t\}$, i.e. the joint characteristic function of the random variables $X_{t_1}, \dots, X_{t_n}$. Exactly in the same way we obtain from (1) and (5) the following difference equation for $g_{t_1,\dots,t_n}(\lambda_1, \dots, \lambda_n)$ at $t_1 < t_2 < \dots < t_{n-1} < t_n$

$$g_{t_1,\dots,t_{n-1},t_n+1}(\lambda_1, \dots, \lambda_n) =$$

$$= M \exp\left\{i \sum_{k=1}^{n-1} \lambda_k^T X_{t_k} + i\lambda_n^T \varphi_{t_n}(X_{t_n}, V_{t_n})\right\}' \tag{6}$$

with the evident initial condition

$$g_{t_1,\dots,t_{n-1},t_{n-1}}(\lambda_1, \dots, \lambda_n) = g_{t_1,\dots,t_{n-1}}(\lambda_1, \dots, \lambda_{n-1} + \lambda_n). \tag{7}$$

The right member of (6) is completely determined by $g_{t_1,\dots,t_n}(\lambda_1, \dots, \lambda_n)$, the distribution of the random variable V_{t_n} being given. So (6) represents a difference equation for $g_{t_1,\dots,t_n}(\lambda_1, \dots, \lambda_n)$ at fixed $t_1, \dots, t_{n-1}$. After finding $g_{t_1,\dots,t_n}$ for $t_1 < t_2 < \dots < t_n$ it is easily found for any set $\{t_1, \dots, t_n\}$ by symmetry:

$$g_{t_1,\dots,t_n}(\lambda_1, \dots, \lambda_n) = g_{t_{s_1},\dots,t_{s_n}}(\lambda_{s_1}, \dots, \lambda_{s_n}) \tag{8}$$

where $(s_1, \dots, s_n)$ is such a permutation of the numbers $(1, \dots, n)$ for which $t_{s_1} < t_{s_2} < \dots < t_{s_n}$ [1].

3. Case of a linear difference equation

In the case of a linear function φ_t in (1),

$$\varphi_t(x, v) = a_t x + b_t v + c_t \tag{9}$$

equation (4) becomes in virtue of independence of X_t and V_t

$$g_{t+1}(\lambda) = \exp\{i\lambda^T c_t\} h_t(b_t^T \lambda) g_t(a_t^T \lambda) \tag{10}$$

where $h_t(\mu)$ is the characteristic function of the random variable V_t. This equation is easily solved yielding

$$g_{t+1}(\lambda) = g_1(a_1^T \dots a_t^T \lambda) \exp\{i\lambda^T c_t\} h_t(b_t^T \lambda) \times$$
$$\times \prod_{s=1}^{t-1} \exp\{i\lambda^T a_t \dots a_{s+1} c_s\} h_s(b_s^T a_{s+1}^T \dots a_t^T \lambda). \tag{11}$$

Similarly substituting (9) into (6) we get the equation for the n-dimensional characteristic function

$$g_{t_1,\dots,t_{n-1} t_n+1}(\lambda_1, \dots, \lambda_n) =$$
$$= \exp\{i\lambda_n^T c_{t_n}\} h_{t_n}(b_{t_n}^T \lambda_n) g_{t_1,\dots,t_{n-1},t_n}(\lambda_1, \dots, \lambda_{n-1}, a_{t_n}^T \lambda_n). \tag{12}$$

Solving this equation with the initial condition (7) in the same way as before we obtain

$$g_{t_1,\dots,t_{n-1},t_n+1}(\lambda_1, \dots, \lambda_n) =$$
$$= g_{t_1,\dots,t_{n-1}}(\lambda_1, \dots, \lambda_{n-1} + a_{t_{n-1}}^T \dots a_{t_n}^T \lambda_n) \exp\{i\lambda_n^T c_{t_n}\} \times$$
$$\times h_{t_n}(b_{t_n}^T \lambda_n) \prod_{s=t_{n-1}}^{t_n-1} \exp\{i\lambda_n^T a_{t_n} \dots a_{s+1} c_s\} h_s(b_s^T a_{s+1}^T \dots a_{t_n}^T \lambda_n). \tag{13}$$

Formula (11) and the recursive formula (13) give explicit expressions for all the finite-dimensional characteristic functions of the random sequence $\{X_t\}$ in the case of a linear equation (1).

4. Approximate methods for the case of non-linear difference equations

The general approach to solving equations (4), (6)–(7) approximately consists in the parametrization of the unknown finite-dimensional distributions and derivation of the corresponding equations for the parameters of distributions. Namely, we replace unknown finite-dimensional characteristic functions $g_{t_1,\dots,t_n}(\lambda_1, \dots, \lambda_n)$ (and the respective densities, if they exist) by some known functions $g_n^*(\lambda_1, \dots, \lambda_n; \theta^{(n)})$ depending on a finite-dimensional vector parameter $\theta^{(n)}$ in such a way that $\theta^{(n)}$ includes all the components of the parameter $\theta^{(n-1)}$ of the function $g_{n-1}^*(\lambda_1, \dots, \lambda_{n-1}; \theta^{(n-1)})$ and then derive from (4) and (6) the difference equations for all the components of the vectors $\theta^{(1)}, \theta^{(2)}, \dots$ successively. The more components the vectors $\theta^{(n)}$ contain, the higher is the accuracy of approximating $g_{t_1,\dots,t_n}(\lambda_1, \dots, \lambda_n)$ by $g_n^*(\lambda_1, \dots, \lambda_n; \theta^{(n)})$ (in principle).

The method of normal approximation. The simplest way to parametrize a distribution is to approximate it by a normal one reducing the problem to finding the expectations $m_t = MX_t$ covariance and cross-covariance matrices $K_t = M(X_t - m_t)X_t^T$, $K_{t_1,t_2} = M(X_{t_1} - m_{t_1})X_{t_2}^T$ of the random variables X_t [1]. The equations for these parameters follow immediately from (1):

$$m_{t+1} = M\varphi_t(X_t, V_t), \tag{14}$$

$$K_{t+1} = M\varphi_t(X_t, V_t)\varphi_t(X_t, V_t)^T - M\varphi_t(X_t, V_t)M\varphi_t(X_t, V_t)^T, \tag{15}$$

$$K_{t_1,t_2+1} = MX_{t_1}\varphi_{t_2}(X_{t_2}, V_{t_2})^T - m_{t_1}M\varphi_{t_2}(X_{t_2}, V_{t_2})^T. \tag{16}$$

Evaluating the right members of (14) and (15) for the normal distribution of X_t instead of the true one we obtain them as known functions of m_t and K_t. So (14) and (15) represent a set of approximate difference equations for m_t and K_t. These equations with given m_1 and K_1 determine m_t and K_t for all $t > 1$.

Similarly replacing the joint distribution of X_{t_1}, X_{t_2} by a normal one while calculating the expectations in (16) we get the difference equation for K_{t_1,t_2} at $t_2 > t_1$. This equation with m_t, K_t found before and $K_{t_1,t_1} = K_{t_1}$ determines K_{t_1,t_2} at all $t_2 > t_1$. To find K_{t_1,t_2} at $t_2 < t_1$ it is sufficient to use the symmetry $K_{t_1,t_2} = K_{t_2,t_1}^T$.

Equations (14)–(16) may also be obtained from (4) and (6) by differentiating twice with respect to $i\lambda$ and then putting $\lambda = 0$.

The method of moments. If the components of the parameters $\theta^{(n)}$ of the approximating functions $g_n^*(\lambda_1, \dots, \lambda_n; \theta^{(n)})$ $(n = 1, 2, \dots)$ represent the moments of $X_{t_1}, \dots, X_{t_n}$ up to a given order N,

$$\alpha_{t_1,\dots,t_n}^{r_1,\dots,r_n} = MX_{t_1 1}^{r_{11}} \dots X_{t_1 p}^{r_{1p}} \dots X_{t_n 1}^{r_{n1}} \dots X_{t_n p}^{r_{np}}$$

or

$$\mu_{t_1,\ldots,t_n}^{r_1,\ldots,r_n} = M(X_{t_1 1} - m_{t_1 1})^{r_{11}} \ldots (X_{t_1 p} - m_{1_1 p})^{r_{1p}} \ldots$$
$$\ldots (X_{t_n 1} - m_{t_n 1})^{r_{n1}} \ldots (X_{t_n p} - m_{t_n p})^{r_{np}},$$

$r_1, \ldots, r_n$ being vector indices $r_k = [r_{k1} \ \ldots \ r_{kp}]^T$ $(k=1, \ldots, n)$, $\sigma(r_k) = r_{k1} + \ \ldots \ + r_{kp}$, then these parameters are determined by

$$\alpha_{t+1}^{r_1} = M\varphi_{t1}^{r_{11}}(X_t, V_t) \ldots \varphi_{tp}^{r_{1p}}(X_t, V_t)$$
$$(r_{11}, \ldots, r_{1p} = 0, 1, \ldots, N; \quad \sigma(r_1) = r_{11} + \ \ldots \ + r_{1p} = 1, \ldots, N) \tag{17}$$

$$\alpha_{t_1,\ldots,t_{n-1},t_n+1}^{r_1,\ldots,r_n} = MX_{t_1 1}^{r_{11}} \ldots X_{t_{n-1},p}^{r_{n-1,p}} \varphi_{t_n 1}^{r_{n1}}(X_{t_n}, V_{t_n}) \ldots$$
$$\ldots \varphi_{t_n p}^{r_{np}}(X_{t_n}, V_{t_n})$$
$$(t_1 < \ \ldots \ < t_{n-1} \leqq t_n; \quad r_{11}, \ldots, r_{np} = 0, 1, \ldots, N-n+1;$$
$$\sigma(r_1), \ldots, \sigma(r_n) = 1, \ldots, N-n+1;$$
$$\sigma(r_1) + \ \ldots \ + \sigma(r_n) = n, \ldots, N; \quad n = 2, \ldots, N) \tag{18}$$

or respectively

$$\mu_{t+1}^{r_1} = M[\varphi_{t1}(X_t, V_t) - M\varphi_{t1}(X_,, V_t)]^{r_{11}} \times \ldots \times$$
$$\times [\varphi_{tp}(X_t, V_t) - M\varphi_{tp}(X_t, V_t)]^{r_{1p}}$$
$$(r_{11}, \ldots, r_{1p} = 0, 1, \ldots, N; \quad \sigma(r_1) = r_{11} + \ \ldots \ + r_{1p} = 1, \ldots, N) \tag{19}$$

$$\mu_{t_1,\ldots,t_{n-1},t_n+1}^{r_1,\ldots,r_n} = M(X_{t_1 1} - m_{t_1 1})^{r_{11}} \ldots (X_{t_{n-1} p} - m_{t_{n-1} p})^{r_{n-1,p}} \times$$
$$\times [\varphi_{t_n 1}(X_{t_n}, V_{t_n}) - M\varphi_{t_n 1}(X_{t_n}, V_{t_n})]^{r_{n1}} \times \ldots \times$$
$$\times [\varphi_{t_n p}(X_{t_n}, V_{t_n}) - M\varphi_{t_n p}(X_{t_n}, V_{t_n})]^{r_{np}}$$
$$(t_1 < \ldots < t_{n-1} \leqq t_n; \quad r_{11}, \ldots, r_{np} = 0, 1, \ldots, N-n+1;$$
$$\sigma(r_1), \ldots, \sigma(r_n) = 1, \ldots, N-n+1;$$
$$\sigma(r_1) + \ldots + \sigma(r_n) = n, \ldots, N; \quad n = 2, \ldots, N). \tag{20}$$

Replacing unknown true distributions in the expressions of the expectations in (17)–(20) by their respective approximating functions we get closed sets of approximate difference equations successively determining the moments of the finite-dimensional distributions of the sequence $\{X_t\}$. Equations (17) or (19) with a given α_1^r (respectively μ_1^r) determine α_t^r or μ_t^r for all $t > 1$. Equations (18) or (20) with the initial condition

$$\alpha_{t_1, \ldots, t_{n-1}, t_{n-1}}^{r_1 \ldots, r_n} = \alpha_{t_1, \ldots, t_{n-1}}^{r_1, \ldots, r_{n-1}+r_n} \qquad (n=2, \ldots, N) \tag{21}$$

and

$$\mu_{t_1, \ldots, t_{n-1}, t_{n-1}}^{r_1, \ldots, r_n} = \mu_{t_1, \ldots, t_{n-1}}^{r_1, \ldots, r_{n-1}+r_n} \qquad (n=2, \ldots, N) \tag{22}$$

determine $\alpha_{t_1, \ldots, t_n}^{r_1, \ldots, r_n}$ or $\mu_{t_1, \ldots, t_n}^{r_1, \ldots, r_n}$, respectively for all $t_n > t_{n-1}$.

Equations (17)–(20) follow immediately from (1) or may be derived from (4) and (6) by differentiations with respect to $i\lambda, i\lambda_1, \ldots, i\lambda_n$ and putting $\lambda = \lambda_1 = \ldots = \lambda_n = 0$.

As functions approximating the distributions, truncated orthogonal expansions may be used for densities, in particular Hermite polynomial expansions or truncated Edgworth series.

The methods based on orthogonal expansions. The coefficients of truncated orthogonal expansions of the densities,

$$f_n^*(x_1, \ldots, x_n;\ \theta^{(n)}) = w(x_1, \ldots, x_n, m, K) \times$$

$$\times \left\{ 1 + \sum_{k=3}^{N} \sum_{\sigma(\nu_1) + \ldots + \sigma(\nu_n) = k} c_{\nu_1, \ldots, \nu_n}^{t_1, \ldots, t_n} p_{\nu_1 \ldots, \nu_n}(x_1, \ldots, x_n) \right\}$$

where

$$m = [m_{t_1}^T \ldots m_{t_n}^T]^T,$$

$$K = \begin{bmatrix} K_{t_1} & K_{t_1, t_2} & \ldots & K_{t_1, t_n} \\ K_{t_1, t_2}^T & K_{t_2} & \ldots & K_{t_2, t_n} \\ \ldots & \ldots & \ldots & \ldots \\ K_{t_1, t_n}^T & K_{t_2, t_n}^T & \ldots & K_{t_n} \end{bmatrix} \tag{24}$$

$$c_{\nu_1, \ldots, \nu_n}^{t_1, \ldots, t_n} = \left[q_{\nu_1, \ldots \nu_n}\left(\frac{\partial}{i\partial\lambda_1}, \ldots, \frac{\partial}{i\partial\lambda_n} \right) g_{t_1, \ldots, t_n}(\lambda_1, \ldots \lambda_n) \right]_{\lambda_1 = \ldots = \lambda_n = 0}, \tag{25}$$

$\nu_k = [\nu_{k1} \ldots \nu_{kp}]^T$ $(k = 1, \ldots, n)$, $\sigma(\nu_k) = \nu_{k1} + \ldots + \nu_{kp}$, $\{p_{\nu_1, \ldots, \nu_n}, q_{\nu_1, \ldots, \nu_n}\}$ is a set of biorthonormal polynomials,

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} w(x_1, \ldots, x_n; m, K) p_{\nu_1, \ldots \nu_n}(x_1, \ldots, x_n) \times$$

$$\times q_{\mu_1, \ldots, \mu_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n = \delta_{\nu_1\mu_1} \ldots \delta_{\nu_n\mu_n} \tag{26}$$

may be taken as parameters $\theta^{(n)}$ of the approximating functions.

From (4) and (25) with $n=1$ follows the relation

$$c_\nu^{t+1} = \left[q_\nu\left(\frac{\partial}{i\partial\lambda}\right) M \exp\{i\lambda^T \varphi_t(X_t, V_t)\}\right]_{\lambda=0}$$

or

$$c_\nu^{t+1} = M q_\nu(\varphi_t(X_t, V_t))$$

$$(\nu_1, \ldots, \nu_p = 0, 1, \ldots, N; \quad \sigma(\nu) = 3, \ldots, N). \tag{27}$$

Similarly we get from (6) and (25)

$$c_{\nu_1,\ldots,\nu_n}^{t_1,\ldots,t_{n-1},t_n+1} = M q_{\nu_1,\ldots,\nu_n}(X_{t_1}, \ldots, X_{t_{n-1}}, \varphi_{t_n}(X_{t_n}, V_{t_n}))$$

$$(\nu_{1k}, \ldots, \nu_{nk} = 0, 1, \ldots, N; \quad k = 1, \ldots, p; \quad \sigma(\nu_1), \ldots, \sigma(\nu_n) =$$

$$= 1, \ldots, N-n+1; \quad \sigma(\nu_1) + \ldots + \sigma(\nu_n) = n, \ldots, N). \tag{28}$$

After substituting into (14), (15) and (27), expression (23) for $f_1^*(x, \theta^{(1)})$ instead of the unknown true density $f_1(x; t)$ we get a closed set of equations determining m_t, K_t and all the coefficients c_ν in formula (23) at $n=1$.

After substituting into (16) and (28) at $n=2$, expression (23) for $f_2^*(x_1, x_2; \theta^{(2)})$ instead of the unknown true density $f_2(x_1, x_2; t_1, t_2)$ we get a closed set of equations for K_{t_1,t_2} and all the coefficients c_{ν_1,ν_2} in formula (23) at $n=2$. Equations (28) at any $n>2$ represent a closed set of equations for all the coefficients $c_{\nu_1,\ldots,\nu_n}$ in formula (23) at $n>2$. So equations (14), (15), (27); (16), (28) at $n=2$; and (28) at $n>2$ determine successively approximate expressions (23) for all the finite-dimensional distributions of the sequence $\{X_t\}$.

Other methods widely used for stochastic systems governed by differential equations, namely methods based on approximate equations for the semi-invariants of the random processes ([1], § 6) are hardly applicable to difference equations owing to the difficulties connected with the derivation of equations for the semi-invariants in this case.

To reduce the number of equations determining parameters of the finite-dimensional distributions in high-dimensional problems, the approach initiated by Malchikov [2] and developed in [3] may be recommended.

This approach is based upon taking into account only such dependences among the components of a vector random process and their values at different time instants which are characterized by the second-order semi-invariants and neglecting the mixed semi-invariants of higher orders. This leads to the expressions for all mixed moments of orders higher than 2 in terms of moments of the components taken separately and the second-order mixed moments. As a result only those of equations (17), (19) and (27), remain for which only one of the components of the vector index r differs from zero, and equations (18), (20) and (28) become superfluous. To avoid the evaluation of higher-order mixed moments in terms of the moments of the components taken separately and the second-order mixed moments the Mal'chikov's approximation of multidimensional densities

$$f(x_1, \ldots, x_r) \approx \prod_{k=1}^{r} f_k(x_k) \prod_{p=2}^{r} \prod_{q=1}^{p-1} \left\{1 + \frac{k_{pq}}{k_{pp}k_{qq}}(x_p - m_p)(x_q - m_q)\right\}$$

with approximations (23) for each $f_k(x_k)$ may be used, where $m_p = MX_p$, $k_{pq} = M(X_p - m_p)(X_q - m_q)$ $(p, q = 1, \ldots, r)$ [2].

5. Conclusions

The methods proposed may naturally be applied to statistical analysis of nonlinear sampled data systems described by difference equations. They may also be applied to problems of designing conditionally optimal filters and extrapolators for discrete-time random processes determined by difference equations [4–6].

One of the most important problems of further research in the field is the problem of automatic deriving equations (17)–(20), (27), (28) for specific practical problems by computers. The programs for automatic derivation of these equations will considerably promote the practical use of the methods proposed for various problems of the statistical analysis and synthesis of sampled data systems.

References

1. *Pugachev, V. S., Sinitsyn, I. N.*, Stochastic Differential Systems (Stokhasticheskie differentsialnye sistemy). M., Nauka, 1985 (in Russian).
2. *Mal'chikov, S. V.*, Determination of distribution output variables of multidimensional non-linear system. Autom. and Remote Control, **34** (1973), 1724–1729.
3. *Kashkarova, A. G., Shin, V. I.*, Modified semi-invariant methods for analysis of nonlinear stochastic systems. Autom. and Remote Control, **46**, *4* (1986).
4. *Pugachev, V. S.*, Recursive estimation of variables and parameters in stochastic systems described by autoregression equations. Soviet Math. Dokl., **19**, *4* (1978), 991–995.
5. *Pugachev, V. S.*, Recursive estimation of variables and parameters in stochastic systems described by difference equations. Soviet. Math. Dokl., **19**, *6* (1978), 1495–1497.
6. *Pugachev, V. S.*, Estimation of variables and parameters in discrete-time nonlinear systems. Autom. and Remote Control, **40** (1979), 512–521.

Приближенные методы определения конечномерных распределений случайных последовательностей, описываемых разностными уравнениями

В. С. ПУГАЧЕВ

(Москва)

Предложены приближенные методы нахождения конечномерных распределений случайных последовательностей, определяемых разностными уравнениями. Эти методы аналогичны методам, применяемым к случайным процессам, определяемым стохастическими дифференциальными уравнениями [1]. Для последовательности случайных величин $\{X_t\}$ со значениями в R^p, определяемой уравнением (1), где $\{V_t\}$ — последовательность случайных величин со значениями в R^q, а $\varphi_t(x, v)$ — известные функции, выводятся разностные уравнения (4) и (6) для конечномерных характеристических функций $g_{t_1,\ldots,t_n}(\lambda_1, \ldots, \lambda_n)$ $(n=1, 2, \ldots)$ последовательности $\{X_t\}$. Эти уравнения вместе с заданной характеристической функцией $g_1(\lambda)$ первого члена последовательности X_1 предполагаемого независимым от $\{V_t\}$, начальным условием (7) полностью определяют конечномерные распределения последовательности $\{\dot{X}_t\}$.

Для случая линейного уравнения (1) дается явное решение (11), (13) уравнений (4), (6) и (7).

Для случая нелинейного уравнения (1) предлагаются приближенные методы решения уравнений (4), (6) и (7), основанные на параметризации распределений.

В качестве простейшего метода предлагается метод нормальной аппроксимации распределений, сводящий задачу к решению приближенных разностных уравнений (14)–(16) для математического ожидания m_t, ковариационной матрицы K_t и ковариационной функции K_{t_1,t_2} последовательности $\{X_t\}$ при аппроксимации распределений нормальными правые части уравнений (14)–(16) представляют собой известные функции m_t, K_t и K_{t_1,t_2}.

В качестве методов, которые принципиально решают уравнения (4), (6) и (7) с любой степенью точности, предлагается метод моментов, приводящий задачу к решению приближенных разностных уравнений (17), (18), (21) или (19), (20) и (22), и методы ортогональных разложений, приводящие задачу к решению приближенных разностных уравнений (14)–(16), (27) и (28) и к применению приближенной формулы (23).

Для уменьшения числа уравнений для моментов или коэффициентов ортогональных разложений предлагается применять прием, предложенный в [2] для аналогичной задачи для процессов, определяемых стохастическими дифференциальными уравнениями, и аппроксимацию Мальчикова [3] для многомерных распределений.

Отмечается, что предложенные методы применимы для статистического анализа нелинейных дискретных систем, описываемых разностными уравнениями и для фильтров и экстраполяторов для процессов, определяемых разностными уравнениями [4–6].

В. С. Пугачев
Институт проблем информатики АН СССР
СССР, 117900 Москва ГСП-1,
Вавилова, 30

Problems of Control and Information Theory, Vol. 15 (2), pp. 111—127 (1986)

ROBUST ESTIMATION IN DISCRETE AND CONTINUOUS FAMILIES BY MEANS OF A MINIMUM CHI-SQUARE METHOD

I. VAJDA

(*Prague*)

(Received April 15, 1985)

Sharper variants of former results concerning minimum divergence estimators are presented. The class of minimum chi-square estimators, which can be defined as L_2-projections of generalized empirical d.f.'s into families of generalized theoretical d.f.'s is studied in more detail. Influence curves are evaluated and asymptotic normality is established. Examples concerning parameters of a homogeneous Markov chain and concerning parameters of location and scale are presented. Other examples concerning parameters of communication channels have alredy been presented in [19].

1. Introduction and basic results

The basic concepts and notations of this paper are as follows. $\mathscr{R}$ is the real line, $\mathscr{N}$ the set of natural numbers, $(\mathscr{X}, \mathscr{A})$ a measurable sample space, $\mathscr{P}$ the class of all probability distributions on $(\mathscr{X}, \mathscr{A})$ and $\mathscr{P}_* \subset \mathscr{P}$ the subclass of all empirical distributions defined by

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \text{ for every } \mathbf{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n,\ n \in \mathscr{N}, \tag{1.1}$$

where $\delta_x \in \mathscr{P}$ has all probability concentrated at $x \in \mathscr{X}$. Θ is a locally compact Hausdorff parameter space with countable base and Borel σ-algebra $\mathscr{B}$, $\mathscr{P}_{\boldsymbol{\Theta}} = \{P_\theta : \theta \in \Theta\} \subset \mathscr{P}$ is a "theoretical family" and $D: \mathscr{P}_{\boldsymbol{\Theta}} \times \mathscr{P} \mapsto [-\infty, \infty]$ an arbitrary function. We write simply $D(\theta, P)$ instead of $D(P_\theta, P)$ and $D(\theta, \mathbf{x})$ instead of $D(\theta, P_n)$ for $\mathbf{x} \mapsto P_n$ in the sense of (1.1). A *D-estimator* T is a mapping $\mathscr{P} \mapsto \Theta$ such that

$$T(P) \in \arg\min_{\boldsymbol{\Theta}} D(\theta, P) \subset \Theta \qquad \text{for every} \quad P \in \mathscr{P}, \tag{1.2}$$

$$T(\mathbf{x}) = T(P_n) \text{ for } \mathbf{x} \mapsto P_n \text{ is } (\mathscr{A}^n, \mathscr{B})\text{-measurable for every } n \in \mathscr{N}. \tag{1.3}$$

Note that in the case of particular functions D considered in [15, 16], condition (1.2) was replaced by $T(P) = \tau(\arg\min_{\boldsymbol{\Theta}} D(\theta, P))$ where $\tau(B) \in B$ for $B \subset \Theta$ is a fixed rule of choice satisfying some continuity condition. This specification proved to be

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

inconvenient. In particular it requires to modify the topology on $\exp \Theta$ considered in [15, 16] in order to meet the continuity conditions on a sufficiently general class of parameter spaces. Therefore we first establish the existence and equivariance results analoguous to what has been presented in [15, 16] for the present less restricted variants of D-estimators.

Theorem 1.1. Let $D(\theta, P)$ be continuous on Θ for every $P \in \mathscr{P}$ and $D(\theta, \mathbf{x})$ $\mathscr{A}^n$-measurable on $\mathscr{X}^n$ for every $\theta \in \Theta$, $n \in \mathscr{N}$, and let

$$D(\theta, P) < D(\theta^*, P) \quad \text{for some} \quad \theta = \theta(P, \theta^*) \in \Theta \quad \text{and all } P \in \mathscr{P},\ \theta^* \in \Theta^* - \Theta, \tag{1.4}$$

where Θ^* is a compact extension of Θ with countable base such that

$$D(\theta^*, P) = \lim_{\theta \to \theta^*} D(\theta, P) \qquad \text{for every} \quad P \in \mathscr{P},\ \theta^* \in \Theta^* - \Theta. \tag{1.5}$$

Then the D-estimator exists.

Proof. Let $P \in \mathscr{P}$ be arbitrary fixed. By (1.4),

$$\arg\min_{\Theta^*} D(\theta, P) = \arg\min_{\Theta} D(\theta, P). \tag{1.6}$$

We next prove that subset (1.6) is non-empty. Since Θ is σ-compact, there exist compacts $B_1 \subset B_2 \subset \ldots \subset \Theta$ with $\bigcup_{j=1}^{n} B_j = \Theta$ and parameters $\theta_j \in B_j$ such that $D(\theta_j, P) = \inf_{B_j} D(\theta, P)$. By Theorem 5 of Chapter 5 of [4], there exists a limit point θ_* of $\{\theta_j : j \in \mathscr{N}\} \subset \Theta$ and a subsequence $\theta_{j_k} \to \theta_*$ as $k \to \infty$. The identity

$$\lim_{k \to \infty} \inf_{B_{j_k}} D(\theta, P) = \inf_{\Theta} D(\theta, P)$$

implies $D(\theta_{j_k}, P) \to \inf_{\Theta} D(\theta, P)$. By assumptions, $D(\theta, P)$ is continuous on Θ^*. Consequently, the last convergence implies the identity $D(\theta_*, P) = \inf_{\Theta} D(\theta, P)$. This and (1.6) imply $\theta_* \in \arg\min_{\Theta^*} D(\theta, P)$ which proves the non-emptiness of (1.6). Thus there exists a mapping $T: \mathscr{P} \to \Theta$ satisfying (1.2) and it remains to prove that (1.3) can be satisfied as well. Let $n \in \mathscr{N}$ be arbitrary fixed. By the assumptions it holds that (1) $D(\theta, \mathbf{x})$ is continuous on Θ^* for all $\mathbf{x} \in \mathscr{X}^n$. (1) together with the $\mathscr{A}^n$-measurability assumption and with the last assertion of Theorem 1.9 of Pfanzagl [13] implies that (2) $\inf_B D(\theta, \mathbf{x})$ is $\mathscr{A}^n$-measurable for every compact $B \subset \Theta^*$. (1), (2) and the above established non-emptiness of (1.6) together with Theorem 3.10 of Pfanzagl [13] imply that there exists a mapping T_n with $T_n(\mathbf{x}) \in \arg\min_{\Theta^*} D(\theta, \mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}^n$ which is measurable w.r.t. $\mathscr{A}^n$ and the Borel σ-algebra on Θ^*. However, by (1.6), $T_n(\mathbf{x}) \in \arg\min_{\Theta} D(\theta, \mathbf{x})$ holds for all $\mathbf{x} \in \mathscr{X}^n$ so that T_n is $(\mathscr{A}^n, \mathscr{B})$-measurable. Q.E.D.

Let $\mathscr{X}$ be finite, $\mathscr{A} = \exp \mathscr{X}$, λ be the counting measure on $\mathscr{X}$, $p_\theta = dP_\theta / d\lambda$, $p_n = dP_n / d\lambda$, $p = dP / d\lambda$, and let

$$D(\theta, P) = E_\lambda p f(p_\theta / p) \qquad \text{for every} \quad \theta \in \Theta, P \in \mathscr{P} \tag{1.8}$$

be the f-divergence (Csiszár [3]; cf. also Perez [12]) where $f\colon [0, \infty] \mapsto [-\infty, \infty]$ is continuous on and twice differentiable in $[0, \infty]$ with $f''(u) \geqq 0$, $f''(1) > 0$, $f(1) = 0$, and where $0f(0/0) = 0$, $0f(p/0) = p \lim_{u\to\infty} f(u)/u$ for $p > 0$. The corresponding D-estimator will be called simply an *f-estimator*. The f-estimators have been first studied by Rao [14]. The next result seems to be new.

Corollary 1.1. If Θ is compact and $p_\theta(x)$ continuous on Θ for every $x \in \mathscr{X}$, then all f-estimators exist.

Proof. Since $\mathscr{A} = \exp \mathscr{X}$ and Θ is compact, all assumptions of Theorem 1.1 hold for $\Theta^* = \Theta$. Q.E.D.

Let for arbitrary $(\mathscr{X}, \mathscr{A})$ and Θ under consideration $\mathscr{P}_\Theta$ be dominated by a σ-finite measure λ on $(\mathscr{X}, \mathscr{A})$ with $p_\theta = dP_\theta/d\lambda$ and let for $\alpha \in (0, 1)$

$$D(\theta, P) = E_P(1 - p_\theta^\alpha)/\alpha \qquad \text{for every} \quad \theta \in \Theta,\ P \in \mathscr{P} \tag{1.9}$$

be the directed α-divergence (cf. [15, 16]). The corresponding M.D.E. will be called simply an *α-estimator*. The α-estimators are closely related to the minimum contrast estimators of Pfanzagl [13].

Corollary 1.2. If $p_\theta(x)$ is continuous on Θ with $\lim_{\theta\to\theta^*} p_\theta(x) = 0$ for every $x \in \mathscr{X}$ and $\theta^* \in \Theta^* - \Theta$, where Θ^* is a compact extension of Θ with countable base, and if $p_\theta(x)$ is positive and uniformly bounded on $\mathscr{X}$ for all $\theta \in \Theta$ then all α-estimators exist.

Proof. The assumption that $p_\theta(x)$ vanishes in a neighborhood of θ^* and the Lebesgue bounded convergence theorem imply that $D(\theta, P)$ is continuous on Θ and $D(\theta^*, P) = 1/\alpha$ for every $P \in \mathscr{P}$. Since $p_\theta > 0$, it holds $E_P p_\theta^\alpha > 0$ for every $\theta \in \Theta$, $P \in \mathscr{P}$, i.e. (1.4) holds as well. Finally, since

$$D(\theta, \mathbf{x}) = \frac{1}{\alpha}\left(1 - \frac{1}{n}\sum_{i=1}^{n} p_\theta(x_i)\right) \qquad \text{for every} \quad \mathbf{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n,\ n \in \mathscr{N},$$

the $\mathscr{A}^n$-measurability condition of Theorem 1.1 holds as well. Q.E.D.

Let Θ be arbitrary, let $\mathscr{X}$ be a Hausdorff topological space and let us consider the product topology on $\mathscr{X}^n$, $n \in \mathscr{N}$. We say that $\mathscr{A}_* = \{A_x\colon x \in \mathscr{X}\}$ is a $(\mathscr{P}, \mathscr{A})$-uniformity class on $(\mathscr{X}, \mathscr{A})$ if $\mathscr{A}_* \subset \mathscr{A}$, if the generalized d.f.'s $F(x) = P(A_x)$ are $\mathscr{A}$-measurable and if $\mathscr{A}_*$ is the $(P, \mathscr{A})$-uniformity class in the sense of Gaensler and Stute [6] for all $P \in \mathscr{P}$ (i.e. the Glivenko–Cantelli theorem: $\lim_{n\to\infty} \sup_{\mathscr{X}} |F_n(x) - F(x)| = 0$ a.s. $[P^\infty]$ holds for all $P \in \mathscr{P}$). Let further $\mathscr{F}$ be the class of all generalized d.f.'s, $\mathscr{F}_\Theta = \{F_\theta(x) = P_\theta(A_x)\colon \theta \in \Theta\} \subset \mathscr{F}$, and let for $\alpha \in [1, \infty)$

$$D(\theta, F) = E_{W_\theta} |F_\theta - F|^\alpha \qquad \text{for every} \quad \theta \in \Theta,\ F \in \mathscr{F} \tag{1.10}$$

be the weak χ^α-divergence (cf. [15, 16]), where $\{W_\theta\colon \theta \in \Theta\}$ is a family of measures on $(\mathscr{X}, \mathscr{A})$. The corresponding D-estimator will be called simply a *χ^α-estimator*. Millar [10] and Boos [1] formerly considered the χ^2-estimators of parameters of continuous families on $\mathscr{X} = \mathscr{R}$.

Corollary 1.3. Let $\{W_\theta: \theta\in\Theta\}$ be dominated by a σ-finite measure λ with $w_\theta = dW_\theta/d\lambda$ such that for every $\tilde\theta\in\Theta$ there exists an open neighborhood $B\subset\Theta$ with λ-integrable function $\tilde w(x)=\sup_B w_\theta(x)$, $x\in\mathscr{X}$. Further let $F_\theta(x)$ be continuous on Θ for every $x\in\mathscr{X}$ and $\mathbf{x}\mapsto F_n(x)$ continuous on $\mathscr{X}^n$ a.s. $[W_\theta]$ for every $x\in\mathscr{X}$, $\theta\in\Theta$ and $n\in\mathscr{N}$. Finally, let all χ^α-divergences (1.10) satisfy conditions (1.4), (1.5). Then all χ^α-estimators exist.

Proof. It holds $\sup_B w_\theta(x)|F_\theta(x)-F(x)|^\alpha \leqq \tilde w(x)$ for every $\tilde\theta\in\Theta$ and some open neighborhood $B\subset\Theta$ of $\tilde\theta$. Hence the Lebesgue bounded convergence theorem and the continuity assumptions imply that $D(\theta, F)$ is continuous on Θ for every $F\in\mathscr{F}$ and $D(\theta,\mathbf{x})$ is continuous and consequently $\mathscr{A}^n$-measurable on $\mathscr{X}^n$ for every $\theta\in\Theta$. Therefore all assumptions of Theorem 1.1 hold. Q.E.D.

Next we formulate conditions under which there exist equivariant versions of D-estimators. Let $\mathscr{P}_\Theta$ be structural with a parent $P_e\in\mathscr{P}$, i.e. let Θ be a group with unit e homomorphic with a group $[\Theta]$ of $\mathscr{A}$-measurable bijections $[\theta]: \mathscr{X}\mapsto\mathscr{X}$ such that $P_\theta = P_e[\theta]^{-1}$ for all $\theta\in\Theta$. Denote $\mathscr{P}_\Theta = P_{e/\Theta}$ ($\mathscr{F}_\Theta = F_{e/\Theta}$ for $F_e(x)=P_e(A_x)$ in the model of Corollary 1.3) and notice that $F_\theta(x)=P_\theta(A_x)=P_e([\theta]^{-1}(A_x))=F_e([\theta]^{-1}(x))$ for all $\theta\in\Theta$, $x\in\mathscr{X}$, iff

$$[\theta]A_x = A_{[\theta](x)} \qquad \text{for all} \quad \theta\in\Theta,\ x\in\mathscr{X}. \tag{1.11}$$

In models with structural families we define $\theta B=\{\theta\tilde\theta: \tilde\theta\in B\}\subset\Theta$ for all $\theta\in\Theta$, $B\subset\Theta$ and we say that an estimator $T: \mathscr{P}\mapsto\Theta$ is *equivariant* if

$$T(P[\theta]^{-1}) = \theta T(P) \qquad \text{for every} \quad \theta\in\Theta,\ P\in\mathscr{P}. \tag{1.12}$$

Let $\Theta\subset\mathscr{R}^m$ for some $m\in\mathscr{N}$ be alphabetically ordered according to the magnitude of coordinates $\theta_1, \ldots, \theta_m$ of $\theta\in\Theta$ (i.e. $\theta<\tilde\theta$ is $\theta_1<\tilde\theta_1$ or $\theta_1=\tilde\theta_1$ and $\theta_2<\tilde\theta_2, \ldots$). Let further

$$\theta<\tilde\theta \quad \text{iff} \quad \theta_*\theta<\theta_*\tilde\theta \qquad \text{for all} \quad \theta_*\in\Theta, \tag{1.13}$$

and let, for a compact subset $B\subset\Theta$, $\tau(B)$ denotes the unique element of B with the property $\tau(B)<\theta$ for all $\theta\in B$, $\theta\neq\tau(B)$ (cf. Lemma 1 in [18]).

Theorem 1.2. If D satisfies the conditions of Theorem 2.1 and the above considered conditions and if, moreover, for some $\Phi: \Theta\mapsto(0,\infty)$, $\Psi: \Theta\mapsto\mathscr{R}$

$$D(\tilde\theta, P[\theta]^{-1}) = \Phi(\theta)D(\theta^{-1}\tilde\theta, P)+\Psi(\theta) \quad \text{for every} \quad \theta,\tilde\theta\in\Theta \quad \text{and} \quad P\in\mathscr{P}, \tag{1.14}$$

then

$$\tilde T(P) = \tau(\arg\min_\Theta D(\theta, P)) \quad \text{for every} \quad P\in\mathscr{P} \tag{1.15}$$

is an equivariant version of the D-estimator T satisfying (1.2).

Proof. Since $D(\theta, P)$ is continuous on the compact Θ^*, $B=\arg\min_{\Theta^*} D(\theta, P) = \arg\min_\Theta D(\theta, P)$ (cf. (1.6)) is a compact subset of Θ. Therefore (1.15) defines a mapping $\tilde T: \mathscr{P}\mapsto\Theta$. We shall prove that this mapping satisfies (1.12). Let $\theta\in\Theta$ and $P\in\mathscr{P}$ be arbitrary fixed. By (1.14), $\tilde{\theta}_* \in \arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, P[\theta]^{-1})$ iff $\theta^{-1}\tilde{\theta}_* \in \arg\min_{\boldsymbol{\theta}} D(\theta, P)$ i.e. iff $\tilde{\theta}_* = \theta\bar{\theta}$ for $\bar{\theta} \in \arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, P)$. This implies the identity

$$\arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, P[\theta]^{-1} = \theta \arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, P).$$

Thus, by (1.13), $\tau(\arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, P[\theta]^{-1})) = \theta\tau(\arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, P))$ so that, by (1.15), (1.12) holds. Q.E.D.

2. χ^2-estimators for discrete families

Throughout this section we assume $\Theta \subset \mathscr{R}^m$ for some $m \in \mathscr{N}$. Let us first consider the f-estimators under the assumptions of Corollary 1.1. The function $f(u) = -\ln u$ defines the well-known M.L.E. The minimum divergence principle was first applied by Cramér [4] and Neyman [11]. They introduced the f-estimator $f(u) = (1-u)^2/u$ which minimizes the χ^2-statistic $E_\lambda[(p_n - p_\theta)^2/p_\theta]$. Later Rao [14] extended the approach of Cramér and Neyman to arbitrary differentiable f. It follows from the results of Rao that all f-estimators have essentially the same asymptotic properties as the M.L.E. The same conclusion applies to the robustness as well.

Indeed, suppose that p_θ is twice differentiable and let $p'_\theta = (d/d\theta)p_\theta$, $p''_\theta = (d/d\theta)^T p'_\theta$ for $(d/d\theta)^T = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_m)$. By Theorem 2.1 of [17], the influence curves $\Omega_\theta(x) = \lim_{\varepsilon\to 0}[T((1-\varepsilon)P_\theta + \varepsilon\delta_x) - T(P_\theta)]/\varepsilon$ of all f-estimators coincide and are given by

$$\Omega_\theta = \left[E_{P_\theta}\left(\left(\frac{p'_\theta}{p_\theta}\right)\left(\frac{p'_\theta}{p_\theta}\right)^T\right)\right]^{-1} \frac{p'_\theta}{p_\theta} \qquad \text{on} \quad \mathscr{X} \times \operatorname{int} \Theta. \tag{2.1}$$

Thus, in order to obtain robust alternatives to M.L.E.'s of parameters of discrete families, one has in fact to look after estimators outside the class of all f-estimators.

In the present section we analyze from this point of view the χ^2-estimators under the conditions of Corollary 1.3 specified further as follows: $\mathscr{X}$ is countable and $\mathscr{A} = \exp \mathscr{X}$, W_θ are identical for all $\theta \in \Theta$, Θ is compact and $\mathscr{A}_*$ arbitrary (each $\mathscr{A}_*$ is a $(\mathscr{P}, \mathscr{A})$-uniformity class for $(\mathscr{P}, \mathscr{A})$ under consideration). While p'_θ/p_θ in (2.1) is typically unbounded on $\mathscr{X}$ for every $\theta \in \operatorname{int}\Theta$, F'_θ in Theorem 2.2 below is typically bounded. Hence, by Theorem 2.2, the χ^2-estimators of the present section meet the Hampel-type criteria of robustness [7]. A deeper insight into the efficiency and robustness of these estimators and into the applicability of theorems proved in this section is provided by examples presented in the next two sections.

Thus let us consider a χ^2-estimator T considered in Corollary 1.3, where $\mathscr{X} = \{0, 1, \ldots\}$, $\mathscr{A} = \exp \mathscr{X}$, $A_x = \{0, 1, \ldots, a(x)\}$ for $x \in \mathscr{X}$ and for a non-decreasing $a(x)$ with $a^{-1}(x) = \inf\{\tilde{x} \in \mathscr{X} : a(\tilde{x}) = x\}$, $\mathscr{F}$ is the class of all generalized d.f.'s $F(x) = P(A_x)$, $F_\theta(x) = P_\theta(A_x)$ and $W_\theta = W$. Let us suppose that Θ is compact and $F_\theta(x)$ continuous on

Θ for every $x\in\mathscr{X}$ so that T exists (cf. Corollary 1.3). Finally, let us define for every $x\in\mathscr{X}$ and $\theta\in\Theta$ the influence curve

$$\Omega_\theta(x)=\lim_{\varepsilon\to 0}\frac{T(G_\varepsilon)-T(F_\theta)}{\varepsilon}\quad\text{for}\quad G_\varepsilon=(1-\varepsilon)F_\theta+\varepsilon\chi_{[a^{-1}(x),\infty)}\in\mathscr{F},\tag{2.2}$$

where $\chi_{[a^{-1}(x),\infty)}\in\mathscr{F}$ is the generalized d.f. of $\delta_x\in\mathscr{P}$.

Theorem 2.1. If, for some $F\in\mathscr{F}$, $\{T(F)\}=\arg\min_{\boldsymbol{\theta}}D(\theta,F)$ then $\lim_{n\to\infty}T(F_n)=T(F)$ a.s. $[P^\infty]$.

Proof. Let $\theta\in\Theta$ be a limit point of $\{T(F_n):n\in\mathscr{N}\}\subset\Theta$, let $T_j=T(F_{n_j})\to\theta_*$ as $j\to\infty$, and let $\theta\in\Theta$ be arbitrary fixed. Since $|E_W(F_\theta-\tilde{F})^2-E_W(F_\theta-F)^2|\leqq 2E_W|\tilde{F}-F|$ for all $F,\tilde{F}\in\mathscr{F}$, it holds

$$|D(\theta,\tilde{F})-D(\theta,F)|\leqq 2\sup\nolimits_{\mathscr{X}}|\tilde{F}(x)-F(x)|.\tag*{(2)3.}$$

Hence, by the Glivenko–Cantelli theorem, $D(\theta,F_{n_j})\to D(\theta,F)$ as $j\to\infty$ a.s. $[P^\infty]$. Since at the same time $D(T_j,F_{n_j})\leqq D(\theta,F_{n_j})$, it holds

$$\limsup_{j\to\infty}D(T_j,F_{n_j})\leqq D(\theta,F)\quad\text{a.s.}\quad[P^\infty].\tag{2.4}$$

Further it holds

$$|D(T_j,F_{n_j})-D(\theta_*,F)|\leqq|D(T_j,F_{n_j})-D(T_j,F)|+|D(T_j,F)-D(\theta_*,F)|$$

where the first right-hand term tends to zero a.s. $[P^\infty]$ (cf. (2.3) and the Glivenko–Cantelli theorem) and the second term does the same (cf. the assumed continuity and boundedness of the generalized d.f.'s). Therefore $\lim_{j\to\infty}D(T_j,F_{n_j})=D(\theta_*,F)$ a.s. $[P^\infty]$ which together with (2.4) yields $D(\theta_*,F)\leqq D(\theta,F)$. Since θ was arbitrary, this inequality implies $\theta_*\in\arg\min_{\boldsymbol{\theta}}D(\theta,F)$, i.e. $\theta_*=T(F)$. Hence $T(F)$ is the a.s. $[P^\infty]$ unique limit point of $\{T_n:n\in\mathscr{N}\}$. Q.E.D.

Corollary 2.1. If for no $\tilde{\theta}\neq\theta$ $F_{\tilde{\theta}}=F_\theta$ a.s. $[W]$ then $\lim_{n\to\infty}T(F_n)=\theta$ a.s. $[P_\theta^\infty]$ for all $\theta\in\Theta$.

Proof. By the assumption, every $F=F_\theta$, $\theta\in\Theta$, satisfies the unambiguity assumption of Theorem 2.1. with $T(F)=\theta$. Q.E.D.

Theorem 2.2. Let the assumption of Corollary 2.1 hold, let* $F'_\theta=(d/d\theta)F_\theta$, $F''_\theta=(d/d\theta)^TF'_\theta$ be continuous on intΘ for every $x\in\mathscr{X}$ and bounded on $\mathscr{X}$ for every $\theta\in\text{int}\Theta$ and let $E_W(F'_\theta F'^T_\theta)>0$ on intΘ. Then the influence curve (2.2) is given by

$$\Omega_\theta(x)=[E_W(F'_\theta F'^T_\theta)]^{-1}E_W[\chi_{[a(x)^{-1},\infty)}-F_\theta)F'_\theta]\qquad\text{for all}\quad x\in\mathscr{X},\ \theta\in\text{int}\Theta.\tag{2.5}$$

* Here and in the sequel, the superscript T denotes transposed vectors or matrices.

Proof. First we shall prove that $\lim_{\varepsilon\to 0} T(G_\varepsilon)=\theta$ for arbitrary fixed $x\in\mathscr{X}$, $\theta\in\operatorname{int}\Theta$. By (2.3)

$$|D(\tilde\theta, G_\varepsilon)-D(\tilde\theta, F_\theta)|\leqq 2\varepsilon \sup\nolimits_{\mathscr{X}}|F_\theta-\chi_{[a^{-1}(x),\infty)}|\leqq 2\varepsilon \qquad \text{for every} \quad \tilde\theta\in\Theta \qquad (2.6)$$

so that $\lim_{\varepsilon\to 0} D(\theta, G_\varepsilon)=D(\theta, F_\theta)$. This and the inequality $D(T(G_\varepsilon), G_\varepsilon)\leqq D(\theta, G_\varepsilon)$ imply

$$\limsup_{\varepsilon\to 0} D(T(G_\varepsilon), G_\varepsilon)\leqq D(\theta, F_\theta). \qquad (2.7)$$

On the other hand, the inequality

$$|D(T(G_\varepsilon), G_\varepsilon)-D(\theta^*, F_\theta)|\leqq|D(T(G_\varepsilon), G_\varepsilon)-D(T(G_\varepsilon), F_\theta)|+$$
$$+|D(T(G_\varepsilon), F_\theta)-D(\theta^*, F_\theta)|$$

together with (2.6) and with the continuity of $\tilde\theta\mapsto D(\tilde\theta, F_\theta)$ implies $\lim_{j\to\infty} D(T(G_{\varepsilon_j}), G_{\varepsilon_j})=D(\theta^*, F_\theta)$ for every limit $\theta^*\in\Theta$ of $T(G_{\varepsilon_j})$ for some $\varepsilon_j\to 0$. Hence, by (2.7), $D(\theta^*, F_\theta)\in D(\theta, F_\theta)$, i.e. $\theta^*=\theta=T(F_\theta)$ (cf. the proof of Corollary 2.1).

The rest of the proof is now easy. The existence and boundedness of F'_θ, F''_θ implies for every $\theta\in\Theta$, $F\in\mathscr{F}$

$$D'(\Theta, F)=(d/d\theta)D(\theta, F)=2E_W(F_\theta-F)F'_\theta,$$
$$D''(\theta, F)=(d/d\theta)^T D(\theta, F)'=2E_W(F'_\theta F'^T_\theta)+2E_W((F_\theta-F)F''_\theta). \qquad (2.8)$$

Since $T(F_\theta)=\theta$, (2.8) implies $D'(T(F_\theta), G_\varepsilon)=E_W((F_\theta-\chi_{[a^{-1}(x),\infty)})F'_\theta)$ for all $\theta\in\operatorname{int}\Theta$, $x\in\mathscr{X}$, $\varepsilon\in(0,1)$. By the first assertion of this proof, $D'(T(G_\varepsilon), G_\varepsilon)=0$ for all sufficiently small ε. Therefore

$$D'(T(F_\theta), G_\varepsilon)-D'(T(G_\varepsilon), G_\varepsilon)=2\varepsilon E_W((F_\theta-\chi_{[a^{-1}(x),\infty)})F'_\theta). \qquad (2.9)$$

The mean value theorem applied coordinatewise to

$$\Delta(u)=D'(uT(F_\theta)+(1-u)T(G_\varepsilon), G_\varepsilon),\ u\in[0,1],$$

implies

$$D'(T(F_\theta), G_\varepsilon)-D'(T(G_\varepsilon), G_\varepsilon)=D''(\theta_\varepsilon, G_\varepsilon)(T(F_\theta)-T(G_\varepsilon)) \qquad (2.10)$$

where $\theta_\varepsilon\in\operatorname{int}\Theta$ (possibly different for different coordinates) tends to $T(F_\theta)$ as $\varepsilon\to 0$. Since $G_\varepsilon\to F_\theta$ pointwise on $\mathscr{X}$ as $\varepsilon\to 0$ and both integrands in the second equality of (2.8) are bounded, it holds $\lim_{\varepsilon\to 0} D''(\theta_\varepsilon, G_\varepsilon)=D''(T(F_\theta), F_\theta)=2E_W(F'_\theta F'^T_\theta))$. (2.5) now follows from (2.2), (2.9), (2.10). Q.E.D.

Theorem 2.3. If the assumptions of Theorem 2.2 hold then $\sqrt{n}(T(x)-\theta)$ tends P_θ^∞-weakly to $N(E_{F_\theta}\Omega_\theta, \operatorname{Var}_{F_\theta}\Omega_\theta)$ for all $\theta\in\operatorname{int}\Theta$.

Proof. Let $\theta \in \operatorname{int} \Theta$ be arbitrary fixed. By (2.8) it holds for every $\mathbf{x}= =(x_1, \ldots, x_n) \mapsto F_n$

$$D'(\theta, F_n)=2E_W((F_\theta - F_n)F'_\theta)=2E_W\left(\left(F_\theta - \frac{1}{n}\sum_{i=1}^{n} \chi_{[a^{-1}(x_i), \infty)}\right)F'_\theta\right)=$$

$$=\frac{1}{n}\sum_{i=1}^{n} E_W((F_\theta - \chi_{[a^{-1}(x_i), \infty)})F'_\theta)= -E_W(F'_\theta F'^T_\theta)\frac{1}{n}\sum_{i=1}^{n} \Omega_\theta(x_i). \tag{2.11}$$

Further, by Corollary 2.1, for all sufficiently large n it holds $D'(T(F_n), F_n)=0$ a.s. $[P_\theta^\infty]$. Hence, by (2.11),

$$D'(T(F_\theta), F_n)-D'(T(F_n), F_n)= -E_W(F'_\theta F'^T_\theta)\frac{1}{n}\sum_{i=1}^{n} \Omega_\theta(x_i). \tag{2.12}$$

On the other hand, by the mean value theorem applied coordinatewise to the function $\Delta(u)=D'(uT(F_\theta)+(1-u)T(F_n), F_n)$, $u \in [0, 1]$, it holds for all sufficiently large $n \in \mathcal{N}$ that the linear segments connecting $T(F_\theta)$ and $T(F_n)$ contain some θ_n (possibly different for different coordinates) such that, coordinatewise,

$$D'(T(F_\theta), F_n)-D'(T(F_n), F_n)=D''(\theta_n, F_n)(T(F_\theta)-T(F_n))=$$
$$= -D''(\theta_n, F_n)(T(\mathbf{x})-\theta).$$

Moreover, by the definition of θ_n, by Corollary 2.1, by the Glivenko–Cantelli theorem and by (2.8),

$$D''(\theta_n, F_n) \to D''(\theta, F_\theta)=E_W(F'_\theta F'^T_\theta) \quad \text{a.s.} \quad [P_\theta^\infty].$$

The desired assertion now follows from (2.12), from the multidimensional central limit theorem and from the Cramér–Slutskij theorem (cf. Fuller [5], pp. 140–145).

3. χ^2-estimators for geometrical family

Let in the model of Section 2 $\Theta=[0, 1]$, $p_\theta(0)=0$ for all $\theta \in [0, 1]$ and

$$p_\theta(x)=\begin{cases}(1-\theta)\theta^{x-1} & \text{if } \theta \in [0, 1) \\ \chi_{\{x\}}(k) & \text{if } \theta=1, \quad k \in \mathcal{N} \quad \text{fixed}\end{cases}$$

for $x \geqq 1$. Let us consider the χ^2-estimators T_r defined by* $A_x=\{0, 1, \ldots, r\}$ for all $x \in \mathcal{X}$ where $0<r \leqq k$. Since $F_\theta(x)=1-\theta^r$ for all $x \in \mathcal{X}$, $\theta \in \Theta$, by the definition in Section 1,

* Since the sets A_x are fixed for all $x \in \mathcal{X}$, all generalized d.f.'s considered below are constants independent of $x \in \mathcal{X}$.

$D(\theta, F)=(1-\theta^r-F(r))^2$ for all $\theta\in\Theta$, $F\in\mathscr{F}$, so that $T_r(F)=(1-F(r))^{1/r}$ unambiguously for all $F\in\mathscr{F}$ (cf. Corollary 1.3). By Corollary 2.1, every T_r is strongly consistent. Since $F'_\theta(x)=r\theta^{r-1}$, $a^{-1}(x)=0$ for all $x\in\mathscr{X}$, the influence curve of T_r is given by

$$\Omega_\theta(x)=\begin{cases}\theta/r & \text{if } \ 0\leqq x\leqq r\\ -(1-\theta^r)/(r\theta^{r-1}) & \text{if } \ x>r\end{cases}\tag{3.1}$$

for everv $\theta\in\operatorname{int}\Theta=(0, 1)$. Moreover, $\sqrt{n}(T_r(\mathbf{x})-\theta)$ is asymptotically normal with the asymptotic mean zero and the asymptotic variance

$$\mathscr{V}_{T_r}(\theta)=\left(\frac{\theta}{r}\right)^2\frac{1-\theta^r}{\theta^r}\tag{3.2}$$

(cf. Theorems 2.2, 2.3).

Let us now compare the χ^2-estimator with the M.L.E. which is here defined for all $P\in\mathscr{P}$ by

$$\tilde{T}(P)=\frac{\mu(P)}{1+\mu(P)} \qquad \text{where} \quad \mu(P)=\sum_{x\in\mathscr{X}}xP(\{x\}).$$

The influence curve of $\tilde{T}$(as well as of all other regular f-estimators, cf. (2.1)) is given by

$$\tilde{\Omega}_\theta(x)=\left\langle\begin{matrix}0 & \text{if} \ \ x=0\\ (1-\theta)^2(x-1)-\theta(1-\theta) & \text{if} \ \ x\geqq 1\end{matrix}\right.$$

for every $\theta\in(0, 1)$. Curve (3.1) is bounded on $\mathscr{X}$ (on Θ as well provided $r=1$) while $\tilde{\Omega}_\theta(x)$ is unbounded. Applying the central limit theorem to the identity

$$\tilde{T}(\mathbf{x})=\frac{\bar{\mathbf{x}}}{1+\bar{\mathbf{x}}} \qquad \text{for all} \quad \mathbf{x}\in\mathscr{X}^n,\ n\in\mathscr{N},$$

one obtains that $\sqrt{n}(\tilde{T}(\mathbf{x})-\theta)$ is asymptotically normal with the asymptotic mean zero and the asymptotic variance $\mathscr{V}_{\tilde{T}}(\theta)=\theta(1-\theta)^2$. Figure 3.1 presents the square root of this variance as well as that of (3.2) for several $r\in\mathscr{N}$. We see that for every $\theta\in\Theta$ one can choose among T_1, T_2, T_4, T_{10} a χ^2-estimator with a deficiency between 10–20%. Some knowledge about prior probability of intervals [0, 0.35), [0.35, 0.6), [0.6, 0.8), [0.8, 1] is needed if the best T_r has to be used.*

Another comparison of robustness of T_r and T can be done by means of the upper and lower bounds $U_r(\theta)$, $L_r(\theta)$ and $\tilde{U}(\theta)$, $\tilde{L}(\theta)$ for $T((1-\varepsilon)P_\theta+\varepsilon P^*)$ yield by the least

* In applications like that of [19], where the sample size is very large (e.g. $n\approx 10^6$), the asymptotic variances of Fig. 3.1 have a practical meaning.

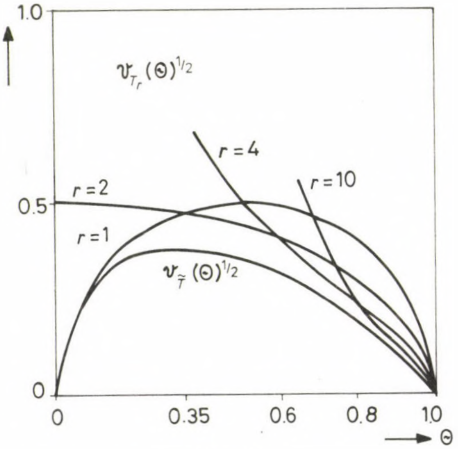


Fig. 3.1

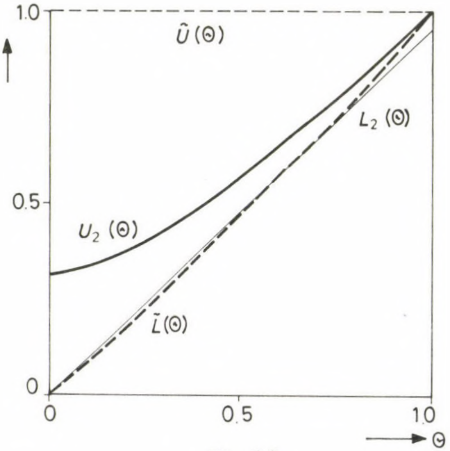


Fig. 3.2

favourable contaminations $P^* \in \mathscr{P}$. Put $Q_\varepsilon = (1-\varepsilon)P_\theta + \varepsilon P^*$. By the above found explicit expressions for T_r, $\tilde{T}$, it holds

$$T_r(Q_\varepsilon) = [\theta^r(1-\varepsilon) + \varepsilon(1-F^*(r))]^{1/r},$$

$$\tilde{T}(Q_\varepsilon) = 1 - \frac{1-\theta}{1+\varepsilon\mu(P^*)+\varepsilon\theta(1+\mu(P^*))}.$$

From here we find for all $\theta \in \Theta$

$$U_r(\theta) = (\theta^r(1-\varepsilon)+\varepsilon)^{1/r}, \qquad L_r(\theta) = \theta(1-\varepsilon)^{1/r},$$

$$\tilde{U}(\theta) = 1, \qquad \tilde{L}(\theta) = \frac{\theta(1-\varepsilon)}{1-\theta\varepsilon}.$$

These bounds display a principal difference between the robust χ^2-estimators and the non-robust M.L.E. This difference can be seen from Fig. 3.2, where all four bounds $\tilde{U}$, $\tilde{L}$, U_2, L_2 are drawn for $\varepsilon = 0.1$.

4. χ^2-estimators of parameters of Markov chains

The χ^2-estimators of Section 2 can estimate parameters of random sequences with discrete state spaces in a manner permitting to employ the general results of Section 2 to obtain the consistency, asymptotic normality and the influence curve. This is illustrated by an example that follow.

Let us consider the class of all Markov chains with two states 0 and 1, with the initial state 1 and with the probabilities μ, σ of transitions $0\to 0$, $1\to 0$ respectively, where $\theta=(\mu,\sigma)\in[0,1]^2=\Theta$. Let $\Theta_1=\{1\}\times(0,1]\subset\Theta$, $\Theta_0=\Theta-\Theta_1$. Obviously, each $\theta\in\Theta$ defines a probability distribution ω_θ on the σ-algebra generated by cylinders in $S=\{0,1\}^\infty$ such that

$$\omega_\theta(S_0)=\begin{cases}1 & \text{if}\quad \theta\in\Theta_1\\ 0 & \text{if}\quad \theta\in\Theta_0\end{cases}$$

where

$$S_0=\{(s_1,s_2,\dots)\in S:\limsup_{i\to\infty} s_i=0\}.$$

Let us define r.v.'s $x_1, x_2, \dots$ on S by $x_i=x_i(s)=k\in\mathcal{N}$ fixed or $x_i=x_i(s)=$ number of zeros between the i-th and $(i+1)$-st unit in the sequence $1, s_1, s_2, \dots$, depending on whether $s=(s_1,s_2,\dots)\in S_0$ or $S-S_0$ respectively. The r.v.'s $x_1, x_2, \dots$ are i.i.d. with a distribution P_θ on the same $(\mathcal{X},\mathcal{A})$ as in the preceding sections, where

$$P_\theta(\{x\})=\begin{cases}\chi_{\{k\}}(x) & \text{for}\quad \theta\in\Theta_1,\\ \begin{cases}1-\sigma & \text{if}\quad x=0\\ \sigma(1-\mu)\mu^{x-1} & \text{if}\quad x\geqq 1\end{cases} & \text{for}\quad \theta\in\Theta_0.\end{cases}$$

Thus for $\theta\in\Theta_0$ it holds $P_\theta(\{0, 1, \dots, x\})=1-\sigma+\sigma(1-\mu^x)=1-\sigma\mu^x$. Let us now consider the χ^2-estimator T_r defined for a fixed $0<r\leqq k$ by $A_0=\{0\}$, $A_x=\{0,1,\dots,r\}$ if $x\geqq 1$. The generalized d.f.'s defined by P_θ are given as follows

$$F_\theta(x)=\begin{cases}1-\sigma & \text{if}\quad x=0\\ 1-\sigma\mu^r & \text{if}\quad x\geqq 1\end{cases}\quad \text{for every}\quad \theta=(\mu,\sigma)\in\Theta,\ x\in\mathcal{X}. \tag{4.1}$$

By (4.1) and (1.10), $D(\theta, F)=w(1-\sigma-F(\theta))^2+(1-w)(1-\sigma\mu^r-F(r))^2$ for every $\theta\in\Theta$, $F\in\mathcal{F}$, where $w=W(\{0\})$ is assumed to be from $[0, 1)$. Hence $T_r(F)$ is for every $F\in\mathcal{F}$ unambiguously given by

$$T_r(F)=(\mu_r(F),\sigma_r(F))=\left(\left[\frac{1-F(r)}{1-F(0)}\right]^{1/r},\quad 1-F(0)\right). \tag{4.2}$$

The strong consistency of all estimators T_r follows from (4.1) and Corollary 2.1.

The two-dimensional influence curve, the asymptotic normality and the asymptotic mean and variance explicitly follow from (4.1) and Theorems 2.2, 2.3 — details are space-consuming and are thus omitted here. Comparing the influence curves T_r with the influence curve of the M.L.E. $\tilde{T}$ one encounters a situation analogical to that discussed in Section 3.

It is seen from what has been said above that the χ^2-estimation method is nothing but a least square (L_2-projection) method applied to the empirical and theoretical d.f.'s rather than, as usual, to the data and their linear statistical models.

5. χ^2-estimators for continuous families

Hereafter we consider the χ^2-estimators T under the conditions of Corollary 1.3 specified further as follows: $\mathscr{X} = \mathscr{R}^k$ for $k \in \mathscr{N}$, $\mathscr{A}$ is the corresponding Borel σ-algebra, $A_x = (-\infty, x)$ are the products of intervals upperbounded by the corresponding coordinates of $x \in \mathscr{R}^k$ so that the generalized d.f.'s reduce to the ordinary multidimensional d.f.'s ($\mathscr{A}_*$ is then a $(\mathscr{P}, \mathscr{A})$-uniformity class on $(\mathscr{X}, \mathscr{A})$), $\mathscr{F}_{\boldsymbol{\Theta}}$ is structural (i.e. $\mathscr{F}_{\boldsymbol{\Theta}} = F_{e/\boldsymbol{\Theta}}$ where $F_e \in \mathscr{F}$ is a parent of the family $\mathscr{F}_{\boldsymbol{\Theta}}$) and $\{W_\theta : \theta \in \Theta\} = W_{e/\boldsymbol{\Theta}} \subset \mathscr{P}$ is structural as well with a parent $W_e = W$. Note that under the present assumptions, (1.11) holds so that $F_\theta = F_e[\theta]^{-1}$. Let $F_e(x)$ be continuous on $\mathscr{X}$ and $[\theta]^{-1}(x)$ continuous on Θ for every $x \in \mathscr{X}$. Let further λ be the Lebesgue measure on $\mathscr{X}^k$ and denote $w = w_e$. The substitution rule for integrals and (1.10) imply under the present assumptions

$$D(\theta, F) = E_W(F_e - F[\theta])^2 \qquad \text{for every} \quad \theta \in \Theta,\ F \in \mathscr{F}. \tag{5.1}$$

Then, by Corollary 1.3, the estimator T exists. Moreover, it follows from (5.1) and from the identity $F_\theta = F_e[\theta]^{-1}$ that (1.14) holds for $\Phi(\theta) = 1$, $\Psi(\theta) = 0$ on Θ. Therefore, by Theorem 1.2, there exists an equivariant version of T provided the parameter space satisfies the conditions $\Theta \subset \mathscr{R}^m$ and (1.13).

Theorem 5.1. If $T(F)$ is unambiguous for some $F \subset \mathscr{F}$ in the sense $\{T(F)\} = \arg\min_{\boldsymbol{\theta}} D(\theta, F)$, then $\lim_{n\to\infty} T(F_n) = T(F)$ a.s. $[P^\infty]$.

Proof. Since each W_θ is supposed to be a probability measure, inequality (2.3) holds. Hence the proof of this statement is essentially the same as the proof of Theorem 2.1. The only difference is that one has to prove that the limit point $\theta_* \in \Theta^*$ is not from $\Theta^* - \Theta$. This can be proved by contradiction. If $\theta_* = \theta^* \in \Theta^* - \Theta$ then, by the definition in (1.5), $D(\theta^*, F) = \lim_{j\to\infty} D(T_j, F)$ for arbitrary fixed $F \in \mathscr{F}$. Since by (2.3) it holds $|D(T_j, F) - D(T_j, F_{n_j})| \to 0$ a.s. $[P^\infty]$, the last inequality implies $D(\theta^*, F) = \lim_{j\to\infty} D(T_j, F_{n_j})$ a.s. $[P^\infty]$. Hence by (2.4) it holds $D(\theta^*, F) \leqq D(\theta, F)$ for all $\theta \in \Theta$ which contradicts assumption (1.11). Q.E.D.

Corollary 5.1. If for no $\theta \in \Theta$, $\theta \neq e$, $F_\theta = F_e$ a.s. $[W]$, then $\lim_{n\to\infty} T(F_n) = \theta$ a.s. $[P_\theta^\infty]$ for all $\theta \in \Theta$.

Proof. If $F = F_\theta$ for arbitrary fixed $\theta \in \Theta$, then it follows from (5.1) and from the identities $[\theta]^{-1}([\theta]) = [\theta^{-1}]([\theta]) = [\theta^{-1}\theta] = [e]$ that θ belongs to $\arg\min_{\boldsymbol{\theta}} D(\tilde{\theta}, F)$ and that no other $\tilde{\theta} \in \Theta$ belongs to this set as long as the Corollary condition holds. Thus $T(F_\theta) = \theta$ is unambiguous as required in Theorem 5.1. Q.E.D.

In the rest of this paper we assume $\Theta \subset \mathscr{R}^m$, $m \in \mathscr{N}$. Let us define for every $x \in \mathscr{X}$ and $\theta \in \Theta$ the influence curve

$$\Omega_\theta(x) = \lim_{\varepsilon \to 0} \frac{T(G_\varepsilon) - T(F_\theta)}{\varepsilon} \qquad \text{where} \quad G_\varepsilon = (1-\varepsilon)F + \varepsilon\chi_{(x,\infty)} \in \mathscr{F}$$

and where $\chi_{(x,\infty)} \in \mathscr{F}$ is the d.f. of $1_{\{x\}} \in \mathscr{P}$ (analogously as $(-\infty, x)$, (x, ∞) denotes the product of intervals on $\mathscr{R}$ all elements of which dominate the corresponding coordinates of $x \in \mathscr{R}^k$).

In the next theorem we write simply $w_\theta = J(\theta^{-1})w[\theta]^{-1}$, $F_\theta = F_e[\theta]^{-1}$, $w'_\theta = (d/d\theta)w_\theta$, $w''_\theta(d/d\theta)^T w'_\theta$, $F'_\theta = (d/d\theta)F_\theta$, $F''_\theta = (d/d\theta)^T F'_\theta$. We also consider

$$p(x) = \frac{d}{dx} F_e(x), \quad t(x) = \left(\frac{d}{d\theta}[\theta]^{-1}(x)\right)_{\theta=e} \quad \text{for all} \quad x \in \mathscr{X} \tag{5.2}$$

where $d/d\theta$ is the same as in Section 2 and d/dx is a similar $(k \times 1)$-matrix type differential operator.

Theorem 5.2. Let the condition of Corollary 5.1 hold, let all elements of matrices $w'_\theta w''_\theta$, F'_θ, F''_θ be continuous on Θ for every $x \in \mathscr{X}$ and the elements of F'_θ, F''_θ bounded on $\Theta \times \mathscr{X}$. Finally let the elements of w'_θ, w''_θ satisfy the same λ-integrability condition as w_θ in Corollary 1.3 and let $\infty > E_W(tpp^Tt^T) > 0$. Then for every $x \in \mathscr{X}$ and $\theta \in \operatorname{int} \Theta$

$$\Omega_\theta(x) = [E_W(tpp^Tt^T)]^{-1} E_W((\chi_{x,\infty)}[\theta] - F_e)pt). \tag{5.3}$$

Proof. As said in the proof of Theorem 5.1, (2.3) is true for the divergences D under consideration as well. Hence, by the first part of the proof of Theorem 2.2, $\lim_{\varepsilon \to 0} T(G_\varepsilon) = \theta$ for all $x \in \mathscr{X}$, $\theta \in \operatorname{int}\Theta$. Further, by the assumptions of Theorem 5.2 and 3° in 7.2.C of Loève [9],

$$D'(\theta, F) = \frac{d}{d\theta} d(\theta, F) = E_\lambda[w'_\theta(F_\theta - F)^2 + 2w_\theta(F_\theta - F)F'_\theta]$$

$$D''(\theta, F) = \left(\frac{d}{d\theta}\right)^T D'(\theta, F) = \tag{5.4}$$

$$= E_\lambda[w''_\theta(F_\theta - F)^2 + 2w'_\theta(F_\theta - F)F'^{\,T}_\theta + 2w_\theta F'_\theta F'^{\,T}_\theta + 2w_\theta(F\theta - F)F''_\theta],$$

for every $\theta \in \operatorname{int} \Theta$, $F \in \mathscr{F}$. Since $T(F_\theta) = \theta$, (5.4) implies

$$D'(T(F_\theta), G_\varepsilon) = \varepsilon^2 E_\lambda w'_\theta(F_\theta - \chi_{(x,\infty)})^2 + 2\varepsilon F_{w_\theta}((F_\theta - \chi_{(x,\infty)})F'_\theta).$$

Employing the substitution $\tilde{x} \mapsto [\theta]^{-1}(\tilde{x})$ for $\tilde{x} \in \mathscr{X}$ we obtain

$$\lim_{\varepsilon \to 0} \frac{D'(T(F_\theta), G_\varepsilon)}{\varepsilon} = 2E_W((F_e - \chi_{(x,\infty)}[\theta])F'_\theta). \tag{5.5}$$

Using the same argument as presented in the end of the proof of Theorem 2.2, we obtain from (5.4)–(5.5)

$$\Omega_\theta(x) = [E_{W_\theta}(F'_\theta F'^T_\theta)]^{-1} E_W((\chi_{(x,\infty)}[\theta] - F_e)F'_e) \qquad \text{for every} \quad x \in \mathscr{X},\ \theta \in \operatorname{int} \Theta.$$

Since $E_{W_\theta}(F'_\theta F'^T_\theta) = E_W(F'_e F'^T_e)$, it remains to apply (5.2) to this identity in order to obtain (5.3). Q.E.D.

Corollary 5.2. If the assumptions of Theorem 5.2 hold for $k=1$ and if $x \leqq \tilde{x}$ iff $[\theta]\,(x) \leqq [\theta]\,(\tilde{x})$ for every $\theta \in \Theta$, then the influence curve is given by (5.3) with

$$E_W((\chi_{(x,\infty)}[\theta] - F_e)pt) = -(\psi([\theta]^{-1}(x)) - E_{F_e}\psi) \tag{5.6}$$

where

$$\psi(y) = \int_{-\infty}^{y} w(x)p(x)t(x)dx \qquad \text{for every} \quad y \in \mathscr{X}. \tag{5.7}$$

Proof. Let $x \in \mathscr{X}$ and $\theta \in \operatorname{int} \Theta$ be arbitrary fixed and put $y = [\theta]^{-1}(x) \in \mathscr{X}$. It holds

$$\psi(\infty) = \varphi(\infty)F_e(\infty) - \psi(-\infty)F_e(-\infty) = E_\lambda d(\psi F_e)/dx =$$
$$= E_\lambda(wptF_e) + E_\lambda(\psi p) = E_W ptF_e + E_{F_e}\psi,$$

so that

$$E_W ptF_e = \psi(\infty) - E_{F_e}\psi. \tag{5.8}$$

Further it holds $\chi_{(x,\infty)}[\theta] = \chi_{(y,\infty)}$ so that

$$E_W(pt\chi_{(x,\infty)}[\theta]) = E_\lambda(wpt\chi_{(y,\infty)}) =$$
$$= E_\lambda(wpt) - \int_{-\infty}^{y} w(x)p(x)t(x)dx = \psi(\infty) - \psi(y).$$

Since by assumptions $\psi(\infty) = E_W pt = E_W F'_e < \infty$, (5.6) follows from here and from (5.8). Q.E.D.

It follows from (5.3) and (5.6) that all χ^2-estimators satisfying the conditions of Corollary 5.2 are robust in the sense that their gross-error and local-shift sensitivities (cf. Hampel [7]) are bounded.

Theorem 5.3. If the assumptions of Corollary 5.2 hold and $\sup_{\mathscr{X}} |w'_\theta / w_\theta| < \infty$ then $\sqrt{n}(T(\mathbf{x}) - \theta)$ tends P_θ^∞-weakly to $N(0, \operatorname{Var}_{F_\theta}\Omega_\theta)$.

Proof. Let $\theta \in \operatorname{int} \Theta$ be arbitrary fixed. Then $D'(\theta, F_n)$ formally differs from (2.11) by the term

$$|E_\lambda w'_\theta(F_\theta - F_n)^2| \leqq \sup_{\mathscr{X}} \left|\frac{w'_\theta}{w_\theta}\right| E_{W^\theta}(F_\theta - F_n)^2 \leqq$$
$$\leqq \sup_{\mathscr{X}} \left|\frac{w'_\theta}{w_\theta}\right| (\sup_{\mathscr{X}} |F_\theta - F_n|)^2. \tag{5.9}$$

Moreover, by Corollary 5.2, $E_{F_\theta}\Omega_\theta=0$. Hence, if we prove

$$\lim_{n\to\infty} \sqrt{n}E_\lambda w_\theta'(F_\theta-F_n)^2=0 \qquad \text{in} \quad P_\theta^\infty\text{-probability} \tag{5.10}$$

then, by the Cramér–Slutskij theorem (Fuller [5], pp. 140–145), the same argument as used in the proof of Theorem 2.3 yields the desired result. But (5.10) follows from (5.9) and from the Kolmogorov–Smirnov theorem. Q.E.D.

The results of the present section considerably sharpen the exposition presented in [17, 18]. In the next section we concretize these results for the well-known location and scale model.

6. χ^2-estimators of location and scale

Let the model of Section 5 be specified as follows: $k=1$, $\theta=(\mu,\sigma)\in\mathscr{R}\times\times[\delta,\delta^{-1}]=\Theta$, where $\delta\in(0,1)$ is arbitrary fixed, $F_e(x)=\int\limits_{-\infty}^{x}(\exp(-x^2/2)/\sqrt{2\pi})dx$, $F_\theta(x)=F_e((x-\mu)/\sigma)$ for every $\theta=(\mu,\sigma)\in\Theta$ and let W be an arbitrary probability distribution on $\mathscr{X}=\mathscr{R}$ with unimodal and twice continuously differentiable density w symmetric about 0 such that $w(x)$, $w'(x)$, $xw(x)$, $w''(x)$, $x^2w''(x)$ are bounded on $\mathscr{X}$. Finally let, in addition to all $(\mu,\sigma)\in\Theta$, Θ^* contain the points $\theta^*=(-\infty,\sigma)$ and (∞,σ) for all $\sigma\in[\delta,\delta^{-1}]$. By (5.1), (1.5) holds with

$$D(-\infty,\sigma)=E_WF_e^2, \qquad D(\infty,\sigma)=E_W(1-F_e)^2 \qquad \text{for all} \quad \sigma\in[\delta,\delta^{-1}].$$

The validity of (1.4) for all $\theta^*\in\Theta^*-\Theta$ thus follows e.g. from what is stated in the proof of Lemma 2.1 of Boos [1]. Since $(\mu_*,\sigma_*)(\mu,\sigma)=(\mu_*+\sigma_*\mu,\sigma_*\sigma)$, (1.13) holds for Θ too. Therefore we may conclude that in the present situation all assumptions of Section 5 hold and

$$p(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \qquad t(x)=-\begin{bmatrix}1\\x\end{bmatrix} \quad \text{(cf. (5.2))}. \tag{6.1}$$

Thus all χ^2-estimators of location and scale with estimates

$$(M(F),S(F))\in\arg\min\nolimits_{\mathscr{R}\times[\delta,\delta^{-1}]}E_W(F_e-F[\mu,\sigma])^2 \qquad \text{for all} \quad F\in\mathscr{F}$$

possess the following properties: (1) existence (cf. Corollary (1.3), (2) equivariant versions (cf. Theorem 1.3), (3) strong consistency (Corollary 5.1), (4) influence curves of the form

$$\Omega_{\mu,\sigma}(x)=\begin{bmatrix}(\psi_1((x-\mu)/\sigma)-E_{F_e}\psi_1(x))/E_{F_e}\tilde{\psi}_1(x)\\(\psi_2((x-\mu)/\sigma)-E_{F_e}\psi_2(x))/E_{F_e}(x\tilde{\psi}_2(x))\end{bmatrix}, \tag{6.2}$$

where $\psi_1(x)$ or $\psi_2(x)$ are primitive to $\tilde{\psi}_1(x)=p(x)w(x)$ or $\tilde{\psi}_2(x)=xp(x)w(x)$ respectively, and (5) the asymptotic normality with the asymptotic mean zero and the asymptotic variance $E_{F_e}(\Omega_{\mu,\sigma}\Omega^T_{\mu,\sigma})$. Expression (6.2) follows from Corollary 5.2, from (5.7) and (6.1) and from the fact that

$$E_W(p^2tt^T)=E_W\begin{bmatrix} p(x)^2 & xp(x)^2 \\ xp(x)^2 & x^2p(x)^2 \end{bmatrix}=$$

$$=\begin{bmatrix} E_Wp(x)^2 & 0 \\ 0 & E_Wx^2p(x)^2 \end{bmatrix}=\begin{bmatrix} E_{F_e}\tilde{\psi}_1(x) & 0 \\ 0 & E_{F_e}x\tilde{\psi}_2(x) \end{bmatrix}$$

For $\sigma=1$ the results (1) and (3)-(5) agree with the former results of Millar [10] and Boos [1] who applied the present least square method to d.f.'s of the location model (for the location and scale see also Boos [2]).

References

1. *Boos, D. D.*, Minimum distance estimators for location and goodness of fit. JASA **76**, 663–670 (1981).
2. *Boos, D. D.*, Minimum Anderson–Darling estimation. A preprint (1981).
3. *Csiszár, I.*, Eine Informationstheoretische Ungleichung und Ihre Anwendung auf der Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci., Ser. A, **8**, 85–108 (1963).
4. *Cramér, H.*, Mathematical Methods of Statistics, Princeton Univ. Press (1946).
5. *Fuller, W.*, Introduction to Statistical Time Series, New York, J. Wiley (1976).
6. *Gaenssler, P.*, *Stute, W.*, Empirical Distributions and Processes. In Lecture notes *566*, Springer, Berlin (1976).
7. *Hampel, F. R.*, The influence curve and its role in robust estimation. JASA **69**, 383–393 (1974).
8. *Kelley, J. L.*, General Topology, Princeton, Van Nostrand (1957).
9. *Loève, M.*, Probability Theory, Princeton, Van Nostrand (1960).
10. *Millar, P. W.*, Robust estimation via minimum distance methods. Zeitschr. Wahrsch. **55**, 73–89 (1981).
11. *Neyman, J.*, Contributions to the theory of χ^2-test. Proc. 1st Berkeley Symp. on Math. Statist., ..., 239–273, Berkeley, Univ. of Calif. Press (1949).
12. *Perez, A.*, Information-theoretic risk estimates in statistical decisions. Kybernetika **3**, 1–21 (1967).
13. *Pfanzagl, J.*, On the measurability and consistency of minimum contrast estimates. Metrika **14**, 249–272 (1969).
14. *Rao, C. R.*, Asymptotic efficiency and limiting information. Proc. 4th Berkeley Symp. on Math. Statist., ..., Vol. **1**, 531–546, Berkeley, Univ. of Calif. Press (1961).
15. *Vajda, I.*, Minimum divergence principle in statistical estimation. Recent Results in Estimation Theory (Ed. P. Sen, J. Dudewicz, D. Plachky), München, Oldenburg (1984).
16. *Vajda, I.*, Motivation, existence and equivariance of *D*-estimators. Kybernetika **20**, 189–208 (1984).
17. *Vajda, I.*, Asymptotic efficiency and robustness of D-estimators. Kybernetika **20**, 358–375 (1984).
18. *Vajda, I.*, Minimum weak divergence estimators of structural parameters. The 3rd Prague Symp. on Asympt. Statist. (Ed. P. Mandl and M. Hušková), 417–423, Amsterdam, Elsevier (1984).
19. *Vajda, I.*, Minimum chi-square estimates of multistate communication channel models. Problems of Control and Inf. Th. **13**, 343–356 (1984).

Робастные оценки в дискретных и непрерывных семействах методом минимального хи-квадрат

И. ВАЙДА

(Прага)

Уточнение предшествующих результатов относительно общих оценок с минимальным расхождением приводится в начале статьи. Потом исследуется класс оценок с минимальным хи-квадрат расхождением, которые соответствуют L_2-проекциям обобщенных эмпирических ф. р. в семействах обобщенных теоретических ф. р. Для векторных параметров вычислены кривая влияния и асимптотическое среднее значение и ковариационная матрица. Рассмотрены примеры с параметрами однородной цепи Маркова и параметрами сдвига и масштаба. Примеры относительно параметров каналов передачи информации рассматривались в этом журнале ранее.

I. Vajda
Institute of Information Theory and Automation,
Czechoslovak Academy of Sciences,
Pod vodárenskou věží 4, 18208 Prague, Czechoslovakia

Problems of Control and Information Theory, Vol. 15 (2), pp. 129—147 (1986)

EXPURGATED ERROR BOUNDS FOR CONCATENATED CODES

G. DUECK, V. MÖLLER

(Bielefeld)

(Received February 1, 1985)

Achievable error exponents for concatenated codes are studied. Bounds given by Forney [1] are improved in two aspects. First we can sharpen equierror superchannel arguments by proving a convexity property of the expurgated bound. Then it is shown that a direct analysis of superchannels of good codes gives even better results.

I. Introduction

In [1] Forney analyzed the theoretical performance of concatenated codes. He proved that the random coding technique he used performs worst if the error probability matrix of the inner code forms a so-called equierror superchannel. Since this symmetric superchannel is mathematically easy to handle this argument results in rather immediate upper bounds on the probability of eı ror for concatenated codes. In this paper we show that the same reasoning can be used also in connection with expurgated bounding techniques. This follows readily from our Convexity Lemma where we demonstrate that the familiar expurgated bound function for the discrete memoryless channel is a convex function in the channel. For low rates we can achieve this way considerably better bounds than the ones obtained just by random coding. In the other part of this paper we follow a different concept. We do not try to get bounds by the usual equierror superchannel arguments. We start in a new approach with an error probability matrix of an appropriate inner code which we choose to meet the conditions stated in the Packing Lemma of [2, see also 3]. A careful evaluation of the expurgated bound for this particular superchannel gives us new (and stronger) bounds for the error exponent for concatenated codes.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

II. Formal statement of the problem and main results

We shall consider concatenated codes for the discrete memoryless channel (DMC).

A DMC is given by a finite input alphabet $\mathscr{X}$, a finite output alphabet $\mathscr{Y}$, and a set $\{W(y|x)|x \in \mathscr{X}, y \in \mathscr{Y}\}$ of transition probabilities. For n-words $x^n = (x_1, \ldots, x_n) \in \mathscr{X}^n$, $y^n = (y_1, \ldots, y_n) \in \mathscr{Y}^n$ the transition probability is defined by

$$W^n(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i).$$

An n-block code $\mathscr{C}$ for the DMC consists of a set $\{u_1, \ldots, u_N\} \subset \mathscr{X}^n$ of code words and of a decoder

$$\varphi : \mathscr{Y}^n \to \mathscr{M},$$

where $\{u_1, \ldots, u_N\} \subset \mathscr{M}$. The average error probability of such a code is defined by

$$P_e(\mathscr{C}) = \frac{1}{N} \sum_{i=1}^{N} W^n(\varphi^{-1}(u_i)^c | u_i).$$

Now let an n-block code $\mathscr{C} = (u_1, \ldots, u_N, \varphi)$ and a positive integer k be given. An $n \cdot k$-block code for the DMC W in which the code words have the structure

$$u_{i_1} \cdot u_{i_2} \ldots u_{i_k} \qquad (\cdot \text{ means concatenation})$$

is called a *concatenated code* built from the *inner code* $\mathscr{C}$. Here, the problem is to build large codes (codes with a large block length) from smaller ones.

What is the proper choice of an inner code and what is the best way to form a so-called *outer code* (i.e. the best way to define a concatenation device)? A systematic analysis in this direction was given by Forney in [1]. A key argument in his approach led to the notion of an *equierror superchannel.* To explain this look for a given inner code $\mathscr{C} = (u_1, \ldots, u_N, \varphi)$ with $\varphi : \mathscr{Y}^n \to \{u_1, \ldots, u_N\}$ at the $N \times N$ matrix formed by the entries

$$W^n(\varphi^{-1}(u_j)|u_i)$$

for $i = 1, \ldots, N$; $j = 1, \ldots, N$. This matrix is called the *superchannel* transition probability matrix defined by the code $\mathscr{C}$.

Compare this matrix with

$$\begin{pmatrix} p & q & q & q & \cdots \\ q & p & & & \\ q & & p & & \\ & & & & p \end{pmatrix},$$

where

$$p=\frac{1}{N}\sum_{i=1}^{N} W^n(\varphi^{-1}(u_i)|u_i)$$

and

$$q=\frac{1}{N^2-N}\sum_{\substack{i=1\\j=1\\i\neq j}}^{N} W^n(\varphi^{-1}(u_j)|u_i).$$

Observe that this matrix is a symmetrized map of the former one. It is called the *equierror superchannel* matrix connected with $\mathscr{C}$. Suppose there exists a code $\mathscr{C}'$, whose superchannel equals the equierror superchannel given by $\mathscr{C}$. Then one might think that the problem to find good outer codes using $\mathscr{C}$ is easier than to find those using $\mathscr{C}'$. The reason is that the capacity of the superchannel connected with $\mathscr{C}$ is larger than the one of the equierror superchannel. However, there does not exist a general result of this type. On the other hand, Forney [1] could show that the random coding bound of the equierror superchannel is less than, or equal to, the random coding bound of the original superchannel. This result yields random coding exponents for concatenated codes. We state the upper bound of Forney in the following.

Let $\mathscr{C}=(u_1, \ldots, u_N, \varphi)$ be an inner n-block code. $R=\frac{1}{n}\log N$ is called the rate of $\mathscr{C}$. From this inner code we can build an outer k-block code such that the total number of code words of block length $n\cdot k$ is given by $\exp\{n\cdot k\cdot r\cdot R\}$, $r\in(0,1]$. r determines the rate of the outer code. $R_0=r\cdot R$ is the total rate of the so-called concatenated code.

Define $E_c(R_0)$ to be the reliability function for concatenated codes.

Theorem (Forney [1]). For every $0<R_0<C$, C being the capacity of W

$$E_c(R_0)\geqq \max_{r\cdot R=R_0}(1-r)\frac{E_x(R)+\min\{E_x(R),R\}}{2}$$

and

$$E_c(R_0)\geqq \max_{r\cdot R=R_0}(1-r)\frac{E_r(R)+\min\{E_r(R),R\}}{2}$$

where the maximal range over $r\in(0,1]$ and $0<R<C$.

In the Theorem $E_r(R)$ denotes the *random coding bound* and $E_x(R)$ the *expurgated bound* of W. We recall the exact definitions below.

In this paper we proceed in two different ways. In a first approach we prove that the so-called *erasure equierror superchannels* are worst in the expurgated bound sense. This is done in Section III. We state the final result here.

For a distribution P on $\mathscr{X}$ and $\lambda\geqq 0$ we define

$$E_{r\lambda}(R,P,W):=\min_V(D(V||W|P)+\lambda|I(P,V)-R|^+).$$

The min is taken over all DMC's V with input $\mathscr{X}$ and output $\mathscr{Y}$. $D(\cdot)$ denotes the Kullback–Leibler information or I-divergence. $I(P, V)$ stands for mutual information. $|t|^{+} = \max\{0, t\}$ for real numbers t. (This notation is adopted from [3]). We set

$$E_r(R, P, W) = E_{r1}(R, P, W).$$

$$E_r(R) = \max_{P} E_r(R, P, W)$$

is the well-known random coding bound.

Theorem 1. For every $0 < R_0 < C$

$$E_c(R_0) \geqq \max_{r \cdot R = R_0} \max_{P} G(r, R, P), \qquad \text{where} \qquad 0 < r \leqq 1 \qquad \text{and}$$

$$G(r, R, P) = \begin{cases} -rR + E_r(R, P, W), & \text{if} \quad R \geqq E_r(R, P, W) \\ (1-r) \displaystyle\max_{\substack{\lambda \geqq 1 \\ R \leqq \tilde{R} \leqq I(P, W)}} \min\left\{\frac{1}{2}(E_{r\lambda}(R, P, W) + \tilde{R}), \quad E_{r1/\lambda}(\tilde{R}, P, W)\right\} \\ \text{else.} \end{cases}$$

Corollary 1. If $R \leqq E_r(R, P, W)$, then

$$G(r, R, P) \geqq (1-r)\frac{2E_r(R, P, W) + R}{3}$$

In Section IV we leave the equierror superchannel approach. We start with results of Csiszár, Körner and Marton ([2], see [3]). They showed that code word sets in n-space are good codes if they satisfy certain distance properties. If we choose such a code as an inner code then we have some detailed knowledge about the structure of the superchannel matrix given by this code. It turns out that this knowledge presented by those distance properties enables us to derive an expurgated-type bound on the function $E_c(R_0)$.

For $0 \leqq s \leqq 1$ and $x, x' \in \mathscr{X}$ we define

$$d_{W,s}(x, x') = -\log \sum_{y \in \mathscr{Y}} W(y|x)^s \cdot W(y|x')^{1-s}.$$

If X, X' is a pair of random variables on $\mathscr{X}$ with a joint distribution, then we write $\mathbf{E}d_{W,s}(X, X')$ for the expected value of $d_{W,s}$. For $\rho > 0$ set

$$E^{s}_{x,\rho}(R, P, W) = \min_{\substack{P_X = P_{X'} = P \\ I(X \wedge X') \leqq R}} \frac{1}{2\rho} \mathbf{E}d_{W,s}(X, X') + I(X \wedge X') - R,$$

where $I(\cdot)$ denotes the mutual information. For $\rho=1/2$ we write $E^s_x(R, P, W)$ instead of $E^s_{x,1/2}(R, P, W)$. For $s=1/2$ we write $E_x(R, P, W)$ for $E^{1/2}_x(R, P, W)$.

$$E_x(R) := \max_P E_x(R, P, W)$$

is the expurgated bound (compare [3]).

Theorem 2. $E_c(R_0) \geqq \max_{r\cdot R=R_0} \max_P F(r, R, P)$, where $0<R_0\leqq C$, $0<r\leqq 1$ and

$$F(r, R, P) = \max_{\rho\geqq 1, K\geqq 0} \Bigg\{(1-r)\rho R - \Bigg| \max \Bigg\{\rho R - \max_{0\leqq s\leqq 1} (E^s_x(R, P, W) - sk),$$
$$-\rho \max_{0\leqq s\leqq 1} \left(E^s_{x,\rho}(R, P, W) + \frac{sk}{2\rho}\right)\Bigg\}\Bigg|^+\Bigg\}.$$

We shall further investigate the function $F(r, R, P)$ in Section IV. Especially we shall prove:

Corollary 2. If $R\leqq E_x(R, P, W)$, then

$$F(r, R, P) \geqq (1-r)\frac{2E_x(R, P, W)+R}{3}.$$

Corollaries 1 and 2 allow a direct comparison with Forney's bounds. However, in our paper we allow also erasure decoding. Therefore one might think that our improvement results from the fact that we consider a larger class of decoding procedures. We show in Section V that our bounds are even better than Forney's if we restrict our attention only to procedures without erasure decoding.

III. Convexity of the expurgated bound and proof of Theorem 1

Convexity Lemma. Let $W_1, W_2 : \mathscr{X} \to \mathscr{Y}$ two DMC's. Then for every P on $\mathscr{X}$, $R\geqq 0$, $s\in(0, 1)$, $\alpha\in[0, 1]$:

$$\alpha E^s_x(R, P, W_1) + (1-\alpha)E^s_x(R, P, W_2) \geqq E^s_x(R, P, W_0)$$

where $W_0 = \alpha W_1 + (1-\alpha)W_2$.

Proof. Let (X_1, X_1') and (X_2, X_2') be pairs of random variables with joint distributions $P_{X_1X_1'}$, $P_{X_2X_2'}$ which yield the min in the definition of $E^s_x(R, P, W_1)$ resp. $E^s_x(R, P, W_2)$. Then

$$P_{X_1} = P_{X_1'} = P_{X_2} = P_{X_2'} = P$$

and

$$I(X_i \wedge X_i') = 2H(P) - H(X_i, X_i') \leqq R \qquad \text{for} \quad i=1, 2.$$

Set $W_0 = \alpha W_1 + (1-\alpha) W_2$ and let (X_0, X'_0) be a pair with joint distribution $P_{X_0 X'_0} = = \alpha P_{X_1 X'_1} + (1-\alpha) P_{X_2 X'_2}$.

We shall write $\Pr(x_1, x'_1)$ instead of $\Pr(X_1 = x_1, X'_1 = x'_1)$ etc. We derive quickly a variant of Hölder's inequality:

Let $a_1, a_2, b_1, b_2 \geqq 0$ be given. Recall that for $c_1, c_2, d_1, d_2 \geqq 0$, $p, q > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ by Hölder's inequality:

$$c_1 d_1 + c_2 d_2 \leqq (c_1^p + c_2^p)^{1/p} (d_1^q + d_2^q)^{1/q},$$

and consequently, by setting $s = 1/p$, $c_1 = a_1^s$, $c_2 = a_2^s$, $d_1 = b_1^{1-s}$, $d_2 = b_2^{1-s}$

$$a_1^s b_1^{1-s} + a_2^s b_2^{1-s} \leqq (a_1 + a_2)^s (b_1 + b_2)^{1-s}.$$

This estimate gives us for $x, x' \in \mathscr{X}$, setting $\alpha_1 = \alpha$, $\alpha_2 = 1 - \alpha$:

$$\sum_{i=1}^{2} \alpha_i \sum_{y \in \mathscr{Y}} W_i(y|x)^s W_i(y|x')^{1-s}$$

$$\sum_{i=1}^{2} \sum_{y \in \mathscr{Y}} (\alpha_i W_i(y|x))^s (\alpha_i W_i(y|x'))^{1-s} \leqq \sum_{y \in \mathscr{Y}} W_0(y|x)^s W_0(y|x')^{1-s}.$$

We abbreviate: $g_{W,s}(x, x') = \sum_{y \in \mathscr{Y}} W(y|x)^s \cdot W(y|x')^{1-s}$. Using the log-sum inequality [3]:

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geqq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}, \quad a_i, b_i \geqq 0,$$

we get straightforward

$$\sum_{i=1}^{2} \alpha_i E_x^s(R, P, W_i) =$$

$$= 2H(P) - R + \sum_{i=1}^{2} \sum_{x_i, x'_i} \alpha_i \Pr(x_i, x'_i) \log \frac{\alpha_i \Pr(x_i, x'_i)}{\alpha_i g_{W_i, s}(x_i, x'_i)} \geqq$$

$$\geqq 2H(P) - R + \sum_{x_0, x'_0} \Pr(x_0, x'_0) \log \frac{\Pr(x_0, x'_0)}{g_{W_0, s}(x_0, x'_0)} =$$

$$= 2H(P) - R - H(X_0, X'_0) + \mathbf{E} d_{W_0, s}(X_0, X'_0) =$$

$$= I(X_0 \wedge X'_0) - R + \mathbf{E} d_{W_0, s}(X_0, X'_0) \geqq E_x^s(R, P, W_0),$$

where the last inequality is justified by the concavity argument

$$I(X_0 \wedge X'_0) = 2H(P) - H(X_0, X'_0) \leqq 2H(P) - \sum_{i=1}^{2} \alpha_i H(X_i, X'_i) =$$

$$= \sum_{i=1}^{2} \alpha_i I(X_i \wedge X'_i) \leqq R.$$

Consider now a DMC $W:\mathscr{X}\to\mathscr{X}\cup\{E\}$ with $|\mathscr{X}|\geqq 2$. $E\notin\mathscr{X}$ is called an erasure symbol.

We associate a DMC $\tilde{W}$ to W. Let

$$p:=1-\frac{1}{|\mathscr{X}|}\sum_{x\in\mathscr{X}}W(x|x)$$

and

$$p_E:=\frac{1}{|\mathscr{X}|}\sum_{x\in\mathscr{X}}W(E|x).$$

Now set for $x\in\mathscr{X}$, $x'\in\mathscr{X}$, $x\neq x'$:

$$\tilde{W}(x'|x):=\frac{1}{|\mathscr{X}|-1}(p-p_E)$$

and

$$\tilde{W}(x|x):=1-p,$$

and

$$\tilde{W}(E|x):=p_E.$$

We call $\tilde{W}$ erasure equierror superchannel. We show in the next Lemma that indeed $\tilde{W}$ is worse than W in the expurgated bounded sense.

Worst Channel Lemma. $E_x(R, W)\geqq E_x(R, \tilde{W})$ for every R, where, for instance,

$$E_x(R, W)=\max_P E_x(R, P, W).$$

Proof. Let P_{equ} be the equidistribution on $\mathscr{X}$. Since $\tilde{W}$ is an equidistant channel, we have that

$$E_x(R, \tilde{W})=E_x(R, P_{\text{equ}}, \tilde{W})$$

(cf. [3], page 194). Of course we can estimate

$$E_x(R, W)\geqq E_x(R, P_{\text{equ}}, W).$$

Now we need a small permutation trick which was useful also in [1]. Let π be a permutation on $\{1, \dots, |\mathscr{X}|\}$. Define a new channel W^π by

$$W^\pi(x'|x)=W(\pi(x')|\pi(x)), \qquad x, x'\in\mathscr{X}$$

and

$$W^\pi(E|x)=W(E|\pi(x)); \qquad x\in\mathscr{X}.$$

Tracing back to the definitions it is easy to see that $E_x(R, P_{\text{equ}}, W)=$ $=E_x(R, P_{\text{equ}}, W^\pi)$ for every permutation π. Also it is a routine matter to show that

$$\tilde{W}=\frac{1}{|\mathscr{X}|!}\sum_\pi W^\pi.$$

This gives us the estimate:

$$\begin{aligned} E_x(R, W) &\geqq E_x(R, P_{\text{equ}}, W) = \\ &= \frac{1}{|\mathcal{X}|!}\sum_{\pi} E_x(R, P_{\text{equ}}, W^{\pi}) \geqq \\ &\geqq E_x\Big(R, P_{\text{equ}}, \frac{1}{|\mathcal{X}|!}\sum_{\pi} W^{\pi}\Big) = \\ &= E_x(R, P_{\text{equ}}, \tilde{W}) = E_x(R, \tilde{W}), \end{aligned}$$

where we used the Convexity Lemma in the key step.

We start the proof of Theorem 1 by the definition of a suited inner code.

We consider n-block codes $\mathcal{C}$ with decoders

$$\varphi : \mathcal{Y}^n \to \{\text{set of code words}\} \cup \{x_E\},$$

where $\varphi(y^n) = x_E$ means the "erasure". If $\{u_1, \ldots, u_N\}$ is the code word set, then we call

$$W^n(\varphi^{-1}(x_E)|u_i)$$

the probability of a *detected* error given that u_i is sent. Analogously

$$W^n\Big(\bigcup_{j\neq i} \varphi^{-1}(u_j)|u_i\Big)$$

is called the probability of an *undetected* error.

Theorem 5.11 [3] gives achievable bounds for these error probabilities. We recall this result here in a somewhat weaker form:

Theorem. For every $\tilde{R} \geqq R > 0$, $\lambda \geqq 1$, $\delta > 0$ and every distribution P on $\mathcal{X}$ the following is true for n large enough:

There exists an n-block code $\mathcal{C}$ consisting of a set $\{u_1, \ldots, u_N\}$ of code words and a decoder

$$\varphi : \mathcal{Y}^n \to \{u_1, \ldots, u_N\} \cup \{x_E\},$$

x_E not a code word such that $\frac{1}{n}\log N \geqq R - \delta$ and

$$W^n\Big(\bigcup_{j\neq i} \varphi^{-1}(u_j)|u_i\Big) \leqq \exp\{-n(E_{r\lambda}(R, P, W) + \tilde{R} - R - \delta)\}$$

$$W^n(\varphi^{-1}(x_E)|u_i) \leqq \exp\{-n(E_{r1/\lambda}(\tilde{R}, P, W) - \delta)\}$$

for every $i \in \{1, \ldots, N\}$.

We fix now R, $\tilde{R} \geqq R$, $\lambda \geqq 1$, $\delta > 0$, P.

For an n large enough we choose a code $\mathscr{C}$ with code word set $\{u_1, \ldots, u_N\} \subset \mathscr{X}^n$ and decoder φ satisfying the claim of the Theorem.

Let $W^n(R, \tilde{R}, P, \lambda, \delta)$ be the superchannel $\{u_1, \ldots, u_N\} \to \{u_1, \ldots, u_N\} \cup \{x_E\}$ connected with $\mathscr{C}$ and let $\tilde{W}^n(R, \tilde{R}, P, \lambda, \delta)$ be the associated erasure equierror superchannel.

Then we can conclude from the Worst Channel Lemma that for any $r \in (0, 1]$

$$\frac{1}{n} E_x(r \cdot n(R-\delta), \tilde{W}^n(R, \tilde{R}, P, \lambda, \delta))$$

is an achievable error exponent for concatenated codes with rate $r \cdot (R-\delta)$. We estimate this exponent in the following.

Since $\tilde{W}^n(\ldots)$ is equidistant, we have to consider again the equidistribution P_{equ}, this time on $\{u_1, \ldots, u_N\}$.

Thus, setting

$$g_{\tilde{W}}(u_i, u_j) = \sum_{u \in \{u_1, \ldots, u_N\} \cup \{x_E\}} \sqrt{\tilde{W}^n(u|u_i)\tilde{W}^n(u|u_j)}$$

(we wrote $\tilde{W}^n$ instead of $\tilde{W}^n(R, \tilde{R}, P, \lambda, \delta)$ for convenience) we get the following estimate:

$$E_x(r \cdot n(R-\delta), \tilde{W}^n(R, \tilde{R}, P, \lambda, \delta)) =$$

$$= \max_{\rho \geqq 1} \left\{ -r \cdot n \cdot \rho(R-\delta) - \rho \log \sum_{u_i, u_j} P_{\text{equ}}(u_i) \cdot P_{\text{equ}}(u_j) \cdot g_{\tilde{W}}(u_i, u_j)^{1/\rho} \right\} \geqq$$

$$\geqq \max_{\rho \geqq 1} \left\{ (1-r)n\rho(R-\delta) - \rho \log \sum_{u_i} P_{\text{equ}}(u_i) \sum_{u_j} g_{\tilde{W}}(u_i, u_j)^{1/\rho} \right\} =$$

$$= \max_{\rho \geqq 1} \left\{ (1-r)n\rho(R-\delta) - \rho \log \left(1 + \sum_{j \neq 1} g_{\tilde{W}}(u_1, u_j)^{1/\rho} \right) \right\}$$

where we used in the last step the symmetry of $\tilde{W}^n$. We have to upper bound $g_{\tilde{W}}(\cdot)$. Use the definition of an associated superchannel $\tilde{W}^n$ to obtain for $u_j \neq u_1$:

$$g_{\tilde{W}}(u_1, u_j) \leqq \sqrt{\tilde{W}^n(\varphi^{-1}(u_1)|u_j)} + \sqrt{\tilde{W}^n(\varphi^{-1}(u_j)|u_1)} +$$

$$+ \sum_{\substack{u \in \{u_1, \ldots, u_N\} \\ u \notin \{u_1, u_j\}}} \sqrt{\tilde{W}^n(\varphi^{-1}(u)|u_1) \cdot \tilde{W}^n(\varphi^{-1}(u)|u_j)} +$$

$$+ \sqrt{\tilde{W}^n(\varphi^{-1}(x_E)|u_1) \cdot \tilde{W}^n(\varphi^{-1}(x_E)|u_j)} \leqq$$

$$\leqq 2\sqrt{\tilde{W}^n(\varphi^{-1}(u_j)|u_1)} + \tilde{W}^n\left(\bigcup_{j \neq 1} \varphi^{-1} u_j)|u_1 \right) +$$

$$+ \tilde{W}^n(\varphi^{-1}(x_E)|u_1).$$

For the first summand we use again the definition of $\tilde{W}^n$, the other two summands are the error probabilities being bounded by our assumption on the code we started with. Hence especially

$$2\sqrt{\tilde{W}^n(\varphi^{-1}(u_j)|u_1)} \leqq 2[\exp\{n(R-\delta)\}-1]^{-1/2}\,\tilde{W}^n\Big(\bigcup_{j\neq 1}\varphi^{-1}(u_j)|u_1\Big).$$

Straightforward analysis leads now to the result that the function

$$G'(r,R,P)=\max_{\substack{\lambda,\rho\geqq 1\\ R\leq\tilde{R}\leq I(P,W)}}\Big\{(1-r)\rho R-$$
$$-\Big|\rho R-\min\Big\{\frac{1}{2}(E_{r\lambda}(R,P,W)+\tilde{R}),\,E_{r1/\lambda}(\tilde{R},P,W)\Big\}\Big|^{+}\Big\}$$

is an achievable error exponent for concatenated codes with inner code rate R and outer code rate r.

If we choose in the definition of $G'(r,R,P)$ particularly $\lambda=1$, $\rho=1$, $\tilde{R}=R$ then we can conclude that

$$G'(r,R,P)\geqq -rR+E_r(R,P,W)$$

if $R\geqq E_r(R,P,W)$. On the other hand,

$$G'(r,R,P)\leqq \max_{\substack{\rho,\lambda\geqq 1\\ R\leq\tilde{R}\leq I(P,W)}}\{(1-r)\rho R-|\rho R-E_{r1/\lambda}(\tilde{R},P,W)|^{+}\}=$$
$$=\max_{\rho\geqq 1}\{(1-r)\rho R-|\rho R-E_r(R,P,W)|^{+}\}$$

because of the monotonicity properties of $E_{r1/\lambda}(\tilde{R},P,W)$ in $\tilde{R}$ and λ. Thus, if $R\geqq E_r(R,P,W)$,

$$G'(r,R,P)\leqq -rR+E_r(R,P,W)$$

and, using the opposite inequality above, the equality holds. Further, by observing $|t|^{+}\geqq t$ and $(1-r)\leqq 1$ we get easily

$$G'(r,R,P)\leqq(1-r)\max_{\substack{\lambda\geqq 1\\ R\leq\tilde{R}\leq I(P,W)}}\min\Big\{\frac{1}{2}(E_{r\lambda}(R,P,W)+\tilde{R}),\quad E_{r1/\lambda}(\tilde{R},P,W)\Big\}.$$

Let now $\bar{R}$ be the rate satisfying $\bar{R}=E_r(\bar{R},P,W)$. Assume $R\leqq E_r(R,P,W)$ and hence $R\leqq\bar{R}$. Then also (choose $\lambda=1$, $\tilde{R}=\bar{R}$)

$$\max_{\substack{\lambda\geqq 1\\ R\leq\tilde{R}\leq I(P,W)}}\min\Big\{\frac{1}{2}(E_{r\lambda}(R,P,W)+\tilde{R}),\quad E_{r1/\lambda}(\tilde{R},P,W)\Big\}\geqq$$
$$\geqq E_r(\bar{R},P,W)=\bar{R}\geqq R\,.$$

Thus we can choose for $R \leqq E_r(R, P, W)$ a $\rho \geqq 1$ such that ρR equals the left hand side of the last inequality. For this choice we see that

$$G'(r, R, P) \geqq (1-r) \max_{\substack{\lambda \geqq 1 \\ R \leqq \tilde{R} \leqq I(P,W)}} \min \left\{ \frac{1}{2}(E_{r\lambda}(R, P, W) + \tilde{R}), \quad E_{r1/\lambda}(\tilde{R}, P, W) \right\}.$$

Note now that we have shown that G' equals the function G defined in Theorem 1.

Proof of Corollary 1. Let $R \leqq E_r(R, P, W)$. In the max min expression defining $G(r, R, P)$ we look at the case $\lambda = 1$. We shall see that it is possible to choose $\tilde{R}$ such that

$$\frac{1}{2}(E_r(R, P, W) + \tilde{R}) = E_r(\tilde{R}, P, W).$$

$\frac{1}{2}(E_r(R, P, W) + \tilde{R})$ is strictly increasing with $\tilde{R}$ and equals $\frac{1}{2} E_r(R, P, W)$ for $\tilde{R} = 0$. $E_r(\tilde{R}, P, W)$ is strictly decreasing in $\tilde{R} \in [0, I(P, W)]$ and $\frac{1}{2} E_r(R, P, W) < E_r(0, P, W)$.

Hence, there is a unique $\tilde{R}$ satisfying the above equality. Moreover, $\tilde{R} \geq R$ for this choice because if $\tilde{R} < R$ then the equality yields

$$\tilde{R} = 2E_r(\tilde{R}, P, W) - E_r(R, P, W) > E_r(\tilde{R}, P, W).$$

However, this and the assumption $R \leqq E_r(R, P, W)$ would imply $\tilde{R} > R$, a contradiction.

Since $E_r(\cdot)$ as a function of R has slope at least -1 we have that

$$E_r(R, P, W) - E_r(\tilde{R}, P, W) \leqq \tilde{R} - R.$$

Together with $\tilde{R} = 2E_r(\tilde{R}, P, W) - E_r(R, P, W)$ this gives

$$E_r(\tilde{R}, P, W) \geqq 1/3\,(2E_r(R, P, W) - R).$$

Observe that this proves Corollary 1.

IV. Proof of Theorem 2

For n-sequences $x^n = (x_1, \ldots, x_n)$, $\bar{x}^n = (\bar{x}_1, \ldots, \bar{x}_n) \in \mathscr{X}^n$ we call the empirical distribution on $\mathscr{X} \times \mathscr{X}$ defined by the pairs $(x_i, \bar{x}_i)$ the (joint) type $P_{x^n\bar{x}^n}$ of $(x^n, \bar{x}^n)$.

P_{x^n} and $P_{\bar{x}^n}$ denote the empirical distributions induced by the components of x^n and $\bar{x}^n$, respectively. A distribution P on $\mathscr{X}$ is called n-type if $P = P_{x^n}$ for a $x^n \in \mathscr{X}^n$.

Lemma 5.1 of [3] asserts now:

For every $R>0$, $\delta>0$ the following holds for large n: For every n-type P there exists a system $\{u_1, \ldots, u_N\} \subset \mathscr{X}^n$ of n-sequences having type P such that

$$\exp\{nR\} \geqq N \geqq \exp\{n(R-\delta)\}$$

and such that for every pair (X, X') of random variables on $\mathscr{X} \times \mathscr{X}$ we have: For any $i \in \{1, \ldots, N\}$ there are at most $\lfloor \exp\{n(R-I(X \wedge X'))\}\rfloor$ indices $j \in \{1, \ldots, N\}$ such that $P_{u_i u_j}$ is the distribution of (X, X'). (Note that the latter is possible only if $P_{u_j} = P_{u_i} = P$).

One can show (see [3], pp. 185–186) that using the maximum likelihood decoder φ such a system forms a code which has a maximal error probability bounded by $\exp\{-n(E_x(R, P, W)-\delta)\}$, n large. A key estimate in this proof is

$$W^n(\varphi^{-1}(u_j)\,|\,u_i) \leqq \sum_{y^n : W^n(y^n|u_j) \geqq W^n(y^n|u_i) \text{ for some } j} W^n(y^n\,|\,u_i)$$

$$\leqq \sum_{y^n} \sqrt{W^n(y^n|u_j)\, W^n(y^n|u_i)}\,.$$

In fact in the last step we could estimate also

$$\leqq \sum_{y^n} W^n(y^n|u_j)^s W^n(y^n|u_i)^{1-s}$$

for every $s \in [0, 1]$. Actually in the DMC coding theorem $s = 1/2$ is best. Here, we get sharper bounds using a general s. If one observes this difference and if one additionally uses an erasure maximum likelihood decoder the same proof gives the following result.

Theorem. For every $R>0$, $\delta>0$ and $K \geqq 0$ the following is true for sufficiently large n:

For every n-type P satisfying $H(P)>R$ there is an n-block code $\mathscr{C}$ given by a set $\{u_1, \ldots, u_N\}$ of code words having type P and a decoder $\varphi: \mathscr{Y}^n \to \{u_1, \ldots, u_N\} \cup \{x_E\}$, $x_E \notin \{u_1, \ldots, u_N\}$ defined by

$$\varphi(y^n) = \begin{cases} u_i & \text{if for all } \ j \in \{1, \ldots, N\} \\ & W^n(y^n|u_i) \cdot 2^{-nK} > W^n(y^n|u_j) \\ x_E & \text{else} \end{cases}$$

satisfying the following conditions:

(I) $\exp\{nR\} \geqq N \geqq \exp\{n(R-\delta)\}$.

(II) For every $i \in \{1, \ldots, N\}$ and every pair of random variables (X, X') on $\mathscr{X} \times \mathscr{X}$ there are at most $\lfloor \exp\{n(R-I(X \wedge X'))\}\rfloor$ indices $j \in \{1, \ldots, N\}$ such that $P_{u_i u_j}$ is the distribution of (X, X').

(III) For every pair $i, j \in \{1, \ldots, N\}$, $i \neq j$,

$$W^n(\varphi^{-1}(u_j)|u_i) \leqq \exp\left\{-n\left(\max_{0 \leqq s \leqq 1} \mathbf{E} d_{W,s}(X, X') + sK\right)\right\},$$

where (X, X') has distribution $P_{u_i u_j}$.

(IV) For every $i \in \{1, \ldots, N\}$,

$$W^n(\varphi^{-1}(x_E)|u_i) \leqq \exp\left\{-n\left(\max_{0 \leqq s \leqq 1} E_x^s(R, P, W) - sK - \delta\right)\right\}.$$

(V) For every $i \in \{1, \ldots, N\}$,

$$W^n\left(\bigcup_{j \neq i} \varphi^{-1}(u_j)|u_i\right) \leqq \exp\left\{-n\left(\max_{0 \leqq s \leqq 1} E_x^s(R, P, W) + sK - \delta\right)\right\}.$$

For given $R > 0$, $\delta > 0$, $K \geqq 0$, n large enough, and for an n-type P we choose now an inner code $\mathscr{C}$ as in the Theorem. We analyze this code to prove Theorem 2/Corollary 2.

Let $W_{\mathscr{C}}^n$ be the superchannel given by the probabilities $W^n(u|u_i)$, $i = 1, \ldots, N$, $u \in \{u_1, \ldots, u_N, x_E\}$.

Then $\frac{1}{n} E_x(r \cdot n(R-\delta), W_{\mathscr{C}}^n)$ is an achievable error exponent for $(r, R-\delta)$-rate concatenated codes. We estimate this exponent in the following. We lower bound first by assuming equidistribution on the code words.

$$E_x(r \cdot n(R-\delta), W_{\mathscr{C}}^n) \geqq \max_{\rho \geqq 1}\left\{-\rho r n(R-\delta) - \rho \log \frac{1}{N}\frac{1}{N}\sum_{i,j} g_{W_{\mathscr{C}}^n}(u_i, u_j)^{1/\rho}\right\} \geqq$$

$$\geqq \max_{\rho \geqq 1}\left\{(1-r)n\rho(R-\delta) - \rho \log \frac{1}{N}\sum_{i,j} g_{W_{\mathscr{C}}^n}(u_i, u_j)^{1/\rho}\right\}.$$

As in the preceding section we estimate

$$g_{W_{\mathscr{C}}^n}(u_i, u_j) \leqq \sqrt{W^n(\varphi^{-1}(u_j)|u_i)} + \sqrt{W^n(\varphi^{-1}(u_i)|u_j)} +$$

$$+ \sum_{\substack{u \in \{u_1, \ldots, u_N\} \\ u \notin \{u_i, u_j\}}} \sqrt{W^n(\varphi^{-1}(u)|u_i)\, W^n(\varphi^{-1}(u)|u_j)} +$$

$$+ \sqrt{W^n(\varphi^{-1}(x_E)|u_i)\, W^n(\varphi^{-1}(x_E)|u_j)}\,.$$

By the particular choice of our code (see (IV) and (V) in the Theorem)

$$g_{W^n_{\mathscr{C}}}(u_i, u_j)^{1/\rho} \leqq (\sqrt{W^n(\varphi^{-1}(u_j)|u_i)})^{1/\rho} + (\sqrt{W^n(\varphi^{-1}(u_i)|u_j)})^{1/\rho} +$$

$$+\exp\left\{-\frac{n}{\rho}\left(\max_{0\leqq s\leqq 1} E^s_x(R, P, W) - sK - \delta\right)\right\} +$$

$$+\exp\left\{-\frac{n}{\rho}\left(\max_{0\leqq s\leqq 1} E^s_x(R, P, W) + sK - \delta\right)\right\}.$$

Note that $g_{W^n_{\mathscr{C}}}(u_i, u_i) = 1$ for all i. Observe that on the right hand side of the last inequality the last term is bounded by the preceding one. Further, by (II) and (III) of the Theorem:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}((\sqrt{W^n(\varphi^{-1}(u_j)|u_i)})^{1/\rho} + (\sqrt{W^n(\varphi^{-1}(u_i)|u_j)})^{1/\rho}) \leqq$$

$$\leqq 2\max_{i}\sum_{j\neq i}(\sqrt{W^n(\varphi^{-1}(u_j)|u_i)})^{1/\rho} \leqq$$

$$\leqq 2\cdot(n+1)^{|\mathscr{X}|^2} \max_{\substack{P_{XX'}\, n\text{-type} \\ P_X = P_{X'} = P \\ I(X\wedge X')\leqq R}} \exp\{n(R - I(X\wedge X')) - \frac{n}{2\rho}\max_{0\leqq s\leqq 1}(\mathbf{E}d_{W,s}(X, X') + sK)\} \leqq$$

$$\leqq \max_{\substack{P_X = P_{X'} = P \\ I(X\wedge X')\leqq R}} \exp\left\{-n\max_{0\leqq s\leqq 1}\left(E^s_{x,\rho}(R, P, W) + \frac{sK}{2\rho} - \delta\right)\right\},$$

where the last step is valid for large n because of the continuity of all involved functions. Note that the union of all n-types on $\mathscr{X}\times\mathscr{X}$, $n = 1, 2, \ldots$ is a dense set in the set of all probability measures on $\mathscr{X}\times\mathscr{X}$. Altogether we have now for large n:

$$\log\frac{1}{N}\sum_{i,j} g_{W^n_{\mathscr{C}}}(u_i, u_j)^{1/\rho} = \log\left(1 + \frac{1}{N}\sum_{\substack{i,j \\ i\neq j}} g_{W^n_{\mathscr{C}}}(u_i, u_j)^{1/\rho}\right) \leqq$$

$$\leqq \log\left(1 + \exp\left\{-n\max_{0\leqq s\leqq 1}\left(E^s_{x,\rho}(R, P, W) + \frac{sK}{2\rho} - \delta\right)\right\} +\right.$$

$$\left. + 2\exp\left\{-n/\rho\max_{0\leqq s\leqq 1}(E^s_x(R, P, W) - sK - \delta)\right\}\exp\{nR\}\right).$$

Inserting this into the estimate of $E_x(rn(R-\delta), W^n_{\mathscr{C}})$ we see that the exponent $F(r, R, P)$ defined in Theorem 2 is achievable for (r, R)-rate concatenated codes.

In the sequel we give a further study of $F(r, R, P)$. Note that

$$F(r, R, P) \leqq \max_{\rho \geqq 1, K \geqq 0} \left\{(1-r)\rho R - \left|\rho R - \max_{0 \leqq s \leqq 1} (E_x^s(R, P, W) - sK)\right|^+\right\} =$$

$$= \max_{\rho \geqq 1} \left\{(1-r)\rho R - \left|\rho R - \max_{0 \leqq s \leqq 1} E_x^s(R, P, W)\right|^+\right\} =$$

$$= \max_{\rho \geqq 1} \{(1-r)\rho R - |\rho R - E_x(R, P, W)|^+\} =$$

$$= \begin{cases} -rR + E_x(R, P, W) & \text{if} \quad R \geqq E_x(R, P, W) \\ (1-r)E_x(R, P, W) & \text{if} \quad R \leqq E_x(R, P, W)\,. \end{cases}$$

For a lower bound we prove the next result.

Lemma 1. For any $R \geqq 0$, $\rho \geqq 1$

$$-\rho E_{x,\rho}^{1/2}(R, P, W) \leqq \rho R - 1/2(E_x(R, P, W) + R)\,.$$

Furthermore, strict inequality holds if $E_x(0, P, W) > E_x(R, P, W) + R$.

Proof. Let (X, X') be a minimizing pair of random variables in the definition of $E_{x,\rho}^{1/2}(R, P, W)$. Then

$$-\rho E_{x,\rho}^{1/2}(R, P, W) = -\rho\left(\frac{1}{2\rho}\,\mathbf{E}\, d_{W,1/2}(X, X') + I(X \wedge X') - R\right)$$

$$\leqq \rho R - 1/2\,\mathbf{E}\, d_{W,1/2}(X, X') - 1/2\, I(X \wedge X')$$

$$\leqq \rho R - 1/2\, R - 1/2\, E_x(R, P, W)\,.$$

In the first inequality we used $\frac{1}{2} I(X \wedge X') \leqq I(X \wedge X')$. This inequality is strict unless $I(X \wedge X') = 0$. Therefore, we have a strict estimate unless all (X, X') achieving the min in the definition of $E_{x,\rho}^{1/2}(R, P, W)$ satisfy $I(X \wedge X') = 0$, or, what is the same, $\mathbf{E}\, d_{W,1/2}(X, X') = E_x(0, P, W)$.

This condition implies

$$-\rho E_{x,\rho}^{1/2}(R, P, W) = \rho R - 1/2\, E_x(0, P, W)\,,$$

and together with the assertion of the Lemma we get

$$E_x(0, P, W) \leqq E_x(R, P, W) + R\,.$$

Remark. Observe that the condition for strict inequality is satisfied for most channels W and small R.

We apply this Lemma in the following estimate (set $\rho=1$, $K=0$)

$$F(r, R, P) \geqq (1-r)R - \left| \max\left\{R - E_x(R, P, W), \; - \max_{0 \leqq s \leqq 1} E^s_{x,1}(R, P, W)\right\}\right|^+ \geqq$$

$$\geqq (1-r)R - |\max\{R - E_x(R, P, W), \; -E^{1/2}_{x,1}(R, P, W)\}|^+ =$$

$$= (1-r)R - R + E_x(R, P, W), \qquad \text{if} \quad R \geqq E_x(R, P, W).$$

In summary, we have now

$$F(r, R, P) = -rR + E_x(R, P, W), \qquad \text{if} \quad R \geqq E_x(R, P, W).$$

We discuss succintly the case $R \leqq E_x(R, P, W)$. Fix r and R. For a maximizing pair ρ, K in the definition of $F(r, R, P)$ at least one of the two expressions

$$\rho R - \max_{0 \leqq s \leqq 1} (E^s_x(R, P, W) - \rho K); \qquad - \max_{0 \leqq s \leqq 1} \left(E^s_{x,\rho}(R, P, W) + \frac{sK}{2\rho}\right)$$

is nonnegative. As a function of K, the first is increasing with K, the second one decreasing with K. Further, there is a K such that the first expression is larger than, or equal to, the second one. Now one can see that for a maximizing pair ρ, K either $K=0$ or K is such that the two above expressions are equal.

These considerations lead to the following formulation for $F(r, R, P)$. Let

$$E_\rho(R, P, W) = \rho R - \max_{0 \leqq s \leqq 1} (E^s_x(R, P, W) - sK(\rho)),$$

where $K(\rho) \geqq 0$ is such that

$$0 \leqq \rho R - \max_{0 \leqq s \leqq 1} (E^s_x(R, P, W) - sK(\rho)) = -\rho \max_{0 \leqq s \leqq 1} \left(E^s_{x,\rho}(R, P, W) + \frac{sK(\rho)}{2\rho}\right).$$

If there is no such $K(\rho)$, set $E_\rho(R, P, W) = (1-r)\rho R$. Then,

$$F(r, R, P) = \begin{cases} -rR + E_x(R, P, W) & \text{if} \quad R \geqq E_x(R, P, W) \\ (1-r)E_x(R, P, W) & \text{if} \quad R \leqq E_x(R, P, W), \quad R \in \mathscr{R} \\ \max\limits_{\rho \geqq 1} \{(1-r)\rho R - E_\rho(R, P, W)\} & \text{else}, \end{cases}$$

where $\mathscr{R} = \{R \mid R \leqq E_x(R, P, W)$, max in the formula for $F(r, R, P)$ is achieved for $K=0\}$.

Proof of Corollary 2. We use Lemma 1. In the first step consider the case $s=1/2$.

$$F(r, R, P) \geqq$$

$$\geqq \max_{\rho \geqq 1, K \geqq 0} \left\{ (1-r)\rho R - \left| \max \left\{ \rho R - E_x(R, P, W) - \frac{K}{2}, \; -\rho E_{x,\rho}^{1/2}(R, P, W) + \frac{K}{4} \right\} \right|^{+} \right\}$$

$$\geqq \max_{\rho \geqq 1, K \geqq 0} \left\{ (1-r)\rho R - \left| \max \left\{ \rho R - E_x(R, P, W) - \frac{K}{2}, \rho R - 1/2\, E_x(R, P, W) - \right.\right.\right.$$

$$\left.\left.\left. - R/2 + \frac{K}{4} \right\} \right|^{+} \right\}.$$

If $R \leqq E_x(R, P, W)$, then we can choose

$$K = 2/3\, E_x(R, P, W) - 2/3 R \geqq 0$$

on the right hand side to obtain

$$F(r, R, P) \geqq \max_{\rho \geqq 1} \{(1-r)\rho R - |\rho R - (2/3\, E_x(R, P, W) + 1/3\, R)|^{+}\}$$

$$\geqq (1-r) \frac{2E_x(R, P, W) + R}{3}, \qquad \text{if} \quad R \leqq E_x(R, P, W)\,.$$

V. Discussion

In this paper we discussed decoders which are allowed to use an erasure symbol. This is a difference to the approach of Forney in [1]. Forney restricted his view to inner decoders which decode just an inner code word and which do not provide any further information being contained in the channel output. Thus, when we were able to give better bounds than Forney's, it might have been caused by the fact that we allow a larger class of decoders. Here, we show that our bound given in Theorem 2 is even better than Forney's result, if we restrict our attention to the case $K=0$.

If one considers in the formula defining $F(r, R, P)$ only the case $K=0$, $s=1/2$, we see that

$$H(r, R, P) = \max_{\rho \geqq 1} \{(1-r)\rho R - |\max\{\rho R - E_x(R, P, W); -\rho E_{x,\rho}^{1/2}(R, P, W)\}|^{+}\}$$

is an achievable error exponent for concatenated codes which do not use erasure decoding.

Lemma 2

$$H(r, R, P) > (1-r)\frac{E_x(R, P, W) + R}{2}$$

if $R < E_x(R, P, W)$ and if the condition of Lemma 1 is satisfied, i.e. if $E_x(0, P, W) > E_x(R, P, W) + R$.

Proof. Choose R as in the statement of the Lemma. In this case, Lemma 1 gives (set $\rho = 1$!)

$$-E^{1/2}_{x,1}(R, P, W) < R - 1/2(E_x(R, P, W) + R) = \frac{1}{2}(R - \mathrm{E}_x(R, P, W)) < 0\,.$$

Therefore,

$$\max\{R - \mathrm{E}_x(R, P, W); \quad -E^{1/2}_{x,1}(R, P, W)\} < 0\,.$$

We conclude: There exists a $\rho^* \in [1, \infty)$ such that

$$\max\{\rho^* R - E_x(R, P, W); \quad -\rho^* E^{1/2}_{x,\rho^*}(R, P, W)\} < 0\,.$$

If now $\rho^* R - E_x(R, P, W) \geqq -\rho^* E^{1/2}_{x,\rho^*}(R, P, W)$, then by the definition of ρ^*

$$\rho^* R = E_x(R, P, W) > \frac{E_x(R, P, W) + R}{2}\,,$$

and we get

$$H(r, R, P) > (1-r)\rho^* R > (1-r)\frac{E_x(R, P, W) + R}{2}\,.$$

If, however, $\rho^* R - E_x(R, P, W) < -\rho^* E^{1/2}_{x,\rho^*}(R, P, W)$, then the definition of ρ^* together with Lemma 1 yields $0 = -\rho^* E^{1/2}_{x,\rho^*}(R, P, W) < \rho^* R - 1/2(E_x(R, P, W) + R)$ which is the same inequality as in the first case.

References

1. *Forney, G. D.*, Concatenated Codes, MIT Press, Cambridge, Mass., 196.
2. *Csiszár, I., Körner, J., Marton, K.*, "A new look at the error exponent of a discrete memoryless channel, preprint, presented at the IEEE International Symposium on Information Theory, Oct. 1977, Cornell, N. Y.
3. *Csiszár, I., Körner, J.*, Information Theory: Coding Theorems for Discrete Memoryless Systems, Akadémiai Kiadó, Budapest 1981.

Граница вероятности ошибки с выбрасыванием для каскадных кодов

Г. ДЮК, В. МОЛЛЕР

(Билефелд)

Приведенные исследования позволяют получить новую, более точную, чем у Форни, верхнюю гранипу для вероятности ошибки при декодировании каскадного кода в дискретном канале без памяти.

В работе сделан подробный анализ и показано, что наихудшим расширенным каналом со стиранием как с точки зрения экспоненты случайного кодирования, так и с точки зрения экспоненты с выбрасыванием, является канал с равновероятными ошибками и стиранием.

G. Dueck
V. Möller
Universität Bielefeld
Fakultät für Mathematik
Postfach 8640
4800 Bielefeld
FRG

Problems of Control and Information Theory, Vol. 15 (2), pp. 149—158 (1986)

ON THE SEPARATION PRINCIPLE IN THE PROBLEM OF ENSURED CONTROL-ESTIMATION

S. V. KRUGLIKOV

(Sverdlovsk)

(Received February 8, 1985)

A minimax problem [1, 2] of optimal control is considered for a linear observed system with uncertain squarely bounded parameters. A quality index is formed by an integral over a tube of informational sets consistent with available observations [3]. Sufficient conditions for an optimal control to be generated through the superposition of solutions of two independent problems are given. These are problems of ensured mean square estimation and of control with complete measurement, respectively.

1. Introduction

The types of control observation problems under discussion in modern control theory depend upon the type of treatment of uncertainty in the mathematical system model. These problems may be considered in terms of either stochastic [4, 5] or ensured (guaranteed) control and estimation theory [1–3]. In particular, in case of control with incomplete information caused by imperfect measurement of the state space variables the solution of a fairly general feedback control problem requires a rather cumbersome procedure. However, for a rather broad class of problems of stochastic feedback control the so-called separation principle [4–6] is true. The latter allows us to reduce the solutions to a separate treatment of the state estimation problem and the feedback control problem with perfect information on the state space variables.

A similar question may be raised for problems of control and observation in a deterministic (guaranteed) setting. In the present paper the ensured control-estimation problem with the perfomance criterion of special type is considered. Sufficient conditions are given for the optimal control to be constructed according to the separation principle. The obtained assertion is based on results of [7].

Let us explain the chosen notations. Suggest that a certain finite interval $[t_0, t_1]$ is given. Then a symbol $f(\cdot)$ denotes an element of a corresponding space of functions defined on $[t_0, t_1]$; $f(t)$ is a value of $f(\cdot)$ at the point t; and $f^t(\cdot)$ is a function on $[t_0, t] \subseteq [t_0, t_1]$, whose values satisfy the equality $f^t(\tau) = f(\tau)$ a.e. on $[t_0, t]$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. Problem statement

Consider the observed uncertain linear dynamic system of n-th order

$$\dot{x}=A(t)x+B(t)u+C(t)v, \quad x(t_0)=x_0, \tag{2.1}$$

$$y=G(t)x+D(t)v, \quad t\in[t_0,t_1]. \tag{2.2}$$

Here u is an m-vector control, y is a k-dimensional observation signal for which all the measured values are accumulated. The initial state $x_0\in R^n$ and the disturbances $v(\cdot)\in L_2^r[t_0,t_1]$ have been unknown in advance, but their possible realisations are restricted by a quadratic inequality

$$(x_0-\bar{x})^T M(x_0-\bar{x})+\int_{t_0}^{t_1}(v(\tau)-\bar{v}(t))^T R(\tau)(v(t)-\bar{v}(\tau))\,d\tau\leqq\mu^2, \tag{2.3}$$

where $\mu=\mathrm{const}$ and $(\bar{x},\bar{v}(\cdot))\in R^n\times L_2^r[t_0,t_1]$ are fixed.

For all $t\in[t_0,t_1]$ the matrices M, $R(t)$ are positive-definite, and $D(t)$ is of full rank (in the columns), $r(D(t))=k$. In particular, it means that the dimensions of the observation and disturbance vectors $y(t)$ and $v(t)$ are interconnected by an inequality $k\leqq r$. The matrix functions $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $D(\cdot)$, $G(\cdot)$, $R(\cdot)$ are continuous on $[t_0,t_1]$.

Let us give some definitions and notations. Further it will be assumed that the control is generated through a feedback procedure on the basis of the whole available information on the system perfomance. Therefore the term "control" here designates an operator $u=u(\eta^t(\cdot))$ that maps the variety of all pairs $\eta^t(\cdot)=\{t,\varphi^t(\cdot)\}$, $t\in[t_0,t_1]$, $\varphi(\cdot)\in L_2^k[t_0,t_1]$ into R^m.

Definition 2.1. The control $u=u(\eta^t(\cdot))$ is said to be admissible if for all $\varphi_i(\cdot)\in L_2^k[t_0,t_1]$, $i=1,2$, the functions $u_i[\cdot]$: $u_i[t]=u(\{t,\varphi_i^t(\cdot)\})$, $t\in[t_0,t_1]$, are square-integrable on the interval $[t_0,t_1]$ and a Lipschitz inequality is valid for each $t\in[t_0,t_1]$, namely

$$\|u_1^t[\cdot]-u_2^t[\cdot]\|_{L_2}\leqq K\|\varphi_1^t(\cdot)-\varphi_2^t(\cdot)\|_{L_2},$$

where the constant $K\geqq 0$ does not depend on the choice of t, $\varphi_i(\cdot)$.

For example, these conditions are satisfied by each operator u_I of the form

$$u_I(\{t,\varphi^t(\cdot)\})=\int_{t_0}^{t}X(t,\tau)\varphi(\tau)\,d\tau, \tag{2.4}$$

$$X(\cdot,\cdot)\in L_2^{m\times k}([t_0,t_1]\times[t_0,t_1]).$$

Definition 2.2. A vector function $x[\cdot]=x(\cdot;u\cdot,x_0,v(\cdot))$ absolutely continuous in t is said to be the solution of system (2.1) with the admissible control $u=u(\eta^t(\cdot))$ and with the pair $(x_0,v(\cdot))$ satisfying (2.3), if there exists $u[\cdot]\in L_2^m[t_0,t_1]$ such that a substitution $x = x[\cdot]$, $u = u[\cdot]$ converts expression (2.1) into equality a.e. on $[t_0, t_1]$, moreover $u[t] = u(\{t, y^t(\cdot)\})$ for a.e. $t \in [t_0, t_1]$, where

$$y^t[\tau] = G(\tau)x[\tau] + D(\tau)v(\tau), \quad \tau \in [t_0, t].$$

A standard application of the contractive transformations principle shows that for each $u = u(\eta^t(\cdot))$ and $(x_0, v(\cdot))$ the solution $x(\cdot; u, x_0, v(\cdot))$ exists and is unique.

Let a certain admissible control $u = u(\eta^t(\cdot))$ be given. Then it is convenient to define the functional position of system (2.1) as a pair $\zeta^t(\cdot) = (t, y^t(\cdot))$, where $t \in [t_0, t_1]$, and $y^t(\cdot)$ is the signal observed on the interval $[t_0, t]$. In the sequel the set of all possible positions with $u = u(\eta^t(\cdot))$ given, is denoted by the symbol $\Xi(t, u)$; $\Delta^{s,p}(u)$ is the operator set,

$$\Delta^{s,p}(u) = \{\sigma_u = \sigma_u(t, \kappa | \zeta^{t_1}(\cdot)) | \sigma_u: [t_0, t_1] \times R^s \times \Xi(t_1, u) \to R^p\}.$$

The elements $\chi_u \in \Delta^{s,p}(u)$ denoted by $\chi_u = \chi_u(\kappa | \zeta^t(\cdot))$ are nonanticipative. This means that for any $\zeta_i(\cdot) \in \Xi(t_1, u)$, $i = 1, 2$, an equality $\zeta_1^\vartheta(\cdot) = \zeta_2^\vartheta(\cdot)$ at $\vartheta \in [t_0, t_1]$ is followed by

$$\chi_u(\kappa | \zeta_1^\tau(\cdot)) = \chi_u(\kappa | \zeta_2^\tau(\cdot)) \qquad \text{a.e. on} \quad [t_o, \vartheta]. \tag{2.5}$$

Moreover, each operator $\sigma_u \in \Delta^{s,p}(u)$ generates a mapping $\sigma_u = \sigma_u[t, \kappa]$ from $[t_0, t_1] \times R^s$ into R^p when a position $\zeta^{t_1}(\cdot) \in \Xi(t_1, u)$ is given.

Having known an admissible control u and a signal realization $y^t(\cdot)$, an informational domain [2, 3] $\mathscr{X}(u, \zeta^t(\cdot))$, $\zeta^t(\cdot) = (t, y^t(\cdot)) \in \Xi(t, u)$ of states consistent with $y(\cdot)$ may be constructed. Among its elements there is a vector $x(t)$ characterizing the actual current state of object. Note that a more exact description of $x(t)$ is impossible.

Consider the following performance criterion for the system (2.1)–(2.2) on the set of available controls

$$J(u) = \sup\{I(u, \zeta^{t_1}(\cdot)) | \zeta^{t_1}(\cdot) \in \Xi(t_1, u)\}, \tag{2.6}$$

$$I(u, \zeta^{t_1}(\cdot)) = \int\limits_{\mathscr{X}(u, \zeta^{t_1}(\cdot))} L_0(\kappa)\gamma_u(\kappa | \zeta^{t_1}(\cdot))\, d\kappa +$$

$$\int\limits_{t_0}^{t_1} \int\limits_{\mathscr{X}(u, \zeta^{t_1}(\cdot))} L_u(t, \kappa | \zeta^{t_1}(\cdot))\gamma_u(\kappa) | \zeta^t(\cdot))\, d\kappa\, dt.$$

Here for arbitrary $u = u(\eta^t(\cdot))$ and $\zeta^{t_1}(\cdot) \in \Xi(t_1, u)$ the nonnegative operators $L_0 = L_0(\kappa)$, $L_u = L_u(t, \kappa | \zeta^{t_1}(\cdot))$, $\gamma_u = \gamma_u(\kappa | \zeta^t(\cdot))$; $L_0 : R^n \to R^1$; $L_u, \gamma_u \in \Delta^{n,1}(u)$ generate the functions $l_0 = L_0(\kappa)\ \gamma_u[t_1, \kappa]$, $l_u = L_u[t, \kappa]\gamma_u[t, \kappa]$ which are integrable correspondently on the domain $\mathscr{X}(u, \zeta^{t_1}(\cdot))$ and on the tube of informational sets

$$X(u, \zeta^{t_1}(\cdot)) = \bigcup(\{t\} \times \mathscr{X}(u, \zeta^t(\cdot)) | t_0 < t \leqq t_1).$$

Problem 2.1. Find an admissible control $u^* = u^*(\eta^t(\cdot))$ such that $J(u^*) \leqq J(u)$ for all admissible $u = u(\eta^t(\cdot))$.

In the sequel we discuss a probabilistic interpretation of this problem. This is possible due to a structure coincidence of the ensured problem 2.1 and the stochastic problem of control with incomplete observation [4]. Thus, $I(u, \zeta^{t_1}(\cdot))$ is analogous to a conditional expectation of an integral functional with respect to an observed signal. The multiplier $\gamma_u = \gamma_u(\kappa | \zeta^t(\cdot))$ corresponds to a density of a conditional distribution.

Definition 2.3. The vector $\hat{x} = \hat{x}(u, \zeta(\cdot)) \in R^n$ is said to be the optimal state estimate at the moment $t \in [t_0, t_1]$ for system (2.1), provided that the following condition holds

$$\max_x \{\|x - \hat{x}\| \,|\, x \in \mathscr{X}(u, \zeta^t(\cdot))\} =$$

$$\min_z \max_x \{\|x - z\| \,|\, x, z \in \mathscr{X}(u, \zeta^t(\cdot))\}.$$

An informational domain $\mathscr{X}(u, \zeta^t(\cdot))$ is known [7] to be an ellipsoid,

$$\mathscr{X}(u, \zeta^t(\cdot)) =$$

$$= \{x \colon (x - \hat{x}(u, \zeta^t(\cdot)))^T P^{-1}(t)(x - \hat{x}(u, \zeta^t(\cdot))) \leqq \mu^2 - h^2(\zeta^t(\cdot))\}.$$

Here for all $u = u(\eta^t(\cdot))$ and $\zeta^{t_1}(\cdot) = (t_1, y(\cdot)) \in \Xi(t_1, u)$ an inequality $h^2(\zeta^\tau(\cdot)) \leqq \mu^2$ is correct and the functions $P(\cdot)$, $\hat{x}[\cdot]$, $h^2[\cdot]$ satisfy the system

$$\dot{P} = AP + PA^T + CR^{-1}C^T - (PG^T + CR^{-1}D^T) \times$$

$$\times (DR^{-1}D^T)^{-1}(PG^T + CR^{-1}D^T)^T, \quad P(t_0) = M^{-1}. \tag{2.7}$$

$$\dot{\hat{x}} = A\hat{x} + Bu + C\bar{v} + Sw(t), \quad \hat{x}[t_0] = \bar{x}, \tag{2.8}$$

$$h^2[t] = \int_{t_0}^{t} w^T(\tau)(D(\tau)R^{-1}(\tau)D^T(\tau))^{-1} w(\tau)\, d\tau,$$

where

$$S(t) = (P(t)G^T(t) + C(t)R^{-1}(t)D^T(t))(D(t)R^{-1}(t)D^T(t)^{-1},$$

$$w(t) = y(t) - G(t)\hat{x}[t] - D(t)\bar{v}(t), \; t \in [t_0, t_1].$$

Equations (2.7)–(2.8) partly coincide in their structure with the formulas of Kalman–Bucy filtration for the case of correlated noises. The function $w(\cdot)$ is analogous to a stochastic innovation process [4]. Moreover, according to (2.1)–(2.2) and (2.8) its values are completely determined if the pair $(x_0, v(\cdot))$ is given.

Lemma 2.1. For every admissible control $u = u(\eta^t(\cdot))$ there exists an operator ψ_u, generating a one-to-one correspondence between elements $v(\cdot) \in W$,

$$W = \{v(\cdot) \colon \int_{t_0}^{t_1} v^T(\tau)(D(\tau)R^{-1}(\tau)D^T(\tau))^{-1} v(\tau))\; d\tau \leqq \mu^2\},$$

and positions $\zeta^{t_1}(\cdot)\in\Xi(t_1,u)$. Furthermore, if $v_i(\cdot)\in W$, $i=1$, 2, are such that $v_1^{\vartheta}(\cdot)=v_2^{\vartheta}(\cdot)$, $\vartheta\in[t_0,t_1]$, then $\zeta_1^{\vartheta}(\cdot)=\zeta_2^{\vartheta}(\cdot)$, $\zeta_i(\cdot)=\psi_u(v_i(\cdot))$.

In fact, the integral Volterra equation of the form

$$z(\tau)-\int_{t_0}^{\tau} G(\tau)\Phi(\tau,s)B(s)u(\{s,z^s(\cdot)\})\,ds+$$

$$\int_{t_0}^{\tau} G(\tau)\Phi(\tau,s)S(s)z(s)\,ds=f(\tau), \tag{2.9}$$

$\tau\in[t_0,\ t_1]$, for each right-hand side $f(\cdot)\in L_2^k[t_0,\ t_1]$ has the unique solution $z_f(\cdot)\in L_2^k[t_0,t_1]$. It follows from the properties of an admissible control. Here $\Phi(t,\tau)$ is the state transition matrix associated with the system matrix $A(t)-S(t)G(t)$. In addition, note that the set W is the totality of $w(\cdot)$ being available with respect to (2.1)–(2.3). Hence the mapping ψ_u is defined by $\psi_u(v(\cdot))=(t_1,z_v(\cdot))$, where $z_v(\cdot)$ is the solution of equation (2.9) with $f(\cdot)=v(\cdot)+\bar{f}(\cdot)$,

$$\bar{f}(t)=G(t)\Phi(t,t_0)\bar{x}+D(t)\bar{v}(t)+\int_{t_0}^{t} G(t)\Phi(t,\tau)(C(\tau)-S(\tau))\bar{v}(\tau)\,d\tau.$$

Moreover, a similar reasoning is possible for equation (2.9) defined on the interval $[t_0,\vartheta]$.

Thus one can use rather a pair $\zeta^t(\cdot)\in\Xi(t,u)$ or element $v^t(\cdot)$, $v(\cdot)\in W$ as a functional position when a control $u=u(\eta^t(\cdot))$ is given. The set W does not depend on the choice of $u=u(\eta^t(\cdot))$. So a notion of admissible control can be defined as an operator $\bar{u}=\bar{u}(v^t(\cdot))$ by the corresponding superposition of the maps u and $\psi_u=\psi_u(v(\cdot))$.

Further it is supposed that the transit to positions $v^t(\cdot)$ has been done and all the above-mentioned notations remained. In particular,

$$J(u)=J(\bar{u})=\sup\{I(u,v^{t_1}(\cdot))\,|\,v^{t_1}(\cdot)\in W\}.$$

According to the form of informational domains $\mathscr{X}(u,\zeta^t(\cdot))$ one can consider the functional $J(u)$ as a perfomance criterion for system (2.8) describing a minimax estimate. Thereby the separation property for problems of control and observation holds, provided that a solution $u^*=u^*(\eta^t(\cdot))$ of problem 2.1 is defined by the equality

$$u^*[t]=\bar{\Psi}(t,\hat{x}[t]) \qquad \text{a.e. on} \quad [t_0,t_1],$$

where

$$\hat{x}[t]=\hat{x}(u^*,\zeta^t(\cdot)),\ \bar{\Psi}\colon [t_0,t_1]\times R^n\to R^m.$$

3. Separation principle

Later, together with problem 2.1, the following one is considered for a system with complete measurement.

Problem 3.1. Find an admissible control $\bar{u}_* = \bar{u}_*(v^t(\cdot))$ such that $\bar{J}(\bar{u}_*) \leqq \bar{J}(\bar{u})$ for all $\bar{u} = \bar{u}(v^t(\cdot))$, where

$$\bar{J}(\bar{u}) = \sup_{W^0} \Bigg\{ \int\limits_{\Sigma(v^{t_1}(\cdot))} L_0(\kappa + z[t_1])\gamma_u(\kappa + z[t_1] \,|\, v^{t_1}(\cdot))\, d\kappa +$$

$$+ \int\limits_{t_0}^{t_1} \int\limits_{\Sigma(v^{\tau}(\cdot))} L_u(\tau, \kappa + z[\tau] \,|\, v^{t_1}(\cdot))\gamma_u(\kappa + z[\tau] \,|\, v^{\tau}(\cdot))\, d\kappa\, d\tau \Bigg\}, \tag{3.1}$$

$W^0 = W \backslash \partial W$ is an interior of W. An ellipsoid $\Sigma(v^\tau(\cdot))$ and a function $z[\cdot]$, $z[\tau] = z(\bar{u}, v^\tau(\cdot))$ for all $v(\cdot) \in W$ are defined by

$$\Sigma(v^\tau(\cdot)) = \{\kappa\colon \kappa^T P^{-1}(\tau)\kappa \leqq \mu^2 - h^2(v^\tau(\cdot))\}, \tag{3.2}$$

$$\dot{z} = A(t)z + B(t)\bar{u} + C(t)\bar{v}(t) + S(t)v(t), \quad z[t_0] = \bar{x}. \tag{3.3}$$

Since by lemma 2.1, system (3.3) is equivalent to (2.8), we have $z(\bar{u}, v^t(\cdot)) = \hat{x}(\bar{u}, v^t(\cdot))$, and $\bar{J}(\bar{u}) \leqq J(\bar{u})$. Actually, the elements $v(\cdot) \in W^0$ correspond to the signals which do not provide the exact state description for system (2.1) and the reduction of the domains $\Sigma(v^t(\cdot))$, $\mathscr{X}(\bar{u}, v^t(\cdot))$ to singleton sets.

Further on some concrete conditions on the functionals L_0, L_u, γ_u are given under which the values of controls u^*, $\bar{u}_*$ solving problems 2.1, 3.1, respectively, coincide and are defined by

$$\bar{u}_*[t] = \hat{\psi}(t, \hat{x}[t]). \tag{3.4}$$

The operator $\hat{\psi}\colon [t_0, t_1] \times R^n \to R^m$ is of the form

$$\hat{\psi}(t, z) = -Q_2^{-1}(t)B^T(t)Q(t)z, \tag{3.5}$$

$$-\dot{Q} = A^T Q + QA + Q_1 - QBQ_2^{-1}B^T Q, \quad Q(t_1) = Q_0, \tag{3.6}$$

$Q_0 \geqq 0$; $Q_1(t) \geqq 0$, $Q_2(t) > 0$ at every $t \in [t_0, t_1]$. The matrix functions $Q_1(\cdot)$, $Q_2(\cdot)$ are continuous. By the Cauchy formula and (2.4), expression (3.4) generates the admissible control, and therefore in such a case the separation principle may be formulated.

Suggestion 3.1. Let a representation

$$L_0(\kappa) = \kappa^T Q_0 \kappa, \qquad L_u(\tau, \kappa \,|\, v^{t_1}(\cdot)) = \kappa^T Q_1(\cdot)\kappa + \lambda_u(\tau \,|\, v^{t_1}(\cdot)),$$

$$\gamma_u(\kappa \,|\, v^\tau(\cdot)) = \Gamma(\kappa - \hat{x}[\tau] \,|\, v^\tau(\cdot)), \qquad \hat{x}[\tau] = \hat{x}(\bar{u}, v^\tau(\cdot)),$$

be valid on W for all $\bar{u}=\bar{u}(v^t(\cdot))$ and $(\tau, \kappa)\in[t_0, t_1]\times R^n$. Given on W, the nonnegative functionals $\lambda_u=\lambda_u(\tau|v^{t_1}(\cdot))$ and $\Gamma=\Gamma(\kappa|v^\tau(\cdot))$ are such that $\lambda_u=\lambda_u[\cdot]\in L_1[t_0, t_1]$ and for arbitrary $v^{t_1}(\cdot)\in W^0$

$$\int\limits_{\Sigma(v^{t_0}(\cdot))} \Gamma(\kappa|v^t(\cdot))d\kappa=1 \quad \text{a.e.} \quad t\in[t_0, t_1].$$

The conditions so formulated allow us to define the functionals L_u, γ_u through the choice of λ_u and Γ. Note that the character of dependence of γ_u on admissible control is strictly determined. But by a concrete selection of Γ the analogies of various conditional densities may be constructed. For example, if Γ:

$$\Gamma(\kappa|v^\tau(\cdot))=\chi^n(v^\tau(\cdot))\Gamma_0^{-1}(\tau)\exp\{-\chi(v^\tau(\cdot))\kappa^T P^{-1}(\tau)\kappa\},$$

where $v^{t_1}(\cdot)\in W^\circ$, $\chi(v^\tau(\cdot))=(\mu^2-h^2(v^\tau(\cdot)))^{-1}$,

$$\Gamma_0(t)=\int\limits_{\{z:z^TP^{-1}(t)z\leqq 1\}} \exp\{-z^TP^{-1}(t)z\}\,dz,$$

then the functional γ_u on W^0 coincides in the structure with a density of the normal distribution. Furthermore, at every $v_*^{t_1}(\cdot)\in\partial W$ there exists a moment $t_*\in[t_0, t_1]$ such that

$$\forall\ \bar{t}\in[t_0, t_*)\ \exists\ \bar{v}(\cdot)\in W^0 : \bar{v}^{\bar{t}}(\cdot)=v_*^{\bar{t}}(\cdot)\ \&$$
$$\Sigma(v_*^t(\cdot))=\{0\} \quad \forall t\in[t_*, t_1]. \tag{3.7}$$

Therefore the values $\gamma_u(\kappa|v_*^\tau(\cdot))$ are described by (2.5) interior to the interval $[t_0, t_*)$ and may be arbitrary outside it.

Lemma 3.1. If suggestion 3.1 is fulfilled everywhere on W^0, such a representation should be correct

$$I(\bar{u}, v^{t_1}(\cdot))=\hat{x}^T[t_1]Q_0\hat{x}[t_1]+2\hat{x}^T[t_1]Q_0Z(v^{t_1}(\cdot))+\alpha(v^{t_1})+$$
$$+\int\limits_{t_0}^{t_1}\{\hat{x}^T[t]Q_1(t)\hat{x}[t]+2\hat{x}^T[t]Q_1(t)Z(v^t(\cdot))+\lambda_u(t|v^{t_1}(\cdot))\}\,dt,$$

where $\hat{x}[t]=\hat{x}(\bar{u}, v^t(\cdot))$,

$$Z(v^t(\cdot))=\int\limits_{\Sigma(v^t(\cdot))} \kappa\Gamma(\kappa|v^t(\cdot))\,d\kappa\in R^n,$$

$$\alpha(v^{t_1}(\cdot))=\int\limits_{\Sigma(v^{t_1}(\cdot))} \kappa^TQ_0\kappa\Gamma(\kappa|v^{t_1}(\cdot))\,d\kappa+$$
$$+\int\limits_{t_0}^{t_1}\int\limits_{\Sigma(v^\tau(\cdot))} \kappa^TQ_1(\tau)\kappa\Gamma(\kappa|v^\tau(\cdot))\,d\kappa\,d\tau.$$

Provided that for a certain $v^{t_1}(\cdot)\in W^0$ the function $\Gamma=\Gamma[\tau,\kappa]$ is even in $\kappa\in R^n$ the vector $Z(v^{\tau}(\cdot))=0$ and a minimax estimate $\hat{x}(\bar{u},v^t(\cdot))$ may be interpreted as a conditional mean value of a state vector.

Define on W the operators $\Lambda_0=\Lambda_0(\tau,\xi|v^{t_1}(\cdot))$ and $\varphi=\varphi(\tau|v^{t_1}(\cdot))$, $(\tau,\xi)\in\in[t_0,t_1]\times R^m$, interconnected by

$$\begin{aligned}\Lambda_0(\tau,\xi|v^{t_1}(\cdot))=(\xi-Q_2^{-1}(\tau)B^T(\tau)\varphi(\tau|v^{t_1}(\cdot)))^TQ_2(\tau)\\(\xi-Q_2^{-1}(\tau)B^T(\tau)\varphi(\tau|v^{t_1}(\cdot))).\end{aligned}\tag{3.8}$$

If $v^{t_1}(\cdot)\in W$ is fixed, the function $\varphi=\varphi[\tau]$ satisfies the equation

$$-\dot{\varphi}=A^T\varphi+Q(C\bar{v}+Sv)+Q_1Z[t],\quad \varphi[t_1]=Q_0Z[t_1],$$

where $Z[t]=Z(v^t(\cdot))$, and $Q(\cdot)$ is the solution of (3.6).

Suggestion 3.2. Let there exist a functional

$$\Lambda=\Lambda(\tau,\xi|v^{t_1}(\cdot)),(\tau,\xi)\in[t_0,t_1]\times R^m,$$

satisfying for all $v(\cdot)\in W$ and a.e. $\tau\in[t_0,t_1]$ the conditions

1) for each admissible control $\bar{u}=\bar{u}(v^t(\cdot))$

$$\lambda_u(\tau|v^{t_1}(\cdot))=\Lambda(\tau,\bar{u}[\tau]|v^{t_1}(\cdot)),$$

2) the function $s_\tau(\cdot)$, $s_\tau(\xi)=\Lambda[\tau,\xi]-\Lambda_0[\tau,\xi]$, attains minima with respect to $\xi\in R^m$ at the point $\xi_*=\bar{u}_*[\tau]$ (3.4),

3) at any $v_*^{t_1}\in\partial W$ (3.7), the following property holds

$$\forall\,\varepsilon\succ 0\;\exists\,\delta>0\;\forall\,\tilde{t}\in[t_*-\delta,t_*)$$

$$\exists\,v_\varepsilon(\cdot)\in W^0\colon v_\varepsilon^{\tilde{t}}(\cdot)=v_*^{\tilde{t}}(\cdot)\,\&$$

$$|\Lambda(\tau,\xi|v_\varepsilon^{t_1}(\cdot))-\Lambda(\tau,\xi|v_*^{t_1}(\cdot))|<\varepsilon\hat{\Lambda}(\tau,\xi|v_*^{t_1}(\cdot)).$$

Here $\hat{\Lambda}=\hat{\Lambda}(\tau,\xi|v^{t_1}(\cdot))$ is such that for all $\bar{u}=\bar{u}(v^t(\cdot))$ the functions $\hat{\Lambda}=\hat{\Lambda}[\tau,\bar{u}[\tau]]$ are integrable on $[t_0,t_1]$.

Note that for $\Lambda(\tau,\xi|v^{t_1}(\cdot))=\Lambda_0(\tau,\xi|v^{t_1}(\cdot))$ (3.8), suggestion 3.2 follows 3.1.

Theorem 3.1. Let suggestions 3.1–3.2 be fulfilled. Then $\bar{u}_*=\bar{u}_*(v^t(\cdot))$ (3.4) solves problem 3.1.

Proof. Consider on W^0 the functional V:

$$V(\bar{u},v^{\tau}(\cdot))=\hat{x}^T(\bar{u},v^{\tau}(\cdot))Q(\tau)\hat{x}(\bar{u},v^{\tau}(\cdot)),$$

$\tau\in[t_0,t_1]$. The matrix function $Q(\cdot)$ is the solution of Riccati equation (3.6). Having integrated the full derivative of $V=V[t]$ on the interval $[t_0,t_1]$, one obtains for $I(\bar{u},v^{t_1}(\cdot))$, $v(\cdot)\in W^0$, the following representation

$$I(\bar{u},v^{t_1}(\cdot))=\bar{x}^TQ(t_0)\bar{x}+2\bar{x}^T\varphi[t_0]+\beta(v^{t_1}(\cdot))+$$

$$+\int_{t_0}^{t_1}(\bar{u}[\tau]-\hat{\psi}[t,\hat{x}[t]])^T Q_2(t)(\bar{u}[\tau]-\hat{\psi}(t,\hat{x}[t])\,dt+$$

$$+\int_{t_0}^{t_1}(\Lambda[t,\bar{u}[t]]-\hat{\Lambda}[t,\bar{u}[t]])\,dt,$$

where $\bar{u}[t]=\bar{u}(v^t(\cdot))$, $\hat{x}[t]=\hat{x}(\bar{u}, v^t(\cdot))$, and the operator $\hat{\psi}$ is defined by (3.5). Therefore

$$I(\bar{u}, v^{t_1}(\cdot))\geqq I(\bar{u}_x, v^{t_1}(\cdot)), \quad v^{t_1}(\cdot)[t_0, t_1\rfloor\in W^0.$$

Condition 3) of suggestion 3.2 is not used in the above given reasoning.

Lemma 3.2. If suggestions 3.1–3.2 hold then $J(\bar{u})=\bar{J}(\bar{u})$ for all $\bar{u}=\bar{u}(v^t(\cdot))$.

Actually, the inequation $J(\bar{u})>\bar{J}(\bar{u})$ is not valid. Otherwise there exists $v_*(\cdot)\in\partial W$ for which

$$I(\bar{u}, v_*(\cdot))=\sup\{I(\bar{u}, v(\cdot))\,|\,v(\cdot)\in W^0\}+\alpha, \quad \alpha>0.$$

But according to (3.7), 3) and the absolute continuity of the Lebesgue integral of $\rho(\cdot)$,

$$\rho(\tau)=\int_{\mathscr{X}(\bar{u},\, v_*^\tau(\cdot))} L_u(\tau,\kappa\,|\,v_*^{t_1}(\cdot))\gamma_u(\kappa\,|\,v_*^\tau(\cdot))\,d\kappa,$$

one can find $\bar{v}(\cdot)\in W^0$ such that $I(\bar{u}, \bar{v}(\cdot))>I(\bar{u}, v_*(\cdot))-\alpha/2$.

Thus, problems 2.1 and 3.1 are equivalent and the next result holds.

Theorem 3.2. (The separation principle). Let suggestions 3.1–3.2 be fulfilled for linear system (2.1)–(2.2) and perfomance criterion (2.6). Then the admissible control $u^*=u^*(\zeta^t(\cdot))$ solving problem 2.1 exists and its current values are defined by

$$u^*[t]=\psi(t, x[t]), t\in[t_0, t_1].$$

Here $\hat{x}[t]=\hat{x}(u^*, \zeta^t(\cdot))$ is the optimal minimax state estimate for system (2.1) at the point t with respect to the measured realization of signal (2.2). The operator $\hat{\psi}$ of form (3.5) generates a feedback control, solving problem 3.2 with the complete observation. Moreover,

$$J(u^*)=\min_{u}\ \sup_{\Xi(t_1,u)} I(u, \zeta^{t_1}(\cdot))=\sup_{W}\ \min_{\bar{u}} I(\bar{u}, v^{t_1}(\cdot)).$$

Acknowledgements

The author is grateful to A. B. Kuržanskii for posing this problem and for the useful discussions.

References

1. *Krasovskii, N. N.*, Game problems on the encounter of motions, Moscow, "Nauka", 1970.
2. *Kuržanskii, A. B.*, Control and observation under uncertainty, Moscow, "Nauka", 1977.
3. *Kuržanskii, A. B.*, Dynamical problems of decision making under uncertainty. In: Modern condition of operation theory, Moscow, "Nauka", 1979.
4. *Fleming, W. H.*, *Rishel, R. W.*, Deterministic and stochastic optimal control, Springer-Verlag, 1975.
5. *Bryson, A. E.*, *Ho, Yu-Chi*, Applied optimal control, Blaisdell Publ. Co., 1969.
6. *Wonham, W. M.*, On the separation theorem of stochastic control. SIAM J. Control, 1968, vol. **6,** No. *2*, pp. 312–326.
7. *Kruglikov, S. V.*, On separation of problems of control and observation under uncertainty. Different. Uravn., 1985, vol. **21,** No. *3*, pp. 398–404.

О принципе разделения в задаче гарантированного управления-оценивания

С. В. КРУГЛИКОВ

(Свердловск)

В стохастической теории оптимального управления для систем с неполным составом измерения в широком классе случаев справедлив, так называемый, принцип разделения. А именно, решение общей задачи управления строится в виде суперпозиции решений двух независимых задач: наблюдения и управления при полной информации о состоянии объекта. В данной работе для линейной наблюдаемой системы с неопределенными заранее параметрами, ограниченными квадратичным неравенством, рассматривается задача позиционного управления в минимаксной постановке. При этом показателем качества является интеграл по трубке информационных областей, совместимых с результатами наблюдения. Приведены достаточные условия на подинтегральную функцию, при которых оптимальное управление строится согласно принципу разделения.

С. В. Кругликов
Институт математики и механики УНЦ АН СССР,
СССР, 620219, Свердловск ГСП-384, ул. С. Ковалевской, 16.

Problems of Control and Information Theory, Vol. 15 (2), pp. 159—166 (1986)

KNAPSACK-TYPE CRYPTOSYSTEMS AND ALGEBRAIC CODING THEORY

H. NIEDERREITER

(*Vienna*)

(Received April 16, 1985)

Recently Chor and Rivest proposed a knapsack-type public-key cryptosystem for low-weight message vectors. We introduce cryptosystems of this type involving public keys with fewer bits and yielding a higher information rate than the Chor–Rivest cryptosystem. The design of these cryptosystems is based on techniques from algebraic coding theory.

1. Introduction

In the last decade the field of cryptography, which is concerned with the design of systems for the communication of secret information, has undergone a dramatic development stimulated by the introduction of public-key cryptosystems in the fundamental paper of Diffie and Hellman [3]. In a *public-key cryptosystem*, the encryption keys of all correspondents are available in a public directory, whereas each correspondent keeps his decryption key secret. If correspondent *B* wants to send a confidential message to correspondent *A*, he looks up *A*'s encryption key in the directory, uses this key to encipher the message, and transmits the resulting ciphertext to *A*. If the cryptosystem is well designed, then only *A* can recover the original message in a reasonable amount of time by applying his decryption key.

Various types of public-key cryptosystems are known today. The principles of most of them can be traced back to the RSA cryptosystem of Rivest, Shamir, and Adleman [13] and the knapsack cryptosystem of Merkle and Hellman [11]. As a representative sample of recent proposals we mention the Massey–Omura lock (see [18]), the public-key cryptosystem of ElGamal [4], the FSR cryptosystems of the author [12], and the knapsack-type cryptosystem of Chor and Rivest [2]. Knapsack-type cryptosystems are based on the difficulty of recovering the summands from the value of their sum. Information on cryptosystems can be found in the books of Beker and Piper [1] and Lidl and Niederreiter [8].

In this paper we introduce a class of knapsack-type public-key cryptosystems that are based on devices from algebraic coding theory. We recall that a *linear* (n, k) *code*

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

C over the finite field F_q of order q is a k-dimensional linear subspace of the n-dimensional vector space F_q^n over F_q, where $1 \leqq k < n$. Thus C contains exactly q^k code words. For a row vector $\mathbf{y} \in F_q^n$ we define the *weight* $w(\mathbf{y})$ to be the number of nonzero coordinates of $\mathbf{y}$, and for $\mathbf{x}, \mathbf{y} \in F_q^n$ we let $d(\mathbf{x}, \mathbf{y}) = w(\mathbf{x}-\mathbf{y})$ be the Hamming distance. The *minimum distance* of C is defined as the smallest weight of a nonzero code word of C. If t is a positive integer, then C is called *t-error-correcting* if for any $\mathbf{y} \in F_q^n$ there is at most one $\mathbf{c} \in C$ such that $d(\mathbf{y}, \mathbf{c}) \leqq t$. If C has minimum distance d, then the largest error-correcting capability of C is $t = \lfloor (d-1)/2 \rfloor$. For a general background on algebraic coding theory we refer to the books of MacWilliams and Sloane [9] and van Lint [17].

We will compare our cryptosystems with the recent knapsack-type cryptosystem of Chor and Rivest [2], and so we briefly describe the latter cryptosystem. Let F_q be a publicly known finite field with $q = p^t$, p prime, $t \geqq 2$, and choose a primitive element β of F_q at random. The field F_q should be such that it is feasible to calculate the discrete logarithm ind (α) of any nonzero element α of F_q, where $a = \operatorname{ind}(\alpha)$ is the unique integer with $\alpha = \beta^a$ and $0 \leqq a \leqq q-2$. For well-chosen (but secret) elements $\alpha_1, \ldots, \alpha_p$ of F_q calculate the integers $a_i = \operatorname{ind}(\alpha_i)$, $1 \leqq i \leqq p$, and scramble them by applying a random permutation and a random shift. The resulting integers $c_1, \ldots, c_p$ form the public key, the random data are kept secret. With this cryptosystem we encipher binary messages $\mathbf{m} = (m_1, \ldots, m_p) \in F_2^p$ of length p and weight $< t$. The ciphertext corresponding to $\mathbf{m}$ is the integer $E(\mathbf{m})$ with $0 \leqq E(\mathbf{m}) \leqq p^t - 2$ and

$$E(\mathbf{m}) \equiv \sum_{i=1}^{p} m_i c_i \bmod (p^t - 1).$$

A crucial lemma (in a corrected form given in Lidl and Niederreiter [8, Ch. 9]) shows that distinct message vectors of the type above are mapped into distinct ciphertexts. The decryption of ciphertexts is possible on the basis of the secret information (compare with [2], [8, Ch. 9]).

2. The cryptosystems

We first describe a *conventional cryptosystem* which is a prototype of our public-key cryptosystem. Choose a t-error-correcting linear (n, k) code C over F_q. Good choices for C will be discussed in Section 3. Let H be a parity-check matrix of C, i.e. H is an $(n-k) \times n$ matrix over F_q of rank $n-k$ such that C consists exactly of all $\mathbf{c} \in F_q^n$ with $H\mathbf{c}^T = \mathbf{0}$, where $\mathbf{c}^T$ denotes the transpose of $\mathbf{c}$. The cryptosystem depends on the simple but crucial fact that the matrix H yields a mapping from F_q^n to F_q^{n-k} that is one-to-one when restricted to vectors of weight $\leqq t$.

Lemma. If $H\mathbf{y}^T = H\mathbf{z}^T$ for some $\mathbf{y}, \mathbf{z} \in F_q^n$ with $w(\mathbf{y}) \leqq t$ and $w(\mathbf{z}) \leqq t$, then $\mathbf{y} = \mathbf{z}$.

Proof. From $H\mathbf{y}^T = H\mathbf{z}^T$ we get $H(\mathbf{y}-\mathbf{z})^T = \mathbf{0}$, hence $\mathbf{y}-\mathbf{z}=\mathbf{c}$ for some $\mathbf{c} \in C$. Now $d(\mathbf{y}, \mathbf{0}) = w(\mathbf{y}) \leqq t$ and $d(\mathbf{y}, \mathbf{c}) = w(\mathbf{y}-\mathbf{c}) = w(\mathbf{z}) \leqq t$, and so we must have $\mathbf{c}=\mathbf{0}$ by the definition of a t-error-correcting code. Therefore $\mathbf{y}=\mathbf{z}$.

In the conventional cryptosystem we keep H secret and encipher a plaintext message $\mathbf{y} \in F_q^n$ of weight $\leqq t$ as the ciphertext $H\mathbf{y}^T$. Upon receipt of this ciphertext, we can recover $\mathbf{y}$ uniquely. Note that in the language of algebraic coding theory $H\mathbf{y}^T$ is the syndrome of $\mathbf{y}$ with respect to the code C. Since $d(\mathbf{y},\mathbf{0}) = w(\mathbf{y}) \leqq t$, we may view $\mathbf{y}$ as an error vector relative to the code word $\mathbf{0}$. Therefore, an application of the decoding algorithm of C to the syndrome $H\mathbf{y}^T$ will yield the error vector $\mathbf{y}$.

To obtain a *public-key cryptosystem* from this conventional cryptosystem, we form a scrambled version of the matrix H and take it as a public key. This can be done in various ways. For instance, we may use the following scrambling device employed in the Goppa-code cryptosystem (see [8, Ch. 9]): premultiply H by a randomly chosen nonsingular $(n-k) \times (n-k)$ matrix M over F_q and postmultiply by a randomly chosen $n \times n$ matrix P over F_q that is obtained by permuting the rows of a nonsingular diagonal matrix. The matrices M, H, and P are kept secret, whereas the $(n-k) \times n$ matrix $K = MHP$ serves as the public key. A plaintext message $\mathbf{y} \in F_q^n$ of weight $\leqq t$ is now enciphered as $K\mathbf{y}^T$. Thus the ciphertexts are column vectors over F_q of length $n-k$. Upon receipt of the ciphertext $K\mathbf{y}^T = MHP\mathbf{y}^T$, we premultiply it by M^{-1} to get $HP\mathbf{y}^T = H(\mathbf{y}P^T)^T$. Note that $\mathbf{y}P^T$ is again a vector of weight $\leqq t$. Therefore we can obtain $\mathbf{y}P^T$ by the same method as in the conventional cryptosystem, namely by applying the decoding algorithm of C. Postmultiplying by $(P^T)^{-1}$, we recover the original message $\mathbf{y}$.

In the binary case $q=2$ this cryptosystem is of the classical knapsack type. The ciphertext $K\mathbf{y}^T$ is then just a sum of at most t column vectors of the public-key matrix K, and determining $\mathbf{y}$ is equivalent to deciding which column vectors of K yield the given sum $K\mathbf{y}^T$. For general q, the ciphertext $K\mathbf{y}^T$ is a linear combination of at most t column vectors of K with coefficients from F_q, and finding $\mathbf{y}$ is equivalent to determining this linear combination explicitly.

3. The choice of codes

A brief inspection of the construction of our cryptosystems shows that a code C suitable for our purposes should satisfy the following requirements: (i) C should have a relatively large error-correcting capability (or equivalently a large relative distance d/n) so that a reasonable number of message vectors can be used; (ii) C should allow an efficient decoding algorithm so that the decryption can be carried out with a short run time. Further analysis reveals that the dimension k of C should be in a medium range relative to the length n. For if k is too small, then there are relatively few good codes of dimension k, which makes it easier to break the cryptosystem. On the other hand, if k is too close to n, this results in short ciphertexts, which again imperils the security of the cryptosystem.

There are various classes of linear codes that satisfy the criteria above. A benchmark for the quality of a code is provided by the Gilbert–Varshamov bound (see [9, Ch. 17], [17, Ch. 5]). A family of good codes should meet this bound, at least asymptotically. A well-known family of codes that meet this bound is given by alternant codes, and these codes also allow an efficient decoding algorithm (see [9, Ch. 12]). An important subclass of alternant codes is formed by Goppa codes, which have the advantage that they can be described quite easily in terms of a suitable polynomial. Goppa codes still meet the Gilbert–Varshamov bound (see [17, Ch. 8]). According to a result of Sarwate [14], a t-error-correcting Goppa code of length n can be decoded in $0(n \log^2 n)$ arithmetic operations for fixed t/n. For a detailed description of the decoding algorithm for Goppa codes see Lidl and Niederreiter [8, Ch. 8] and McEliece [10, Ch. 8]. There is another family of good linear codes introduced recently by Goppa [5], namely his algebraic geometry codes. This family contains codes that even go beyond the Gilbert–Varshamov bound for sufficiently large q (see [15], [16]), but on the other hand the decoding problem has not yet been solved satisfactorily.

Another suitable class of codes is given by Reed–Solomon codes (see [9, Ch. 10]). These codes are of interest since they are maximum distance separable, i.e. they achieve equality in the Singleton bound $d \leqq n-k+1$ for linear codes. A Reed–Solomon code is a cyclic code of length $n=q-1$ over F_q with generator polynomial

$$g(x)=(x-\beta^b)(x-\beta^{b+1}) \ldots (x-\beta^{b+d-2}),$$

where the integer $b \geqq 0$ is arbitrary, β is a primitive element of F_q, and any d with $2 \leqq d \leqq n$ may be prescribed. The minimum distance of this code is d and its dimension is $k=n-d+1$. Reed–Solomon codes belong to the well-known family of BCH codes and thus allow an efficient decoding algorithm. In fact, Justesen [6] has shown that a t-error-correcting Reed–Solomon code of length n can be decoded in $0(n \log^2 n)$ arithmetic operations for fixed t/n. Reed–Solomon codes are also instrumental in building concatenated codes of excellent quality, such as the Justesen codes (see [9, Ch. 10]).

4. Discussion

The most desirable property of a public-key cryptosystem is of course its ability to withstand attempts at breaking it. Possible attacks can be directed against two targets, either the deciphering of a specific ciphertext without knowledge of the secret keys, or the more ambitious goal of determining the secret keys M, H, and P. The latter task is complicated by the fact that the "factorization" $K=MHP$ of the public key is by no means unique. This is in marked contrast to the RSA cryptosystem, where one can at least rely on the unique factorization of integers.

Even if k and $n-k$ are only moderately large, a brute-force attack on the secret keys based on trying all possibilities is hopeless. Note that the number of possibilities for M is equal to the number of nonsingular $(n-k)\times(n-k)$ matrices over F_q, which is

$$q^{(n-k)^2}\prod_{j=1}^{n-k}(1-q^{-j})$$

by a well-known formula (see [7, p. 401]). The number of possibilities for the code C is equal to the number of k-dimensional linear subspaces of F_q^n, which is

$$\prod_{j=0}^{k-1}(q^{n-j}-1)(q^{k-j}-1)^{-1}$$

by [7, p. 456]. The number of possibilities for P is $n!\,(q-1)^n$. A brute-force attack on deciphering a specific ciphertext would be based on trying all possible message vectors $\mathbf{y}\in F_q^n$ of weight $\leq t$. The number of such vectors is

$$\sum_{j=0}^{t}\binom{n}{j}(q-1)^j .$$

Our public-key cryptosystem shares a common feature with that of Chor and Rivest in Section 1, namely that it works with low-weight message vectors. A comparison with the Chor–Rivest cryptosystem immediately shows one drawback of this cryptosystem, namely a longer setup time due to time-consuming calculations of discrete logarithms. Another aspect which favors our cryptosystem is the information rate. Let S be the number of possible messages and T the number of possible ciphertexts in a cryptosystem. Then the *information rate* of the cryptosystem is defined by

$$R=\frac{\log_2 S}{\log_2 T},$$

where $\log_2$ denotes the logarithm to the base 2. Thus R may be viewed as the amount of information contained per bit of ciphertext. To have a fair comparison, we consider both the Chor–Rivest cryptosystem and our cryptosystem for binary message vectors of length n and weight $\leqq t$. For the Chor–Rivest cryptosystem we have then

$$S=\sum_{j=0}^{t}\binom{n}{j}\leqq 2^n,\quad T=n^{t+1}-1\geqq n^t .$$

Let $t=\theta n$ with $0<\theta<1$. Then

$$R\leqq\frac{n}{\theta n\log_2 n},$$

so that asymptotically (i.e. as $n\to\infty$) the information rate R tends to 0.

The information rate of our cryptosystem shows a completely different behavior. In the binary case $q=2$ we have here

$$S=\sum_{j=0}^{t}\binom{n}{j}, \quad T=2^{n-k}.$$

Choose a family of codes that meet the Gilbert–Varshamov bound. According to [9, p. 557, Theorem 30] this means that for given $0<\delta<\frac{1}{2}$ we can achieve relative distance $d/n\geqq\delta$ and asymptotically

$$\frac{k}{n}\gtrsim 1-H_2\left(\frac{d}{n}\right),$$

where $H_2(x)=-x\log_2 x-(1-x)\log_2(1-x)$ is the binary entropy function. Let again $t=\theta n$. Then

$$\log_2 S=\log_2\sum_{j=0}^{\theta n}\binom{n}{j}\geqq H_2(\theta)n-\frac{1}{2}\log_2 n-\log_2 c(\theta)$$

for some constant $c(\theta)>0$ by [9, p. 310, Corollary 9]. Since $\log_2 T=n-k$, we get

$$R\gtrsim\frac{H_2(\theta)}{1-(k/n)}\gtrsim\frac{H_2(\theta)}{H_2(d/n)}.$$

Since $t=\lfloor(d-1)/2\rfloor$, we have $d\sim 2\theta n$ asymptotically, hence

$$R\gtrsim\frac{H_2(\theta)}{H_2(2\theta)}.$$

This means that asymptotically R stays above a positive lower bound. If for instance $\theta\to\frac{1}{4}$, then

$$R\gtrsim H_2\left(\frac{1}{4}\right)\approx 0.81.$$

To have some concrete examples, consider first the binary concatenated code mentioned in [9, p. 308] with parameters $n=104$, $k=24$, $d=32$, $t=15$. This code is obtained by concatenation of the (8, 4) binary extended Hamming code of minimum distance 4 with a (13, 6) punctured Reed–Solomon code over F_{16} of minimum distance 8. In this case the public key K contains $80\cdot 104=8320$ bits. This compares favorably with the Chor–Rivest cryptosystem where for the suggested parameters the key requires about 36 000 bits. With this binary concatenated code, the number of possible message vectors is about $5 \cdot 10^{17}$, and this yields an information rate $R \approx 0.73$. For the parameters $n=104$ and $k=24$, the number of possible binary linear codes is greater than 10^{570} and the number of nonsingular $(n-k) \times (n-k)$ matrices over F_2 is greater than 10^{1900}.

As a second example, consider a Reed–Solomon code over F_{31} with $n=30$, $k=12$, $d=19$, $t=9$. The number of bits in the public key K is then $18 \cdot 30 \cdot \lceil \log_2 30 \rceil = 2700$. This is comparable to the size of the RSA public key, which according to current recommendations requires about 1200 bits. With this Reed–Solomon code, the number of possible message vectors is about $3 \cdot 10^{20}$, and this yields an information rate $R \approx 0.76$. For the parameters $n=30$ and $k=12$, the number of possible linear codes over F_{31} is greater than 10^{320} and the number of nonsingular $(n-k) \times (n-k)$ matrices over F_{31} is greater than 10^{480}.

References

1. *Beker, H., Piper, F.*, Cipher Systems. The Protection of Communications, Northwood, London, 1982.
2. *Chor, B., Rivest, R. L.*, A knapsack type public key cryptosystem based on arithmetic in finite fields, Proc. CRYPTO '84, to appear.
3. *Diffie, W., Hellman, M. E.*, New directions in cryptography, IEEE Trans. Information Theory **22,** 644–654 (1976).
4. *ElGamal, T.*, A public key cryptosystem and a signature scheme based on discrete logarithms, IEEE Trans. Information Theory, to appear.
5. *Goppa, V. D.*, Algebraic-geometric codes (Russian), Izv. Akad. Nauk SSSR Ser. Mat. **46,** 762–781 (1982).
6. *Justesen, J.*, On the complexity of decoding Reed–Solomon codes, IEEE Trans. Information Theory **22,** 237–238 (1976).
7. *Lidl, R., Niederreiter, H.*, Finite Fields, Encyclopedia of Math. and Its Appl., Vol. **20,** Addison-Wesley, Reading, Mass., 1983.
8. *Lidl, R., Niederreiter, H.*, Introduction to Finite Fields and Their Applications, Cambridge Univ. Press, Cambridge, 1985.
9. *MacWilliams, F. J., Sloane, N. J. A.*, The Theory of Error-Correcting Codes, North-Holland, Amsterdam, 1977.
10. *McEliece, R. J.*, The Theory of Information and Coding, Encyclopedia of Math. and Its Appl., Vol. **3,** Addison-Wesley, Reading, Mass., 1977.
11. *Merkle, R. C., Hellman, M. E.*, Hiding information and signatures in trapdoor knapsacks, IEEE Trans. Information Theory **24,** 525–530 (1978).
12. *Niederreiter, H.*, A public-key cryptosystem based on shift register sequences, Proc. EUROCRYPT '85, to appear.
13. *Rivest, R. L., Shamir, A., Adleman, L.*, A method for obtaining digital signatures and public-key cryptosystems, Comm. Assoc. Comput. Mach. **21,** 120–126 (1978).
14. *Sarwate, D. V.*, On the complexity of decoding Goppa codes, IEEE Trans. Information Theory **23,** 515–516 (1977).
15. *Tsfasman, M. A.*, Goppa codes that are better than the Varshamov–Gilbert bound (Russian), Problemy Peredači Informacii **18,** *3*, 3–6 (1982).
16. *Tsfasman, M. A., Vlădut, S. G., Zink, T.*, Modular curves, Shimura curves, and Goppa codes, better than Varshamov–Gilbert bound, Math. Nachr. **109,** 21–28 (1982).
17. *van Lint, J. H.*, Introduction to Coding Theory, Springer, New York, 1982.
18. *Wah, P. K. S., Wang, M. Z.*, Realization and application of the Massey–Omura lock, Proc. Internat. Seminar on Digital Communications (Zürich, 1984), pp. 175–182.

Криптосистемы типа рюкзака и алгебраическая теория кодирования

Х. НИДЕРРЕЙТЕР

(Вена)

В работе проводится описание новой криптосистемы с общим ключом, надежность которой основана на сложности задачи вычисления вектора ошибки e в линейном коде по синдрому He^T, где H — проверочная матрица кода. Описано асимптотическое поведение параметров системы, приведены примеры.

H. Niederreiter
Mathematical Institute
Austrian Academy of Sciences
Dr. Ignaz-Seipel-Platz 2
A-1010 Vienna
Austria

Problems of Control and Information Theory, Vol. 15 (2), pp. 167—177 (1986)

SEQUENTIAL DESIGN OF ROBUST DECENTRALIZED CONTROLLERS FOR SERIALLY INTERCONNECTED SYSTEMS

L. BAKULE J. LUNZE

(Prague) *(Dresden)*

(Received June 10, 1985)

A sequential control design procedure is derived for decentralized controller of serially interconnected systems. The global closed-loop system has to satisfy given dynamical requirements which cannot be decomposed into independent subsystem requirements. The model uncertainties are respected by upper model error bounds. An illustrative example is supplied.

1. Introduction

Feedback control problems for large-scale plants must often be solved in a decentralized way, i.e. the controller consists of several control stations which operate independently on the corresponding subsystems of the global plant. The design of such controllers represents a rather complex problem because the solution must satisfy the structural constraint on the control law, which is imposed by this decentralization. This is true even if the plant is completely known. The complexity of the design problem grows significantly, if there are uncertainties in the plant model.

Therefore, the reduction of complexity and dimensionality is one of the main problems in decentralized control design. If the plant has a hierarchical structure, the complete design problem can be decomposed into subproblems, so that the control stations can be calculated nearly independently by means of a model of the corresponding subsystems only.

Such a decomposition has been described for instance in [4, 8] under the assumption that the plant is completely known and the design requirements are given by the stability and optimality of the closed-loop system according to a quadratic performance index.

This paper deals with the more general problem, where the fulfilment of requirements for the dynamical I/0-behaviour as well as the robustness of the decentralized controller according to model uncertainties represent the dominating difficulties of the design task. As it is explained in Chapter 3.1, because of these design requirements — and in contrast to stability considerations — the parameters of all

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

control stations are mutually dependent, although the plant is assumed to possess only a unidirectional coupling between subsystems. Nonetheless, this paper is aimed at a complete decomposition of the design task, so that the different control stations are designed independently. Concerning the authors' knowledge, a similar solution has been described only in [2] under the assumption that the dynamical requirements can be decomposed into independent requirements on the subsystems. Here, such an assumption is not imposed.

The base of presented procedure is the design of the control stations by means of the approximate model of the isolated subsystems and the robustness evaluation of the closed-loop subsystems outlined in Chapters 3.3, 3.4. An illustrative example is supplied in Chapter 4.

2. Problem formulation

Let us consider a continuous-time plant S composed of N subsystems in the form

$$\begin{aligned} S\colon \quad & \dot{x}_i = A_i x_i + B_i u_i + F_i p_{i-1}\,, && x_i(0) = x_{i0}\,,\\ & y_i = C_i x_i + D_i u_i + H_i p_{i-1}\,, && \\ & p_i = \underline{C}_i x_i + \underline{D}_i u_i + \underline{H}_i p_{i-1}\,, && i = 1, \ldots, N \\ & p_0 = 0\,, && \qquad (1) \end{aligned}$$

where x_i, u_i, y_i, p_i denote the i-th subsystem state vector, respectively, control vector, output vector, interconnection output vector, respectively, see Fig. 1. $A_i, \ldots, H_i$ are constant matrices of appropriated dimensions.

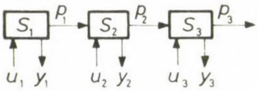


Fig. 1

The objective is to find a decentralized controller such that the closed-loop system satisfies the following requirements:

1) The closed-loop system must be stable.
2) The asymptotic regulation occurs in the closed-loop system for step commands $w_i = \bar{w}_i \sigma(t)$ and step disturbances, i.e. $y_i - \bar{w}_i \underset{t \to \infty}{\to} 0$. w_i is a constant vector, $\sigma(t)$ is unit step function.
3) The dynamical I/0-behaviour must satisfy given requirements.

4) The controller is robust, i.e. requirements 1)–3) are satisfied even if the model used in the design step does not exactly describe the original plant (1).

Because of requirement 2), PI-controllers are used

$$\dot{x}_{ri}=e_i, \quad e_i=y_i-w_i, \quad u_i=K_{1i}e_i+K_{2i}x_{ri}, \tag{2}$$

where $K_i=(K_{1i}, K_{2i})$ is a constant matrix. The asymptotic regulation occurs if the closed-loop system using (2) is stable [3].

Item 3) includes a lot of different requirements on the I/0-behaviour of the closed-loop system. We refer mainly to such typical features as bounds of the overshoot and settling time of the step response y_i for step inputs $w_j=\bar{w}_j\sigma(t)$ at the same subsystem $(j=i)$ or at one of the preceding subsystem $(j<i)$.

3. Solution

3.1 Decomposition of the design problem

Let us describe at the beginning the motivation of our sequential design approach to solve the given design problem.

Because the subsystems of the plant are coupled only in one direction, we try to decompose the design problem in such a way that the different control stations can be designed by means of the model of isolated subsystem.

Concerning requirements 1), 2) the following Lemma can be used.

Lemma 1. The closed-loop global system (1), (2) satisfies requirements 1), 2), iff all isolated closed-loop subsystems, i.e. eqs. (1), (2) for $p_i=0$, $i=2, \ldots, N$, are stable.

Proof. The chain structure of the global feedback system corresponds to a lower block triangular system matrix with diagonal blocks corresponding to free closed-loop subsystems. Therefore, requirement 1) is satisfied iff all free closed-loop subsystems are stable. The fulfilment of requirement 2) follows from the internal model principle [3].

Lemma 1 points out that requirements 1), 2) can be tested on the isolated subsystems only.

A similar complete decomposition of the dynamical requirements is in general impossible. For instance, the I/0-behaviour of subsystem 2 for step inputs w_1 is influenced by both control stations 1 and 2. A very small dependence of y_2 upon w_1 can only be produced by designing control station 1 so as to receive a low dependence of the coupling signal p_1 upon w_1 as well as by choosing control station 2 such that the "disturbance" p_1 is diminished as far as possible.

These considerations motivate a design procedure in which the control stations are parametrized in a sequential way. At first, control station 1 is fitted so as to satisfy the stability as well as the dynamical requirements on subsystem 1. Then, for step inputs $w_1 = \bar{w}_1 \sigma(t)$ the coupling signal $p_1(t)$ is calculated. At second, control station 2 is designed by means of the model of subsystem 2. To check whether a given controller meets the dynamical requirements concerning output y_2 in relation to step input $w_2 = \bar{w}_2 \sigma(t)$, only the model of the closed-loop subsystem 2, i.e. eqs. (1), (2) for $i=2$, $p_1 = 0$, is necessary. To investigate the behaviour of subsystem 2 in relation to command step at subsystem 1, the closed-loop subsystem 2 is considered for the signal $p_1(t)$, which has been received in the first design step. At third, subcontroller 3 is designed similarly to control station 2, etc.

Obviously, in such a sequential design procedure only the model and the design specifications related to a certain subsystem must be considered. Hence, the complexity and dimensionality of our control problem are significantly reduced. Moreover, robustness requirement 4) can be easily included in the design of this decentralized controller. As in each step only the model of the corresponding subsystem is used, the uncertainties of only that part of the model must be considered.

Based on this way of the solution, we have to tackle the problems of *designing the control station i for the approximate model* of free subsystem i as well as of analysing the closed-loop subsystem i for given command steps w_i and interconnection signals p_{i-1}. The second problem must refer to *the model uncertainties of subsystem i.*

3.2 Model of the system

Consider the i-th subsystem model description in the form

$$\begin{aligned} \dot{x}_i &= \hat{A}_i x_i + \hat{B}_i u_i + \hat{E}_i s_i + \hat{F}_i p_{i-1}, \qquad x_i(0) = x_{i0}, \\ y_i &= \hat{C}_i x_i + \hat{D}_i u_i + \hat{G}_i s_i + \hat{H}_i p_{i-1}, \\ z_i &= \hat{\hat{C}}_i x_i + \hat{\hat{D}}_i u_i + \hat{\hat{G}}_i s_i + \hat{\hat{H}}_i p_{i-1}, \\ p_i &= \underline{\hat{C}}_i x_i + \underline{\hat{D}}_i u_i + \underline{\hat{G}}_i s_i + \underline{\hat{H}}_i p_{i-1}, \end{aligned} \tag{3}$$

and

$$|s_i| = V_i \ast |z_i| + r_{0i}(x_{0i}, t), \qquad \forall t \geqq 0, \tag{4}$$

where s_i, z_i, V_i, r_{0i} denote interconnection input vector to the error model (4), interconnection output vector to the error model (4), impulse response matrix and free motion of the model error, respectively. Equation (4) describes an upper bound of the model error [2, 5, 6]. The approximate model is given by eq. (3) for $s_i = 0$. $\ast$ denotes the convolution operation.

3.3 Design of controller for the approximate model

The controller for the i-th subsystem, see eq. (2), is designed by means of LQ-procedures for the approximate model, i.e. for $s_i = 0$, $p_{i-1} = 0$ in eq. (3). Therefore, this system has the form

$$\begin{aligned}\dot{x}_i &= \hat{A}_i x_i + \hat{B}_i u_i\,,\\ \hat{y}_i &= \hat{C}_i x_i + \hat{D}_i u_i\,,\end{aligned} \tag{5}$$

where $\hat{y}_i$ denotes the approximate model output.

The integral part of controller is considered as a part of an extended subsystem in the form

$$\begin{aligned}\dot{\tilde{x}}_i &= \begin{pmatrix}\dot{x}_i\\ \dot{x}_{ri}\end{pmatrix} = \begin{pmatrix}\hat{A}_i & 0\\ \hat{C}_i & 0\end{pmatrix}\begin{pmatrix}x_i\\ x_{ri}\end{pmatrix} + \begin{pmatrix}\hat{B}_i\\ \hat{D}_i\end{pmatrix}u_i + \begin{pmatrix}0\\ -I\end{pmatrix}w_i\,,\\ \tilde{y}_i &= \begin{pmatrix}\hat{y}_i\\ \hat{e}_i\end{pmatrix} = \begin{pmatrix}\hat{C}_i & 0\\ 0 & I\end{pmatrix}\begin{pmatrix}x_i\\ x_{ri}\end{pmatrix} + \begin{pmatrix}\hat{D}_i\\ 0\end{pmatrix}u_i\,.\end{aligned} \tag{6}$$

Then the controller

$$u_i = (K_{1i}, K_{2i})\tilde{y}_i \tag{7}$$

is designed as the solution of the problem

$$J_i = \int_0^\infty (\tilde{y}_i^T q_i Q_i \tilde{y}_i + u_i^T R_i u_i)\, dt \to \min \tag{8}$$

subject to eq. (6). q_i is a scalar used to satisfy robustness requirement 4), see [5], also Chapter 3.5. This solution can be obtained using procedures [1, 5]. The matrices Q_i, R_i must be iteratively designed so that the closed-loop subsystem i (5), (7) satisfies requirements 1)–3).

3.4 Evaluation of robustness

We have to check that the original closed-loop subsystem (2), (3), (4) for fixed i satisfies requirements 1)–4). This system is described in the form

$$\begin{aligned}\dot{\tilde{x}}_i &= \tilde{A}_i \tilde{x}_i + \tilde{B}_i w_i + \tilde{E}_i s_i + \tilde{F}_i p_{i-1}\,,\\ y_i &= \tilde{C}_i \tilde{x}_i + \tilde{D}_i w_i + \tilde{G}_i s_i + \tilde{H}_i p_{i-1}\,,\\ z_i &= \tilde{\tilde{C}}_i \tilde{x}_i + \tilde{\tilde{D}}_i w_i + \tilde{\tilde{G}}_i s_i + \tilde{\tilde{H}}_i p_{i-1}\,,\\ p_i &= \underline{\tilde{C}}_i \tilde{x}_i + \underline{\tilde{D}}_i w_i + \underline{\tilde{G}}_i s_i + \underline{\tilde{H}}_i p_{i-1}\end{aligned} \tag{9}$$

and eq. (4).

System (9) must be investigated concerning its I/0-behaviour with inputs w_i, p_{i-1} and outputs y_i, p_i. We assume that step inputs w_i are used. Information on p_{i-1} is received from the design of the $(i-1)$-th control station. Supposing that $p_{i-1}(t)$ is approximated by a function $\tilde{p}_{i-1}(t)$, the following Lemma is used for the analysis.

Lemma 2. (I.) The sufficient condition for the stability of system (4), (9) is:

1) The approximate closed-loop system (9) is stable.

2) The following inequality is satisfied

$$\lambda_M \left[\int_0^\infty V_i \, dt \int_0^\infty (|\tilde{\tilde{G}}_i| \delta(t) + |\tilde{\tilde{C}}_i e^{\tilde{A}_i t} \tilde{E}_i|) \, dt \right] < 1 \,. \tag{10}$$

(II.) The I/0-behaviour of system (4), (9) can be approximated by eq. (9) for $s_i = 0$. If eq. (11) is satisfied, the upper bound of the model error is given by inequalities

$$\begin{aligned} |y_i - \hat{y}_i| &\leqq V_{ysi} * \bar{V}_i * V_{zwi} * |w_i| + V_{ysi} * \bar{V}_i * V_{zpi} * |\tilde{p}_{i-1}| \,, \\ |p_i - \hat{p}_i| &\leqq V_{psi} * \bar{V}_i * V_{zwi} * |w_i| + V_{psi} * \bar{V}_i * V_{zpi} * |\tilde{p}_{i-1}| \,, \end{aligned} \tag{11}$$

where

$$\begin{aligned} &V_{ysi} = |\tilde{G}_i| \delta(t) + |\tilde{C}_i e^{\tilde{A}_i t} \tilde{E}_i|, && V_{psi} = |\underline{\tilde{G}}_i| \delta(t) + |\underline{\tilde{C}}_i e^{\tilde{A}_i t} \tilde{E}_i| \,, \\ &\bar{V}_i = V_i + V_i * V_{zsi} * \bar{V}_i, && V_{zpi} = |\tilde{\tilde{H}}_i| \delta(t) + |\tilde{\tilde{C}}_i e^{\tilde{A}_i t} \tilde{G}_i| \,, \\ &V_{zsi} = |\tilde{\tilde{G}}_i| \delta(t) + |\tilde{\tilde{C}}_i e^{\tilde{A}_i t} \tilde{E}_i|, && \tilde{G}_i = \begin{pmatrix} \underline{\hat{B}}_i (I - K_{1i} \hat{D}_i)^{-1} \hat{H}_i + \hat{F}_i \\ (I - \hat{D}_i K_{1i})^{-1} \hat{F}_i \end{pmatrix} . \\ &V_{zwi} = |\tilde{\tilde{D}}_i| \delta(t) + |\tilde{\tilde{C}}_i e^{\tilde{A}_i t} \tilde{B}_i| \,, && \end{aligned} \tag{12}$$

λ_M denotes the maximal eigenvalue of the corresponding matrix.

Proof. Lemma 2 is the result of the direct application of Theorems 3, 4 in [6] to system (3), (9).

The results of Lemma 2 are used in two ways:

1) The command tracking of subsystem i is investigated for $w_i = \bar{w}_i \sigma(t)$, $p_{i-1} = 0$. The results are tolerance bands for y_i, p_i, see Example, Fig. 2.
2) The couplings between command steps in the preceding subsystems and y_i, p_i are investigated using $w_i = 0$, $p_{i-1} = \tilde{p}_{i-1}$. The results are new tolerance bands for y_i, p_i.

The function p_i should describe the influence of command steps at subsystems $1, \ldots, i$ on subsystem $i+1$. Because of the uncertainties, we have got two tolerance bands which must be added. As the analysis of the i-th closed-loop subsystem for a band of possible inputs is very complicated, we propose the following choice of $\tilde{p}_{i-1}$

$$\tilde{p}_i = \hat{p}_i + \bar{p}_i \sigma(t) \,, \tag{13}$$

where $\hat{p}_i$ is the function representing the middle of the tolerance band and $\bar{p}_i$ is the maximum width of the band. To use this step as a representation of the uncertainties of the signal p_i, $\tilde{p}_i$ can be considered as the implementation of the worst possible inteconnection signal for PI controller in the i-th subsystem.

3.5 Summary of the design procedure

The derived procedure can be summarized in the following Algorithm.

Algorithm.

1) Initiate N, x_0, $p_0=0$, $\hat{A}_i$, $\hat{B}_i$, $\hat{C}_i$, $\hat{D}_i$, $\hat{E}_i$, $\hat{F}_i$, $\hat{G}_i$, $\hat{H}_i$, $\hat{\bar{\bar{C}}}_i$, $\hat{\bar{\bar{D}}}_i$, $\hat{\bar{\bar{G}}}_i$, $\hat{\bar{\bar{H}}}_i$, $\underline{\hat{C}}_i$, $\underline{\hat{D}}_i$, $\underline{\hat{G}}_i$, $\underline{\hat{H}}_i$, V_i, $q_i=1$, $\forall i$.
2) Test controllability and observability $(\hat{A}_i, \hat{B}_i)$, $(\hat{A}_i, \hat{C}_i)$, $\forall i$. If it is not satisfied, goto 9).
3) $i=1$.
4) Specify Q_i, R_i and solve (6), (8) using standard LQ-procedure. If requirements 1)–3) on the behaviour are not satisfied modify Q_i, R_i and solve (6), (8).
5) Test condition (10). If it is not satisfied decrease q_i and goto 4).
6) Evaluate the closed-loop error bound (11) for step input w_i and interconnection input $\tilde{p}_{i-1}$. If the bands are too broad so that the requirements 3), 4) are not satisfied, modify q_i and goto 4).
7) Determine $\tilde{p}_i$ (cf. eq. (13)).
8) $i=i+1$. If $i \leqq N$ then goto 4).
9) End.

4. Example

Formulation of the control problem

Consider the string of vehicles described in [8]. Find a decentralized controller such that the velocity of the string and the distance between the vehicles have given values. Moreover, setting changes of the velocity should be followed without overshoot, the distance between the vehicles should at no time instant be smaller than a given value and the controller has to satisfy these requirements for a given range of vehicle load.

Solution

The string of vehicles is described in the form [8]

$$\dot{x}_1 = -\frac{1}{m_1} x_1 + \frac{1}{m_1} u_1 ,$$
$$y_1 = x_1 ,$$
$$p_1 = x_1 ,$$

and

$$\dot{x}_i = \begin{pmatrix} -\dfrac{1}{m_i} & 0 \\ -1 & 0 \end{pmatrix} x_i + \begin{pmatrix} \dfrac{1}{m_i} \\ 0 \end{pmatrix} u_i + \begin{pmatrix} 0 \\ 1 \end{pmatrix} p_{i-1} ,$$
$$y_i = (0 \quad 1) x_i ,$$
$$p_i = (1 \quad 0) x_i , \qquad i = 2, \ldots, N \tag{14}$$

where m_i is the mass of the vehicle. x_{1i}, x_{2i} is the velocity of the i-th vehicle, the distance between vehicle i and $i-1$, respectively.

Because the load of the vehicle may change during the performance of the controller, the mass is assumed to vary in the interval $m_i \in \langle 0.8, 1.2 \rangle$. Then the model of the plant is given by the relations for $N = 3$.

$$\dot{x}_1 = -1.042 x_1 + 1.042 u_1 + s_1 ,$$
$$y_1 = x_1 ,$$
$$z_1 = x_1 + u_1 ,$$
$$p_1 = x_1$$

and

$$\dot{x}_i = \begin{pmatrix} -1.042 & 0 \\ -1 & 0 \end{pmatrix} x_i + \begin{pmatrix} 1.042 \\ 0 \end{pmatrix} u_i + \begin{pmatrix} 0 \\ 1 \end{pmatrix} p_{i-1} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} s_i ,$$
$$y_i = (0 \quad 1) x_i ,$$
$$p_i = (1 \quad 0) x_i ,$$
$$z_i = (-1 \quad 0) x_i + u_i , \qquad i = 2, 3 \tag{15}$$

with

$$s_i = h_i z_i \qquad \text{for} \quad |h_i| \leqq 0.208, \qquad i = 1, 2, 3 . \tag{16}$$

Note that these equations describe the plant in the best possible way, because for each value of h_i it represents the plant in a possible mode of performance.

Controller structure: As the command signals can be approximated by steps signals, decentralized PI-controllers are used. To solve the control task, the first vehicle controls its velocity, while the other vehicles control the distance to the preceding vehicle.

Results. The design problem is solved using Algorithm proposed in Chapter 3.5. Because the model error is described by a static single-input single-output system (16), the robustness can be analyzed by determining the smallest possible tolerance bands of the step response of the closed-loop subsystems [7].

The first control station represents a state feedback of model (6). Applying $Q_1 = 0.1\, I_2$, $R_1 = 1$ the controller parameters $k_{11} = -0.316$, $k_{12} = -0.316$ are computed using optimal state feedback control design procedure.

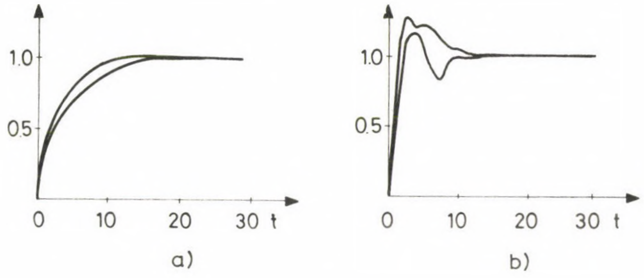


Fig. 2

Figure 2a illustrates the tolerance band of the step response of the closed-loop subsystem 1, i.e. control of velocity for $w_1 = \sigma(t)$. The middle of this band is used as function $\hat{p}_1$ of eq. (13) for the analysis of the coupling between subsystems 1 and 2 in connection with a step of amplitude $\bar{p}_1 = 0.05$.

The second controller is designed as optimal output feedback for (6) with $Q_2 = 0.1\, I_2, R_2 = 1$. It results in $k_{21} = -0.9, k_{22} = -0.17$. The analysis of the closed-loop subsystem 2 must be done with respect to command steps $w_2 = \bar{w}_2 \sigma(t)$ as well as to the coupling signal $\tilde{p}_1$. Figure 2b illustrates the tolerance band — control of distance between vehicle 1 and vehicle 2 — of the response of closed-loop subsystem 2 to command step $\sigma(t)$. A similar band can be obtained for the behaviour of y_2 subject to the "disturbance" p_1.

Finally, Fig. 3 illustrates the behaviour of the global decentralized closed-loop system for all $\hat{m}_i = 1$, i.e. for the approximate global closed-loop model, to the command step at vehicle 1: 1 — velocity of vehicle 1; 2 — distance between vehicles 1 and 2; 3 — distance between vehicles 2 and 3.

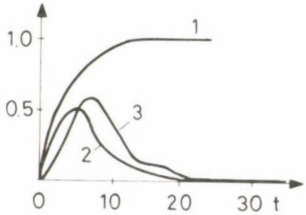


Fig. 3

5. Conclusion

The sequential design procedure derived for decentralized control of serially interconnected subsystems leads to considerable reductions of the complexity and dimensionality of the design problem. Each design step necessitates only the knowledge of an approximate model and a model error bound for the isolated subsystems.

In contrast to approaches known from the literature, our solution refers to the requirements on the dynamical I/0-behaviour and it does not assume that these requirements can be completely decomposed to the subsystems, as for instance in [2].

The presented algorithm can be considered as a principle which underlines the sequential design of the decentralized controllers. The modification of the algorithm can be motivated by specific features of dynamical requirements, e.g. it might be possible to include the approximate models of the closed-loop subsystems $1, \ldots, i-1$ into the approximate model of subsystem i, to look for the smallest possible tolerance band, see [7], or refer to step disturbances in the analysis of the closed-loop system. In any case the decentralized controller is designed at the subsystem level by means of reduced order design subproblems.

References

1. *Levine, W. S., Athans, M.*, On the Determination of the Optimal Constant Output Feedback Gains for Linear Multivariable Systems. IEEE Trans. Automat. Control, vol. **AC-15,** No. *1*, 1970, pp. 44–48.
2. *Bakule, L., Lunze, J.*, Completely Decentralized Design of Decentralized Controllers for Serially Interconnected Systems. 30th Internat. Wissenschaft. Kolloquium, Ilmenau (GDR), *1*, 1985, pp. 15–18.

3. *Davison, E. J.*, The Robust Decentralized Control of a General Servomechanism Problem. IEEE Trans. Automat. Control, vol. **AC-21,** No. *1*, 1976, pp. 14–24.
4. *Hodžič, M., Šiljak, D. D.*, An LQG Design for Large Sparse Systems. Proc. 9th World IFAC Congress, vol. **X,** Budapest, 1984, pp. 293–298.
5. *Lunze, J.*, The Design of Robust Feedback Controllers for Partly Unknown Systems by Optimal Control Procedures. Internat. J. Control, vol. **36,** No. *4*, 1982, pp. 611–630.
6. *Lunze, J.*, The Design of Robust Feedback Controllers in the Time Domain. Internat. J. Control, vol. **39,** No. *6*, 1984, pp. 1243–1260.
7. *Lunze, J., Zscheile, E.*, Analyse der Stabilität und des Übergangsverhaltens linearer Systeme mit Parameterunsicherheit. Messen Steuern Regeln, Berlin 28, 1985 (in print).
8. *Özgüner, Ü., Perkins, W. R.*, Optimal Control of Multilevel Large-Scale Systems. Internat. J. Control, vol. **28,** No. *6*, 1978, pp. 967–980.

Последовательное проектирование серийно соединенных устойчивых систем децентрализованного управления

Л. БАКУЛЕ Я. ЛУНЗЕ

(Прага) (Дрезден)

В работе излагается одна процедура расчета децентрализованного регулятора для системы, состоящей из последовательно соединенных частей. Регулятор должен быть выбран таким, чтобы замкнутая система удовлетворяла ряду требований (устойчивость, робастность и др.). Предполагается алгоритм расчета такого регулятора.

L. Bakule
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Prague, Pod vodárensku věží 4
Czechoslovakia

J. Lunze
Zentralinstitut für Kybernetik und Informationsprozesse
Akademie der Wissenschaften der DDR
1086 Berlin, Kurstrasse 33
DDR

Problems of Control and Information Theory, Vol. 15 (2), pp. 179—183 (1986)

SIXTH INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY (Tashkent, September 18–23, 1984)

The Symposium held in Tashkent in 1984 was attended by 232 participants from the Soviet Union, by 38 from the capitalist countries (Belgium, Brasil, FRG, Finland, France, Italy, Japan, the Netherlands, Norway, Sweden, Switzerland and the USA) and by 56 scientists from the socialist countries (Bulgaria, Czechoslovakia, GDR, Hungary, Poland and Yugoslavia).

The main topics were: I. Mathematical Problems of Information Theory; II. Source Coding, Information-Theoretic Aspects of Coding and Speech; III. Error-correcting Codes; IV. Statistical Theory of Signals and Noise; V. Multicomponent Random Systems; VI. Source and Channel Networks; VII. Information Theoretic Aspects of System Modeling and Optimization.

The most interesting contributions to subject area I were papers on multi-user channels by R. Ahlswede (FRG) "The distortion region for multiple descriptions without excess rate" which proved numerous theorems on multi-user and multi-access channels and by I. Csiszár (Hungary) and R. Ahlswede "Hypothesis testing with exogenous information" which was concerned with extensions of the problem of testing two hypotheses; an exponent of a second type error was obtained with a fixed probability of a first type error.

The paper by De Bruyn and Van der Meulen (Belgium), "Some results on the discrete memoryless multiple-access channel with feedback" determined throughput ranges with feedback transmission for some multiple-access channels.

M. Hellman and J. Peyneri (USA) proved in their paper "Channels and the coding standard" the asymptotic behavior of the size distribution function for the largest component of a random graph and described cryptographic applications.

A. G. Dyachkov (USSR) obtained in his paper, "On random coding bound for multi-access channel" an ensemble-averaged error probability in maximal likelihood decoding and general-purpose encoding for the case of a multiple-access channel and two ensembles of random codes.

Much attention was given to the classical Shannon information theory which at present deals with very sophisticated channels as was the case in S. I. Gelfand's and M. S. Pinker's (USSR) paper "On Gaussian channels with random parameters", G. Sh. Poltyrev (USSR) in his lecture "The expurgation error probability bounds for broadcast channels" discussed transmission in a discrete broadcasting memoryless

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

channel with two receivers whereby general or specific information is fed into one of the receivers and error upperbounds are obtained for an ensemble of expurgation codes.

The papers by R. D. Davisson (USA), "The entropy rate of sliding block coders for Markov sources", by K. Marton (Hungary), "The non-existence of asymptotic isomorphism between discrete memoryless stationary correlated sources", A. K. Gorbunov (USSR), "Epsilon-entropy with delay of Gaussian message", and M. S. Pinsker and L. V. Sofman "ε, σ entropy of completely ergodic processes" display an increased interest in mathematical aspects of source encoding and associated problems of the ergodic theory".

Research is underway in estimation of the throughput and error probability in various communication channels (papers by E. A. Harutyunyan, V. V. Prelov, and V. B. Balakirskii, USSR).

The papers by A. G. Vinck (Netherlands) "Constructive superposition coding for the binary erasure multiple-access channel", Y. Sugiyama (Japan) "An algorithm for solving Toeplitz system based on Euclid algorithms", T. Ericson (Sweden) "A combinatorial problem in the arbitrary varying channel", and G. A. Margulis (USSR) "Arithmetic groups and graps without short cycles" were concerned with various algebraic and combinatorial aspects of data transmission in communication systems.

Subject area II dealt with the methods to encode semi-infinite sequences at the output of a Markov source whose statistics were unknown. New results in universal coding of sources with a finite or infinite alphabet of fixed or variable length blocks were reported in the papers "Complexity of a string relative to the class of Markov sources" by J. Rissanen (Finland), "The Slepian–Wolf theorem for individual sequences" by G. Dueck and L. Wolters (FRG), "On universal coding with given fidelity for continuous sources" by V. F. Babkin and M. M. Gange (USSR), "List source coding and information reconstruction" by R. E. Krichevsky (USSR), and "On Gilbert–Moore codes realization" by Yu. M. Shtar'kov (USSR).

Methods to obtain optimal quantizers were proposed by Z. Györfi and G. Szekeres (Hungary) in their paper "On the structure of recursive vector quantizers" and A. V. Trushkin (USSR) in his paper "On optimal construction".

Applications of information theoretic methods to studies of biological systems were described in the papers "Information theory analysis of ant language" by G. N. Reznikova and B. Ya. Ryabko (USSR) and "Information storage and coding problems in neutron networks" by V. L. Dunin-Barkovski.

Considerable interest was stirred by papers on the processing of elementary particle track patterns by R. A. Pose, U. Bar and A. Schwind (GDR) and by A. Hubler (GDR) and on adaptive coding of geometric figures by V. G. Polyakov, E. A. I. Aidu, V. S. Nagornov, and V. G. Trunov (USSR).

Information theoretic aspects of speech recognition and simulation were treated in the papers by V. A. Abramov and V. S. Dubrovin (USSR), "Unified image and speech recognition procedures in a microprocessor system", A. A. Nekhaev and I. V. Sitnyakovskii (USSR), "A speed signal model for investigation of·speed coding methods", and V. M. Son (USSR), "Switching of speech packets".

In subject area III numerous fundamental results were obtained in discussion of classical problems in the coding theory such as the upperbound on length of optimal codes. 'I nese subjects were discussed by M. A. Tsfasman and S. G. Vladuts (USSR), "Good algebra geometric codes'', and L. A. Bassalygo, E. M. Gabidulin and V R. Sidorenko (USSR), "The Varshamov–Gilbert bound may be asymptotically improved for burst correcting codes".

E. M. Gabidulin's (USSR) survey, "Optimal codes for correcting lattice pattern errors" covered coding theory applications to non-Hamming metrics.

Of the numerous papers on convolutional codes the most interesting were "Synchronization of convolutional codes" by J. Massey (Switzerland), "A low complexity stack decoder for a class of binary rate $(n-1)/n$ convolutional codes" by A. G. Vinck (Netherlands), "Subspaces of $CF(q)^w$ and convolutional codes" by L. Steiger (GDR), "A nearest patch principle for convolution code decoding" by K. Sh. Zigangirov (USSR), and "Lower bounds for the distance of cascade convolutional codes" by V. B. Afanasyev.

Practical methods of using error-correcting codes including new modulations and demodulations were discussed in the papers by V. V. Zyablov and S. V. Shavgulidze (USSR), "Incorrect decoding exponent for a class of convolutional block concatenated codes" and V. I. Korzhik (USSR) "Generalization of Viterbi algorithm for the channel with additive noise described by Markov chain".

In subject area IV the increasing interest in nonparametric static information-theoretic models was obvious in the lively response to papers by R. Yu. Bentkus (USSR) on asymptotic estimation of minimal, in some sense mean square risk, L. Györfi (Hungary) on some unexpected effects in nonparametric estimation of the distribution density, and M. Nussbaum (GDR) who determined the minimax constant for nonparametric regression estimation.

The papers by L. Izzo and L. Paura (Italy), "Multistatic radar detection of fluctuating targets by binary integrators", G. Lukatela (Yugoslavia), "On binary signaling above the Nyquist rate", D. D. Klavsky and S. M. Shirokov (USSR) "Identification of models of random fields in communication channels on the basis of stochastic differential equations", and Yu. G. Sosulin (USSR) "Signal detection in the presence of noise and filtering with bounded complexity" suggest a trend to studying more realistic static models in signal detection and recognition.

The role of statistical methods is on the increase in specific fields of the information theory such as noise-immune radio communication system (paper by M. G. Vasilyeva, A. P. Galieva, and L. N. Goldfeld (USSR), digital broadcasting (paper by L. M. Fink, M. U. Bank, A. S. Grudinin and M. Ya. Lesman (USSR)), earthquake forecasting (paper by A. F. Kushvir and V. I. Pinsky, USSR) and digital synchronization (paper by V. V. Shakhgildyan, USSR).

The papers of subject area V show a trend to a qualitative study of the so-called Pirogov–Sinai theory (K. Kuroda, Japan, S.-E. Pfister, Switzerland, J. Bricmont, Belgium, R. Koteoki, SCCR and M. Zahradnik, CSSR).

The paper obtained numerous equivalent formulations on the properties spin glass models in an "abstract" field.

The meeting was impressed by the paper "Stochastic vortex theory" by M. Pulvirenti (Italy), who proposed a vortex model as a tool to analyze Euler and Navier–Stokes equations in two-dimensional hydrodynamics. This approach may yield solutions to many difficult problems in the dynamics of complex systems.

An important line of research was represented in papers by A. Krámli and D. Szász (Hungary) on translation in some caricature models, J. Fritz (Hungary) "Stochastic gradient dynamics of infinite particle systems", and B. Tót and D. Szász on random hunting with memory in a random environment that concentrated on various models representing deterministic and stochastic dynamics of complex systems such as the Lorenz gas, gradient models, and interacting Markov processes.

Great attention was aroused by papers presented by Ya. G. Sinai and E. B. Vul, "Tori and quanto-tori", Yu. M. Sukhov and A. G. Shukhov, "Limit theorems for linear canonic transformations on C^+-algebra of GAS", S. B. Shlosman and E. A. Pecherski, "Low-temperature non-uniqueness in models with non-degenerate ground state", and R. L. Dobrushin, "Constructive criterion of uniqueness of Gibbs random field" which reported the recent findings of Soviet researchers in multi-component random fields.

The papers in subject area VI confirmed the need for studying multiple access systems in the networks with topology such as those of mobile broadcasting stations. Research efforts in different countries reveal the advantages of the stack-algorithm in local computer networks. In Csibi's (Hungary) paper on stability of random multiple data access during time intervals free of speech transmission, speech packets had priority 1 and data packets priority 2. The throughput of speech transmission channels can be improved by adding data packet transmission. A specific algorithm was characterized in terms of throughput and speech and data packet delays.

Packet data transmission in radio networks was the subject of the paper "ALOHA in communication networks" by B. S. Tsybakov and V. L. Bakyrov (USSR).

Cs. Szabó's (Hungary) paper, "Analysis of asynchronous ALOHA type random cases" concentrated on the dynamic behavior and determined the throughput and packet delay.

I. Kerekes (Hungary) and V. A. Mikhailov (USSR) devoted their paper "Stack-algorithm of RMA for channel with symbol synchronization" to a new model where every symbol as well as packets were synchronized so as to improve the channel throughput.

Problems of Control and Information Theory, Vol. 15 (2), p. 191 (1986)

EXHIBITION ON "INDUSTRIAL ROBOTS AND THEIR COMPONENTS"

An exhibition was organized on the "Industrial Robots and Their Components" in the same time when the Third All-Union Conference on Robotics was held. More than 30 organizations from 17 cities and six Soviet Republics displayed their designs. The exhibition presented more than 90 exhibits including 50 operating full-scale displays. Some exhibits were made in the form of models, brassboard operating models or panels displaying robotic complexes, flexible manufacturing systems and their components such as computerized transportation and warehouse systems, automated modules of mechanical processing, die forging, etc. A large section of the exhibition showed robot sensors and pick-ups of diverse purposes — visual, locational, tactile, heading etc. Samples of industrial robots of various designs and orientation were presented, such as balanced, pneumatic, electromechanical etc. Drivers and actuators, control and diagnosing devices, control units and their elements were displayed as well.

V. M. Nazazetov

РУССКИЙ ПЕРЕВОД

Проблемы управления и теории информации, том **15**, *номер 2*

О ПРИНЦИПЕ РАЗДЕЛЕНИЯ В ЗАДАЧЕ ГАРАНТИРОВАННОГО УПРАВЛЕНИЯ-ОЦЕНИВАНИЯ

С. В. Кругликов

(*Свердловск*)

Для линейной наблюдаемой системы с неопределенными параметрами, ограниченными квадратичным неравенством, рассматривается задача позиционного управления в минимаксной постановке [1, 2]. При этом показателем качества является интеграл по трубке информационных областей [3], совместимых с результатами наблюдения. Приведены достаточные условия на подинтегральную функцию, при которых оптимальное управление определяется суперпозицией решений двух независимых задач: гарантированного среднеквадратического оценивания и управления с полной информацией.

1. Введение

Типичным свойством многих реальных управляемых объектов является неполнота информации о текущем состоянии и внешних возмущениях. При этом можно полагать, что значения всех измеряемых параметров запоминаются и накапливаются. Математическое описание таких объектов часто проводится в рамках стохастической теории управления [4, 5]. Однако весьма распространена ситуация, когда статистические характеристики возмущений неизвестны, но существует некоторое априорное ограничение на их значения. Тогда применим гарантированный подход, основные понятия и методы которого развиты в работах [1, 2].

Независимо от того, выбрана ли для описания объекта с неопределенными параметрами гарантированная или стохастическая модель, можно изучать различные постановки задач управления и наблюдения. В частности, если о состоянии системы поступает некоторая дополнительная информация, то можно исследовать процессы управления, основанные на принципе обратной связи. Однако реализация общих процедур [1–4] решения позиционных задач управления-оценивания достаточно затруднительна. Поэтому представляет интерес выделение класса задач, решение которых может быть получено в явном виде. Для стохастических систем в широком классе случаев справедлив, так называемый, принцип (теорема) разделения [4–6]. А именно, в указанных

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

случаях совпадает вид решений задачи управления при неполном составе измерения и соответствующей ей задачи с полным наблюдением для системы, моделирующей оптимальную оценку. Аналогичный вопрос может рассматриваться для задач позиционного управления-наблюдения в гарантированной постановке.

В данной работе рассматривается задача гарантированного управления для линейной системы с квадратично-ограниченными параметрами. Приведены достаточные условия на критерий качества специального вида, при которых решение общей задачи управления-оценивания строится согласно принципу разделения. Полученное утверждение является развитием результата работы [7].

Поясним принятые обозначения. Предполагается, что задан некоторый конечный интервал времени $[t_0, t_1]$. Символом $f(\cdot)$ обозначается элемент соответствующего пространства функций, определенных на $[t_0, t_1]$, $f^t(\cdot)$ — ограничение $f(\cdot)$ на $[t_0, t] \subseteq [t_0, t_1]$, $f^{t_1}(\cdot) = f(\cdot)$, и $f(t)$ — значение функции $f(\cdot)$ в т. t.

2. Постановка задачи

Рассматривается линейная n-мерная система с наблюдением

$$\dot{x} = A(t)x + B(t)u + C(t)v, \; x(t_0) = x_0, \tag{2.1}$$

$$y = G(t)x + D(t)v, \quad t \in [t_0, t_1]. \tag{2.2}$$

Здесь u m-вектор управляющих воздействий, y k-мерный наблюдаемый сигнал, все значения которого запоминаются по ходу процесса. Неизвестные заранее реализации начальных данных, $x_0 \in R^n$, и внешних возмущений, $v(\cdot) \in L_2^r[t_0, t_1]$, стеснены совместным квадратичным ограничением

$$(x_0 - \bar{x})^T M (x_0 - \bar{x}) + \int_{t_0}^{t_1} (v(\tau) - \bar{v}(\tau))^T R(\tau) (v(\tau) - \bar{v}(\tau)) \, d\tau \leqq \mu^2, \tag{2.3}$$

где $\mu = \mathrm{const}$ и $(\bar{x}, \bar{v}(\cdot)) \in R^n \times L_2^r[t_0, t_1]$ фиксированы.

При всех $t \in [t_0, t_1]$ матрицы M, $R(t)$ положительно определены, а $D(t)$ имеет полный ранг по строкам, $r(D(t)) = k$. Это, в частности, означает, что размерности векторов наблюдения $y(t)$ и помех $v(t)$ связаны неравенством $k \leqq r$. Матричные функции $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $D(\cdot)$, $G(\cdot)$, $R(\cdot)$ непрерывны на $[t_0, t_1]$.

Введем некоторые определения и обозначения. Далее предполагается, что управляющее воздействие формируется согласно принципу обратной связи по всей доступной информации о состоянии объекта. При этом под управлением понимается оператор $u=u(\eta^t(\cdot))$, который каждой паре $\eta^t(\cdot)=\{t,\varphi^t(\cdot)\}$, $t\in[t_0,t_1]$, $\varphi^t(\cdot)\in L_2^k[t_0,t_1]$, ставит в соответствие вектор из пространства R^m.

Определение 2.1. Управление $u=u(\eta^t(\cdot))$ называется допустимым, если для всех $\varphi_i(\cdot)\in L_2^k[t_0,t_1]$, $i=1,\ 2$, функции $u_i[\cdot]$, $u_i[t]=u(\{t,\varphi_i^t(\cdot)\})$, $t\in[t_0,t_1]$, суммируемы с квадратом на интервале $[t_0,t_1]$, и справедливо неравенство Липшица: при всех $t\in[t_0,t_1]$

$$\|u_1^t[\cdot]-u_2^t[\cdot]\|_{L_2}\leqq k\|\varphi_1^t(\cdot)-\varphi_2^t(\cdot)\|_{L_2},$$

где постоянная $k\geqq 0$ от выбора t и $\varphi_i(\cdot)$ не зависит.

Этим условиям удовлетворяет, в частности, любой оператор u_I следующего вида

$$u_I{}^{(}\{t,\varphi^t(\cdot)\})=\int_{t_0}^{t}X(t,\tau)\varphi(\tau)\,d\tau, \tag{2.4}$$

$$X(\cdot,\cdot)\in L_2^{m\times k}[t_0,t_1]\times[t_0,t_1]).$$

Определение 2.2. Абсолютно непрерывная вектор-функция $x[\cdot]=x(\cdot;u,x_0,v(\cdot))$ называется решением системы (2.1) при допустимом управлении u и паре $(x_0,v(\cdot))$, удовлетворяющей ограничению (2.3), если существует $u[\cdot]\in L_2^m[t_0,t_1]$ такая, что при подстановке $x=x[\cdot]$, $u=u[\cdot]$ соотношение (2.1) п. в. на $[t_0,t_1]$ обращается в равенство, и $u[t]=u(\{t,y^t[\cdot]\}$ при п. в. $t\in[t_0,t_1]$, где

$$y^t[\tau]=G(\tau)x[\tau]+D(\tau)v(\tau),\quad \tau\in[t_0,t].$$

Стандартное применение принципа сжимающих отображений позволяет показать, что при любых $x_0, v(\cdot)$ и допустимом $u=u(\eta^t(\cdot))$ решение $x(\cdot;u,x_0,v(\cdot))$ существует и единственно.

Пусть $u=u(\eta^t(\cdot))$ — некоторое заданное допустимое управление. Тогда под функциональной позицией системы (2.1)–(2.2) удобно понимать пару $\zeta^t(\cdot)=(t,y^t(\cdot))$, где $t\in[t_0,t_1]$, и $y^t(\cdot)$ — реализация наблюдаемого сигнала (2.2) на интервале $[t_0,t]$. Множество всех позиций $\zeta^t(\cdot)$ возможных при $u=u(\eta^t(\cdot))$ обозначается далее символом $\Xi(t,u)$, а $\Delta^{s,p}(u)$ — семейство операторов,

$$\Delta^{s,p}(u)=\{\sigma_u=\sigma_u(t,\kappa|\zeta^{t_1}(\cdot))|\sigma_u:[t_0,t_1]\times R^s\times\Xi(t_1,u)\to R^p\}.$$

При этом элементы $\chi_u\in\Delta^{s,p}(\mathrm{u})$, обладающие свойством неупреждаемости, записываются в виде $\chi_u=\chi_u(\kappa|\zeta^t(\cdot))$. Это означает, что для любых $\zeta_i(\cdot)\in\Xi(t_1,u)$, $i=1,2$, таких, что $\zeta_1^\theta(\cdot)=\zeta_2^\theta(\cdot)$ при некотором $\theta\in[t_0,t_1]$, п. в. на интервале $[\mathrm{t}_0,\theta]$ выполняется равенство

$$\chi_u(\kappa|\zeta_1^\tau(\cdot))=\chi_u(\kappa|\zeta_2^\tau(\cdot)). \tag{2.5}$$

Отметим, что если задана позиция $\zeta^{t_1}(\cdot)\in\Xi(t_1,u)$, то каждый оператор $\sigma_u\in\Delta^{s,p}(u)$ порождает функцию $\sigma_u=\sigma_u[t,\kappa]$, отображающую $[t_0,t_1]\times R^s$ в R^p.

При каждом допустимом управлении u, зная реализацию $y^t(\cdot)$ сигнала (2.2), можно определить информационную область [2–3] $\mathscr{X}(u, \zeta^t(\cdot))$, $\zeta^t(\cdot) = (t, y^t(\cdot)) \in \Xi(t, u)$, (множество состояний совместимых с $y^t(\cdot)$). Одним из его элементов является вектор $x(t)$, характеризующий действительное положение объекта. Причем более точное описание $x(t)$ невозможно.

Рассмотрим для системы (2.1)–(2.2) на множестве допустимых управлений следующий критерий качества

$$J(u) = \sup\{I(u, \zeta^{t_1}(\cdot)) \mid \zeta^{t_1}(\cdot) \in \Xi(t_1, u)\}, \tag{2.6}$$

$$I(u, \zeta^{t_1}(\cdot)) = \int\limits_{\mathscr{X}(u, \zeta^{t_1}(\cdot))} L_0(\kappa)\gamma_u(\kappa \mid \zeta^{t_1}(\cdot))\, d\kappa +$$

$$\int\limits_{t_0}^{t_1} \int\limits_{x(u, \zeta^t(\cdot))} L_u(t, \kappa \mid \zeta^{t_1}(\cdot))\gamma_u(\kappa \mid \zeta^t(\cdot))\, d\kappa\, dt.$$

Здесь при всех $u = u(\eta^t(\cdot))$ неотрицательные функционалы $L_0 = L_0(\kappa)$, $L_u = L_u(t, \kappa \mid \zeta^{t_1}(\cdot))$ и $\gamma_u = \gamma_u(\kappa \mid \zeta^t(\cdot))$; $L_0 : R^n \to R^1$; $L_u, \gamma_u \in \Delta^{s,1}(u)$, таковы, что для любой $\zeta^{t_1}(\cdot) \in \Xi(t_1, u)$ функции $l_0 = L_0(\kappa)\gamma_u[t_1, \kappa]$, $l_u = L_u[\tau, \kappa]\gamma_u[\tau, \kappa]$ интегрируемы соответственно на $\mathscr{X}(u, \zeta^{t_1}(\cdot))$ и трубке $X(u, \zeta^{t_1}(\cdot))$,

$$X(u, \zeta^{t_1}(\cdot)) = \bigcup(\{t\} \times \mathscr{X}(u, \zeta^t(\cdot)) \mid t \in [t_0, t_1](\cdot)).$$

Задача 2.1. Найти допустимое управление $u^* = u^*(\eta^t(\cdot))$ для которого $J(u^*) \leqq J(u)$ при всех допустимых $u = u(\eta^t(\cdot))$.

В дальнейшем используется вероятностная интерпретация, основанная на сопоставлении структуры задач 2.1 и стохастического управления с неполным наблюдением [4]. При этом $I(u, \zeta^{t_1}(\cdot))$ является аналогом условного ожидания интегрального функционала относительно наблюдаемого сигнала. Множитель $\gamma_u = \gamma_u(\kappa \mid \zeta^t(\cdot))$ соответствует плотности условного распределения вероятности.

Определение 2.3. Оптимальной минимаксной оценкой состояния системы (2.1) в момент t называется вектор $\hat{x} = \hat{x}(u, \zeta^t(\cdot))$, $\hat{x} \in R^n$, удовлетворяющий условию

$$\max_x \{\|x - \hat{x}\| \mid x \in \mathscr{X}(u, \zeta^t(\cdot))\} =$$

$$\min_z \max_x \{\|x - z\| \mid x, z \in \mathscr{X}(u, \zeta^t(\cdot))\}.$$

Как известно [7], информационная область $\mathscr{X}(u, \zeta^t(\cdot))$ представляет собой эллипсоид,

$$\mathscr{X}(u, \zeta^t(\cdot)) =$$

$$= \{x : (x - \hat{x}(u, \zeta^t(\cdot)))^T P^{-1}(t)(x - \hat{x}(u, \zeta^t(\cdot))) \leqq \mu^2 - h^2(\zeta^t(\cdot))\}.$$

Здесь при всех u и $\zeta^{t_1}(\cdot)=(t_1, y(\cdot))\in\Xi(t_1, u)$ справедливо неравенство $h^2(\zeta^{t_1}(\cdot))\leqq\mu^2$, и функции $P(\cdot)$, $\hat{x}[\cdot]$, $h^2[\cdot]$ удовлетворяют системе

$$\dot{P}=AP+PA^T+CR^{-1}C^T-(PG^T+CR^{-1}D^T)\times$$
$$\times(DR^{-1}D^T)^{-1}(PG^T+CR^{-1}D^T)^T,\ P(t_0)=M^{-1}, \tag{2.7}$$

$$\dot{\hat{x}}=A\hat{x}+Bu+C\bar{v}+Sw[t],\quad \hat{x}[t_0]=\bar{x}, \tag{2.8}$$

$$h^2[t]=\int_{t_0}^{t} w^T[\tau]\,(D(\tau)R^{-1}(\tau)D^T(\tau))^{-1}w[\tau]\,d\tau,$$

где

$$S(t)=(P(t)G^T(t)+C(t)R^{-1}(t)D^T(t))\,(D(t)R^{-1}(t)D^T(t))^{-1},$$

$$w[t]=y(t)-G(t)\hat{x}[t]-D(t)\bar{v}(t).$$

Уравнения (2.7)–(2.8) по форме совпадают с соотношениями фильтра Калмана–Бьюси при коррелированных помехах. При этом функция $w[\cdot]$ является аналогом стохастического процесса обновления [4]. В силу (2.1)–(2.2) и (2.8) ее значения полностью определяются заданием пары $(x_0, v(\cdot))$.

Лемма 2.1. Пусть $u=u(\eta^t(\cdot))$ произвольное допустимое управление. Тогда существует оператор ψ_u, определяющий взаимно однозначное соответствие между элементами $v(\cdot)\in W$,

$$W=\{v(\cdot):\int_{t_0}^{t_1} v^T(\tau)\,(D(\tau)R^{-1}(\tau)D^T(\tau))^{-1}v(\tau)\,d\tau\leqq\mu^2\},$$

и позициями $\zeta^{t_1}(\cdot)\in\Xi(t_1, u)$. Причем, если $v_i(\cdot)\in W$, $i=1,\ 2$, таковы, что $v_1^\theta(\cdot)=v_2^\theta(\cdot)$, $\theta\in[t_0, t_1]$, то $\zeta_1^\theta(\cdot)=\zeta_2^\theta(\cdot)$, где $\zeta_i(\cdot)=\psi_u(v_i(\cdot))$.

Действительно, в силу свойств допустимого управления интегральное уравнение Вольтерра 2-го рода

$$z(\tau)-\int_{t_0}^{\tau} G(\tau)\Phi(\tau, s)B(s)u(\{s, z^s(\cdot)\})\,ds+$$
$$+\int_{t_0}^{\tau} G(\tau)\Phi(\tau, s)S(s)z(s)\,ds=f(\tau), \tag{2.9}$$

$\tau\in[t_0, t_1]$, при любой правой части $f(\cdot)\in L_2^k[t_0, t_1]$ имеет единственное решение $z_f(\cdot)\in L_2^k[t_0, t_1]$. Здесь $\Phi(t,\tau)$ — фундаментальная матрица для системы $\dot{z}=(A(t)-S(t)G(t))z$. Кроме того, можно показать, что W совпадает со множеством всех реализаций функции $w[\cdot]$. Поэтому искомое отображение ψ_u определяется условием $\psi_u(v(\cdot))=(t_1, z_v(\cdot))$, где $z_v(\cdot)$ — решение уравнения (2.9) при $f(\cdot)= =v(\cdot)+\bar{f}(\cdot)$,

$$\bar{f}(t)=G(t)\Phi(t,t_0)\bar{x}+D(t)\bar{v}(t)+$$

$$+\int_{t_0}^{t} G(t)\Phi(t,\tau)(C(\tau)-S(\tau)D(\tau))\bar{v}(\tau)\,d\tau.$$

Отметим, что аналогичное рассуждение справедливо, если соотношение (2.9) рассматривается на интервале $[t_0, \theta]$.

Приведенное утверждение позволяет при каждом заданном допустимом управлении $u=u(\eta^t(\cdot))$ использовать в качестве функциональной позиции наряду с парой $\zeta^t(\cdot)\in\Xi(t,u)$ элемент $v^t(\cdot)$, $v(\cdot)\in W$. Множество W от выбора управления не зависит. При этом с помощью соответствующей суперпозиции отображений $u=u(\eta^t(\cdot))$ и $\psi_u=\psi_u(v(\cdot))$ допустимое управление можно определить в виде оператора $\bar{u}=\bar{u}(v^t(\cdot))$.

Далее предполагается, что произведен переход к позициям $v^t(\cdot)$. При этом сохраняются ранее принятые обозначения. В частности, справедливо соотношение

$$J(u)=J(\bar{u})=\sup\{I(\bar{u}, v^{t_1}(\cdot))\,|\,v^{t_1}(\cdot)(\cdot)\in W\}.$$

Кроме того, в силу вида информационной области $\mathscr{X}(u,\zeta^t(\cdot))$ функционал $J(u)$ можно рассматривать для системы (2.8), моделирующей минимаксную оценку. Поэтому, если решение $u^*=u^*(\eta^t(\cdot))$ задачи 2.1 допускает представление

$$u^*[t]=\bar{\psi}(t,\hat{x}[t]),$$

где

$$\hat{x}[t]=\hat{x}(u,\zeta^t(\cdot)),\quad \bar{\psi}:[t_0,t_1]\times R^n\to R^m,$$

то в данном случае выполнено свойство разделения задач управления и наблюдения.

3. Принцип разделения

В дальнейшем наряду с 2.1 рассматривается следующая задача для системы с полным составом измерения.

Задача 3.1. Среди всех допустимых управлений $\bar{u}=\bar{u}(v^t(\cdot))$ найти $\bar{u}_*= =\bar{u}_*(v^t(\cdot))$ такое, что $\bar{J}(\bar{u}_*)\leqq\bar{J}(\bar{u})$ $\forall\bar{u}$, где

$$\bar{J}(u)=\sup_{W^0}\Bigg\{\int\limits_{\Sigma(v^{t_1}(\cdot))} L_0(\kappa+z[t_1])\gamma_u(\kappa+z[t_1]\,|\,v^{t_1}(\cdot))\,d\kappa+$$

$$+\int_{t_0}^{t_1}\int\limits_{\Sigma(\overline{v}(\cdot))} L_u(\tau,\kappa+z[\tau]\,|\,v^{t_1}(\cdot))\gamma_u(\kappa+z[\tau]\,|\,v^{\tau}(\cdot))\,d\kappa\,d\tau\Bigg\}, \tag{3.1}$$

$W^0 = W \backslash \partial W$ — внутренность множества W. Эллипсоид $\Sigma(v^t(\cdot)) \subseteq R^n$ и функция $z[\cdot]$, $z[t] = z(\bar{u}, v^t(\cdot))$, при каждом $v(\cdot) \in W$ определяются соотношениями

$$\Sigma(v^t(\cdot)) = \{\kappa : \kappa^T P^{-1}(t)\kappa \leqq \mu^2 - h^2(v^t(\cdot))\}\,, \tag{3.2}$$

$$\dot{z} = A(t)z + B(t)\bar{u} + C(t)\bar{v}(t) + S(t)v(t), \quad z(t_0) = \bar{x}\,. \tag{3.3}$$

Согласно лемма 2.1 системы (3.3) и (2.8) эквивалентны, следовательно, $z(\bar{u}, v^t(\cdot)) = \hat{x}(\bar{u}, v^t(\cdot))$, и $\bar{J}(\bar{u}) \leqq J(\bar{u})$. Действительно, элементы $v(\cdot) \in W^0$ соответствуют сигналам, не допускающим точного определения состояния системы (2.1) и вырождения областей $\Sigma(v^t(\cdot))$, $\mathscr{X}(\bar{u}, v^t(\cdot))$.

Далее приведены конкретные условия на функционалы L_0, L_u, γ_u, при которых значения управлений u^* и $\bar{u}_*$, решающих задачи 2.1 и 3.1, совпадают и определяются условием

$$\bar{u}_*[t] = \hat{\psi}(t, \hat{x}[t])\,. \tag{3.4}$$

Причем оператор $\hat{\psi}: [t_0, t_1] \times R^n \to R^m$ имеет вид

$$\hat{\psi}(t, z) = -Q_2^{-1}(t)B^T(t)Q(t)z\,, \tag{3.5}$$

$$-\dot{Q} = A^T Q + QA + Q_1 - QBQ_2^{-1}B^T Q, \quad Q(t_1) = Q_0\,. \tag{3.6}$$

Здесь $Q_0 \geqq 0$, $Q_1(t) \geqq 0$, $Q_2(t) > 0$ при всех $t \in [t_0, t_1]$ и $Q_1(\cdot)$, $Q_2(\cdot)$ — непрерывные матричные функции. В силу формул Коши и (2.4) соотношение (3.4) определяет допустимое управление, поэтому в данном случае можно формулировать принцип разделения.

Предположение 3.1. Пусть при всех $u = u(v^t(\cdot))$ и $(\tau, \kappa) \in [t_0, t_1] \times R^n$ на W справедливо представление

$$L_0(\kappa) = \kappa^T Q_0 \kappa\,,$$

$$L_u(\tau, \kappa | v^{t_1}(\cdot)) = \kappa^T Q_1(\tau)\kappa + \lambda_u(\tau | v^{t_1}(\cdot))\,,$$

$$\gamma_u(\kappa | v^\tau(\cdot)) = \Gamma(\kappa - \hat{x}[\tau] | v^\tau(\cdot)), \qquad \hat{x}[\tau] = \hat{x}(\bar{u}, v^\tau(\cdot))\,.$$

Заданные на W неотрицательные функционалы $\lambda_u = \lambda_u(\tau | v^{t_1}(\cdot))$ и $\Gamma = \Gamma(\kappa | v^\tau(\cdot))$ таковы, что $\lambda_u = \lambda_u[\cdot] \in L_1[t_0, t_1]$, и при каждой $v^{t_1}(\cdot) \in W^0$ п.в. на $[t_0, t_1]$

$$\int\limits_{\Sigma(v^\tau(\cdot))} \Gamma(\kappa | v^\tau(\cdot))\, d\kappa = 1\,.$$

Приведенные условия позволяют за счет выбора λ_u и Γ определять функционалы L_u и γ_u. При этом жестко оговаривается характер зависимости γ_u

от допустимого управления. Однако, задавая конкретный вид Γ, можно получать аналоги различных условных плотностей. В частности, если Γ:

$$\Gamma(\kappa \,|\, v^t(\cdot)) = \chi^n(v^t(\cdot))\,\Gamma_0^{-1}(t)\exp\{-\chi(v^t(\cdot))\kappa^T P^{-1}(t)\kappa\},$$

где $v^{t_1}(\cdot) \in W^0$, $\chi(v^t(\cdot)) = (\mu^2 - h^2(v^t(\cdot)))^{-1}$,

$$\Gamma_0(t) = \int\limits_{\{z : z^T P^{-1}(t) z \leqq 1\}} \exp\{-z^T P^{-1}(t) z\}\, dz,$$

то функционал γ_u на W^0 близок по структуре к плотности нормального распределения. Кроме того, для каждого $v_*^{t_1}(\cdot) \in \partial W$ существует момент $t_* \in [t_0, t_1]$ такой, что

$$\begin{gathered} \forall \bar{t} \in [t_0, t_*)\; \exists \bar{v}(\cdot) \in W^0 : \bar{v}^{\bar{t}}(\cdot) = v_*^{\bar{t}}(\cdot) \\ \&(v_*^t(\cdot)) = \{0\} \quad \forall t \in [t_*, t_1]\,. \end{gathered} \tag{3.7}$$

Поэтому на интервале $[t_0, t_*)$ значения $\gamma_u(\kappa \,|\, v_*^\tau(\cdot))$ определяются согласно (2.5), а вне его могут быть произвольными.

Лемма 3.1. Если выполнено предположение 3.1, то всюду на W^0 справедливо представление

$$\begin{aligned} I(\bar{u}, v^{t_1}(\cdot)) &= \hat{x}^T[t_1] Q_0 \hat{x}[t_1] + 2\hat{x}^T[t_1] Q_0 Z(v^{t_1}(\cdot)) + \alpha(v^{t_1}(\cdot)) \\ &+ \int\limits_{t_0}^{t_1} \{\hat{x}^T[t] Q_1(t) \hat{x}[t] + 2\hat{x}^T[t] Q_1(t) Z(v^t(\cdot)) + \lambda_u(t \,|\, v^{t_1}(\cdot)\}\, dt, \end{aligned}$$

где $\hat{x}[t] = \hat{x}(\bar{u}, v^t(\cdot))$,

$$Z(v^t(\cdot)) = \int\limits_{\Sigma(v^t(\cdot))} \kappa \Gamma(\kappa \,|\, v^t(\cdot))\, d\kappa \in R^n,$$

$$\begin{aligned} \alpha(v^{t_1}(\cdot)) &= \int\limits_{\Sigma(v^{t_1}(\cdot))} \kappa^T Q_0 \kappa \Gamma(\kappa \,|\, v^{t_1}(\cdot))\, d\kappa + \\ &+ \int\limits_{t_0}^{t_1} \int\limits_{\Sigma(v^\tau(\cdot))} \kappa^T Q_1(\tau) \kappa \Gamma(\kappa \,|\, v^\tau(\cdot))\, d\kappa\, d\tau\,. \end{aligned}$$

Отметим, что если при некотором $v^{t_1}(\cdot) \in W^0$ функция $\Gamma = \Gamma[\tau, \kappa]$ четная по $\kappa \in R^n$, то $Z(v^\tau(\cdot)) = 0$, и минимаксную оценку $\hat{x}(\bar{u}, v^\tau(\cdot))$ можно истолковать как условное среднее вектора состояния.

Определим на W операторы $\Lambda_0 = \Lambda_0(\tau, \xi \,|\, v^{t_1}(\cdot))$ и $\varphi = \varphi(\tau \,|\, v^{t_1}(\cdot))$, $(\tau, \xi) \in [t_0, t_1] \times R^m$, связанные соотношением

$$\begin{gathered} \Lambda_0(\tau, \xi \,|\, v^{t_1}(\cdot)) = (\xi - Q_2^{-1}(\tau) B^T(\tau) \varphi(\tau \,|\, v^{t_1}(\cdot))^T \\ Q_2(\tau)\,(\xi - Q_2^{-1}(\tau) B^T(\tau) \varphi(\tau \,|\, v^{t_1}(\cdot)))\,. \end{gathered} \tag{3.8}$$

Здесь при каждом $v^{t_1}(\cdot)\in W$ функция $\varphi=\varphi[\tau]$ удовлетворяет уравнению

$$-\dot{\varphi}=A^T\varphi+Q(C\bar{v}+Sv)+Q_1Z[t],\quad \varphi[t_1]=Q_0Z[t_1],$$

где $Z[t]=Z(v^t(\cdot))$, и $Q(\cdot)$ — решение (3.6).

Предположение 3.2. Пусть существует функционал $\Lambda=\Lambda(\tau,\xi|v^{t_1}(\cdot))$, $(\tau,\xi)\in[t_0,t_1]\times R^m$, такой, что при любом $v^{t_1}(\cdot)\in W$ и п.в. $\tau\in[t_0,t_1]$

1) для каждого допустимого управления $\bar{u}=\bar{u}(v^t(\cdot))$

$$\lambda_u(\tau|v^{t_1}(\cdot))=\Lambda(\tau,\bar{u}[\tau]|v^{t_1}(\cdot)),$$

2) функция $s_\tau(\cdot)$, $s_\tau(\xi)=\Lambda[\tau,\xi]-\Lambda_0[\tau,\xi]$, достигает минимального по $\xi\in R^m$ значения в т. $\xi_*=\bar{u}_*[\tau]$ (3.4),

3) при любом $v_*^{t_1}(\cdot)\in\partial W$ (3.7) выполняется условие

$$\forall\varepsilon>0\exists\delta>0\forall\tilde{t}\in[t_*-\delta,t_*)$$
$$\exists v_\varepsilon(\cdot)\in W^0: v_\varepsilon^{\tilde{t}}(\cdot)=v_*^{\tilde{t}}(\cdot)\&$$
$$|\Lambda(\tau,\xi|v_\varepsilon^{t_1}(\cdot))-\Lambda(\tau,\xi|v_*^{t_1}(\cdot))|<\varepsilon\hat{\Lambda}(\tau,\xi|v_*^{t_1}(\cdot)).$$

Здесь $\hat{\Lambda}=\hat{\Lambda}(\tau,\xi|v^{t_1}(\cdot))$ такой, что при всех $\bar{u}=\bar{u}(v^t(\cdot))$ функция $\hat{\Lambda}=\hat{\Lambda}[\tau,\ \bar{u}[\tau]]$ интегрируема на $[t_0,\ t_1]$.

Отметим, что для $\Lambda(\tau,\ \xi|v^{t_1}(\cdot))=\Lambda_0(\tau,\ \xi|v^{t_1}(\cdot))$ (3.8) предположение 3.1 влечет 3.2.

Теорема 3.1. Пусть выполнены предположения 3.1–3.2. Тогда $\bar{u}_*=\bar{u}_*(v^t(\cdot))$ (3.4) решает задачу 3.1.

Доказательство. Рассмотрим на множестве W^0 функционал

$$V(\bar{u},\ v^\tau(\cdot))=\hat{x}^T(\bar{u},\ v^\tau(\cdot))Q(\tau)\hat{x}(\bar{u},\ v^\tau(\cdot)),$$

$\tau\in[t_0,t_1]$. Здесь $Q(\cdot)$ — решение уравнения Риккати (3.6). Интегрируя полную производную функции $V=V[t]$ на $[t_0,\ t_1]$, получим для $I(\bar{u},\ v^{t_1}(\cdot))$, $v(\cdot)\in W^0$, следующее представление

$$I(\bar{u},\ v^{t_1}(\cdot))=\bar{x}^TQ(t_0)\bar{x}+2\bar{x}^T\varphi[t_0]+\beta(v^{t_1}(\cdot))+$$
$$+\int_{t_0}^{t_1}(\bar{u}[t]-\hat{\psi}(t,\hat{x}[t]))^TQ_2(t)(\bar{u}[t]-\hat{\psi}(t,\hat{x}[t]))\,dt+$$
$$+\int_{t_0}^{t_1}(\Lambda[t,\bar{u}[t]]-\hat{\Lambda}_0[t,\bar{u}[t]])\,dt,$$

где $\bar{u}[t]=\bar{u}(v^t(\cdot))$, $\hat{x}[t]=\hat{x}(\bar{u},\ v^t(\cdot))$, и оператор $\hat{\psi}$ определяется соотношением (3.5). Поэтому

$$I(\bar{u},v^{t_1}(\cdot))\geqq I(\bar{u}_*,v^{t_1}(\cdot)),\quad v^{t_1}(\cdot)\in W^0.$$

Условие 3) предположения 3.2 в приведенном рассуждении не использовалось.

Лемма 3.2. Если выполняются предположения 3.1–3.2, то при всех $\bar{u} = \bar{u}(v^t(\cdot))$ верно равенство $J\bar{u}) = \bar{J}(u)$.

Действительно, неравенство $J(\bar{u}) > \bar{J}(\bar{u})$ невозможно. Иначе, найдется $v_*(\cdot) \in \partial W$ такой, что

$$I(\bar{u}, v_*(\cdot)) = \sup\{I(\bar{u}, v(\cdot)) \mid v(\cdot) \in W^0\} + \alpha, \quad \alpha > 0 .$$

Однако, в силу (3.7), условия 3) и абсолютной непрерывности интеграла Лебега от функции $\rho(\cdot)$,

$$\rho(\tau) = \int\limits_{\mathscr{X}(\bar{u}, v_*^{\tau}(\cdot))} L_u(\tau, \kappa \mid v_*^{t_1}(\cdot))\gamma_u(\kappa \mid v_*^{\tau}(\cdot))\, d\kappa ,$$

существует $\bar{v}(\cdot) \in W^0$ такой, что $I(\bar{u}, \bar{v}(\cdot)) > I(\bar{u}, v_*(\cdot)) - \dfrac{\alpha}{2}$.

Таким образом, задачи 2.1 и 3.1 эквивалентны и можно сформулировать следующее утверждение.

Теорема 3.2. (Принцип разделения). Пусть для линейной системы (2.1)–(2.2) и критерия качества (2.6) выполняются предположения 3.1–3.2. Тогда существует допустимое управление $u^* = u^*(\zeta^t(\cdot))$, решающее задачу 2.1, и его текущие значения определяются условием

$$u^*[t] = \hat{\psi}(t, \hat{x}[t]), \qquad t \in [t_0, t_1] .$$

Здесь $\hat{x}[t] = \hat{x}(u^*, \xi^t(\cdot))$ — оптимальная минимаксная оценка состояния системы (2.1) в момент t по измеренной реализации сигнала (2.2). Оператор $\hat{\psi}$ вида (3.5) определяет управление по принципу обратной связи, решающее задачу 3.2 с полным наблюдением. Кроме того,

$$J(u^*) = \min_{u} \sup_{\Xi(t_1, u)} I(u, \zeta^{t_1}(\cdot)) = \sup_{W} \min_{\bar{u}} I(\bar{u}, v^{t_1}(\cdot)) .$$

Автор благодарит А. Б. Куржанского за постановку задачи и постоянное внимание к работе.

Литература

1. *Красовский Н. Н.* Игровые задачи о встрече движений. М., Наука, 1970, 420 с.
2. *Куржанский А. Б.* Управление и наблюдение в условиях неопределенности. М., Наука, 1977, 392 с.
3. *Куржанский А. Б.* Динамические задачи принятия решений в условиях неопределенности. В кн.: Современное состояние теории исследования операций (Под редакцией Н. Н. Моисеева). М., Наука, 1979, с. 197–235.
4. *Флеминг У., Ришел Р.* Оптимальное управление детерминированными и стохастическими системами. М., Мир, 1978, 314 с.
5. *Брайсон, А., Хо Ю-Ши.* Прикладная теория оптимального управления. М., Мир, 1972, 544 с.
6. *Wonham, W. M.*, On the separation theorem of stochastic control. SIAM J. Control, 1968, vol. **6,** No. *2*, pp. 312–326.
7. *Кругликов С. В.* О разделении задач управления и наблюдения в условиях неопределенности. Дифференц. уравн., 1985, т. **21,** № *3*, с. 398–404.

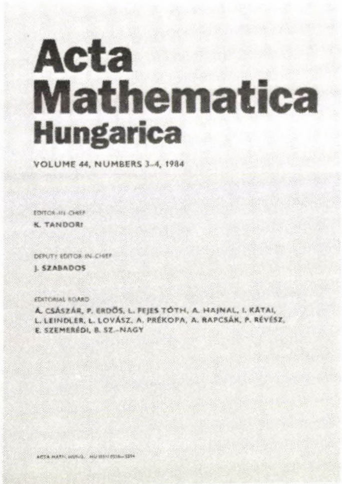
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda, Budapest

Acta Mathematica Hungarica

(Formerly: Acta Mathematica Academiae Scientiarum Hungaricae)

Editor in Chief:
K. Tandori

Deputy Editor in Chief:
J. Szabados



Acta Mathematica Hungarica

VOLUME 44, NUMBERS 3–4, 1984

EDITOR-IN-CHIEF
K. TANDORI

DEPUTY EDITOR-IN-CHIEF
J. SZABADOS

EDITORIAL BOARD
A. CSÁSZÁR, P. ERDŐS, L. FEJES TÓTH, A. HAJNAL, I. KÁTAI, L. LEINDLER, L. LOVÁSZ, A. PRÉKOPA, A. RAPCSÁK, P. RÉVÉSZ, E. SZEMERÉDI, B. SZ.-NAGY

ACTA MATH. HUNG. HU ISSN 0236–5294

The journal covers a wide scope in the field of mathematics. It comprises theory of sets, mathematical logic, classical and modern analysis, algebra, number theory, geometry, topology, combinatorics, mathematical statistics, probability theory, as well as information theory.

Founded 1950
Papers in English, German, French and Russian
Publication: two volumes annually — one volume contains two issues
Price per volume: $ 44.00; DM 99,—
Size: 17 × 25 cm
ISSN 0236–5294

Order form
to be returned to
KULTURA
Hungarian Foreign Trading Company
P.O. Box 149, H-1389 Budapest, Hungary

□ Please enter my/our subscription for ACTA MATHEMATICA HUNGARICA for one year

□ Please enter my/our standing order for ACTA MATHEMATICA HUNGARICA starting with

Name: ____________________

Address: ____________________

Date and signature: ____________________

Contents of Volume 42. Numbers 1–2

Invitation for papers
Manuscripts should be sent to
Acta Mathematica Hungarica
P.O. Box 127
H-1364 Budapest
Hungary

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

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). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as diag $(d_1 , d_2 , \; . \; . \; . \; d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 129090 Москва И-90, ул. Щепкина, 8. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 208-95-01).

Объём 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 языке в трех экземплярах, 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 и 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 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

HU ISSN 0370-2529 Index: 26 660

VOL. / ТОМ 15 · NUMBER / НОМЕР 3

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 СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1986

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
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория ипформации;
- теория ипформации в области сетей с множественным доступом.

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 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1986 Subscription Rate US Dollars 175.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

1828

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

Problems of Control and Information Theory, Vol. 15 (3), pp. 203—218 (1986)

CONTROL WITH INFORMATION DEFICIT

N. N. Krasovskii, S. I. Tarasova, V. E. Tret'jakov, G. I. Shishkin

(Sverdlovsk)

(Received June 5, 1985)

An algorithm of optimal control of dynamic system in the presence of dynamic and informative disturbances is presented.

1. Introduction

We consider some class of minimum problems of optimal control according to the feedback principle for ensured result under conditions of uncontrollable dynamic disturbance and in the presence of deficit of information on initial and current phase states of the controlled object. The deficit lies in the fact that we have not information on the phase vector over all coordinates and, moreover, this information is distorted. It is shown that for the considered class of problems the algorithm developed in [1] with complete information on phase states of an object is preserved.

2. Formulation of the problem

In [1] two of the present authors solved the minimum problem of optimal control for the ensured result over the index

$$\gamma_x(x[\cdot], u[\cdot], v[\cdot]) = \\ = |x[\vartheta]| + \int_{t_0}^{\vartheta} (u'[t]\Phi(t)u[t] - v'[t]\Psi(t)v[t])dt \tag{2.1}$$

for the object described by the differential equation

$$\dot{x} = A(t)x + B(t)u + C(t)v, \qquad t_0 \leqq t \leqq \vartheta. \tag{2.2}$$

Here the phase vector $x \in R^n$, the control $u \in R^r$ and the unknown dynamic disturbance $v \in R^s$ are treated as column vectors; the prime denotes transposition; $u'\Phi(t)u$, $v'\Psi(t)v$ are positive definite quadratic form with piecewise continuous coefficients; $A(t)$, $B(t)$, $C(t)$, $t_0 \leqq t \leqq \vartheta$ are piecewise continuous matrix-valued functions, t_0 and ϑ are fixed instants of time; $|x|$ is the Euclidean norm of vector x. The realizations

Akadémiai Kiadó, Budapest

of control $u[\cdot]=\{u[t], t_0<t\leqq\vartheta\}$ and disturbance $v[\cdot]=\{v[t], t_0<t\leqq\vartheta\}$ are measurable and bounded (each in its own way).

In [1] the optimal control $u^0[t]$ was formed according to the feedback principle on the basis of precisely measured current position $\{t_i, x[t_i]\}$ at any needed instant t_i. In the present paper we assume that the data on phase vector $x[t]$ are fed to us only over part of the coordinates $p\leqq n$ and, besides, this information is distorted. Namely, the current information on the state of x-object (2.2) is delivered by the p-dimensional information variable

$$q[t]=K(t)x[t]+\Delta q[t], \tag{2.3}$$

where $K(t)$, $t_0\leqq t\leqq\vartheta$ is a given piecewise continuous $(p\times n)$ matrix-function, $\Delta q[t]$ is an information disturbance. We shall also accept that the initial phase state $x_0=x[t_0]$ is informed with the distortion Δx_0 in the form of false value $x_0^*=x_0+\Delta x_0$.

Assuming that memorizing of the values $q[t]$ (2.3) and control $u[t]$ generated by us is possible, we concentrate all informations obtained in the instant t in the information element $Y[t]$ consisting of three components

$$Y[t]=\{x_0^*,\ q[t_0[\cdot]t],\ \tilde{y}[t_0[\cdot]t]\} \tag{2.4}$$

where $q[t_0[\cdot]t]=\{q[v],\ t_0\leqq v\leqq t\}$ is the information history which we suppose to be piecewise continuous;

$$\tilde{y}'[t_0[\cdot]t]=\{y'[t_0[\cdot]t],\ \tilde{y}_{n+1}[t_0[\cdot]t]\}$$

is known $n+1$-dimensional controlled vector-valued function evolving in accordance with the differential equations

$$\dot{y}[t]=X[\vartheta, t]B(t)u[t], \qquad \dot{\tilde{y}}_{n+1}[t]=u'[t]\Phi(t)u[t], \tag{2.5}$$

$$y[t_0]=\{0, \ldots, 0\}, \qquad \tilde{y}_{n+1}[t_0]=0. \tag{2.6}$$

Here $X[\vartheta, t]$, $t_0\leqq t\leqq\vartheta$ is the solution of matrix differential equation

$$dX[\vartheta, t]/dt=-X[\vartheta, t]A(t)$$

with the boundary condition $X[\vartheta, \vartheta]=E$, where E is the unit matrix. In the considered formulation the information state $\{t, Y[t]\}$ will play the same role that the current position $\{t, x[t]\}$ played in [1]. Therefore, the next definitions repeat the definitions from [1] with replacement of $\{t, x[t]\}$ by $\{t, Y[t]\}$.

Namely, an arbitrary function

$$u(\cdot)=\{u(t, Y, \varepsilon),\quad t_0\leqq t\leqq\vartheta,\quad \varepsilon>0\} \tag{2.7}$$

which is determined for all possible values of the information element Y is called a strategy $u(\cdot)$.

Collection of three components

$$U=\{u(\cdot),\ \varepsilon,\ \Delta\{t_i\}\}, \tag{2.8}$$

where $\Delta\{t_i\}$, $t_1=t_*$, $t_i<t_{i+1}$, $i=1,\ldots,l$, $t_{l+1}=\vartheta$ is the division of segment $[t_*,\vartheta]$ is called control law U for the segment $[t_*,\vartheta]\subset[t_0,\vartheta]$. Absolutely continuous solution of the stepwise differential equation

$$\begin{gathered}\dot{x}[t]=A(t)x[t]+B(t)u(t_i,\ Y[t_i],\varepsilon)+C(t)v[t],\\ t_i\leqq t<t_{i+1},\qquad i=1,\ldots,l\end{gathered} \tag{2.9}$$

with initial state $x[t_*]=x_*=x[t_0[t_*]t_*]$ unknown to us is called the motion

$$x[t_*[\cdot]\vartheta]=\{x[t],\ t_*\leqq t\leqq\vartheta\}$$

of the given x-object (2.2) generated by the control law U (2.8) and some unknown disturbance $v(t_*[\cdot]\vartheta]$. Here

$$x[t_0[\cdot]t_*]=\{x[v],\ t_0\leqq v\leqq t_*\}$$

is the history of the x-object motion which is formed to the instant $t=t_*$ in some way or other, the x-object starting from the true initial state $\{t_0,\ x_0\}$ unknown to us.

Side by side with real motion $x[t_*[\cdot]\vartheta]$, law (2.8) forms an imaginary one of the controlled component $\tilde{y}[t]$ in system (2.5), (2.6) as a solution of the stepwise differential equations

$$\dot{y}[t]=X[\vartheta,t]B(t)u(t_i,\ Y[t_i],\varepsilon) \tag{2.10}$$

$$\dot{y}_{n+1}[t]=u'(t_i,\ Y[t_i],\varepsilon)\Phi(t)u(t_i,\ Y[t_i],\varepsilon) \tag{2.11}$$

at known initial condition $\tilde{y}[t_*]=\tilde{y}_*=\tilde{y}[t_0[t_*]t_*]$ which is determined by information element $Y[t_*]$.

Thus, side by side with real motion $x[t_0[\cdot]\vartheta]$ unknown to us from the beginning to the end, the imaginary one $Y[t_0[\cdot]\vartheta]$ of the information Y-system is developed, and the states of this system are characterized by the information element $Y[t_i]$ which is known to us at any needed instant $t=t_i$. We shall consider that the information history $q[t_*[\cdot]t_i]$ being a part of the element $Y[t_i]$ is formed by some second player irrespective of our purposes. In this situation we take the role of the first player who forms control

$$u[t]=u[t_i]=u(t_i,\ Y[t_i],\varepsilon),\quad t_i<t\leqq t_{i+1},\quad i=1,\ldots,l$$

with the purpose of ensuring of the smallest possible value of the following quality index for the information Y-system

$$\gamma_*(Y[\vartheta]) = \sup_{\{x_0,\, v[\cdot]\}} [\gamma_x(x[\cdot], u[\cdot], v[\cdot]) - \\ - \int_{t_0}^{\vartheta} (\Delta q[t])' Q(t) \Delta q[t] dt - (\Delta x_0)' P \Delta x_0]. \tag{2.12}$$

Here $(\Delta q)' Q(t) \Delta q$ and $(\Delta x_0)' P \Delta x_0$ are positive definite quadratic forms, and for some x_0, $v[\cdot]$ and known $Y[\vartheta]$ values $x[t]$, $t_0 \leqq t \leqq \vartheta$ are determined by the Cauchy formula

$$x[t] = X[t, t_0]x_0 + X[t, \vartheta]y[t] + \\ + \int_{t_0}^{t} X[t, v]C(v)v[v]dv. \tag{2.13}$$

Further formalization of minimum problem for the ensured result with respect to the index γ_* (2.12) for the information Y-system with complete information on its states $Y[t]$, $t_0 \leqq t \leqq \vartheta$ follows a general plan developed for abstract Y-systems in [2] and realized for specific cases in [1–4]. In this problem a conditional second player deals with choosing of continuations $q(t_*[\cdot]\vartheta]$ of the information history $q[t_0[\cdot]t_*]$ which is known to the instant t_* (the second player may act with intentions unfavorable for us in the sense of index γ_*). Therefore, in accordance with [2] the number

$$\rho^0(t_*, Y[t_*]) = \min_{u(\cdot)} [\overline{\lim_{\varepsilon \to 0}} \lim_{\delta \to 0} \sup_{U_\delta} \sup_{q(t_*[\cdot]\vartheta]} \gamma_*(Y[\vartheta])] \tag{2.14}$$

is called optimal ensured result for information state $\{t_*, Y[t_*]\}$. Here, the least upper bounds are calculated over all possible continuations $q(t_*[\cdot]\vartheta]$ of information history $q[t_0[\cdot]t_*]$, which is a component of the known information element $Y[t_*]$ and according to all laws U_δ corresponds to a chosen strategy $u(\cdot)$, fixed ε and divisions $\Delta\{t_i\}$ such that $t_{i+1} - t_i \leqq \delta$.

Strategy $u^0(\cdot)$, which is supplied minimum in (2.14), is called an optimal one. In the same way as in [1–4], it is proved that such strategy exists. What guarantee will the strategy $u^0(\cdot)$ supply for us during the control of the real x-object based on it, whose motion evolves in accordance with equations (2.9) at $u(\cdot) = u^0(\cdot)$? It follows from the form of indices γ_x (2.1), γ_* (2.12) and definition of $\rho^0(t_*, Y[t_*])$ that value $\rho^0(t_*, Y[t_*])$ (2.14) is optimal ensured result for the real x-object over the index

$$\gamma(x_0^*, q[t_0[\cdot]t_*], u(t_0[\cdot]\vartheta]; \quad x_0, v(t_0[\cdot]\vartheta], q(t_*[\cdot]\vartheta]) = \\ = |x[\vartheta]| + \int_{t_0}^{\vartheta} (u'[t]\Phi(t)u[t] - v'[t]\Psi(t)v[t])dt - \int_{t_0}^{\vartheta} (q[t] - \\ - K(t)x[t])' Q(t) (q[t] - K(t)x[t]) dt -- (x_0^* - x_0')' P (x_0^* - x_0) \tag{2.15}$$

with respect to combined disturbance $\{x_0, v(t_0[\cdot]\vartheta], q(t_*[\cdot]\vartheta]\}$ which is chosen by the second player irrespective of our purposes. Strictly speaking it means the following

Theorem 2.1 For any arbitrarily small number $\zeta > 0$ one can find a number $\varepsilon(\zeta)$ and value $\delta(\varepsilon, \zeta) > 0$ so that at $t_* \in [t_0, \vartheta)$, $u(\cdot) = u^0(\cdot)$, $\varepsilon \leqq \varepsilon(\zeta)$, $\delta \leqq \delta(\varepsilon, \zeta)$ the control law U_δ will form in system (2.9) such motion $x[t_*[\cdot]\vartheta]$ that for any initial state x_0, any possible disturbance $v(t_0[\cdot]\vartheta]$ and any possible continuation $q(t_*[\cdot]\vartheta]$ of the information history $q[t_0[\cdot]t_*]$ the following inequality holds

$$\gamma(x_0^*, q[t_0[\cdot]t_*], u^0(t_0[\cdot]\vartheta]; \quad x_0, v(t_0[\cdot]\vartheta), q(t_*[\cdot]\vartheta]) \leqq \rho + \zeta \tag{2.16}$$

whatever the initial state $\{t_*, Y[t_*]\} = \{x_0^*, q[t_0[\cdot]t_*], \tilde{y}[t_0[\cdot]t_*]\}$ is, and the value $\rho = \rho^0(t_*, Y[t_*])$ is the least one from the numbers ρ which satisfy such conditions. Note that in index γ (2.15) the weight matrices $\Phi(t)$, $\Psi(t)$, $K(t)$, $Q(t)$ and P determine the variation limits of realizations of optimal control $u^0(t_*[\cdot]\vartheta]$ and the disturbances $\{x_0, v(t_0[\cdot]\vartheta], q(t_*[\cdot]\vartheta]\}$ most unfavourable for us, not permitting them to be arbitrarily large. Negative terms in (2.15) estimate compensation which the second player gives us for the dynamic disturbance formed by him and the distortion of information on the current and initial states $\{t, x[t]\}$ of the real x-object. Term $|x[\vartheta]|$ in (2.15) can be treated as a penalty imposed on the second player for deviation of the x-object from the origin of coordinates at instant $t = \vartheta$, and the second positive term can be considered as a cost of resources spent on the control process.

3. Basic result

Thus, it is required to construct control law $U_\delta^0 = \{u^0[t_i] = u^0(t_i, Y[t_i], \varepsilon), i = 1, \ldots, l\}$ optimal with respect to index γ_* (2.12) for information Y-system introduced in Section 2. In accordance with theorem 2.1 for sufficiently small ε and δ such control law U_δ^0 will ensure the best result for the x-object in the sense of index γ (2.15).

If we use the method of extremal displacement of a system onto concomitant point described for an abstract Y-system in Section 7 of paper [2] and utilized for problem (2.1), (2.2) in case of $Y[t] = x[t]$ in [1], then we obtain the following algorithm in the considered problem (see, for instance, formula (7.14) in [1] on p. 19)

$$u^0[t_i] = -\frac{1}{2}\Phi^{-1}(t_i)B'(t_i)X'[\vartheta, t_i]m^0[t_i], \tag{3.1}$$

where

$$\begin{aligned} m^0[t_i] = m^0(t_i, Y[t_i], \varepsilon) = \arg\{\max_{|m| \leq 1} [-\eta[t_i](1 + |m|^2)^{1/2} + \\ + m'h(t_i, Y[t_i]) + m'(F(t_i) - \lambda(t_i)E)m]\}. \end{aligned} \tag{3.2}$$

Here $h(t_i, Y[t_i])$ is an n-dimensional vector calculated in the course of receipt of information. $F(t)$ is an $(n \times n)$-dimensional matrix which can be computed a priori with respect to initial data of the problem for all $t \in [t_*, \vartheta]$ at once;

$$\lambda(t_i) = \max_{t_i \leqq t \leqq \vartheta} \max_{|a|=1} a'F(t)a, \tag{3.3}$$

$$\eta(t_i) = [\varepsilon + \varepsilon(t_i - t_0)]^{1/2} > 0. \tag{3.4}$$

Taking into account (3.3) and (3.4) we see that problem (3.2) has the unique solution $m^0[t_i]$ like a maximum problem from strictly concave function.

The optimal ensured result is calculated according to the formula

$$\begin{aligned}\rho^0(t_*, Y[t_*] = \max_{|m| \leq 1} [m'h(t_*, Y[t_*]) + \\ + m'(F(t_*) - \lambda(t_*)E)m + \chi(t_*, Y[t_*])],\end{aligned} \tag{3.5}$$

where $\chi(t_*, Y[t_*])$ is some number.

Structure of relations (3.1)-(3.5) is typical for sufficiently wide range of problems. Only components of the information element $Y[t]$ and concrete expressions for the values $F(t)$, $h(t, Y[t]\}$, $\chi(t, Y[t])$ will be different in these relations. For example, for problems (2.1), (2.2) with $Y[t] = x[t]$ which was solved in [1], we have

$$h(t, x[t] \equiv X[\vartheta, t]x[t], \tag{3.6}$$

$$\begin{aligned}F(t) = -\frac{1}{4}\int_t^\vartheta (X[\vartheta, \tau][B(\tau)\Phi^{-1}(\tau)B'(\tau) - \\ - C(\tau)\Psi^{-1}(\tau)C'(\tau)]X'[\vartheta, \tau])d\tau.\end{aligned} \tag{3.7}$$

To obtain the expressions of values $h(t, Y[t])$ and $F(t)$ (see Section 5) participating in algorithm (3.1), (3.2) it is sufficient to have concrete expression for linearly quadratic function $H(t_*, Y[t_*], m)$ with respect to m standing under the maximum sign in formula (3.5). But in accordance with the method of stochastic program synthesis [1-4] the optimal ensured result $\rho^0(t_*, Y[t_*])$ (3.5) coincides with the value of stochastic program maximum which is constructed in a proper way. Therefore let us pass to consideration of this value.

4. Stochastic program maximin

Later on instead of variable $q[t]$ (2.3) it is convenient to consider the following observed variable

$$r[t]=q[t]-K(t)\int_{t_0}^{t} X[t,v]B(v)u[v]dv=$$
$$=q[t]-K(t)X[t,\vartheta]y[t], \tag{4.1}$$

whose history $r[t_0[\cdot]t]=\{r[v], t_0\leqq v\leqq t\}$ at current instant of time t is known to us according to the conditions of the problem.

To the information Y-system with replacement in (2.4) of the component $q[t_0[\cdot]t]$ by $r[t_0[\cdot]t]$ we assign a stochastic Z-model. This model is constructed as follows. Let us fix an instant $\tau_*\in[t_0,\vartheta)$ and some initial history $r[t_0[\cdot]\tau_*]$ of the observed variable $r[t]$ (4.1). We set a division $\Delta\{\tau_j\}, j=1,\ldots,k$ for the segment $[\tau_*,\vartheta]$, $\tau_1=\tau_*$, $\tau_j<\tau_{j+1}$, $\tau_k=\vartheta$. Let us assume the probability space $\{\Omega,\mathscr{B},\mathscr{P}\}$ where Ω is a k-dimensional unit cube

$$\Omega=\{\omega=\{\xi_1,\ldots,\xi_k\}, 0\leqq\xi_j<1, j=1,\ldots,k\},$$

$\mathscr{B}$ is a Borel σ-algebra, $\mathscr{P}$ is a Lebesgue measure [5], as a basis of stochastic program construction. The nonanticipating functions ([5], p. 100)

$$u_*(\tau,\omega)=u_*[\tau,\xi_1,\ldots,\xi_j],\quad \tau_j<\tau\leqq\tau_{j+1},\quad j=1,\ldots,k-1 \tag{4.2}$$

$$r_*(\tau,\omega)=r_*[\tau,\xi_1,\ldots,\xi_j],\quad \tau_j<\tau\leqq\tau_{j+1},\quad j=1,\ldots,k-1 \tag{4.3}$$

are called stochastic programs.

State of the Z-model at a current instant $\tau\geqq\tau_*$ is defined by the phase element

$$Z(\tau,\omega)=\{x_0^*, r[t_0[\cdot]\tau;\omega], \tilde{z}(\tau,\omega)\} \tag{4.4}$$

where the component $r[t_0[\cdot]\tau;\omega]=\{r(v,\omega), t_0\leqq v\leqq\tau\leqq\vartheta, \omega\in\Omega\}$ at $\tau=\tau_*$ coincides with the determinated history $r[t_0[\cdot]\tau_*]$ and at $\tau>\tau_*$ its random realizations $r(\tau,\omega)$ are determined by the program $r_*(\tau,\omega)$ (4.3). The vectorial $n+1$-dimensional component

$$\tilde{z}'(\tau,\omega)=\{z'(\tau,\omega),\quad \tilde{z}_{n+1}(\tau,\omega)\}=\tilde{z}'[\tau,\xi_1,\ldots,\xi_j],$$
$$\tau_j<\tau\leqq\tau_{j+1},\quad j=1,\ldots,k-1$$

from $Z(\tau,\omega)$ (4.4) is defined as a solution of equations

$$\dot{z}(\tau,\omega)=X[\vartheta,\tau]B(\tau)u_*(\tau,\omega),$$
$$\dot{\tilde{z}}_{n+1}(\tau,\omega)=u'_*(\tau,\omega)\Phi(\tau)u_*(\tau,\omega) \tag{4.5}$$

at some initial state $\tilde{z}(\tau_*,\omega)=\tilde{z}_*$.

In stochastic variant phase element $Z(\tau, \omega)$ (4.4) imitates information element $Y[\tau]$, and below introduced n-dimensional random variable $w(\cdot)=\{w(\omega), \omega \in \Omega\}$ and s-dimensional random function $v(\cdot)=\{v(\tau, \omega), t_0<\tau \leqq \vartheta, \omega \in \Omega\}$ imitate the disturbances x_0 and $v(t_0[\cdot]\vartheta]$ from the Y-system.

The quantity

$$e(\tau_*, Z[\tau_*], \Delta)= \sup_{\|l(\cdot)\| \leqq 1} \varkappa(\tau_*, Z[\tau_*], \Delta, l(\cdot)) \tag{4.6}$$

is called a program extremum. Here

$$\Delta=\Delta\{\tau_j\}, \quad Z[\tau_*]=Z(\tau_*, \omega)=$$
$$=\{x_0^*, r[t_0[\cdot]\tau_*], \tilde{z}_*\}; \quad l(\cdot)=\{l(\omega), \omega \in \Omega\}$$

is an n-dimensional random variable, $\|l(\cdot)\|=(M\{|l(\omega)|^2\})^{1/2}$, M is the mathematical expectation and the value $\varkappa=\varkappa(\tau_*, Z[\tau_*], \Delta, l(\cdot))$ is determined by the relation

$$\begin{gathered}\varkappa= \sup_{r_*(\cdot)} \inf_{u_*(\cdot)} \sup_{v(\cdot)} \sup_{w(\cdot)} \sigma(\tau_*, Z[\tau_*], \Delta, l(\cdot); \\ r_*(\cdot), u_*(\cdot), v(\cdot), w(\cdot)).\end{gathered} \tag{4.7}$$

Here the function σ is constructed according to the form of index γ (2.15) with regard to the introduced stochastic analogies in the following way.

$$\begin{aligned}\sigma= & M\{l'(\omega)\,[X[\vartheta, t_0]w(\omega)+z_*+\int_{\tau_*}^{\vartheta} X[\vartheta, \tau]B(\tau)u_*(\tau, \omega)d\tau+ \\ & +\int_{t_0}^{\vartheta} X[\vartheta, \tau]C(\tau)v(\tau, \omega)d\tau]+\tilde{z}_{*n+1}+\int_{\tau_*}^{\vartheta} u_*'(\tau, \omega)\Phi(\tau)u_*(\tau, \omega)d\tau- \\ & -\int_{t_0}^{\vartheta} [v'(\tau, \omega)\Psi(\tau)v(\tau, \omega)+[r(\tau, \omega)-K(\tau)\,(X[\tau, t_0]w(\omega)+ \\ & +\int_{t_0}^{\vartheta} X[\tau, v]C(v)v(v, \omega)dv)]'Q(\tau)\,[r(\tau, \omega)-K(\tau)\,(X[\tau, t_0]w(\omega)+ \\ & +\int_{t_0}^{\tau} X[\tau, v]C(v)v(v, \omega)dv)]]d\tau-[x_0^*-w(\omega)]'P[x_0^*-w(\omega)].\end{aligned} \tag{4.8}$$

The connection between the optimal ensured result $\rho^0(t_*, Y[t_*])$ (2.14) and program extremum (4.6) is established by the following theorem.

Theorem 4.1. If in (4.6) we set

$$\tau_* = t_*, \quad \tilde{z}_* = \tilde{y}_*, \quad r[v] = q[v] - K(v)X[v, \vartheta]y[v]$$

at $t_0 \leqq v \leqq \tau_*$, then

$$\rho^0(t_*, Y[t_*]) = \sup_{\Delta} e(\tau_*, Z[\tau_*], \Delta). \tag{4.9}$$

Thus, to search vector $h(t_*, Y[t_*])$ and matrix $F(t_*)$ it is necessary to bring the expression of stochastic program maximin (4.9) to form (3.5)

5. Calculation of program extremum

At fixed τ_*, $Z[\tau_*]$, Δ, $l(\cdot)$ we compute variations of the value σ (4.8) with respect to $w(\cdot)$, $v(\cdot)$, $u_*(\cdot)$ and $r_*(\cdot)$. Making the partial variations with respect to $w(\cdot)$ and $v(\cdot)$ equal to zero, we obtain the following two integral equations correspondingly.

$$\begin{gathered} X'[\vartheta, t_0]l(\omega) + 2\int_{t_0}^{\vartheta} X'[\tau, t_0]K'(\tau)Q(\tau)\,[r(\tau, \omega) - \\ - K(\tau)\,(X[\tau, t_0]w(\omega) + \int_{t_0}^{\tau} X[\tau, v]C(v)v(v, \omega)dv)]d\tau + \\ + 2P(x_0^* - w(\omega)) = 0, \qquad \omega \in \Omega \end{gathered} \tag{5.1}$$

$$\begin{gathered} C'(\tau)\,(X'[\vartheta, \tau]l(\omega) + 2\int_{\tau}^{\vartheta} X'[\eta, \tau]K'(\eta)Q(\eta)\,[r(\eta, \omega) - \\ - K(\eta)\,(X[\eta, t_0]w(\omega) + \int_{t_0}^{\eta} X[\eta, v]C(v)v(v, \omega)dv)]d\eta - \\ - 2\Psi(\tau)v(\tau, \omega) = 0, \qquad t_0 \leqq \tau \leqq \vartheta, \quad \omega \in \Omega. \end{gathered} \tag{5.2}$$

Setting the partial variations of the value σ (4.8) with respect to nonanticipating programs $u_*(\cdot)$ (4.2), $r_*(\cdot)$ (4.3) equal to zero we obtain the following relations which are valid for $\tau_j < \tau \leqq \tau_{j+1}$, $j = 1, \ldots, k-1$

$$u[\tau, \xi_1, \ldots, \xi_j] = -\frac{1}{2}\Phi^{-1}(\tau)B'(\tau)X'[\vartheta, \tau]M\{l(\omega)/\xi_1, \ldots, \xi_j\} \tag{5.3}$$

$$\begin{gathered} r[\tau, \xi_1, \ldots, \xi_j] = K(\tau)\,(X[\tau, t_0]M\{w(\omega)/\xi_1, \ldots, \xi_j\} - \\ - \int_{t_0}^{\tau} X[\tau, v]C(v)M\{v(v, \omega)/\xi_1, \ldots, \xi_j\}dv). \end{gathered} \tag{5.4}$$

Here, the symbol $M\{\ \ldots / \ldots\ \}$ stands for conditional expectation. Equations (5.1)–(5.4) express the necessary extremal conditions for the value σ (4.8). These equations have the unique solution with respect to $w(\cdot)$, $v(\cdot)$, $u_*(\cdot)$, $r_*(\cdot)$. We assume that the extremal elements $w^0(\cdot)$, $v^0(\cdot)$, $u_*^0(\cdot)$, $r_*^0(\cdot)$, which are the solution of equations (5.1)–(5.4), are found in some way or other. Substituting these elements into the expression for σ (4.8) we obtain some linearly quadratic functional of $l(\cdot)$. In addition one can obtain like in [1] that the upper bound in problem (4.6) is achieved in the class of random variables $l(\cdot)$ of the form

$$l(\omega)=l[\xi]=m+a[\xi], \quad \xi\in[0,1), \quad M\{a[\xi]\}=0. \tag{5.5}$$

Going over to calculation of stochastic program maximin in (4.9) we get that the least upper bound is achieved, like in [1, 2], on the division $\Delta:\tau_1=\tau_*$, $\tau_2=\tilde{\tau}$, $\tau_3=\vartheta$, $\tau_*\leqq\tilde{\tau}\leqq\vartheta$ where instant τ is determined from the condition

$$\max_{|a|=1} a'F(\tilde{\tau})a=\max_{\tau_*\leqq\tau\leqq\vartheta}\ \max_{|a|=1} a'F(\tau)a. \tag{5.6}$$

Here $F(\tau)$ is the same matrix that figures in (3.2), (3.3), (3.5). As a result we come to the needed equality

$$\sup_{\Delta} e(\tau_*, Z[\tau_*], \Delta)=\max_{|m|\leqq 1} H(\tau_*, Z[\tau_*], m) \tag{5.7}$$

where in accordance with Theorem 4.1 the function H will be just the same function linearly quadratic in m which stands under the maximum sign in formula (3.5).

Thus, the problem of determination of extremal elements from equations (5.1)–(5.4) turns out to be a central computing problem when searching vector $h(t, Y[t])$ and matrix $F(t)$. Analysis of these equations falls outside the limits of the present paper. Equations similar to (5.1)–(5.4) are investigated in [4] in the special case when $C(t)=c(t)$ is an n-dimensional vector, $\Psi(t)=\psi(t)$ is a scalar and the dynamic disturbance is a scalar function represented by cut off Fourier series

$$v[t]=\sum_{s=1}^{N}\alpha_s g_s(t)=\alpha' g(t), \qquad t_0\leqq t\leqq\vartheta, \tag{5.8}$$

where $\alpha'=\{\alpha_1, \ldots, \alpha_N\}$ is an arbitrary numerical vector (unknown to the first player). In this special case, the expression for the matrix $F(t)$ has the form

$$\begin{aligned} F(t)=-\frac{1}{4}\Big[\int_t^{\vartheta} X[\vartheta,\tau]B(\tau)\Phi^{-1}(\tau)B'(\tau)X'[\vartheta,\tau]d\tau-\\ -G[\vartheta,t_0]W^{-1}(t)G'[\vartheta,t_0]. \end{aligned} \tag{5.9}$$

Here

$$G[\vartheta,t_0]=(V'[\vartheta,t_0], X[\vartheta,t_0]), \tag{5.10}$$

$$V[\vartheta, t_0] = \int_{t_0}^{\vartheta} g(t)c'(t)X'[\vartheta, t]dt \tag{5.11}$$

and matrix $W(t)$ has the following block structure

$$W(t) = \begin{pmatrix} W_{11}(t) & W_{12}(t) \\ W'_{12}(t) & W_{22}(t) \end{pmatrix} \tag{5.12}$$

where

$$\begin{aligned} W_{11}(t) = & \int_{t_0}^{\vartheta} \psi(t)g(t)g'(t)dt + \\ & + \int_{t_0}^{t} V[v, t_0]K'(v)Q(v)K(v)V'[v, t_0]dv \end{aligned} \tag{5.13}$$

$$W_{12}(t) = \int_{t_0}^{t} V[v, t_0]K'(v)Q(v)K(v)X[v, t_0]dv \tag{5.14}$$

$$W_{22}(t) = \int_{t_0}^{t} X'[v, t_0]K'(v)Q(v)K(v)X[v, t_0]dv + P\,. \tag{5.15}$$

Expression for the vector $h(t, Y[t])$ has the form

$$\begin{aligned} h(t, Y[t]) = y[t] + G[\vartheta, t_0]W^{-1}(t)\Bigg[& \int_{t_0}^{t} G'[v, t_0] \times \\ & \times K'(v)Q(v)\,(q[v] - K(v)X[v, \vartheta]y[v])dv + P_0 x_0^* \Bigg] \end{aligned} \tag{5.16}$$

where $P_0 = (0, P)'$ and 0 is null $N \times n$-matrix.

In conclusion we notice that the problem of searching of extremal element has an auxiliary character. However, if various realizations of random extremal elements $w^0(\cdot)$, $v^0(\cdot)$, $r_*^0(\cdot)$ are fed to the information element $Y[t]$, $t \geqq t_*$ and on the real x-object (2.2) which is controlled by the optimal law U_δ^0 (3.1) with small ε and δ, then each time value of the index (2.15) will be obtained arbitrarily close to the value of optimal ensured result $\rho^0(t_*, Y[t_*])$. This remark can be found useful in computing experiments during tests on stability of the work of optimal control law U_δ^0.

6. Example

Let system (2.2) and index γ (2.15) have the form

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = u + v, \qquad 0 = t_0 \leqq t \leqq \vartheta = 3 \tag{6.1}$$

$$\begin{gathered}\gamma = |x[\vartheta]| + \int_{t_0}^{\vartheta} \varphi(t) u^2[t]dt - \int_{t_0}^{\vartheta} \psi(t) v^2[t]dt - \\ - \int_{t_0}^{\vartheta} p(q[t] - x_1[t])^2 dt - p[(x_{01} - x_{01}^*)^2 + (x_{02} - x_{02}^*)^2].\end{gathered} \tag{6.2}$$

The minimum problem of ensured result with respect to the index γ (6.2) for system (6.1) can be interpreted as a problem of straight-line motion of a controlled carriage under the action of a tractive force u and a wind force v from unknown initial state $x_0 = \{x_{01}, x_{02}\}$ into the origin $\{0, 0\}$ of the phase plane (x_1, x_2). To derive some benefit from the disturbance v the carriage is equipped with a wind generator producing energy whose cost is estimated by the first negative term in (6.2). The rest of negative terms in (6.2) estimate the compensation for instrument distortion of information on the current state $\{t, x[t]\}$ of the carriage and for falsely reported initial state $x_0^* = \{x_{01}^*, x_{02}^*\}$. The positive terms in (6.2) determine the penalty for inaccuracy of carriage arrival into the origin of coordinates and the cost of the energy which is spent to produce the tractive force u.

The optimal law U_δ^0 (3.1), (3.2) of carriage control was tested with the help of electronic computer in the case when the disturbance had the form $v[t] = \alpha_1 + \alpha_2 t$, where α_1 and α_2 are arbitrary numbers. The following parameters were chosen

$$\varphi(t) \equiv 0.25, \qquad \psi(t) \equiv 1, \qquad p = 1.0, 1.5, 2.5 \tag{6.3}$$

and the following false initial state is taken

$$x_0^* = \{x_{01}^* = -1, \quad x_{02}^* = 0\}. \tag{6.4}$$

Figure 1 shows the realization of carriage motion on the phase plane provided that the initial state x_0, numbers α_1 and α_2 and the observed instrument readings $q[t], 0 \leqq t \leqq 3$ were the most unfavourable for us in the sense of index γ (6.2), i.e. they were determined by some realizations of extremal elements $w^0(\cdot)$, $v^0(\cdot)$ and $r_*^0(\cdot)$. In Fig. 1 solid lines correspond (from the left to the right) to different values of p from (6.3). In addition we obtained the following values of the index $\gamma = 1.30, 0.85, 0.55$, each of them with chosen accuracy coincides with the value of the corresponding optimal ensured result $\rho^0(t_0, Y[t_0])$ which was calculated a priori by formula (3.5). Dotted line in Fig. 1 corresponds to $\rho = 2.5$ and initial state $x_0 = x_0^*$ (6.4) when other conditions are invariable. Here in accordance with the theory we obtained that $\gamma = 0.23 < \rho^0(t_0, Y[t_0]) = 0.55$. Besides, in Fig. 1 the realization $u^0[t], 0 \leqq t \leqq 3$ of optimal law of control and the

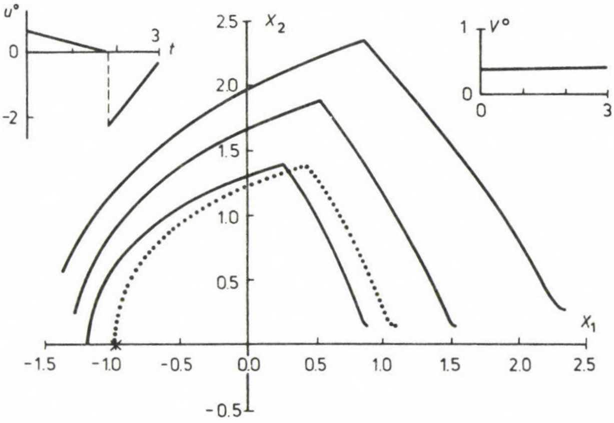


Fig. 1

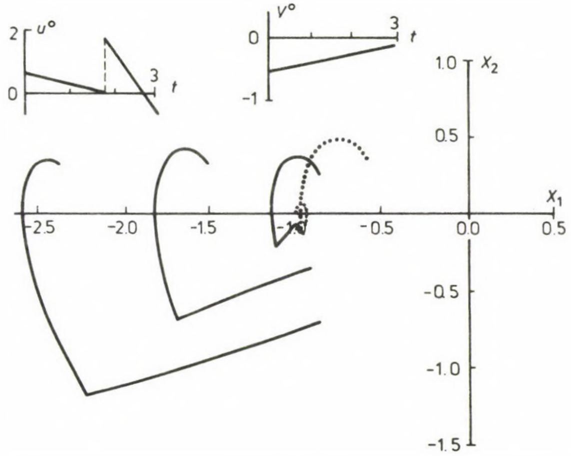


Fig. 2

realization $v^0[t]$, $0 \leqq t \leqq 3$ of extremal element $v^0(\cdot)$ obtained here are shown for $p=2.5$. Since extremal elements $w^0(\cdot)$, $v^0(\cdot)$, $r_*^0(\cdot)$ have random character, in the same situation as for Fig. 1, another picture is possible. It is shown in Fig. 2.

Fig. 3 shows time variation of the coordinate $x_1[t]$, $0 \leqq t \leqq 3$ (solid line) at $\varphi(t) \equiv 1$, $\psi(t) \equiv 10$, $p=2.5$ and at extremal values x_0^0, α_1^0, α_2^0 and $q^0[t]$, $0 \leqq t \leqq 3$. In Fig. 3, the

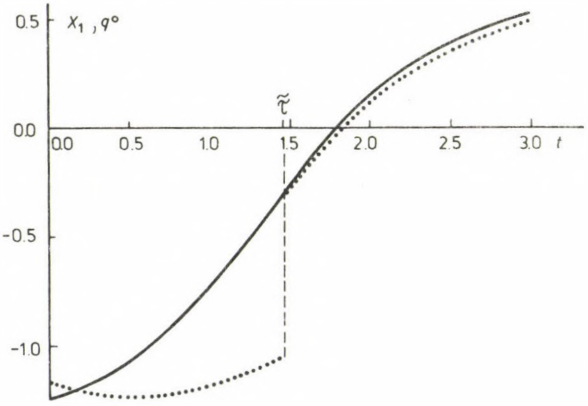


Fig. 3

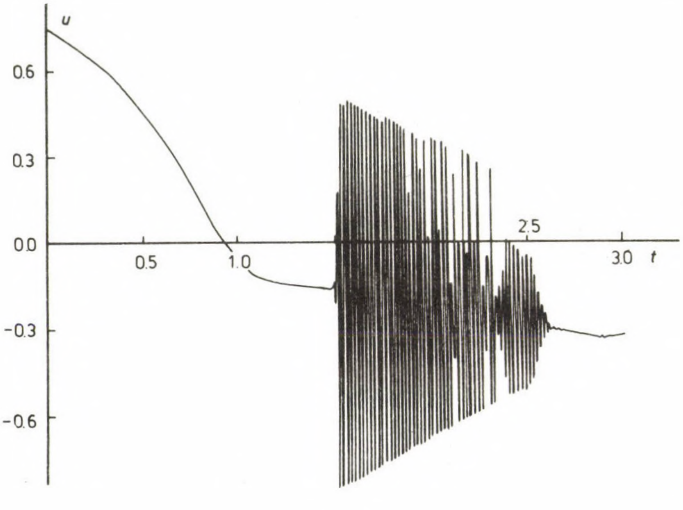


Fig. 4

dotted line represents extremal observed instrument readings $q^0[t]$ which on the segment $[\tilde{\tau}, \vartheta]$ coincided with the true values of coordinate $x_1[t]$. The situation shown in Fig. 4 stresses that in real calculation the accuracy parameter ε in the law U_δ^0, $\delta = t_{i+1} - t_i$ at fixed δ cannot be taken arbitrarily small because of the risk of appearance of sliding regime which is represented in Fig. 4. It happened in the present

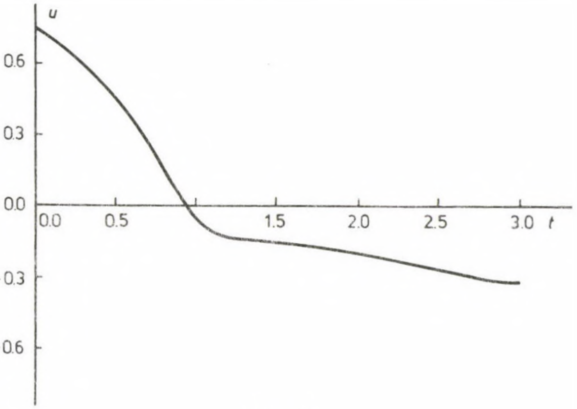


Fig. 5

example at

$$\varphi(t)\equiv 1,\quad \psi(t)\equiv 10,\quad p=2.5,\quad x_0=x_0^*=\{-1,0\},\quad q[t]=x_1[t],$$

$$0\leqq t\leqq 3,\quad \varepsilon=0.001,\quad \delta=0.01.$$

And at the expense of unwarranted increase of the second term in γ (6.2) we obtained that $\gamma=0.66>\rho^0(t_0, Y[t_0])=0.63$, although in accordance with the theory the opposite inequality must be fulfilled in this case. If at the same $\delta=0.01$ we take $\varepsilon=0.01$, then the sliding regime in realization of optimal control disappears (see Fig. 5) and we obtain

$$\gamma=0.37<\rho^0(t_0,\ Y[t_0])=0.63$$

as it should be.

References

1. *Krasovskii, N. N., Tret'jakov, V. E.*, One problem of optimal control on minimum of the ensured result. Izvestija Akad. Nauk SSSR. Tehn. Kibernetika, 1983, *2*, pp. 6–23 (in Russian).
2. *Krasovskii, N. N.*, Extremal aiming and extremal displacement in a game-theoretical control. Probl. Control and Inform. Theory, 1984, **13** *(5)*, pp. 287–302.
3. *Krasovskii, N. N.*, Control problem under conditions of incomplete information. Prikl. mat. meh., 1984, **48** *(4)*, pp. 533–539 (in Russian).
4. *Krasovskii, N. N., Tarasova, S. I., Tret'jakov, V. E., Shishkin, G. I.*, Control problem with incomplete information: Preprint Ural Scientific Center Akad. Nauk SSSR. Sverdlovsk, 1984, p. 63 (in Russian).
5. *Liptser, R. Sh., Shirjaev, A. N.*, Statistics of random processes. Moscow: Nauka, 1974, 696 pp. (in Russian); English transl. Vols. **I, II.** Springer-Verlag, 1977.

Управление при дефиците информации

Н. Н. КРАСОВСКИЙ, С. И. ТАРАСОВА, В. Е. ТРЕТЬЯКОВ, Г. И. ШИШКИН
(Свердловск)

В статье обсуждается задача оптимального управления на минимум гарантированного результата по показателю

$$\gamma = |x[\vartheta]| + \int_{t_0}^{\vartheta} (u'[t]\Phi(t)u[t] - v'[t]\Psi(t)v[t])\, dt$$

$$- \int_{t_0}^{\vartheta} (q[t] - K(t)x[t])' Q(t)(q[t] - K(t)x[t])\, dt -$$

$$- (x_0^* - x_0)' P(x_0^* - x_0)$$

для управляемой вектором u системы

$$\dot{x} = A(t)x + B(t)u + C(t)v, \quad t_0 \leqq t \leqq \vartheta, \quad x[t_0] = x_0,$$

подверженной влиянию неконтролируемой помехи v.

Начальное состояние x_0, текущая позиция $\{t, x[t]\}$ и помеха $v(t_0[\cdot]\vartheta]$ неизвестны в течение всего процесса. Закон управления U строится по принципу обратной связи на основе сведений о ложном начальном состоянии объекта x_0^*, истории некоторого информационного сигнала $q[\tau] = K(\tau)x[\tau] + \Delta q[\tau]$, $t_0 \leqq \tau \leqq t$ и истории вырабатываемого нами управления $u(t_0[\cdot]t]$. Влияние того или иного фактора на введенный показатель процесса характеризуется весовыми определенно-положительными матрицами $\Phi(t)$, $\Psi(t)$, $Q(t)$, P. Показано, что алгоритм оптимального управления совпадает с разработанным ранее в [1] для аналогичной задачи алгоритмом в случае точных сведений о начальной и текущих позициях объекта. При этом роль текущей позиции $\{t, x[t]\}$ в условиях дефицита информации исполняет состояние $\{t, Y[t]\}$ некоторой информационной y-системы, вбирающей в себя все доступные для наблюдения сведения о процессе.

Н. Н. Красовский
Институт математики и механики
Уральского научного центра АН СССР
СССР, 620219 Свердловск, ГСП-384
ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 15 (3), pp. 219—230 (1986)

INTEGRAL ESTIMATIONS OF ROBOTS MOBILE POSSIBILITIES

E. P. Popov, V. G. Tikhomirov, V. I. Ushakov

(*Moscow*)

(Received May 22, 1985)

Several new integral estimates of the robots mobile possibilities essentially extending and supplementing the integral service coefficient of manipulator in the service area are presented. The first group of integral estimates referred to the moment characteristics of the manipulator service distribution in the service area. The second group of estimates is based on the statistical interpretation of the manipulator functional properties in the service area. Two new functions for all integral characteristics calculations are introduced, namely, manipulator service distribution function and density of the manipulator service distribution function, in the service area.

1. In many tasks connecting with a robot control it is important to know the geometrical functional (or mobile) possibilities of manipulator, which essentially depend on the type of manipulator device kinematical structure, its constant geometrical parameters, its degrees of freedom limits, and mutual situation of robot and external obstacles existing in the robot service space.

The robot mobile possibilities usually mean the robot service envelope, in every point of which the segments of the service angle of the manipulator hand are defined. These segments are composed in totality by the manipulator service angle $\psi(x, y, z)$ [1]. The service coefficient in the point of the robot service envelope is a ratio

$$\theta(x, y, z) = \psi(x, y, z)/4\pi\,.$$

Such information is particularly valuable at the phase of planning the motion trajectories of manipulator device, so far as it gives a possibility to estimate a set of all permissible trajectories of manipulator terminal link.

Usually, the analysis of robots mobile possibilities is connected with a constructing of the quality pictures of the manipulator service distribution in its service envelope and a computing of some integral estimations of this distribution.

Now the most widespread integral estimation of the mobile possibilities of manipulators is an integral service coefficient. Let $\Omega \subset \mathbf{R}^3$ be a set of points belonging to the robot service envelope. Than the integral service coefficient may be defined,

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

according to [1], as

$$\Theta = \frac{1}{V} \iiint\limits_{\Omega} \theta(x, y, z) dx\, dy\, dz\,, \tag{1}$$

where

$$V = \iiint\limits_{\Omega} dx\, dy\, dz$$

s a volume of the manipulator service envelope Ω.

The integral service coefficient characterising an average mobility of manipulator in the service envelope is an important quantitative estimation of its mobile possibilities. It allows us to compare the mobile possibilities of manipulators having different kinematical structures and schemes. However, the integral service coefficient, which is an arithmetic mean of the distribution of $\theta(x, y, z)$ in Ω, by no means takes into account the shape of this distribution, although for some kinematical schemes of manipulators their functional possibilities in the middle of the service envelope and in its outlying area are essentially different. At the same time there are schemes of manipulators having sufficiently homogeneous distribution of $\theta(x, y, z)$ in Ω. Integral service coefficient doubtfully defines functional mobile possibilities of the first type manipulators, but for the second type manipulators it is practically reliable.

The purpose of this paper is a construction of some new integral estimations of the robots mobile possibilities, which will essentially define more precisely and will supply the integral service coefficient. These estimations will take into account the shape of the distribution of $\theta(x, y, z)$ in Ω. They may be considered as the original characteristics of "confidence" to value Θ too.

First of all, so far as a speech will be about characteristics of relative distribution of $\theta(x, y, z)$ in Ω, let us introduce two new functions describing this distribution quantitatively.

The function of the manipulator service distribution in the service envelope. Let this function be marked as $W = W(\vartheta)$ and defined as a relative volume of the service envelope part, in every point of which $\theta(x, y, z) \leqq \vartheta$. Thus, $W(\vartheta) = W(\theta \leqq \vartheta)$.

It is obvious that $\forall \vartheta \leqq 0$ $W(\vartheta) = 0$, so far as $\forall (x, y, z) \in \Omega$ $\theta(x, y, z) \geqq 0$. In turn, $\forall \vartheta \geqq 1$ $W(\vartheta) = 1$, so far as $\forall (x, y, z) \in \Omega$ $\theta(x, y, z) \leqq 1$. Thus, $0 \leqq W(\vartheta) \leqq 1$. Absolute volume of the service envelope part, in every point of which $\theta(x, y, z) \leqq \vartheta$, is $U(\vartheta) = V W(\vartheta)$.

It is necessary to note that $W(\vartheta)$ is a non-descending function. It may be proved: the service envelope part in every point of which $\theta(x, y, z) \leqq \vartheta$ is completely contained in the service envelope part, where $\theta(x, y, z) \leqq \vartheta + \Delta\vartheta\, (\Delta\vartheta \geqq 0)$, that is $W(\theta \leqq \vartheta) \leqq$

$\leqq W(\theta \leqq \vartheta + \Delta\vartheta)$. Relative volume of the service envelope part in which $\vartheta \leqq \theta(x, y, z) \leqq \leqq \vartheta + \Delta\vartheta$, is

$$\Delta W(\vartheta \leqq \theta \leqq \vartheta + \Delta\vartheta) = W(\vartheta + \Delta\vartheta) - W(\vartheta).$$

Additionally, it may be said about function $W(\vartheta)$ that in the common case $W(\vartheta)$ may be a discontinuous function having finite set of gap points of the first kind. On the sections of continuity, function $W(\vartheta)$ is a differentiable everywhere except in the common case, in a finite number of points. Exemplary graph of function $W(\vartheta)$ is shown in Fig. 1.

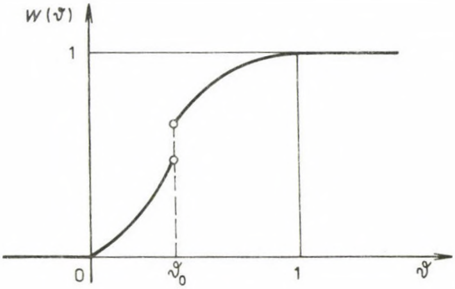


Fig. 1

The function of the density of the manipulator service distribution in the service envelope. Let this function be marked as $\omega(\vartheta)$ and be defined as

$$\omega(\vartheta) = \lim_{\Delta\vartheta \to 0} \frac{W(\vartheta + \Delta\vartheta) - W(\vartheta)}{\Delta\vartheta} = \lim_{\Delta\vartheta \to 0} \frac{\Delta W(\vartheta)}{\Delta\vartheta}. \tag{2}$$

If $\vartheta = \vartheta_0$ is a gap point of the first kind of function $W(\vartheta)$, or a point where function $W(\vartheta)$ is not differentiable, it is nessesary to calculate the left and right limits (2) separately. These limits exist, but, in the common case, are not equal to each other.

Function $\omega(\vartheta) \geqq 0$, so far as $W(\vartheta)$ is a non-descending function. It is obvious that $\forall\vartheta \leqq 0$ and $\forall\vartheta \geqq 1$ $\omega(\vartheta) = 0$. Proceeding from the definition of function $\omega(\vartheta)$, it may be written

$$W(\vartheta_1) = \int_0^{\vartheta_1} \omega(\vartheta) d\vartheta.$$

Also it is obvious that

$$\int_0^1 \omega(\vartheta) d\vartheta = W(1) - W(0) = 1 - 0 = 1.$$

Exemplary graph of function $\omega(\vartheta)$ is shown in Fig. 2.

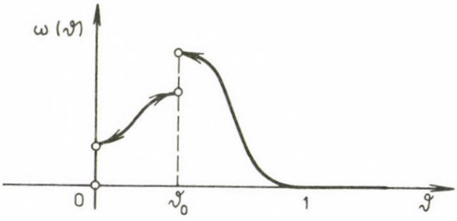


Fig. 2

2. Examined functions $W(\vartheta)$ and $\omega(\vartheta)$ are very convenient, because they characterise distribution and density of distribution of manipulator service in the service envelope in nondimensional form. These functions do not depend on the sizes of region Ω and its shape, and therefore they give a possibility to compare functional properties of different kinematical schemes of manipulator devices. At the same time, as it will be showed in the further account, by means of functions $W(\vartheta)$ and $\omega(\vartheta)$ it is very convenient to compute different robots mobile possibilities integral estimations having a form of

$$\frac{1}{V}\iiint\limits_{\Omega}\varphi(\theta)dx\,dy\,dz\,,$$

where $\varphi(\theta)$ is a continuous function of $\theta(x, y, z)$ argument, defined in the region Ω.

Let us show in a general form that

$$\frac{1}{V}\iiint\limits_{\Omega}\varphi(\theta)dx\,dy\,dz=\int\limits_{0}^{1}\varphi(\vartheta)\omega(\vartheta)d\vartheta\,. \tag{3}$$

For this purpose let the triple integral standing in the left part of equality (3) be replaced by the limit of the integral sum. In this case, region Ω is supposed to be divided into N equal parts having a volume ΔV:

$$\frac{1}{V}\iiint\limits_{\Omega}\varphi(\theta)dx\,dy\,dz=\lim_{\substack{\Delta V\to 0\\ N\to\infty}}\frac{1}{V}\sum_{i=1}^{N}\varphi(\theta_i)\Delta V. \tag{4}$$

Let then be divided the interval $\theta\in[0;1]$ into M equal parts and sort $\varphi(\theta_i)$ values standing under the symbol of sum in expression (4) into received categories, according to the values of arguments θ_i, taking into account, that m_j of elements of region Ω get into the category number j. Summary number of elements must be unchanged

$$\sum_{j=1}^{M}m_j=N\,.$$

Then

$$\lim_{\substack{\Delta V\to 0\\ N\to\infty}} \frac{1}{V}\sum_{i=1}^{N} \varphi(\theta_i)\Delta V = \lim_{\substack{\Delta V\to 0\\ M\to\infty}} \frac{1}{V}\sum_{j=1}^{M} \varphi(\theta_j)m_j\Delta V =$$

$$= \lim_{\substack{\Delta V\to 0\\ M\to\infty}} \sum_{j=1}^{M} \varphi(\theta_j)\frac{\Delta V_j}{V} = \lim_{\substack{\Delta W\to 0\\ M\to\infty}} \sum_{j=1}^{M} \varphi(\theta_j)\Delta W(\theta_j) =$$

$$= \lim_{\substack{\Delta\theta\to 0\\ M\to\infty}} \sum_{j=1}^{M} \varphi(\theta_j)\omega(\theta_j)\Delta\theta = \int_0^1 \varphi(\vartheta)\omega(\vartheta)d\vartheta.$$

Thus, formula (3) is proved. Let us find with its help an expression for integral service coefficient believing, that $\varphi(\theta)=\theta(x, y, z)$:

$$\Theta = \frac{1}{V}\iiint_{\Omega} \theta(x, y, z)dx\,dy\,dz = \int_0^1 \vartheta\omega(\vartheta)d\vartheta. \qquad (5)$$

Expression (5), which is a convenient form for Θ computing, may be transformed by means of function $W(\vartheta)$:

$$\Theta = \int_0^1 \vartheta\omega(\vartheta)d\vartheta = \int_0^1 \vartheta\, dW(\vartheta) = \vartheta W(\vartheta)|_0^1 - \int_0^1 W(\vartheta)d\vartheta = 1 - \int_0^1 W(\vartheta)d\vartheta. \qquad (6)$$

According to formula (6), integral service coefficient is an area of figure, limited by straights: $x=0$, $x=1$, $y=1$ and by graph of function $W=W(\vartheta)$.

3. Let us introduce some new integral estimations of the robots mobile possibilities defining more precisely and supplying integral service coefficient. By analogy to mechanics and probabilities theory [2] these integral estimations will be defined through so-called moments, so further let them be named as moment characteristics.

Dispersion and root-mean-square deviation of the manipulator service in service envelope relative to the arithmetical mean. Dispersion of the manipulator service in the service envelope is

$$R = \frac{1}{V}\iiint_{\Omega} (\theta(x, y, z) - \Theta)^2 dx\,dy\,dz = \int_0^1 (\vartheta - \Theta)^2\omega(\vartheta)d\vartheta.$$

It is defined as a central moment of the second order and characterises the dispersion of the service coefficient $\theta(x, y, z)$ in the service envelope Ω.

The smaller the value of R, the larger "confidence" is entailed by integral service coefficient.

Root-mean-square deviation of the service coefficient relative to the arithmetical mean may be defined as

$$r=\sqrt{R}\,.$$

Asymmetry coefficient of distribution of the manipulator service in the service envelope. This coefficient is defined by the central moment of the third order

$$S=\frac{1}{r^3}\frac{1}{V}\iiint\limits_{\Omega}(\theta(x,y,z)-\Theta)^3dx\,dy\,dz=\frac{1}{r^3}\int\limits_0^1(\vartheta-\Theta)^3\omega(\vartheta)d\vartheta$$

and characterises (with sign and value) the asymmetry of distribution of $\theta(x, y, z)$ in Ω, which is particularly obviously displaying on the graph of function $\omega=\omega(\vartheta)$. In Fig. 3 two asymmetrical distributions $\omega_1(\vartheta)$ and $\omega_2(\vartheta)$ are shown. One of them (ω_1) has a positive asymmetry $(S_1>0)$, another one (ω_2) has a negative asymmetry $(S_2<0)$. Absolute values of $|S_1|$ and $|S_2|$ are equal.

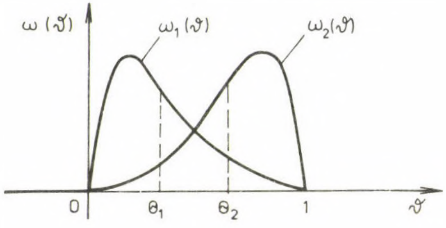


Fig. 3

Excess of the manipulator service distribution in the service envelope. This value characterises sharptoping or planetoping of graph of function $\omega=\omega(\vartheta)$ and is defined through the central moment of the fourth order

$$E=\frac{1}{r^4}\frac{1}{V}\iiint\limits_{\Omega}(\theta(x,y,z)-\Theta)^4dx\,dy\,dz=\frac{1}{r^4}\int\limits_0^1(\vartheta-\Theta)^4\omega(\vartheta)d\vartheta\,.$$

Most sharptoping distribution of $\omega=\omega(\vartheta)$ have the largest value of E. In Fig. 4, three graphs of $\omega=\omega(\vartheta)$ are shown. They have identical arithmetical mean, but most "confidence" to this value is entailed by the first curve, because $E_1>E_2>E_3$.

Thus, the introduced integral estimations of the robot mobile possibilities allow us essentially to define more precisely the distribution shape of $\theta(x, y, z)$ in Ω (dispersion, asymmetry, excess), and give possibility to compare different distributions having identical values of integral service coefficient.

4. Now let us consider one integral estimation of the robot mobile possibilities which in the most general form takes into account the shape of the distribution of $\theta(x, y, z)$ in Ω a *shape coefficient of the distribution of $\theta(x, y, z)$ in Ω.*

In [3] the integral service coefficient is interpreted as a probability of the next situation: in the accidental point of the service envelope the terminal link of manipulator may be arbitrary oriented. Such statistical interpretation, according to [3], permits us to connect functional properties of manipulator, expressed by Θ, with a probability of the mobile tasks execution.

In this paper original development of statistical interpretation of the manipulator service considered in [3] will be given.

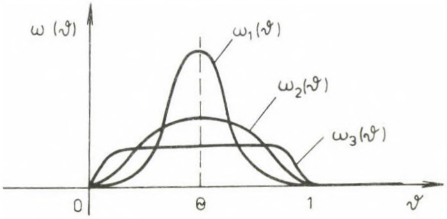


Fig. 4

In the information theory as a measure of uncertainty (or variety) of possible states of some system, defined by continuous accidental value X having function of distribution density $f(x)$, a measure, elaborated by N. Wiener and K. Shannon and named entropy, is used. This measure is widespread used in statistical thermodynamics too. This value is marked by H and is calculated in the form of

$$H = -\int_{-\infty}^{\infty} f(x)\log_2 f(x)dx\,, \tag{7}$$

under condition, that

$$\int_{-\infty}^{\infty} f(x)dx = 1\,. \tag{8}$$

For example, the properties of function H are described in detail in [2, 4], however, in this case, the most important peculiarity of entropy is the following: the function $f(x)$ is of an equipartitional character, the function H is large. In the case of very irregular distribution of the probability density $f(x)$ the value of H will tend to zero.

So it is very temptly to use a formula, similar to (7), for integral estimation of the distribution shape of $\theta(x, y, z)$ in Ω. The values of $\theta(x, y, z)$ in Ω are grouped more densitively relative to Θ, the value of the shape coefficient must be larger. However, in such organization the variety of the mobile possibilities of manipulator will be minimum. It will be a difference between the shape coefficient and the entropy H as a measure of variety of possible states of system, described by the accidental value X.

It is necessary to note, that the value of the shape coefficient must not depend on the sizes and shape of the service envelope Ω.

The first attempt of the writing of the shape coefficient expression may serve, obviously, such that

$$R_\theta = - \iiint\limits_{\Omega} \theta(x, y, z) \log_2 \theta(x, y, z) dx\, dy\, dz\,, \tag{9}$$

where R_θ is a shape coefficient of distribution of $\theta(x, y, z)$ in Ω. However, formula (9) is unhappy, firstly, because the condition of kind (8) is not executed:

$$\iiint\limits_{\Omega} \theta(x, y, z) dx\, dy\, dz = V\Theta \neq 1\,, \tag{10}$$

and, secondly, because value of R_θ depends on the volume V of the region Ω. So it is necessary to modify formula (9), however, its structure must not be changed.

First of all let us reflect the region Ω into region Γ, which is similar to region Ω, by some similar transformation, for example, homotety H_0^k with some centre $O = (x_0, y_0, z_0)$ and similar coefficient K, which is computed from condition that

$$V_\Gamma = \iiint\limits_{\Gamma} d\xi\, d\eta\, d\zeta = 2\,,$$

where V_Γ is a volume of the region Γ.

It is known, that

$$k^3 = V_\Gamma / V, \tag{11}$$

then

$$k = \sqrt{2/V}.$$

Therefore

$$\forall A(x, y, z) \in \Omega\ \exists (\xi, \eta, \zeta) \in \Gamma \quad H_0^k : (x, y, z) \mapsto (\xi, \eta, \zeta)\,, \tag{12}$$

and

$$\begin{pmatrix} \xi \\ \eta \\ \zeta \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + k \begin{pmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{pmatrix}. \tag{13}$$

Besides that, additionally let us demand a condition execution

$$\theta_\Gamma(\xi, \eta, \zeta) = k\theta(x, y, z), \tag{14}$$

if $(\xi, \eta, \zeta) \in \Gamma$ is an image of point $(x, y, z) \in \Omega$ (12).

For the condition of kind (8) execution, let us normalize a value of $\theta_\Gamma(\xi, \eta, \zeta)$:

$$\kappa_\Gamma(\xi, \eta, \zeta) = \theta_\Gamma(\xi, \eta, \zeta) / \iiint\limits_{\Gamma} \theta_\Gamma(\xi, \eta, \zeta) d\xi\, d\eta\, d\zeta\,. \tag{15}$$

In this case

$$\iiint\limits_{\Gamma} \kappa_\Gamma(\xi, \eta, \zeta) d\xi\, d\eta\, d\zeta = 1\,.$$

Using distribution of $\kappa_\Gamma(\xi, \eta, \zeta)$ in Γ formula (9) for R_θ computing may be transformed as

$$R_\theta = -\iiint\limits_\Gamma \kappa_\Gamma(\xi, \eta, \zeta) \log_2 \kappa_\Gamma(\xi, \eta, \zeta) d\xi\, d\eta\, d\zeta\,. \tag{16}$$

Formula (16) is free from drawbacks which related to formula (9), however, it is rather difficult to use this expression practically. So let us return to the parameters of region Ω not analysing the properties of expression (16) here. Let us change variables in the integral (15), taking into account formulas (12), (13), (14), (15):

$$d\xi = kdx, \qquad d\eta = kdy, \qquad d\zeta = kdz\,; \tag{17}$$

$$\kappa_\Gamma(\xi, \eta, \zeta) = \frac{\theta_\Gamma(\xi, \eta, \zeta)}{\iiint\limits_\Gamma \theta_\Gamma(\xi, \eta, \zeta) d\xi\, d\eta\, d\zeta} =$$

$$= \frac{k\theta(x, y, z)}{k^3 \iiint\limits_\Omega k\theta(x, y, z) dx\, dy\, dz} = \frac{\theta(x, y, z)}{k^3 \iiint\limits_\Omega \theta(x, y, z) dx\, dy\, dz} = \frac{\theta(x, y, z)}{k^3 \Theta V}\,. \tag{18}$$

Rewriting formula (16) and taking into account (10), (11), (17) and (18), it is received:

$$R_\theta = -\iiint\limits_\Omega \frac{\theta(x, y, z)}{k^3 \Theta V} \log_2 \left(\frac{\theta(x, y, z)}{k^3 \Theta V}\right) k^3 dx\, dy\, dz =$$

$$= -\frac{1}{\Theta V} \iiint\limits_\Omega \theta(x, y, z) [\log_2 \theta(x, y, z) - \log_2 \Theta - \log_2 (k^3 V)] dx\, dy\, dz =$$

$$= -\frac{1}{\Theta V} \iiint\limits_\Omega \theta(x, y, z) \log_2 \theta(x, y, z) dx\, dy\, dz +$$

$$+ \frac{1}{\Theta V} \iiint\limits_\Omega \theta(x, y, z) dx\, dy\, dz\, [\log_2 \Theta + \log_2 V_\Gamma] =$$

$$= -\frac{1}{\Theta V} \iiint\limits_\Omega \theta(x, y, z) \log_2 \theta(x, y, z) dx\, dy\, dz + \log_2 \Theta + 1\,. \tag{19}$$

Using formula (3) and believing, that $\varphi(\theta) = \theta(x, y, z) \log_2 \theta(x, y, z)$, let be received the final expression for the shape coefficient of the manipulator service distribution in the service envelope computing:

$$R_\theta = 1 + \log_2 \Theta - \frac{1}{\Theta} \int_0^1 \vartheta \log_2 \vartheta \omega(\vartheta) d\vartheta . \tag{20}$$

Let us analyse the properties of receiving expression (20) for value of R_θ computing. Firstly, according to formula (20), R_θ does not depend on the sizes and shape of the service envelope Ω, and it is completely defined through the function of density of distribution of manipulator service in region Ω $\omega = \omega(\vartheta)$. Secondly, according to formula (19), R_θ has maximum value, when $\theta(x, y, z) = \Theta = \text{const}$. In this case, as it follows from direct substitution, $R_\theta = 1$. At the same time, $R_\theta \geqq 0$. It may be proved from the analysis of formula (16). Indeed, $0 \leqq \kappa_\Gamma(\xi, \eta, \zeta) \leqq 1$, therefore $\log_2 \kappa_\Gamma(\xi, \eta, \zeta) \leqq 0$; sign minus standing before the integral in (16) turns all expression into positive value. Thus, for any service envelope Ω with arbitrary distribution of $\theta(x, y, z)$ in Ω, $0 \leqq R_\theta \leqq 1$. The distribution of $\theta(x, y, z)$ in Ω has more homogeneous (equipartitional) character, the value of R_θ tends to 1. With such organization the value of integral service coefficient is unimportant. However, it is necessary to note specially that variety of mobile possibilities of manipulator is minimum in this case.

Thus, introduced shape coefficient of the distribution of $\theta(x, y, z)$ in Ω jointly with moment characteristics of distribution of the manipulator service in the service envelope is original indicator of "confidence" to integral service coefficient. At the same time all considered estimations are independence values. Each of them characterises the distribution of $\theta(x, y, z)$ in Ω defining more precisely its individual peculiarities.

5. In Figs 5 and 6 the examples of distributions of manipulator service in the service envelope are shown. In Fig. 5 plane manipulator with three rotate degrees of freedom is shown. All lengths of links of this manipulator are equal to 1. The limits in all joints are symmetrical and equal to ±90°. In Fig. 6 plane manipulator with three degrees of freedom is shown too. It has two prismatic joints and one rotate joint. The gripper of this manipulator is made in the form of a special pantograph mechanism. This manipulator provides practically homogeneous distribution of the manipulator service in all points of the service envelope. The areas of the service envelopes of two manipulators are equal. The limits in the rotate wrist degree of freedom of the second manipulator (Fig. 6) are selected in such a way that integral service coefficient of two manipulators are equal too. The means of other integral estimations are given in Figs 5 and 6.

In conclusion it is necessary to note that all reasonings concerning to mobile possibilities of manipulators in their service envelope Ω, generally speaking, will be fair for any arbitrary region Ω, which even may not include the service envelope of

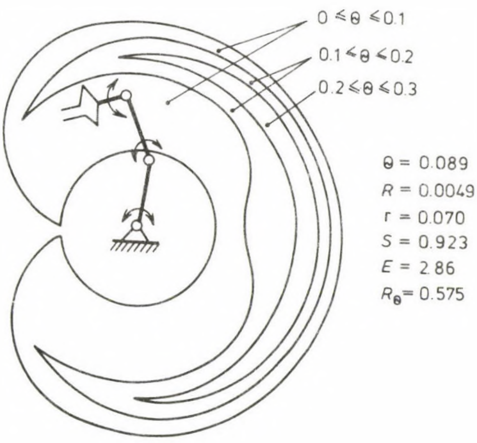


Fig. 5

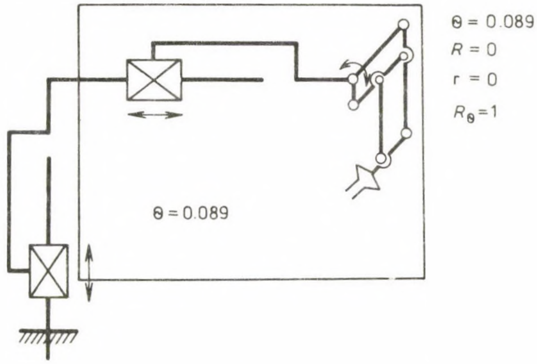


Fig. 6

manipulator analysed. Such widening of the application of examined integral estimations is especially actual in those cases when local properties of manipulator, intended for executing strictly defined class of operations, have particular interest.

References

1. *Vinogradov, I. B., Kobrinsky, A. E., Stepanenko, Y. A., Tyves, L. I.,* Particularities of manipulators kinematics and volumes method. Machines mechanics. M., Nauka, 1971, vol. **27/28,** pp. 5–15.
2. *Ventcel, E. S.,* Theory of probabilities. M., Nauka, 1969, 573 pp.
3. *Kozlov, V. V., Makarychev, V. P., Tymofeev, A. V., Yurevich, E. I.,* Dynamics of robots control. M., Nauka, 1984, 336 pp.
4. *Stratonovich, R. L.* Information theory. M., Sov. radio, 1975, 420 pp.

Интегральные оценки двигательных возможностей манипуляционных роботов

Е. П. ПОПОВ, В. Г. ТИХОМИРОВ, В. И. УШАКОВ
(Москва)

В работе дано развитие метода объемов для оценки геометрических функциональных (или двигательных) возможностей манипуляционных роботов на основе использования современного аппарата теории информации. Дифференциальное (или качественное) распределение функциональных свойств манипуляторов в зоне обслуживания оценивается с помощью системы интегральных характеристик, существенно уточняющих и дополняющих наиболее распространенную в настоящее время величину подобного рода — интегральный коэффициент сервиса.

В представляемой работе функциональные свойства манипулирующих устройств трактуются с вероятностной точки зрения, поэтому наиболее общей интегральной оценкой этих свойств, учитывающей специфику их распределения в зоне обслуживания, может служить энтропийная оценка, названная в работе интегральным коэффициентом формы распределения сервиса манипулятора в зоне обслуживания. Вместе с моментными характеристиками распределения сервиса эта величина позволяет оценивать степень «доверия» к интегральному коэффициенту сервиса.

Практическое приложение представленных в статье результатов актуально в задачах анализа двигательных возможностей манипулирующих устройств, функционирующих в среде с внешними препятствиями — ограничениями, существенно ограничивающими в ряде случаев эти возможности. Использование рассмотренных в работе интегральных оценок позволяет провести сравнительный анализ различных кинематических схем манипуляторов на предмет выявления среди них таких, которые обеспечивают наилучшие двигательные возможности в рамках рассматриваемой среды внешних препятствий — ограничений. Дальнейшее приложение полученных результатов возможно непосредственно в задачах планирования траекторий манипуляторов и построения алгоритмов их управления.

Е. П. Попов
Научно-учебный центр «Робототехника»
СССР, 105037 Москва,
Измайловская пл., 7

Problems of Control and Information Theory, Vol. 15 (3), pp. 231—238 (1986)

NEARNESS ESTIMATIONS FOR OBJECTS WITH COMPLEX STRUCTURE

S. I. Goldberg, O. A. Skripochenko

(Sverdlovsk)

(Received February 8, 1985)

A number of pseudometrics for pattern nearness estimations is introduced. It is implied that there exists a partial order on the set of elements the patterns consist of. At present pseudometrics are applied to solve some problems of medical diagnostics.

1. Introduction

It is well known that a concept of distance between sets of objects has a wide range of application in the problems of control, classification, pattern recognition. Usually, as this takes place, one considers a set X and a system of its subsets $\mathcal{U}$ such that $X \in \mathcal{U}$ and from the sets A and B belonging to the system $\mathcal{U}$ it follows that their unification $A \cup B$, intersection $A \cap B$ and difference $A \backslash B$ belong to this system. In other words, $\mathcal{U}$ is an algebra of sets with unit X [1]. Commonly, one uses a measure of symmetrical difference between A and B as a distance $d(A, B)$ $A \in \mathcal{U}$ $B \in \mathcal{U}$

$$d(A, B) = \mu(A \triangle B)$$

$$\mu(A) \geqq 0 \quad \forall A \in \mathcal{U} \tag{1.1}$$

$$A_1 \cap A_2 = \emptyset \Rightarrow \mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$$

[3], [5].

The function $d(A, B)$ is a pseudometric as the following conditions hold

$$d(A, B) \geqq 0 \tag{1.2}$$

$$d(A, B) = d(B, A) \tag{1.3}$$

$$d(A, B) + d(B, C) \geqq d(A, C) \tag{1.4}$$

$$d(A, A) = 0. \tag{1.5}$$

The important properties of $d(A, B)$ are

$$A \cap C \subseteq B \subseteq A \cup C \Rightarrow d(A, B) + d(B, C) = d(A, C) \tag{1.6}$$

$$A \cap C = B \cap C = \emptyset \Rightarrow d(A, B) = d(A \cup C, B \cup C)\,. \tag{1.7}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

One implies in such pseudometrics that all elements of the set X are independent of each other. And this property is used really in a great number of works in information theory [3], psychological measurements [2], in the study of forest fire propagation [4] etc. However, there is a large range of real life problems with complex relationship between elements of X. For example, consider the choice of the year's best sportsmen on the basis of questionings carried out among sport reviewers (experts) from different countries. The importance of each expert's opinion may differ according to an experts' council. So, if the group consists of the most part of adherents for the same trend as our concrete expert, the importance of his opinion will be lower than if he represents this tradition alone. Besides, other respected experts' opinions may influence him. Thus, it is necessary to take into account possible dependence of experts' opinions to get an objective picture.

Problems analogous to the given one also arise in a great number of medicine diagnostics problems. The authors are connected with the problems of such kind by the character of their work. In the present article we describe pseudometrics suitable for such problems.

We should like to pay attention to just one more aspect of distance between sets, viz. property (1.7) of distance (1.1). It guarantees that the common part does not influence the value of distance. Due to this property, the distance of such kind can be called "distance by distinction". On the other hand, in medicine one regards diseases close to one another (in the sense of similarity of clinical pictures) if most of their typical representatives are similar. Nearness estimation with the help of distance by distinction compares only diseases differing representatives factually. Therefore, e.g. grippe may prove to be equally unlike chill and such rare pathology as jungle fever. In this way, it is reasonable to find the so-called "distance by community", i.e. a function which shortens distance between sets with common part. The present work is an attempt of distance functions' introduction using considerations given above.

2. Basic definitions and notations

Consider a partially ordered set X and a system of its subsets $\mathscr{U}$ which is, as well as the one in [1], an algebra with unit X.

Denote by $R^+(A)$ a set of all maximum (with respect to the given partial order relation) elements of a set A, i.e.

$$R^+(A)=\{a^+\in A\forall a\in A: a\ngeq a^+\}.$$

Analogously, $R^-(A)$ is the set of all minimum elements of the set A

$$R^-(A)=\{a^-\in A\forall a\in A: a\nleq a^-\}.$$

Introduce the functions

$$d(A,B)=\mu(R^{+}(A\cup B))-\mu(R^{+}(A)\cap R^{+}(B)) \tag{2.1}$$

$$\rho(A,B)=\mu(R^{+}(A\triangle B)) \tag{2.2}$$

defined on the elements of the system $\mathcal{U}$.[1]

Here $\mu(A)$ is a non-negative measure possessing the properties;

1. $\mu(A)+\mu(B)=\mu(A\cup B)$ when $A\cap B=\emptyset$
2. $\mu(R^{+}(A))\geqq\mu(R^{-}(A))$.

Show that the functions given by formulae (2.1) and (2.2) are pseudometrics and consider their properties.

3. Function $d(A, B)$ and its properties

Statement 3.1. The function $d(A, B)$ defined by formula (2.1) satisfies conditions (1.2)–(1.5).

Proof. The verity of (1.2)–(1.4) is evident. Consider (1.5). $d(A, B)$ may be represented as

$$d(A,B)=\mu(\tilde{A}_{B})+\mu(\tilde{B}_{A}) \tag{3.1}$$

where

$$\tilde{A}_{B}=\{a^{+}\in R^{+}(A)\,\forall\, b\in B\quad b\ngeqq a^{+}\}$$

$$\tilde{B}_{A}=\{b^{+}\in R^{+}(B)\,\forall\, a\in A\quad a\ngeqq b^{+}\}$$

because $\tilde{A}_{B}$ and $\tilde{B}_{A}$ taken together contain all maximum elements for $A\cup B$ except the ones which belong to A and B simultaneously. Then (1.5) will be written as

$$\mu(\tilde{A}_{C})+\mu(\tilde{C}_{A})\leqq\mu(\tilde{A}_{B})+\mu(\tilde{B}_{A})+\mu(\tilde{C}_{B})+\mu(\tilde{B}_{C}). \tag{3.2}$$

Show the verity of inequalities

$$\mu(\tilde{A}_{C})\leqq\mu(\tilde{A}_{B})+\mu(\tilde{B}_{C}) \tag{3.3}$$

$$\mu(\tilde{C}_{A})\leqq\mu(\tilde{C}_{B})+\mu(\tilde{B}_{A}). \tag{3.4}$$

Take an arbitrary $a^{*}\in\tilde{A}_{C}$. Two variants are possible for it:

(a) $\forall\, b\in B\quad b\ngeqq a^{*}$
(b) $\exists\, b^{0}\in B\quad b^{0}\geqq a^{*}$.

[1] Here and below all the sets we come across belong to the given system of subsets.

In case (a), $a^* \in \tilde{A}_B$ and, consequently, (3.3) is correct. If (b) is true, then taking into account that $a^* \in \tilde{A}_C$, we have $c \ngeqq b^0$ for any $c \in C$. By the definition of $R^+(B)$ there exists $b^+ \in R^+(B)\, b^+ \geqq b^0$ and therefore $c \ngeqq b^+$, i.e. $b^+ \in \tilde{B}_C$. Consider $\{a^*\}$ a set of all a^* meeting variant (b). Let $\{b^+\}$ be a set of all b^+ corresponding to it. According to the 2-d measure property $\mu(\{b^+\}) \geqq \mu(\{a^*\})$. Thus, (3.3) is correct in variant (b), too. (3.4) is proved in the same way. If we add (3.3) and (3.4) term to term, we get the sought inequality of a triangle.

Statement 3.2. Inequality (1.5) turns into equality if the following conditions are fullfilled

$$R^+(A) \cap R^+(B) \subset R^+(C) \subset R^+(A) \cup R^+(B) \tag{3.5}$$

$$\forall\, a \in (R^+(A) \cup R^+(B)) \backslash R^+(A \cup B) \exists\, c \in R^+(C), \qquad a \leqq C. \tag{3.6}$$

Proof. With regard to (1.5) it is sufficient to show that any element from $\tilde{A}_C$, $\tilde{C}_A$, $\tilde{C}_B$, $\tilde{B}_C$ is contained in the sets $\tilde{A}_B$, $\tilde{B}_A$ only once

$$\tilde{A}_C = \{a^+ \in R^+(A)\, \forall\, c \in C \quad c \ngeqq a^+\}$$

$$\tilde{C}_B = \{c^+ \in R^+(C)\, \forall\, b \in B \quad b \ngeqq c^+\}.$$

As $R^+(C)^- \subset R^+(A) \cup R^+(B)$ and with $c \in R^+(B)$ there is $b \in B\; b = C$, then $\tilde{C}_B \subset R^+(A)$. $\tilde{C}_B \cap \tilde{A}_C = \emptyset$ (for $a \in \tilde{A}_C\, \forall\, c \in C\; c \neq a$). Now consider $a \in R^+(A)$ and $a \notin \tilde{A}_B$. In this case for $a \in R^+(A)$ there exists $b \in B$, $b \geqq a$ and so $a \notin \tilde{C}_B$ by the definition of $\tilde{C}_B$, and, besides, either $a \in (R^+(A) \cup R^+(B)) \backslash R^+(A \cup B)$ and according to the condition there is $c \in R^+(C) a \leqq c\; a \notin \tilde{A}_C$, or $a \in R^+(A) \cap R^+(B)$ and there exists $c \in R^+(C) a = c a \notin \tilde{A}_C$. In the same manner we obtain $\tilde{B}_C \cup \tilde{C}_A = \tilde{B}_A$, $\tilde{B}_C \cap \tilde{C}_A = \emptyset$.

Statement 3.3. The following inequality is correct for the function defined by formula (2.1)

$$d(A \cup C, B \cup C) \leqq d(A, B) \quad \forall\, c \quad A \cap C = B \cap C = \emptyset. \tag{3.7}$$

The reader will restore the proof easily, following the logic of the proof of statement (3.1). Obviously, the following statement is valid.

Statement 3.4. Inequality (3.7) is executed strictly, if for some element $p \in R^+(A \cup B) \backslash (R^+(A) \cap R^+(B))$ there exists $c \in C$, connected with it by correlation $c \geqq b$. In other words the set A disagrees with B in the sense of pseudometrics d and at least one element from C is a maximum one for the differing parts of A and B.

4. Function $\rho(A, B)$ and its properties

Statement 4.1. Function $\rho(A, B)$ defined by formula (2.2) is pseudometric.

Proof. The fact that $\rho(A, B)$ satisfies axioms (1.2)–(1.4) is trivial. Let us prove triangle inequality

$$\rho(A, B)+\rho(B, C)\geqq\rho(A, C). \tag{4.1}$$

Otherwise

$$\mu(R^+(A \triangle B))+\mu(R^+(B \triangle C))\geqq\mu(R^+(A \triangle C)). \tag{4.2}$$

Transform the left part using the first property of the measure

$$\mu(R^+(A \triangle B))+\mu(R^+(B \triangle C))\geqq\mu(R^+((A \triangle B)\cup(B \triangle C)))$$

$$\mu(R^+((A \triangle B)\cup(B \triangle C)))=\mu(R^+((A\cup C)\backslash B)\cup(B\backslash(A\cap C)))\geqq$$

$$\geqq\mu(R^+((A\cup C)\backslash(A\cap C)))$$

because

$$(A\cup B\cup C)\backslash(A\cap B\cap C)\supseteq(A\cup C)\backslash(A\cap C).$$

And for arbitrary K and N the inequality $\mu(R^+(K\cup N))\geqq\mu(R^+(K))$ is equivalent to the 2nd property of the measure. One can directly verify

Statement 4.2. Triangle inequality turns into equality, if one of the following limitations on C is fullfilled

$$A\subseteq C\subseteq A\cup B$$

or

$$B\subseteq C\subseteq A\cup B.$$

From formula (2.2) directly follows

Statement 4.3. The following property is correct for the function $\rho(A, B)$, given by formula (2.2)

$$\rho(A\cup C, B\cup C)=\rho(A, B) \quad \forall c \quad A\cap C=B\cap C=\emptyset.$$

5. Estimating of distance between objects with the help of distance between their patterns

One may consider $g(f(A), f(B))$, where f is a function transferring $\mathscr{U}$ in a space with metrics or pseudometrics g, as nearness estimation of the sets A and B.

Consider e.g. $\mathscr{L}$ is a set of all n-tuples filled up with symbols 0 and 1. The line $a\in\mathscr{L}, a=(a_1, a_2, \ldots, a_n)$ is less than the line $b\in\mathscr{L}, b=(b_1, b_2, \ldots, b_n)$ if $a_i\cap b_i=a_i\;\forall\, i$.

For an arbitrary $A \subseteq \mathscr{L}$ consider

$$f(A)=\{\alpha \in \mathscr{L} \quad \alpha \nleqq a \quad \forall\, a \in A\}$$

$$f^k(A)=\{\alpha \in \mathscr{L} \quad \alpha \nleqq a \quad \forall\, a \in A \quad \sum_{i=1}^{n} \alpha_i = k\}.$$

In [6] and [7] we considered

$$d(A, B)=\mu(R^-(f(A) \cup f(B))-\mu(R^-(f(A)) \cap R^-(f(B)))$$

$$\mu(A)+\mu(B)=\mu(A \cup B) \Leftarrow A \cap B=\emptyset$$

$$\mu(R^-(A)) \geqq \mu(R^+(A))$$

$$\bar{\rho}(A, B)=\mu(f^k(A) \triangle f^k(B)).$$

Note that properties of the function $g=g(f(A), f(B))$ are not always the same as the corresponding properties of the analogous distance given on the sets themselves.

Thus, e.g. for $\bar{\rho}=\mu(f^k(A) \triangle f^k(B))$ the property

$$\bar{\rho}(A \cup C, B \cup C)=\bar{\rho}(A, B) \qquad A \cap C=B \cap C=\emptyset$$

is not correct. Really, let

$$\begin{array}{rrr} 01101 & 10100 & \\ A=01011 & B=10011 & C=11100 \\ 00111 & 01101 & \end{array}$$

$$\bar{\rho}(A, B)=5 \qquad \bar{\rho}(A \cup C, B \cup C)=4$$

$$\bar{\rho}(A, \cup C, B \cup C)<\bar{\rho}(A, B).$$

6. Application of suggested pseudometrics

At present, pseudometrics given by formula (2.1) is applied to the problem of differential diagnostics of acute ishaemic heart disease (IHD). Each nosological form is characterized by a collection of typical syndromes.

A concrete patient features a definite collection of syndromes as well. The conclusion about diagnosis is drawn depending on the nearness of the collection of syndromes which characterizes the patient to some syndrome collections, characterizing different pathology. There exists partial order on the set of syndromes. Consider, for example, three syndromes: stenocardia syndrome (an ache behind the sternum, the dull character of an ache), the syndrome of intermediate forms of IHD (an ache behind the sternum, the dull character of an ache, unfamiliar ache as compared to past one) and heartmuscle one (an ache behind the sternum, the dull character of an ache, high transaminase). Stenocardia syndrome may be put into the syndrome of intermediate forms of IHD and into heartmuscle one and so on.

The choice of concrete measure is carried out in the following way. For some collections of syndromes A, B, C, D, the inequalities

$$d(A, B) < d(C, D)$$

$$\mu(R^+(A \cup B)) - \mu(R^+(A) \cap R^+(B)) < \mu(R^+(C \cup D)) - \mu(R^+(C) \cap R^+(D))$$

are defined by the experts. Each inequality of such type is linear with respect to measures of some syndromes $\mu_i(a)$ $a \in A \cup B \cup C \cup D$. Solving the obtained system of linear inequalities with respect to μ_i and in accordance to the 2nd property of the measure, we get the opportunity to build up concrete type of pseudometrics for differential diagnostics. A system for defining ishaemic heart disease of acute forms is not finished yet but its first variant is tested on the basis of experiences in Sverdlovsk specialized clinics. At the same time pseudometrics [7] is used in a system defining acute cerebro-vascular disorder character, while [6] is applied to define prognosis of bone tissue regeneration in osteosynthesis. The two last systems are industrially exploited.

References

1. *Kolmogorov, A. I., Fomin, S. V.*, Elements of functions theory and functional analysis. M., Nauka, 1972, p. 249.
2. *Luce, R. D., Galanter, E.*, Psychophysical scales. In: Psychological measurement. M., Mir, 1967 (pp. 111–195), p. 180.
3. *Nevjo, J.*, Mathematical basis of information theory. M., Mir, 1969, pp. 29–30.
4. *Vorobiov, O. I.*, Conditions models of some distributed probable processes. The news of SD AS USSR, Series of technical sciences No. 3, vol. **1**, Novosibirsk, 1981 (pp. 105–113), p. 107.
5. *Orlov, A. I.*, Stability in social-economical models. M., Nauka, 1979.
6. *Goldberg, S. I.*, On distance on bi-tables sets and its application for the purpose of medical diagnostics. In: Natural sciences serve to the care of public health. Novosibirsk, 1980, pp. 79–80.
7. *Goldberg, S. I.*, About one distance on bi-tables. In: Algorithms of analysis of data of social-economical research works. Novosibirsk, 1982, pp. 44–60.

Оценки близости объектов со сложной структурой

С. И. ГОЛЬДБЕРГ, О. А. СКРИПОЧЕНКО
(Свердловск)

В статье вводятся расстояния между множествами элементов. Считается, что элементы связаны друг с другом (в нашем случае это отражено в задании отношения частичного порядка на элементах).

Рассматриваются как псевдометрики, которые реагируют на добавление к множествам одинаковых частей, так и псевдометрики, значение которых уменьшается при данной процедуре.

В качестве примера и использования изучаемых конструкций предлагаются задачи медицинской диагностики, в частности задача определения форм ишемической болезни сердца в острый период.

С. И. Гольдберг
Областной медицинский информационно-вычислительный центр
СССР Свердловск 620102,
Волгоградская ул., 189

Problems of Control and Information Theory, Vol. 15 (3), pp. 239—251 (1986)

SYNTHESIS OF FEEDBACK CONTROL AND FINITE DIMENSIONAL MODELS

D. A. Serkov

(Sverdlovsk)

(Received March 29, 1985)

The paper is concerned with feedback control problem for a parabolic system. The control and the disturbance are held in a finite set of inner points. An optimal feedback strategy is constructed on the basis of the same one for the finite dimensional approximate model. Examples of such models are adduced.

Introduction

The feedback control problem [1, 2] for a process described by a parabolic equation is studied. As example of the process, heating of heterogeneous body or diffusion in heterogeneous medium may be considered. For the given initial state, the evolution of the process is determined by fixed finite set of point sources (they are point heaters or point sources of diffusing substance in examples). It is supposed that all of them fall into two classes: the sources of the first class have controllable intensity (they create control influence), the rest are uncontrollable (they produce disturbance). The whole process is considered during the finite time interval. The index estimating quality of the control consists of three addends: deviation of process state at terminal moment from the given state, control sources consumption and uncontrollable sources consumption with minus sign. It takes to form control influence to minimise the index. It is allowed to use only the information about process state at the moment under the forming of the control influence, i.e. control formed according to the feedback principle. Thus, the problem is to build up a strategy function that for every moment and for every process state put down control influence at that moment, and it takes among all of the strategies to choose the optimal one, i.e. the greatest of quality index which may occur under the different realizations of disturbance, was minimum. It takes also that the strategy was universal [2], i.e. it must be independent of an initial state of the process; this requirement causes including of some additional precision parameter into the strategy.

The strategy was constructed [3] by means of stochastic program maximin method [2]. However, the values of the optimal strategy were expressed in terms of

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

infinite dimensional functional minimums, that is obstacle to numeral realization of the control. So, in the paper we build new optimal strategy based on the finite dimensional structures. Namely, it is the introduced sequence of the models described by ordinary differential equations that approximate the process in special sense. Proceeding from the initial problem auxiliary feedback control problem is formulated for every model, and it is constructed corresponding to universal optimal strategy by means of stochastic program maximin method. The strategies include no operations with any infinite dimensional element. The optimal feedback control for the initial problem is constructed on the basis of the strategies in the following way. With respect to the state of the process at some moment the control object calculates a phasic state of the model that is chosen according to the precision parameter. Then with respect to the phasic state it calculates the value of the optimal strategy corresponding to the model and puts the value into the process as the control influence at the moment. It is established that under suitable choice of all elements of the scheme the control will be optimal.

1. Statement of the problem

Consider the controllable system described by the equation

$$\begin{cases} \dfrac{\partial \xi}{\partial t}(t,x)-(\mathscr{A}\xi)(t,x)=\sum\limits_{i=1}^{k}\delta(x-x_{u_i})u_i[t]+\sum\limits_{i=1}^{m}\delta(x-x_{v_i})v_i[t], \\ ((t,x)\in[t_*,\vartheta]\times Q), \quad \xi(t,x)=0\,((t,x)\in[t_*,\vartheta]\times\partial Q), \\ \xi(t_*,\cdot)=\xi_*(\cdot)\in L_2(Q), \quad t_*\in[t_0,\vartheta], \end{cases} \tag{1.1}$$

here Q is limitary domain in $R^{\bar{n}}$, $\bar{n}\in\{1,2,3\}$, with properties (i) domain Q satisfies conditions 1), 2) [8, p. 212]; (ii) domain Q has property $\mathscr{R}$ [8, p. 222];

$$(\mathscr{A}f)(x)=\sum_{i,j=1}^{\bar{n}}\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial f}{\partial x_j}\right)(x)+a(x)f(x),$$

$$v_1\sum_{i=1}^{\bar{n}}y_i^2\leqq\sum_{i,j=1}^{\bar{n}}a_{ij}(x)y_iy_j\leqq v_2\sum_{i=1}^{\bar{n}}y_i^2,$$

$$a_{ij}(x)=a_{ji}(x),$$

$$\left|\frac{\partial a_{ij}}{\partial x_k}(\cdot)\right|_{L_\infty(Q)}\leqq v_2 \quad (i,j,k\in\overline{1,\bar{n}}),$$

$$0\leqq a(x)\leqq v_2, \qquad 0<v_1\leqq v_2<+\infty$$

almost everywhere on Q, $a(\cdot)\in L_\infty(Q)$;

$$x_{u_i},x_{v_j}\in \operatorname{Int}(Q)\,(i\in\overline{1,k},j\in\overline{1,m}),u[\cdot]=(u_1[\cdot],\,...\,u_k[\cdot])\in$$
$$\in\mathscr{U}_{[t_*,\vartheta]}=\{u[\cdot]:u_i[\cdot]\in L_\infty([t_*,\vartheta])\,(i\in\overline{1,k}),u[t]\in S_k(\mu)\,(t\in[t_*,\bar{\vartheta}]),$$
$$u[t]\in S_k(0)\,(t\in(\bar{\vartheta},\vartheta])\},v[\cdot]=(v_1[\cdot],\,...,v_m[\cdot])\in\mathscr{V}_{[t_*,\vartheta]}=$$
$$=\{v[\cdot]:v_i[\cdot]\in L_\infty([t_*,\vartheta])\,(i\in\overline{1,m}),v[t]\in S_m(\mu)\,(t\in[t_*,\bar{\vartheta}]),v[t]\in$$
$$\in S_m(0)\,(t\in(\bar{\vartheta},\vartheta]\},S_l(\beta)=\{z\in R^l:|z|\leqq\beta\}\,(l\in N,\beta\geqq 0),$$

$|\cdot|$ is euclidean norm, $\bar{\vartheta}\in(t_0,\vartheta)$, $\vartheta-\vartheta=\varepsilon_\vartheta$. An element $\xi(\cdot)\in L_2([t_*,\vartheta]\times Q)$ which satisfies

$$\int\limits_{[t_*,\vartheta]}\int\limits_Q\xi(\tau,x)\left[-\frac{\partial\psi}{\partial t}(\tau,x)-(\mathscr{A}\psi)(\tau,x)\right]dx\,d\tau=$$

$$=\int\limits_{[t_*,\vartheta]}\left(\sum_{i=1}^k\psi(\tau,x_{u_i})u_i[\tau]+\sum_{i=1}^m\psi(\tau,x_{v_i})v_i[\tau]\right)d\tau+\int\limits_Q\xi_*(x)\psi(t_*,x)\,dx$$

for every $\psi(\cdot)\in\{\psi(\cdot)\in W_2^{1,2}([t_*,\vartheta]\times Q):\psi(t,x)=0\ ((t,x)\in([t_*,\vartheta]\times\partial Q)\cup(\{\vartheta\}\times Q))\}$ will be called [7] a solution of equation (1.1) (partial derivatives and set $W_2^{1,2}([t_*,\vartheta]\times Q)$ are defined in [8]).

Proposition 1.1. For every $t_*\in[t_0,\vartheta]$, $\xi_*(\cdot)\in L_2(Q)$, $u[\cdot]\in\mathscr{U}_{[t_*,\vartheta]}$, $v[\cdot]\in\mathscr{V}_{[t_*,\vartheta]}$ the unique solution $\xi(\cdot)=\xi(\cdot\,|\,t_*,\xi_*(\cdot),u[\cdot],v[\cdot])$ of equation (1.1) exists and may be represented in form

$$\xi(t,x)=\int\limits_Q G(x,z,t-t_*)\xi_*(z)dz+$$
$$+\int\limits_{[t_*,t]}\left(\sum_{i=1}^k G(x,x_{u_i},t-\tau)u_i[\tau]+\sum_{i=1}^{rn}G(x,x_{v_i},t-\tau)v_i[\tau]\right)d\tau$$
$$((t,x)\in(\bar{\vartheta},\vartheta]\times Q)),$$

where

$$G(\cdot)\in L_2(Q\times Q\times[a,b])\cap C^0(\bar{Q}\times\bar{Q}\times[a,b])\,([a,b]\subset(0,+\infty)).$$

Further $\mathscr{M}$-limitary set in $L_2(Q)$;

$$\mathscr{M}(t_*)=\{\xi(t_*,\cdot\,|\,t_0,\xi_0(\cdot),u[\cdot],$$
$$v[\cdot]):\xi_0(\cdot)\in\mathscr{M},u[\cdot]\in\mathscr{U}_{[t_*,\vartheta]},v[\cdot]\in\mathscr{V}_{[t_*,\vartheta]}\}\,(t_*\in[t_0,\vartheta]);$$
$$D=\{(t_*,\xi_*(\cdot)):t_*\in[t_0,\vartheta],\xi_*(\cdot)\in\mathscr{M}(t_*)\};$$

$$\xi^{[t,\vartheta]}(x)=\int\limits_Q G(x,z,\vartheta-t)\xi(z)dz(\xi(\cdot)\in L_2(Q),x\in Q,t\in[t_0,\vartheta]);$$

$\langle \cdot, \cdot \rangle_{L_2(Q)}$ and $\langle \cdot, \cdot \rangle$ are products in $L_2(Q)$ and R^n respectively; $| \cdot |_{L_2(Q)}$ — norm in $L_2(Q)$; $| A |_{m \times n}$ is norm of any $m \times n$-matrix A associated with norms in R^m and R^n; $\Phi(\cdot)$ and $\Psi(\cdot)$ are measurable (respectivelyl $k \times k$- and $m \times m$-matrix-functions with properties

$$C_1\langle u, u\rangle \leqq \langle \Phi(\tau)u, u\rangle \leqq C_2\langle u, u\rangle,$$

$$C_1\langle v, v\rangle \leqq \langle \Psi(\tau)v, v\rangle \leqq C_2\langle v, v\rangle \qquad (u \in R^k, v \in R^m, \tau \in [t_0, \vartheta]);$$

$$\beta(\tau, u, v) = \langle \Phi(\tau)u, u\rangle - \langle \Psi(\tau)v, v\rangle \quad (u \in R^k, v \in R^m, \tau \in [t_0, \vartheta]);$$

$$\begin{aligned} \varphi(t_*, \xi_*(\cdot), u[\cdot], v[\cdot]) = {} & | \xi(\vartheta, \cdot \mid t_*, \xi_*(\cdot), u[\cdot], v[\cdot]) |_{L_2(Q)} + \\ & + \int\limits_{[t_*, \vartheta]} \beta(\tau, u[\tau], v[\tau]) d\tau \quad ((t_*, \xi_*(\cdot)) \in [t_0, \vartheta] \times \\ & \times L_2(Q), u[\cdot] \in \mathscr{U}_{[t_*, \vartheta]}, v[\cdot] \in \mathscr{V}_{[t_*, \vartheta]}). \end{aligned} \tag{1.2}$$

Consider the problem on minimum of ensured result for the functional (1.2) and system (1.1) in formalisation [2, 3]. Any function $U(\cdot): [t_0, \vartheta] \times L_2(Q) \times (0, 1) \to S_k(\mu)$ will be called the *strategy of the first player*. The set of all the strategies will be denoted by $\mathscr{U}$. The triplet $\{U(\cdot), \varepsilon, \Delta\}$, where $U(\cdot) \in \mathscr{U}, \Delta$ — partition of the segment $[t_*, \vartheta]$ $(t_* \in [t_0, \vartheta])$ and $\varepsilon \in (0, 1)$ will be called the *first player control law* on $[t_*, \vartheta]$. The function

$$\xi(\cdot) = \xi(\cdot \mid t_*, \xi_*(\cdot), \{U(\cdot), \varepsilon, \Delta\}, v[\cdot]) \in L_2([t_*, \vartheta] \times Q)$$

will be called the motion of the system from the position $(t_*, \xi_*(\cdot)) \in [t_0, \vartheta] \times L_2(Q)$ produced by the first player control law $\{U(\cdot), \varepsilon, \Delta\}$ $(\Delta = (\tau_i)\, i \in \overline{1, n})$ and the realization $v[\cdot] \in \mathscr{V}_{[t_*, \vartheta]}$ of the second player control, if

$$\xi(t_*, \cdot) = \xi_*(\cdot)$$

$$\xi(t, \cdot) = \xi(t, \cdot \mid \tau_i, \xi(\tau_i, \cdot), u_i[\cdot], v[\cdot]) \quad (t \in (\tau_i, \tau_{i+1}] \cap [t_0, \bar{\vartheta}], i \in \overline{1, n-1})$$

$$u_i[t] = U(\tau_i, \xi(\tau_i, \cdot), \varepsilon) \qquad (t \in [\tau_i, \tau_{i+1}), i \in \overline{1, n-1})$$

$$\xi(t, \cdot) = \xi(t, \cdot \mid \bar{\vartheta}, \xi(\bar{\vartheta}, \cdot), u[\cdot], v[\cdot]) \qquad (t \in (\bar{\vartheta}, \vartheta], u[\cdot] \in \mathscr{U}_{[\bar{\vartheta}, \vartheta]});$$

while the function

$$u[\cdot] = u(\cdot \mid t_*, \xi_*(\cdot), \{U(\cdot), \varepsilon, \Delta\}, v[\cdot]) \in \mathscr{U}_{[t_*, \vartheta]},$$

$$u[t] = u_i[t]\, (t \in [\tau_i, \tau_{i+1}) \cap [t_0, \bar{\vartheta}]), u[t] \in S_k(0)\, (t \in (\bar{\vartheta}, \vartheta])$$

will be called the *realization of the first player control under the motion*

$\xi(\cdot\mid t_*, \xi_*(\cdot), \{U(\cdot), \varepsilon, \Delta\}, v[\cdot])$. For all $U(\cdot)\in\mathscr{U}$, $(t_*, \xi_*(\cdot))\in[t_0, \vartheta]\times L_2(Q)$ denote

$$\gamma_1(U(\cdot), t_*, \xi_*(\cdot)) = \overline{\lim_{\varepsilon\to 0}}\ \overline{\lim_{\delta\to 0}}\ \sup_{\delta(\Delta)\leqq\delta}\ \sup_{v[\cdot]\in\mathscr{V}_{[t_*, \vartheta]}} \varphi(t_*, \xi_*(\cdot),$$
$$u(\cdot\mid t_*, \xi_*(\cdot), \{U(\cdot), \varepsilon, \Delta\}, v[\cdot]), v[\cdot])$$

where $\delta(\Delta)=\max\{\tau_{i+1}-\tau_i : i\in\overline{1, n-1}\}$. The strategy $U_0(\cdot)\in\mathscr{U}$ will be called *D-optimal*, if

$$\gamma_1(U_0(\cdot), t_*, \xi_*(\cdot)) = \inf\{\gamma_1(U(\cdot), t_*, \xi_*(\cdot)) : U(\cdot)\in\mathscr{U}\}\ ((t_*, \xi_*(\cdot)\in D).$$

By the same manner we define the strategy set $\mathscr{V}$ and the control law of the second player, the motion $\xi(\cdot\mid t_*, \xi_*(\cdot), u[\cdot], \{V(\cdot), \varepsilon, \Delta\})$ from the initial position $(t_*, \xi_*(\cdot))$ produced by the second player control law $\{V(\cdot), \varepsilon, \Delta\}$ and the realization $u[\cdot]\in\mathscr{U}_{[t_*, \vartheta]}$ of the first player control on the segment $[t_*, \vartheta]$, as well as the realization $v(\cdot\mid t_*, \xi_*(\cdot), u[\cdot], \{V(\cdot), \varepsilon, \Delta\})$ of the second player control under the motion. For all $V(\cdot)\in\mathscr{V}$, $(t_*, \xi_*(\cdot))\in[t_0, \vartheta]\times L_2(Q)$ denote

$$\gamma_2(V(\cdot), t_*, \xi_*(\cdot)) = \underline{\lim_{\varepsilon\to 0}}\ \underline{\lim_{\delta\to 0}}\ \inf_{\delta(\Delta)\leqq\delta}\ \inf_{u[\cdot]\in\mathscr{U}_{[t_*, \vartheta]}} \varphi(t_*, \xi_*(\cdot), u[\cdot],$$
$$v(\cdot\mid t_*, \xi_*(\cdot), u[\cdot], \{V(\cdot), \varepsilon, \Delta\})).$$

The strategy $V_0(\cdot)$ will be called *D-optimal*, if

$$\gamma_2(V_0(\cdot), t_*, \xi_*(\cdot)) = \inf\{\gamma_2(V(\cdot), t_*, \xi_*(\cdot)) : V(\cdot)\in\mathscr{V}\}\ ((t_*, \xi_*(\cdot))\in D).$$

The couple of the problems to find the above D-optimal strategies of the players consists of the *differential game* [1, 2], that will be denoted by Γ. The function $\rho(\cdot)$: $[t_0, \vartheta]\times L_2(Q)\to R^1$ will called the *value* of the game Γ, if

$$\rho(t_*, \xi_*(\cdot)) = \inf\{\gamma_1(U(\cdot), t_*, \xi_*(\cdot)) : U(\cdot)\in\mathscr{U}\} =$$
$$= \sup\{\gamma_2(V(\cdot), t_*, \xi_*(\cdot)) : V(\cdot)\in\mathscr{V}\}\ ((t_*, \xi_*(\cdot))\in[t_0, \vartheta]\times L_2(Q)).$$

It is established [3], that the value $\rho(\cdot)$ of the game Γ exists.

2. Auxiliary optimal control problems

For every $p\in N$ fix $n(p)\in N$ and consider the system described by equation

$$\begin{cases} \dot{y}_p(t) = B_p(t)u[t] + C_p(t)v[t] & (t\in[t_*, \vartheta)], \\ y_p(t_*) = y_*\in R^{n(p)}, & t_*\in[t_0, \vartheta], \end{cases} \tag{2.1}$$

here $y\in R^{n(p)}$, $B_p(\cdot)$, $C_p(\cdot)$ are continuous on $[t_0,\vartheta]$ (respectively) $n(p)\times k$- and $n(p)\times m$-matrix-functions; $u[\cdot]\in\mathscr{U}_{[t_*,\vartheta]}$, $v[\cdot]\in\mathscr{V}_{[t_*,\vartheta]}$. Solution (according to Carathéodory [12]) of equation (2.1) will be denoted $y_p(\cdot\mid t_*,y_*,u[\cdot],v[\cdot])$; for every

$$\varphi_p(t_*,y_*,u[\cdot],v[\cdot])=k_p\cdot|y_p(\vartheta\mid t_*,y_*,u[\cdot],v[\cdot])|+\int_{[t_*,\vartheta]}\beta(\tau,u[\tau],v[\tau])d\tau\,,$$
$$k_p\in(0,1],\quad (t_*\in[t_0,\vartheta],\,y_*\in R^{n(p)},\,u[\cdot]\in\mathscr{U}_{[t_*,\vartheta]},\,v[\cdot]\in\mathscr{V}_{[t_*,\vartheta]})\,.\tag{2.2}$$

Consider the differential game for system (2.1) with payoff (2.2) in formalisation [2] (denote it Γ_p). The proof of the next two results is similar to reasoning from [2]:

Proposition 2.1. For every $p\in N$ there exists the value $\rho_p(\cdot)$ of the differential game Γ_p. The strategy $U_p(\cdot)$ of the first player in the game Γ_p defined by

$$\begin{gathered}U_p(t,y,\varepsilon)=\arg\min\{k_p^2\langle y-w_0,B_p(t)u\rangle-c_0\langle\Phi(t)u,u\rangle:u\in S_k(\mu)\}\,,\\(w_0,c_0)=\arg\min\{\rho_p(t,w)+c:k_p^2|y-w|^2+c^2\leqq\varepsilon^2\}\\((t,y,\varepsilon)\in[t_0,\bar\vartheta]\times R^{n(p)}\times(0,1))\,,\\U_p(t,y,\varepsilon)\in S_k(0)\,((t,y,\varepsilon)\in(\bar\vartheta,\vartheta]\times R^{n(p)}\times(0,1))\end{gathered}\tag{2.3}$$

is optimal.

Proposition 2.2. For every $p\in N$, if

$$\mu\geqq(2C_1)^{-1}\sup\{k_p|B_p(\tau)|_{n(p)\times k},\,k_p|C_p(\tau)|_{n(p)\times m}:\tau\in[t_0,\bar\vartheta]\}$$

then for all $(t,y,\varepsilon)\in[t_0,\bar\vartheta]\times R^{n(p)}\times(0,1)$

$$\begin{aligned}&U_p(t,y,\varepsilon)=-(k_p/2)\Phi^{-1}(t)B_p'(t)l_p(t,y,\varepsilon)\,,\\&l_p(t,y,\varepsilon)=\arg\max\{k_p\langle y,l\rangle+\langle N_p(t)l,l\rangle-\lambda_p(t)\langle l,l\rangle-\\&\qquad-\varepsilon\sqrt{\langle l,l\rangle+1}:|l|\leqq1\}\,,\\&N_p(t)=\begin{cases}(k_p^2/4)\int_{[t,\bar\vartheta]}[C_p(\tau)\Psi^{-1}(\tau)C_p'(\tau)-B_p(\tau)\Phi^{-1}(\tau)B_p'(\tau)]d\tau & (t\in[t_0,\bar\vartheta])\\0\quad(t\in(\bar\vartheta,\vartheta])\,,\end{cases}\\&\lambda_p(t)=\max\{\langle N_p(\tau)l,l\rangle:|l|\leqq1,\,\tau\in[t,\vartheta]\}\quad(t\in[t_0,\vartheta])\,.\end{aligned}\tag{2.4}$$

Proposition 2.3. Under conditions of Proposition 2.2 there exists $C>0$ such that for all $t\in[t_0,\vartheta]$, $y_1,y_2\in R^{n(p)}$, $\varepsilon\in(0,1)$

$$|U_p(t,y_1,\varepsilon)-U_p(t,y_2,\varepsilon)|\leqq C\cdot k_p\cdot|B_p(t)|_{n(p)\times k}\cdot|y_1-y_2|/(2C_1\varepsilon)\,.\tag{2.5}$$

Scheme of the proof. Denote

$$F_p(t,y,l,\varepsilon)=k_p\langle y,l\rangle+\langle N_p(t)l,l\rangle-\lambda_p(t)\langle l,l\rangle-\varepsilon\sqrt{k_p^2\langle l,l\rangle+1}\,.$$

Define the sets

$$\begin{aligned} M_{1p}(t,\varepsilon)&=\{y\in R^{n(p)}: l_p(t,y,\varepsilon)=\arg\max\{F_p(t,y,l,\varepsilon): l\in R^{n(p)}\}\},\\ M_{2p}(t,\varepsilon)&=\{y\in R^{n(p)}: |\arg\max\{F_p(t,y,l,\varepsilon): l\in R^{n(p)}\}|\geqq 1\}. \end{aligned} \tag{2.6}$$

It follows from the convexity of the function $F_p(\cdot)$ at the third variable that

$$l_p(t,y,\varepsilon)=\begin{cases}\arg\max\{F_p(t,y,l,\varepsilon): l\in R^{n(p)}\} & (y\in M_{1p}(t,\varepsilon))\\ \arg\max\{F_p(t,y,l,\varepsilon): |l|=1\} & (y\in M_{2p}(t,\varepsilon))\end{cases} \tag{2.7}$$

$$(p\in N, t\in[t_0,\vartheta], y\in R^{n(p)}, \varepsilon\in(0,1)).$$

In view of (2.7) and $F_p(t,\cdot,\varepsilon)\in C^\infty(R^{n(p)}\times R^{n(p)})$ we have

$$\begin{aligned} F_{1p}(t,y,l_p(t,y,\varepsilon),\varepsilon)&=\frac{\partial F_p}{\partial l}(t,y,l_p(t,y,\varepsilon),\varepsilon)=0 \qquad (y\in M_{1p}(t,\varepsilon))\\ F_{2p}(t,y,l_p(t,y,\varepsilon),\varepsilon)&=\frac{\partial F_p}{\partial l}(t,y,l_p(t,y,\varepsilon),\varepsilon)+\\ &+2\mu(l_p(t,y,\varepsilon))l_p(t,y,\varepsilon)=0 \qquad (y\in M_{2p}(t,\varepsilon)). \end{aligned} \tag{2.8}$$

Thus, the vector $l_p(t,y,\varepsilon)$ is described by implicit function on every set (2.6). Then the inequality

$$|l_p(t,y_1,\varepsilon)-l_p(t,y_2,\varepsilon)|\leqq C\cdot|y_1-y_2|/\varepsilon$$

where C does not depend on t, y_1, y_2, ε may be obtained using the theorem on derivative of the implicit function [11]. The inequality and (2.4) cause (2.5).

Proposition 2.4. For every $p\in N$

$$|\rho_p(t,y_1)-\rho_p(t,y_2)|\leqq k_p|y_1-y_2| \quad ((t,y_i)\in[t_0,\vartheta]\times R^{n(p)}, i\in\overline{1,2}). \tag{2.9}$$

Scheme of the proof. Represent the value $\rho_p(\cdot)$ in form of the stochastic program maximin [2]:

$$\rho_p(t,y)=\varlimsup_{\delta\to 0}\sup_{\delta(\Delta)\leqq\delta}\sup_{v\in V(\Delta)}\inf_{f\in F(\Delta)}\int_\Omega\varphi_p(t,y,f(\cdot,\omega),v(\cdot,\omega))\,d\omega,$$

here $F(\Delta)$, $V(\Delta)$ — the sets of all non-anticipatory program of the first and second players (respectively) corresponding to the partition Δ [3]. It follows from the structure of $\varphi_p(\cdot)$ (2.2), that

$$\varphi_p(t,y_1,f(\cdot,\omega),v(\cdot,\omega))\leqq\varphi_p(t,y_2,f(\cdot,\omega),v(\cdot,\omega))+k_p|y_1-y_2|$$

$$(t\in[t_0,\vartheta], y_1,y_2\in R^{n(p)}, f(\cdot)\in F(\Delta), v(\cdot)\in V(\Delta))$$

for all $\omega \in \Omega' \subset \Omega$, where Ω' has full Lebesgue measure in Ω. The inequality produces the sequence of another one

$$\int_{\Omega} \varphi_p(t, y_1, f(\cdot, \omega), v(\cdot, \omega))d\omega \leqq \int_{\Omega} \varphi_p(t, y_2, f(\cdot, \omega), v(\cdot, \omega))d\omega + k_p|y_1 - y_2|$$

$$(t \in [t_0, \vartheta], y_1, y_2 \in R^{n(\rho)}, f(\cdot) \in F(\Delta), v(\cdot) \in V(\Delta)),$$

$$\inf_{f \in F(\Delta)} \int_{\Omega} \varphi_p(t, y_1, f(\cdot, \omega), v(\cdot, \omega))\, d\omega \leqq \inf_{f \in F(\Delta)} \int_{\Omega} \varphi_p(t, y_2, f(\cdot, \omega), v(\cdot, \omega))\, d\omega +$$

$$+ k_p|y_1 - y_2| \qquad ((t \in [t_0, \vartheta], y_1, y_2, \in R^{n(\rho)}, v(\cdot) \in V(\Delta)),$$

$$\sup_{v \in V(\Delta)} \inf_{f \in F(\Delta)} \int_{\Omega} \varphi_p(t, y_1, f(\cdot, \omega), v(\cdot, \omega))\mathrm{d}\omega \leqq \sup_{v \in V(\Delta)} \inf_{f \in F(\Delta)} \int_{\Omega} \varphi_p t, y_2, f(\cdot, \omega), v(\cdot, \omega))\, d\omega +$$

$$+ k_p|y_1 - y_2| \qquad (t \in [t_0, \vartheta], y_1, y_2 \in R^{n(\rho)}),$$

$$\cdots\cdots$$

$$\rho_p(t, y_1) \leqq \rho_p(t, y_2) + k_p|y_1 - y_2| \qquad (t \in [t_0, \vartheta], y_1, y_2 \in R^{n(p)}).$$

In the same manner we obtain

$$\rho_p(t, y_1) \geqq \rho_p(t, y_2) - k_p|y_1 - y_2| \qquad (t \in [t_0, \vartheta], y_1, y_2 \in R^{n(p)}).$$

The last two inequalities cause (2.9).

3. Synthesis of optimal control

For every $p \in N$ fix the linear mappings $\eta_p(\cdot)$: $L_2(Q) \to R^{n(p)}$ and $\chi_p(\cdot): R^{n(p)} \to L_2(Q)$ such that $|\chi_p(y)|_{L_2(Q)} = k_p|y|$ $(y \in R^{n(p)})$, $k_p \in (0, 1]$Denote

$$a(p) = \sup\{|\xi^{[t,\vartheta]}(t, \cdot | t_*, \xi_*(\cdot), u[\cdot], v[\cdot]) - \chi_p(y_p(t | t_*, \eta_p(\xi_*^{[t_*,\vartheta]}(\cdot)),$$

$$u[\cdot], v[\cdot]))|_{L_2(Q)} : t \in [t_*, \vartheta], (t_*, \xi_*(\cdot)) \in D, u[\cdot] \in \mathscr{U}_{[t_*,\vartheta]}, v[\cdot] \in \mathscr{V}_{[t_*,\vartheta]}\}.$$

Proposition 3.1. For all $p \in N$, $(t_*, \xi_*(\cdot)) \in D$, $t \in [t_*, \vartheta]$, $u[\cdot] \in \mathscr{U}_{[t_*,\vartheta]}$, $v[\cdot] \in \mathscr{V}_{[t_*,\vartheta]}$

$$|\eta_p(\xi^{[t,\vartheta]}(t, \cdot | t_*, \xi_*(\cdot), u[\cdot], v[\cdot])) -$$

$$- y_p(t | t_*, \eta_p(\xi_*^{[t_*,\vartheta]}(\cdot)), u[\cdot], v[\cdot])| \leqq 2a(p).$$

Proof. Denote $\xi(t, \cdot) = \xi(t, \cdot | t_*, \xi_*(\cdot), u[\cdot], v[\cdot])$. Then

$$|\eta_p(\xi^{[t,\vartheta]}(t, \cdot)) - y_p(t | t_*, \eta_p(\xi_*^{[t_*,\vartheta]}(\cdot)), u[\cdot], v[\cdot])| =$$

$$= \chi_p(\eta_p(\xi^{[t,\vartheta]}(t, \cdot)) \pm \xi^{[t,\vartheta]}(t, \cdot) - \chi_p(y_p(t | t_*, \eta_p(\xi_*^{[t_*,\vartheta]}(\cdot)), u[\cdot], v[\cdot]))|_{L_2(Q)} \leqq$$

$$\leqq |\chi_p(y_p(t | t, \eta_p(\xi^{[t,\vartheta]}(t, \cdot)), u[\cdot], v[\cdot])) - \xi^{[t,\vartheta]}(t, \cdot | t, \xi(t, \cdot), u[\cdot], v[\cdot])|_{L_2(Q)} +$$

$$+ |\xi^{[t,\vartheta]}(t, \cdot | t_*, \xi_*(\cdot), u[\cdot], v[\cdot]) - \chi_p(y_p(t | t_*, \eta_p(\xi_*^{[t_*,\vartheta]}(\cdot)), u[\cdot], v[\cdot]))|_{L_2(Q)} \leqq 2a(p).$$

Proposition 3.2. For all $p \in N$, $(t_*, \xi_*(\cdot)) \in D$

$$|\rho(t_*, \xi_*(\cdot)) - \rho_p(t_*, \eta_p(\xi_*^{[t_*, \vartheta]}(\cdot)))| \leqq a(p).$$

The proof is similar to Proposition 2.4.

Let be $K = \sup\{k_p |B_p(t)|_{n(p) \times k},\ \ k_p |C_p(t)|_{n(p) \times m} : t \in [t,\ \ \vartheta],\ \ p \in N\}$ and $p(\cdot):(0, 1) \to N$. Define the strategy $U_0(\cdot) \in \mathscr{U}$:

$$U_0(t, \xi(\cdot), \varepsilon) = U_{p(\varepsilon)}(t, \eta_{p(\varepsilon)}(\xi^{[t, \vartheta]}(\cdot)), \varepsilon) \quad ((t, \xi(\cdot), \varepsilon) \in [t_0, \vartheta] \times L_2(Q) \times (0, 1)).$$

Theorem 3.1. If $K < +\infty$, $\mu \geqq K/2C_1$ and $\overline{\lim\limits_{\varepsilon \to 0}}\, a(p(\varepsilon))/\varepsilon = 0$, then the strategy $U_0(\cdot)$ will be D-optimal.

Scheme of the proof. Let $\xi(\cdot)$ be the motion of the system from the position $(t_*, \xi_*(\cdot)) \in D$ produced by the first player control law $\{U_0(\cdot), \varepsilon, \varDelta\}$ $(\varepsilon \in (0, 1)$, $\varDelta = (\tau_i) i \in \overline{1, n})$ and the realization $v[\cdot] \in \mathscr{V}_{[t_*, \vartheta]}$ of the second player control, $u[\cdot]$—realization of the first player control under the motion,

$$y(t) = y_{p(\varepsilon)}(t | t_*,\ \ \eta_{p(\varepsilon)}(\xi_*^{[t_*, \vartheta]}(\cdot)), u[\cdot], v[\cdot]),$$

$$\dot{c}(t) = \beta(t, u[t], v[t]), \quad c(t_*) = 0$$

and

$$(\omega_i, \mathrm{c}_i) = \arg\min\{\rho_{p(\varepsilon)}(\tau_i, \omega) + c : |w - y(\tau_i)|^2 + |c - c(\tau_i)|^2 \leqq \varepsilon^2\} \ (t \in [t_*, \vartheta], i \in \overline{1, n}).$$

Verify that for $i \in 1, n$

$$\begin{aligned} \rho_{p(\varepsilon)}(\tau_i, w_i) + c_i \leqq \rho_{p(\varepsilon)}(t_*, \eta_{p(\varepsilon)}(\xi_*^{[t_*, \vartheta]}(\cdot))) + \\ + (\tau_i - t_*)(K_1 a(p(\varepsilon))/\varepsilon + f_1(\delta(\varDelta), \varepsilon)) \end{aligned} \tag{3.1}$$

where K_1 does not depend on p, ε and $\overline{\lim\limits_{\varepsilon \to 0}}\ \overline{\lim\limits_{\delta \to 0}}\ f_1(\delta, \varepsilon) = 0$. The base of induction follows immediately from the definition of (w_1, c_1). Verify inductive step. Denote

$$y_1(t) = y_{p(\varepsilon)}(t | \tau_i, y(\tau_i), u_1[\cdot], v[\cdot]),$$

$$y_2(t) = y_{p(\varepsilon)}(t | \tau_i, w_i, u_2[\cdot], v[\cdot]),$$

$$\dot{c}_j(t) = \beta(t, u_j[t], v[t]) \quad (j \in \overline{1, 2}),$$

$$c_1(\tau_i) = c(\tau_i),$$

$$c_2(\tau_i) = c_i, u_1[t] = U_{p(\varepsilon)}(\tau_i, y(\tau_i), \varepsilon) \quad (t \in [\tau_i, \tau_{i+1}]).$$

The form (2.3) of the strategy $U_{p(\varepsilon)}(\cdot)$ causes

$$\begin{aligned} (k_{p(\varepsilon)}^2 |y_1(\tau_{i+1}) - y_2(\tau_{i+1})|^2 + (c_1(\tau_{i+1}) - c_2(\tau_{i+1}))^2)^{\frac{1}{2}} \leqq \\ \leqq \varepsilon + (\tau_{i+1} - \tau_i) f_2(\delta(\varDelta), \varepsilon), \end{aligned} \tag{3.2}$$

where $\overline{\lim\limits_{\varepsilon\to 0}}\,\overline{\lim\limits_{\delta\to 0}}\, f_2(\delta,\varepsilon)=0$. It follows from propositions 3.1, 2.3, and from the conditions of the theorem that

$$\begin{aligned}(k_{p(\varepsilon)}^2|y(\tau_{i+1})-y_1(\tau_{i+1})|^2+(c(\tau_{i+1})-c_1(\tau_{i+1}))^2)^{\frac{1}{2}}\leqq\\ \leqq(\tau_{i+1}-\tau_i)K_2\cdot a(p(\varepsilon))/\varepsilon,\end{aligned}\tag{3.3}$$

where K_2 does not depend on p, ε. Inequalities (3.2), (3.3), and proposition 2.4 provide

$$\begin{aligned}\rho_{p(\varepsilon)}(\tau_{i+1},w_{i+1})+c_{i+1}\leqq\rho_{p(\varepsilon)}(\tau_{i+1},y_2(\tau_{i+1}))+\\ +c_2(\tau_{i+1})+(\tau_{i+1}-\tau_i)\cdot(K_1 a(p(\varepsilon))/\varepsilon+f_1(\delta(\Delta),\varepsilon)).\end{aligned}\tag{3.4}$$

On the other hand [2], $u_2[\cdot]$ may be chosen such that

$$\rho_{p(\varepsilon)}(\tau_{i+1},y_2(\tau_{i+1}))+c_2(\tau_{i+1})\leqq\rho_{p(\varepsilon)}(\tau_i,w_i)+c_i.\tag{3.5}$$

Inequalities (3.4), (3.5), and inductive assumption complete the induction. The definition of (w_n, c_n) and (3.1) (where $i=n$) give

$$\begin{aligned}k_{p(\varepsilon)}|y(\vartheta)|+c(\vartheta)\leqq\rho_{p(\varepsilon)}(t_*,\eta_{p(\varepsilon)}(\xi_*^{[t_*,\vartheta]}(\cdot)))+\\ +c_3\varepsilon+(\vartheta-t_0)(K_1 a(p(\varepsilon))/\varepsilon+f_1(\delta(\Delta),\varepsilon)),\end{aligned}$$

whence (see proposition 3.2, and definitions of $\chi_p(\cdot)$, and $a(p)$)

$$|\xi(\vartheta)|_{L_2(Q)}+c(\vartheta)\leqq\rho(t_*,\xi_*(\cdot))+f_3(\delta(\Delta),\varepsilon),$$

where $\overline{\lim\limits_{\varepsilon\to 0}}\,\overline{\lim\limits_{\delta\to 0}}\, f_3(\delta,\varepsilon)=0$. The inequality is equivalent to the conclusion of the theorem.

Example 4.1. Let $\{\lambda_i : i\in N\}$ be the set of all eigenvalues of the operator $\mathscr{A}$, $\mathscr{A}g_i(\cdot)=\lambda_i g_i(\cdot)$, $g_i(x)=0\ (x\in\partial Q, i\in N)$ and $\{g_i(\cdot):i\in N\}$ is orthonormed set in $L_2(Q)$. For every $p\in N$ let $n(p)=p$

$$\chi_p(y)=\sum_{i=1}^{p} y_i g_i(\cdot)\quad (y\in R^p),\tag{4.1}$$

$$\eta'_p(\xi(\cdot))=(\langle\xi(\cdot),g_1(\cdot)\rangle_{L_2(Q)},\ldots,\langle\xi(\cdot),g_p(\cdot)\rangle_{L_2(Q)})\quad (\xi(\cdot)\in L_2(Q))$$

(hence $k_p=1$) and system (2.1) has the form

$$\begin{cases}\dot y_{pi}(t)=\exp(\lambda_i(t-t_*))(\sum\limits_{j=1}^{k} g_i(x_{u_j})u_j[t]+\sum\limits_{j=1}^{m} g_i(x_{v_j})v_j[t])\\ \qquad\qquad (t\in[t_*,\vartheta], i\in\overline{1,p}),\\ y_p(t_*)=y_*\in R^p,\qquad t_*\in[t_0,\vartheta].\end{cases}\tag{4.2}$$

Let be $K_3, K_4 \in R^1$, such that [8, p. 222], [10, p. 98]

$$|\psi(\cdot)|_{W^2_{2,0}(Q)} \leqq K_3 |(\mathscr{A}\psi)(\cdot)|_{L_2(Q)}, |\psi(\cdot)|_{C^0(Q)} \leqq K_4 |\psi(\cdot)|_{W^2_{2,0}(Q)}$$

$$(\psi(\cdot) \in W^2_{2,0}(Q)),$$

$\alpha > 0$ and $p(\varepsilon) = \inf\{p \in N : p \geqq \varepsilon^{-1}\}$. If

$$\mu \geqq \frac{K_3 \cdot K_4}{2C_1}(m+k) \sum_{i=1}^{\infty} \lambda_i \exp(\lambda_i \varepsilon_\vartheta),$$

than the conditions of theorem 3.1 will be fulfilled.

Remark 4.1. The statement of theorem 3.1 may be improved for the auxiliary systems of the form (4.2) and the mappings (4.1).

It follows [2] from the definition of the first player optimal strategy $U_p(\cdot)$ in the game Γ_p, that for every $p \in N$ there exist $f_p(\cdot):(0,1)\times(0,1)\to(0,1)$ with properties

$$\overline{\lim_{\varepsilon\to 0}}\, \overline{\lim_{\delta\to 0}}\, f_p(\delta, \varepsilon) = 0$$

$$\varphi_p(t_*, y_*, u(\cdot\,|t_*, y_*, \{U_p(\cdot), \varepsilon, \Delta\}, v[\cdot]), v[\cdot]) \leqq \rho_p(t_*, y_*) + f_p(\delta(\Delta), \varepsilon)$$

$$(t_* \in [t_0, \vartheta], y_* \in R^{n(p)}, \varepsilon \in (0,1), v[\cdot] \in \mathscr{V}_{[t_*, \vartheta]})$$

where Δ — any partition of the segment $[t_*, \vartheta]$, $u(\cdot\,|t_*, y_*, \{U_p(\cdot), \varepsilon, \Delta\}, v[\cdot])$ — first player control realization under the motion $y_p(\cdot\,|t_*, y_*, \{U_p(\cdot), \varepsilon, \Delta\}, v[\cdot])$. Define the functions $p(\cdot):(0,1)\to N$, $\alpha(\cdot):(0,1)\to(0,1)$, such that

$$\lim_{\varepsilon\to 0} p(\varepsilon) = +\infty \qquad \overline{\lim_{\varepsilon\to 0}}\, \overline{\lim_{\delta\to 0}}\, f_{p(\varepsilon)}(\alpha(\varepsilon), \delta) = 0.$$

Then the strategy $\tilde{U}(\cdot) \in \mathscr{U}$ of the form

$$\tilde{U}(t, \xi(\cdot), \varepsilon) = U_{p(\varepsilon)} = U_{p(\varepsilon)}(t, \eta_{p(\varepsilon)}(\xi^{[t,\vartheta]}(\cdot)), \alpha(\varepsilon))$$

$$((t, \xi(\cdot), \varepsilon) \in [t_0, \vartheta] \times L_2(Q) \times (0,1))$$

is D-optimal.

Example 4.2. For every $p \in N$ define $r_p > 0$, $\lim\limits_{p\to\infty} r_p = 0$

$$\delta_p(\cdot) \in c^\infty R^{\bar{n}}, \operatorname{supp}(\delta_p(\cdot)) = \{x \in R^{\bar{n}} : |x| \leqq r_p\},$$

$$\int_{R\bar{n}} \delta_p(x)\,dx = 1, \quad \delta_p(x) \geqq 0, \quad \delta_p(x) = \delta_p(-x) \qquad (x \in R^{\bar{n}}),$$

as well $h_p > 0$, $\{0\omega_p(i) : i \in 1, n(p)\}$ — (in terms of [9]) partition of the domain Q on rectangles with edges h_p, $\Delta p = \mathrm{mes}(\omega_p(i))\,(i = \overline{1, n(p)})$,

$$\chi_p(y)(x) = y_i \qquad (x \in \omega_p(i),\ y \in R^{n(p)}),$$

$$\eta'_p(\xi(\cdot)) = \Big(\Delta_p^{-1} \int\limits_{\omega_p(1)} \xi(x)dx, \ldots, \Delta_p^{-1} \int\limits_{\omega_p(n(p))} \xi(x)dx\Big) \qquad (\xi(\cdot) \in L_2(Q))$$

(hence $k_p = \Delta p$), $\mathscr{A}p$ is difference operator under the network Q_{hp} [9]:

$$(\mathscr{A}_p y_p)(k) = \sum_{i,j=1}^{\bar{n}} (a_{pij} y_{x_j})_{\bar{x}_i}(k) + a_p(k) y_p(k)$$

$$a_{pij}(k) = \Delta_p^{-1} \int\limits_{\omega_p(k)} a_{ij}(x)dx\,,$$

$$a_p(k) = \Delta_p^{-1} \int\limits_{\omega_p(k)} a(x)\,dx \quad (k \in \overline{1, n(p)},\ i, j \in \overline{1, \bar{n}}).$$

For every $p \in N$ define system (2.1) by the following way:

$$\begin{cases} \dot{y}_p(t) = X_p(\vartheta, t)(B_p u[t] + C_p v[t]) & (t \in [t_*, \vartheta]), \\ y_p(t_*) = y_* \in R^{n(p)}, & t_* \in [t_0, \vartheta], \end{cases}$$

where $B_p = [b_{1p}, \ldots, b_{kp}]$, $C_p = [c_{1p}, \ldots, c_{mp}]$,

$$b'_{ip} = \Big(\Delta_p^{-1} \int\limits_{\omega_p(1)} \delta_p(x - x_{u_i})\,dx, \ldots, \Delta_p^{-1} \int\limits_{\omega_p(n(p))} \delta_p(x - x_{u_i})\,dx\Big),$$

$$c'_{jp} = \Big(\Delta_p^{-1} \int\limits_{\omega_p(1)} \delta_p(x - x_{v_j})\,dx, \ldots, \Delta_p^{-1} \int\limits_{\omega_p(n(p))} \delta_p(x - x_{vj})\,dx\Big),$$

$(i \in \overline{1, k}, j \in \overline{1, m})$, $X_p(\cdot, \cdot)$ — $n(p) \times n(p)$-matrix-function described by the matrix equation

$$\dot{X}_p(t, \tau) = \mathscr{A}_p X_p(t, \tau) \quad (t \in R^1), \qquad X_p(\tau, \tau) = E \quad (r \in R^1).$$

For the given sequence $(\delta_p(\cdot))$ there may be chosen the sequence (h_p) and the low boundary for μ such that the conditions of theorem 3.1 will be fulfilled.

Acknowledgement

The author is grateful to N. N. Krasovskii for the problem has been given and attention was paid, A. V. Krjažimskii, Ju. S. Osipov, V. E. Tret'jakov for the discussions and fruitful comments.

References

1. *Krasovskii, N. N.*, *Subbotin, A. I.*, Closed-loop differential games. Nauka, Moscow, 1974 (in Russian).
2. *Krasovskii, N. N.*, *Tret'jakov, V. E.*, One optimal control problem with ensured result. Izv. Akad. Nauk SSSR, Ser. Tehn. kibernet., 1983, *2*, pp. 6–23 (in Russian).
3. *Serkov, D. A.*, Synthesis of optimal feedback control and stochastic maximin. Probl. Control and Inform. Theory, 1985, **14,** *4.*
4. *Krasovskii, N. N.*, On differential evolutionary systems. Prikl. Mat. i Meh., 1977, **41,** *5*, pp. 774–782 (in Russian).
5. *Korotkii, A. I.*, *Osipov, Ju. S.*, Approximation in problems of feedback control for a parabolic systems. Prikl. Mat. i Meh., 1978, **42,** 4, pp. 599–605 (in Russian).
6. *Korotkii, A. I.*, Feedback control problems for systems with distributed parameters. Diss. Sverdlövsk, 1981 (in Russian).
7. *Lions, J. L.*, Control optimal de systèmes gouvernés par des équations aux dérivées partielles. Dunod Gauthier-Villars, Paris, 1968 (in French).
8. *Ladyženskaja, O. A.*, *Ural'ceva, N. N.*, Linear and quasilinear equations of elliptic type. Nauka, Moscow, 1973 (in Russian).
9. *Ladyženskaja, O. A.*, Boundary value problems of mathematical physics. Nauka, Moscow, 1973 (in Russian).
10. *Adams, R. A.*, Sobolev spaces. Acad. Press, New York, 1975.
11. *Schwartz, L.*, Analysis. Vol. **1.** Mir, Moscow, 1972 (in Russian).
12. *Sansone, G.*, Ordinary differential equations. Vol. **1.** Inostr. Liter., Moscow, 1954 (in Russian).

Синтез позиционного управления параболической системой и конечномерные модели

Д. А. СЕРКОВ

(Свердловск)

Рассматривается задача построения позиционного управления для параболической системы. Управление и помеха сосредоточены в конечном числе внутренних точек области изменения пространственных переменных. Функционал платы имеет терминальную и (квадратичную) интегральную части. В работе строится универсальная оптимальная позиционная стратегия первого (минимизирующего) игрока на основе таких же стратегий во вспомогательных дифференциальных играх для конечномерных аппроксимирующих систем. Приведены два способа построения подобных аппроксимирующих моделей: первый основывается на методе Бубнова–Галеркина, второй — на методе прямых.

Д. А. Серков
Институт математики и механики УНЦ АН СССР,
СССР, 620219, Свердловск,
ГСП-384, ул. С. Ковалевской, 16.

Problems of Control and Information Theory, Vol. 15 (3), pp. 253—264 (1986)

SOME CONSTRUCTIONS OF SIGNATURE SEQUENCES FOR *T*-USER MULTIPLE-ACCESS ADDER CHANNELS WITH CORRELATION TYPE RECEIVER

NGUYEN QUANG A

(*Hanoi*)

(Received May 5, 1985)

The signature problem for a *T*-user multiple-access adder channel is the following: to each of *T* users in the network a ±1-valued sequence of length *N*, the *signature sequence*, is assigned. If the user is active he sends his sequence, otherwise he keeps silent. The common receiver has to identify the active users based on the output of the channel which is the sum of the sequences of active users. It is assumed that the number of active users is at most *M* at any time. Some constructions of signature sequences for *T*-user multiple-access adder channel are given for bit and frame synchronous cases with correlation type receivers. Examples of constructions based on the Kerdock and dual BCH codes are given.

1. Introduction

Consider the following model of a *T*-user multiple access adder channel. The channel has *T* inputs and one output, if $z_1, z_2, \ldots, z_T$ stand for the inputs then the output of the channel is

$$y = \sum_{i=1}^{T} z_i \, .$$

The system is assumed to be in bit and block synchronization. Let $\beta = (\beta_1, \beta_2, \ldots, \beta_T)$ be the activity vector of the network, i.e. β_i is 1 if the *i*-th user is active, otherwise it is 0. The signature problem can be stated simply as follows. To each of *T* users in the network a ±1-valued sequence of length *N*, the *signature sequence*, is assigned. If the user is active he sends his sequence, otherwise he keeps silent. The common receiver has to identify the active users based on the output of the channel which is the sum of the sequences of active users. It is assumed that the number of active users is at most *M* at any time.

If $X_i = (X_{i1}, X_{i2}, \ldots, X_{iN})$ denotes the signature sequence of the *i*-th user then the output sequence over a frame is the following

$$Y = \sum_{i=1}^{T} \beta_i X_i \, . \tag{1}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

The common receiver knowing all signature sequences has the task to estimate the activity vector based on the output sequence Y. If the receiver is of correlation type then it has to estimate β based only on the correlations between the output Y and the signature sequences.

The signature problem appears in many communication tasks, for example:

— alarming,
— monitoring,
— telemetring,
— dialing,
— demans transmission for random access with reservation.

Signature problem for binary adder channel, where X_{ij} takes values 1 or 0, has been treated at some length in [1], that for OR channel and collision channel without feedback has been investigated by several authors [2], [3], [4], [5], [6]. In [1] we have shown how to use signature sequences for both identification of active users as well as for transmission of messages of active users.

Some constructions of signature sequences for T-user multiple access adder channel are presented. The common receiver, or separated autonomous receivers, are assumed to be of correlation type.

2. Construction of signature sequences

The correlation type receiver performs

$$R_i = \langle Y, X_i \rangle \tag{2}$$

where $\langle\ .\ ,\ .\ \rangle$ denotes the inner product. The decoder has to make decision based only on the correlation vector $R=(R_1, \ldots, R_T)$. Let A denote the set of active users, i.e. A is a subset of

$$(1, 2, \ldots, T)\,.$$

By assumption $|A| \leqq M$. From (1) it follows that

$$R_i = \sum_{j \in A} \langle X_j, X_i \rangle\,. \tag{3}$$

The *decision rule* reads as follows.

$$\beta'_i = 1 \qquad \text{if} \quad R_i \geqq K; \qquad \text{otherwise} \quad \beta'_i = 0 \tag{4}$$

where β'_j is the estimate of β_i and K is the threshold level to be chosen according to some optimal criteria.

There are two types of errors:

(i) false alarm occurs when $\beta_i=0$ but $R_i \geqq K$;
(ii) misdetection occurs when $\beta_i=1$ but $R_i<K$.

Let

$$\mu_{ij}=\langle X_i, X_j\rangle \tag{5}$$

$$\mu_m=\min_{i\neq j} \mu_{ij} \tag{6}$$

$$\mu_M=\max_{i\neq j} \mu_{ij}\,. \tag{7}$$

Our construction is based on the following simple fact.

Lemma. The sequences of distinct X_i, $1\leqq i\leqq T$, are error free signature sequences of length N for T users and

$$M=\min\left\{\left\lceil \frac{K}{\mu_M}\right\rceil-1, \left\lfloor \frac{N-K}{|\mu_m|}\right\rfloor+1\right\} \tag{8}$$

where K is being chosen to maximize M, $\lfloor x\rfloor$ is the greatest integer less than or equal to x, $\lceil x\rceil$ is the smallest integer greater than or equal to x.

Proof. We shall show that neither false alarm nor misdetection can occur if the conditions of the lemma are satisfied. First, suppose that i-th user is nonactive, i.e. $i\notin A$, then

$$R_i=\sum_{j\in A}\mu_{ij}\leqq \mu_M|A|\leqq \mu_M M\,.$$

By (8) it follows that $R_i<K$ so false alarm cannot occur.

Now suppose that i-th user is active then

$$R_i=N+\sum_{j\in A,\, j\neq i}\mu_{ij}\geqq N+(M-1)\mu_m\geqq N-(M-1)\,|\mu_m|\,.$$

This and (8) imply that $R_i\geqq K$ so misdetection cannot occur. The proof is complete.

Our construction can be described as follows. Let **C** be a binary error-correcting code of length N, consisting of T code words. The binary code word $x=(x_{i1}, \ldots, X_{iN})$ is transformed into sequence $X_i=(X_{i1}, \ldots, X_{iN})$ as follows

$$X_{ij}=1-2x_{ij}\,.$$

If d_{ij} denotes the Hamming distance between x_i and x_j then it can be seen easily that

$$\mu_{ij}=N-2d_{ij}\,. \tag{9}$$

Let d_m and d_M denote the minimal and maximal Hamming distances of the code **C** then

$$\mu_M = N - 2d_m \tag{10}$$

$$\mu_m = N - 2d_M\,. \tag{11}$$

From (8), (10) and (11) we see that a binary code for which $d_m < N/2$ and both d_m and d_M are near to $N/2$ can be used to generate a good signature code. Some examples are given in the next section.

3. Examples

Our goal is to find a large set of sequences with length as short as possible that ensures a large number of simultaneous active users, i.e. we look for sets of T sequences of length N ensuring M simultaneous active users. Such a set is denoted by $S(T, M, N)$, and for fixed T and M, it is desired that N is as small as possible. By (8) this problem is equivalent to find a set of sequences with good auto- and crosscorrelation properties [8]. In the sequel we investigate some examples of such sequences.

Example 1. By (8), (10), (11), if we use a Hadamard matrix of size $T \times T$ [7, pp. 44–54] then $\mu_m = \mu_M = 0$, i.e. it is an $S(T, T, T)$ signature code. We can use distinct shift of an m-sequence [7, pp. 406–412], [8] to get an $S(T, T, T)$ code for any $T = 2^m - 1$, in this case $\mu_m = \mu_M = -1$ and K can be chosen to be $-N$.

In practice we have a large population of T potential users and at any given time only a small percentage of users are active, $M \ll T$, in such a situation N can be much smaller than T. It is worth to note that if the receiver is not required to be of correlation type then even for the case $M = T$ the length N of sequence can be much shorter than T; in fact, Lindström [1] has shown that

$$N \sim 2T/\log T$$

where log denotes logarithm of base two. It is not clear yet whether this remains true for a correlation type receiver.

Example 2. Let $\mathbf{K}(2m)$ be the nonlinear Kerdock code of length $N = 2^{2m}$ [7, pp. 453–465]. Let **C** be subset of **K** such that if a code word is in **C** then its complement does not belong to **C**. Then **C** generates an $S(T, M, N)$ signature code with

$$\begin{aligned} T &= 2^{4m-1} \\ M &= 2^{m-1} \\ N &= 2^{2m} \end{aligned} \tag{12}$$

for all $m \geq 2$.

This can be seen as follows. The code **K** contains 2^{4m} code words and its distance distribution is given by [7, p. 456]

i	A_i
0	1
$2^{2m-1}-2^{m-1}$	$2^{2m}(2^{2m-1}-1)$
2^{2m-1}	$2^{2m+1}-2$
$2^{2m-1}+2^{m-1}$	$2^{2m}(2^{2m-1}-1)$
2^{2m}	1

where A_i is the ratio of the number of pairs of code words with distance i to the number of code words. Since if $d_{ij}=N$ then X_i and X_j must be complement of each other and there are exactly 2^{4m} pairs having distance N in **K**, no pair of code words in **C** can have distance N. In other words in **C**

$$d_M=2^{2m-1}+2^{m-1}; \qquad d_m=2^{2m-1}-2^{m-1}.$$

Choosing $K=d_M$ we get (12). Parameters of $S(T, M, N)$ for some m are given below.

m	2	3	4	5
T	128	2048	32 768	334 288
M	2	4	8	16
N	16	64	256	1024

It is worthwhile to mention that the construction of this example works also by using the Delsarte–Goethals codes [7, pp. 461–465] which are generalized Kerdock codes.

Example 3. Let **B** be a binary BCH code of length $N=2^m-1$ with designed minimal distance

$$\begin{gathered} d=2t+1 \\ t<2^{\lceil m/2\rceil-1}+1. \end{gathered} \tag{18}$$

Let **C** be the dual code of **B** with all-zero code word deleted. Then **C** generates signature code $S(T, M, N)$ for synchronous adder channel, where

$$\begin{aligned} T&=2^{tm}-1 \\ M&=\left\lfloor \frac{2^{m-1}+(t-1)2^{m/2}}{(t-1)2^{m/2+1}-1} \right\rfloor \\ N&=2^m-1. \end{aligned} \tag{14}$$

This can be seen as follows.

By Carlitz–Uchiyama bound [7, p. 280] for any $x \in \mathbf{C}$ the weight $wt(x)$ satisfies

$$2^{m-1}-(t-1)2^{m/2} \leqq wt(x) \leqq 2^{m-1}+(t-1)2^{m/2} . \tag{15}$$

Since the dual code of a BCH code is linear code, the weight distribution and distance distribution are the same; so the maximal and minimal distance of $\mathbf{C}$ satisfy

$$\begin{aligned} d_M &\leqq 2^{m-1}+(t-1)2^{m/2} \\ d_m &\geqq 2^{m-1}-(t-1)2^{m/2} . \end{aligned} \tag{16}$$

If $t=1$, choose $K=0$, otherwise choose K to maximize M in (8) we get (14). The number of code words of dual code is [7, p. 263] $2^{tm}=T+1$. For some t the bounds can be sharpened [7, pp. 449–452, 869] and for such cases, $t=2$ or $t=3$, M may be larger than given by (14). For example, if $t=2$ then M is given by

$$M=2^{\lfloor (m-3)/2 \rfloor} . \tag{17}$$

Parameters of $S(T, M, N)$ for some m are given below for $t=2$.

m	5	6	7	8	9
T	1023	4095	16 383	65 535	262 143
M	2	2	4	4	8
N	31	63	127	255	511

Remarks

1. The Lemma gives sufficient conditions for a set of sequences to be an $S(T, M, N)$ code, but these conditions are not necessary. Taking this into account and the fact that bounds on maximal and minimal distances might be not tight, in general, the number of simultaneous active users M is larger than given by (8).

2. All sequences having good correlation properties as Gold sequences, Kasami sequences [8] can be used as signature codes for asynchronous adder channel. For example, the large set of Kasami sequences of length 255 generated by polynomial 6 031 603 [8, Table II] has 4111 sequences and $\mu_M=31$, $\mu=-33$. Choosing, for example, $K=124$ we have a signature code S (4111, 4, 255).

3. The following simple procedure shows how to use signature sequences for solving both switching (active user identification) and information transmission functions [1]. Suppose there are T potential users in the network and its is known that at any given time there are at most M active users. Consider a signature code $S(nT, M, N)$ and distribute to each user n code words, i.e. the source alphabet of each

user consists of n letters. By increasing the size of source alphabet n it has been shown that the cost of switching function can be made as small as one wishes compared with the cost of information transmission.

Acknowledgement

The author would like to thank L. Györfi for useful discussions and to a reviewer for valued suggestions.

References

1. *Nguyen, Q. A., Györfi, L.*, On Signature Coding for Multiple Access Binary Adder Channel of Changing Population, to be published.
2. *Massey, J. L.*, The Capacity of the Collision Channel without Feedback, IEEE Int. Symp. Inf. Th., Les Arcs, France, 1982.
3. *Massey, J. L., Mathys, P.*, *M*-user Collision Channel without Feedback, Technical Report, ETH Zurich, April 1983.
4. *Tybakov, B. S., Likhanov, N. B.*, Packet Communication on a Channel without Feedback, Problems of Information Transmission, vol. **19,** No. *2*, pp. 69–84, 1983.
5. *Bassalygo, L. A., Pinsker, M. S.*, Restricted Multiple-Access of a Nonsynchronous Channel, Problems of Information Transmission, vol. **19,** No. *4*, pp. 92–96, 1983.
6. *Nguyen, Q. A., Györfi, L., Massey, J. L.*, Some Construction of Protocol Sequences for Collision Channel and a Class of Optimal Cyclic Constant Weight Codes, IEEE Int. Symp. Inf. Theory, Brighton, England 1985.
7. *MacWilliams, F. J., Sloane, N. J. A.*, The Theory of Error-Correcting Codes, North-Holland Publishing Company, Amsterdam 1977.
8. *Sarwate, D. W., Pursley, M. B.*, Crosscorrelation Properties of Pseudorandom and Related Sequences, Proc. IEEE, vol. **68,** No. *5*, pp. 593–619, May 1980.

Некоторые конструкции сигнатурных кодов для суммирующего канала с T пользователями

НГУЕН КВАНГ А
(Ханой)

Рассмотрена задача выделения активных источников в аддитивном суммирующем канале. В системе с T источниками одновременно ведут передачу M или менее источников, $M \leqq T$. Работающие источники передают фиксированные последовательности длины $N \pm 1$ (сигнатурных кодов). В канале производится сложение в поле действительных чисел предаваемых последовательностей, а приемник осуществляет корреляционный прием с целью выделения номеров i_1, i_2, i_s всех $s \leqq n$ работающих источников. Рассмотрена задача построений набора из T последовательностей, позволяющего максимизировать число M одновременно работающих источников.

Nguyen Quang A
Institute of Communication and Electronics,
Technical University of Budapest,
H-1111 Budapest, XI. Stoczek u. 2.
Hungary
on leave from the Institute of Computer Science and Cybernetics,
Vietnamese Academy of Sciences

Problems of Control and Information Theory, Vol. 15 (3), pp. 261—264 (1986)

AN EVALUATION OF ADAPTIVE MTI FILTERS

NGUYEN TAI CAN

(*Hanoi*)

(Received April 10, 1985)

The evaluation of adaptive MTI filters, the adaptive cancellers using close-loop is dealt with. The evaluation is based on the consideration of autocorrelation function of the clutters at the output of the filter.

Introduction

The signal processing of the correlated clutters requires an operation of decorrelation of clutters. A decorrelation equipment such as the conventional decorrelator, which separates target signal from the interfering clutters using multiple-delay lines canceller, has low effectiveness because in such equipment statistical characteristics of the clutters are not taken into acount. Conventional MTI (Moving Target Indicator) systems have been designed on the basis of average properties of the clutters. Statistical characteristics of clutters are often not known apriori, so such clutter can be suppressed only adaptively.

In this paper the evaluation of the correlation function of clutter at the output of the adaptive filter using close-loop is described.

Correlation function of the clutter at the filter output

The filter using close-loop (the canceller with correlation feedback) is demonstated in Fig. 1, where 1 is a delay line of length T (length of repetition period); 2 is the multiplier; 3 is the comulative adder; 4 is the adder; * denotes the complex conjugation; $X_{i,\,j}$ is the clutter backscatter from the i-th range cell in the j-th radiation period.

Clutters backscatter is assumed to have normal distribution with zero mean which can be justified by the argument of the central limit theorem. The filter works with the following algorithm [1]:

$$\begin{aligned} Y_i &= X_{i,1} - w_{i-1} X_{i,2} \\ w_{i-1} &= w_{i-2} + \gamma X^*_{i-1,2}\, Y_{i-1} \end{aligned} \tag{1}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

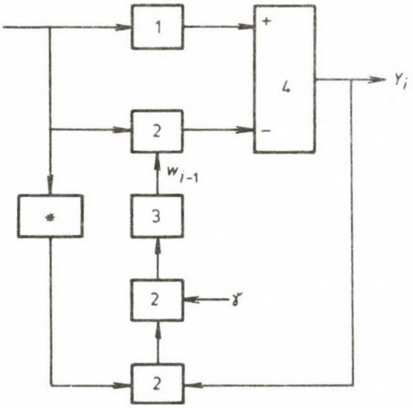


Fig. 1

where Y_i is the clutter at the filter output in the i-th adaptive step,
w_i is the weight coefficient used in the i-th adaptive step,
γ is the adaptive coefficient chosen to be [1] $0<\gamma<\alpha^{-2}$,
α^2 is the power of the input clutter,
* denotes the complex conjugation.

Assume that $w_0=0$, we have the following general expression for the weight coefficient:

$$w_{i-1}=\gamma\left[X_{i-1,1}X^*_{i-1,2}+\sum_{n=1}^{i-2}X_{n,1}X^*_{n,2}\prod_{k=n+1}^{i-2}(1-\gamma|X_{k,2}|^2)\right]. \tag{2}$$

Correlation function of clutter at the filter output in the i-th adaptive step through m radiation periods is as follows:

$$B(m)=E[Y_i Y^*_{i/m}] \tag{3}$$

where

$$Y_{i/m}=X_{i,m+1}-w_{i/m-1}X_{i,m+2} \tag{4}$$

$$w_{i/m-1}=\left[X_{i-1,m+1}X^*_{i-1,m+2}+\sum_{n=1}^{i-2}X_{n,m+1}X^*_{n,m+2}\prod_{k=n+1}^{i-1}(1-\gamma|X_{k-2}|^2)\right]$$

and $E[\cdot]$ denotes the expectation.

From (1), (2), (3) we find that $B(m)$ depends on the couple of $X_{i,j}$, i.e., on $X_{i,j}$ and $X_{i,j+1}$. These couples have the following probability density:

$$p(X_{i,j},X_{i,j+1})=\frac{1}{2\pi\alpha^2(1-r^2)^{1/2}}\exp\left[\frac{(X_{i,j}-rX_{i,j+1})^2}{2\alpha^2(1-r^2)}-\frac{X^2_{i,j+1}}{2\alpha^2}\right]$$

where r is the clutter correlation coefficient.

As clutter backscatters from various range cells can be considered statistically independent so the clutter probability density from N range cells in the two consecutive moments j and $j+1$ is as follows:

$$p(u)=\prod_{i=1}^{2N}\frac{1}{2\pi\alpha^2(1-r^2)^{1/2}}\exp\left[\frac{(X_{i,j}-rX_{i,j+1})^2}{2\alpha^2(1-r^2)}-\frac{X^2_{i,j+1}}{2\alpha^2}\right] \tag{5}$$

where α is the vector of sequences $X_{i,j}$ backscattering from N range cells.

Suppose that fluctuations of input clutter and weight coefficient are independent, from (1)–(5) we have

$$B(m)=\alpha^2[R(m)-rF(1-p^{i-1})+R(m)V(1-q^{i-1})/(1-q)+$$
$$+Z(1-q^{i-2})/(1-q)\beta^{-1}-Zp\beta^{-1}(p^{i-2}-q^{i-2})/(p-q)] \tag{6}$$

where

$\beta=\gamma\alpha^2$ is the reduced adaptive coefficient;

$R(m)$ is the clutter correlation coefficient through m period;

$$F=R(m+1)+R(m-1)$$

$$p=1-\beta$$

$$V=\beta^2[r^2+R^2(m)]$$

$$q=1-2\beta+\beta^2[1+R(m)]$$

$$Z=R(m)\beta^2[2r^2(1-\beta)-rR(m)\beta F].$$

The normalized correlation function of clutter at the filter output is as follows:

$$r(m)=B(m)/B(0). \tag{7}$$

It is worthwhile to mention that by setting $m=0$ from (6) we have

$$B(0)=\alpha^2\left\{1-\frac{2r^2-\beta(1+r^2)}{2(1-\beta)}[1-(1-2\beta+2\beta^2)^{i-1}]\right\}. \tag{8}$$

This is just the average power of the output clutter, an earlier result of [1].

From (6), (7), (8) we have

$$r(m)=\frac{R(m)-rF(1-p^{i-1})+R(m)V\dfrac{1-q^{i-1}}{1-q}+Z\dfrac{1-q^{i-2}}{(1-q)\beta}-Z\dfrac{N(p^{i-2}-q^{i-2})}{\beta(p-q)}}{1-\dfrac{2r^2-\beta(1+r^2)}{2(1-\beta)}[1-(1-2\beta+2\beta^2)^{i-2}]}.$$

The normalized correlation functions are presented in Fig. 2 for various values of i, m and β.

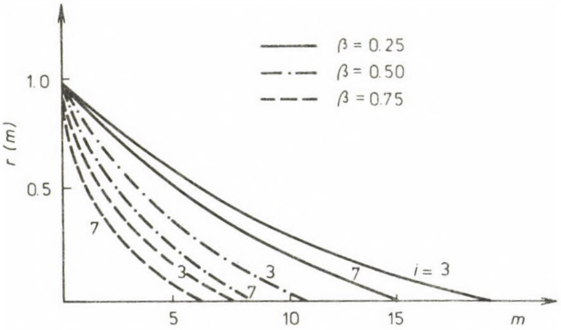


Fig. 2

Based on our analytic results and the numerical results the following conclusions can be made.

The normalized correlation function of the output clutter depends on (cf. (9))

— the reduced adaptive coefficient,

— the processing time *m*,

— the amount of informations processed *i*, i.e., the number of range cells.

The numerical results show that the reduced adaptive coefficient has strong effect on the whitening ability of the input clutter. As β increases the whitening ability of the filter increases. On the other hand from Eq. (8) the clut0ter supression ability of the filter [1] and the dependency of this ability on the reduced adaptive coefficient can be determined. There is trade-off between the whitening and clutter supression abilitiies of the filter, i.e., there exists an optimal reduced adaptive coefficient if both whitening and clutter supression abilities of the filter are considered.

References

1. *Likharev, V. A., Panov, V. M.*, Analiz effektivnosti avtokompensatora pomekh *s* KOC po koefficienty podavlenia, Radioelektronika **XXV**, *11*, 1982, p. 75.
2. *Levin, B. R.*, Teoretitseskie osnovy statitseskoi radiotekhniki, M., Sov. Radio, 1969

Анализ адаптивных MTI фильтров

НГУЕН ТАИ КАН

(Ханой)

Рассмотрены анализы адаптивных MTI (Moving Target Indicator) фильтров. Анализ основывается на автокорреляционой функции выходного сигнала — сигнала фильтра.

Nguyen Tai Can
Technical University of Hanoi
SRV

РУССКИЙ ПЕРЕВОД

Проблемы Управления и Теории Информации, том **15**, *номер 3 (1986)*

УПРАВЛЕНИЕ ПРИ ДЕФИЦИТЕ ИНФОРМАЦИИ

Н. Н. Красовский, С. И. Тарасова, В. Е. Третьяков, Г. И. Шишкин

(*Свердловск*)

Указывается алгоритм оптимального управления динамической системой при наличии динамических и информационных помех.

1. Введение

Рассматривается один класс задач оптимального управления по принципу обратной связи на минимум гарантированного результата в условиях неконтролируемой динамической помехи и при дефиците информации о начальном и текущем фазовом состоянии управляемого объекта. Этот дефицит состоит в том, что сведения о фазовом векторе доставляются не по всем координатам и притом еще с искажением. Показывается, что для рассматриваемого класса задач сохраняется алгоритм, разработанный в [1] при полной информации о фазовых состояниях объекта.

2. Постановка задачи

В [1] решена задача оптимального управления на минимум гарантированного результата по показателю

$$\begin{aligned}&\gamma_x(x[\cdot], u[\cdot], v[\cdot]) = \\ &= |x[\vartheta]| + \int_{t_0}^{\vartheta} (u'[t]\Phi(t)u[t] - v'[t]\Psi(t)v[t])dt\end{aligned} \tag{2.1}$$

для объекта, описываемого дифференциальным уравнением

$$\dot{x} = A(t)x + B(t)u + C(t)v, \qquad t_0 \leqq t \leqq \vartheta. \tag{2.2}$$

Здесь $x \in R^n$ — фазовый вектор, $u \in R^r$ — управление, $v \in R^s$ — неизвестная динамическая помеха — трактуются как векторы-столбцы, знак штрих означает транспонирование; $u'\Phi(t)u$, $v'\Psi(t)v$ — определенно-положительные квадратичные формы с кусочно-непрерывными коэффициентами; $A(t)$, $B(t)$, $C(t)$, $t_0 \leqq t \leqq \vartheta$ — кусочно-непрерывные матрицы-функции, t_0 и ϑ — фиксированные моменты

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

времени; $|x|$ — евклидова норма вектора x. Реализации управления $u[\cdot]=\{u[t], t_0<t\leqq\vartheta\}$ и помехи $v[\cdot]=\{v[t], t_0<t\leqq\vartheta\}$ измеримы и ограничены (каждая по своему).

Оптимальное управление $u^0[t]$ формировалось в [1] по принципу обратной связи на основе точно измеренной в любой нужный момент t_i текущей позиции $\{t_i, x[t_i]\}$. В этой статье предполагается, что сведения о фазовом векторе $x[t]$ поступают лишь по части $p\leqq n$ координат и притом еще с искажением. Именно, текущая информация о состоянии x-объекта (2.2) доставляется p-мерной информационной переменной

$$q[t]=K(t)x[t]+\Delta q[t], \tag{2.3}$$

где $K(t)$, $t_0\leqq t\leqq\vartheta$ — заданная кусочно-непрерывная $(p\times n)$-матрица-функция; $\Delta q[t]$ — информационная помеха. Примем еще, что начальное фазовое состояние $x_0=x[t_0]$ также сообщается с искажением Δx_0 в виде ложного значения $x_0^*=x_0+\Delta x_0$.

Предполагая возможным запоминание значений $q[t]$ (2.3) и вырабатываемого нами управления $u[t]$, сосредоточим всю полученную к моменту t информацию в информационном элементе $Y[t]$, состоящем из трех компонент

$$Y[t]=\{x_0^*,\ q[t_0[\cdot]t],\ \tilde{y}[t_0[\cdot]t]\}, \tag{2.4}$$

где $q[t_0[\cdot]t]=\{q[v],\ t_0\leqq v\leqq t\}$ — информационная история, которую полагаем кусочно-непрерывной функцией;

$$\tilde{y}'[t_0[\cdot]t]=\{y'[t_0[\cdot]t],\ \tilde{y}_{n+1}[t_0[\cdot]t]\}$$

— известная $n+1$-мерная управляемая вектор-функция, эволюционирующая в соответствии с дифференциальными уравнениями

$$\dot{y}[t]=X[\vartheta,t]B(t)u[t], \qquad \dot{\tilde{y}}_{n+1}[t]=u'[t]\Phi(t)u[t], \tag{2.5}$$

$$y[t_0]=\{0,\dots,0\}, \quad \tilde{y}_{n+1}[t_0]=0. \tag{2.6}$$

Здесь $X[\vartheta,t]$, $t_0\leqq t\leqq\vartheta$ — решение матричного дифференциального уравнения $dX[\vartheta,t]/dt=-X[\vartheta,t]A(t)$ с краевым условием $X[\vartheta,\vartheta]=E$, где E — единичная матрица. Информационному состоянию $\{t, Y[t]\}$ предстоит в рассматриваемой постановке сыграть ту же роль, какую в [1] играла текущая позиция $\{t, x[t]\}$. Поэтому последующие определения повторяют определения из [1] с заменой $\{t, x[t]\}$ на $\{t, Y[t]\}$.

Именно, стратегией $u(\cdot)$ называется произвольная функция

$$u(\cdot)=\{u(t,Y,\varepsilon),\quad t_0\leqq t\leqq\vartheta,\quad \varepsilon>0\}, \tag{2.7}$$

определенная для всех возможных значений информационного элемента Y.

Законом управления U для отрезка $[t_*, \vartheta] \subset [t_0, \vartheta]$ назовем совокупность трех компонент

$$U = \{u(\cdot),\ \varepsilon,\ \Delta\{t_i\}\}, \tag{2.8}$$

где $\Delta\{t_i\}$, $t_1 = t_*$, $t_i < t_{i+1}$, $i = 1, \ldots, l$, $t_{l+1} = \vartheta$ есть разбиение отрезка $[t_*, \vartheta]$. Движением $x[t_*[\cdot]\vartheta] = \{x[t], t_* \leqq t \leqq \vartheta\}$ данного x-объекта (2.2), порожденным законом управления U (2.8) и какой-то неизвестной помехой $v(t_*[\cdot]\vartheta]$, называется абсолютно-непрерывное решение пошагового дифференциального уравнения

$$\begin{gathered} \dot{x}[t] = A(t)x[t] + B(t)u(t_i, Y[t_i], \varepsilon) + C(t)v[t], \\ t_i \leqq t < t_{i+1}, \qquad i = 1, \ldots, l \end{gathered} \tag{2.9}$$

при неизвестном нам исходном состоянии

$$x[t_*] = x_* = x[t_0[t_*]t_*],$$

где

$$x[t_0[\cdot]t_*] = \{x[\nu],\ t_0 \leqq \nu \leqq t_*\}$$

есть сложившаяся так или иначе к моменту $t = t_*$ история движения x-объекта, стартовавшего из некоторого неизвестного нам истинного начального состояния $\{t_0, x_0\}$.

Параллельно с действительным движением $x[t_*[\cdot]\vartheta]$ закон U (2.8) формирует в системе (2.5), (2.6) воображаемое движение управляемой компоненты $\tilde{y}[t]$ как решение пошаговых дифференциальных уравнений

$$\dot{y}[t] = X[\vartheta, t]B(t)u(t_i, Y[t_i], \varepsilon), \tag{2.10}$$

$$\dot{\tilde{y}}_{n+1}[t] = u'(t_i, Y[t_i], \varepsilon)\Phi(t)u(t_i, Y[t_i], \varepsilon) \tag{2.11}$$

при известном исходном условии $\tilde{y}[t_*] = \tilde{y}_* = \tilde{y}[t_0[t_*]t_*]$, определяемом по информационному элементу $Y[t_*]$.

Таким образом, наряду с реальным, но неизвестным нам от начала до конца движением $x[t_0[\cdot]\vartheta]$ развертывается воображаемое движение $Y[t_0[\cdot]\vartheta]$ информационной Y-системы, состояние которой характеризуется известным нам в любой нужный момент $t = t_i$ информационным элементом $Y[t_i]$. При этом будем считать, что информационная история $q[t_*[\cdot]t_i]$ в составе $Y[t_i]$ формируется некоторым вторым игроком (независимо от наших целей). В этой картине берем на себя роль первого игрока, формирующего управление

$$u[t] = u[t_i] = u(t_i,\ Y[t_i], \varepsilon),\ t_i < t \leqq t_{i+1}, \qquad i = 1, \ldots, l$$

с целью гарантировать возможно меньшее значение следующего показателя качества для информационной Y-системы.

$$\gamma_*(Y[\vartheta]) = \sup_{\{x_0,\, v[\cdot]\}} [\gamma_x(x[\cdot], u[\cdot], v[\cdot]) - \\ - \int_{t_0}^{\vartheta} (\Delta q[t])' Q(t) \Delta q[t] dt - (\Delta x_0)' P \Delta x_0]. \tag{2.12}$$

Здесь $(\Delta q)'Q(t)\Delta q$ и $(\Delta x_0)'P\Delta x_0$ — определенно-положительные квадратичные формы, и при каких-то x_0, $v[\cdot]$, а также при известном $Y[\vartheta]$ значения $x[t]$, $t_0 \leqq t \leqq \vartheta$ определяются формулой Коши

$$x[t] = X[t, t_0]x_0 + X[t, \vartheta]y[t] + \int_{t_0}^{t} X[t, \nu]C(\nu)v[\nu]d\nu. \tag{2.13}$$

Дальнейшая формализация задачи на минимум гарантированного результата по показателю γ_* (2.12) для информационной Y-системы с полными сведениями о ее состояниях $Y[t]$, $t_0 \leqq t \leqq \vartheta$ следует общему плану, разработанному для абстрактных Y-систем в [2], и в конкретных случаях — реализованному в [1–4]. В этой задаче выбором продолжений $q(t_*[\cdot]\vartheta]$ известной к моменту t_* информационной истории $q[t_0[\cdot]t_*]$ распоряжается условный второй игрок (возможно с неблагоприятными для нас в смысле показателя γ_* (2.12) намерениями). Поэтому и в соответствии с [2] оптимальным гарантированным результатом для информационного состояния $\{t_*, Y[t_*]\}$ называется число

$$\rho^0(t_*, Y[t_*]) = \min_{u(\cdot)} \left[\overline{\lim_{\varepsilon \to 0}} \lim_{\delta \to 0} \sup_{U_\delta} \sup_{q(t_*[\cdot]\vartheta]} \gamma_*(Y[\vartheta])] \right], \tag{2.14}$$

где точные верхние грани вычисляются по всем возможным продолжениям $q(t_*[\cdot]\vartheta]$ информационной истории $q[t_0[\cdot]t_*]$, являющейся компонентой известного информационного элемента $Y[t_*]$, и по всем законам U_δ, отвечающим выбранной стратегии $u(\cdot)$, зафиксированному ε и разбиениям $\Delta\{t_i\}$, для которых $t_{i+1} - t_i \leqq \delta$.

Стратегия $u^0(\cdot)$, доставляющая в (2.14) минимум, называется оптимальной. Как и в [1–4], доказывается, что такая стратегия существует. Какую же гарантию доставит нам эта стратегия $u^0(\cdot)$ при управлении на ее основе реальным x-объектом, движение которого развертывается в соответствии с уравнениями (2.9) при $u(\cdot) = u^0(\cdot)$? Из вида показателей γ_x (2.1), γ_* (2.12) и определения $\rho^0(t_*, Y[t_*])$ (2.14) следует, что величина $\rho^0(t_*, Y[t_*])$ (2.14) является оптимальным гарантированным результатом для реального x-объекта по показателю

$$\gamma(x_0^*, q[t_0[\cdot]t_*], \quad u(t_0[\cdot]\vartheta]; \quad x_0, v(t_0[\cdot]\vartheta], \quad q(t_*[\cdot]\vartheta]) = \\ = |x[\vartheta]| + \int_{t_0}^{\vartheta} (u'[t]\Phi(t)u[t] - v'[t]\Psi(t)v[t])dt - \int_{t_0}^{\vartheta} (q[t] - \\ - K(t)x[t])'Q(t)\,(q[t] - K(t)x[t])dt - (x_0^* - x_0)'P(x_0^* - x_0) \tag{2.15}$$

относительно комбинированной помехи $\{x_0, v(t_0[\cdot]\vartheta], q(t_*[\cdot]\vartheta]\}$, выбираемой вторым игроком независимо от наших целей. Строго говоря, это означает следующее

Теорема 2.1. Для любого сколь угодно малого числа $\zeta > 0$ можно указать число $\varepsilon(\zeta)$ и величину $\delta(\varepsilon, \zeta) > 0$ так, что закон управления U_δ при $t_* \in [t_0, \vartheta]$, $u(\cdot) = u^0(\cdot)$, $\varepsilon \leqq \varepsilon(\zeta)$, $\delta \leqq \delta(\varepsilon, \zeta)$ сформирует в системе (2.9) такое движение $x[t_*[\cdot]\vartheta]$, что при любом начальном состоянии x_0, при любой возможной помехе $v(t_0[\cdot]\vartheta]$ и при любом возможном продолжении $q(t_*[\cdot]\vartheta]$ информационной истории $q[t_0[\cdot]t_*]$ будет иметь место неравенство

$$\gamma(x_0^*, q[t_0[\cdot]t_*]\ \ u^0(t_0[\cdot]\vartheta];\quad x_0,\ v(t_0[\cdot]\vartheta],\ q(t_*[\cdot]\vartheta]) \leqq \rho + \zeta, \tag{2.16}$$

каково бы ни было исходное состояние $\{t_*, Y[t_*] = \{x_0^*, q[t_0[\cdot]t_*], \tilde{y}[t_0[\cdot]t_*]\}$; и значение $\rho = \rho^0(t_*, Y[t_*])$ является наименьшим из чисел ρ, которые удовлетворяют такому условию.

Заметим, что весовые матрицы $\Phi(t)$, $\Psi(ttt)$, $K(t)$, $Q(t)$ и P в показателе γ (2.15) регулируют по величине реализации оптимального управления $u^0(t_*[\cdot]\vartheta]$ самых неблагоприятных для нас помех $\{x_0, v(t_0[\cdot]\vartheta], q(t_*[\cdot]\vartheta]\}$, не позволяя им быть сколь угодно большими. Отрицательные слагаемые в (2.15) оценивают компенсацию, предоставляемую нам вторым игроком за формируемую им динамическую помеху и за искажение информации о текущем и начальном состояниях $\{t,\ x[t]\}$ реального x-объекта. Слагаемое $|x[\vartheta]|$ в (2.15) можно трактовать как штраф, накладываемый на первого игрока, за отклонение x-объекта в момент $t = \vartheta$ от начала координат, а второе положительное слагаемое — как стоимость затраченных на процесс управления ресурсов.

3. Основной результат

Итак, требуется построить оптимальный по показателю γ_* (2.12) закон управления $U_\delta^0 = \{u^0[t_i] = u^0(t_i, Y[t_i],\ \ \varepsilon),\ \ i = 1, \ldots, l\}$ для введенной в § 2 информационной Y-системы. Такой закон управления U_δ^0 в согласии с теоремой 2.1 при достаточно малых ε и δ будет гарантировать для x-объекта результат $\rho^0(t_*,\ Y[t_*]) + \zeta$, наилучший в смысле показателя γ (2.15).

Если воспользоваться методом экстремального сдвига системы на сопутствующую точку, описанным для абстрактной Y-системы в § 7 работы [2] и использованным для задачи (2.1), (2.2) в случае $Y[t] = x[t]$ в [1], то придем в рассматриваемой задаче к следующему алгоритму (см., например, формулу (7.14) в [1] на стр. 19)

$$u^0[t_i] = -\frac{1}{2}\Phi^{-1}(t_i)B'(t_i)X'[\vartheta, t_i]m^0[t_i], \tag{3.1}$$

где

$$m^0[t_i]=m^0(t_i,\, Y[t_i],\, \varepsilon)=\arg\{\max_{|m|\leqq 1}[-\eta[t_i]\,(1+|m|^2)^{1/2}+ \\ +m'h(t_i,\, Y[t_i])+m'(F(t_i)-\lambda(t_i)E)m]\}. \tag{3.2}$$

Здесь $h(t_i, Y[t_i])$ — n-мерный вектор, вычисляемый по ходу дела; $F(t)$ — $(n\times n)$-матрица, которая может быть сосчитана априори по исходным данным задачи сразу для всех $t\in[t_*, \vartheta]$;

$$\lambda(t_i)=\max_{t_i\leqq t\leqq\vartheta}\ \max_{|a|=1} a'F(t)a, \tag{3.3}$$

$$\eta(t_i)=[\varepsilon+\varepsilon(t_i-t_0)]^{1/2}>0. \tag{3.44}$$

Учитывая (3.3) и (3.4), видим, что задача (3.2) имеет единственное решение $m^0[t_i]$, как задача на максимум от строго вогнутой функции.

Оптимальный гарантированный результат вычисляется по формуле

$$\rho^0(t_*,\, Y[t_*])=\max_{|m|\leqq 1}[m'h(t_*,\, Y[t_*])+ \\ +m'(F(t_*)-\lambda(t_*)E)m+\chi(t_*,\, Y[t_*])], \tag{3.5}$$

где $\chi(t_*, Y[t_*])$ — некоторое число.

Структура соотношений (3.1)–(3.5) характерна для довольно широкого круга задач. Различными в этих соотношениях в зависимости от задачи будут лишь компоненты информационного элемента $Y[t]$ и конкретные выражения для величин $F(t)$, $h(t, Y[t])$, $\chi(t, Y[t])$. Так, например, для задачи (2.1), (2.2) при $Y[t]=x[t]$, решенной в [1], имеем

$$h(t,\, x[t])\equiv X[\vartheta, t]x[t], \tag{3.6}$$

$$F(t)=-\frac{1}{4}\int_t^{\vartheta}(X[\vartheta,\tau]\,[B(\tau)\Phi^{-1}(\tau)B'(\tau)- \\ -C(\tau)\Psi^{-1}(\tau)C'(\tau)]X'[\vartheta,\tau])d\tau. \tag{3.7}$$

Для того, чтобы в рассматриваемой задаче получить выражения величин $h(t, Y[t])$ и $F(t)$, участвующих в алгоритме (3.1), (3.2) (см. по этому поводу ниже § 5), достаточно иметь конкретное выражение для линейно-квадратичной относительно m функции $H(t_*, Y[t_*], m)$, стоящей под знаком максимума в формуле (3.5). Но в соответствии с методом стохастического программного синтеза [1–4] оптимальный гарантированный результат $\rho^0(t_*, Y[t_*])$ (3.5) совпадает с величиной должным образом сконструированного стохастического программного максимина. Поэтому перейдем к этой величине.

4. Стохастический программный максимин

В дальнейшем вместо переменной $q[t]$ (2.3) удобно рассматривать следующую наблюдаемую переменную

$$r[t]=q[t]-K(t)\int_{t_0}^{t}X[t,v]B(v)u[v]dv= \\ =q[t]-K(t)X[t,\vartheta]y[t], \tag{4.1}$$

история которой $r[t_0[\cdot]t]=\{r[v],t_0\leqq v\leqq t\}$ в текущий момент времени t по условиям задачи нам известна.

Информационной Y-системе при замене в (2.4) компоненты $q[t_0[\cdot]t]$ на $r[t_0[\cdot]t]$ поставим в соответствие стохастическую Z-модель. Она строится так. Зафиксируем некоторый момент $\tau_*\in[t_0,\vartheta)$ и какую-то исходную историю $r[t_0[\cdot]\tau_*]$ наблюдаемой переменной $r[t]$ (4.1). Назначим разбиение $\Delta\{\tau_j\}$, $j=1,\ldots,k$ для отрезка $[\tau_*,\vartheta]$, $\tau_1=\tau_*$, $\tau_j<\tau_{j+1}$, $\tau_k=\vartheta$. В основу стохастической программной конструкции положим вероятностное пространство $\{\Omega,\mathscr{B},\mathscr{P}\}$, где Ω есть k-мерный единичный куб

$$\Omega=\{\omega=\{\xi_1,\ldots,\xi_k\},\quad 0\leqq\xi_j<1,\quad j=1,\ldots,k\};$$

$\mathscr{B}$ — борелевская σ-алгебра, $\mathscr{P}$ — лебегова мера [5]. Назовем стохастическими программами неупреждающие функции ([5], стр. 100)

$$u_*(\tau,\omega)=u_*[\tau,\xi_1,\ldots,\xi_j],\quad \tau_j<\tau\leqq\tau_{j+1},\quad j=1,\ldots,k-1, \tag{4.2}$$

$$r_*(\tau,\omega)=r_*[\tau,\xi_1,\ldots,\xi_j],\quad \tau_j<\tau\leqq\tau_{j+1},\quad j=1,\ldots,k-1. \tag{4.3}$$

Состояние Z-модели в текущий момент $\tau\geqq\tau_*$ определим фазовым элементом

$$Z(\tau,\omega)=\{x_0^*,r[t_0[\cdot]\tau;\omega],\tilde{z}(\tau,\omega)\}, \tag{4.4}$$

где компонента $r[t_0[\cdot]\tau;\omega]=\{r(v,\omega),t_0\leqq v\leqq\tau\leqq\vartheta,\omega\in\Omega\}$ при $\tau=\tau_*$ совпадает с детерминированной историей $r[t_0[\cdot]\tau_*]$, а при $\tau>\tau_*$ ее случайные реализации $r(\tau,\omega)$ определяются программой $r_*(\tau,\omega)$ (4.3). Векторная $n+1$-мерная компонента

$$\tilde{z}'(\tau,\omega)=\{z'(\tau,\omega),\quad \tilde{z}_{n+1}(\tau,\omega)\}=\tilde{z}'[\tau,\xi_1,\ldots,\xi_j], \\ \tau_j<\tau\leqq\tau_{j+1},\quad j=1,\ldots,k-1$$

в составе $Z(\tau,\omega)$ (4.4) определяется как решение уравнений

$$\dot{z}(\tau,\omega)=X[\vartheta,\tau]B(\tau)u_*(\tau,\omega), \\ \dot{\tilde{z}}_{n+1}(\tau,\omega)=u_*'(\tau,\omega)\Phi(\tau)u_*(\tau,\omega) \tag{4.5}$$

при некотором исходном состоянии $\tilde{z}(\tau_*, \omega) = \tilde{z}_*$. Фазовый элемент $Z(\tau, \omega)$ (4.4) имитирует в стохастическом варианте информационный элемент $Y[\tau]$, а вводимые ниже n-мерная случайная величина $w(\cdot) = \{w(\omega), \omega \in \Omega\}$ и s-мерная случайная функция $v(\cdot) = \{v(\tau, \omega),\ t_0 < \tau \leqq \vartheta,\ \omega \in \Omega\}$ имитируют помехи x_0 и $v(t_0[\cdot]\vartheta]$ из Y-системы.

Назовем программным экстремумом величину

$$e(\tau_*, Z[\tau_*], \Delta) = \sup_{\|l(\cdot)\| \leqq 1} \varkappa(\tau_*, Z[\tau_*], \Delta, l(\cdot)), \tag{4.6}$$

где

$$\Delta = \Delta\{\tau_j\}, \quad Z[\tau_*] = Z(\tau_*, \omega) =$$
$$= \{x_0^*, r[t_0[\cdot]\tau_*], \tilde{z}_*\}; \quad l(\cdot) = \{l(\omega), \quad \omega \in \Omega\}$$

— n-мерная случайная величина, $\|l(\cdot)\| = (M\{|l(\omega)|^2\})^{1/2}$, M — математическое ожидание, и величина $\varkappa = \varkappa(\tau_*, Z[\tau_*], \Delta, l(\cdot))$ определяется соотношением

$$\begin{gathered}\varkappa = \sup_{r_*(\cdot)} \inf_{u_*(\cdot)} \sup_{v(\cdot)} \sup_{w(\cdot)} \sigma(\tau_*, Z[\tau_*], \Delta, l(\cdot); \\ r_*(\cdot),\ u_*(\cdot),\ v(\cdot),\ w(\cdot)).\end{gathered} \tag{4.7}$$

Здесь функция σ строится по виду показателя γ (2.15) с учетом введенных стохастических аналогов следующим образом

$$\begin{aligned}\sigma = M\{&l'(\omega)\,[X[\vartheta, t_0]w(\omega) + z_* + \int_{\tau_*}^{\vartheta} X[\vartheta, \tau]B(\tau)u_*(\tau, \omega)d\tau + \\ &+ \int_{t_0}^{\vartheta} X[\vartheta, \tau]C(\tau)v(\tau, \omega)d\tau] + \tilde{z}_{n+1*} + \int_{\tau_*}^{\vartheta} u'_*(\tau, \omega)\Phi(\tau)u_*(\tau, \omega)d\tau - \\ &- \int_{t_0}^{\vartheta} [v'(\tau, \omega)\Psi(\tau)v(\tau, \omega) + [r(\tau, \omega) - K(\tau)\,(X[\tau, t_0]w(\omega) + \\ &+ \int_{t_0}^{\tau} X[\tau, v]C(v)v(v, \omega)dv)]'Q(\tau)\,[r(\tau, \omega) - K(\tau)\,(X[\tau, t_0]w(\omega) + \\ &+ \int_{t_0}^{\tau} X[\tau, v]C(v)v(v, \omega)dv)]]d\tau - [x_0^* - w(\omega)]'P[x_0^* - w(\omega)]\}.\end{aligned} \tag{4.8}$$

Связь между оптимальным гарантированным результатом $\rho^0(t_*, Y[t_*])$ (2.14) и программным экстремумом (4.6) устанавливается следующей теоремой.

Теорема 4.1. Если в (4.6) положить $\tau_* = t_*$, $\tilde{z}_* = \tilde{y}_*$, $r[v] = q[v] - K(v)X[v, \vartheta]y[v]$ при $t_0 \leqq v \leqq \tau_*$, то

$$\rho^0(t_*, Y[t_*]) = \sup_{\Delta} e(\tau_*, Z[\tau_*], \Delta). \tag{4.9}$$

Таким образом, для отыскания вектора $h(t_*, Y[t_*])$ и матрицы $F(t_*)$ необходимо выражение стохастического программного максимина (4.9) привести к виду (3.5).

5. Вычисление программного экстремума

При фиксированных τ_*, $Z[\tau_*]$, Δ, $l(\cdot)$ вычислим вариации величины σ (4.8) по $w(\cdot)$, $v(\cdot)$, $u_*(\cdot)$ и $r_*(\cdot)$. Приравнивая нулю частные вариации по $w(\cdot)$ и $v(\cdot)$ получим соответственно следующие два интегральных уравнения

$$X'[\vartheta, t_0]l(\omega) + 2\int_{t_0}^{\vartheta} X'[\tau, t_0]K'(\tau)Q(\tau)\,[r(\tau, \omega) -$$

$$- K(\tau)\,(X[\tau, t_0]w(\omega) + \int_{t_0}^{\tau} X[\tau, v]C(v)v(v, \omega)dv)]d\tau +$$

$$+ 2P(x_0^* - w(\omega)) = 0, \qquad \omega \in \Omega, \tag{5.1}$$

$$C'(\tau)\,(X'[\vartheta, \tau]l(\omega) + 2\int_{\tau}^{\vartheta} X'[\eta, \tau]K'(\eta)Q(\eta)\,[r(\eta, \omega) -$$

$$- K(\eta)\,(X[\eta, t_0]w(\omega) + \int_{t_0}^{\eta} X[\eta, v]C(v)v(v, \omega)dv)]d\eta) -$$

$$- 2\Psi(\tau)v(\tau, \omega) = 0, \qquad t_0 \leqq \tau < \vartheta, \quad \omega \in \Omega. \tag{5.2}$$

Приравнивая нулю частные вариации величины σ (4.8) по неупреждающим программам $u_*(\cdot)$ (4.2), $r_*(\cdot)$ (4.3), получим следующие соотношения, справедливые при $\tau_j < \tau \leqq \tau_{j+1}$, $j = 1, \ldots, k-1$

$$u[\tau, \xi_1, \ldots, \xi_j] = -\frac{1}{2}\Phi^{-1}(\tau)B'(\tau)X'[\vartheta, \tau]M\{l(\omega)\,|\,\xi_1, \ldots, \xi_j\}, \tag{5.3}$$

$$r[\tau, \xi_1, \ldots, \xi_j] = K(\tau)\,(X[\tau, t_0]M\{w(\omega)\,|\,\xi_1, \ldots, \xi_j\} -$$

$$- \int_{t_0}^{\tau} X[\tau, v]C(v)M\{v(v, \omega)\,|\,\xi_1, \ldots, \xi_j\}dv). \tag{5.4}$$

Здесь символ $M\{\ldots|\ldots\}$ означает условное математическое ожидание. Уравнения (5.1)–(5.4) выражают необходимые условия экстремальности для величины σ (4.8). Эти уравнения имеют единственное решение относительно $w(\cdot)$, $v(\cdot)$, $u_*(\cdot)$, $r_*(\cdot)$. Предположим, что экстремальные элементы $w^0(\cdot)$, $v^0(\cdot)$, $u_*^0(\cdot)$, $r_*^0(\cdot)$ — решение уравнений (5.1)–(5.4) — так или иначе найдены. Подставляя их в выражение для σ (4.8), получим некоторый линейноквадратичный функционал от $l(\cdot)$. Кроме того, как и в [1], получается, что верхняя грань в задаче (4.6) достигается в классе случайных величин $l(\cdot)$ вида

$$l(\omega)=l[\xi]=m+a[\xi], \quad \xi\in[0,1), \quad M\{a[\xi]\}=0. \tag{5.5}$$

Переходя затем к вычислению стохастического программного максимина в (4.9), получим еще, что точная верхняя грань достигается, как и в [1, 2], на разбиении $\Delta:\tau_1=\tau_*, \tau_2=\tilde{\tau}, \tau_3=\vartheta, \tau_*\leqq\tilde{\tau}\leqq\vartheta$, где момент $\tilde{\tau}$ определяется из условия

$$\max_{|a|=1} a'F(\tilde{\tau})a=\max_{\tau_*\leqq\tau\leqq\vartheta}\max_{|a|=1} a'F(\tau)a. \tag{5.6}$$

Здесь $F(\tau)$ есть та самая матрица, которая фигурирует в (3.2), (3.3), (3.5). На таком пути в итоге придем к нужному равенству

$$\sup_{\Delta} e(\tau_*, Z[\tau_*], \Delta)=\max_{|m|\leqq 1} H(\tau_*, Z[\tau_*], m), \tag{5.7}$$

где в соответствии с теоремой 4.1 функция H и будет как раз той линейно-квадратичной относительно m функцией, которая стоит под знаком максимума в формуле (3.5).

Таким образом, центральной вычислительной проблемой при отыскании вектора $h(t, Y[t])$ и матрицы $F(t)$ оказывается задача определения экстремальных элементов из уравнений (5.1)–(5.4). Анализ этих уравнений выходит за рамки данной статьи. Уравнения, аналогичные уравнениям (5.1)–(5.4), исследованы в [4] в частном случае, когда $C(t)=c(t)$ есть n-мерный вектор, $\Psi(t)=\psi(t)$-скаляр и динамическая помеха есть скалярная функция, представленная отрезком ряда Фурье

$$v[t]=\sum_{s=1}^{N}\alpha_s g_s(t)=\alpha' g(t), \quad t_0<t\leqq\vartheta, \tag{5.8}$$

где $\alpha'=\{\alpha_1, \ldots, \alpha_N\}$ — произвольный (неизвестный первому игроку) числовой вектор. В этом частном случае выражение для матрицы $F(t)$ имеет вид

$$\begin{gathered} F(t)=-\frac{1}{4}\Bigg[\int_t^{\vartheta} X[\vartheta,\tau]B(\tau)\Phi^{-1}(\tau)B'(\tau)X'[\vartheta,\tau]d\tau- \\ -G[\vartheta,t_0]W^{-1}(t)G'[\vartheta,t_0]\Bigg]. \end{gathered} \tag{5.9}$$

Здесь

$$G[\vartheta,t_0]=(V'[\vartheta,t_0],\ X[\vartheta,t_0]), \tag{5.10}$$

$$V[\vartheta,t_0]=\int_{t_0}^{\vartheta} g(t)c'(t)X'[\vartheta,t]dt \tag{5.11}$$

и матрица $W(t)$ имеет следующую блочную структуру

$$W(t)=\begin{pmatrix} W_{11}(t) & W_{12}(t) \\ W'_{12}(t) & W_{22}(t) \end{pmatrix}, \tag{5.12}$$

где

$$\begin{aligned} W_{11}(t)= \int\limits_{t_0}^{\vartheta} \psi(t)g(t)g'(t)dt+ \\ + \int\limits_{t_0}^{t} V[v,t_0]K'(v)Q(v)K(v)V'[v,t_0]dv, \end{aligned} \tag{5.13}$$

$$W_{12}(t)= \int\limits_{t_0}^{t} V[v,t_0]K'(v)Q(v)K(v)X[v,t_0]dv, \tag{5.14}$$

$$W_{22}(t)= \int\limits_{t_0}^{t} X'[v,t_0]K'(v)Q(v)K(v)X[v,t_0]dv+P. \tag{5.15}$$

Выражение для вектора $h(t,\ Y[t])$ имеет вид

$$\begin{aligned} h(t,\ Y[t])=y[t]+G[\vartheta,t_0]W^{-1}(t)\Bigg[\int\limits_{t_0}^{t} G'[v,t_0]\times \\ \times K'(v)Q(v)\,(q[v]-K(v)X[v,\vartheta]y[v])dv+P_0x_0^*\Bigg], \end{aligned} \tag{5.16}$$

где $P_0=(0,\ P)'$ и 0 есть нулевая $N\times n$-матрица.

Заметим в заключение, что задача отыскания экстремальных элементов носит вспомогательный характер. Однако, если на реальный x-объект (2.2), управляемый оптимальным законом U_δ^0 (3.1) с малыми ε и δ, и в информационный элемент $Y[t]$, $t\geqq t_*$ будут подаваться те или иные реализации случайных экстремальных элементов $w^0(\cdot)$, $v^0(\cdot)$, $r_*^0(\cdot)$, то значения показателя γ (2.15) будут получаться каждый раз сколь угодно близкими к величине оптимального гарантированного результата $\rho^0(t_*,\ Y[t_*])$. Это замечание может оказаться полезным в вычислительных экспериментах при испытании на устойчивость работы оптимального закона управления U_δ^0.

6. Пример

Пусть система (2.2) и показатель γ (2.15) имеют вид

$$\dot{x}_1=x_2, \qquad \dot{x}_2=u+v, \qquad 0=t_0\leqq t\leqq\vartheta=3, \tag{6.1}$$

$$\begin{aligned} \gamma=|x[\vartheta]|+\int\limits_{t_0}^{\vartheta}\varphi(t)u^2[t]dt-\int\limits_{t_0}^{\vartheta}\psi(t)v^2[t]dt- \\ -\int\limits_{t_0}^{\vartheta}p(q[t]-x_1[t])^2dt-p[(x_{01}-x_{01}^*)+(x_{02}-x_{02}^*)^2]. \end{aligned} \tag{6.2}$$

Задачу на минимум гарантированного результата по показателю γ (6.2) для системы (6.1) можно интерпретировать как задачу прямолинейного перемещения управляемого экипажа под действием тяги u и силы ветра v из неизвестного начального состояния $x_0=\{x_{01}, x_{02}\}$ в начало координат $\{0,0\}$ фазовой плоскости (x_1, x_2). С целью извлечения некоторой выгоды из помехи v экипаж снабжен ветровым генератором, вырабатывающим энергию, стоимость которой оценивается первым в (6.2) отрицательным слагаемым. Остальные отрицательные слагаемые в (6.2) оценивают компенсацию за искажение прибором информации о текущем состоянии $\{t, x[t]\}$ экипажа и за неправильно сообщенное начальное состояние $x_0^*=\{x_{01}^*, x_{02}^*\}$. Положительные слагаемые в (6.2) определяют штраф за недоезд экипажа до начала координат и стоимость энергии, идущей на выработку тяги u.

Оптимальный закон U_δ^0 (3.1), (3.2) управления экипажем был испытан на ЭВМ в случае, когда помеха имеет вид $v[t]=\alpha_1+\alpha_2 t$, где α_1 и α_2 — произвольные числа. Были выбраны следующие параметры

$$\varphi(t)\equiv 0.25, \qquad \psi(t)\equiv 1, \qquad p=1.0,\ 1.5,\ 2.5 \tag{6.3}$$

и принято следующее ложное начальное состояние

$$x_0^*=\{x_{01}^*=-1, \quad x_{02}^*=0\}. \tag{6.4}$$

На рис. 1 даны изображения на фазовой плоскости (x_1, x_2) реализаций движения экипажа при условии, что начальные состояния x_0, числа α_1 и α_2 и показания прибора $q[t]$, $0\leqq t\leqq 3$ являлись для нас самыми неблагоприятными в смысле показателя γ (6.2), т. е. определялись некоторыми реализациями экстремальных элементов $w^0(\cdot)$, $v^0(\cdot)$ и $r_*^0(\cdot)$. Сплошные линии на рис. 1 соответствуют (слева направо) различным значениям p из (6.3). При этом получились следующие значения показателя $\gamma=1.30$, 0.85, 0.55, каждое из которых с выбранной точностью совпадает с величиной соответствующего оптимального гарантированного результата $\rho^0(t_0,\ Y[t_0])$, подсчитанного априори по формуле (3.5). Точечная линия на рис. 1 отвечает значению $p=2.5$ и при прочих неизменных условиях — начальному состоянию $x_0=x_0^*$ (6.4). Здесь в согласии с теорией получилось, что $\gamma=0.23<\rho^0(t_0,\ Y[t_0])=0.55$. Кроме того, на рис. 1 для $p=2.5$ изображены реализация $u^0[t]$, $0\leqq t\leqq 3$ оптимального закона управления и осуществившаяся здесь реализация $v^0[t]$, $0\leqq t\leqq 3$ экстремального элемента $v^0(\cdot)$.

Поскольку экстремальные элементы $w^0(\cdot)$, $v^0(\cdot)$, $r_*^0(\cdot)$ носят случайный характер, то в той же ситуации, что и для рис. 1, возможна и другая картина. Она изображена на рис. 2.

На рис. 3 изображено изменение во времени координаты $x_1[t]$, $0\leqq t\leqq 3$ (сплошная линия) при $\varphi(t)\equiv 1$, $\psi(t)\equiv 10$, $p=2.5$ и при экстремальных значениях x_0^0,

α_1^0, α_2^0 и $q^0[t]$, $0 \leqq t \leqq 3$. Точечной линией на рис. 3 изображены экстремальные показания прибора $q^0[t]$, которые на отрезке $[\tilde{\tau}, \vartheta]$ получились совпадающими с истинными значениями координаты $x_1[t]$. Ситуация, представленная на рис. 4, подчеркивает, что в реальном счете значение параметра точности ε в законе U_δ^0, $\delta = t_{i+1} - t_i$ при фиксированном δ не может быть взято сколь угодно малым из-за опасности возникновения в реализациях $u^0[t]$ (3.1) скользящего режима, который и представлен на рис. 4. Это случилось в данном примере при

$$\varphi(t) \equiv 1, \quad \psi(t) \equiv 10, \quad p = 2.5, \quad x_0 = x_0^* = \{-1, 0\}, \quad q[t] = x_1[t],$$
$$0 \leqq t \leqq 3, \quad \varepsilon = 0.001, \quad \delta = 0.01.$$

При этом за счет неоправданного увеличения второго слагаемого в γ (6.2) вышло, что $\gamma = 0.66 > \rho^0(t_0, Y[t_0]) = 0.63$, хотя в согласии с теорией должно в данном случае выполняться противоположное неравенство. Если при том же $\delta = 0.01$ взять $\varepsilon = 0.01$, то скользящий режим в реализации оптимального управления исчезает (см. рис. 5) и при этом, как и должно быть, получается

$$\gamma = 0.37 < \rho^0(t_0, Y[t_0]) = 0.63.$$

Литература

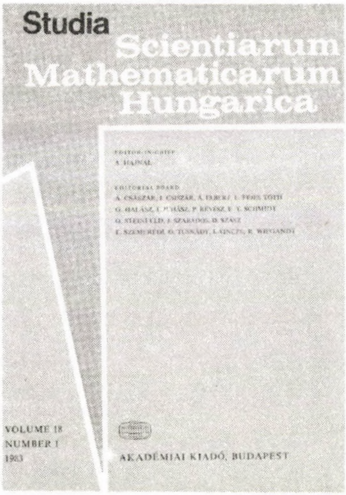
1. *Красовский Н. Н., Третьяков В. Е.* Одна задача оптимального управления на минимум гарантированного результата. Изв. АН СССР, Техническая кибернетика, № *2*, 1983, с. 6–23.
2. *Красовский Н. Н.* Экстремальное прицеливание и экстремальный сдвиг в игровом управлении. Проблемы управления и теории информации, т. **13***(5)*, 1984, с. 287–302.
3. *Красовский Н. Н.* Задача об управлении в условиях неполной информации. Прикл. матем. и мех., т. **48,** вып. *4*, 1984, с. 533–539.
4. *Красовский Н. Н., Тарасова С. И., Третьяков В. Е., Шишкин Г. И.* Задача управления при неполной информации. Препринт, УНЦ АН СССР, Свердловск, 1984, 63 с.
5. *Липцер Р. Ш., Ширяев А. Н.* Статистика случайных процессов. М., Наука, 1974, 696 с.

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda, Budapest

Studia Scientiarum Mathematicarum Hungarica

Editor in Chief:
A. Hajnal



The journal publishes original research papers on mathematics and its diverse fields of application in science, technology, economics, etc. Articles on theory of probability, on mathematical statistics, numerical and graphic methods as well as differential equation are also presented.

Founded 1966
Papers mainly in English, but also in French, German or Russian
Publication: semiannually — one volume of two double issues annually
Price per volume: $44.00; DM 99,–
Size: 17 × 25 cm
ISSN 0081–6906

Order form
to be returned to
KULTURA
Hungarian Foreign Trading Company
P.O. Box. 149, H-1389 Budapest, Hungary

□ Please enter my/our subscription for STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA for one year

□ Please enter my/our standing order for STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA starting with

Name: ______________________

Address: ______________________

Date and signature: ______________________

STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA

Contents of Volume 16. Numbers 1–2

1828

Akadémiai
Kiadó

Publishing House
of the Hungarian Academy of Sciences
Budapest

Invitation for papers
Manuscripts should be sent to
J. Merza
Managing Editor
H-1053 Budapest
Reáltanoda u. 13–15.
Hungary

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

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). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as diag $(d_1, d_2, \dots d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 129090 Москва И-90, ул. Щепкина, 8. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 208-60-19).

Объём 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 языке в трех экземплярах, 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 и 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 оттисков их статей.

Рукописи непринятых статей возвращаю гся авторам.

CONTENTS · СОДЕРЖАНИЕ

HU ISSN 0370-2529 Index: 26 660

VOL. / TOM 15 · NUMBER / НОМЕР 4

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

1986

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
— theory of adaptive systems
— theory of estimation and identification
— theory of controlling robot-technologic and flexible manufacturing systems
— information theory
— information-theoretic aspects of multiple access networks.

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;
— теория адаптивных систем;
— теория оценивания и идентификации;
— теория управления робототехническими и гибкими производственными системами;
— теория информации;
— теория информации в области сетей с множественным доступом.

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 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1986 Subscription Rate US Dollars 175.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

ON THE PROBLEM OF MATRIX PARAMETER IDENTIFICATION

A. M. Ustyužanin

(Sverdlovsk)

(Received June 27, 1985)

An information set for a linear discrete-time problem of estimating a constant matrix parameter under conditions of uncertainty has been described by using a support function. The conditions of precise identification are given when input disturbance belongs to a convex compact set. Precise identification means that the information set consists of a single matrix. Examples with additional assumptions on disturbance have been considered. An example is given for the case when the disturbance is stochastic.

Introduction

(*A*) *Problem formulation*

Let us consider the set $\{z_k\}_{k\in I}$, where z_k belongs to finite-dimensional Euclidean space $\mathbf{R}^n$, I is some set of positive integers. Let $\{z_k\}_{k=1}^N$ be $\{z_k\}_{k\in I}$, when $k\in\overline{1,N}$.

For a matrix $C\in\mathbf{R}^{m\times n}$, we consider the observation equation

$$y_k=Cx_k+\xi_k, \qquad k\in\overline{1,N}\,. \tag{1}$$

Vectors $x_k\in\mathbf{R}^n$, $y_k\in\mathbf{R}^m$ are set to be inputs and outputs respectively. Vectors $\xi_k\in\mathbf{R}^m$ are uncertain disturbances, information being available

$$\xi_k\in\Xi_k, \qquad k\in\overline{1,N} \tag{2}$$

where Ξ_k are convex compacts in $\mathbf{R}^m$ with non-empty interior.

The scalar product in $\mathbf{R}^{m\times n}$ is given by $(C_1,C_2)=\operatorname{tr}(C_1^TC_2)$, where $\operatorname{tr}(\cdot)$ is the trace of the matrix; the symbol ^{T} denotes transposition.

Following [1], we introduce

Definition. The information set $\mathscr{S}$ is the set of all matrices $C\in\mathbf{R}^{m\times n}$ which satisfy equation (1) and condition (2) with given $\{x_k\}_{k=1}^N$, $\{y_k\}_{k=1}^N$, $\{\Xi_k\}_{k=1}^N$.

The identification problem consists in describing the set $\mathscr{S}$.

Let C_* be the matrix, for which given inputs x_k and realizations $\bar{\xi}_k\in\Xi_k$ generate outputs y_k according to (1). Sets such as $\{x_k\}_{k=1}^N$ and $\{y_k\}_{k=1}^N$ will be considered. Hence, $C_*\in\mathscr{S}$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

(*B*) *Notations*

$$\mathscr{L}\{z_k\}_{k\in I}\triangleq\{z:z=\sum_{k\in I}\alpha_k z_k,\quad \alpha_k\in\mathbf{R}^n\},$$

$$\mathscr{L}^+\{z_k\}_{k\in I}\triangleq\{z:z=\sum_{k\in I}\alpha_k z_k,\quad \alpha_k\geqq 0\},$$

$\partial\Xi$ is the boundary of set Ξ.

For $\varepsilon>0$ and realization of disturbance $\bar{\xi}_k$, define the following sets

$$U_k(\varepsilon)\triangleq\{\xi\in\mathbf{R}^m:\|\xi-\bar{\xi}_k\|\leqq\varepsilon,\quad \xi\in\partial\Xi_k\},$$

$$\Lambda_k(\varepsilon)\triangleq\{\lambda\in\mathbf{R}^m:\|\lambda\|=1,\quad \exists\xi\in U_k(\varepsilon),\quad \rho(-\lambda|\Xi_k)=-\lambda^T\xi\} \tag{3}$$

where $\rho(\cdot|\Xi_k)$ is a support function of the set Ξ_k [2], $\|\cdot\|$ is Euclidean norm. Let us note that for $\bar{\xi}_k\notin\partial\Xi_k$, sets (3) are empty for sufficiently small ε.

Precise identification

Theorem 1. If the set of inputs $\{x_k\}_{k=1}^N$ satisfies the condition

$$\mathscr{L}\{x_k\}_{k=1}^N=\mathbf{R}^n,$$

then the information set $\mathscr{S}$ is a convex compact with support function

$$\rho(L|\mathscr{S})=\inf_{\lambda_k\in\mathbf{R}^m}\left\{\sum_{k=1}^N\lambda_k^T y_k+\sum_{k=1}^N\rho(-\lambda_k|\Xi_k):\sum_{k=1}^N\lambda_k x_k^T=L\right\}. \tag{4}$$

Proof. Let $\mathscr{S}_k$ be a set of matrices C which satisfy the inclusion $y_k-Cx_k\in\Xi_k$. The convexity and the closure of $\mathscr{S}_k$ are verified directly. The support function of $\mathscr{S}_k$ is

$$\rho(L|\mathscr{S}_k)=\begin{cases}\lambda_k^T y_k+\rho(-\lambda_k|\Xi_k),\ \lambda_k x_k^T=L_k\\ +\infty,\qquad \text{otherwise}.\end{cases} \tag{5}$$

The set $\mathscr{S}$ is the intersection of $\mathscr{S}_k$, $k\in\overline{1,N}$, therefore the support function of $\mathscr{S}$ is the infimal convolution [2] of functions (5)

$$\rho(L|\mathscr{S})=\begin{cases}\inf\limits_{L_k}\left\{\sum_{k=1}^N\rho(L_k|\mathscr{S}_k):\sum_{k=1}^N L_k=L,\ \lambda_k x_k^T=L_k\right\}\\ +\infty,\ \forall\lambda_k\in\mathbf{R}^m,\ \sum_{k=1}^N\lambda_k x_k^T\neq L.\end{cases} \tag{6}$$

The assumption of the theorem holds iff any matrix $L\in\mathbf{R}^{m\times n}$ may be represented as $L=\sum_{k=1}^N\lambda_k x_k^T$. The necessity of this assumption may be proved by contradiction. In fact, if $\mathscr{L}\{x_k\}_{k=1}^N \neq \mathbf{R}^n$, then there exists a vector $x_0 \neq 0$, for which $(x_0, x_k)=0$, $k \in \overline{1, N}$. Denote $L^T \triangleq (x_0, 0, \dots, 0)$, then $L=\sum_{k=1}^N \lambda_k x_k^T$. Hence, $x_0=\sum_{k=1}^N \lambda_k^1 x_k$ and $\|x_0\|=$ $=\left(x_0, \sum_{k=1}^N \lambda_k^1 x_k\right)=0$, a contradiction.

Sufficiency of the assumption arises from the possibility to represent each j-th row of matrix L as $L_j=\sum_{k=1}^N \lambda_k^j x_k^T$.

Thus, the equality $\rho(L|\mathscr{S})=+\infty$ is impossible, and formula (6) may be rewritten in form (4) which implies the boundedness of $\mathscr{S}$. The convexity and the closure of $\mathscr{S}$ result from the corresponding properties of $\mathscr{S}_k$ and those of intersection operation. This completes the proof.

We shall use another convenient formula for $\rho(L|\mathscr{S})$

$$\rho(L|\mathscr{S})=\inf_{\substack{\|\lambda_k\|=1 \\ \alpha_k \geqq 0}} \left\{\sum_{k=1}^N \alpha_k \lambda_k^T y_k+\sum_{k=1}^N \rho(-\alpha_k \lambda_k|\Xi_k): \sum_{k=1}^N \alpha_k \lambda_k x_k^T=L\right\}.$$

Corollary 1. Precise identification theorem.

Let us assume that the following conditions hold:

$$\forall \varepsilon>0, \quad \forall L \in \mathbf{R}^{m \times n}, \quad \exists I \subseteq \overline{1, N}, \quad \exists \lambda_k \in \Lambda_k(\varepsilon), \quad \exists \alpha_k \geqq 0:$$

$$: \sum_{k \in I} \alpha_k \lambda_k x_k^T=L, \tag{7}$$

$$\lim_{\varepsilon \to 0} \sum_{k \in I} \alpha_k \lambda_k^T (\bar{\xi}_k-\xi_k)=0, \tag{8}$$

where the vector $\xi_k \in U_k(\varepsilon)$ corresponds to vector λ_k from (7) according to the definition of $\Lambda_k(\varepsilon)$.

Then the information set $\mathscr{S}$ contains the single matrix C_*.

We remind that $y_k=C_* x_k+\bar{\xi}_k$, $k \in \overline{1, N}$.

Proof. The idea of the proof is to obtain the equality $\rho(L|\mathscr{S})=(L, C_*)$ for any matrix $L \in \mathbf{R}^{m \times n}$.

Let $L \in \mathbf{R}^{m \times n}$ be a fixed matrix. For any $\varepsilon>0$ hold

$$\rho(L|\mathscr{S}) \leqq \sum_{k=1}^N \alpha_k \lambda_k^T y_k+\sum_{k=1}^N \rho(-\alpha_k \lambda_k|\Xi_k)=$$

$$=\left(\sum_{k=1}^N \alpha_k \lambda_k x_k^T, C_*\right)+\sum_{k \in I} \alpha_k \lambda_k^T \bar{\xi}_k+\sum_{k \in I}(-\alpha_k \lambda_k^T \xi_k)=$$

$$=\left(\sum_{k \in I} \alpha_k \lambda_k x_k^T, C_*\right)+\sum_{k \in I} \alpha_k \lambda_k^T(\bar{\xi}_k-\xi_k) \triangleq \varphi(L, \varepsilon),$$

where $\alpha_k=0$, if $k \notin I$. If $k \in I$, then α_k, λ_k, ξ_k satisfy (7), (8). Hence,

$$\rho(L|\mathscr{S}) \leqq \inf_{\varepsilon>0} \{\varphi(L,\varepsilon)\} \leqq \lim_{\varepsilon\to 0} \varphi(L,\varepsilon)=(L, C_*) .$$

We have $C_* \in \mathscr{S}$, therefore the inequality $\rho(L|\mathscr{S}) \geqq (L, C_*)$ is true. Thus, $\rho(L|\mathscr{S})= =(L, C_*)$.

Corollary 2. If each set Ξ_k is a sphere with its centre in the origin:

$$\Xi_k=\{\xi : \|\xi\| \leqq v_k,\, v_k>0\}, \qquad k \in \overline{1, N}$$

and for some subset of indices $I \subseteq \overline{1, N}$ the conditions

$$\bar{\xi}_k \in \partial \Xi_k, \qquad k \in I , \tag{9}$$

$$\forall L \in \mathbf{R}^{m\times n}, \quad \exists \alpha_k \geqq 0, \quad \sum_{k\in I} \alpha_k \bar{\xi}_k x_k^T=L \tag{10}$$

hold, then the information set $\mathscr{S}$ contains the single matrix C_*.

Proof. Let $L \in \mathbf{R}^{m\times n}$ be a fixed matrix. Define $Q=-L$. From condition (10) one can find $\alpha_k \geqq 0$, for which $\sum\limits_{k\in I} \alpha_k \bar{\xi}_k x_k^T=Q$. For $k \notin I$, let $\alpha_k=0$. Denote $\lambda_k \triangleq -\alpha_k \bar{\xi}_k$, $k \in \overline{1, N}$, then $\sum\limits_{k=1}^{N} \lambda_k x_k^T=L$.

For the sets Ξ_k mentioned above, the equality $\rho(-\lambda_k|\Xi_k)=\|\lambda_k\| \cdot v_k$ is true. Therefore, using (4), (9), we obtain

$$\rho(L|\mathscr{S}) \leqq \sum_{k=1}^{N} \lambda_k^T y_k + \sum_{k=1}^{N} \|\lambda_k\| \cdot v_k=$$

$$=\sum_{k=1}^{N} (-\alpha_k \bar{\xi}_k x_k^T, C_*)+\sum_{k=1}^{N} (-\alpha_k \bar{\xi}_k^T \bar{\xi}_k)+\sum_{k=1}^{N} \|\alpha_k \bar{\xi}_k\| \cdot v_k=(L, C_*) .$$

Since $C_* \in \mathscr{S}$, then $\rho(L|\mathscr{S}) \geqq (L, C_*)$. Hence, $\rho(L|\mathscr{S})=(L, C_*)$.

Examples

Consider the examples of inputs and disturbances for which formula (4) permits us to identify matrix C precisely.

In all the examples the following type of constraints is used:

$$\Xi_k \equiv \Xi \triangleq \{\xi : \|\xi\| \leqq 1\} .$$

Besides, the validity of condition (9) is assumed for any $k \in \overline{1, N}$.

Example 1. Let the following conditions hold:

1) input x_k is periodic with period $R \geqq n+1$,
2) $\mathscr{L}^+\{x_r\}_{r=1}^R = \mathbf{R}^n$,
3) disturbance $\bar{\xi}_k$ is periodic with period $P \geqq m$,
4) $\mathscr{L}\{\bar{\xi}_p\}_{p=1}^P = \mathbf{R}^m$,
5) P and R are relative prime and $N = P \times R$.

Then formula (4) allows us to identify matrix C precisely. It follows from conditions 1), 3), 5) that there exists a one-to-one correspondence between pairs (r, p), $r \in \overline{1, R}$, $p \in \overline{1, P}$ and numbers $k \in \overline{1, N}$, such that $x_k = x_r$, $\bar{\xi}_k = \bar{\xi}_p$. For the proof of this fact we fix k, $k \in \overline{1, N}$ and find r and p such that $x_k = x_r$, $\bar{\xi}_k = \bar{\xi}_p$, i.e. $k = r + \alpha_1 R$, $k = $ $= p + \beta_1 P$. For some $l \in \overline{1, N}$ assume that there are vectors $x_l = x_r$, $\bar{\xi}_l = \bar{\xi}_p$, i.e. $l = r + \alpha_2 R$, $l = p + \beta_2 P$. Without loss of generality let $l \geqq k$, then

$$l - k = (\alpha_2 - \alpha_1) R$$

$$l - k = (\beta_2 - \beta_1) P\,.$$

Therefore, the difference $l - k$ is divided by N without a remainder. Since $l - k < N$, then $l = k$.

Thus, for each number k we find at least one pair (r, p), and for each pair (r, p) we find no more than one number k. Since the number of pairs (r, p) is equal to N, then one-to-one correspondence in question takes place. We shall denote this correspondence as $k = k(r, p)$.

Let $L \in \mathbf{R}^{m \times n}$ be a fixed matrix. Condition (4) implies the existence of set $\{z_p : z_p \in \mathbf{R}^n\}_{p=1}^P$ with the property

$$\sum_{p=1}^{P} \bar{\xi}_p z_p^T = L.$$

It follows from conditions 1), 2) and the correspondence mentioned above that there exist non-negative scalars $\beta_{k(r,p)}$ for any $p \in \overline{1, P}$ which satisfy

$$\sum_{r=1}^{R} \beta_{k(r,p)} x_{k(r,p)} = z_p\,.$$

Thus,

$$L = \sum_{p=1}^{P} \bar{\xi}_p \sum_{r=1}^{R} \beta_{k(r,p)} x_{k(r,p)} = \sum_{k=1}^{N} \xi_k \beta_k x_k^T\,,$$

i.e. condition (10) holds.

Example 2. Let the following conditions hold for $N=2n$:

1) $\bar{\xi}_k \equiv \bar{\xi} \in \partial \Xi$, $k \in \overline{1, N}$,
2) set of inputs is $\{x_1, \ldots, x_n, -x_1, \ldots, -x_n\}$,
3) $\mathscr{L}\{x_k\}_{k=1}^n = \mathbf{R}^n$.

Then formula (4) allows us to identify matrix C precisely.

We note in this example that the interval of observation is independent on m (m is dimension of observation vector), and what is more, conditions of corollary 2 are invalid for $m \geqq 2$.

The proof in outline is the following: for a fixed $L \in \mathbf{R}^{m \times n}$ and $\varepsilon > 0$ we find $\lambda_k \in \mathbf{R}^m$ and $\alpha_k \geqq 0$, which satisfy (r), and we prove that $\Sigma \alpha_k \lambda_k^T (\bar{\xi} - \psi_k) = B \cdot \varepsilon$, where B does not depend on ε and the vector $\psi_k \in U_k(\varepsilon)$ corresponds to $\lambda_k \in \Lambda_k(\varepsilon)$.

Let $L \in \mathbf{R}^{m \times n}$ be a fixed matrix, and $\varepsilon > 0$. Condition 3) implies the existence of the set $\{z_k : z_k \in \mathbf{R}^m\}_{k=1}^n$ with the property

$$\sum_{k=1}^{n} z_k x_k^T = L .$$

For each $k \in \overline{1, n}$ we choose λ_k, λ_{n+k}, α_k, α_{n+k}, ψ_k, ψ_{n+k} as follows:

a) If $|\bar{\xi}^T z_k| = \|z_k\|$, then $\lambda_k = \lambda_{n+k} = -\xi$; if $z_k = 0$, then $\alpha_k = \alpha_{n+k} = 0$; if $z_k = -\bar{\xi} \cdot \|z_k\|$, then $\alpha_k = \|z_k\|$, $\alpha_{n+k} = 0$; if $z_k = \xi \cdot \|z_k\|$, then $\alpha_k = 0$, $\alpha_{n+k} = \|z_k\|$. We note that these α_k and λ_k, $k \in \overline{1, N}$ do not depend on ε.

b) If $|\bar{\xi}^T z_k| < \|z_k\|$, then we denote by P_k the intersection of boundary of Ξ with the plane, which contains points ξ, z_k and the origin of $\mathbf{R}^m$. Hence, P_k is the circle, which contains points $\bar{\xi}$, $v_k \triangleq z_k \cdot \|z_k\|^{-1}$. Without loss of generality assume that $\varepsilon < \|v_k - \bar{\xi}\|$. Hence, there exists a point ψ_{n+k} and numbers γ_k, β_k, such that

$$\|\psi_{n+k} - \bar{\xi}\| = \varepsilon ,$$

$$0 < \gamma_k, \quad \beta_k < \frac{1}{\sin \varphi_{n+k}},$$

$$v_k = \gamma_k(-\bar{\xi}) + \beta_k \psi_{n+k} ,$$

where φ_{n+k} is the angle between $\bar{\xi}$ and ψ_{n+k}. Here we denote $\alpha_k = \gamma_k \cdot \|z_k\|$, $\alpha_{n+k} = \beta_k \cdot \|z_k\|$, $\lambda_k = -\bar{\xi}$, $\lambda_{n+k} = -\psi_{n+k}$.

In cases a) and b) when $\lambda_k = -\bar{\xi}$, $k \in \overline{1, N}$, let $\psi_k = \bar{\xi}$, $\varphi_k = \varphi$, $0 < \varphi < \frac{\pi}{2}$. For $k \in \overline{1, n}$, let $z_{n+k} = z_k$.

For these α_k, λ_k, ψ_k, $k \in \overline{1,N}$ the following equalities are true:

$$\sum_{k=1}^{N} \alpha_k \lambda_k x_k^T = L\,,$$

$$\rho(-\alpha_k \lambda_k | \Xi) = -\alpha_k \lambda_k^T \psi_k\,,$$

where $\lambda_k \in \Lambda_k(\varepsilon)$, $\alpha_k \leqq \dfrac{\|z_k\|}{\sin \varphi_k}$. Since $0 \leqq \lambda_k^T(\xi - \psi_k) < \varepsilon \sin \varphi_k$, then

$$\sum_{k=1}^{N} \alpha_k \lambda_k^T(\xi - \psi_k) < \sum_{k=1}^{N} \frac{\|z_k\|}{\sin \varphi_k} \cdot \varepsilon \sin \varphi_k = \sum_{k=1}^{N} \|z_k\| \cdot \varepsilon\,,$$

where $\sum_{k=1}^{N} \|z_k\|$ does not depend on ε.

Thus, conditions of corollary 1 hold.

Example of precise identification in the case of uncertain noise. The above results can be applied in the case when the disturbance is stochastic. For illustration consider the following example.

Example 3. Let the following conditions hold:

1) $\{\xi_k\}_{k=1}^{\infty}$ is the realization of sequence of independent stochastic variables with identical density functions, when density functions are nontrivial at any point $\xi \in \Xi$,

2) input x_k is periodic with period $R \geqq n$,

3) $\mathscr{L}\{x_r\}_{r=1}^{R} = \mathbf{R}^n$.

Then, the equality

$$\mathbf{P}\left[\lim_{N\to\infty} \mathscr{S}_N = \{C_*\}\right] = 1 \tag{11}$$

holds. Here $\mathbf{P}[A]$ denotes the probability of event A, $\mathscr{S}_N$ is the information set for $k \in \overline{1,N}$, and $\lim\limits_{N\to\infty} \mathscr{S}_N = \{C_*\}$ means convergence in Hausdorf metric [1].

Let $\{\psi_p : \psi_p \in \partial\Xi\}_{p=1}^{P}$ be a fixed set of vectors with the property $\mathscr{L}^+\{\psi_p\}_{p=1}^{P} = \mathbf{R}^m$. Without loss of generality assume that $R = n$, $P = m+1$.

Let ε be a fixed scalar, for which $0 < \varepsilon < \dfrac{1}{2} \min\limits_{\substack{p,t \in \overline{1,P} \\ p \neq t}} \{\|\psi_p - \psi_t\|\}$. According to Borel–Cantelli lemma, for each vector ψ_p and number $r \in \overline{1,R}$, with probability 1 there exists the number $k = k(r, p, \varepsilon)$ with properties $\|\psi_p - \xi_k\| \leqq \varepsilon$ and $x_k = x_r$. Further, for each triplet (r, p, ε) we shall consider a minimal number k with these properties. Hence, $\psi_p \in U_k(\varepsilon)$, and the corresponding vector $\lambda_k \in \Lambda_k(\varepsilon)$ is $\lambda_k = -\psi_p$. By $I(\varepsilon)$ we denote the set of numbers $k = k(r, p, \varepsilon)$, $r \in \overline{1,R}$, $p \in \overline{1,P}$, and we denote $N(\varepsilon) = \max\{k : k \in I(\varepsilon)\}$.

Let $L \in \mathbf{R}^{m\times n}$ be a fixed matrix. According to conditions 2), 3) one may select the set of vectors $z_r \in \mathbf{R}^m$, such that $\sum_{r=1}^{R} z_r x_r^T = -L$.

For each vector z_r the representation $z_r = \sum_{p=1}^{P} \beta_{rp}\psi_p$ holds, where the coefficients $\beta_{rp} \geqq 0$ do not depend on ε. Denote $\alpha_{k(r,p,\varepsilon)} = \beta_{rp}$. Then, with probability 1, we have

$$\sum_{k\in I(\varepsilon)} \alpha_k \lambda_k x_k^T = \sum_{r=1}^{R}\sum_{p=1}^{P}(-\beta_{rp})\psi_p x_r^T = L\,,$$

$$\left|\sum_{k\in I(\varepsilon)} \alpha_k \lambda_k^T(\xi_k - \psi_p)\right| \leqq \sum_{r=1}^{R}\sum_{p=1}^{P} \beta_{rp}\varepsilon\,,$$

where $B \triangleq \sum_{r=1}^{R}\sum_{p=1}^{P}\beta_{rp}$ does not depend on ε.

Consider the sets $\mathscr{S}_N^*$ given by

$$\mathscr{S}_N^* = \{C : C = C_1 - C_*,\ C_1 \in \mathscr{S}_N\}\,.$$

For all N, matrix C_* belongs to $\mathscr{S}_N$, therefore

$$0 \leqq \rho(L|\mathscr{S}_{N(\varepsilon)}^*) \leqq \sum_{k\in I(\varepsilon)} \alpha_k\lambda_k^T y_k + \sum_{k\in I(\varepsilon)} \rho(-\alpha_k\lambda_k|\Xi) - (L, C_*) =$$

$$= \sum_{k\in I(\varepsilon)} \alpha_k\lambda_k^T(C_* x_k + \xi_k) - \sum_{k\in I(\varepsilon)} \alpha_k\lambda_k^T\psi_p - (L, C_*) =$$

$$= \sum_{k\in I(\varepsilon)} \alpha_k\lambda_k^T(\xi_k - \psi_p) \leqq B\cdot\varepsilon\,,$$

i.e. with probability 1 the equality

$$\lim_{\varepsilon\to 0}\rho(L|\mathscr{S}_{N(\varepsilon)}^*) = 0 \tag{12}$$

holds.

Since equality (12) takes place for an arbitrary matrix $L \in \mathbf{R}^{m\times n}$, equality (11) is true.

References

1. *Куржанский А. Б.* Управление и наблюдение в условиях неопределенности. М., Наука, 1977.
2. *Rockafellar, R. T.*, Convex Analysis. Princeton Univ. Press, 1970.

О задаче идентификации матричного параметра

А. М. УСТЮЖАНИН

(Свердловск)

Рассматривается задача идентификации постоянного матричного параметра в уравнении наблюдения

$$y_k = Cx_k + \xi_k, \quad k \in \overline{1, N}, \qquad \text{где} \quad y_k, \xi_k \in \mathbf{R}^m, C \in \mathbf{R}^{m \times n}, x_k \in \mathbf{R}^n.$$

x_k и y_k при $k \in \overline{1, N}$ считаются известными. Относительно помехи ξ_k предполагается, что $\xi_k \in \Xi_k$ при $k \in \overline{1, N}$, где Ξ_k — заданные выпуклые компакты в $\mathbf{R}^m$ с непустой внутренностью.

Получен вид опорной функции информационного множества (т.е. множества матриц, совместимых с уравнением наблюдения и ограничением на помеху) и условие его ограниченности. Приведены условия точной идентификации, т.е. вырождения информационного множества в точку.

Результаты проиллюстрированы примерами для случаев периодической, постоянной и случайной помехи ξ_k.

А. М. Устюжанин
Институт математики и механики
Уральского научного центра АН СССР
ГСП-384, Свердловск, ул. С. Ковалевской, 16

*Problems of Control and Information Theory, Vol. 15 (4), pp. 275—286 (*1986*)*

ON USING OF BUNDLE METHODS IN NONDIFFERENTIABLE OPTIMAL CONTROL PROBLEMS

J. V. OUTRATA, Z. SCHINDLER

(*Prague*)

(Received September 1, 1985)

The paper contains three general assertions concerning the computation of points from the generalized gradients of objectives in nonsmooth optimal control problems with respect to the control variable. They depend on various rules of the calculus of generalized gradients and the implicit function theorem of Clarke. With the help of them we may compute the above mentioned points in the same way as gradients, i.e. by means of suitable adjoint equations. Consequently, effective bundle or subgradient methods may be used to the numerical solution of the given nonsmooth optimal control problems.

Introduction

Optimization problems with nondifferentiable functions in objectives and/or constraints appear frequently in many situations starting with exact penalties over Chebyshev approximation to various threshold or saturation phenomena etc. Concerning optimal control problems, we may encounter these difficulties especially in modern economic models. They are due to e.g. the fixed proportions law, known from the classical theory of capital growth or various nonsmooth dependences appearing in these models (unit price of the inventory space on the overall hired space, unit production cost on the overall production). Also the quantity of products which may be sold by a producer is given by

$$\min\{q(p), Q\},$$

where Q is the state of the inventory and q is the *demand* function depending on the price p.

Provided these models are sufficiently simple, they may be solved analytically with the help of the "nonsmooth" variant of the Pontryagin maximum principle, derived by Clarke in [1]. Examples of this elegant way can be found in papers [3], [4]. But with the increasing number of state and control variables and the increasing complexity of the problem the analytical solution becomes impossible and then we have to be satisfied with a numerical solution. To obtain it, we recommend a bundle method of the type [6], in which, however, we must be able to compute at every admissible control at least one element of the generalized gradient of Clarke of the

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

objective (or eventually augmented objective in the case of state-space constraints) with respect to the control variable.

The aim of the present paper is to give certain general guide-lines concerning the computation of points of generalized gradients in optimal control problems. An adequate knowledge of the nonsmooth analysis is assumed; in this respect we refer to [2]. Section 2 contains the description of two important economic optimization models which may be numerically solved using the presented approach.

In what follows, ∂f is the generalized gradient of a scalar-valued function or the generalized Jacobian according to the nature of f (in the sense of Clarke). $\partial_H f$ is the generalized derivative introduced in [5] in the following way: We assume that $f[X \to Y]$ is locally Lipschitz, X is a separable Banach space and Y is a reflexive separable Banach space. Let H be such a subset of X that f is Gâteaux differentiable over H and its complement $X \backslash H$ is of Haar measure zero.

Then, for $x_0 \in X$

$$\partial_H f(x_0) = \overline{\mathrm{co}} \left\{ \lim_{i \to \infty} \nabla f(x_i),\ x_i \to x_0,\ x_i \in H \right\},$$

where the limit and the closure of the convex hull are taken in the space $L_S(X, Y_\sigma)$, which means the space of continous linear maps endowed with the simple convergence topology with respect to the $\sigma(Y, Y^*)$ topology of Y. For $T \subset L_S(X, Y_\sigma)$, plen $\{T\}$ means the *plenary hull* of T, i.e. the set

$$\{A \in L_S(X, Y_\sigma) \mid \langle y^*, Ax \rangle \leqq \sup_{B \in T} \langle y^*, Bx \rangle \forall x \in X, y^* \in Y^*\}.$$

$\mathfrak{a}_x(a)$ denotes the filter of all (norm) neighbourhoods of a in a normed space X, $f^0(x, h)$ is the directional derivate of f at x in the direction h (in the sense of Clarke).

1. Generalized gradients of the objectives in optimal control problems

We consider the following general model

$$\begin{array}{ll} & I(x, u) \to \inf \\ \text{subj. to} & A(x, u) = 0 \\ & u \in \omega \subset U \\ & x \in X, \end{array} \tag{$\wp$}$$

where X, U is the state, control space respectively, $I[X \times U \to \mathbf{R}]$ is the objective, $A[X \times U \to X]$ defines the system equation $A(x, u) = 0$ and ω is the set of admissible controls. We suppose that incidental state-space constraints have been already augmented to the objective by a suitable penalty. It is also assumed that X and U are Banach, I and A are locally Lipschitz and the system equation defines a unique implicit function μ on U such that $A(\mu(u), u)=0$ on U.

With respect to the numerical solution of $(\mathscr{p})$ we rewrite it into the mathematical programming form

$$\Theta(u)=I(\mu(u), u)\rightarrow \inf \qquad (\mathscr{mp})$$

subj. to

$$u \in \omega$$

and our task here is the computation of a vector $\xi \in \partial\Theta(u)$ for all $u \in \omega$.

Proposition 1.1. Let A be a C_1 map over $X \times U$, $\bar{u} \in \omega$ and $\bar{x}=\mu(\bar{u})$. Let μ be a C_1 map over U, $(\bar{\xi}, \bar{\eta}) \in \partial I(\bar{x}, \bar{u})$ and λ^* be a solution of the adjoint equation

$$(\nabla_x A(\bar{x}, \bar{u}))^* \lambda^* + \bar{\xi} = 0 . \qquad (1.1)$$

Then

$$(\nabla_u A(\bar{x}, \bar{u}))^* \lambda^* + \bar{\eta} \in \partial\Theta(\bar{u}) , \qquad (1.2)$$

provided either I (or $-I$) is regular at $(\bar{x}, \bar{u})$ or μ maps every neighbourhood of $\bar{u}$ onto a set which is dense in some neighbourhood of $\bar{x}$ (in particular, if the derivative $\nabla\mu(\bar{u})$ is surjective).

Proof. Θ is clearly locally Lipschitz over U so that we are entitled to compute $\partial\Theta$ at any $\bar{u} \in U$. Let us denote $F[U \rightarrow X \times U]$ the map given by

$$F: u \mapsto (\mu(u), u) .$$

This map is continuously Fréchet differentiable over U; hence we may utilize the well-known chain-rule II of [2] which gives under our assumptions immediately

$$\partial\Theta(\bar{u}) = (\nabla F(\bar{u}))^* \partial I(\bar{x}, \bar{u}) \ni (\nabla\mu(\bar{u}))^* \bar{\xi} + \bar{\eta} .$$

It remains to express the operator $(\nabla\mu(\bar{u}))^*$ by means of derivatives $\nabla_x A(\bar{x}, \bar{u}), \nabla_u A(\bar{x}, \bar{u})$.

Clearly, for any $u \in U$ $A(\mu(u), u)=0$, so that

$$\nabla_x A(\bar{x}, \bar{u}) \nabla\mu(\bar{u}) + \nabla_u A(\bar{x}, \bar{u}) = 0 .$$

Thus, on using the adjoint equation, one has for $k \in U$

$$\langle(\nabla\mu(\bar{u}))^* \bar{\xi}, k\rangle = \langle -(\nabla_x A(\bar{x}, \bar{u}))^* \lambda^*, \nabla\mu(\bar{u})k\rangle =$$
$$= \langle \lambda^*, \nabla_u A(\bar{x}, \bar{u})k\rangle = \langle(\nabla_u A(\bar{x}, \bar{u}))^* \lambda^*, k\rangle . \qquad \square$$

Remark. In many practical cases $\partial_x I(\bar{x}, \bar{u}) \times \partial_u I(\bar{x}, \bar{u}) \subset \partial I(\bar{x}, \bar{u})$, then it suffices to take $\bar{\xi} \in \partial_x I(\bar{x}, \bar{u})$, $\bar{\eta} \in \partial_u I(\bar{x}, \bar{u})$.

The following assertion hinges on a complement to the implicit function theorem of Clarke, cf. [2]. We suppose that $X=\mathbf{R}^n$, $U=\mathbf{R}^m$ $(A[\mathbf{R}^n\times\mathbf{R}^m\rightarrow\mathbf{R}^n])$ and denote for a fixed couple $(\bar{x},\bar{u})$ by $\pi_x\partial A(\bar{x},\bar{u})$ the set of all $[n\times n]$ matrices M such that, for some $[n\times m]$ matrix N, the $[n\times(n+m)]$ matrix $[M,N]\in\partial A(\bar{x},\bar{u})$. Furthermore, we denote by Γ_A the open (possibly empty) set of couples (x,u) that admit a neighborhood $\mathcal{O}$ where the gradient ∇A is continous.

Proposition 1.2. Let I be a C_1 function, $\bar{u}\in\omega$ and $\bar{x}=\mu(\bar{u})$. Let all matrices of $\pi_x\partial A(\bar{x},\bar{u})$ be nonsingular and $(\bar{x},\bar{u})\in\bar{\Gamma}_A$. We assume, furthermore, that $\{(x_i,u_i)\}$ is a sequence converging to $(\bar{x},\bar{u})$,

$$(x_i,u_i)\in\Gamma_A\cap\text{graph }\mu\text{ for all }i \tag{1.3}$$

and $\{(x_{i'},u_{i'})\}$ is a subsequence of $\{(x_i,u_i)\}$ that $\lim\limits_{i'\rightarrow\infty}\nabla A(x_{i'},u_{i'})$ exists. Finally, let λ^* be the solution of the adjoint equation

$$\left(\lim_{i'\rightarrow\infty}\nabla_x A(x_{i'},u_{i'})\right)^*\lambda^*+\nabla_x I(\bar{x},\bar{u})=0\,. \tag{1.4}$$

Then

$$\left(\lim_{i'\rightarrow\infty}\nabla_u A(x_{i'},u_{i'})\right)^*\lambda^*+\nabla_u I(\bar{x},\bar{u})\in\partial\Theta(\bar{u})\,. \tag{1.5}$$

Proof. Note that

$$\lim_{i'\rightarrow\infty}\nabla A(x_{i'},u_{i'})=\left[\lim_{i'\rightarrow\infty}\nabla_x A(x_{i'},u_{i'}),\ \lim_{i'\rightarrow\infty}\nabla_u A(x_{i'},u_{i'})\right]$$

exists due to the Lipschitz continuity and belongs to $\partial A(\bar{x},\bar{u})$ by the definition; hence the matrix $\lim\limits_{i'\rightarrow\infty}\nabla_x A(x_{i'},u_{i'})$ is nonsingular because of our assumption. The continuity of ∇A on Γ_A implies furthermore the existence of a natural number k_0 such that for $i'>k_0$ $\nabla_x A(x_{i'},u_{i'})$ is nonsingular. For these i'

$$\nabla\mu(u_{i'})=-(\nabla_x A(x_{i'},u_{i'}))^{-1}\nabla_u A(x_{i'},u_{i'})$$

due to the classical implicit function theorem and, consequently

$$\lim_{i'\rightarrow\infty}[-(\nabla_x A(x_{i'},u_{i'}))^{-1}\,\nabla_u A(x_{i'},u_{i'})]=$$

$$=-\,(\lim_{i'\rightarrow\infty}\nabla_x A(x_{i'},u_{i'}))^{-1}\,\lim_{i'\rightarrow\infty}\nabla_u A(x_{i'},u_{i'})\in\partial\mu(\bar{u})\,.$$

The derivation of (1.4), (1.5) proceeds now as in the proof of Proposition 1.1. An infinite-dimensional version of the above statement presented below is substantionally more restrictive due to the absence of a suitable implicit function theorem and difficulties with the extension of generalized Jacobians to infinite dimensions. We assume that U is separable, X is reflexive separable and A possesses a special structure

$$A(x, u) = A_1(x, B(u))\,, \tag{1.6}$$

where $B[U \to Z]$ is locally Lipschitz and Z is a reflexive separable Banach space.

Proposition 1.3. Let I and A_1 be C_1 over $X \times U$, $X \times Z$, respectively, $\bar{u} \in \omega$ and $\bar{x} = \mu(\bar{u})$. Let $\nabla_x A_1(\bar{x}, \bar{u})$ be a linear homeomorphism of X onto X and B be Gâteaux differentiable on a set $H \subset U$ with $U \backslash H$ being of Haar measure zero. Finally, assume that λ^* solves the adjoint equation

$$(\nabla_x A_1(\bar{x}, \bar{v}))^* \lambda^* + \nabla_x I(\bar{x}, \bar{u}) = 0\,, \tag{1.7}$$

where $\bar{v} = B(\bar{u})$ $(v = B(u))$. Then

$$M^* (\nabla_v A_1(\bar{x}, \bar{v}))^* \lambda^* + \nabla_u I(\bar{x}, \bar{u}) \in \partial \Theta(\bar{u}) \tag{1.8}$$

for all $M \in \operatorname{plen}\{\partial_H B(\bar{u})\}$.

Proof. According to the classical implicit function theorem there exists a neighbourhood $\mathcal{O} \in \mathfrak{a}_Z(\bar{v})$ and a unique map $\kappa[\mathcal{O} \to X]$ such that $\kappa(\bar{v}) = \bar{x}$, $A_1(\kappa(v), v) = 0$ on $\mathcal{O}$ and

$$\nabla \kappa(\bar{v}) = -(\nabla_x A_1(\bar{x}, \bar{v}))^{-1}\, \nabla_v A_1(\bar{x}, \bar{v})\,.$$

Using the generalized derivative chain rule of [9] we obtain that

$$\partial_H \mu(\bar{u}) = \nabla \kappa(\bar{v}) \circ \partial_H B(\bar{u})\,.$$

Furthermore, by the corollary of the above-mentioned rule for $h \in U$

$$\begin{aligned}
\Theta^0(\bar{u}, h) = {} & \max_{R \in \partial_H \mu(\bar{u})} \langle \nabla_x I(\bar{x}, \bar{u}), Rh \rangle + \langle \nabla_u I(\bar{x}, \bar{u}), h \rangle = \\
= {} & \max_{S \in \partial_H B(\bar{u})} \langle -[\nabla_v A_1(\bar{x}, \bar{v})]^* ([\nabla_x A_1(\bar{x}, \bar{v})]^*)^{-1} \nabla_x I(\bar{x}, \bar{u}), Sh \rangle + \langle \nabla_u I(\bar{x}, \bar{u}), h \rangle = \\
= {} & \max_{S \in \partial_H B(\bar{u})} \langle [\nabla_v A_1(\bar{x}, \bar{v})]^* \lambda^*, Sh \rangle + \langle \nabla_u I(\bar{x}, \bar{u}), h \rangle \geqq \\
\geqq {} & \langle M^* [\nabla_v A_1(\bar{x}, \bar{v})]^* \lambda^* + \nabla_u I(\bar{x}, \bar{u}), h \rangle
\end{aligned}$$

for any $M \in \operatorname{plen}\{\partial_H B(\bar{u})\}$ by definition. □

The above assertion can be applied e.g. to optimal control of a linear plant with a dead band. A theoretical analysis of this problem can be found in [7]; here we bring some numerical results for the discrete-time version of this problem. These results have been obtained using the code described in [6], and for the computation of elements from $\partial\Theta$ Proposition 1.3 has been applied.

Example (discrete-time optimal control problem with a dead band).

$$\frac{1}{2}\sum_{i=0}^{23}\sum_{j=1}^{2}(u_i^j)^2 \to \min$$

subj. to

$$x_{i+1}=\begin{bmatrix} 0.905 & 0.092 & 0 & 0 \\ 0 & 0.932 & 0 & 0 \\ 0 & 0 & 0.97 & 0.095 \\ 0 & 0 & 0 & 0.932 \end{bmatrix} x_i +$$

$$+\begin{bmatrix} 0.095u_i^1+0.005u_i^2 \\ 0.097u_i^2 \\ 0.005u_i^1+\alpha(u_i^2) \\ 0.097u_i^1 \end{bmatrix}, \qquad i=0,1,\ldots,23,$$

where $\alpha[\mathbf{R}\to\mathbf{R}]$ is given by

$$\alpha(\xi)=\begin{cases} 0.1(\xi-d) & \text{if } \xi\geqq d \\ 0 & \text{if } \xi\in(-d,d) \\ 0.1(\xi+d) & \text{if } \xi\leqq -d \end{cases}$$

and $2d$ is the width of the dead band,

$$x_{24}=0,$$

$$-1.5\leqq u_i^j\leqq 1.5, \qquad i=0,1,\ldots,23, \quad j=1,2.$$

We have treated the terminal state constraint by means of the classical exterior quadratic penalty with a properly chosen penalty parameter and solved the problem three times for $d=0$, $d=0.5$ and $d=1.0$. The results are interesting and confirm the well-known fact that a suitably chosen nonlinearity can improve the properties of the control loop.

d	Optimal cost value
0	31.93726
0.5	29.01725
1.0	29.98727

The approximate optimal controls for $d=0$ and $d=1.0$ are depicted in Figs. 1 and 2, respectively.

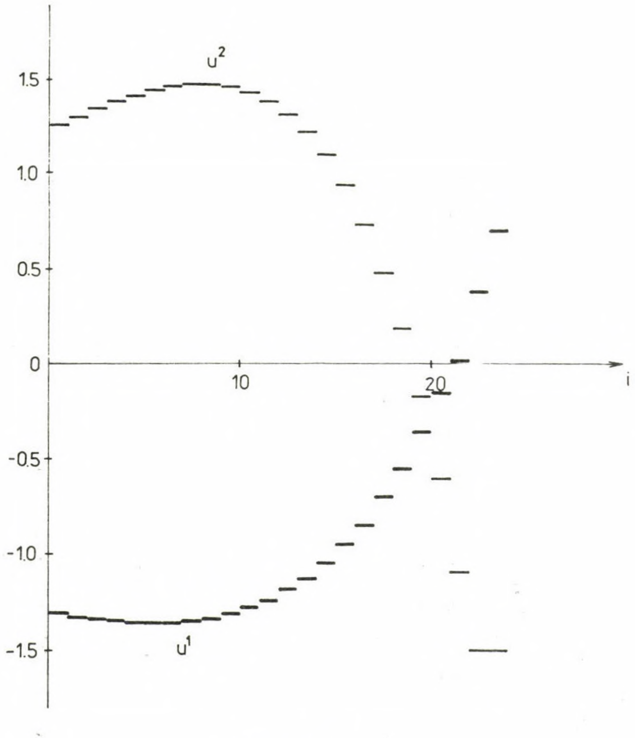


Fig. 1. d=0

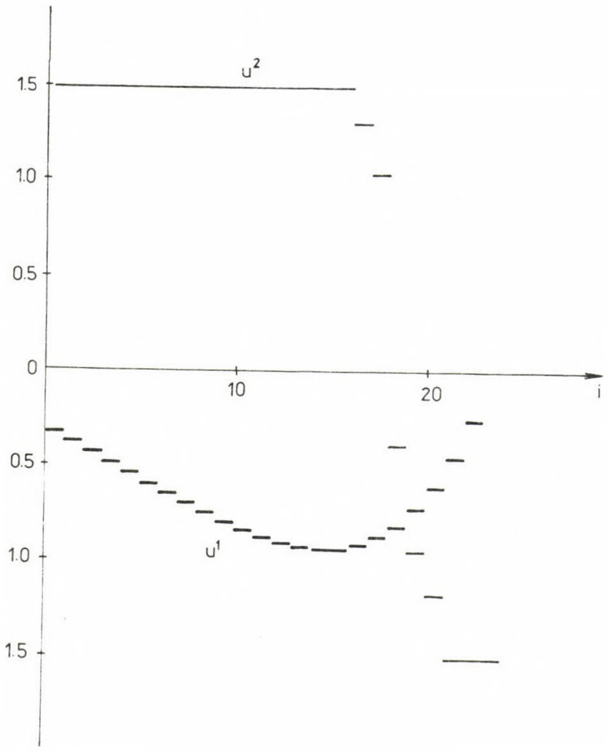


Fig. 2. d=1.0

2. Economic applications

In this section we describe briefly two economic models, solvable numerically by means of a bundle algorithm. The first is an employment-production-inventory model

$$e^{-\vartheta T}(S_1 x(T)+S_2 y(T))+\int_0^T e^{-\vartheta t}[c(f(y))+wy+k(u)+\psi(x)+\varphi(\eta(p)-o)-po]\,dt \to \inf$$

subj. to

$$\begin{aligned}
&\dot{x}=f(y)-o && \text{a.e. on } [0,T],\, x(0)\geqq 0 \text{ given}\\
&\dot{y}=u-\sigma(w)y && \text{a.e. on } [0,T],\, y(0)\geqq 0 \text{ given}\\
&x\geqq 0 &&\\
&y\geqq 0 &&\\
&w\geqq \bar{w}` &&\\
&p\geqq 0 &&\\
&0\leqq o\leqq \eta(p) && \text{for all } t\in[0,T],
\end{aligned} \tag{2.1}$$

where T is a given finite horizon, ϑ is an interest rate assumed to be constant and positive through time, x is the inventory, y is the level of employment assumed to be homogeneous, w is the wage, u is the rate of recruitment or discharge, p is the selling price charged by the firm, o is the actual supplied output, S_1, S_2 are the (constant) salvage values of the inventory and employment level, respectively, f is the production function, $c\circ f$ are the production costs, k is the labour adjustment cost function, ψ is the holding cost, η is the demand function, φ is the shortage cost and σ is the voluntary decrease of employment.

Problem (2.1) has two state variables (x, y) and four controls (w, u, p, o). We will assume that $\sigma, c, f, k, \eta[\mathbf{R}\to\mathbf{R}]$ are C_1 functions,

$$\psi(x)=\begin{cases} 0 & \text{for } x<0\\ h_1 x & \text{for } 0\leqq x<x_1\\ h_2 x-(h_2-h_1)x_1 & \text{for } x\geqq x_1, \end{cases} \tag{2.2}$$

with h_1, h_2 given positive constants and

$$\varphi(\kappa)=d(\kappa)^+ \tag{2.3}$$

with d being a given positive constant. The piecewise linear form of the holding costs was taken from [4] and corresponds to a certain warehousing constraint x_1. For inventory levels $x>x_1$ an additional space has to be rented at a unit cost of $h_2>h_1$.

It remains to augment suitably the inequality state-space and control constraints which may be done e.g. by replacing ψ by $\tilde{\psi}$ and φ by $\tilde{\varphi}$, where

$$\tilde{\psi}(t, x)= \begin{cases} \psi(x) & \text{for } \; x \geqq 0 \\ -re^{\vartheta t}x & \text{for } \; x<0 , \end{cases} \tag{2.4}$$

$$\tilde{\varphi}(t, \kappa)= \begin{cases} d\kappa & \text{for } \; \kappa \geqq 0 \\ -re^{\vartheta t}\kappa & \text{for } \; \kappa<0 , \end{cases} \tag{2.5}$$

and adding the penalty term $r(-y)^+$ to this new objective. r is a suitable penalty parameter. Proposition 1.1 may now be utilized with $U=L_\infty[0, T, \mathbf{R}^4]$ and $X=C_0[0, T, \mathbf{R}^2]$.

Proposition 2.1. Let $(\bar{w}, \bar{u}, \bar{p}, \bar{o}) \in L_\infty[0, T, \mathbf{R}^4]$ and $(\bar{x}, \bar{y})$ be the corresponding trajectory. Let (s, q) be the solution of the adjoint differential equations

$$\begin{aligned} &\dot{s}(t)=e^{-\vartheta t}\beta_1(t, \bar{x}(t)) \\ &\dot{q}(t)= - \nabla f(\bar{y}(t))s(t)+\sigma(\bar{w}(t))q(t)+e^{-\vartheta t}\beta_2(\bar{y}(t))+\beta_3(\bar{y}(t)) \qquad \text{a.e.} \end{aligned} \tag{2.6}$$

on $[0, T]$ with the terminal conditions

$$s(T)= -e^{-\vartheta T}S_1 , \; q(T)= -e^{-\vartheta T}S_2 .$$

Functions $\beta_1[[0, T] \times \mathbf{R} \to \mathbf{R}]$, $\beta_2[\mathbf{R} \to \mathbf{R}]$ and $\beta_3[\mathbf{R} \to \mathbf{R}]$ are given by

$$\beta_1(t, \bar{x}(t))= \begin{cases} h_2 & \text{for } \; t \in N_2=\{t \in [0, T] \mid \bar{x}(t) \geqq x_1\} \\ h_1 & \text{for } \; t \in N_1=\{t \in [0, T] \mid 0 \leqq \bar{x}(t) < x_1\} \\ -re^{\vartheta t} & \text{for } \; t \in [0, T] \backslash (N_1 \cup N_2) , \end{cases}$$

$$\beta_2(\bar{y}(t))= \nabla c(f(\bar{y}(t))) \nabla f(\bar{y}(t))+\bar{w}(t) ,$$

$$\beta_3(\bar{y}(t))= \begin{cases} -r & \text{for } \; \bar{y}(t)<0 \\ 0 & \text{otherwise} . \end{cases}$$

Then the function $\gamma[[0, T] \to \mathbf{R}^4]$ given by

$$\gamma(t)= \begin{bmatrix} \bar{y}(t) \nabla \sigma(\bar{w}(t))q(t) \\ -q(t) \\ 0 \\ s(t) \end{bmatrix} + e^{-\vartheta t} \begin{bmatrix} \bar{y}(t) \\ \nabla k(\bar{u}(t)) \\ \beta_4(t, \bar{p}(t), \bar{o}(t)) - \bar{o}(t) \\ \beta_5(t, \bar{p}(t), \bar{o}(t)) - \bar{p}(t) \end{bmatrix}$$

with

$$\beta_4(t, \bar{p}(t), \bar{o}(t))= \begin{cases} d\nabla \eta(\bar{p}(t)) & \text{for } \; t \in N_3=\{t \in [0, T] \mid \eta(\bar{p}(t)) \geqq \bar{o}(t)\} \\ -re^{\vartheta t} \nabla \eta(\bar{p}(t)) & \text{for } \; t \in [0, T] \backslash N_3 , \end{cases}$$

$$\beta_5(t, \bar{p}(t), \bar{o}(t)) = \begin{cases} -d & \text{for} \quad t \in N_3 \\ re^{\vartheta t} & \text{for} \quad t \in [0, T] \backslash N_3 , \end{cases}$$

belongs to $\partial \Theta(\bar{w}, \bar{u}, \bar{p}, \bar{o})$.

Proof. The integrand in the augmented objective possesses three nonsmooth terms $\tilde{\psi}(t, x)$, $\tilde{\varphi}(t, \eta(p) - o)$ and $r(-y)^+$, all of them being regular. Hence, the whole integrand is regular and Proposition 1.1 may be applied. For the derivation of the adjoint equation (2.6) the standard way may be used as it is described e.g. in [9].□

The second economic problem comes from resource economics. In fact, it is a time-discretization and a slight modification of the problem which may be found in [2]. Here it attains the following from:

$$-y_k + \sum_{j=0}^{k-1} (\alpha_j v_j + \beta_j w_j) \to \inf$$

subj. to

$$\begin{aligned} & y_{j+1} = y_j + e^{-\vartheta j} \Pi_j \min(z_j, v_j) r(Q - q_j), \quad y_0 = 0 \\ & q_{j+1} = q_j + \min(z_j, v_j) r(Q - q_j), \quad q_0 = 0 \\ & z_{j+1} = z_j + w_j, \quad z_0 \geqq 0 \quad \text{given} \\ & \quad v_j \geqq 0 \\ & \quad w_j \geqq 0 \\ & \alpha_j v_j + \beta_j w_j \leqq U_j \\ & \quad q_k \leqq Q, \qquad j = 0, 1, \ldots, k-1, \end{aligned} \tag{2.7}$$

where $k \geqq 1$ is a given finite horizon, y_j is the discounted return up to time j, Π_j is the estimated price of the extracted raw in the time interval $[j, j+1)$, Q is the overall estimated stock of the resource at $j = 0$, q_j is the amount of extracted raw up to time j, v_j is the labour force usable in $[j, j+1)$, z_j is the number of machines usable at time j, w_j is the increase of the machine number in $[j, j+1)$, α_j is the unit labour cost in $[j, j+1)$, β_j is the unit cost of the machine equipment in $[j, j+1)$, U_j is the upper bound of expenses in $[j, j+1)$, r reflects the change of the extraction rate with the level of the remaining stock and ϑ is again the interest rate.

Problem (2.7) has three state variables (y, q, z) and two controls (v, w). We may include the terminal state constraint to the objective in some smooth way, e.g. by the exterior quadratic penalty or augmented Lagrangian technique. Then, Proposition 1.2 can be easily applied with $U = \mathbf{R}^{2k}$, $X = \mathbf{R}^{3k}$ which enables finally to solve the problem (2.7) numerically by a bundle method. This use of Proposition 1.2 is described in [7] in detail.

Remark. In optimal control problems, where nonsmooth terms appear in the objective as well as in the system dynamics, we may often transfer them from the objective to the system equation by introducing new state-variables. Thus, Proposition 1.2 concerns a fairly large class of problems.

References

1. *Clarke, F. H.*, The maximum principle under minimal hypotheses. SIAM J. Contr., Vol. **14**, 1976, pp. 1076–1091.
2. *Clarke, F. H.*, Optimization and Nonsmooth Analysis. Wiley, New York, 1983.
3. *Feichtinger, G.*, *Luptáčik, M.*, Optimal employment and wage policies of a monopolistic firm. Research Report No. *62*, Inst. for Econ. and Oper. Res., TU Vienna, 1983.
4. *Hartl, R. F.*, *Sethi, S. P.*, Optimal control problems with differential inclusions: Sufficiency conditions and an application to a production-inventory model. Optimal Contr. Appl. and Meth., Vol. **5**, 1984, pp. 289–307.
5. *Hiriart-Urruty, J.-B.*, *Thibault, L.*, Existence et caractérisation de differentielles généralisées d'applications localement lipschitziennes d'un espace de Banach séparable dans un espace de Banach réflexif séparable. C. R. Acad. Sc. Paris **290-A**, pp. 1091–1094.
6. *Lemaréchal, Cl.*, *Strodiot, J. J.*, *Bihain, A.*, On a bundle algorithm for nonsmooth optimization. NPS 4, Madison, 1980.
7. *Outrata, J. V.*, On the numerical solution of optimal control problems with nondifferentiable system equations. To appear.
8. *Polak, E.*, Computational Methods in Optimization. Acad. Press, New York, 1971.

Об использовании методов связки в негладких задачах оптимального управления

Й. В. ОУТРАТА, З. ШИНДЛЕР

(Прага)

В статье три общих утверждения, касающихся нахождения точек обобщенных градиентов критериев оптимальности в негладких задачах оптимального управления. Они зависят от правил вычисления обобщённых градиентов и связаны с теоремой о неявной функции Кларка. На основе этих предложений можно вычислять точки обобщённых градиентов с помощью сопряжённых уравнений аналогично гладкому случаю. Следовательно, для численного решения этих негладких задач можно использовать эффективные методы связки.

J. Outrata, Z. Schindler
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou Věží 4
182 08 Praha 8
Czechoslovakia

Problems of Control and Information Theory, Vol. 15 (4), pp. 287—298 (1986)

STOCHASTIC HEREDITARY CONTROL SYSTEMS

I. N. SINITSYN

(*Moscow*)

(Received April 24, 1985)

The equations of the statistical dynamics for stochastic hereditary control systems described by linear and non-linear stochastic integro-differential Ito equations are obtained. For deriving the equations of the statistical dynamics the approximation kernels method by the functions satisfying some ordinary differential equations is used. As a result by means of extending the state vector the initial stochastic integro-differential equation is reduced to the corresponding stochastic Ito differential equation to which the equations of the statistical dynamics are known and the software is provided. Some examples are given.

1. Introduction

Many problems of the theory of hereditary control systems (for instance, the systems which contain the elements described by hereditary mechanics [1]) under the conditions of random disturbances lead to the study of the following stochastic integro-differential system (SIDS):

$$\dot{X}=a(X,t)+\int_{t_0}^{t} a_1(X(\tau),\tau,t)d\tau+$$

$$+\left[b(X,t)+\int_{t_0}^{t} b_1(X(\tau),\tau,t)d\tau\right]\dot{W},\qquad X(t_0)=X_0\,, \tag{1.1}$$

which is considered in Ito sense. Here $X\in R^p$ is the state vector; $\dot{W}\in R^q$ is the vector white noise of the intensity $\nu=\nu(t)$ independent of the initial condition X_0; $W=W(t)$ is the random second-order process whith independent increments (non-obligatory Wiener process);

$$a=a(X,t), a: R^p\times R\to R^p;\qquad a_1=a_1(X(\tau),\tau,t), a_1: R^p\times R\times R\to R^p;$$

$$b=b(X,t), b: R^p\times R\to R^{pq};\qquad b_1=b_1(X(\tau),\tau,t), b_1: R^p\times R\times R\to R^{pq}\,.$$

In technical applications the functions a_1 and b_1 are represented in the form

$$a_1=A(t,\tau)\varphi(X(\tau),\tau)\,,\qquad b_1=\sum_{h=1}^{N} B_h(t,\tau)\psi_h(X(\tau),\tau)\,, \tag{1.2}$$

where hereditary kernels $A(t,\tau)$ and $B_h(t,\tau)$ are $(p\times p)$- and $(p\times q)$-matrix functions, $\psi_h(X(\tau),\tau)$ is the $(q\times q)$-matrix function, $\varphi(X(\tau),\tau)$ is the p-dimensional vector function.

Let us suppose that the hereditary kernels $A(t,\tau)=\{A_{ij}(t,\tau)\}$, $i,j=\overline{1,p}$ and $B_h(t,\tau)=\{B_{hij}(t,\tau)\}$ $h=\overline{1,N}$, $i=\overline{1,p}$, $j=\overline{1,q}$ satisfy the conditions of the non-anticipativeness and the conditions of dying hereditary (or asymptotical stability)

$$A_{ij}(t,\tau)=0\,, \qquad B_{hij}(t,\tau)=0\,, \qquad \forall\tau>t\,, \tag{1.3}$$

$$\int_{-\infty}^{\infty} |\,A_{ij}(t,\tau)\,|\,d\tau<\infty\,, \qquad \int_{-\infty}^{\infty} |\,B_{hij}(t,\tau)\,|\,d\tau<\infty\,. \tag{1.4}$$

In the case where the hereditary kernels satisfy the conditions

$$A_{ij}(t,\tau)=\tilde{A}_{ij}(\xi)\,, \qquad B_{hij}(t,\tau)=\tilde{B}_{hij}(\xi)\,, \qquad \xi=t-\tau\,, \tag{1.5}$$

one say about the invariance or the stationarity of the hereditary. The important class of hereditary kernels are singular kernels

$$A_{ij}(t,\tau)=A_{ij}^{+}(t)A_{ij}^{-}(\tau)\,, \qquad B_{hij}(t,\tau)=B_{hij}^{+}(t)B_{hij}^{-}(\tau)\,. \tag{1.6}$$

The simplest examples of the functions satisfying conditions (1.3)–(1.6) are the functions $\exp(-\alpha\,|\,\xi\,|)\mathbf{1}(\xi)$, $\exp(-\alpha\,|\,\xi\,|)\,[\cos\omega\xi+\gamma\sin\omega\,|\,\xi\,|]\mathbf{1}(\xi)$, $\mathbf{1}(\xi)$ is the unit step function.

One of the most effective ways of obtaining the statistical dynamics equations for SIDS is their reduction to stochastic differential equations by means of the approximation of the real hereditary kernels by the functions satisfying some ordinary differential equations. This paper is devoted to obtaining the equations of statistical dynamics for such SIDS.

2. Problem statement

Let us consider a non-linear SIDS (1.1) under condition (1.2)

$$\dot{X}=a(X,t)+\int_{t_0}^{t} A(t,\tau)\varphi(X(\tau),\tau)d\tau+$$

$$+[b(X,t)+\sum_{h=1}^{N}\int_{t_0}^{t} B_h(t,\tau)\psi_h(X(\tau),\tau)d\tau]\,\dot{W}, \quad X(t_0)=X_0. \tag{2.1}$$

We shall determine a class of admissible kernels by conditions (1.3), (1.4) and by the order and the form of linear ordinary equations to which the kernels as the functions of t and τ must satisfy. We shall consider that the one-dimensional characteristic function (c.f.) $h_1(\mu;t)$ of the process $W$is a differentiable time function and that its logarithmic derivative $\chi(\mu;t)=(\partial/\partial t)\ln h_1(\mu;t)$ exists. We denote by $g_0(\lambda_1)=M\exp(i\lambda_1^T X_0)$, $\lambda_1=[\lambda_{11}\ \ldots\ \lambda_{1p}]^T$ c.f. of X_0.

Let a random process $X_t=X(t)$ represent a strong solution of (2.1). The problem of obtaining the general equations of statistical dynamics i.e. the equations for the finite-dimensional c.f. of X_t under the condition that the c.f. of X_0 and the one-dimentional c.f. of $W$are known is stated. We shall give also the generalization for the case where $W$is the random second-order process with independent increments and with random jumps at some fixed time moments $\{\tau_k\}$, $k=1, 2, \ldots$.

3. Main results

Let the kernels $A_{ij}(t, \tau)$ and $B_{hij}(t, \tau)$ in (2.1) satisfy (1.3), (1.4) and at fixed τ be the solutions of linear differential equations

$$\sum_{i=1}^{p} F^{t}_{ri} A_{ij}(t, \tau) = H^{t}_{rj} \delta(t-\tau) ; \qquad r, j = \overline{1, p} ,$$

$$\sum_{i=1}^{p} \Phi^{t}_{hri} B_{hij}(t, \tau) = \Psi^{t}_{hrj} \delta(t-\tau) , \qquad r = \overline{1, p} , \quad j = \overline{1, q} , \tag{3.1}$$

at a fixed t the kernels are determined by the formulae

$$A_{lj}(t, \tau) = \sum_{s=1}^{p} H^{*\vartheta}_{sj} A'_{ls}(t, \tau) , \qquad l, j = \overline{1, p} ,$$

$$B_{hij}(t, \tau) = \sum_{s=1}^{p} \Psi^{*\tau}_{hsj} B'_{hls}(t, \tau) , \qquad l = \overline{1, p} , \quad j = \overline{1, q} , \tag{3.2}$$

and the functions $A'_{ls}(t, \tau)$ and $B'_{hls}(t, \tau)$ entering into (3.2) at a fixed t are the solutions of linear differential equations

$$\sum_{s=1}^{p} F^{*\tau}_{sr} A'_{ls}(t, \tau) = \delta_{rl} \delta(t-\tau) , \qquad l, r = \overline{1, p} ,$$

$$\sum_{s=1}^{p} \Phi^{*\tau}_{hsr} B'_{hls}(t, \tau) = \delta_{rl} \delta(t-\tau) , \qquad r = \overline{1, p} , \quad l = \overline{1, q} . \tag{3.3}$$

Here $F_{rl} = F_{rl}(t, D)$, $H_{rl} = H_{rl}(t, D)$, $\Phi_{hrl} = \Phi_{hrl}(t, D)$, $\Psi_{hrl} = \Psi_{hrl}(t, D)$ are some linear differential operators, the order of the operators H_{rl} and Ψ_{hrl} denoted by m, the order of F_{rl} and Ψ_{hrl} is denoted by n, $n \geqq m$, the superscript t at the operator indicates that the operator acts on the function considered as the function of t at a fixed τ, by asterisk the adjoint operators are marked, δ_{rl} is the Kronecker symbol, $\delta(t-\tau)$ is the Dirac delta-function.

As it is known from the linear differential systems theory [2] the integral terms which enter into (2.1)

$$Y_i = \int_{t_0}^{t} \sum_{k=1}^{p} A_{ik}(t, \tau) \varphi_k(X(\tau), \tau) d\tau , \qquad i = \overline{1, p} ,$$

$$U_{hij}=\int_{t_0}^{t}\sum_{k=1}^{a} B_{hik}(t,\tau)\psi_{hkj}(X(\tau),\tau)d\tau\,,\qquad i=\overline{1,p}\,,\quad j=\overline{1,q}\,,\tag{3.4}$$

represent the solutions of the following equations:

$$\begin{aligned}\sum_{k=1}^{p} F_{rk}Y_k&=\sum_{k=1}^{p} H_{rk}\varphi_k\,, && r=\overline{1,p}\,,\\ \sum_{l=1}^{q}\sum_{k=1}^{p}\Phi_{hrk}U_{hkl}&=\sum_{l=1}^{q}\sum_{k=1}^{q}\Psi_{hrk}\psi_{hkl}\,, && r=\overline{1,p}\,.\end{aligned}\tag{3.5}$$

if their kernels satisfy (3.1)–(3.3).

It is clear that here the non-linear functions $\varphi_k=\varphi_k(X,t)$ and $\psi_{hkl}(X,t)$ must be differentiable not less than m times.

It is evident that if we introduce the vectors Z' and ε' of dimension $p(1+qN)\times 1$ and $(p+q^2N)\times 1$ and the vectors U and ε of dimension $pqN\times 1$ and $q^2N\times 1$ assuming

$$\begin{gathered}Z'=[Y^TU^T]^T\,,\qquad Y=[Y_1\,\ldots\,Y_p]^T,\qquad U=[U_1^T\,\ldots\,U_N^T]^T\,,\\ U_h=[U_{h1}^T\,\ldots\,U_{hp}^T]^T\,,\quad h=\overline{1,N}\,,\qquad \varepsilon'=[\varphi^T\varepsilon^T]^T\,,\\ \varphi=[\varphi_1\,\ldots\,\varphi_p]^T\,,\qquad \varepsilon=[\psi_1^T\,\ldots\,\psi_N^T]^T\,.\end{gathered}\tag{3.6}$$

Then the scalar equations (3.5) may be written in the form

$$F'Z'=H'\varepsilon'\,.\tag{3.7}$$

Here the matrix differential operators F' and H' are

$$\begin{aligned}F'&=\{F'_{rl}(t,D)\}=\sum_{k=0}^{n}\alpha_kD^k\,, && r,l=\overline{1,p(1+qN)}\,,\\ H'&=\{H'_{rl}(t,D)\}=\sum_{k=0}^{m}\beta_kD^k\,, && r,l=\overline{1,(p+q^2N)}\,,\end{aligned}\tag{3.8}$$

where $\alpha_k=\alpha_k(t)$ and $\beta_k=\beta_k(t)$ are the $(p(1+qN)\times p(1+qN))$ and $(p+q^2N)\times(p+q^2N)$ matrix coefficients. Further we shall replace (3.7) by the set of equations in Cauchy form according to the known rules of [2]:

$$\begin{aligned}\dot Z''_k&=Z''_{k+1}+q_k\varepsilon'\,, && k=\overline{1,n-1}\,,\\ \dot Z''_n&=-\sum_{l=1}^{n}\alpha_n^{-1}\alpha_{l-1}Z''_l+q_n\varepsilon'\,.\end{aligned}\tag{3.9}$$

Here Z''_h are the $p(1+qN)$-dimensional vectors, $h=\overline{1,n}$, $q_k=q_k(t)-p(1+qN)\times q^2N$ matrices determined by the formulae

$$q_k=\alpha_n^{-1}\left[\beta_{n-1}-\sum_{h=0}^{k-1}\sum_{l=0}^{k-h}C_{n-k-l}^{n-k}\alpha_{n-k+h+l}q_h^{(l)}\right],\qquad k=\overline{1,n-1}\,,$$

$$q_n = \alpha_n^{-1}\left[\beta_0 - \sum_{h=0}^{n-1}\sum_{l=0}^{n-h}\alpha_{h+l}q_h^{(l)}\right]. \tag{3.10}$$

Here the dependence between Z' and ε' will exist

$$Z' = Z_1'' + q_0\varepsilon', \qquad q_0 = \alpha_n^{-1}\beta_n \tag{3.11}$$

By virtue of (3.6) and (3.11), integral terms (3.4) take the form

$$\begin{gathered} Y = \int_{t_0}^{t} A(t,\tau)\varphi(X(\tau),\tau)d\tau = a'(Z_1'' + q_0\varepsilon'), \\ \sum_{h=1}^{N} U_h = \sum_{h=1}^{N}\int_{t_0}^{t} B_h(t,\tau)\psi_h(X(\tau),\tau)d\tau = \left(\sum_{h=1}^{N} b_h'\right)(Z_1'' + q_0\varepsilon'), \end{gathered} \tag{3.12}$$

where the following notations are introduced:

$$a' = [I_p 0], \qquad b_h' = \Lambda_h[0 I_{pqN}], \qquad \Lambda_h = \left[0 \ldots \underset{n}{1} \ldots 0\right], \tag{3.13}$$

and I_p is the $(p \times p)$ unit matrix. Consequently, SIDS (2.1) together with equations (3.9), (3.12) for the extended state vector of the dimension $np(2+qN)$

$$Z = [X^T Z^T]^T, \; Z'' = [Z_1''^T \ldots Z_n''^T]^T$$

is reduced to a SIDS of the form

$$\dot{Z} = c(Z,t) + l(Z,t)\dot{W}, \qquad Z(t_0) = Z_0, \tag{3.14}$$

where

$$c(Z,t) = \begin{bmatrix} a(X,t) + a'[Z_1'' + q_0(t)\varepsilon'(X,t)] \\ \gamma(t)Z'' + \gamma'(t)\varepsilon'(X,t) \end{bmatrix},$$

$$l(Z,t) = \begin{bmatrix} b(X,t) + \left(\sum_{h=1}^{N} b_h'\right)[Z_1'' + q_0(t)\varepsilon'(X,t)] \\ 0 \end{bmatrix},$$

$$Z_0 = [X_0^T 0]^T, \qquad \gamma'(t) = \operatorname{diag}[q_1(t) \ldots q_n(t)],$$

$$\gamma(t) = \begin{bmatrix} I & 0 & \ldots & 0 \\ 0 & I & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & I \\ -\alpha_n^{-1}\alpha_0 & -\alpha_n^{-1}\alpha_1 & \ldots & -\alpha_n^{-1}\alpha_{n-1} \end{bmatrix}, \quad I = I_{p(1+qN)}. \tag{3.15}$$

The matrices a', b'_h which enter into (3.15) are determined by (3.13), q_l, $l=\overline{0,n}$ are determined by (3.10), (3.11), ε' is determined by (3.6). The dimension of $c(Z, t)$ is equal to $np(2+qN)$ and the dimension of $l(Z, t)$ is $np(2+qN)\times q$. Equations (3.14) are linear in Z'' and non-linear in X.

It is shown in [3, 4] that if W in (2.1) represents a continuous second-order process with a differentiable one-dimensional c.f., and the functions $c(Z, t)$ and $l(Z, t)$ satisfy the conditions

$$M\mid c(Z,t)\mid^2<k_1+k_2M\mid Z\mid^2,$$

$$M\mid c(Z',t)-c(Z,t)\mid^2<k_3M\mid Z'-Z\mid^2,$$

$$M\operatorname{tr} l(Z,t)v(t)l(Z,t)^T<k_4+k_5M\mid Z\mid^2,$$

$$M\operatorname{tr}[l(Z',t)-l(Z,t)]v(t)[l(Z',t)^T-l(Z,t)^T]<k_6M\mid Z'-Z\mid^2. \qquad (3.16)$$

for some $k_1, \ldots, k_6$ and any random variables Z and Z' with finite second-order moments then the finite-dimentional c.f.

$$g_n(\lambda_1, \ldots, \lambda_n, t_1, \ldots, t_n)=M\exp\left[i\sum_{k=1}^{n}\lambda_k^TZ_{t_k}\right] \qquad (n=1, 2, \ldots)$$

of Z_t which represents the unique strong solution of (3.14) are determined by the Pugachev equation for c.f. [3, 4]:

$$(\partial/\partial t_n)g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n)=$$

$$=M\{i\lambda_n^Tc(Z_{t_n}, t_n)+\chi(l(Z_{t_n}, t_n)^T\lambda_n; t_n)\}\times$$

$$\times\exp\left[i\sum_{k=1}^{n}\lambda_k^TZ_{t_k}\right] \qquad (n=1, 2, \ldots) \qquad (3.17)$$

with the initial conditions

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_{n-1})=g_{n-1}(\lambda_1, \ldots, \lambda_{n-1}+\lambda_n; t_1, \ldots, t_{n-1}),$$

$$g_1(\lambda_1, t_0)=g_0(\lambda_1). \qquad (3.18)$$

The methods for solving these equations are considered in [4]. The finite-dimensional c.f. practically determine the probabilities of all the events with which we have to deal in technical problems.

Thus we state the following result.

Theorem 1. Let in non-linear SIDS (2.1) Wrepresent a process with independent increments for which the logarithmic derivative $\chi(\mu; t)$ of the one-dimensional c.f. $h_1(\mu; t)$ exists, the initial value X_0 is a random variable independent of $W$with finite second-order moment and the c.f. $g_0(\lambda_1)$, the hereditary kernels $A_h(t,\tau)$ and $B_h(t,\tau)$ satisfy the conditions of the non-anticipativeness (1.3), of asymptotical stability (1.4) and equations (3.1)–(3.3) also, the non-linear functions φ and ψ_h are differentiable m times. Then SIDS (2.1) by means of extending the state vector up to dimension $np(2+qN)$ may be reduced to SDS (3.14). If conditions (3.16) are fulfilled, then the finite-dimensional c.f. of the strong solution of (2.1) satisfies equations (3.17), (3.18).

Let us consider now the class of kernels satisfying the conditions of Theorem 1 and the condition of stationarity (1.5). As it is known from the theory of linear stationary systems [2, 4] in this case it is sufficient to suppose instead of (3.1)–(3.3) that the Laplace transforms of hereditary kernels $\tilde{A}(\xi)$ and $\tilde{B}_h(\xi)$ are rational functions of the scalar variable s, i.e. admit the representation of the form

$$\int_0^\infty \tilde{A}(\xi)e^{-s\xi}d\xi = F(s)^{-1}H(s)\,, \qquad \int_0^\infty \tilde{B}_h(\xi)e^{-s\xi}d\xi = \Phi_h(s)^{-1}\Psi_h(s)\,. \tag{3.19}$$

Here the order of the matrix polynomials $H(s)=\{H_{rl}(s)\}$ and $\Psi_h(s)=\{\Psi_{hrl}(s)\}$ is equal to m and the order of the polynomials $F(s)=\{F_{rl}(s)\}$ and $\Phi_h(s)=\{\Phi_{hrl}(s)\}$ is equal to n, $n\geqq m$. Then the following theorem is valid.

Theorem 2. Let in non-linear SIDS (2.1) W, X_0, φ, ψ_h satisfy the conditions of Theorem 1, the hereditary kernels satisfy conditions (1.3)–(1.5) and (3.19). Then SIDS (2.1) by means of extending the state vector up to dimension $np(2+qN)$ may be reduced to SDS (3.14).

If conditions (3.16) are fulfilled then the finite-dimensional c.f. of the strong solution of (2.1) satisfies equations (3.17), (3.18).

In conclusion we consider the class of the hereditary kernels satisfying the conditions of Theorem 1 and the singularity condition (1.6). In such a case the order of the operators F_{ri} and Φ_{hri} is equal to one, $n=1$, the order of the operators H_{rj} and Ψ_{hrj} is equal to 0, $m=0$. In this special case equations (3.4), (3.5) take the following form:

$$\int_{t_0}^{t} A(t,\tau)\varphi(X(\tau),\tau)d\tau = A^+Y; \quad Y=[Y_1\ \dots\ Y_p]^T\,,$$

$$\int_{t_0}^{t} B_h(t,\tau)\psi_h(X(\tau),\tau)d\tau = B_h^+U_h\,, \quad U_h=[U_{h1}^T\ \dots\ U_{hq}^T]^T\,, \tag{3.20}$$

$$\dot{Y}=A^-Y\,, \quad Y(t_0)=0\,, \quad \dot{U}_h=B_h^-U_h\,, \quad U_h(t_0)=0\,. \tag{3.21}$$

Here $A^\pm=\{A_{ij}^\pm(t)\}$ and $B_h^\pm=\{B_{hij}^\pm(t)\}$ are $(p\times p)$- and $(p\times q)$-matrices, Y and $U_{hi}=[U_{hi}, \dots U_{hip}]^T$ are the p-dimensional vectors. As a result SIDS (2.1) is reduced to SDS (3.14) for the extended $p(2+qN)$-dimensional state vector $Z=[X^TY^TU^T]^T$ if we introduce (pqN)-dimensional vector $U=[U_1^T\ \dots\ U_N^T]^T$, the $p(2+qN)$-dimensional vector function $c(Z,t)$, the $(p(2+qN)\times q)$-dimensional matrix-function $l(Z,t)$ and the initial $p(2+qN)$-dimensional vector Z_0 assuming

$$c(Z,t)=\begin{bmatrix} a(X,t)+A^+Y \\ A^-Y \\ B^-U \end{bmatrix}, \quad l(Z,t)=\begin{bmatrix} l(X,t)+B^+U \\ 0 \\ 0 \end{bmatrix},$$

$$B^\pm=[B_1^\pm \ \dots \ B_N^\pm], \qquad Z_0=[X_0^T \ \ 0 \ \ 0]^T. \tag{3.22}$$

The functions $c(Z,t)$ and $l(Z,t)$ are linear in Y and U and non-linear in X. We emphasize that in the case of singular kernels the differentiability of φ and ψ_h is not obligatory. Then the following theorem is valid.

Theorem 3. Let non-linear SIDS (2.1) W, X_0 satisfy the conditions of Theorem 1, the hereditary kernels satisfy conditions (1.3), (1.4), (1.6). Then SIDS (2.1) by means of extending the state vector up to the $p(2+qN)$-dimension may be reduced to SDS (3.14) under condition (3.22). If conditions (3.16) are fulfilled then the finite-dimensional c.f. of the strong solution of (2.1) satisfy equations (3.17), (3.18).

4. Generalizations

Let W in (2.1) be a random second-order process with independent increments and random jumps at the fixed time moments $\{\tau_k\}$ $(k=1,2,\dots)$. In this case $\nu(t)$ and $\chi(\mu;t)$ will contain the linear combinations of delta-functions $\delta(t-\tau_k)$ $(k=1,2,\dots)$. Then as it is shown in [3] if conditions (3.16) can be fulfilled for all time moments beside $\{\tau_k\}$, then the finite-dimensional c.f. of the strong solution of (2.1) will satisfy (3.17), (3.18). It is not difficult to formulate the statements which correspond to Theorems 1–3.

5. Examples

1. The one-dimensional linear SIDS

$$\dot{X}=a_0+a_1X+A\int_0^t e^{-\alpha(t-\tau)}X(\tau)d\tau+b\dot{W}, \quad X_0=0,$$

(a_0, a_1, A, b, α are constant) is reduced to the following linear two-dimensional SDS:

$$\dot{X}=a_0+a_1X+(A/\alpha)Y+b\dot{W}, \quad \dot{Y}=\alpha(X-Y), \quad X_0=Y_0=0.$$

The equations for the variances and the covariance of X and Y in the case of arbitrary $\dot{W}$ have the form

$$\dot{D}_x = 2[a_1 D_x + (A/\alpha)k_{xy}] + vb^2\,,$$
$$\dot{D}_y = 2\alpha(k_{xy} - D_y)\,,$$
$$\dot{k}_{xy} = \alpha D_x + (a_1 - \alpha)k_{xy} + (A/\alpha)D_y\,.$$

2. The p-dimensional linear SIDS (2.1) at

$$a = a_0 + a_1 X\,, \quad b = b_0\,, \quad \varphi(X, t) = X\,, \quad \psi_h = 0$$

and the hereditary kernels satisfying (3.1)–(3.3) are reduced to linear SDS (3.14) by means of extending the state vector up to the $2np$ dimension. In this case

$$c(Z, t) = c_0 + c_1 Z\,, \qquad l(Z, t) = l_0\,,$$
$$c_1 = \begin{bmatrix} a_1 + q_0 & a'' \\ \gamma' & \gamma \end{bmatrix}, \qquad c_0 = \begin{bmatrix} a_0 \\ 0 \end{bmatrix}, \qquad l_0 = \begin{bmatrix} b_0 \\ 0 \end{bmatrix},$$
$$a'' = [a'\,0\,\ldots\,0]\,.$$

If we denote the solution of the linear system $\dot{u} = c_1 u$ with the initial condition $u(\tau, \tau) = I$ by $u = u(t, \tau)$ then at $t_0 \leqq t_1 \leqq \ldots \leqq t_n$ the finite-dimensional c.f. is determined by the formula [3, 4]:

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n) =$$
$$= g_0\left(\sum_{k=1}^{n} u(t_k, t_0)\lambda_k^T\right) \exp\left\{i \sum_{k=1}^{n} \lambda_k^T \int_{t_0}^{t_k} u(t_k, \tau)c_0(\tau)d\tau + \right.$$
$$\left. + \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \chi(l_0(\tau)^T \sum_{r=k}^{n} u(t_r, \tau)^T \lambda_r; \tau)d\tau\right\} \qquad (n = 1, 2, \ldots)\,.$$

3. The linear SIDS of example 2 with singular hereditary kernel by means of extending the state vector up to the $2p$-dimension is reduced to linear SDS (3.14) at

$$c(Z, t) = c_0 + c_1 Z\,, \qquad l(Z. t) = l_0\,,$$
$$c_0 = \begin{bmatrix} a_0 \\ 0 \end{bmatrix}, \qquad c_1 = \begin{bmatrix} a_1 & \underline{A^+} \\ A & 0 \end{bmatrix}, \qquad l_0 = \begin{bmatrix} b_0 \\ 0 \end{bmatrix}.$$

4. The one-dimensional SIDS with a parametric noise

$$\dot{X} = a_0 + a_1 X + \left[b + B\int_0^t e^{-\beta(t-\tau)} X(\tau)d\tau\right]\dot{W}, \quad X_0 = 0\,,$$

(a_0, a_1, b, B, β are constant) is reduced to the following two-dimensinal SDS with the parametric noise:

$$\dot{X} = a_0 + a_1 X + [b + (B/\beta) Y]\dot{W},$$

$$\dot{Y} = \beta(X - Y), \qquad X_0 = Y_0 = 0.$$

The equations for variances and the covariance of X and Y in accordance with the general formulae [4] have the form

$$\dot{D}_x = 2a_1 D_x + \nu b^2 + (2/\beta) Bb\, \nu m_y + (m_y^2 + D_y)(1/\beta^2) B^2 \nu,$$

$$\dot{D}_y = 2\beta(k_{xy} - D_y)$$

$$\dot{k}_{xy} = (a_1 - \beta) k_{xy} + \beta D_x, \qquad m_y = MY.$$

5. The one-dimensional non-linear SIDS

$$\dot{X} = a_0 + a_1 X + \int_0^t e^{-\alpha(t-\tau)} \varphi(X(\tau)) d\tau + b\dot{W}, \quad X_0 = 0$$

(a_0, a_1, α, b are constant) is reduced to the following non-linear two-dimensional SDS:

$$\dot{X} = a_0 + a_1 X + (1/\alpha) Y + b\dot{W},$$

$$\dot{Y} = \alpha[\varphi(X) - Y], \qquad X_0 = Y_0 = 0.$$

Solving this system by normal approximation method [4] for normal $\dot{W}$ we come to the to the following set of the equations for the expectations, the variances and the covariance of X and Y:

$$\dot{m}_x = a_0 + a_1 m_x + (1/\alpha) m_y, \qquad \dot{m}_y = \alpha(\varphi_0 - m_y),$$

$$\dot{D}_x = 2[a_1 D_x + (1/\alpha) k_{xy}] + b^2 \nu,$$

$$\dot{D}_y = 2\alpha(\varphi_1 k_{xy} - D_y),$$

$$\dot{k}_{xy} = \alpha\varphi_1 D_x + (a_1 - \alpha) k_{xy} + (1/\alpha) D_y.$$

Here $\varphi_0 = \varphi_0(m_x, D_x)$, $\varphi_1 = (\partial/\partial m_x)\varphi_0$ are the coefficients of the statistical linearization of the non-linear function φ.

6. Conclusions

One of the most effective ways of obtaining the statistical dynamics equations for SIDS is their reduction to SDS for which a wide software is developed and first of all the most universal approximate method for non-linear multi-dimensional systems, i.e. the normal approximation method [4]. In this paper first the methods of the reduction of SIDS to SDS by means of the approximation of real hereditary kernels by the functions satisfying the linear ordinary differential equations of given order and also their combination with the singular kernels method are developed for linear and non-linear SIDS and white noises of various types. A special applied Fortran-4 package destined for automatic setting of the equations for the expectation and the covariance matrix of the normal approximation method for SIDS and solving these equations by Runge–Kutta method of the fourth order is developed by V. I. Shin and M. P. Terentjeva for the multi-dimensional SIDS (2.1) with the exponential hereditary kernels and the polynomial non-linearities.

References

1. *Corduneanu, C., Lakshmikantham, V.* Equations with Infinite Delays. Automation and Remote Control, 1985, **7**, 5–44 (in Russian).
2. *Pugachev, V. S.*, Theory of Random Functions and its Application to Control Problems, Pergamon Press, 1965.
3. *Pugachev, V. S.*, Finite-dimensional distributions of a random process determined by a stochastic differential equation and their application to control problems. Problems of Control and Information Theory, **10** (2), 95–114 (1981).
4. *Pugachev, V. S., Sinitsyn, I. N.*, Stochastic Differential Systems, "Nauka", Moscow, 1985.

Стохастические системы управления с памятью

И. Н. СИНИЦЫН

(Москва)

Ставится задача получения уравнений статистической динамики систем управления с памятью, описываемых линейными и нелинейными стохастическими интегро-дифференциальными уравнениями вида (2.1), понимаемые в смысле Ито, где $X = X(t)$ и $W = W(t)$ — случайные процессы со значениями в конечномерных пространствах R^p и R^q, соответственно

$$a = a(X, t), a; \quad R^p \times R \to R^p; \quad b = b(X, t), b : R^p \times R \to R^{pq},$$

$$A(t, \tau) = A, A : R \times R \to R^{pp}; \quad B_n(t, \tau) = B_h, B_h : R \times R \to R^{pq},$$

$$\varphi = \varphi(X, t), \varphi : R^p \times R \to R^p; \quad \psi_h = \psi_h(X, t), \psi_h : R^p \times R \to R^{qq}.$$

Предполагается, что W является произвольным процессом второго порядка с независимыми приращениями (необязательно винеровским).

Для вывода уравнений статистической динамики применяется приближенный метод аппроксимации ядер памяти A и B_h с функциями, удовлетворяющими некоторым обыкновенным дифференциальным уравнениям. В результате путем расширения вектора состояния (2.1) приводится к (3.14), для которого известны уравнения статистической динамики и создано математическое обеспечение. В связи с этим доказывается теорема 1. Пусть в (2.1) W представляет собой случайный процесс с независимыми приращениями, для которого существует логарифмическая производная $X(\mu; t)$ от одномерной характеристической функции $h_1((\mu; t)$, начальное значение X_0 является случайной величиной, не зависимой от W с конечными моментами второго порядка и имеет характеристическую функцию $g_0(\lambda_1)$, ядра памяти A и B_h удовлетворяют условиям физической реализуемости (1.3) асимптотической устойчивости (1.4), а также уравнениям (3.1)–(3.3) (нелинейные функции φ и ψ_h m раз дифференцируемы). Тогда (2.1) путем расширения вектора состояния до размерности $np(2+qN)$ может быть сведена к (3.14). При этом, если выполнены условия (3.16), то конечномерная характеристическая функция сильного решения (2.1) удовлетворяет уравнениям (3.17), (3.18). Теоремы 2 и 3 содержат уточнение теоремы 1 на случай стационарных и вырожденных ядер памяти A и B_h.

Теоремы 1–3 обобщаются, во-первых, на случай процесса W со случайными разрывами первого рода в фиксированных точках, во-вторых, на класс систем (4.1) и, в-третьих, на более общий класс вырожденных ядер (4.2).

Метод вырожденных ядер можно применять для систем с недифференцируемыми нелинейностями, в то время как метод аппроксимации ядер памяти функциями, удовлетворяющими дифференциальным уравнениям, лучше применять для систем с гладкими нелинейностями.

В разделе 5 разобраны 5 типичных систем. Отмечается, что для многомерных систем (2.1), экспоненциальных ядер памяти и полиномиальных нелинейностей разработан специальный пакет прикладных программ на языке фортран-4, предназначенный для автоматического составления уравнений для математического ожидания и ковариационной матрицы метода нормальной аппроксимации для (2.1) и решения этих уравнений методом Рунге-Кутта четвертого порядка.

И. Н. Синицын
Институт проблем информатики АН СССР,
Москва, 117900 ГСП-1, ул. Вавилова, 30.

Problems of Control and Information Theory, Vol. 15 (4), pp. 299—307(1986)

A NOTE ON NUMERICAL ASPECTS OF ROBUST TESTING

J. Á. Víšek

(*Prague*)

(Received September 5, 1985)

Two methods of determination of critical region of robust tests are discussed and references are given. One of them, namely the direct approximation of corresponding quantiles, is illustrated by a numerical example.

AMS subject classification: 62F35, 62F03
Keywords: Robust test, approximation of quantiles.

Introduction and discussion

This note is intended to pay attention to numerical aspects of robust tests and the present author would like to share some experiences from this field with others.

It is well known that for some models of contaminacy, and let us stress that from the practical point of view rather satisfactorily constructed models of contaminacy, the problem of finding the least favourable pair (LFP) was solved. Due to a definition of the LFP and due to the Neyman–Pearson lemma we know that a test based on the likelihood ratio (or, if possible, on the log-likelihood ratio) of LFP is the most powerful one for the corresponding testing problem. But as it was already emphasized at Víšek (1983), to establish a (robust) test it is necessary not only to give a statistic on which the test is to be based, but also to declare its critical region. It is true that in classical cases of testing statistical hypotheses solution was given only by means of the Neyman–Pearson lemma and numerical aspects of determination of a critical region were not usually discussed since it was understood without saying that the distribution functions (*df*) of the corresponding statistics were not difficult to find. Since the statistics are in many cases the sums of log-likelihood ratios computed at the observed points, the *df*'s are convolutions of given "parent" distributions. But for robust statistics the "parent" *df* has only exceptionally a form admitting to evaluate the convolution analytically. A typical form of this *df* is described by means of a partition of real line so that in every member of this partition the *df* possesses a different analytical form (from the form in another member of partition). Since on the other hand the numerical convolution, even for not very large sample sizes, is slow and expensive, we have to try to establish the

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

critical region in another way. Letting aside the possibility of normal approximation which, we shall see below, is not tight, there are at least two straightforward ways how to cope with it. Firstly we may try to find an approximation of convolution of parent *df* by means of known methods or secondly to construct a direct approximation of critical region by means of approximation of quantiles of corresponding *df*. Both possibilities will now be briefly discussed.

Accepting the first way one may use an Edgeworth or saddle-point expansion with rather good results (see Víšek (1983)). Another possibility how to approximate *df* of the log-likelihood ratio of LFP is to utilize the small-sample asymptotic method—for the theory, see Field and Hampel (1982), for the preciseness, see Tables 1 and 2 below. The latter way is better since it is, at least, more rapid than the former ones and this advantage may occur very important in connection with the following considerations.

Recently, some studies of sensitivity of the error probabilities of robust tests with respect to a change of contamination level were performed and characteristics of this sensitivity were proposed (see Víšek (1984a, b and c)). The problem is as follows. Having estimated the level of contaminacy, one establishes a robust test Φ_R in a framework of corresponding model of contaminacy. Since the estimation of level of contaminacy is not a highly reliable procedure, one may want to create an idea about behaviour of this robust test Φ_R under assumption that the "true" level of contaminacy is a little bit different from the initially assumed one. Then one may either evaluate the least favourable *df*, if any, for the test Φ_R in the changed models of contaminacy and so directly find the worst possible values of error probabilities of the test in changed models, or to compute the above-mentioned characteristics of sensitivity. For both cases the small-sample asymptotic method is excellent since for the former it is necessary to compute *df* of Φ_R in many models (see Víšek (1984c)) and for the latter, one needs a rather dense table for *df* of the corresponding statistics (which the Φ_R is based on), not only for given sample size n, but also for $n-1$. (Computer program for all above-mentioned evaluation is available from the author.)

On the other hand, as it was shown in Michálek, Vajda, Víšek (1984) and Víšek (1983) it is more convenient, even when we have at hand a very dense table of *df*, instead of the sensitivity characteristic to utilize some kind of their asymptotic approximations, which are of suitable preciseness in the characterization of error probability sensitivity and are quickly and easily available. And then we are only a step from the idea to avoid completely the evaluation of approximation of *df* of log-likelihood ratio of LFP and to try to produce directly an approximation of the critical region, i.e. we would like to find an approximation of quantiles of the mentioned *df*. A straightforward way how to do it is to take into account an expansion of this *df* and to make an inversion. It is equvalent to looking for a point at which the considered expansion is equal to one minus the size of test.

Let us consider for instance an Edgeworth expansion of *df*. Since its first term is equal to the normal *df* one should start with the normal quantile and try to make a correction of it to arriwe at the desired quantile (compare Hall (1983)). For a normed statistic T_n (i.e. with mean equal to zero and variance equal to one) which is of the sum type we may derive e.g. first and second order approximation of quantile t_α in the form

$$t_\alpha^{(1)} = u_\alpha + n^{-\frac{1}{2}} \frac{\lambda_3}{6} (u_\alpha^2 - 1) \tag{1}$$

and

$$t_\alpha^{(2)} = t_\alpha^{(1)} + n^{-1} \left\{ \frac{\lambda_3^2}{72} (9u_\alpha - 2u_\alpha^3 - u_\alpha^5) + \frac{\lambda_4}{24} (u_\alpha^3 - 3u_\alpha) \right\} \tag{2}$$

where λ_3 and λ_4 are the corresponding third and fourth cumulants, u_α is the standard normal upper α-quantile and then

$$F_n(T_n > t_\alpha^{(i)}) = \alpha + o(n^{-\frac{i}{2}}) .$$

(The derivation of (1) and (2) is simple, hence it is omitted.)

Numerical illustration

Approximation of the above mentioned *df* and of its quantiles was studied for some robust tests which were constructed in a framework of the general contamination neighbourhoods (Rieder's) model of contaminacy (see Rieder (1977)). Let us recall at first this model. Let $\mathfrak{M}$ be the space of all probability measures define on a measurable space $(\mathscr{X}, \mathscr{A})$, $P_0 \in \mathfrak{M}$, $P_1 \in \mathfrak{M}$, $P_0 \neq P_1$, ε_i and δ_i real numbers for $i = 0, 1$ satisfying $0 \leqq \varepsilon_i$, $0 \leqq \delta_i$, $0 < \varepsilon_i + \delta_i < 1$ and put

$$\mathscr{P}_i = \{Q \in \mathfrak{M} : Q(A) \leqq (1 - \varepsilon_i) P_i(A) + \varepsilon_i + \delta_i , \ \forall A \in \mathscr{A}\} .$$

Then under assumption that $\mathscr{P}_0 \cap \mathscr{P}_1 = \emptyset$ there is a LFP$(\mathscr{P}_0, \mathscr{P}_1) = (Q_0, Q_1)$ defined by

$$Q_0(\pi(x) > t) = \sup \{Q'(\pi(x) > t) : Q' \in \mathscr{P}_0\}$$

and

$$Q_1(\pi(x) > t) = \inf \{Q''(\pi(x) > t) : Q'' \in \mathscr{P}_1\} ,$$

where $\pi(x) \in dQ_1/dQ_0$ (see again Rieder (1977)). The test Φ_R^n based on $\prod_{i=1}^{n} \pi(x_i)$ is the most powerful test for testing $H_n : w_n \in \mathscr{P}_0^{(n)}$ against $A_n : w_n \in \mathscr{P}_1^{(n)}$, where w_n denoted a probability measure which generated sample $x_1, \ldots, x_n$ and $\mathscr{P}_i^{(n)} = \left\{ \prod_{j=1}^{n} Q_j : Q_j \in \mathscr{P}_i \right\}$.

In this numerical study such tests were constructed for normal, log-normal, Weibull, double-exponential (Laplace), gamma and approximately Rayleigh (Rice) model (playing the role of P_0 and P_1 — see the description of the tables below) for parameters of contaminacy being equal to $\varepsilon_0 = \varepsilon_1 = .050$ and $\delta_0 = \delta_1 = .025$ (for discussion of these values, see Víšek (1983)). The fraction of results (obtained by means of a special program available also from the author) were arranged into the following tables. Let us briefly explain their content.

At the first two tables an approximation of *df* of the above-mentioned robust test for normal and gamma model are compared with values obtained in simulations. More precisely, in the first table for normal model (i.e. in the above introduced notation $P_0 = N(0, 1)$ and $P_1 = N(1, 1)$) the values of approximation of the probability $Q_0^{\otimes m}\left(\sum_{i=1}^{n} \log \pi(x_i) > 0\right)$ (evaluated by means of small sample asymptotic technique) are compared with simulated values for a few sample sizes. (Sample size n is displayed in the first row, the approximations in the second row and finally the simulated values in the third row.) The second table offers the analogous comparison for gamma model

$$(\text{i.e. } f(x) = a^p \Gamma_{(p)}^{-1} \exp\{-ax\} x^{p-1}, \quad P_0 : a=1, p=3; \quad P_1 : a=2, \quad p=3)$$

but instead of

$$Q_0^{\otimes n}\left(\sum_{i=1}^{n} \log \pi(x_i) > 0\right)$$

the sum of error probabilities

$$Q_0^{\otimes n}\left(\sum_{i=1}^{n} \log \pi(x_i) > 0\right) + Q_1^{\otimes n}\left(\sum_{i=1}^{n} \log \pi(x_i) < 0\right)$$

were considered.

The next tables gather the above-mentioned approximations of quantiles and their precise values for a few probability models and sample sizes 20, 30 and 40. The probability model is specified before every triple of further tables and the sample size n is presented at the head of it. (Parameters ε and δ had been given earlier.) The sense of all columns of tables is indicated by a head of the third table.

Remark. The probabilistic models for the present study were chosen to built up an idea of preciseness of quantile approximation for some of the most frequently used ones. The exception is the last one of them. The reason for including it was its practical importance for processing radar data (for more details see Michálek, Vajda, Víšek (1984)).

NORMAL MODEL

5	10	15	20	25	30	35	40	45	50
.25522	.17654	.12828	.09535	.07188	.05472	.04196	.03236	.02506	.01948
.25692	.17025	.11846	.09128	.07026	.05539	.04103	.03231	.02051	.01641

GAMMA MODEL

5	10	15	20	25	30	35	40	45	50
.42051	.25392	.17330	.10789	.07246	.04921	.03370	.02322	.01608	.01119
.41650	.22900	.15200	.10100	.06950	.04150	.02850	.01850	.01500	.00500

NORMAL MODEL

$$f(x)=(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$P_0:\mu=0\,,\ \sigma^2=1$ $P_1:\mu=1\,,\ \sigma^2=1$

$n=20$

α-size of test	precise value of t_α	normal approximation of t_α	$t_\alpha^{(1)}$	$t_\alpha^{(2)}$
.197	−.268	−.269	−.272	−.267
.109	−.045	−.054	−.048	−.043
.054	.179	.161	.178	.179
.011	.581	.548	.592	.577
		$n=30$		
.197	−.438	−.438	−.441	−.437
.110	−.219	−.225	−.221	−.218
.055	0	−.014	−.001	0
.010	.438	.408	.446	.435
		$n=40$		
.191	−.569	−.569	−.571	−.569
.096	−.316	−.322	−.318	−.316
.051	−.126	−.137	−.127	−.126
.011	.253	.227	.258	.251

LOGNORMAL MODEL

$$f(x)=(2\pi)^{-\frac{1}{2}}b^{-1}(x-a)^{-1}\exp\left\{-\frac{[\log(x-a)-m]^2}{2b^2}\right\}$$

$P_0 : a=0\,,\ b=1\,,\ m=0$ $P_1 : a=0\,,\ b=1\,,\ m=1$

$n=20$

.199	−.224	−.230	−.228	−.224
.097	.045	.044	.041	.044
.053	.224	.230	.222	.224
.010	.626	.661	.640	.629

$n=30$

.195	−.383	−.388	−.387	−.384
.093	−.110	−.109	.112	−.110
.055	0	.003	−.002	0
.012	.438	.463	.447	.440

$n=40$

.187	−.506	−.509	−.508	−.506
.094	−.253	−.252	−.255	−.253
.051	−.063	−.058	−.064	−.063
.011	.316	.337	.323	.317

WEIBULL MODEL

$$f(x)=cpx^{p-1}\exp\{-cx^p\}$$

$P_0 : c=2,\ p=3$ $P_1 : c=1\,,\ p=3$

$n=20$

.196	.045	.046	.043	.045
.098	.179	.170	.177	.179
.058	.268	.252	.268	.269

$n=30$

.196	0	.001	−.001	0
.112	.110	.104	.108	.110
.058	.219	.206	.219	.219
012	.438	.406	.442	.437

$n=40$

.220	−.063	−.061	−.064	−.063
.013	.379	.352	.382	.378
.004	.506	.468	.512	.504
.001	.632	.582	.643	.628

DOUBLE-EXPONENTIAL (LAPLACE) MODEL

$$f(x)=(2b)^{-1}\exp\left\{-\frac{|x-a|}{b}\right\}$$

$P_0: a=2,\ b=1$ $P_1: a=0,\ b=1$

$n=20$

.206	−1.88	−1.87	−1.89	−1.87
.101	−1.39	−1.42	−1.39	−1.38
.050	−.984	−1.05	−.984	−.981
.010	−.224	−.378	−.198	−.261

$n=30$

.202	−2.46	−2.46	−2.47	−2.46
.097	−1.97	−2.00	−1.98	−1.96
.049	−1.59	−1.65	−1.59	−1.59
.011	−.876	−1.00	−.860	−.899

$n=40$

.201	−2.97	−2.97	−2.98	−2.97
.104	−2.53	−2.55	−2.53	−2.52
.047	−2.08	−2.14	−2.09	−2.09
.010	−1.39	−1.50	−1.38	−1.41

GAMMA MODEL

$$f(x)=a^{p}\Gamma^{(1)}_{(p)}\exp\{-ax\}x^{p-1}$$

$P_0: a=1,\ p=3$ $P_1: a=2,\ p=3$

$n=20$

.197	−.537	−.533	−.540	−.532
.096	−.224	−.244	−.227	−.220

.052	0	−.041	0	−.001
.010	.492	.397	.506	.478

$n = 30$

.192	−.769	−.764	−.769	−.763
.101	−.493	−.508	−.495	−.490
.048	−.219	−.255	−.219	−.219
.011	.219	.141	.229	.211

$n = 40$

.210	−1.01	−1.01	−1.01	−1.01
.101	−.696	−.708	−.697	−.694
.050	−.443	−.473	−.443	−.442
.011	0	−.068	.007	.006

APPROX. RAYLEIGH–RICE MODEL

$$f(x) = x \cdot \sigma^{-2} \{\exp 1 - (x^2 + \alpha^2)(2\sigma^2)^{-1}\} [1 + \alpha^2 x^2 (4\sigma^2)^{-1}]$$

$P_0 : \alpha = 0\,,\ \sigma^2 = 1$ $P_1 : \alpha = 1.5\,,\ \sigma^2 = 1$

$n = 20$

.206	−.358	−.348	−.360	−.352
.098	−.089	−.114	−.091	−.085
.048	.134	.073	.136	.134
.010	.537	.395	.551	.518

$n = 30$

.189	−.493	−.488	−.495	−.490
.101	−.274	−.293	−.275	−.271
.049	−.055	−.104	−.053	−.055
.011	.329	.214	.338	.317

$n = 40$

.187	−.632	−.628	−.634	−.630
.109	−.443	−.457	−.444	−.441
.047	−.190	−.235	−.189	−.190
.010	.190	.086	.197	.180

References

Hampel, F. R., Field, Ch. A. Small-sample asymptotic distribution of M-estimates of location, Biometrika **69** (1982), 29–46.

Hall, P., Inverting an Edgeworth expansion, Ann. Statist. **11**, (1983), 569–576.

Michálek, J., Vajda, J., Víšek, J. Á., New topics in robust statistics with applications, Transactions of COMPSTAT **84**, (1984), 73–83, Physica-Verlag, Wien.

Rieder, H., Least favourable pairs for special capacities. Ann. Statist. **5** (1977), 909–922

Víšek, J. Á., On second order efficiency of a robust test and approximations of its error probabilities, Kybernetika **19** (1983), 387–407.

Víšek, J. Á., On the dependence of the test error probabilities on the contamination level, (1984a), to appear in Statistics.

Víšek, J. Á., Influence of contamination level deviations on the test error probabilities, (1984b), submitted to Kybernetika.

Víšek, J. Á., Sensitivity of the test error probabilities with respect to the level of contamination in general model of contaminacy, (1984c), to appear in J. Statist. Plan. Inf.

Замечание о вычислительных проблемах робастных тестов

Й. А. ВИШЕК

(Прага)

В статье обсуждаются вычислительные проблемы, касающиеся построения робастных тестов, именно аппроксимации квантилов тестовых статистик. Нумерическая иллюстрация показывает хорошую точность такой аппроксимации для целого ряда вероятностных моделей.

J. Á. Víšek
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Praha 8 — Libeň
Czechoslovakia

Problems of Control and Information Theory, Vol. 15 (4), pp. 309—318 (1986)

ABOUT ONE CLASS OF GAMES WITH STABLE AND PARETO OPTIMAL SOLUTIONS

V. A. GORELIK

(*Moscow*)

(Received May 15, 1985)

A class of *n*-person games (in particular two-person games) with pay-off functions being transformations by the operation of taking minimum of two criteria one of which describes competition of players in some common sphere of activity and the other describes private achievements of the player is considered. It is shown that under some conditions of monotony of criteria in such games stable and Pareto optimal solutions exist.

1. Formulation of problem

Now theory of games has definite applications in analysis of conflict situations in different areas of human activity. However, expansion of game approach is not very broad and in any case does not answer those expectations which were born by game theory in initial period of its development. Such condition is determined by several reasons, but one of the most fundamental is connected with the principal internal problem of game theory, that is the lack of common principle of optimal behaviour. This means that the reasonable agreement from the point of view of one principle of optimality is not such from the point of view of the other principle of optimality. As a result there is an element of subjectivity in the investigation of real conflicts connected with the choice of conception of solution for their game models, which wrongs the value of mathematical approach. As a way out of such situation it is finding of such classes of game models which have just the same solutions for different principles of optimality. An important class of such models has been sûggested by professor Germeyer at the beginning of the 70s [1]. It serves for the descriptions of conflicts the participants of which have by the side with the private contradictory goals one common goal. Under some assumptions of monotony of goal functions for such conflicts there are compromises (solutions) which satisfy at once the set of principles of optimal behaviour (first of all equilibrium and Pareto optimum).

In connection with the investigation of such models there is a question: how important the presence of a common goal of conflict participants is. Is it possible to

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

have mutually advantageous and stable agreements in situations, when there is no such a common goal? In this paper we shall describe the new class of games the investigatior of which will give us the positive answer to this question.

The game models which are studied here, describe conflict situations characterized by active competition of players. Let us suppose that each player has two criteria, which are transformed in one criterion by the operation of taking minimum ("estimation by worse result"). One criterion describes competition of players in some common area (sphere) of activity (let us name it external sphere). The second criterion describes private achievements of the player in his own sphere of activity (let us name it internal sphere). Strategies of players are distributions of resources among external and internal spheres of activity. The value of the first criterion of each player depends on strategies of all players, because this criterion describes the relative achievements or influence in external (common) sphere. The natural conditions are its monotonously increasing with increase of one's own resources invested in external area and monotonously decreasing with increase of investmens of other participants. The second criterion depends only on value of one's own resources invested in internal area and naturally is monotonously increasing function.

Let us begin our consideration with the simplest case of two participants (players), which has scalar resources (for example, everything in monetary units). Let us denote the general value of resources of participant i by a_i, the value of resources invested in external area by x_i and in internal area by y_i $(i=1, 2)$. The criterion of the first player which describes the achievements in external area is the function $\psi_1(x_1, x_2)$, monotonously increasing by x_1 and monotonously decreasing by x_2. Analogously for the second player $\psi_2(x_1, x_2)$ monotonously increases by x_2 and monotonously decreases by x_1. The criteria of players which describes the achievements in internal area are monotonously increasing functions $h_i(y_i)$. As a condition of balance $x_i+y_i=a_i$ gives us relation $y_i=a_i-x_i$, let us reckon that the player's strategy concludes in choice of x_i from the segment $[0, a_i]$ and let us take as "internal" criterion of each player the function $\varphi_i(x_i)=h_i(a_i-x_i)$ which is correspondingly decreasing by x_i. In accordance with the adopted earlier kind of transformation of criterions the general criterion of each player is

$$u_i(x_1, x_2)=\min\{\psi_i(x_1, x_2), \varphi_i(x_i)\}, \qquad i=1, 2. \tag{1}$$

The sense of relation (1) is that the players estimate their positions by states of internal and external affairs and determing is that area, where the affaire are relatively worse. Certainly, such transformation is not the only possible one, but in this situation it seems to be quite grounded.

About the functions ψ_i, φ_i we shall suppose that they are continuous and

$$\psi_i(0, 0)<\varphi_i(0),\ \varphi_i(a_i)=\psi_1(0, a_2)=\psi_1(a_1, 0)=0$$

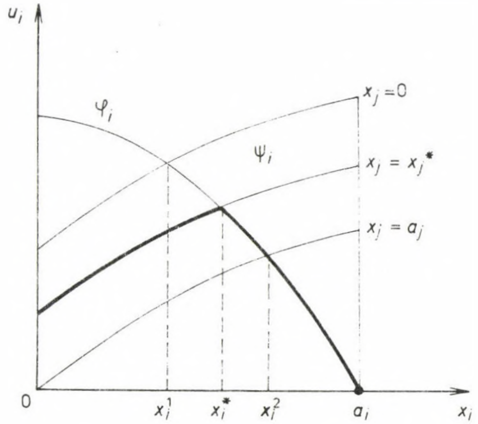


Fig. 1

(these conditions are not essential and serve only for simplification of analyses of the game excluding the cases when the solutions are in the ends of the segments $[0, a_i]$).

Let us imagine graphic sight of criteria (in Figure 1 the graphic of function u_i of argument x_i under fixed $x_j = x_j^*$, $j \neq i$, is drawn by thick line).

2. Investigation of two-person game

Analysis of the game with criterion (1) will be connected with investigation of existence and propeties of equilibrium first of all. Let us remind that the point (situation) of equilibrium is such (x_1^*, x_2^*) that

$$\begin{aligned} u_1(x_1^*, x_2^*) &\geqq u_1(x_1, x_2^*) \qquad \forall x_1 \in [0, a_1]\,, \\ u_2(x_1^*, x_2^*) &\geqq u_2(x_1^*, x_2) \qquad \forall x_2 \in [0, a_2]\,. \end{aligned} \tag{2}$$

If inequalities (2) are strict under $x_1 \neq x_1^*$, $x_2 \neq x_2^*$, then the situation of equilibrium is called strict.

Lemma 1. Necessary and sufficient conditions of equilibrium are equalities

$$\psi_i(x_1^*, x_2^*) = \varphi_i(x_i^*), \qquad i = 1, 2\,. \tag{3}$$

The proof of this fact, taken into account the monotony of function, is rather obvious.

Let us denote

$$v_i(x_j) = \max_{0 \leqq x_i \leqq a_i} u_i(x_1, x_2) \tag{4}$$

and realization of maximum in (4) by $x_i^0(x_j)$.

Lemma 2. The function $v_i(x_j)$ is monotonously decreasing and the function $x_i^0(x_j)$ is single value continuous and monotonously increasing.

The proof of this lemma is also evident.

Theorem 1. In the game with criterions (1) there exists at least one situation of equilibrium and all situations of equilibrium are strict.

Proof. Let us consider the map $(x_1, x_2) \rightarrow (x_1^0(x_2), x_2^0(x_1))$. It is a single value and continuous map of rectangle $[0, a_1] \times [0, a_2]$ in itself, consequently by Brower's theorem it has a fixed point which is evidently the point of equilibrium. It is strict as the function $x_i^0(x_j)$ is single value.

Figure 2 illustrates Theorem 1 (origin of the points x_i^1, x_i^2 is clear from Figure 1). The equilibrium point may be not singular (we shall give corresponding example further).

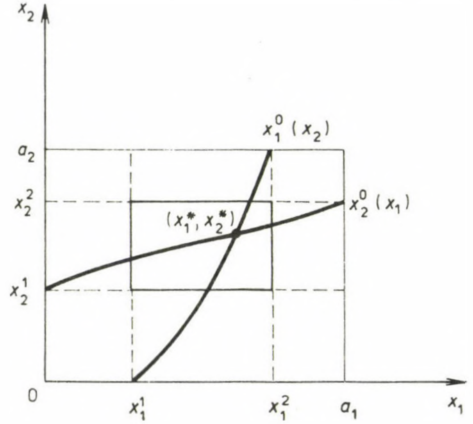


Fig. 2

Lemma 3. If there are several equilibrium points, then among them there is the best one for both participants (with minimal coordinates).

Proof. Let us suppose that there are two equilibrium points (x_1^*, x_2^*) and (x_1^{**}, x_2^{**}). From monotony of $x_i^0(x_j)$ we have $x_1^* < x_1^{**} \Leftrightarrow x_2^* < x_2^{**}$. In the equilibrium point by Lemma 1 $u_i^* = u_i(x_1^*, x_2^*) = \varphi_i(x_i^*)$ and from monotony of function φ_i we have $u_i^{**} < u_i^*$, $i = 1, 2$. From continuity of function $x_i^0(x_j)$ the existence of the best equilibrium follows under arbitrary number of equilibrium points (equilibrium point with minimal coordinates).

For exposition of othr properties of equilibrium we shall remind two important conceptions.

The point (x_1, x_2) is called Pareto point (or pareto-optimal) if there does not exist such point (x_1', x_2'), for which

$$u_i(x_1', x_2') \geqq u_i(x_1, x_2), \qquad i = 1, 2, \tag{5}$$

and at least one of the inequalities (5) is strict.

For definition of the second conception let us consider the game with criterion (1) where one player (leader) plays first (chooses x_i) and communicates his choice to the other player (partner). Let us call this game Γ^1. As the rational choice of the partner in this game is determined $x_j^0(x_i)$, then it is easy to find the best result of the leader. For the first player

$$\gamma_1 = \max_{0 \leqq x_1 \leqq a_1} u_1(x_1, x_2^0(x_1)), \tag{6}$$

and for the second player

$$\gamma_2 = \max_{0 \leqq x_2 \leqq a_2} u_2(x_1^0(x_2), x_2). \tag{7}$$

The optimal strategy of the first player (if he is a leader) is x_1^0 which realizes maximum in (6), analogously for the second player being leader, it is x_2^0, realizing maximum in (7).

The set of Pareto points where player i gains no less than γ_i $(i=1, 2)$ is called γ-core.

Lemma 4. In the game Γ^1 with any leader (under arbitrary sequence of moves) both players gain no less than in the best equilibrium point.

Proof. In the game Γ^1 the leader can guarantee himself a gain no less than in any strict equilibrium point (by the choice of equilibrium strategy). So the proof needs only for the partner. If (x_1^*, x_2^*) is the best equilibrium, then for the optimal strategy of the leader in Γ^1 it fulfils $x_i^0 \leqq x_i^*$ (if $x_i^0 > x_i^*$ then $u_i^0 \leqq \varphi_i(x_i^0) < \varphi_i(x_i^*) = u_i^*$ and we come to the contradiction with the statement of this lemma for the leader). As $v_j(x_i)$ decreases by x_i, then the partner in Γ^1 gains also no less than u_j^*.

The next result is the most interesting one.

Theorem 2. If the function

$$\psi(x_1, x_2) = \psi_1(x_1, x_2) + \psi_2(x_1, x_2)$$

does not decrease monotonously by x_1, x_2, in particular $\psi(x_1, x_2) \equiv \text{const}$, then the situation of equilibrium in the game with criterion (1) is unique, pareto-optimal and is a unique point of γ-core.

Proof. Let us suppose that there are two equilibrium points (x_1^*, x_2^*) and (x_1^{**}, x_2^{**}), and $x_i^{**} > x_i^*$, $i=1, 2$. Then

$$u_i^{**} = u_i(x_1^{**}, x_2^{**}) = \varphi_i^{**} < \varphi_i^* = u_i^*, \qquad i=1,2.$$

On the other hand,

$$u_1^{**} + u_2^{**} = \psi_1^{**} + \psi_2^{**} = \psi^{**} \geqq \psi^* = \psi_1^* + \psi_2^* = u_1^* + u_2^*$$

and we come to the contradiction, consequantly, the equilibrium is unique. Further, if $x_i > x_i^*$, then $u_i(x_1, x_2) \leqq \varphi_i(x_i) < \varphi_i^* = u_i^*$, $i=1, 2$. If $x_i \leqq x_i^*$, $i=1, 2$, then

$$u_1(x_1, x_2) + u_2(x_1, x_2) \leqq \psi_1(x_1, x_2) + \psi_2(x_1, x_2) \leqq \psi_1^* + \psi_2^* = u_1^* + u_2^*,$$

so if $u_i(x_1, x_2) > u_i^*$, then $u_j(x_1, x_2) < u_j^*$, that is the equilibrium point (x_1^*, x_2^*) is pareto-optimal.

Let us show now that (x_1^*, x_2^*) is a unique solution of the game Γ^1 under arbitrary sequence of moves. In fact, if $x_i > x_i^*$, then $u_i(x_1, x_2) < u_i^*$; if $x_i < x_i^*$ then the partner in virtue of monotonous decreasing of function $v_j(x_i)$ gains more than u_j^*, as the point (x_1^*, x_2^*) pareto-optimal the leader gains less than u_i^*. Consequently, for the optimal strategy in the game Γ^1 it fulfils $x_i^0 = x_i^*$ and the partner chooses x_j^*. So under arbitrary sequence of moves in the game Γ^1 players gain u_i^* (it means that the sequence of moves is indifferent for them), and in any other point evidently at least one player gains less than u_i^*, i.e. (x_1^*, x_2^*) is a unique point of γ-core. The theorem is proved.

For the illustration of the results let us consider some examples.

Example 1

$$\psi_i = a + x_i - x_j, \quad \varphi_i = 2(a - x_i), \quad 0 \leqq x_i \leqq a, \quad i, j = 1, 2.$$

Here $\psi \equiv 2a$, i.e. the condition of Theorem 2 is fulfilled. Solving the system of equations (3) we get $x_1^* = x_2^* = \frac{a}{2}$, $u_1^* = u_2^* = a$. In virtue of Theorem 2 $\left(\frac{a}{2}, \frac{a}{2}\right)$ is unique equilibrium point, pareto-optimal and constitute γ-core (it is a solution of the game Γ^1 under arbitrary sequence of moves). Let us mark that in spite of the fact that (because $\psi \equiv \text{const}$) "external" interests of players are antagonistic, the game with criterions $u_i = \min\{\psi_i, \varphi_i\}$ is not antagonistic

$$\left(\min_{x_i} \min_{x_j} u_i = \min_{x_j} \max_{x_i} u_i = \frac{2}{3} a\right).$$

Example 2.

$$\psi_1 = 2a + x_1 - 2x_2, \quad \psi_2 = 2a + \frac{1}{2} x_2 - 2x_1, \quad \varphi_i = 3(a - x_i), \quad 0 \leqq x_i \leqq a, \quad i = 1, 2.$$

Here $\psi = 4a - x_1 - \frac{3}{2} x_2$, i. e. monotonously decreases by x_1, x_2 and the condition of Theorem 2 is not fulfilled. We have $x_1^* = \frac{11}{20} a$, $x_2^* = \frac{3}{5} a$, $u_1^* = \frac{27}{20} a$, $u_2^* = \frac{6}{5} a$, the equilibrium point $\left(\frac{11}{20} a, \frac{3}{5} a\right)$ is not pareto-optimal. The solutions of game Γ^1 correspondingly with the first and the second player in the role of leader are $x_1^0 = 0$, $u_1^0 = \frac{10}{7} a$ $\left(\text{for the partner } \hat{x}_2 = \frac{2}{7} a,\ \hat{u}_2 = \frac{15}{7} a\right)$ and $x_2^0 = 0$, $u_2^0 = \frac{3}{2} a$ $\left(\text{for the partner } \hat{x}_1 = \frac{1}{4} a,\ \hat{u}_1 = \frac{9}{4} a\right)$, they are not also Pareto points. The inequalities $\hat{u}_i > u_i^0 > u_i^*$ are fulfilled, i.e. here the sequence of moves in the game Γ^1 is essential (it is profitable to be a partner), the equilibrium for both players is worse than the role of leader or partner, but Pareto point (0, 0), where $u_1 = u_2 = 2a$, is better for both players.

This example shows that the condition of Theorem 2 is essential.

Example 3.

$$\psi_i = 1 + x_i - 2x_j, \quad \varphi_i = 1 - x_i, \quad 0 \leqq x_i \leqq 1, \quad i, j = 1, 2.$$

Here $\psi = 2 - x_1 - x_2$, i.e. monotonously decreases. All the points of diagonal $x_i = x_j$ of the square $[0, 1] \times [0, 1]$ are equilibriums. The best equilibrium for both players (0, 0) is a şolution of the game Γ^1 under arbitrary sequence of moves and Pareto point, i.e. some properties formulated in Theorem 2 are fulfilled.

Example 4.

$$\psi_i = \rho a + x_i - x_j, \quad \varphi_1 = \rho a - x_1, \quad \varphi_2 = K(a - x_2),$$

$$0 \leqq x_1 \leqq \rho a, \quad 0 \leqq x_2 \leqq a, \quad \rho > 1, \quad K > 1, \quad i, j = 1, 2.$$

There is a certain sense in this example (though we want to emphasize that the general game model is interesting also by its sapid aspect). Parameter ρ can be interpreted here as a rate of superiority of the first player in resources and parameter K as a rate of insensibility of the second player to internal affairs (if K is greater than he attaches more importance to external area and invests the greater part of resources here).

In dependence of relation between K and ρ we have

$$x_1^* = \begin{cases} \dfrac{(K-\rho)a}{2K+1}, & K > \rho, \\ 0, & K \leqq \rho, \end{cases} \qquad x_2^* = \begin{cases} \dfrac{2(K-\rho)a}{2K+1}, & K > \rho, \\ 0, & K \leqq \rho, \end{cases}$$

$$u_1^* = \begin{cases} \dfrac{(2K\rho + 2\rho - K)a}{2K+1}, & K > \rho, \\ \rho a, & K \leqq \rho \end{cases} \qquad u_2^* = \begin{cases} \dfrac{(2K\rho + K)a}{2K+1}, & K > \rho, \\ Ka, & K \leqq \rho, \end{cases}$$

i.e. $u_1^* > u_2^*$ under $\rho > K$, $u_1^* = u_2^*$ under $\rho = K$, $u_1^* < u_2^*$ under $\rho < K$.

Thus a deficiency of resource can be compensated in certain degree by decreasing of importance of internal area, as a result a weaker participant can resist to a stronger one rather successfully.

3. Investigation of n-person game

The obtained results can be generalized in the case of vector resources, counteraction of players in two fields and so on. Here we shall consider a generalization on the case of n persons. We shall remind one more conception. Strong equilibrium in the game of n persons is such a point that for any coalition (subset of all players) it is not profitable to decline from it, i.e. there does not exist such set of strategies of coalition that using these strategies (other players use equilibrium strategies) all players of coalition gain no less than in the equilibrium and at least one player of coalition gains strictly more.

Strong equilibrium is also usual equilibrium for which in the definition it is necessary to take coalitions with single element and is Pareto point to which coalition of all players (big coalition) answers.

Let us consider a game of n persons with criterions

$$u_i(x_1, \dots, x_n) = \min\{\psi_i(x_1, \dots, x_n), \varphi_i(x_i)\}, \quad 0 \leqq x_i \leqq a_i, \quad i = \overline{1, n}\,. \tag{9}$$

Functions ψ_i, φ_i are supposed to be continuous, φ_i monotonously decreases by x_i, ψ_i monotonously increases by x_i and decreases by other arguments,

$$\varphi_i(a_i) = \psi_i(a_1, \dots, a_{i-1}, 0, a_{i+1}, \dots, a_n) = 0\,,$$

$$\varphi_i(0) > \psi_i(0, \dots, 0)\,.$$

Theorem 3. For n-person game with criterions (9) the next properties are valid:

1) necessary and sufficient conditions of equilibrium are

$$\psi_i(x_1^*, \dots, x_n^*) = \varphi_i(x_i^*), \qquad i = \overline{1, n}\,;$$

2) at !eas: one situation of equilibrium exists and each situation of equilibrium is strict;

3) if function $\psi(x) = \sum_{i=1}^{n} \psi_i(x)$, where $x = (x_1, \dots, x_n)$, does not decrease monotonously by all arguments, in particular $\psi(x) \equiv \text{const}$, then each equilibrium is strong (and consequently pareto-optimal);

4) if function $\psi_i(x)$ can be represented in the form $\psi_i\left(x_i, \sum_{j \neq i} x_j\right)$, then the best for all players situation of equilibrium exists (equilibrium point with minimal coordinates) and if simultaneously condition 3) is fulfilled, then equilibrium is unique.

Proof. The first and second properties are analogous to those for the game of two persons and are rather evident. So we shall begin with the third property. Let $(x_1^*, \dots, x_n^*)$ be equilibrium point and coalition S declines from it. If at least one $x_i > x_i^*$, then $u_i \leqq \varphi_i(x_i) < \varphi_i^* = u_i^*$, i.e. for coalition S on the whole it is not profitable to decline from the equilibrium $(x_1^*, \dots, x_n^*)$. If $x_i < x_i^*$, $i \in S$, then in a new point in virtue of monotony $\sum_{i \notin S} \psi_i > \sum_{i \notin S} \psi_i^*$, and as $\sum_{i=1}^{n} \psi_i \leqq \sum_{i=1}^{n} \psi_i^*$, then $\sum_{i \in S} \psi_i < \sum_{i \in S} \psi_i^*$. It means that there exists $i_0 \in S$ such that $\psi_{i_0} < \psi_{i_0}^*$ and $u_{i_0} \leqq \psi_{i_0} < \psi_{i_0}^* = u_{i_0}^*$, i.e. again for coalition S on the whole it is not profitable to decline from the equilibrium $(x_1^*, \dots, x_n^*)$. As coalition S is arbitrary we have that equilibrium is strong and consequently pareto-optimal.

Let us pass to property 4). Suppose that there are two equilibriums $(x_1^*, \ldots, x_n^*)$ and $(x_1^{**}, \ldots, x_n^{**})$. As functions ψ_i, φ_i are monotonous, realization of maximum u_i by x_i under other fixed arguments is a single-value monotonously increasing function $x_i^0\left(\sum_{j \neq i} x_j\right)$. So if $\sum_{i=1}^n x_i^{**} = \sum_{i=1}^n x_i^*$, then $x_i^{**} = x_i^*$, $i = \overline{1, n}$ (if $x_i^{**} > x_i^*$, then $\sum_{j \neq i} x_j^{**} <$ $< \sum_{j \neq i} x_i^*$, but $x_i^* = x_i^0\left(\sum_{j \neq i} x_j^*\right)$, we come to a contradiction). If $\sum_{i=1}^n x_i^{**} > \sum_{i=1}^n x_i^*$, then $x_i^{**} > x_i^*$, $i = \overline{1, n}$ (otherwise if $x_i^{**} \leqq x_i^*$, then $\sum_{j \neq i} x_j^{**} > \sum_{j \neq i} x_i^*$ and we come again to a contradiction). It means that among two different equilibriums one of them has all coordinates less than the other one, and such equilibrium is better for all players, because $u_i^* = \varphi_i^* > \varphi_i^{**} = u_i^{**}$, $i = \overline{1, n}$. Now it is evident that the best equilibrium for all players (with minimal coordinates) exists. Finally if property 3) is fulfilled simultaneously then each equilibrium is pareto-optimal, but only the best equilibrium can be pareto-optimal, consequently the equilibrium is unique.

Example 5.

$$\psi_i = c x_i \left(\sum_{j=1}^n x_j\right)^{-1}, \quad \psi_i(0) = 0,$$

$$\varphi_i = a_i - x_i, \quad 0 \leqq x_i \leqq a_i, \quad i = \overline{1, n}, \quad c < \sum_{j=1}^n a_j .$$

All conditions of Theorem 3 are fulfilled (discontinuity in zero is not essential), so there exists an unique strict and strong equilibrium. To find it we shall use the necessary and sufficient conditions

$$c x_i \left(\sum_{j=1}^n x_j\right)^{-1} = a_i - x_i, \qquad i = \overline{1, n},$$

from which we have

$$x_i^* = \frac{a_i \left(\sum_{j=1}^n a_j - c\right)}{\sum_{j=1}^n a_j}, \qquad i = \overline{1, n} .$$

Reference

1. *Germeyer, U. B., Vatel, I. A.,* Games with hierarchical vector of interests. Izv. AN SSSR, Tekhnicheskaya kibernetika, 1974, *3*, pp. 54–69.

Об одном классе игр с устойчивыми и паретооптимальными решениями

В. А. ГОРЕЛИК

(Москва)

Рассматривается класс игр с произвольным числом игроков, характеризующийся тем, что функции выигрыша всех участников представляют собой свертку с помощью операции взятия минимума двух критериев, один из которых описывает результаты соревнования для данного игрока в некоторой общей сфере деятельности и зависит от стратегий всех игроков, а второй описывает результаты деятельности данного игрока в его внутренней сфере деятельности и зависит только от его стратегии. Стратегиями игроков являются распределения ресурсов между общей и личной (внутренней) сферами деятельности.

Сначала исследуются игры двух лиц описанного вида. При некоторых естественных условиях монотонности доказана теорема существования равновесия и получены необходимые и достаточные условия равновесия. Получено условие, связанное с видом критериев, описывающих общую сферу деятельности, при выполнении которого равновесие является единственным и одновременно паретооптимальным. Приведены примеры, показывающие существенность данного условия.

Полученные результаты обобщаются на игры *n* лиц, для которых установлен факт существования равновесия, выведены необходимые и достаточные условия равновесия и найдены условия, при которых равновесие является единственным и сильным (в частности, паретооптимальным).

В. А. Горелик
Вычислительный центр АН СССР
СССР, 117333, Москва, ул. Вавилова, 40.

Problems of Control and Information Theory, Vol. 15 (4), pp. 319—331 (1986)

LOGICAL CONTROL OF ADAPTIVE ROBOT. PRINCIPLES OF CONTROL SYSTEM ORGANIZATION

S. L. ZENKEVICH

(*Moscow*)

(Received September 1, 1985)

The new approach to the problem of control system model forming of robot-based flexible manufacturing module is presented. To construct a robot's model, its control system is structurized. The new level of control system, the so-called logical level is created. The proposed approach is based on the description of all active elements of work cell, in particular manipulator, sensor system, machine tools, as finite automata. An illustrated example is considered.

1. Introduction

Up to now robot may be regarded as inalienable part of manufacturing. Nonetheless, the most part of existing approaches to robot control system design consider robot to be an alone standing unit without any connection with machine tools while it needs to be considered as one of the elements of the manufacturing cell, all components of which are involved in technological operation execution (Fig. 1). Such an approach is espesially well-founded in connection with the development and

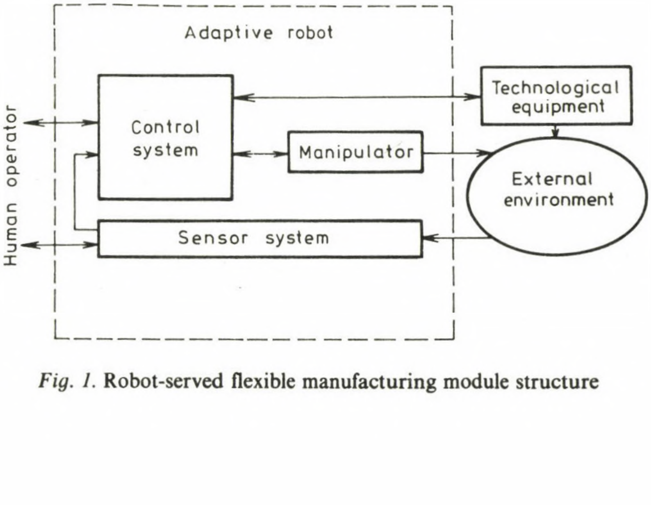


Fig. 1. Robot-served flexible manufacturing module structure

installation of flexible manufacturing systems (FMS), where robot is incorporated as a rule in flexible manufacturing module (FMM) structure [1, 2, 3].

The new approach to robot control system design, based on its structurization and forming its mathematical model as finite automata, is described in this paper on nonformal description level. It is shown that if a robot includes sensor system, its model must be filled up with stochastic automata thus providing the robot with nondeterministic behaviour. To control the robot, we introduce the so-called logical level of control system. This level represents the net of interactive automata performing the function of coordination of operations of all active elements involved in task execution. Possible implementation of such a control system as well as the example of practical utilization of the suggested approach is presented.

2. Structurization of robot control system

To build the mathematical model of a robot incorporated in FMM structure we shall structurize its control system [4, 5] on the basis of the next principles:

1. Every level of control system performs a strictly described function.
2. Data exchange is allowed only for neighbouring levels, and the value of this data needs to be as small as possible.
3. Description of every level of control system is allowed to have the form of any mathematical model used by upper level to choose its own control low.

Here we shall not consider the training mode but we shall briefly discuss how one can use the above-described principles for structurization of robot control system executing the previously formed task. Possible realization of multilevel control system looks as follows (Fig. 2).

Level L_1 (drivers level)

Function: evaluation and generation of actions for control of actuators in the manipulator joints. In the case when position feedback loop includes a computer, than level L_1 performs regulator functions. The computer program ensuring all these operations we call servo-system driver.
Input: desired manipulator state in the space of the angles (discplacements) of degrees of freedom.
Output: control signals for interface module to hardware.
Model: the model of the object to be controlled by this level consists of mathematical models of the mechanical part of the robot coupled with models of the actuators, driving the joints, amplifiers etc.

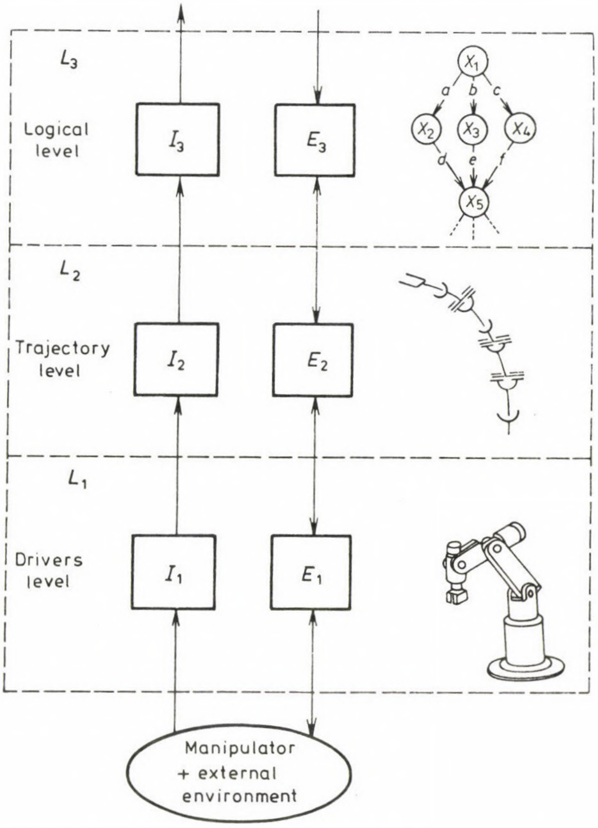


Fig. 2. Multilevel control system of adaptive robot. (E_i — executive modules, I_i — information modules)

Level L_2 (trajectory level)

Function: evaluation of the desired manipulator hand trajectory and conversion of this into the desired angular space trajectory.

Input: name of the task step, which defines the goal position as well as the kind of tracking of the desired trajectory including this position. Actually, it is often impossible for adaptive robots to determine this position while training, thus we shall use the notation "frame", mentioning that some elements of this structure need to be filled in during execution.

Output: coincides with the L_1 input.

Model: the model of the controlled object is assumed to be the kinematic model of the manipulator. In essence, the trajectory level is composite level itself and it may be represented as a set of sublevels, but we shall not go into details.

Level L_3 (logical level)

Function: logical control of all active units involved in technological task execution.
Input: a) from human-operator — the program of operations of the overall robot-served FMM.

b) from active units — information about its current states.

Output: control actions for active units.
Model: all objects to be controlled are considered as initial finite automata.

Let us discuss now this level organization in detail as it is innovation for robot control systems.

3. Finite automata

Recall that a finite automata K may be given by a set of five objects [6]:

$$K=\{U, X, Z, f, h\}$$

where

$U=(u_1, u_2, \ldots, u_l)$ – finite input alphabet,
$X=(x_1, x_2, \ldots, x_n)$ – finite set of automata states,
$Z=(z_1, z_2, \ldots, z_m)$ – finite output alphabet,
$f: X \times U \rightarrow X$ – one-step transfer function,
$h: X \times U \rightarrow Z$ – output function.

A finite automaton may be interpreted as a dynamic system which at time generates output signal z when being in state x. If at this time input signal u is received, then at time $t+1$ the system transfers to state $f(x, u)$, i.e.

$$x(t+1)=f(x(t), u(t))$$

$$z(t)=h(x(t)) \text{ – Moore automata,}$$

or

$$z(t)=h(x(t), u(t)) \text{ – Mealy automata .}$$

The automaton K is called initial finite automata if its description additionally contains initial state $x_0 \in X$.

4. Object models

As it follows from above described principles (Section 2), to construct logical level of control system we need previously to construct mathematical models f objects to be controlled. We shall build these models as initial finite automata.

Let us briefly discuss the way of construction of the models of all active units which are incorporated in FMM structure (Fig. 1).

Technological equipment

All objects of this kind can be easily described as finite automata as they have finite number of states as well as finite number of input and output alphabets elements.

For example, a manipulator hand, which can be associated for convenience with technological units (this approach has a lot of advantages) can be described as a finite automata with the next attributes:

$$U=(c,o)\,,$$
$$X=(x_1,x_2,x_3)\,,$$
$$Z=(z_0,z_c)\,,$$

$$\begin{aligned} f\colon &(x_1,o)\to x_2, & h\colon & x_2\to z_0,\\ &(x_1,c)\to x_3, & & x_3\to z_c,\\ &(x_2,c)\to x_3\,, & & \\ &(x_3,o)\to x_2\,, & & \end{aligned}$$

where letters "c" and "o" denote "closed" and "opened" correspondingly. The graph of automata transfers and outputs is presented in Fig. 3.

Manipulator

Let us consider two types of manipulator, distinguished by sort of drive used to control manipulator joints.

1) "pick-and-place" manipulator

In such a manner we shall call manipulator which is able to achieve a finite number of positions p_j ($j=\overline{1,n}$) in a work space governed by finite number of control

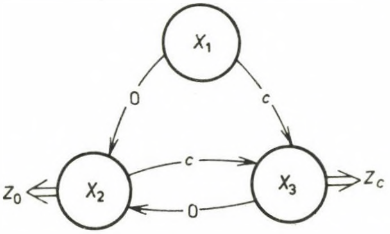


Fig. 3. Transition graph of manipulator hand described as finite automata

actions u_i $(i=\overline{1,l})$. Manipulators of this type are usually equipped with special sensors performing signals $d_k(k=\overline{1,m})$ when terminal state of every joint is achieved.

Then it is convenient to consider manipulator as finite automata with the next attributes:

$$U=\{u_i\}_{i=\overline{1,l}},\quad X=\{p_j\}_{j=\overline{1,n}},\quad Z=\{d_k\}_{k=\overline{1,m}}$$

functions $f(\;)$ and $h(\;)$ need to be constructed in a convenient manner. Kinematic scheme of standard industrial "pick-and-place" manipulator, its work space and transfer graph are given in Figs 4a, 4b, 4c, respectively.

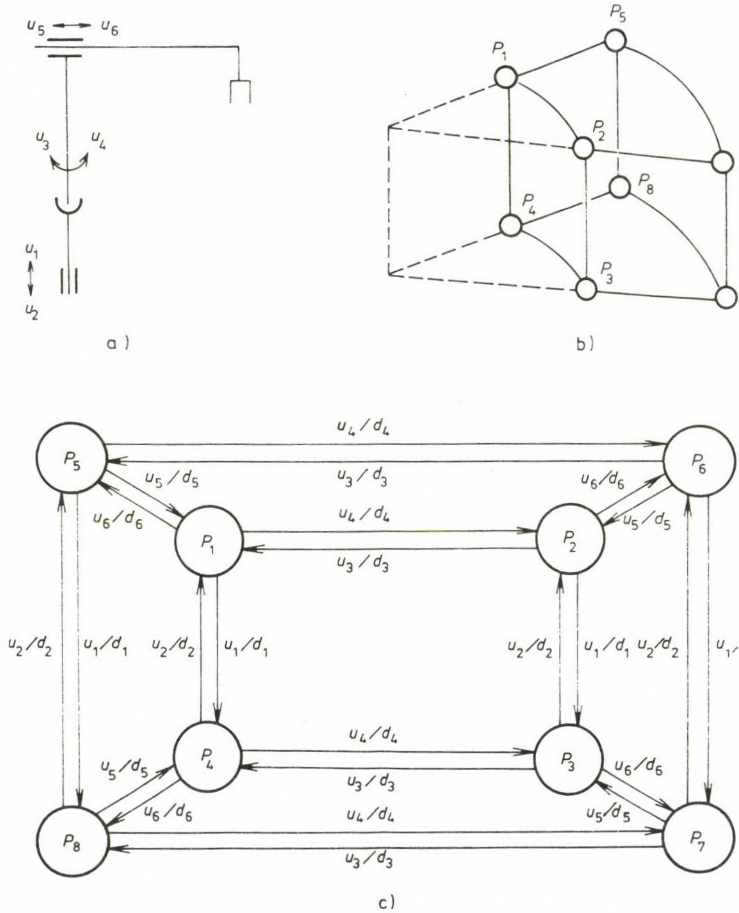


Fig. 4. **Industrial manipulator as finite automata:**
a) — kinematic scheme; b) — work space; c) — transfer graph

2) servosystem manipulator

Let us take the next assumptions:

— control system of robot with manipulator of this type is structurized as it have been described above.

— manipulator is trained.

Let $F=\{f_i\}$ be the set of frame names created in teaching operation. Let $E=\{e\}$ be the set of the only element e, confirming that the part of task described by every frame is executed. Let $X=F\cup\{x_0\}$ be the state set where x_0 is the initial state of the finite automata. Then the object to be considered is described as follows:

$$K=\{F, X, E, f, h\}$$

where productions $f(\ )$ and $h(\ )$ may be chosen in a convenient form. Kinematic scheme of illustrative manipulator and its transfer graph are presented in Fig. 5a and Fig. 5b, respectively.

Sensor system

In the multilevel robot control system description we have not discussed the functions performed by its information modules. When speaking about logical level one can confirm that the information received by this level has the property that the number of external environment descriptors values is finite.

For example let us consider a TV-based sensor system performing classification function. Let it attribute any object of the presented scene to one of the previously defined classes $a, b, c, \ldots, y_1 z$. Then description of the finite automata which is a model

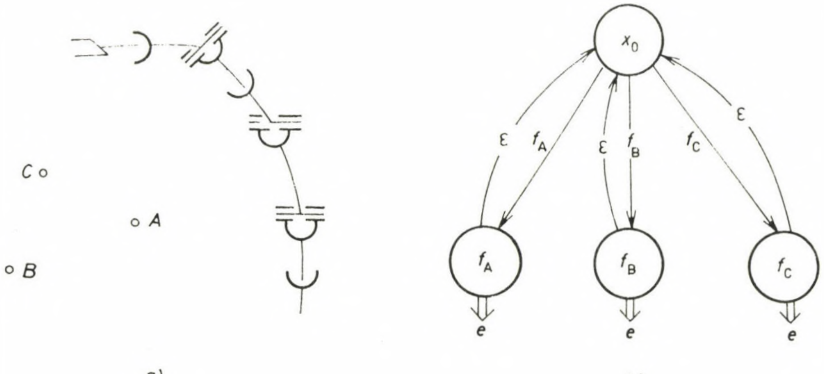


Fig. 5. Servosystem manipulator as finite automata:
a) — kinematic scheme (A, B, C — goal positions); b) — transfer graph

of a sensor system coupled with the part of software, performing this classification, looks as follows:

$$U=(s, \varepsilon),$$

$$X=(x_0, x_1),$$

$$Z=(a, b, c, \ldots, y, z).$$

The difference of the sensor system model from all previously described models is that is represents stochastic automata, i.e. for every element of output alphabet there corresponds the probability of its appearence in the output stream for every element of the output alphabet.

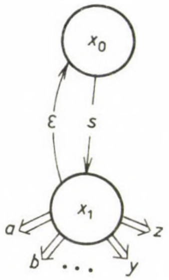


Fig. 6. TV-based sensor system as finite automata (classification mode)

Similarly, we can describe TV-system, performing the visual inspection function, tactile-based sensor system etc. The information from the other sensor systems such as TV-system determining the location of objects in work space, range sensors, force and torque sensors, is not used by logical level of the control system; it is used by lower levels for providing control system with parametrical adaptation.

Thus models of all active elements are built as finite automata.

5. Logical level of control system

Let us consider at first the simple problem of control of one active element. Let

$$K=\{U_K, X_K, Z_K, f_K, h_K\}$$

be the model of this active element (manipulator, for example). Then interpreting this element as object to be controlled, let us construct the automata described as follows (Fig. 7):

$$D=\{U_D, X_D, Z_D, f_D, h_D\},$$

where

$$U_D = Z_K \cup U'_D, \quad Z_D = U_K \cup Z'_D.$$

We shall call automata D regulator or logical driver of automata K. It is easy to see that elements of the output alphabet of automaton K are input signals for driver D and on the other hand, the output alphabet of driver D represents the input alphabet for K, thus providing the required behaviour of controlled object K, that is tracking of the previously defined sequence of states $\{x_i\}$, $x_i \in X$.

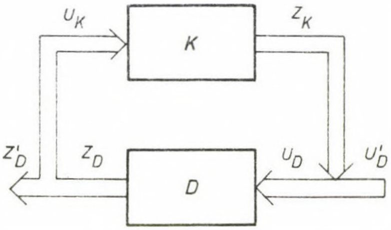


Fig. 7. Object control by means of regulator

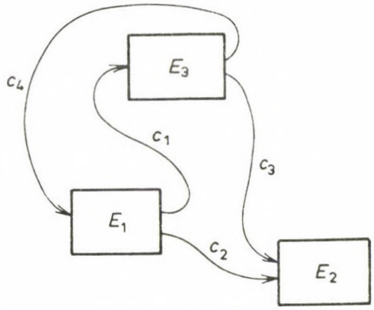


Fig. 8. Logical net

Let the set of controlled objects with models K_i, $i = \overline{1, n}$, be given. Then every object needs to be controlled by regulator D_i. But to coordinate the operations of these regulators, their interaction must be ensured. This interaction is accomplished by means of alphabets U'_D and Z'_D which are the elements of regulator description.

Thus the logical level of control system needs to be realized as the net of interactive finite automata exchanging elements with their output alphabets during work. In detail, it means the following.

The oriented multigraph L is said to be net (Fig. 8) if

$$L = \{F, C\}$$

where

$E=(E_1, E_2, \ldots, E_q)$ — net nodes, further called elements,

$C=(c_1, c_2, \ldots, c_l)$ — net directed arcs, further called channels,

with

$$c_i=(E_j, E_k), \quad E_j, E_k \in E\,.$$

Each net element represents the set of three components as follows:

$$E=\{R, A, S\}$$

where R is the input commutator, A is finite automata, and S is output commutator. Input commutator R can be described as follows:

$$R=\{\bar{U}, X, U, r\}$$

where

$\bar{U}=U^1 \times U^2 \times \ldots \times U^m$ is the set of the element E inputs with

$$\bar{u} \in \bar{U}, \bar{u}=(u_1, u_2, \ldots, u_m), \quad u_i \in U^i\,,$$

$X=(x_1, x_2, \ldots, x_n)$ is the set of the automata's A states,

$U=\bigcup_{i=1}^{m} U^i$ — the set of commutator R outputs,

$r: \bar{U} \times X \to U$ — commutating function of R.

Automata A is described as follows:

$$A=\{U, X, Z, f, h\}\,,$$

where the attributes have the same sense as in Section 3.

Output commutator S is described as follows:

$$S=\{Z, X, \bar{Z}, s\}$$

where

$\bar{Z}=Z \times Z \times \ldots Z=Z^p$ — the set of element E outputs,

$s: X \times Z \to \bar{Z}$ — commutating function of S.

The element E operation consists in joint functioning of automata A, input and output commutator and it includes the execution of the next steps:

— receiving a symbol from one of the input channel of element E in accordance with the state of automata A and sending this symbol to its input;

— transition automata A to new state and generation of the symbol of output alphabet;

— sending this symbol to channels defined by the commutating function of output commutator.

Net L operation consists in the parallel work of its elements E initiated by the symbols of the corresponding alphabets.

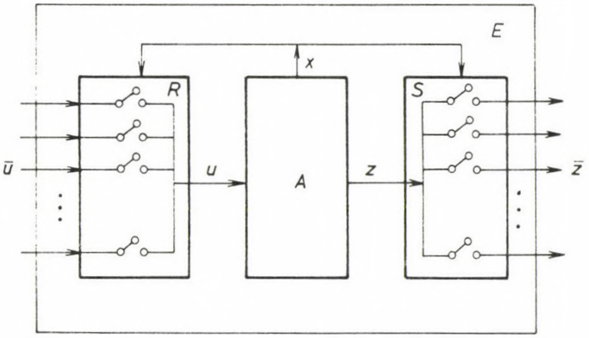


Fig. 9. Element as logical net component

Thus the logical level of a control system is considered to be logical multilevel neι consisting of interacting elements; the lower net level includes logical drivers immediately connected with objects to be controlled, the upper levels ensure the coordination of the FMM subsystems operation controlling its overall behaviour.

6. Example

The approach described above has been implemented in design of the adaptive assembly robot-served cell control system. This work cell contains:

— all-electrical six degree-of-freedom anthropomorphic manipulator;

— industrial control unit modified in the aim of increasing its computing abilities;

— TV-based sensor system;

— machine tools for executing the assembly operation.

The software of this work cell represents a problem-oriented operating system running under RSX-11S Executive. The elements of the net representing the logical level of control system are realized as parallel running tasks (processes). Experimental implementation of this approach have demonstrated high performance indexes of FMM control system.

7. Conclusion

We have briefly described the principles of robot control system forming, based on its structurization. The control system designed by means of the above described approach ensures:

— the control of all active units incorporated in FMM structure, in particular the control of robot, machine tools and sensor system;

— the coordination of all active units' operation;

— the formalization of the analysis and synthesis of the control system on the basis of the modern theory of control;

— the consideration of the designed control system as flexible manufacturing module control system providing its connection with FMS elements.

References

1. *Hatvany, J., Horváth, M., Somló, J.*, The computer-controlled manufacturing cell-achievements, possibilities and perspectives. Preprints IFAC, 9th World Congress, Hungary, 1984, **11**, pp. 93–98.
2. *Ránky, P. G.*, A software library for designing and controlling flexible manufacturing systems. Preprints IFAC, 9th World Congress, Hungary, 1984, **6**, pp. 147–152.
3. *Milačić, V., Kalajdžić, N., Milutinović, D.*, Simulation methods used to analyze a robot's tasks in flexible manufacturing systems based on artificial intelligence. Proc. 2nd Yugoslav–Soviet Symposium on Applied Robotics. Beograd, 1984, pp. 37–51.
4. *Попов Е. П., Верещагин А. Ф., Зенкевич С. Л.*, Манипуляционные роботы. Динамика и алгоритмы. М., Наука, 1978, 398 с.
5. *Попов Е. П.* (ред.), Робототехника. М., Машиностроение, 1984, 288 с.
6. *Arbib, M. A.*, The algebraic theory of machines, languages and semigroups. Academic Press, 1968.

Логическое управление адаптивным роботом. Принципы построения системы управления

С. Л. ЗЕНКЕВИЧ

(Москва)

Принадлежность робота гибкому производственному модулю (ГПМ) предъявляет ряд требований к его системе управления, важнейшим из которых является обеспечение его совместной работы с входящими в состав ГПМ оборудованием, синхронизация их совместных действий, направленных на выполнение технологической операции. По сути дела, все существующие в настоящее время подходы к проектированию системы управления робота рассматривают его как автономно действующий механизм, и лишь в малой степени учитывают его связь с технологическим окружением.

В статье решается задача построения математической модели робота, описание его системы управления и очувствления как конечного автомата со специально выбранными атрибутами, и далее — построение системы управления роботизированного ГПМ.

На основе разработанных моделей строится логический уровень системы управления, в функции которого входит управление всей совокупностью активных устройств, составляющих ГПМ. Мы строим этот уровень как многоуровневую сеть взаимодействующих конечных автоматов, при этом нижний уровень (так называемые логические драйверы) управляет непосредственно активными устройствами, а верхние обеспечивают синхронизацию их совместной работы.

В заключении приведен пример использования разработанных принципов при создании системы управления адаптивного сборочного робототехнического комплекса, построенного на базе аптропоморфного манипулятора и системы технического зрения.

С. Л. Зенкевич
Научно-учебный Центр «Робототехника»
АН СССР и Минвуза СССР,
СССР, 105037 Москва, Измайловская пл. 7.

РУССКИЙ ПЕРЕВОД

Проблемы Управления и Теории Информации, том **15**, *номер 4 (1986)*

О ЗАДАЧЕ ИДЕНТИФИКАЦИИ МАТРИЧНОГО ПАРАМЕТРА

А. М. Устюжанин

(*Свердловск*)

С использованием аппарата опорных функций описаны информационные множества [I] в линейной дискретной задаче оценивания постоянного матричного параметра в условиях неопределенности. Даны условия точной идентификации, т. е. вырождения информационного множества в точку, в предположении, что помеха в уравнении наблюдения принадлежит выпуклому компакту с непустой внутренностью. Рассмотрены примеры при дополнительных предположениях относительно помехи, в частности пример точной идентификации при случайных помехах.

Введение

А. *Постановка задачи*

Обозначим через $\{z_k\}_{k\in I}$ множество векторов z_k конечномерного евклидова пространства с индексами k из некоторого подмножества I неотрицательных целых чисел $\mathbf{N}$. Через $\{z_k\}_{k=1}^N$ обозначим предыдущее множество, если $k\in\overline{1,N}$.

Для матриц C из евклидова пространства $\mathbf{R}^{m\times n}$ рассмотрим уравнение наблюдения:

$$y_k=Cx_k+\xi_k,\qquad k\in\overline{1,N}\,. \tag{1}$$

Заданные вектор-столбцы $x_k\in\mathbf{R}^n$ будем называть входами уравнения (1), а измеряемые величины $y_k\in\mathbf{R}^m$ — выходами. Векторы $\xi_k\in\mathbf{R}^m$ суть неопределенные помехи, информация о которых исчерпывается условием

$$\xi_k\in\Xi_k,\qquad k\in\overline{1,N}\,, \tag{2}$$

где Ξ_k — выпуклые компакты в $\mathbf{R}^m$ с непустой внутренностью.

Скалярное произведение в $\mathbf{R}^{m\times n}$ задается формулой $(C_1, C_2)=\operatorname{tr}(C_1^TC_2)$, где $\operatorname{tr}(\cdot)$ — след матрицы, а ^{T} — символ транспонирования.

Следуя работе [1], дадим

Определение. Информационным множеством $\mathscr{S}$ называется множество матриц C из $\mathbf{R}^{m\times n}$, удовлетворяющих уравнению (1) и условию (2) при заданных $\{x_k\}_{k=1}^N$, $\{y_k\}_{k=1}^N$, $\{\Xi_k\}_{k=1}^N$.

Задача идентификации состоит в описании множества $\mathscr{S}$.

Пусть C_* — матрица, для которой по заданным входам x_k и реализовавшимся помехам $\bar{\xi}_k \in \Xi_k$ в силу (1) получены выходы y_k. Далее всюду будут рассматриваться именно такие множества $\{x_k\}_{k=1}^N$ и $\{y_k\}_{k=1}^N$. Следовательно, имеем $C_* \in \mathscr{S}$.

B. Обозначения

$$\mathscr{L}\{z_k\}_{k\in I} \triangleq \{z : z \sum_{k\in I} \alpha_k z_k, \alpha_k \in \mathbf{R}\} ,$$

$$\mathscr{L}^+\{z_k\}_{k\in I} \triangleq \{z : z = \sum_{k\in I} \alpha_k z_k, \alpha_k \geqq 0\} ,$$

$\partial\Xi$ — граница множества Ξ.

Для $\varepsilon > 0$ и реализовавшейся помехи $\bar{\xi}_k$ рассмотрим следующие множества

$$U_k(\varepsilon) \triangleq \{\xi_k \in \mathbf{R}^m : \|\xi - \bar{\xi}_k\| \leqq \varepsilon, \xi \in \partial\Xi_k\} ,$$

$$\Lambda_k(\varepsilon) \triangleq \{\lambda_k \in \mathbf{R}^m : \|\lambda_k\| = 1, \exists \xi_k \in U_k(\varepsilon)\ \rho(-\lambda_k|\Xi_k) = -\lambda_k^T \xi_k\} , \tag{3}$$

где $\exists$ — квантор существования, а $\rho(\cdot|\Xi_k)$ — опорная функция множества Ξ_k [2]. Здесь и ниже символ $\|\cdot\|$ — евклидова норма. Заметим, что при $\bar{\xi}_k \notin \partial\Xi_k$ множества (3) для достаточно малых ε пустые.

Точная идентификация

Теорема 1. Пусть множество входов $\{x_k\}_{k=1}^N$ удовлетворяет условию

$$\mathscr{L}\{x_k\}_{k=1}^N = \mathbf{R}^n .$$

Тогда информационное множество $\mathscr{S}$ есть выпуклый компакт и опорная функция множества $\mathscr{S}$ имеет вид

$$\rho(L|\mathscr{S}) = \inf_{\lambda_k \in \mathbf{R}^m} \left\{ \sum_{k=1}^N \lambda_k^T y_k + \sum_{k=1}^N \rho(-\lambda_k|\Xi_k) : \sum_{k=1}^N \lambda_k x_k^T = L \right\} . \tag{4}$$

Доказательство. Обозначим через $\mathscr{S}_k$ множества матриц C, удовлетворяющих включению $y_k - Cx_k \in \Xi_k$. Выпуклость и замкнутость $\mathscr{S}_k$ проверяются непосредственно.

Выпишем опорную функцию множества $\mathscr{S}_k$:

$$\rho(L|\mathscr{S}_k) = \begin{cases} \lambda_k^T y_k + \rho(-\lambda_k|\Xi_k), \ \lambda_k x_k^T = L \\ +\infty, \qquad \text{в противном случае} . \end{cases} \tag{5}$$

Множество $\mathscr{S}$ есть пересечение множеств $\mathscr{S}_k$, $k \in \overline{1, N}$, поэтому опорная функция $\mathscr{S}$ равна инфимальной конволюции [2] фынкций (5):

$$\rho(L|\mathscr{S}) = \begin{cases} \inf\limits_{L_k} \left\{ \sum\limits_{k=1}^{N} \rho(L_k|\mathscr{S}_k) : \sum\limits_{k=1}^{N} L_k = L, \quad \lambda_k x_k^T = L_k \right\}, \\ +\infty, \ \forall \lambda_k \in \mathbf{R}^m \ \sum\limits_{k=1}^{N} \lambda_k x_k^T \neq L \, . \end{cases} \tag{6}$$

Условие теоремы является необходимым и достаточным для того, чтобы любая матрица $L \in \mathbf{R}^{m \times n}$ могла быть представлена в виде $L = \sum\limits_{k=1}^{N} \lambda_k x_k^T$. Необходимость этого условия докажем от противного. Пусть $\mathscr{L}\{x_k\}_{k=1}^{N}$ — собственное подпространство в $\mathbf{R}^n$, тогда найдется ненулевой вектор x_0, для которого $(x_0, x_k) = 0$, $k \in \overline{1, N}$. По условию $L^T \triangleq (x_0, 0, \ldots, 0) = \sum\limits_{k=1}^{N} x_k \lambda_k^T$.

Тогда $x_0 = \sum\limits_{k=1}^{N} x_k \lambda_k^1$ и $\|x_0\|^2 = \left(x_0, \sum\limits_{k=1}^{N} x_k^1 \lambda_k \right)$, что противоречит предположению.

Достаточность условия $\mathscr{L}\{x_k\}_{k=1}^{N} = \mathbf{R}^n$ вытекает из возможности разразложить каждую j-ю строку матрицы L по векторам x_k^T с коэффициентами разложения λ_k^j, $j \in \overline{1, m}$, $k \in \overline{1, N}$.

Таким образом, в формуле (6) второй случай невозможен, и она может быть переписана в виде (4). Это одновременно доказывает и ограниченность множества $\mathscr{S}$. Выпуклость и замкнутость его вытекают из соответствующих свойств множеств $\mathscr{S}_k$ и свойств операции пересечения. Теорема доказана.

Далее опорную функцию множества $\mathscr{S}$ будет удобно представить в виде

$$\rho(L|\mathscr{S}) = \inf_{\|\lambda_k\| = 1, \, \alpha_k \geqq 0} \left\{ \sum_{k=1}^{N} \alpha_k \lambda_k^T y_k + \sum_{k=1}^{N} \rho(-\alpha_k \lambda_k | \Xi_k) : \sum_{k=1}^{N} \alpha_k \lambda_k x_k^T = L \right\}.$$

Следствие 1. Теорема о точной идентификации.

Пусть для матрицы $C_* \in \mathscr{S}$ выполнены условия:

$$\forall \varepsilon > 0, \quad \forall L \in \mathbf{R}^{m \times n}, \quad \exists I \subseteq \overline{1, N}, \quad \exists \lambda_k \in \Lambda_k(\varepsilon), \quad \exists \alpha_k \geqq 0 :$$

$$: \sum_{k \in I} \alpha_k \lambda_k x_k^T = L \, , \tag{7}$$

$$\lim_{\varepsilon \to 0} \sum_{k \in I} \alpha_k \lambda_k^T (\bar{\xi}_k - \xi_k) = 0 \, . \tag{8}$$

Здесь ξ_k — точка из $U_k(\varepsilon)$, которая в силу определения множества $\Lambda_k(\varepsilon)$ соответствует вектору λ_k из условия (7).

Тогда информационное множество $\mathscr{S}$ содержит единственную матрицу C_*.

Напомним, что $y_k = C_* x_k + \bar{\xi}_k$, $k \in \overline{1, N}$.

Доказательство. Покажем, что опорная функция множества $\mathscr{S}$ равна (L, C_*), какова бы ни была матрица $L \in \mathbf{R}^{m \times n}$. Это и будет означать, что $\mathscr{S} = \{C_*\}$.

Фиксируем $L \in \mathbf{R}^{m \times n}$. Для каждого $\varepsilon > 0$ имеем

$$\rho(L|\mathscr{S}) \leqq \sum_{k=1}^{N} \alpha_k \lambda_k^T y_k + \sum_{k=1}^{N} \rho(-\alpha_k \lambda_k | \Xi_k) =$$

$$= \sum_{k \in I} \alpha_k \lambda_k^T C_* x_k + \sum_{k \in I} \alpha_k \lambda_k^T \bar{\xi}_k + \sum_{k \in I} (-\alpha_k \lambda_k^T \xi_k) =$$

$$= \left(\sum_{k \in I} \alpha_k \lambda_k x_k^T, C_* \right) + \sum_{k \in I} \alpha_k \lambda_k^T (\bar{\xi}_k - \xi_k) \triangleq \varphi(L, \varepsilon)$$

Здесь $\alpha_k = 0$ при $k \notin I$. При $k \in I$ α_k, λ_k, ξ_k выбираются из условий (7), (8). Следовательно,

$$\rho(L|\mathscr{S}) \leqq \inf_{\varepsilon > 0} \{\varphi(L, \varepsilon)\} \leqq \lim_{\varepsilon \to 0} \varphi(L, \varepsilon) = (L, C_*)\,.$$

Но $C_* \in \mathscr{S}$, поэтому верно обратное неравенство $\rho(L|\mathscr{S}) \geqq (L, C_*)$. Таким образом, $\rho(L|\mathscr{S}) = (L, C_*)$.

Следствие 2. Пусть каждое множество Ξ_k есть шар с центром в нуле:

$$\Xi_k = \{\xi : \|\xi\| \leqq \nu_k,\ \nu_k > 0\}\,, \qquad k \in \overline{1, N}\,,$$

и для некоторого подмножества индексов $I \subseteq \overline{1, N}$ выполнены условия

$$\bar{\xi}_k \in \partial \Xi_k, \qquad k \in I\,, \tag{9}$$

$$\forall L \in \mathbf{R}^{m \times n}; \quad \exists \{\alpha_k : \alpha_k \geqq 0\}_{k \in I}, \quad \sum_{k \in I} \alpha_k \bar{\xi}_k x_k^T = L \tag{10}$$

Тогда информационное множество $\mathscr{S}$ содержит единственную матрицу C_*.

Доказательство. Покажем, что при всех $L \in \mathbf{R}^{m \times n}$ будет справедливо равенство $p(L|\mathscr{S}) = (L, C_*)$. Фиксируем матрицы $L \in \mathbf{R}^{m \times n}$ и $Q = -L$. Из условия (10) найдем $\alpha_k \geqq 0$, для которых $\sum_{k \in I} \alpha_k \bar{\xi}_k x_k^T = Q$. Для $k \notin I$ полагаем $\alpha_k = 0$. Пусть $\lambda_k \triangleq -\alpha_k \bar{\xi}_k$, $k \in \overline{1, N}$. Тогда $\sum_{k=1}^{N} \lambda_k x_k^T = L$. Учитывая (4), (9) и то, что для указанных множеств Ξ_k верно равенство $\rho(-\lambda_k|\Xi_k) = \|\lambda_k\| \cdot \nu_k$, получаем

$$\rho(L|\mathscr{S}) \leqq \sum_{k=1}^{N} \lambda_k^T y_k + \sum_{k=1}^{N} \|\lambda_k\| \nu_k =$$

$$= \sum_{k=1}^{N} (-\alpha_k \bar{\xi}_k x_k^T, C_*) + \sum_{k=1}^{N} (-\alpha_k \bar{\xi}_k^T \bar{\xi}_k) + \sum_{k=1}^{N} \|\alpha_k \bar{\xi}_k\| \cdot \nu_k = (L, C_*) .$$

Но $C_* \in \mathscr{S}$, поэтому следствие доказано.

Примеры

Приведем примеры помех и входных воздействий, для которых формула (4) позволяет однозначно идентифицировать матрицу C_*.

Во всех примерах ограничения на помеху выберем в следующем виде:

$$\Xi_k \equiv \Xi \triangleq \{\xi : \|\xi\| \leqq 1\} .$$

Кроме того, будем считать, что выполнено условие (9) для всех $l = \overline{1, N}$.

Пример 1. Пусть выполнены условия:

1) входное воздействие x_k периодическое с периодом $\mathbf{R} \geqq n+1$,
2) $\mathscr{L}^+\{x_r\}_{r=1}^R = \mathbf{R}^n$,
3) помеха $\bar{\xi}_k$ периодическая с периодом $P \geqq m$,
4) $\mathscr{L}\{\bar{\xi}_p\}_{p=1}^P = \mathbf{R}^m$
5) P и R взаимно просты и $N = P \cdot R$.

Тогда формула (4) позволяет однозначно идентифицировать матрицу C_*.

Из условий 1), 3), 5) следует, что существует взаимнооднозначное соответствие между парами (r, p), $r \in \overline{1, R}$, $p \in \overline{1, P}$ и номерами $k \in \overline{1, N}$ такое, что $x_k = x_r$, $\bar{\xi}_k = \bar{\xi}_p$. Для доказательства этого факта фиксируем $k \in \overline{1, N}$ и находим r и p такие, что $x_k = x_r$, $\bar{\xi}_k = \bar{\xi}_p$, т. е. $k = r + \alpha_1 R$, $k = p + \beta_1 P$. Пусть l — номер, для которого $x_l = x_r$, $\bar{\xi}_l = \bar{\xi}_p$, $l \in \overline{1, N}$, т.е. $l = r + \alpha_2 R$, $l = p + \beta_2 P$. Не ограничивая общности, считаем $l \geqq k$, тогда

$$l - K = (\alpha_2 - \alpha_1) R ,$$
$$l - k = (\beta_2 - \beta_1) P .$$

Следовательно, разность $l - k$ делится на N нацело. Но $l - k < N$, поэтому $l = k$. Таким образом, каждому номеру k соответствует по крайней мере одна пара (r, p), каждой паре (r, p) соответствует не более одного номера k, а число пар совпадает с числом номеров. Значит имеет место указанное взаимнооднозначное соответствие. Его будем обозначать через $k = k(r, p)$.

Фиксируем $L \in \mathbf{R}^{m \times n}$. Из условия 4) следует, что найдется набор векторов $\{z_p : z_p \in \mathbf{R}^n\}_{p=1}^{P}$ таких, что

$$\sum_{p=1}^{P} \xi_p z_p^T = L .$$

Из установленного соответствия и условий 1), 2) следует, что для любого $p \in \overline{1, P}$ найдутся неотрицательные скаляры $\beta_{k(r,p)}$, для которых

$$\sum_{r=1}^{R} \beta_{k(r,p)} x_{k(r,p)} = z_P .$$

Таким образом, имеем

$$L = \sum_{p=1}^{P} \xi_p \sum_{r=1}^{R} \beta_{k(r,p)} x_{k(r,p)} = \sum_{k=1}^{N} \xi_k \beta_k x_k^T$$

и условие (10) выполнено.

Пример 2. Пусть для $N = 2n$ выполнены условия:

1) $\xi_k \equiv \xi \in \partial \Xi$, $k \in \overline{1, N}$,
2) множество входов есть $\{x_1, \ldots, x_n, -x_1, \ldots, -x_n\}$,
3) $\mathscr{L}\{x_k\}_{k=1}^{n} = \mathbf{R}^n$.

Тогда формула (4) обеспечивает точную идентификацию матрицы C_*.

Заметим, что в данном примере интервал наблюдения не зависит от числа m — размерности вектора наблюдения, причем при $m \geqq 2$ условия следствия 2 не выполнены.

Для доказательства при фиксированных $L \in \mathbf{R}^{m \times n}$ и $\varepsilon > 0$ выберем $\lambda_k \in \mathbf{R}^m$ и $\alpha_k \geqq 0$ так, чтобы выполнялось условие (7) и укажем оценку типа $K \cdot \varepsilon$ выражения $\Sigma \alpha_k \lambda_k^T (\xi_k - \psi_k)$, где вектор $\psi_k \in U_k(\varepsilon)$ соответствует λ_k в силу определения $\Lambda_k(\varepsilon)$.

Фиксируем $L \in \mathbf{R}^{m \times n}$ и $\varepsilon > 0$. По условию 3) найдется набор векторов $\{z_k : z_k \in \mathbf{R}^m\}_{k=1}^{n}$ таких, что

$$\sum_{k=1}^{n} z_k x_k^T = L .$$

Для каждого $k \in \overline{1, n}$ выберем величины λ_k, λ_{n+k}, α_k, α_{n+k}, ψ_k, ψ_{n+k} следующим образом:

а) если $|\xi^T z_k| = \|z_k\|$, то $\lambda_k = \lambda_{n+k} = -\xi$; если $z_k = 0$, то $\alpha_k = \alpha_{n+k} = 0$; если $z_k = -\xi \cdot \|z_k\|$, то $\alpha_k = \|z_k\|$, $\alpha_{n+k} = 0$; если $z_k = \xi \cdot \|z_k\|$, то $\alpha_k = 0$, $\alpha_{n+k} = \|z_k\|$. Заметим, что выбранные таким образом величины не зависят от ε;

б) если $|\xi^T z_k| < \|z_k\|$, то пересечение границы Ξ_k и плоскости, проходящей через начало координат и точки ξ, z_k, обозначим через P_k. P_k есть окружность, которой принадлежат точки $\bar{\xi}$, $v_k \triangleq z_k \cdot \|z_k\|^{-1}$. Не ограничивая общности, считаем $\varepsilon < \|z_k \cdot \|z_k\|^{-1} - \bar{\xi}\|$. Тогда на P_k на расстоянии ε от точки $\bar{\xi}$ найдется точка ψ_{n+k} и скаляры γ_k и β_k такие, что

$$\|\psi_{n+k} - \xi\| = \varepsilon\,,$$

$$0 < \gamma_k, \beta_k < \frac{1}{\sin \varphi_{n+k}}\,,$$

$$z_k \cdot \|z_k\|^{-1} = \gamma_k(-\bar{\xi}) + \beta_k \psi_{n+k}\,,$$

где φ_k — угол между $\bar{\xi}$ и ψ_{n+k}. Полагаем $\alpha_k = \gamma_k \cdot \|z_k\|$, $\alpha_{n+k} = \beta_k \cdot \|z_k\|$, $\lambda_k = -\bar{\xi}$, $\lambda_{n+k} = -\psi_{n+k}$.

В случае а) и б), когда $\lambda_k = -\bar{\xi}$, $k \in \overline{1, N}$, полагаем $\psi_k = \bar{\xi}$, $\varphi_k = \varphi$, $0 < \varphi < \dfrac{\pi}{2}$. Для $k \in \overline{1, n}$ полагаем $z_{n+k} = z_k$.

Для указанных α_k, λ_k, ψ_k, $k \in \overline{1, N}$, имеем

$$\sum_{k=1}^{N} \alpha_k \lambda_k x_k^T = L\,,$$

$$\rho(-\alpha_k \lambda_k | \Xi) = -\alpha_k \lambda_k \psi_k\,,$$

причем $\lambda_k \in \Lambda_k(\varepsilon)$ и $\alpha_k \leqq \dfrac{\|z_k\|}{\sin \varphi_k}$. Кроме того, $0 \leqq \lambda_k^T(\bar{\xi} - \psi_k) < \varepsilon \cdot \sin \varphi_k$, поэтому

$$\sum_{k=1}^{N} \alpha_k \lambda_k^T(\bar{\xi} - \psi_k) < \sum_{k=1}^{N} \frac{\|z_k\|}{\sin \varphi_k} \varepsilon \sin \varphi_k = \sum_{k=1}^{N} \|z_k\| \cdot \varepsilon\,,$$

где $\sum\limits_{k=1}^{N} \|z_k\|$ не зависит от ε.

Таким образом, условия следствия 1 выполнены.

Пример точной идентификации при случайных помехах.

Вышеприведенные результаты могут быть применены в случае, когда помеха ξ_k в уравнении (1) имеет вероятностную природу. Для иллюстрации приведем

Пример 3. Пусть выполнены условия:

1) $\{\xi_k\}_{k=1}^{\infty}$ — реализация последовательности независимых одинаково распределенных случайных величин, плотность распределения которых не равна 0 в любой точке Ξ;

2) входное воздействие x_k периодическое с периодом $R \geqq n$;

3) $\mathscr{L}\{x_r\}_{r=1}^{R} = \mathbf{R}^n$.

Тогда имеет место равенство

$$\mathbf{P}\left[\lim_{N\to\infty} \mathscr{S}_N = \{C_*\}\right] = 1\,. \tag{11}$$

Здесь $\mathbf{P}[A]$ означает вероятность события A, $\mathscr{S}_N$ — информационное множество для $k \in \overline{1, N}$, $\lim\limits_{N\to\infty} \mathscr{S}_N = \{C_*\}$ означает сходимость в метрике Хаусдорфа [1].

Фиксируем набор векторов $\psi_p \in \partial\Xi$, $p \in \overline{1, P}$ со свойством $\mathscr{L}^+\{\psi_p\}_{p=1}^P = \mathbf{R}^m$. Не ограничивая общности, считаем, что $R = n$, $P = m + 1$.

Фиксируем ε, $0 < \varepsilon < \dfrac{1}{2} \min\limits_{\substack{t,\, p \in \overline{1,\, P} \\ t \neq p}} \{\|\psi_p - \psi_t\|\}$. По лемме Бореля–Кантелли для любой точки $\psi_p \Lambda$ любого номера $r \in \overline{1, R}$ с вероятностью 1 найдется номер $k = k(r, p, \varepsilon)$ со свойством: $\|\psi_p - \bar{\xi}_k\| \leqq \varepsilon$ и $x_k = x_r$. Далее будем рассматривать для каждой тройки $(r,\ p,\ \varepsilon)$ наименьший номер k с таким свойством. При этом $\psi_p \in U_k(\varepsilon)$, а вектор λ_k из $\Lambda_k(\varepsilon)$, соответствующий точке ψ_p, есть $\lambda_k = -\psi_p$. Множество указанных индексов $k = k(r, p, \varepsilon)$ при $r \in \overline{1, R}$, $p \in \overline{1, P}$ обозначим через $I(\varepsilon)$ и обозначим $N(\varepsilon) \triangleq \max\{k : k \in I(\varepsilon)\}$.

Фиксируем $L \in \mathbf{R}^{m \times n}$. В силу условий 2), 3) выберем набор векторов $z_r \in \mathbf{R}^m$, чтобы $\sum\limits_{r=1}^{R} z_r x_r^T = -L$. Для каждого вектора z_r имеет место разложение $z_r = \sum\limits_{p=1}^{P} \beta_{rp}\psi_p$, где коэффициенты β_{rp} не зависят от ε.

Введем обозначение $\alpha_{k(r,\, p,\, \varepsilon)} = \beta_{rp}$. Тогда с вероятностью 1 имеем

$$\sum_{k \in I(\varepsilon)} \alpha_k \lambda_k x_k^T = \sum_{r=1}^{R} \sum_{p=1}^{P} (-\beta_{rp}) \cdot \psi_p x_r^T = L\,,$$

$$\left| \sum_{k \in I(\varepsilon)} \alpha_k \lambda_k^T (\bar{\xi}_k - \psi_p) \right| \leqq \sum_{r=1}^{R} \sum_{p=1}^{P} \beta_{rp}\varepsilon\,,$$

где $B \triangleq \sum\limits_{r=1}^{R} \sum\limits_{p=1}^{P} \beta_{rp}$ не зависит от ε.

Рассмотрим множества

$$\mathscr{S}_N^* \triangleq \{c : c = c_1 - c_*,\, c_1 \in \mathscr{S}_N\}.$$

При всех N матрица $C_* \in \mathscr{S}_N$, поэтому

$$0 \leqq \rho(L | \mathscr{S}^*_{N(\varepsilon)}) \leqq \sum_{k \in I(\varepsilon)} \alpha_k \lambda_k^T y_k + \sum_{k \in I(\varepsilon)} \rho(-\alpha_k \lambda_k | \Xi) - (L, C_*) =$$

$$= \sum_{k \in I(\varepsilon)} \alpha_k \lambda_k^T (C_* x_k + \bar{\xi}_k) - \sum_{k \in I(\varepsilon)} \alpha_k \lambda_k^T \psi_p - (L, C_*) =$$

$$= \sum_{k \in I(\varepsilon)} \alpha_k \lambda_k^T (\bar{\xi}_k - \psi_p) \leqq B \cdot \varepsilon .$$

То есть с вероятностью 1

$$\lim_{\varepsilon \to 0} \rho(L | \mathscr{S}^*_{N(\varepsilon)}) = 0 . \tag{12}$$

Поскольку равенство (12) имеет место для произвольной матрицы $L \in \mathbf{R}^{m \times n}$, это означает, что выполнено равенство (11).

Литература

1. *Куржанский А. Б.* Управление и наблюдение в условиях неопределенности. М., Наука, 1977, 392 с.
2. *Рокафеллар Р.* Выпуклый анализ. М., Мир, 1973, 470 с.

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda, Budapest

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

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). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as diag $(d_1, d_2, \ldots d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 129090 Москва И-90, ул. Щепкина, 8. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 208-60-19).

Объём 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 языке в трех экземплярах, 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 и 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 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

HU ISSN 0370-2529 Index: 26 660

VOL. 15 · NUMBER 5
ТОМ 15 · НОМЕР 5

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 СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1986

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
— theory of adaptive systems
— theory of estimation and identification
— theory of controlling robot-technologic and flexible manufacturing systems
— information theory
— information-theoretic aspects of multiple access networks.

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;
— теория адаптивных систем;
— теория оценивания и идентификации;
— теория управления робототехническими и гибкими производственными системами;
— теория информации;
— теория информации в области сетей с множественным доступом.

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 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1986 Subscription Rate US Dollars 175.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

Problems of Control and Information Theory, Vol. 15 (5), pp. 333—343 (1986)

DECOMPOSITION OF THE PROCESS OF CONTROL IN USING ALGORITHMS WITH FORECASTING MODEL

V. N. Bukov, V. F. Chernov, B. G. Chudinova

(*Moscow*)

(Received May 11, 1985)

The task of control of complex multivariable object is considered. In order to form optimum controls the forecasting model with sensibility matrix is used. A disparallelled algorithm as a result of arbitrary division of the initial object into blocks is obtained. Algorithms of control of blocks with various degrees of interaction are given. As an example, control of motion of the cabin of dynamic simulator is obtained.

1. Introduction

One of the characteristic features of modern level of development of the control of dynamic process theory is integration of control systems which solve particular tasks. It is suggested that integration of these systems would provide potential opportunities to improve characteristics of the control system on the whole.

The task of integrated control inevitably faces the problem of control of multivariable processes which can be interconnected. A great number of equations can be a serious obstacle on the way of practical realization of algorithms.

Methods of distributed processing of information can cut the necessity of transmitting all data into one processor and can enable to distribute computer loading between several processors. The main variants of decomposition of control are as follows: hierarchical control in which processors are united in functional hierarchy; decentralized control in which processors interact on the same level. Various combinations of these variants are possible. On the whole, organization of information exchange and calculations in multi-computer control system is closely linked with peculiarities of algorithms of control. Working out the control system requires special algorithms designed to solve this task.

This paper is devoted to the problems of decomposition of control process of non-linear object when the control is formed on the basis of algorithms with forecasting model.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

In [1] it is shown that for the object

$$\dot{x} = F(x, \delta, t), \qquad \dot{\delta} = u \tag{1}$$

the optimum in the sense of minimum of functional

$$I = V_3[x(t_k)] + \int_{t_0}^{t_k} Q(\tau)\, d\tau + \frac{1}{2} \int_{t_0}^{t_k} (u'k^{-1}u + u'_{\text{opt}}k^{-1}u_{\text{opt}})\, d\tau \tag{2}$$

is the equation

$$\begin{aligned} u_{\text{opt}}(t) = -k \Bigg[& Z'(t_k) \frac{\partial V'_3(t_k)}{\partial x_M} + \frac{\partial V'_3(t_k)}{\partial \delta_M} + \\ & + \int_{t_0}^{t_k} \left(Z'(\tau) \frac{\partial Q'(\tau)}{\partial x_M} + \frac{\partial Q'(\tau)}{\partial \delta_M} \right) d\tau \Bigg], \end{aligned} \tag{3}$$

where Z stands for the sensibility matrix $\partial x / \partial \delta$ wich accords equation

$$\frac{d}{d\tau} Z = \frac{\partial F}{\partial x_M} \cdot Z + \frac{\partial F}{\partial \delta_M} \tag{4}$$

with initial condition $Z = 0$; calculations (3) and (4) are performed on the prognosed states of the object

$$\frac{d}{d\tau} x_M = F(x_M, \delta_M, t), \qquad \frac{d}{d\tau} \delta_M = 0, \tag{5}$$

on the interval $[t_0, t_k]$ with initial conditions $x_M(t_0) = x(t_0)$, $\delta_M(t_0) = \delta(t_0)$. Here x is an n-dimensional vector of the state of the object; δ is an m-dimensional vector of the position of the object rudders; F is an n-dimensional vectoral differentiated function; V and Q are non-negative definite scalar functions; $k = \text{diag}\,(k_1, \ldots, k_m)$ is a diagonal matrix of assigned coefficients; sign $'$ denotes transposition.

2. Arbitrary division of objects into blocks

In many applied tasks of equation (5) division of equations into blocks with this or that level of cross ties is permitted. This division, as a rule, should be based either on the analysis of the task or on the basis of experience of working out control systems for the given class of objects. However, for the time being we shall consider that division of equation (5) into blocks is arbitrary.

We introduce the following symbols: x_i — state vector of the i-th block of the object (subvector of state); F_i — vector of the right-hand side of the equations of i-block of the object; δ_i — vector of positions of rudders of i-block of the object; u_i — control of speed of changes of rudder-positions of i-block of the object to be defined. With further consideration that we deal only with two blocks (indices i and j, respectively) we shall write equation (5) as:

$$\begin{aligned} \dot{x}_i &= F_i(x_i, x_j, \delta_i, \delta_j, t), \qquad \dot{\delta}_i = u_i, \\ \dot{x}_j &= F_j(x_i, x_j, \delta_i, \delta_j, t), \qquad \dot{\delta}_j = u_j. \end{aligned} \tag{6}$$

It should be noted that generalization for a greater number of blocks is obtained without considerable difficulties.

In this case when $k = \operatorname{diag}(k_1, \ldots, k_m)$, we have

$$\begin{aligned} u_i(t) = -k_i \Bigg[& Z'_{ii}(t_k) \frac{\partial V'_3(t_k)}{\partial x_i} + Z'_{ji}(t_k) \frac{\partial V'_3(t_k)}{\partial x_j} + \\ & + \frac{\partial V'_3(t_k)}{\partial \delta_i} + \int_t^{t_k} \left(Z'_{ii} \frac{\partial Q'}{\partial x_i} + Z'_{ji} \frac{\partial Q'}{\partial x_j} + \frac{\partial Q'}{\partial \delta_i} \right) d\tau \Bigg]. \end{aligned} \tag{7}$$

Here $Z_{ii} = \partial x_i / \partial \delta_i$ is a sensibility matrix of i-block of the object in regard to rudder positions of the same block; $Z_{ji} = \partial x_i / \partial \delta_j$ is a sensibility matrix of i-block in regard to rudder positions of j-block. Besides, for the sake of brevity, in (7) we omitted indices M which reflect the usage of states and controls of models of blocks of the object. We hope that this reservation will help avoid confusion.

So, (7) is calculated on the states which are modelled on $[t, t_k]$ interval with the help of equations

$$\begin{aligned} \frac{d}{d\tau} x_i &= F_i(x_i, x_j, \delta_i, \delta_j), \qquad \dot{\delta}_i = 0, \\ \frac{d}{d\tau} x_j &= F_j(x_i, x_j, \delta_i, \delta_j), \qquad \dot{\delta}_j = 0 \end{aligned} \tag{8}$$

with initial conditions corresponding to the current state of the object. Sensibility matrices are defined by means of integrating equations on the same modelled states

$$\begin{aligned} \frac{d}{d\tau} Z_{ii} &= \frac{\partial F_i}{\partial x_i} Z_{ii} + \frac{\partial F_i}{\partial x_j} Z_{ji} + \frac{\partial F_i}{\partial \delta_i}, \\ \frac{d}{d\tau} Z_{ji} &= \frac{\partial F_j}{\partial x_i} Z_{ii} + \frac{\partial F_j}{\partial x_j} Z_{ji} + \frac{\partial F_j}{\partial \delta_i}, \end{aligned} \tag{9}$$

with initial conditions $Z_{ii} = 0$, $Z_{ji} = 0$, where $\partial F_i / \partial x_i$ is a Jacobi matrix of i-block of the object; $\partial F_i / \partial x_j$ is a matrix which defines the influence of the state of j-block on the processes of i-block; matrices $\partial F_i/\partial \delta_i$ and $\partial F_j/\partial \delta_i$ define the influence of rudders of i-block on the processes of i-blocks, respectively.

By analogy, it is possible to put down equations to control j-block of object (6). In Fig. 1 the structure of algorithm (7)–(9) for the case $V_3 = 0$ is shown. Two channels are clearly expressed here. Interconnection of channels is manisfested, firstly, in exchange of prognosed states $x_i(\tau)$ and $x_j(\tau)$, and, secondly, in simultaneous calculation of partial derivative of Q-function. In concrete cases, the latter connection can be insignificant. If the total number of blocks is more than two, then in (7), (9) and in Fig. 1 we should foresee the summing up of constituents linked with motion (change of state) of all "external" blocks.

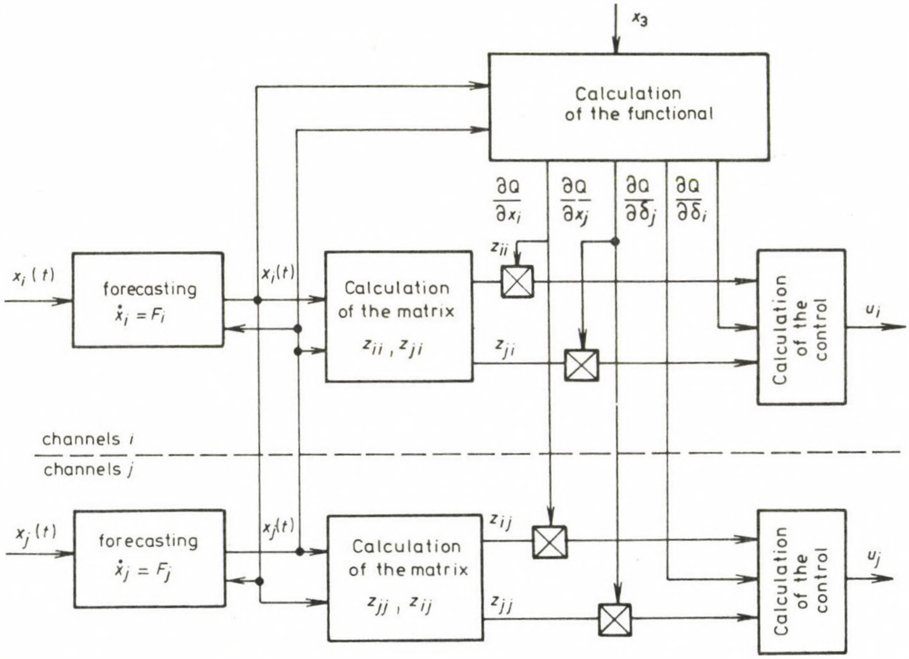

x_3
Calculation of the functional
$\frac{\partial Q}{\partial x_i}$
$\frac{\partial Q}{\partial x_j}$
$\frac{\partial Q}{\partial \delta_j}$
$\frac{\partial Q}{\partial \delta_i}$
z_{ii}
$x_i(t)$
forecasting $\dot{x}_i = F_i$
$x_i(t)$
Calculation of the matrix z_{ii}, z_{ji}
z_{ji}
Calculation of the control
u_i
channels i
channels j
$x_j(t)$
forecasting $\dot{x}_j = F_j$
$x_j(t)$
Calculation of the matrix z_{jj}, z_{ij}
z_{ij}
z_{jj}
Calculation of the control
u_j

Fig. 1

3. Some particular cases

Algorithm (7)–(9) corresponds to the general case and consists in integrating the total volume of initial differential equations of algorithm (3)–(5). Computer loading can be lowered only by means of disparallelling calculations between blocks with cross exchange of information. Possibilities of a further simplification of the algorithm are connected with corresponding division of the initial object into blocks. Let us consider important particular cases.

I. *Autonomous control.* Any of the channels (for example, i-channel) of algorithm (7)–(9) is autonomous in forming control in two cases:

a) vector-valued functions F_j of other blocks of object (6) do not depend on the state and position of rudders of i-block;

b) vectorial function F_i of the given block and scalar functions V_3 and Q of the functional do not depend on state X_j of other blocks.

In the first case $\partial F_j/\partial\delta_i=0$, $\partial F_j/\partial x_i=0$ and, as we see in (9), control for Z_{ji} has zero solution, and algorithm of control looks as

$$u_i(t)=-k_i\left[Z'_{ii}\frac{\partial V'_3(t_k)}{\partial x_i}+\frac{\partial V'_3(t_k)}{\partial\delta_i}+\int\limits_t^{t_k}\left(Z'_{ii}\frac{\partial Q'}{\partial x_i}+\frac{\partial Q'}{\partial\delta_i}\right)d\tau\right], \tag{10}$$

$$\begin{aligned}&\frac{d}{d\tau}x_i=F_i(x_i, x_j, \delta_i, \delta_j), &&\frac{d}{d\tau}\delta_i=0,\\ &\frac{d}{d\tau}x_j=F_j(x_j, \delta_j), &&\frac{d}{d\tau}\delta_j=0,\end{aligned} \tag{11}$$

$$\frac{d}{d\tau}Z_{ii}=\frac{\partial F_i}{\partial x_i}Z_{ii}+\frac{\partial F_i}{\partial\delta_i}. \tag{12}$$

Except of symbols (10) and (12) here coincide with (3) and (4), that is, determine control of isolated object (block). Equation (11) takes into consideration the influence of other blocks as sources of disturbance on the state of the controlled block.

In the second case, we obtain: $\partial F_i/\partial x_j=0$, $\partial Q/\partial x_j=0$, $\partial V_3/\partial x_j=0$, as a result of which the first equation does not depend on the solution of the second equation and correlation (7) does not contain components with matrix Z_{ji}. The algorithm is represented by formula (10) and by equations

$$\frac{d}{d\tau}x_i=F_i(x_i, \delta_i, \delta_j), \qquad \frac{d}{d\tau}\delta_i=0 \tag{13}$$

and by equation (12). Equation (13) takes into consideration the influence of the position of rudders of other blocks on the state of controlled object.

II. *Hierachical control* (control of block which has no rudders of its own). As a result of decomposition equations of the object

$$\dot{x}_i = F_i(x_i, \delta_i), \qquad \dot{\delta}_i = u_i, \qquad \dot{x}_j = F_j(x_i, x_j) \tag{14}$$

is obtained. Then $\partial F_j/\partial\delta_j = 0$, $\partial F_i/\partial\delta_j = 0$. In this case, control of i-block will be determined by means of formula (7), which is calculated with the help of solutions of equations

$$\begin{aligned} &\frac{d}{d\tau} x_i = F_i(x_i, \delta_i), \qquad \frac{d}{d\tau}\delta_i = 0, \\ &\frac{d}{d\tau} x_j = F_j(x_i, x_j), \\ &\frac{d}{d\tau} Z_{ii} = \frac{\partial F_i}{\partial x_i} Z_{ii} + \frac{\partial F_i}{\partial \delta_i}, \\ &\frac{d}{d\tau} Z_{ji} = \frac{\partial F_j}{\partial x_i} Z_{ii} + \frac{\partial F_j}{\partial x_j} Z_{ji}. \end{aligned} \tag{15}$$

In the case of dependence of functions V_3 and Q on the state of j-block the solution of the last equation provides for (7) a consideration of requirements set for the motion of this block in forming control of i-block.

4. Heuristic modifications of algorithm

In paper [2] a variant of forming hierarchical control of object (14) based on division of tasks is treated. We can this approach to make clearer in the following way.

Let us introduce an additional vector of state x_{i*} which will reflect certain prescribed state of i-block. In this case x_{i*} can fail coincide with x_i precisely, however, the number of tasks of control will include the transformation of x_i into $x_{i*}$with sufficient precision and quality. We can suggest that it satisfies equation

$$\dot{x}_{i*} = u_{j*}. \tag{16}$$

Here u_{j*} is vector of additional fictitious controls (pseudocontrols), whose dimension coincides with the dimension of vector x_{i*}. To ensure the process when x_i follows x_{i*} in minimized functional, we additionally introduce positively defined functions $V_{3*}(x_i - x_{i*})$ and $Q_*(x_i - x_{i*})$, which are *summed up* with the corresponding functions V_3 and Q. Besides, into this functional we introduce corresponding constituents with vectors u_{j*} and $u_{j*\text{opt}}$ and matrix K_{j*}.

Further we suggest that in equation of j-block in (14) we have not a true but an introduced vector of state. Then, solving the task of control of object we should consider equations

$$\begin{aligned} \dot{x}_i &= F_i(x_i, \delta_i), \qquad \dot{\delta}_i = u_i, \\ \dot{x}_j &= F_j(x_{i*}, x_j), \quad \dot{x}_{i*} = u_{j*}. \end{aligned} \tag{17}$$

Formally, blocks of this object are mutually independent. Connection between them exists only through a functional which is minimized in the process of control. If we apply algorithm (7)–(9) to object (17) with consideration of change of symbols, zero solution of the second equation (9) with $\partial F_j/\partial x_i = \partial F_j/\partial \delta_i = 0$, zero solution of equation for Z_{ij} in an analogous case and in presence of additional components in the functional, we obtain the following for i-block

$$\begin{aligned} \dot{\delta}_i = u_i = -k_i\Bigg[z'_{ii}(t_k)\left(\frac{\partial V_3(t_k)}{\partial x_i} + \frac{\partial V_{3*}(t_k)}{\partial x_i}\right) + \frac{\partial V'_3(t_k)}{\partial \delta_i} + \\ + \int_t^{t_k} \left(Z'_{ii}\left(\frac{\partial Q'}{\partial x_i} + \frac{\partial Q'_*}{\partial x_i}\right) + \frac{\partial Q'}{\partial \delta_i}\right) d\tau \Bigg], \\ \frac{d}{d\tau} x_i = F_i(x_i, \delta_i), \qquad \frac{d}{d\tau}\delta_i = 0, \\ \frac{d}{d\tau} Z_{ii} = \frac{\partial F_i}{\partial x_i} Z_{ii} + \frac{\partial F_i}{\partial \delta_i}, \end{aligned} \tag{18}$$

and for j-block we obtain:

$$\dot{x}_{i*} = u_{j*} = -k_{j*}\left[Z'_{jj}(t_k)\frac{\partial V'_3(t_k)}{\partial x_j} + \frac{\partial V'_{3*}(t_k)}{\partial x_{i*}} + \int_t^{t_k} \left(Z'_{jj}\frac{\partial Q'}{\partial x_j} + \frac{\partial Q'_*}{\partial x_{i*}}\right) d\tau \right], \tag{19}$$

$$\frac{d}{d\tau} x_j = F_j(x_{i*}, x_j), \qquad \frac{d}{d\tau} x_{i*} = 0,$$

$$\frac{d}{d\tau} Z_{jj} = \frac{\partial F_j}{\partial x_j} Z_{jj} + \frac{\partial F_j}{\partial x_{i*}}.$$

5. Control of the cabin of centrifuge

Equation of motion of the cabin of centrifuge which is used as a simulator together with positioning elements looks as:

$$\dot{x}_l = F_l, \tag{20}$$

where

$$F_1 = k_1 \cdot \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} \left(k_0 \left(ky - k_4 x_2 - k_3 \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} x_1 \right) \right),$$

$$F_2 = k_2 \cdot \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} x_1,$$

$$F_3 = k_2 \cdot \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} x_1 - k_{15} x_6 - k_8 \mathop{\lceil\!\!\rfloor}\limits_{B_2}^{B_2} k_7 x_3,$$

$$F_4 = D \left(k_{10} \cdot 1(R + B_1) - 1(R - B_1) R - k_{13} \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} x_4 \right) + D_5 \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} R - D_7 x_5,$$

$$R = \mathop{\lceil\!\!\rfloor}\limits_{B_2}^{B_2} k_7 x_3 + k_9 (x_2 - k_{15} x_5),$$

$$F_5 = x_6,$$

$$F_6 = D_1 \mathop{\lceil\!\!\rfloor}\limits_{B_1}^{B_1} x_4 + \frac{md}{J} (l \cdot \omega^2 \cdot \cos x_5 - g \cdot \sin x_5).$$

Here $\lceil\!\!\rfloor$ is a symbol of limitation above and below the variable value which stands to the right of the sign; $1(\cdot)$ is the unit step function obtaining zero value when independent variable is negative; $K_v, D_\mu, B_j, m, d, J, l$ are constant parameters reflecting dynamic and static properties of the object of the control; ω is the rotation speed of the cantilever of centrifuge on which the cabin is fixed; g is the acceleration of gravitation; y is the input signal corresponding to the current of control; x_5 is the turn angle of the cabin.

We shall consider the formation of the speed of change of input signal $\dot{y} = u$, which is optimum for (2) with zero subintegral function Q and terminal function

$$V_3[x(t_k)] = \beta_1 [x_4(t_k) - x_{43}]^2 + \beta_2 [x_5(t_k) - x_{53}]^2 + \beta_3 [x_6(t_k) - x_{63}]^2, \tag{21}$$

where x_{43}, x_{53}, x_{63} are the required states of the object which are put into control system with frequency of 10 Hz.

Application of (3) to (20), (21) results in

$$\begin{aligned} u_{\text{iopt}} = -2\{ & \beta_1 Z_4(t_k) [x_4^M(t_k) - x_{43}] + \beta_2 Z_5(t_k) [x_5^M(t_k) - x_{53}] + \\ & + \beta_3 Z_6(t_k) [x_6^M(t_k) - x_{63}] \}, \end{aligned} \tag{22}$$

where $x_j^M(t_k)$ and $Z_j(t_k)$ are solutions for forecasting model (5) and for control (4) in the case when initial conditions are considered.

We decompose the process of control into blocks on analogy with (14):

— switchboard (first pair of equations (20))

— cabin dynamics (third pair of equations (21))

— drive (second pair of equations (20)).

We single out the block of cabin dynamics "heuristically", introducing additional coordinate $y_* = \int_{B_3}^{B_3} x_4$ and considering control $\dot{y}_* = u_*$. Then the forecasting model falls into two formally independent blocks.

Block I

$$\begin{aligned} &\dot{x}_1^M = F_1(x_1^M, x_2^M, y), \quad \dot{y}=0, \quad \dot{y}=0, \quad && \dot{Z}_1 = \frac{\partial F_1}{\partial x_1} Z_1 + \frac{\partial F_1}{\partial x_2} Z_2 + \frac{\partial F_1}{\partial y}, \\ &\dot{x}_2^M = F_2(x_1^M), && \dot{Z}_2 = \frac{\partial F_2}{\partial x_1} Z_1 . \end{aligned} \tag{23}$$

Block II

$$\begin{aligned} &\dot{x}_5^M = x_6^M, && \dot{Z}_5 = Z_6, \\ &\dot{x}_6^M = D_1 y_* + F(x_5^M), && \dot{Z}_6 = \frac{\partial F}{\partial x_5} Z_5 + D_1, \quad \dot{y}_* = 0 \end{aligned} \tag{24}$$

and a block having no controls of its own,

Block III

$$\begin{aligned} &\dot{x}_3^M = F_3(x_3^M, x_1^M, x_6^M), && \dot{Z}_3 = \frac{\partial F_3}{\partial x_3} Z_3 + f_{30}, \\ &\dot{x}_4^M = F_4(x_3^M, x_4^M, x_1^M, x_2^M, x_5^M, x_6^M), && \dot{Z}_4 = \frac{\partial F_4}{\partial x_3} Z_3 + \frac{\partial F_4}{\partial x_4} Z_4 + f_{40}, \end{aligned} \tag{25}$$

where

$$f_{30} = \frac{\partial F_3}{\partial x_1} Z_1 + \frac{\partial F_3}{\partial x_6} Z_6, \qquad f_{40} = \frac{\partial F_4}{\partial x_1} Z_1 + \frac{\partial F_4}{\partial x_2} Z_2 + \frac{\partial F_4}{\partial x_5} Z_5 + \frac{\partial F_4}{\partial x_6} Z_6 .$$

An additional term, $\beta_* \left(\int_{B_3}^{B_3} x_4 - y_* \right)^2$, is joined to the main part of functional (21). Then main control should be calculated in accordance with (18) by means of the following formula

$$u_i = u_{i\text{opt}} - 2\beta_* Z_4(t_k) \left(\int_{B_3}^{B_3} x_4(t_k) - y_* \right)^2 \cdot (1(x_4(t_k) + B_3) - 1(x_4(t_k) - B_3)) . \tag{26}$$

The results of digital modelling of total algorithm and algorithm with decomposition of control system are presented in Fig. 2. The continuous curve shows the required change of the turn angle of the cabin x_{53}; the dashed line shows the turn angle of the cabin controlled with the help of total algorithm; the dashed-and-dotted line shows the turn angle of the cabin controlled with the help of algorithm with decomposition.

As it can be seen from the figure, the decomposition of the control system practically does not decrease the precision of the operation of the control system in comparison with total algorithm.

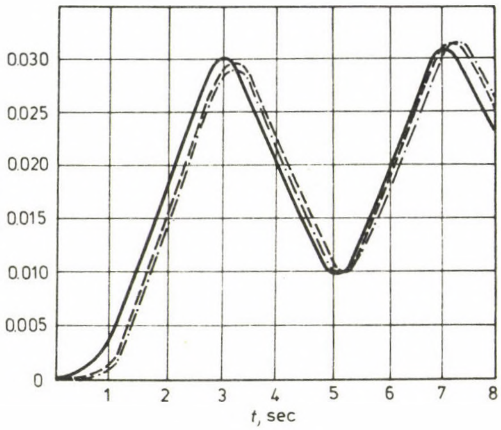


Fig. 2

References

1. *Bukov, V. N.*, Synthesis of controls through a forecasting model in adaptive control system. Problems of Control and Information Theory, **9** (5), 1980, pp. 329–337.
2. *Krasovskii, A. A.*, Decomposition and synthesis of suboptimal systems. Publication of the Academy of Sciences of the USSR. Technical Cybernetics, *2*, 1984, pp. 157–158.

Декомпозиция процессов управления при использовании алгоритмов с прогнозирующей моделью

В. Н. БУКОВ, В. Ф. ЧЕРНОВ, В. Г. ЧУДИНОВА

(Москва)

Рассматривается задача управления сложным многомерным объектом. Для формирования оптимальных управлений используется алгоритм с прогнозирующей моделью и матрицей чувствительности. Получен распараллеленный алгоритм при произвольном разделении исходного объекта на блоки. Строятся алгоритмы управления блоками объекта при различной степени их автономности. Анализируются опубликованные ранее эвристические подходы к декомпозиции в задачах с прогнозирующими моделями. В качестве примера приводится управление движением кабины динамического тренажера.

В. Н. Буков
Научный совет по комплексной проблеме
«Кибернетика»
СССР, 117333 Москва В-333,
ул. Вавилова, 40

Problems of Control and Information Theory, Vol. 15 (5), pp. 345—365 (1986)

ROBUST STABILITY OF DECENTRALIZED OUTPUT CONTROL SYSTEMS WITH APPLICATION TO SINGULAR PERTURBATION THEORY*

D. B. PETKOVSKI

(*Novi Sad*)

(Received August 10, 1985)

In this paper we give a simple approach to determine conditions for stability of linear decentralized control systems subject to model uncertainties and large parameter variations in the system dynamics. As an example of the application of the results, we consider the determination of finite regions of stability for singularly perturbed systems.

1. Introduction

In recent years, there has been an increased interest in the development of satisfactory control design methods implemented in a decentralized way (see e.g. [11]). One of the most basic issues what arises in this class of problems is the robustness of the decentralized design, i.e. its ability to maintain stability and performance in the face of uncertainties. The importance of obtaining robustly stable feedback control systems has long been recognized by designers. The mathematical models of the systems to which they are concerned are idealized, inexactly identified, or the systems themselves are subject to unpredictable changes with time. This is of particular importance to modern control engineers whose assignment is to design highly sophisticated control systems on the basis of such mathematical models.

In this paper we consider the problem of robustness in decentralized control systems subject to model uncertainties and large parameter variations in the system dynamics. Although, in the area of robustness analysis for multivariable systems, a wide choice of measure of the "size" of the perturbations has been suggested, still one of the basic needs is for refine tests and measures of robustness. The measures suggested so far, lead to overly conservative results in many instances, since they characterize model

* This research was partially supported by the U.S.–Yugoslav Scientific and Technological Cooperation under Grant ENERGY-401.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

errors simply by their norms. To overcome these difficulties, we shall combine the information concerning the nature of the perturbations and the mathematical characterization of the perturbation matrices. It is obvious that this approach leads to a less conservative robustness characterization.

A specific instance of the robustness question arises when a system is approximated by making singular perturbation to reduce its order. Severel papers [3, 14, 16] have appeared in the literature on asymptotic stability of singularly perturbed systems, but contain no expression for calculating an upper bound for the perturbation parameter for which the perturbed system remains stable. As an example of the application of the results presented in this paper, we consider the determination of finite regions of stability for singularly perturbed systems. A relationship will be established between the perturbation bounds and the prescribed degree of stability of the nominal system, thus helping a designer to select an appropriate degree of stability to attain a robust design.

2. Decentralized feedback design with a prescribed degree of stability

In this section a brief discussion on an approach for decentralized feedback designs with prescribed degree of stability is given. The approach enables us to control the system by a set of controllers — each having different information and control variables. For more detailed discussion, see [9, 10, 19, 20].

2.1. Problem formulation

Consider a large-scale linear system

$$(S) \qquad \dot{x}(t) = Ax(t) + \sum_{i=1}^{k} B_i u_i(t), \qquad x(t_0) = x_0 \tag{2.1}$$

where $x(t) \in R^n$ is a state vector and $u_i(t) \in R^{m_i}$ is a control vector, $\sum_{i=1}^{k} m_i = m$. The information available to the local controller i is assumed to be

$$y_i(t) = C_i x(t) \tag{2.2}$$

where $y_i(t) \in R^{r_i}$ is a local output vector, $\sum_{i=1}^{k} r_i = r$. The local control u_i is assumed to be a direct feedback from the local output y_i, namely,

$$u_i(t) = \bar{E}_i y_i(t), \qquad i = 1, 2, \ldots, k \tag{2.3}$$

where E_i is a time-invariant gain matrix. The goal is to determine the decentralized control which ensures stability of the closed loop system with a prescribed degree α, $\alpha \in [0, \alpha_{\max})$.

2.2. Design procedure

The approach is based on computation of a complete state feedback and reduction to a specified control with a decentralized structure. We introduce the performance index

$$J = 1/2 \int_0^\infty e^{2\alpha t}(x^T Q x + u^T R u)\, dt \tag{2.4}$$

$$Q = Q^T \geqq 0, \quad R = \operatorname{diag}(R_i) > 0, \qquad i = 1, 2, \ldots, k \tag{2.5}$$

and we seek to determine the optimal control law which minimizes (2.4) subject to the dynamic constraints

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.6}$$

where $B = [B_1\ B_2\ \ldots\ B_k]$, $u^T = [u_1^T\ u_2^T\ \ldots\ u_k^T]$.

The solution is given by [1],

$$u(t) = R^{-1} B^T \hat{K} x(t) \tag{2.7}$$

where $\hat{K}$ is the positive definite solution of the Riccati algebraic equation

$$\hat{A}^T \hat{K} + \hat{K}\hat{A} - \hat{K}BR^{-1}B^T\hat{K} + Q = 0, \qquad \hat{A} = A + \alpha I. \tag{2.8}$$

Having the full state feedback, the next step is to reduce this to a specified decentralized structure, that is a control law given by (2.3). In [9, 10] three different methods have been proposed for this reduction. There, it has been shown that all methods for decentralized control system designs lead to the control law which can be represented in the form

$$u_i(t) = E_i \Lambda_i x(t), \qquad i = 1, 2, \ldots, k \tag{2.9}$$

where the values of the matrices E_i and Λ_i depend on the specific method chosen.

When the decentralized control is applied to system (2.1), the value of the performance index is

$$J = e^{2\alpha t_0} x^T(t_0) P x(t_0) \tag{2.10}$$

where P satisfies the matrix equation

$$\left(A+\alpha I+\sum_{i=1}^{k} B_i E_i \Lambda_i\right)^T P+P\left(A+\alpha I+\sum_{i=1}^{k} B_i E_i \Lambda_i\right)+\sum_{i=1}^{k} \Lambda_i^T E_i^T R_i E_i \Lambda_i+Q=0\ldots \tag{2.11}$$

For the ease of the subsequent calculations, it is assumed that the matrix

$$\bar{T}=\sum_{i=1}^{k} \Lambda_i^T E_i^T R_i E_i \Lambda_i+Q \tag{2.12}$$

is a nonsingular matrix, thus guaranteeing that P is positive definite. In this way, we have a well-defined Lyapunov function. Consequently, this Lyapunov function is a logical choice for evaluating the robustness of the decentralized feedback design.

Implicit in what has been assumed is that the closed loop system

$$\dot{x}(t)=\left(A+\sum_{i=1}^{k} B_i E_i \Lambda_i\right) x(t) \tag{2.13}$$

is exponentially stable with degree α, i.e.

$$\operatorname{Re}\left[\lambda_i\left(A+\sum_{i=1}^{k} B_i E_i \Lambda_i\right)\right]<\alpha. \tag{2.14}$$

However, direct synthesis procedures for decentralized output feedback designs, guaranteeing an arbitrary degree of stability, are unknown at present. If a system is stabilizable by a decentralized control law then there is a maximum degree of stability $\alpha_{\max}$ for this system. Therefore, when we are referring to a decentralized control system with a prescribed degree of stability, we mean the system with

$$0 \leqq \alpha \leqq \alpha_{\max}. \tag{2.15}$$

3. Robustness characterization

3.1. Problem formulation

In this section the attention is focused on the stability properties of the closed loop systems, designed on the basis of decentralized output feedback techniques suggested in the previous section, when the system matrices A and B_i, $i=1, 2, \ldots, k$ undergo large parameter variations.

The class of systems considered here are perturbed version of (S) (see eq. (2.1)) satisfying

$$(\tilde{S}_1)\quad\begin{cases}\dot{\tilde{x}}=A\tilde{x}+\sum\limits_{i=1}^{k}B_i\tilde{u}_i+A^1\tilde{x}+\sum\limits_{i=1}^{k}B_i^1\tilde{u}_i, \qquad \tilde{x}(t_0)=x_0 & (3.1)\\ \tilde{u}_i=E_i\Lambda_i\tilde{x} & (3.2)\end{cases}$$

where the matrices A, B_i, E_i and Λ_i are the same as in the nominal, unperturbed system (S); so all the parameter variations and modelling errors are lumped into the matrices A^1 and B_i^1, $i=1, 2, \ldots, k$. Although we restrict our attention to the case of linear perturbations in the system dynamics, the results of this paper can be applied to the case of nonlinear perturbations, too (see [10]).

3.2. *Characterization of perturbations*

A wide choice of measures of the "size" of the perturbation matrices has been suggested in the literature. So far, a measure of the magnitude of the perturbations has been made possible by the concept of matrix norm, and the robustness properties of a control system have been quantitatively expressed in term of the matrix norm. However, it has been noticed that this approach leads to excessively conservative results, since these measures do not distinguish the "direction" of the perturbations. In practice, many of the perturbations simply cannot occur physically.

This points out that in the area of robustness analysis for multivariable systems, one of the basic needs is for more refine tests and measures of robustness. One direction to mitigate this difficulty would be to combine the information concerning the nature of the perturbation that the system is most sensitive to, and the mathematical characterization of the perturbation matrices. That is, to define the perturbations in certain directions which are more appropriate from a physical standpoint.

In this sense, in what follows, it will be assumed that the perturbation matrices A^1 and B_i^1, $i=1, 2, \ldots, k$ are given in two components, one of which lies along a given direction in the space of all A^1 and B_i^1, $i=1, 2, \ldots, k$, respectively. In particular, define a model perturbation term A^1 of the open loop matrix A as

$$A^1=\gamma(t)\bar{A}+\Delta A \tag{3.3}$$

and a model perturbation terms B_i^1, of the control actuating matrices B_i as

$$B_i^1=\gamma(t)\bar{B}_i+\Delta B_i, \qquad i=1, 2, \ldots, k \tag{3.4}$$

where $\gamma(t)$ is a scalar function, and ΔA and ΔB_i represent the perturbations in the system dynamics, which lie out of the directions $\bar{A}$ and $\bar{B}_i$, respectively.

This points out the fundamental problem of obtaining the characterisation of the uncertainties associated with a given model. It seems that this knowledge can be acquired only by experience with real applications. Therefore, the idea behind the presentation of the perturbation matrices in two components is that a designer following his intuition and experience has enough information on the perturbations (modelling error or parameter variations) to select the most appropriate directions in the space of all perturbation matrices. When the matrices $\bar{A}$ and $\bar{B}_1$, $i=1, 2, \ldots, k$ are selected, it will be shown in the next section that it is possible to calculate a sector $(\gamma_{\min}, \gamma_{\max})$ such that the disturbed system remain stable, for $\gamma(t) \in (\gamma_{\min}, \gamma_{\max})$. The next step is to determine the norm bounds on the matrices ΔA and ΔB_i, $i=1, 2, \ldots, k$, which lie out of the chosen direction. It is obvious that this approach leads to a less conservative robustness characterization.

3.3. Results

Using eqs. (3.3) and (3.4) the perturbed system $(\tilde{S}_1)$ (eqs. (3.1) and (3.2)) becomes

$$(\tilde{S}_2) \quad \begin{cases} \dot{\tilde{x}} = A\tilde{x} + \sum\limits_{i=1}^{k} B_i\tilde{u}_i + \gamma\bar{A}\tilde{x} + \Delta A\tilde{x} + \sum\limits_{i=1}^{k} \gamma\bar{B}_i\tilde{u}_i + \sum\limits_{i=1}^{k} \Delta B_i\tilde{u}_i, \qquad \tilde{x}(t_0) = x_0 & (3.5) \\ \tilde{u}_i = E_i\Lambda_i\tilde{x} & (3.6) \end{cases}$$

Now suppose that the perturbation matrix A^1 lies entirely along the direction $\bar{A}$ and that the matrix B_i^1 also lies entirely along the direction $\bar{B}_i$, that is $\Delta A = 0$ and $\Delta B_i = 0$, $i = 1, 2, \ldots, k$. Then the following theorem gives the sector $(\gamma_{\min}, \gamma_{\max})$, i.e. bounds on the scalar function $\gamma(t)$ such that the perturbed system remains stable.

Theorem 3.1. Let the matrices $\Delta A = 0$ and $\Delta B_i = 0$, $i = 1, 2, \ldots, k$. If the following inequalities are satisfied

$$\lambda_{\min}^{-1}(Y) = \gamma_{\min} < \gamma(t) < \gamma_{\max} = \lambda_{\max}^{-1}(Y) \tag{3.7}$$

where $\gamma(t)$ is memoryless, time-varying nonlinearity, and

$$Y = T^{-1}\left(\left(\bar{A} + \sum_{i=1}^{k} \bar{B}_i E_i \Lambda_i\right)^T P + P\left(\bar{A} + \sum_{i=1}^{k} \bar{B}_i E_i \Lambda_i\right)\right) \tag{3.8}$$

$$T = \left(\sum_{i=1}^{k} \Lambda_i^T E_i^T R_i E_i \Lambda_i + Q + 2\alpha P\right) \tag{3.9}$$

and where the matrix P is the positive definite solution of eq. (2.11), and α is a prescribed degree of stability parameter; for all $t \in [0, \infty)$ then the perturbed system $(\tilde{S}_2)$ remains asymptotically stable.

Proof. See Appendix A.

In essence, Theorem 3.1 shows that if the perturbations in the dynamic system lie entirely along the directions $\gamma(t)\bar{A}$ and $\gamma(t)\bar{B}_i$, $i=1, 2, \ldots, k$ with $\gamma(t)\in$ $\in(\gamma_{\min}, \gamma_{\max})$, for all $t\in[0, \infty)$ then the perturbed system remains stable. Hence, the directions

$$\{\gamma(t)A;\quad \gamma(t)B_i,\quad i=1,2,\ldots,k:\quad \gamma\in(\gamma_{\min},\gamma_{\max})\} \tag{3.10}$$

are termed stability directions.

Although the expressions for the bounds on the nonlinear scalar function $\gamma(t)$ appear to be complicated, they are, in fact, not very difficult to calculate. Once the decentralized control problem is solved a few further computations are needed to carry out the robustness analysis. Remember that matrices P and T are calculated as part of design procedure and the calculation of the suboptimal index performance. In addition, there are straightforward methods for the $\lambda_{\min}$ and $\lambda_{\max}$ calculation.

So far, the robustness analysis has been restricted to the case when the perturbations in the system dynamics lie entirely along the given directions in the space of all perturbation matrices. In [20] the conditions have been established which guarantee that a stable closed loop system with $\gamma(t)\in(\gamma_{\min}, \gamma_{\max})$ for all $t\in[0, \infty)$, will remain stable in face of model perturbations out of the given directions, $\bar{A}$ and $\bar{B}_i$, whenever the norm bounds on ΔA and on ΔB_i, $i=1, 2, \ldots, k$ remain appropriately bounded.

As an example of the applications of the above results, in the next section, we consider the determination of the finite regions of stability for singularly perturbed systems.

4. Application to singular perturbation theory

4.1. Introduction

In the previous section the robustness of the stability property of a decentralized control system to model variations has been considered. In this section we will consider a particular form of model variation due to a singular perturbation.

Consider a dynamic system of the form

$$\dot{x}_1(t)=A_{11}x_1(t)+A_{12}x_2(t)+B^{11}u(t) \tag{4.1}$$

$$\mu\dot{x}_2(t)=A_{21}x_1(t)+A_{22}x_2(t)+B^{22}u(t) \tag{4.2}$$

where μ is a small positive parameter, and it is assumed that A_{22} is stable matrix.

There are two quite distinct motivations for the interest in singular perturbation method. One reason is of course the possible reduction in computation associated with dealing with reduced order models. The reduction of the number of equations is not the only advantage of this approach. It is to be expected that less accurate integration routine or shorter word can be used for solving the corresponding equations. The second motivation is based on the realization that it is usually impractical or impossible to implement feedback controllers based on complete models of the system, and that simplified mathematical models will lead to a simplified control system structure.

Therefore, engineering utilization of a mathematical model of a system requires on the one hand that the model be simple enough to permit a simplified control system structure, and to permit design computations to be carried out by present day computers. However, on the other hand, the model must retain sufficient detail to reflect the significant aspects of system performance accurately. Traditionally, engineers have made the necessary trade-offs between these two conflicting requirements based on intuition and experience. However, in the case of complicated large scale systems the requisite intuition may be difficult to acquire, and at the same time, there may be a compelling need to employ a greatly simplified mathematical model of the system. Therefore, a fundamental problem in large scale system theory is to give conditions for the success of design based on simplified models — this is essentially a robustness problem.

4.2. Design procedure

Among the most actively investigated singularly perturbed optimal control problems[1] is the general optimal-linear quadratic regulator problem, which has been solved in [6–8, 13, 17]. There, the performance index of the form

$$J=1/2\int_{0}^{T}(y^T(t)Q_1 y(t)+u^T(t)Ru(t))\,dt \tag{4.3}$$

where y is the output identical to the substate x_1 and

$$Q=\begin{bmatrix} Q_1 & 0 \\ 0 & 0 \end{bmatrix} \tag{4.4}$$

has been considered. It has been demonstrated that the zeroth order solution can be improved by the method of matched asymptotic expansions. That is, it has been shown that the high-order optimal solution $K(\mu)$ of the corresponding Riccati equation, can be approximated by its truncated Maclaurin series in μ

$$K(\mu)=\tilde{K}=K(0)+\left.\frac{\partial K}{\partial \mu}\right|_{\mu=0}. \tag{4.5}$$

[1] See [5] for an excellent survey of results in singular perturbation theory.

In approximate designs one of the basic issues is the problem of stability. Several papers [3, 14, 16] have appeared in the literature on asymptotic stability of singularly perturbed systems, but contain no expression for calculating an upper bound for μ for which the perturbed system remains stable. In [18] using the contraction mapping technique, the conditions for system stability and an upper bound for the singular parameter μ have been derived. However, these results require rather intensive calculation, and the connection of these results with other possible perturbations in the system input transducers, system dynamics and system sensors is unclear. For frequency domain approach in [12] a bound on the magnitude of the perturbation parameter μ that can be tolerated and still have a stability analysis of the reduced system valid for the full system has been given. Therefore, in this paper a special case, when a reduced system ($\mu=0$) is used in design procedure, has been considered.

In the previous section we have discussed the robustness of the stability property of decentralized feedback designs subject to model variations. Using a similar idea, in what follows, we will consider the problem of stability of singularly perturbed control systems and the problem of finding an upper bound for the parameter variation such that asymptotic stability is ensured.

First of all, notice that by using the performance index

$$J=\int_0^\infty e^{2\alpha t}(y^T(t)Q_1 y(t)+u^T(t)Ru(t))\,dt \tag{4.6}$$

instead of (4.3), and by using the proposed approach for decentralized designs in Section 2, based on mimic the centralized control by decentralized control, we can also employ the perturbation method for decentralized feedback designs. Therefore, we can also use the a priori knowledge, about the dynamics of the subsystems, in determining decentralized control strategies for the large scale interconnected systems.

4.3. *Stability analysis*

Let us present the decentralized control law, for some $\mu=\mu_0$, in the form

$$u_i(t)=-R_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i x(t) \tag{4.7}$$

where

$$B_i=\begin{bmatrix} B_i^{11} \\ \dfrac{B_i^{22}}{\mu_0} \end{bmatrix}, \qquad \tilde{K}=K(\mu_0), \qquad \tilde{\Lambda}_i=\Lambda_i(\mu_0). \tag{4.8}$$

Then the closed loop system becomes

$$\dot{x}(t)=\left(A-\sum_{i=1}^{k} B_i R_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right)x(t). \tag{4.9}$$

Even if the closed loop system (4.9) is stable[2] for $\mu=\mu_0$, the stability of the system is in question if the perturbation parameter μ changes his value during the operation of the system. Thus the estimation of the range of the parameter μ for which the approximate solution can be used to stabilize the perturbed system is of particular interest. To do that we shall apply the results of Section 3. Before that, define a new parameter γ such that

$$\gamma=\frac{1}{\mu}, \qquad \mu\to 0, \quad \gamma\to\infty. \tag{4.10}$$

The following relationship can be established between the perturbed parameter μ and the nominal value μ_0

$$\mu=\mu_0 \Delta\mu \tag{4.11}$$

then

$$\Delta\mu=\frac{1}{\mu_0 \Delta\gamma+1} \tag{4.12}$$

where

$$\Delta\gamma=\gamma-\gamma_0. \tag{4.13}$$

In a similar way if

$$\mu=\mu_0+\Delta\mu \tag{4.14}$$

then

$$\Delta\mu=-\frac{\Delta\gamma}{(\Delta\gamma+\gamma_0)\gamma_0}. \tag{4.15}$$

In both cases $\Delta\mu$ and $\Delta\gamma$ are some nonlinear functions of t. Therefore, if we define the range on $\Delta\gamma$ so that the closed loops system remains stable, then $\Delta\mu$ is uniquely defined by (4.12) or by (4.15). Now, the results of Section 3 can be directly applied.

If the perturbations in the parameter μ are dominant perturbations in the system dynamics, then they uniquely define a direction in the whole space of the perturbation matrices A^1 and B_i^1. That is,

$$A^1=\Delta\gamma\bar{A}+\Delta A, \qquad B_i^1=\Delta\gamma\bar{B}_i+\Delta B_i \tag{4.16}$$

where

$$\bar{A}=\begin{bmatrix} 0 & 0 \\ A_{21} & A_{22} \end{bmatrix} \qquad \bar{B}_i=\begin{bmatrix} 0 \\ B_i^{22} \end{bmatrix} \tag{4.17}$$

and ΔA and ΔB_i cover all perturbations out of these directions.

We give now the following theorem.

[2] Because of continuity, there exists $\mu^*>0$ such that the system is asympto'ically stable for all $\mu_0\in[0,\mu^*]$.

Theorem 4.1. If the closed loop system (4.9) is stable for some $\mu = \mu_0$, it will remain stable for all $\mu = \mu_0 + \Delta\mu$; $[\mu = \mu_0 \Delta\mu]$; where $\Delta\mu$ is given with (4.15) [(4.12)] and $\Delta\gamma$ satisfies the inequalities

$$\lambda_{\min}^{-1}(Y) = \gamma_{\min} < \Delta\gamma < \gamma_{\max} = \lambda_{\max}^{-1}(Y) \tag{4.18}$$

for all $t \in [0, \infty)$, where

$$Y = T^{-1}\left(\left(\bar{A} - \sum_{i=1}^{k} \bar{B}_i R_i^{-1} \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i\right)^T P + P\left(\bar{A} - \sum_{i=1}^{k} \bar{B}_i R_i^{-1} \tilde{B}^T \tilde{K} \tilde{\Lambda}_i\right)\right). \tag{4.19}$$

$T = T(\mu_0)$ is defined by (3.9), $\tilde{B}_i = \tilde{B}_i(\mu_0)$, $\tilde{K} = \tilde{K}(\mu_0)$, $\tilde{\Lambda}_i = \tilde{\Lambda}_i(\mu_0)$, the matrix $P = P(\mu_0)$ is the positive definite solution of the linear matrix equation (2.11), where $\bar{E}_i \Lambda_i = -R_i^{-1} \tilde{B}_i^T K \tilde{\Lambda}_i$, and $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote maximum and minimum eigenvalue of $(\cdot)$, respectively.

Proof. The proof of Theorem 4.1 is based on the proof of Theorem 3.1 and therefore is omitted.

Remark 4.1. Notice that we are only interested in the upper bound of μ as we know that the system is stable for all $\mu \in [0, \mu_0]$.

Remark 4.2. Someone may wonder why we do not compute $\mu = \mu_0 + \Delta\mu$ directly by making the matrix

$$A_c(\mu) = A(\mu) = \sum_{i=1}^{k} B_i(\mu) R^{-1} \tilde{B}_i(\mu_0) \tilde{K}(\mu_0) \tilde{\Lambda}_i(\mu_0) \tag{4.20}$$

stable. The reason is clear. In order to determine the range of μ for which the matrix A_c is stable, the usual way is to use Routh stability criterion or to check the Hurwitz determinants. In either case, some of the conditions will involve high-order polinomials of μ. Solving inequality equations of this kind is formidable, if not impossible.

The alternative expression for the perturbation parameter μ which does not disturb system stability is given in the following lemma.

Lemma. 4.1. If the perturbation parameter $\mu(t)$ satisfies condition (4.12), where

$$|\Delta\gamma(t)| \left\| \bar{A} - \sum_{i=1}^{k} \bar{B}_i R_i^{-1} \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i \right\|_s \leqq \frac{\lambda_{\min}(T)}{2\lambda_{\max}(P)} \tag{4.21}$$

and where the matrix $T = T(\mu_0)$ is defined by $T = \sum_{i=1}^{k} \tilde{\Lambda}_i^T \tilde{K}^T \tilde{B}_i^T R_i \tilde{B}_i \tilde{K} \tilde{\Lambda}_i + Q + 2\alpha P$ and the matrix $P = P(\mu_0)$ is defined by (2.11); for all $t \in [0, \infty)$, then the closed loop system remains stable.

[3] $\|\cdot\|_s$ denotes spectral norm.

Proof. See Appendix B.

The case when the operation of the system the perturbation parameter μ takes different values in different parts of the system, can be also easily included in the proposed formulation. We consider the case when the perturbation matrices can be presented in the form

$$A^1 = \begin{bmatrix} 0 & 0 \\ \dfrac{A_{21}^1}{\mu_1} & \dfrac{A_{22}^1}{\mu_1} \\ \dfrac{A_{21}^2}{\mu_2} & \dfrac{A_{22}^2}{\mu_2} \\ \vdots & \vdots \\ \dfrac{A_{21}^v}{\mu_v} & \dfrac{A_{22}^v}{\mu_v} \end{bmatrix} . \qquad B_i^1 = \begin{bmatrix} 0 \\ \dfrac{(B_i^{22})^1}{\mu_1} \\ \dfrac{(B_i^{22})^2}{\mu_2} \\ \vdots \\ \dfrac{(B_i^{22})^v}{\mu_v} \end{bmatrix} . \tag{4.22}$$

That is,

$$A^1 = \Gamma \bar{A}, \qquad B_i^1 = \Gamma \bar{B}_i \tag{4.23}$$

where

$$\Gamma = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \gamma_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \gamma_v \end{bmatrix}, \qquad \gamma_i = \frac{1}{\mu_i} = \gamma_0 + \Delta\gamma_i, \qquad i = 1, 2, \dots, v \tag{4.24}$$

and (see (4.17))

$$A_{21} = \begin{bmatrix} A_{21}^1 \\ A_{21}^2 \\ \vdots \\ A_{21}^v \end{bmatrix}, \qquad A_{22} = \begin{bmatrix} A_{22}^1 \\ A_{22}^2 \\ \vdots \\ A_{22}^v \end{bmatrix}, \qquad B_i^{22} = \begin{bmatrix} (B_i^{22})^1 \\ (B_i^{22})^2 \\ \vdots \\ (B_i^{22})^v \end{bmatrix} . \tag{4.25}$$

Then the following lemma holds.

Lemma 4.2. The closed loop system

$$\dot{\tilde{x}}(t) = \left(A + A^1 - \cdot\sum_{i=1}^{k} (B_i + B_i^1) R_i^{-1}) R_i^{-1} \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i \right) \tilde{x}(t) \tag{4.26}$$

where the matrices A^1 and B_i^1 are given by (4.23) will remain stable if the following condition holds

$$|\Delta\gamma_i|_{\max}\left(\|\bar{A}\|_s - \sum_{i=1}^{k}\|\bar{B}_i\|_s\|R_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\|_s\right) < \frac{\lambda_{\min}(T)}{2v\lambda_{\max}(P)} \tag{4.27}$$

where the matrix $T = T(\mu_0)$ is defined by (3.9) and the matrix $P = P(\mu_0)$ is defined by (2.11); for all $t \in 0, \infty)$.

Proof. The proof follows directly from Lemma 4.1 and therefore is omitted.

The actual system should suffer from large variations in the perturbation parameter μ, then the control system should naturally be designed to accomodate these perturbations. To appreciate fully the implications of the presented results for providing a basis for the design of robust decentralized multivariable feedback control system, it is instructive to prove the following lemma.

Lemma 4.3. The following relationship between the perturbation bounds on the parameter μ and the dominant closed loop poles of the nominal system exists

$$|\Delta\gamma(t)|\left\|\bar{A} - \sum_{i=1}^{k}\bar{B}_iR_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right\|_s \leqq -\mathrm{Re}\left[\lambda_{\max}\left(A - \sum_{i=1}^{k}B_iR_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right)\right] \tag{4.28}$$

for all $t \in [0, \infty)$. In other words, (4.28) is sufficient for (4.21).

Proof. See Appendix C.

A similar result can be easily derived for the case when during the operation of the system the parameter μ takes different values in different parts of the system.

Remark 4.3. Having in mind that

$$-\mathrm{Re}\left[\lambda_{\max}\left(A - \sum_{i=1}^{k}B_iR_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right)\right] \geqq \alpha, \tag{4.28}$$

inequality (4.28) also establishes a relationship between the allowable perturbations in the parameter μ and the prescribed degree of stability α, helping a designer to use the parameter α as a design parameter.

It is important to emphasize that the decentralized pole-placement approach for large scale control system designs present no difficulty to the theory, and that the results developed, can all be applied directly.

4.4. Examples and discussion

Finally we will give two examples to illustrate some of the ideas of the section.

Example 1. Consider the second order system whose state space representation is

$$\dot{x}_1 = x_1 + x_2 \tag{4.29}$$

$$\dot{x}_2 = -100x_2 + 50u. \tag{4.30}$$

System (4.29), (4.30) can be also defined in the form

$$\dot{x} = x_1 + x_2 \tag{4.31}$$

$$\mu\dot{x}_2 = -x_1 + 0.5u \tag{4.32}$$

where $\mu = 0.01$, i.e.,

$$A = \begin{bmatrix} 1 & 1 \\ \dfrac{-1}{\mu} & \dfrac{0.5}{\mu} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \dfrac{0.5}{\mu} \end{bmatrix}. \tag{4.33}$$

For the high order design we choose the following weighting matrices

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad R = 1 \tag{4.34}$$

and $\alpha = 1$.

Using the truncated Maclaurin series (4.5) for $\mu = 0.01$, the feedback gain matrix (4.7) becomes

$$G(0.01) = -[4.1937 \quad 0.0424] \tag{4.35}$$

where $\Lambda = I$ has been chosen.

The corresponding closed loop system matrix is

$$A_c = \begin{bmatrix} 1 & 1 \\ -209.685 & -102.118 \end{bmatrix} \tag{4.36}$$

which is a stable matrix. Therefore, the closed loop system (4.9) is stable for the nominal value $\mu_0 = 0.01$. However, if during the operation of the system the singular parameter μ changes its value, $\mu \neq \mu_0$, then the estimation of the range of the parameter μ for which the approximate solution (4.35) can be used to stabilize the perturbed system is of interest. To do that, we will apply the results of this section.

Let the perturbations in the parameter μ be only perturbations in the system model. Then the perturbation matrices $\bar{A}$ and $\bar{B}$ are defined by

$$\bar{A} = \begin{bmatrix} 0 & 0 \\ -1 & 0.5 \end{bmatrix}, \qquad \bar{B} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}. \tag{4.37}$$

Now, we can use the results of Theorem 4.1. If the solution of eq. (3.9) is used as a Lyapunov function then from conditions (4.18) it follows that

$$-0.5832 < \Delta\gamma < 0.4195. \tag{4.38}$$

The corresponding variations in the perturbation parameter μ can be calculated from (4.12),

$$0.9958 < \Delta\mu < 1.0059. \tag{4.39}$$

First, notice that we are only interest in the upper bound as we know that the system is stable for all $\mu \in [0, \mu_0]$. Second, the result is very conservative, as the allowable perturbation is only 0.59% of the nominal value μ_0.

In fact the example has been chosen to point out some of the difficulties which can arise in the application of the robustness results of Section 3 to singularly perturbed systems. The main problem is that in case of singularly perturbed systems, the matrix T, eq. (3.3) can be close to singular matrix, which leads to overly conservative results.

In the case of second order system (4.31), (4.32), the corresponding matrix T is given by

$$T = \begin{bmatrix} 18.5871 & 0.1778 \\ 0.1778 & 0.0018 \end{bmatrix} \tag{4.40}$$

and the corresponding eigenvalues are

$$\lambda_1 = 18.5888, \qquad \lambda_2 = 0.0001. \tag{4.41}$$

To overcome these difficulties, we can define a new Lyapunov matrix equation

$$A_c^T P + P A_c + Q = 0 \tag{4.42}$$

where

$$Q = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}. \tag{4.43}$$

In this case,

$$-28.168 < \Delta\gamma < \infty \tag{4.44}$$

that is,

$$\Delta\mu \in 0, (1.3922). \tag{4.45}$$

Therefore, the allowable variations in the perturbation parameter are nearly 40% of its nominal value.

Example 2. As another example we consider a fifth order system, primarily to illustrate the application of the methodology proposed to the case of decentralized control. The system with fast and slow modes is defined in form (4.1) and (4.2), where,

$$A_{11}=\begin{bmatrix} -0.2 & 0.5 \\ 0 & -0.5 \end{bmatrix}, \quad A_{12}=\begin{bmatrix} 0 & 0 & 0 \\ 1.6 & 0 & 0 \end{bmatrix}. \tag{4.46}$$

$$A_{21}=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_{22}=\begin{bmatrix} -14.28 & 85.72 & 0 \\ 0 & -25.0 & 75.0 \\ 0 & 0 & -10.0 \end{bmatrix}. \tag{4.47}$$

$$B^{11}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B^{22}=\begin{bmatrix} 3.0 & 0 & 0 \\ 0 & 3.0 & 0 \\ 0 & 0 & 3.0 \end{bmatrix}. \tag{4.48}$$

For weighting matrices in performance index (4.3)

$$Q=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad R=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.49}$$

are chosen, and the optimal control (4.5) is calculated.

Having the full state feedback the next step was to reduce this to a specified decentralized structure [9, 10], that is the control law given by (2.3), where

$$y_i(t)=C_i x(t), \qquad i=1,2,3 \tag{4.50}$$

where

$$C_1=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad C_2=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$
$$C_3=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.51}$$

For the nominal value of the perturbation parameter $\mu_0=0.1$ is chosen. Then from condition (4.18) it follows that the perturbed system with $\mu=\mu_0 \Delta\mu$ will remain stable for

$$1<\Delta\mu<1.010993. \tag{4.52}$$

In other words, the allowable perturbation is only 1.0993% of the nominal value.

To provide an algorithm for determining computationally the largest posible number $\Delta\mu$, such that the perturbed system remains stable, we use an iterative procedure similar to that proposed in [20]. In this way, after 54 iterations $\Delta\mu_{\max}$ is given with

$$1<\Delta\mu<1.6504. \tag{4.53}$$

Therefore, the allowable perturbation is over 65% of the nominal value $\mu_0=0.1$.

5. Conclusions

In this paper we have considered the robustness of stability property of decentralized control systems to variations in the system model. By decomposition of the perturbation matrices in two components new robustness results have been given which lead to reduction of conservatism in robustness tests.

We then applied the robustness results to the analysis of singularly perturbed systems. An explicit, readily computable bound on magnitude of the perturbation parameter μ which does not disturb system stability has been given. The case when during the operation of the system the perturbation parameter μ takes different values in different parts of the system has been also included in the proposed formulation. Finally, a relationship has been established between the perturbation bounds and the prescribed degree of stability of the nominal closed loop system, thus helping the designer to select an appropriate degree of stability to attain a robust design.

References

1. *Anderson, B. D. O., Moore, J. B.*, Linear Optimal Control. Prentice-Hall, New Jersey, 1971.
2. *Bellman, R.*, Introduction to Matrix Analysis. McGraw-Hill, New York, 1960.
3. *Desoer, C. A., Shensa, M. J.*, Networks with very small and very large parasitics: Natural frequences and stability. Proc. IEEE, **58** (1970), 1933–1938.
4. *Franclin, J. N.*, Matrix Theory. Prentice-Hall, New Jersey, 1968.
5. *Kokotović, P. V., O'Malley, R., Sannuti, P.*, Singular perturbations and order reduction in control theory — An overview. Automatica, **12** (1976), 123–132.
6. *Kokotović, P. V., Sannuti, P.*, Singular perturbation method for reducing the model order in optimal control design. IEEE Trans. Autom. Control, **13** (1968), 377–384.

7. *Kokotović, P. V.*, Feedback design of large linear systems, in Feedback Systems (J. B. Cruz, ed.). McGraw-Hill, New York, 1972.
8. *Kokotović, P. V., Yackel, R. A.*, Singular perturbation of linear regulators: Basic theorems. IEEE Trans. Autom. Control, **17** (1972), 27–37.
9. *Petkovski, Dj., Athans, M.* Robustness of decentralized output control designs with application to an electric power system. Third IMA Conference on Control Theory, Sheffield, September, 1980. Academic Press, London, 1981, 859–880.
10. *Petkovski, Dj.*, Robustness of decentralised control systems subject to sensor perturbations. IEE Proceedings, **132,** Pt. D (1985), 53–60.
11. *Sandell, N. R., Varaiya, P., Athans, M., Safonov, M.*, Survey of decentralized control methods for large scale systems. IEEE Trans. Autom. Control, **23** (1978), 108–128.
12. *Sandell, N. R.*, Robust stability of linear dynamic systems with application to singular perturbation theory. Automatica **15** (1978).
13. *Sannuti, P., Kokotović*, Near optimal design of linear systems by a singular perturbation method. IEEE Trans. Autom. Control, **14** (1969), 15–21.
14. *Šiljak, D. D.*, Singular perturbation of absolute stability. IEEE Trans. Autom. Control, **17** (1972), 720.
15. *Thrall, R. M., Thornheim, L.* Vector Space and Matrices, Wiley, New York, 1957.
16. *Wilde, R. R., Kokotović, P. V.*, Stability of singularly perturbed systems and networks with parasitics. IEEE Trans. Autom. Control, **17** (1972), 245–246.
17. *Yackel, R. A., Kokotović, P. V.*, A boundary layer method for the matrix Riccati equations. IEEE Trans. Autom. Control, **18** (1973), 17–24.
18. *Zien, L.*, An upper bound for the singular parameter in a stable singularly perturbed system. J. Franclin Institute, **295** (1973), 373–381.
19. *Petkovski, Dj., Athans, M.*, Robust decentralized control of multiterminal DC/AC power systems. Electric Power Systems Research, **9** (1985), 253–262.
20. *Petkovski, Dj.*, Robustness of control systems subject to modelling uncertainties, R. A. I. R. O. Automatique/Systems Analysis and Control, **18** (1984), 315–327.

APPENDIX A. *Proof of Theorem 4.1*

The proof is based on Lyapunov theory. Choose the positive definite Lyapunov functions as $V(\tilde{x})=\tilde{x}^T(t)P\tilde{x}(t)$. Taking the time derivative along the solution of $(\tilde{S}_2)$ for $\Delta A=0$, $\Delta B_i=0$, $i=1, 2, \ldots, k$, and using eq. (2.11), $\dot{V}(\tilde{x})$ may be represented as

$$\dot{V}(\tilde{x})=-\tilde{x}^T(t)\left(2\alpha P+Q+\sum_{i=1}^{k}\Lambda_i^T E_i^T R_i E_i \Lambda_i-\gamma(t)\left(\bar{A}+\sum_{i=1}^{k}\bar{B}_i E_i \Lambda_i\right)^T P-\right.$$

$$\left.-\gamma(t)P\left(\bar{A}+\sum_{i=1}^{k}\bar{B}_i E_i \Lambda_i\right)\right)\tilde{x}(t), \qquad \text{where} \quad E_i\Lambda_i=-R_i^{-1}\tilde{\bar{B}}_i^T\tilde{K}\tilde{\Lambda}_i. \tag{A.1}$$

To prove condition (3.7), recall the following lemma.

Lemma A1 [15]. If M and N are symmetric matrices and M is positive definite, there exists a nonsingular matrix S such that

$$S^T(M+N)S=I+G \tag{A.2}$$

where matrix G is a diagonal matrix whose elements are eigenvalues of $M^{-1}N$.

Therefore, using the result of Lemma Al, it can be easily concluded that the perturbed.system $(\tilde{S}_2)$ remains asymptotically stable if the following inequality is satisfied,

$$1-\gamma(t)\lambda_j\left(\left(2\alpha P+Q+\sum_{i=1}^{k}\Lambda_i^T E_i^T R_i E_i \Lambda_i\right)^{-1}\times\right.$$

$$\times\left(\left(\bar{A}+\sum_{i=1}^{k}\bar{B}_iE_i\Lambda_i\right)^TP+P\left(\bar{A}+\sum_{i=1}^{k}\bar{B}_iE_i\Lambda_i\right)\right)>0, \qquad j=1,2,\ldots,n \tag{A.3}$$

i.e.

$$1-\gamma(t)\lambda_j(Y)>0, \qquad j=1,2,\ldots,n \tag{A.4}$$

where the matrix Y is defined by (3.8), for all $t\in[0,\infty)$.

Now, under the assumption that $\lambda_{\max}(Y)>0$, which is a usual case, it follows that

$$\gamma_{\max}=\lambda_{\max}^{-1}(Y). \tag{A.5}$$

In a similar way if $\lambda_{\min}(Y)<0$ then

$$\gamma_{\min}=\lambda_{\min}^{-1}(Y) \tag{A.6}$$

i.e., the perturbed system $(\tilde{S}_2)$ remains stable if

$$\gamma(t)\in(\gamma_{\min},\gamma_{\max}). \tag{A.7}$$

Notice that if $\lambda_{\min}(Y)$ is not negative, then the bound $\gamma_{\min}$ ceases to exist and $(-\infty,\gamma_{\max})$. In a similar way if $\lambda_{\max}(Y)$ is not positive, then the bound $\gamma_{\max}$ ceases to exist and $(\gamma_{\min},\infty)$.

APPENDIX B. *Proof of Lemma 4.1*

The proof is similar to the proof of Theorem 4.1. Choosing $V(\tilde{x})=\tilde{x}^T(t)P\tilde{x}(t)$, where the matrix P is positive definite solution of eq. (2.11) for $\mu=\mu_0$, as a Lyapunov function for system (4.9), where $E_i\Lambda_i=-R_i^{-1}\tilde{B}_i^T\tilde{K}_i\tilde{\Lambda}_i$, $\dot{V}(\tilde{x})$ can be calculated as before to give

$$\dot{V}(\tilde{x})=-\tilde{x}^T(t)\left(2\alpha P+Q+\sum_{i=1}^{k}\tilde{\Lambda}_i^T\tilde{K}\tilde{B}_iR_i\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i-\Delta\gamma(t)\left(\bar{A}-\sum_{i=1}^{k}\bar{B}_iR_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right)^TP-\right.$$

$$\left.-\Delta\gamma(t)P\left(\bar{A}-\sum_{i=1}^{k}\bar{B}_iR^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right)\right)\tilde{x}(t). \tag{B.1}$$

Now, notice that

$$\tilde{x}^T(t)\Delta\gamma(t)P\left(\bar{A}-\sum_{i=1}^{k}\bar{B}_iR_i^{-1}\tilde{B}_i^T\tilde{K}_i^T\tilde{K}\tilde{\Lambda}_i\right)\tilde{x}(t)\leqq\|x\|_s^2\|P\|_s\left\|\bar{A}-\sum_{i=1}^{k}\bar{B}_iR_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right\|_s|\Delta\gamma(t)|. \tag{B.2}$$

Therefore, from (B.1), (B.2) and condition (4.21) it follows that

$$\tilde{x}^T(t)\Delta\gamma(t)P\left(\bar{A}-\sum_{i=1}^{k}\bar{B}_iR_i^{-1}\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i\right)\tilde{x}(t)\leqq\frac{1}{2}\lambda_{\min}(T)\|x\|_s^2 \tag{B.3}$$

and $\dot{V}(x)$ becomes

$$\dot{V}(\tilde{x})=-\tilde{x})=-\tilde{x}^T(t)\left(\sum_{i=1}^{k}\tilde{\Lambda}_i^T\tilde{K}\tilde{B}_iR_i\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i+Q+2\alpha P-\lambda_{\min}(T)\right)\tilde{x}(t). \tag{B.4}$$

It is easy to see that $\dot{V}(x)\leqq0$ for all $\tilde{x}(t)$ and, hence, system (4.9) is stable.

APPENDIX C. *Proof of Lemma 4.3*

The Lyapunov equation (2.11) can be expressed as

$$A_c^TP+PA_c=-\left(\sum_{i=1}^{k}\tilde{\Lambda}_i^T\tilde{K}\tilde{B}_iR_i\tilde{B}_i^T\tilde{K}\tilde{\Lambda}_i+Q+2\alpha P\right) \tag{C.1}$$

where the closed loop matrix A_c is defined by

$$A_c = A - \sum_{i=1}^{k} B_i R_i^{-1} \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i. \tag{C.2}$$

Now, recall the following theorem.

Theorem C1 [2, 4]. If λ_1 and λ_2 are the greatest and smallest roots of matrix M, where M is a Hermitian matrix, then

$$\lambda_1 \geqq \frac{\bar{q}Mq}{\bar{q}q} \geqq \lambda_2 \tag{C.3}$$

where $\bar{q}$ is the complex conjugate of q for any vector q, $q \neq 0$. That is,

$$\lambda_1 = \max_{q \neq 0} |\bar{q}Mq/\bar{q}q| \tag{C.4}$$

$$\lambda_2 = \min_{q \neq 0} |\bar{q}Mq/\bar{q}q|. \tag{C.5}$$

Denote by p the eigenvector corresponding to the greatest root of A_c. Premultiplying and postmultiplying (C.1) by $\bar{p}$ and p it follows that

$$\mathrm{Re}[\lambda_{\max}(A_c)] = -\frac{\bar{p} \sum_{i=1}^{k} \tilde{\Lambda}_i^T \tilde{K} \tilde{B}_i R_i \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i + Q + 2\alpha P)p}{2\bar{p}Pp}. \tag{C.6}$$

Notice that matrices of the right side of eq. (C.6) are symmetric matrices. Therefore, by applying the Theorem C1, it follows

$$\lambda_{\min}\left(\sum_{i=1}^{k} \tilde{\Lambda}_i^T \tilde{K} \tilde{B}_i R_i \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i + Q + 2\alpha P \right) \leqq \bar{p} \left(\sum_{i=1}^{k} \tilde{\Lambda}_i^T \tilde{K} \tilde{B}_i R_i \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i + Q + 2\alpha P \right) p \tag{C.7}$$

$$\lambda_{\min}(P) \leqq \bar{p}Pp \leqq \lambda_{\max}(P) \tag{C.8}$$

and

$$-\mathrm{Re}[\lambda_{\max}(A_c)] \leqq \frac{\lambda_{\min}\left(\sum_{i=1}^{k} \tilde{\Lambda}_i^T \tilde{K} \tilde{B}_i R_i \tilde{B}_i^T \tilde{K} \tilde{\Lambda}_i + Q + 2\alpha P \right)}{2\lambda_{\max}(P)}. \tag{C.9}$$

Therefore, it is easy to see that inequality (4.28) follows directly from (4.21) and (C.9).

Робастная стабильность многомерных систем управлениа применительно к сингулярным возмущениям систем

Д. Б. ПЕТКОВСКИ

(Нови Сад)

В работе предлагается простой подход к определению условий стабильности линейных многомерных систем управления в случае неопределенных или больших вариаций параметров динамической системы. В качестве примера применения результатов рассматривается определение области стабильности для сингулярных возмущений систем.

D. B. Petkovski
Faculty of Technical Sciences
University of Novi Sad
Veljka Vlahovic'a 3
21000 Novi Sad
Yugoslavia

Problems of Control and Information Theory, Vol. 15 (5), pp. 367—375 (1986)

SOME QUASIPERFECT DOUBLE ERROR CORRECTING CODES

S. M. DODUNEKOV

(*Sofia*)

(Received May 22, 1985)

The purpose of this paper is to prove that the following classes of codes are quasiperfect: Dumer–Zinov'ev's cyclic and constacyclic codes over $GF(4)$ with composite generator polynomial, binary reversible Melas' codes and modified cyclic double error correcting BCH codes.

1. Introduction

Let α be a primitive element of the field $GF(2^m)$. Consider the following optimal double error correcting codes:

1. *Dumer–Zinov'ev's cyclic and constacyclic codes over* $GF(4)$ *with composite generator polynomial* [1] (see also [2]). Set $m=2s$, $n=(2^{2s}-1)/3$. Let $g_i(x)$ be the minimal polynomial of α^i over $GF(4)$. Denote by A_S the constacyclic $[n, n-2s, 5]$ code over $GF(4)$ with generator polynomial $g_1(x)g_{-2}(x)$ and by B_S (if $(S, 3)=1$) — the cyclic $[n, n-2s, 5]$ code over $GF(4)$ with generator polynomial $g_3(x)g_{-6}(x)$.

2. *Melas' codes* [3]. Let $m=2s+1$ and let M_S be a binary cyclic reversible $[n=2s+1-1, n-4s-2, 5]$ code with generator polynomial $m_1(x)m_{-1}(x)$, where $m_i(x)$ denotes the minimal polynomial of α^i over $GF(2)$.

3. *Modified BCH codes* [4]. Let $(i, m)=1$, $\sigma=2^i$. Denote by $D\sigma$ the binary cyclic $[n=2^m-1, n-2m, 5]$ code with generator polynomial $m_1(x)m_{\sigma+1}(x)$.

The purpose of this paper is to prove that the codes A_S, B_S, M_S and $D\sigma$ are quasiperfect. In section 2 we give some lemmas, from which we derive the quasiperfectness of the codes A_S, B_S and M_S. In section 3 we establish the statement for $D\sigma$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. Quasiperfectness of the codes A_S, B_S and M_S

We begin with some lemmas. By $GF(2^m)^*$ we will denote the set of all nonzero elements of the field $GF(2^m)$.

Lemma 1. Let $b \in GF(2^m)^*$. Then the system

$$\begin{aligned} x_1+x_2+x_3&=0, \\ x_1^{-1}+x_2^{-1}+x_3^{-1}&=b \end{aligned} \tag{1}$$

has a solution in $GF(2^m)$.

Proof. From (1) we obtain

$$x_1^2(1+bx_2)+x_1x_2(1+bx_2)+x_2^2=0.$$

Setting $x_2=c\neq b^{-1}$, $x_1=cy$ we get the equation

$$y^2+y+\frac{1}{1+cb}=0.$$

The above equation (and hence, system (1)) has a solution in $GF(2^m)$ iff

$$\operatorname{Tr}\left(\frac{1}{1+cb}\right)=0, \qquad (\operatorname{Tr}(x)=x+x^2+\ldots+x^{2^{m-1}}) \tag{2}$$

(see [5]). But $1/1+cb$ covers $GF(2^m)^*$ when c runs over $GF(2^m)\backslash\{b^{-1}\}$ and obviously there exists c for which (2) holds.

Lemma 2. Let $b \in GF(2^m)$, $\operatorname{Tr}(b^{-1})=0$. Then the system

$$\begin{aligned} x_1+x_2&=1, \\ x_1^{-1}+x_2^{-1}&=b \end{aligned} \tag{3}$$

has a solution in $GF(2^m)$.

Proof. From the above equations it follows that x_1 and x_2 are roots of the equation $u^2+u+b^{-1}=0$, which has a solution in $GF(2^m)$ iff $\operatorname{Tr}(b^{-1})=0$.

Lemma 3. Let $b \in GF(2^m)^*$, $\operatorname{Tr}(b)=0$. Then the system

$$\begin{aligned} x_1+x_2+x_3&=1, \\ x_1^{-1}+x_2^{-1}+x_3^{-1}&=b \end{aligned} \tag{4}$$

has a solution in $GF(2^m)$.

Proof. Setting $x_i=y_i+1$, $i=1, 2, 3$, we obtain from (4)

$$1+y_1^2+y_2^2+y_1y_2=b(1+y_1^2+y_2^2+y_1^2y_2+y_1y_2^2+y_1y_2).$$

For arbitrary $y_2=c\in GF(2^m)$, $c\neq 0, 1$ we get the equation

$$\lambda y_1^2+c\lambda y_1+d=0,$$

where

$$\lambda=\lambda(c)=1+b+bc, \qquad d=1+b+(1+b)c^2.$$

Take $c\neq 1+b^{-1}$, i.e. $\lambda(c)=\lambda\neq 0$. Then

$$y_1^2+cy_1+d\lambda^{-1}=0$$

and hence (4) has a solution in $GF(2^m)$ iff the equation

$$u^2+u+d\lambda^{-1}c^{-2}=0 \tag{5}$$

has one, i.e. iff there exists $c\in GF(2^m)$, $c\neq 0$, 1, $1+b^{-1}$ such that

$$\mathrm{Tr}\,(d\lambda^{-1}c^{-2})=0.$$

We shall show that such an element actually exists. Note first that from

$$d\lambda^{-1}c^{-2}=\frac{\lambda+\lambda c+c+c^2}{\lambda c^2}=\frac{1}{c^2}+\frac{1}{c}+\frac{1+c}{\lambda c}$$

and

$$\mathrm{Tr}\,(d\lambda^{-1}c^{-2})=\mathrm{Tr}\left(\frac{1+c}{\lambda c}\right)$$

it follows that (5) has a solution in $GF(2^m)$ iff $\mathrm{Tr}\left(\frac{1+c}{\lambda c}\right)=0$.

Now it is easy to verify that the equation

$$\frac{c+1}{\lambda c}=\frac{b}{b^2+1}$$

has a solution $c\in GF(2^m)$. Since

$$\mathrm{Tr}\left(\frac{b}{b^2+1}\right)=\mathrm{Tr}\left(\frac{1}{b+1}+\frac{1}{b^2+1}\right)=0,$$

the lemma is proved.

Lemma 4. Let $b\in GF(2^m)$, $\mathrm{Tr}\,(b)=\mathrm{Tr}\,(b^{-1})=1$, $b^2+b+1\neq 0$. Then, system (4) has a solution in $GF(2^m)$.

Proof. Take $c=1/b(b+1)$. It is easy to check that $c\neq 1$, $1+b^{-1}$. Since

$$\frac{1+c}{\lambda c}=b+b^{-2}$$

then

$$\operatorname{Tr}\left(\frac{1+c}{\lambda c}\right)=\operatorname{Tr}(b)+\operatorname{Tr}(b^{-2})=0$$

and according to the proof of lemma 3, system (4) has a solution in $GF(2^m)$.

Lemma 5. Let $b \in GF(2^m)$, $\operatorname{Tr}(b)=\operatorname{Tr}(b)=\operatorname{Tr}(b^{-1})=1$, $b^2+b+1=0$, $m>2$. Then system (4) has a solution in $GF(2^m)$.

Proof. Note first that in this case $m=2(2t+1)$. As in the proof of lemma 3 we must show that there exists $c \in GF(2^m)$, $c \neq 0, 1, b$, such that $\operatorname{Tr}\left(\frac{1+c}{\lambda c}\right)=0$. From

$$\frac{c+1}{\lambda c}=\frac{1}{\lambda}+\frac{1}{\lambda c}$$

it follows that this is equivalent to the existence of c, such that

$$\operatorname{Tr}(\lambda^{-1})=\operatorname{Tr}(\lambda^{-1}c^{-1}), \qquad c \neq 0, 1, b.$$

In order to prove this we make some observations. Let c, $f \in GF(2^m)$ and let

$$c\lambda(c)=f\lambda(f), \qquad c \neq f, \tag{6}$$

i.e.

$$c+f+b(c+f)+b(c+f)^2=0.$$

Then

$$1+b=b(c+f), \qquad b^2=b(c+f)$$

and therefore $c+f=b$. Conversely, if $c+f=b$, we get (6).

Take now c, $f \in GF(2^m)$, $c+f=b$. We have

$$\operatorname{Tr}\left(\frac{1}{\lambda(c)}\right)=\operatorname{Tr}\left(\frac{c}{c\lambda(c)}\right)=\operatorname{Tr}\left(\frac{f+b}{f\lambda(f)}\right)=\operatorname{Tr}\left(\frac{1}{\lambda(f)}\right)+\operatorname{Tr}\left(\frac{b}{f\lambda(f)}\right).$$

If we choose f such that

$$\operatorname{Tr}\left(\frac{b}{f\lambda(f)}\right)=1,$$

then either $\operatorname{Tr}\left(\frac{1}{\lambda(c)}\right)$ or $\operatorname{Tr}\left(\frac{1}{\lambda(\mathrm{f})}\right)$ will coincide with ε where

$$\varepsilon=\operatorname{Tr}\left(\frac{1}{c\lambda(c)}\right)=\operatorname{Tr}\left(\frac{1}{f\lambda(f)}\right).$$

Consider now the equation

$$f\lambda(f)=\xi^{-1},$$

where $\xi \in GF(2^m)$, $\xi \neq 1$, $\operatorname{Tr}(b\xi)=1$, i.e.

$$f^2+bf+\xi^{-1}b^{-1}=0.$$

Hence the problem is to find $\xi \in GF(2^m)^*$, such that

$$\operatorname{Tr}(\xi^{-1})=0, \qquad \operatorname{Tr}(b\xi)=1, \qquad \xi \neq 1.$$

But for every $\xi \in GF(2^m)$

$$\operatorname{Tr}(b\xi^2)=\operatorname{Tr}((1+b^2)\xi^2)=\operatorname{Tr}(\xi)+\operatorname{Tr}(b\xi),$$
$$\operatorname{Tr}(\xi^{-1})=\operatorname{Tr}(\xi^{-2})$$

and if there exists $\eta \in GF(2^m)$ such that

$$\operatorname{Tr}(\eta^{-1})=0, \qquad \operatorname{Tr}(\eta)=1$$

then either $\xi=\eta$ or $\xi=\eta^2$ is a solution of the above problem. Suppose that for every $\eta \in GF(2^m)$ with $\operatorname{Tr}(\eta)=0$

$$\operatorname{Tr}(\eta^{-1})=\operatorname{Tr}(\eta)=0.$$

It follows that the set of all $x \in GF(2^m)$, for which $\operatorname{Tr}(x)=0$ is a subfield of $GF(2^m)$ with 2^{m-1} elements, i.e. $m-1$ divides m which is a contradiction.

We shall now prove the main results of this section.

Theorem 1. The codes A_S and B_S are quasiperfect.

Proof. (i) Consider first the code A_S. Its check matrix is

$$\begin{pmatrix} 1 & \alpha & \alpha & \dots & \alpha^{n-1} \\ 1 & \alpha^{-2} & \alpha^{-4} & \dots & \alpha^{-2(n-1)} \end{pmatrix}.$$

Let (S_1, S_2) be an arbitrary nonzero syndrome. We will show that there exist elements $\eta_1, \eta_2, \eta_3 \in GF(4)$ and integers j_1, j_2, j_3, $0 \leqq j_1, j_2, j_3 \leqq n-1$, such that

$$\begin{aligned} \eta_1\alpha^{j_1}+\eta_2\alpha^{j_2}+\eta_3\alpha^{j_3}&=S_1, \\ \eta_1\alpha^{-2j_1}+\eta_2\alpha^{-2j_2}+\eta_3\alpha^{-2j_3}&=S_2. \end{aligned} \tag{7}$$

Case (a). $S_1=0$. We try to find $\eta_1, \eta_2, \eta_3 \in GF(4)^*$. Since $\eta_i=\eta_i^{-2}$, $i=1, 2, 3$, setting $x_i=\eta_i\alpha^{j_i}$ in (7) we get the system

$$x_1+x_2+x_3=0,$$
$$x_1^{-1}+x_2^{-1}+x_3^{-1}=\sqrt{S_2},$$

which has a solution $x_i=\alpha^{k_i}$, $i=1, 2, 3$ in $GF(2^{2s})$ (lemma 1). Let $k_i=l_in+j_i$, $0\leqq j_i\leqq n-1$, $i=1, 2, 3$. Then

$$x_i=\alpha^{k_i}=(\alpha^n)^{l_i}\alpha^{k_i}=\eta_i\alpha^{j_i},$$

where $\eta_i=(\alpha^n)^{l_i}\in GF(4)$. Hence we obtain a solution of (7).

Case (*b*). $S_1\neq 0$. If $\operatorname{Tr}((S_1\sqrt{S_2})^{-1})=0$, we try to find a solution of (7) with $\eta_3=0$, $\eta_1, \eta_2\in GF(4)^*$. After a suitable substitution we obtain a system

$$x_1+x_2=1,$$
$$x_1^{-1}+x_2^{-1}=S_1\sqrt{S_2},$$

which, according to lemma 2, has a solution in $GF(2^{2s})$.

If $\operatorname{Tr}((S_1\sqrt{S_2})^{-1})=1$, using lemmas 3, 4 and 5, we obtain a solution of (7) with $\eta_1, \eta_2, \eta_3\in GF(4)^*$.

(ii) The check matrix α of the code B_S is

$$\begin{pmatrix} 1 & \beta & \beta^2 & \dots & \beta^{n-1} \\ 1 & \beta^{-2} & \beta^{-4} & \dots & \beta^{-2(n-1)} \end{pmatrix}, \qquad \beta=\alpha^3.$$

It is sufficient to prove that for every nonzero syndrome (S_1, S_2) there exist $\eta_1, \eta_2, \eta_3\in GF(4)$ and integers j_1, j_2, j_3, such that $0\leqq j_1, j_2, j_3\leqq n-1$,

$$\begin{aligned} \eta_1\beta^{j_1}+\eta_2\beta^{j_2}+\eta_3\beta^{j_3}&=S_1, \\ \eta_1\beta^{-2j_1}+\eta_2\beta^{-2j_2}+\eta_3\beta^{-2j_3}&=S_2. \end{aligned} \tag{8}$$

Setting $x_i=\eta_i\beta^{j_i}$, $i=1, 2, 3$ we obtain as in (i) systems of the type (1), (3) or (4), which have solutions in $GF(2^{2s})$. But since $(n, 3)=1$, for every $y\in GF(2^{2s})$ we have

$$y=\eta\beta^j, \qquad \eta\in GF(4), \qquad 0\leqq j\leqq n-1,$$

and therefore every solution of such a system yields a solution of (8). The theorem is proved.

Theorem 2. The code M_S is quasiperfect.

The proof is similar to those of theorem 1 and therefore is omitted. Note that the quasiperfectness of the code M_S was recently proved by Moreno [6] (see also [7]).

3. Quasiperfectness of the code $D\sigma$

We first recall the following fact.

Lemma 6. ([4]). The equation

$$n^{2^l}+n+d=0, \qquad (l, m)=1, \qquad d\in GF(2^m), \tag{9}$$

has a solution in $GF(2^m)$ iff $\operatorname{Tr}(d)=0$.

Further we give a simple arithmetical lemma.

Lemma 7. Let $(i, 2s)=1$. Then $(2^i+1, 2^{2s}-1)=3$.

Proof. Let $d=(2^i+, 2^{2s}-1)$ and let 2 belong to exponent t with respect to modulus d (see [8], p. 245). If $t=2^a t_1$, where t_1 is odd, it follows from $t_1 | 2i$ (this means t_1 divides $2i$), $t_1 | 2s$ and $(i, s)=1$ that $a=t_1=1$, i.e. d divides 3. But obviously 3 divides d. Therefore $d=3$.

Theorem 3. The code $D\sigma$ is quasiperfect.

Proof. Since the check matrix of the code $D\sigma$ is

$$\begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ 1 & \alpha^{\sigma+1} & \alpha^{2(\sigma+1)} & \dots & \alpha^{(n-1)(\sigma+1)} \end{pmatrix},$$

it is sufficient to prove that for every nonzero syndrome (a, b), $a, b \in GF(2^m)$ there exist $x_1, x_2, x_3 \in GF(2^m)$, such that

$$x_1+x_2+x_3=a,$$
$$x_1^{\sigma+1}+x_2^{\sigma+1}+x_3^{\sigma+1}=b.$$

Set $x_i=y_i+a$, $i=1, 2, 3$. Since $y_1+y_2+y_3=0$, then

$$x_1^{\sigma+1}+x_2^{\sigma+1}+x_3^{\sigma+1}=y_1^{\sigma+1}+y_2^{\sigma+1}+y_3^{\sigma+1}+a^{\sigma+1}.$$

Hence for y_1, y_2, y_3 we obtain the system

$$\begin{aligned} &y_1+y_2+y_3=0, \\ &y_1^{\sigma+1}+y_2^{\sigma+1}+y_3^{\sigma+1}=c, \qquad c=b+a^{\sigma+1}. \end{aligned} \tag{10}$$

Eliminating y_3 we get

$$y_1^{\sigma} y_2+y_1 y_2^{\sigma}+c=0. \tag{11}$$

If $c=0$, (10) has a solution $y_2=0$, $y_1=y_3$. Let $c \neq 0$. Then $y_2 \neq 0$ and with respect to $u=y_1 y_2$ from (11) we obtain

$$n^{\sigma}+u+\frac{c}{y_2^{\sigma+1}}=0. \tag{12}$$

According to lemma 6, (12) has a solution in $GF(2^m)$ iff $\operatorname{Tr}(c/y_2^{\sigma+1})=0$. Hence the problem is: for every $c \in GF(2^m)^*$ to find $y_2 \in GF(2^m)^*$, such that $\operatorname{Tr}(c/y_2^{\sigma+1})=0$.

Case (a). m is odd. Then the mapping

$$\sigma+1 : GF(2^m) \to GF(2^m), \qquad x \to x^{\sigma+1},$$

is one-to-one (because $(\sigma+1, 2^m-1)=1$ for odd m). Therefore $c/y_2^{\sigma+1}$ cover $GF(2^m)^*$ when y_2 runs over $GF(2^m)^*$ and we can choose y_2 with the above property.

Note that in this case the statement follows from the Delsarte's bound [9] as well, since for odd m the dual code $D_\sigma^\perp$ has exactly three nonzero weights (see [10], p. 252).

Case (*b*). m is even. According to lemma 7, there exist integers v and w, such that

$$(2^i+1)v+(2^m-1)w=3.$$

Let $c=\alpha^{3k+r}$, $r=0, 1, 2$. If $r=0$, we choose $y_2=\alpha^{vk}$. Since

$$(2^i+1)vk\equiv 3k(\operatorname{mod} 2^m-1),$$

$$\beta^{\sigma+1}=\beta^{3v}=\alpha^{3k}=c,$$

and since $\operatorname{Tr}(1)=0$, $y_2=\beta$ is a solution of the problem.

Let $r\neq 0$. Now there exists $\gamma\in GF(2^m)$, such that

$$\gamma=\alpha^{3l+r}, \qquad \operatorname{Tr}(\gamma)=0.$$

Choose $\beta\in GF(2^m)$, $\beta^v=\alpha^{k-l}$. Then

$$\beta^{\sigma+1}=\alpha^{3(k-l)}, \qquad \frac{c}{\beta^{\sigma+1}}=\alpha^{3l+r}$$

and therefore $y_2=\beta$ is a solution. The theorem is proved.

Note that for D_2 (i.e. for the primitive double error correcting BCH code), this theorem was proved by Gorenstein, Peterson and Zierler [11].

References

1. *Dumer, I. I., Zinov'ev, V. A.*, Some new maximal codes over the Galois field GF(4). Problemy Peredachi Informatsii, **14**, *3*, 1978, pp. 24–34.
2. *Gevorkjan, D. N., Avetisjan, A. M., Tigranjan, G. A.*, On the question of construction double error correcting codes in Hamming matrix over Galois fields. Computational technique, Kujbishev, **3**, 1975, pp. 19–21.
3. *Melas, C. M.*, A cyclic code for double error correction. IBM J.Res.Devel., **4**, *3*, 1960, pp. 364–366.
4. *Dumer, I. I.*, Some new uniformly packed codes. Tr. MFTI, ser. "Radiotechnique and electronics", 1976, pp. 72–78.
5. *Berlekamp, E. R., Rumsey, H., Solomon, G.* On the solutions of algebraic equations over finite fields. Inform. Control, **10**, *6*, 1967, pp. 553–564.
6. *Moreno, O.*, Further results on quasiperfect codes related to the Goppa codes. Congressus Numerandium, **40**, 1983, 249–256.
7. *Cohen, G. D., Karpovsky, M. R., Mattson, H. F.*, Jr., *Schatz, J. R.*, Covering radius — survey and recent results. Preprint.
8. *Sierpinski, W.*, Elementary Theory of Numbers. Panstwowe Wydawnictwo Naucowe, Warszawa, 1964.
9. *Delsarte, P.*, Four fundamental parameters of a code and their combinatorial significance. Inform. Control, **23**, *5*, pp. 407–438.
10. *MacWilliams, F. J., Sloane, N. J. A.*, The Theory of Error–Correcting Codes.North-Holland Publishing Company, Amsterdam–New York–Oxford, 1977.
11. *Gorenstein, D. C., Peterson, W. W., Zierler, N.*, Two-error correcting Bose–Chaudhuri codes are quasiperfect. Inform. Control, **3**, *3*, 1960, pp. 291–294.

Некоторые квазисовершенные коды, исправляющие две ошибки

С. М. ДОДУНЕКОВ

(София)

Доказывается, что следующие классы кодов являются квазисовершенными: циклические и констациклические коды Думера-Зиновьева над полем *GF*(4) с разложимым порождающим многочленом, двоичные реверсивные коды Меласа и модифицированные коды БЧХ, исправляющие две ошибки.

С. М. Додунеков
Институт математики с ВЦ БАН
НР Болгария, 1090 София,
ул. «Акад. Г. Бончев», бл. 8

Problems of Control and Information Theory, Vol. 15 (5), pp. 377—384 (1986)

ON CONSTRUCTING CODES WITH GIVEN DISTANCE IN LEE-METRIC

A. RACSMÁNY

(*Budapest*)

(Received September 10, 1985)

Let $\mathscr{C}_L(n, d, q)$ denote the class of n-dimensional codes over an alphabet of size q with minimal code distance d in the Lee-metric. In this paper we propose a procedure for constructing a code $C \in \mathscr{C}_L(n, \min\limits_{1 \leqq i \leqq s} (q_1 q_2 \dots q_{i-1} d_i), q_1 q_2 \dots q_s)$ by combining codes $C_i \in \mathscr{C}_L(n, d_i, q_i)$, $i = 1, \dots, s$. This permits us to find good codes over a large alphabet by making use of good codes over small alphabets.

Introduction

Many negative results are known which claim the non-existence of perfect codes in the Lee-metric for certain parameters, see e.g. [2], [5]. Thus the question arises how can we construct in such a case at least a "good" code. In this paper we propose a way for constructing codes which perform beyond the Varshamov–Gilbert bound.

Definitions and notations

Let q be an arbitrary positive integer. Let Z_q denote the set of the non-negative integers mod q and Z_q^n the set of the n-dimensional vectors over Z_q.

A set C is called an n-dimensional code over Z_q, if $C \subset Z_q^n$. By the Lee-distance of $\mathbf{c} \in C$ and $\mathbf{c}' \in C$ we mean the sum

$$d_{Lq}(\mathbf{c}, \mathbf{c}') = \sum_{i=1}^{n} d_{Lq}(c_i, c_i'), \qquad \text{where} \quad d_{Lq}(c_i, c_i') = \min(|c_i - c_i'|, q - |c_i - c_i'|).$$

In some cases it is profitable to use the following variant of the Lee-distance:

$$d_{Lq}(\mathbf{c}, \mathbf{c}') = \sum_{i=1}^{n} |\tilde{d}_{Lq}(c_i, c_i')|, \qquad \text{where} \quad \tilde{d}_{Lq} \equiv c_i - c_i' \bmod q \quad \text{and} \quad 0 \leqq |\tilde{d}_{Lq}(c_i, c_i')| \leqq \frac{q}{2}.$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

The minimal code distance of C is defined, as usual, by

$$d = \min_{\substack{\mathbf{c}\in C,\, \mathbf{c}'\in C \\ \mathbf{c}\neq \mathbf{c}'}} d_{Lq}(\mathbf{c}, \mathbf{c}').$$

A code $C \subset Z_q^n$ is called e-Lee-error correcting code if $d_{Lq}(\mathbf{c}, \mathbf{c}') \geqq 2e+1$, where $\mathbf{c} \in C$ and $\mathbf{c}' \in C$. Let $\mathscr{C}_L(n, 2e+1, q)$ denote the set of all these codes.

We call the set

$$V_L(n, e, q) = \{\mathbf{y} \mid d_{Lq}(\mathbf{c}, \mathbf{y}) \leqq e, \text{ where } \mathbf{c} \in Z_q^n \text{ is a prescribed vector}\}$$

Lee-sphere of radius e.

It is known [1] that for $q \geqq 2e+1$ we have

$$|V_L(n, e, q)| = \sum_{i=0}^{e} 2^i \binom{n}{i}\binom{e}{i}. \tag{1}$$

We say that a code C is perfect, if $|V_L(n, e, q)| \cdot |C| = q^n$.

Construction

By combination of codes with small code distance over a small alphabet we obtain codes with large code distance over a large alphabet. Let us consider the codes $C_1 \in \mathscr{C}_L^1(n, q_1 d, q_1)$ and $C_2 \in \mathscr{C}_L^2(n, d, q_2)$. The codewords of the form $\mathbf{c} = q_1\mathbf{c}_2 + \mathbf{c}_1$, where $\mathbf{c}_1 \in C_1$ and $\mathbf{c}_2 \in C_2$ constitute a code C over Z_q, $q = q_1 q_2$.

Theorem. The minimal code distance of C in the d_{Lq} metric is at least $q_1 d$.

Proof. Let $\mathbf{c} = q_1\mathbf{c}_2 + \mathbf{c}_1$ and $\mathbf{c}' = q_1\mathbf{c}_2' + \mathbf{c}_1'$ be two different codewords of C. We must show that the condition $d_{Lq_1}(\mathbf{c}_1, \mathbf{c}_1') \geqq q_1 d$ and $d_{Lq_2}(\mathbf{c}_2, \mathbf{c}_2') \geqq d$ implies the inequality

$$d_{Lq}(\mathbf{c}, \mathbf{c}') = \sum_{i=1}^{n} d_{Lq}(c_i, c_i') \geqq q_1 d.$$

If $\mathbf{c}_1 = \mathbf{c}_1'$, we have for each i $d_{Lq}(c_i, c_i') = q_1 d_{Lq_2}(c_{2i}, c_{2i}')$. Summing up these equalities we obtain the desired assertion. Therefore, we may assume that $\mathbf{c}_1 \neq \mathbf{c}_1'$. We shall show that we have for each i the inequality $d_{Lq}(c_i, c_i') \geqq d_{Lq_1}(c_{1i}, c_{1i}')$, which again implies by summation the assertion. Figure 1 helps us to do so. Since $d_{Lq}(c_i, c_i')$ is the length of the shorter circular arc joining the points c_i and c_i', throwing a glance in Fig. 1 we see that

$$d_{Lq}(c_i, c_i') \geqq \min\left(|c_{1i} - c_{1i}'|, q_1 - |c_{1i} - c_{1i}'|\right) = d_{Lq_1}(c_{1i}, c_{1i}').$$

As an obvious consequence we see that the number of elements of C is $|C_1| \cdot |C_2|$.

□

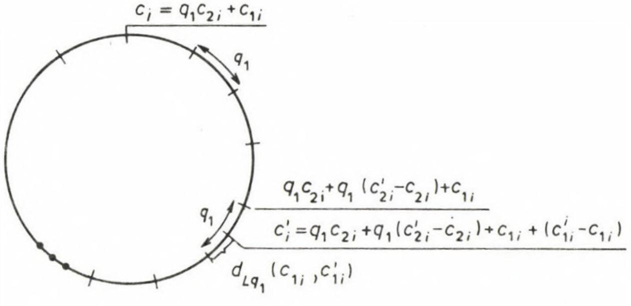

$c_i = q_1 c_{2i} + c_{1i}$
q_1
$q_1 c_{2i} + q_1 (c'_{2i} - c_{2i}) + c_{1i}$
q_1
$c'_i = q_1 c_{2i} + q_1 (c'_{2i} - c_{2i}) + c_{1i} + (c'_{1i} - c_{1i})$
$d_{Lq_1}(c_{1i}, c'_{1i})$

Fig. 1

The following example illustrates the construction.

Example. Let $C_1 = \{(0, 0, \ldots, 0), (1, 1, \ldots, 1)$ and $(2, 2, \ldots, 2)\}$, be a 9-dimensional code with minimal code distance 9 over an alphabet of 3 symbols. Let C_2 be a perfect one-error-correcting code with parameters $n = 9$, $q_2 = 19$. (Such a code exists according to [7].) Then the code obtained by the above construction, which consists of codewords of the form $\mathbf{c} = 3\mathbf{c}_2 + \mathbf{c}_1$, $\mathbf{c}_1 \in C_1$ and $\mathbf{c}_2 \in C_2$, i.e. altogether $3 \cdot 19^8$ codewords. This number is almost 5 times as big as the bound guaranteed by the estimate of Gilbert–Varshamov, and 22-times less than the Hamming bound. We recall the fact that in most of the known constructions the number of codewords is less than the Gilbert–Varshamov bound.

Remark 1. The above construction yields reasonably good codes also in the case of low dimension, even if C_1 consists of a single codeword.

Let $C_2 \in \mathscr{C}_L(n, 3, q_2)$ $(q_2 > 6)$ be a perfect code. According to [7], such a code exists if the dimension n has the form $n = \dfrac{\dfrac{q_2^r}{h} - 1}{2}$ for any natural numbers r and h. Then the above construction yields a code $C \in \mathscr{C}_L(n, 3q_1, q_1 q_2)$ with $\dfrac{q_2^n}{2n+1}$ elements. In the class $\mathscr{C}_L(n, 3q_1, q_1 q_2)$ the Gilbert–Varshamov bound guarantees the existence of a code with at least $\dfrac{(q_1 q_2)^n}{|V_L(n, 3q_1 - 1, q_1 q_2)|}$ codewords. We shall prove that to any dimension $n < 10$ there exists a number k_n such that

$$\frac{q_2^n}{2n+1} \geqq \frac{(q_1 q_2)^n}{|V_L(n, 3q_1 - 1, q_1 q_2)|}, \tag{2}$$

provided $q_1 \geqq k_n$. This will show that the number of elements of the code constructed in this way exceeds the Gilbert–Varshamov bound. Now we shall make use of equality (1):

$$|V_L(n, 3q_1-1, q_1q_2)| = \sum_{i=0}^{3q_1-1} 2^i \binom{n}{i}\binom{3q_1-1}{i}.$$

If $3q_1 - 1 \geqq n$, then we can estimate the sum on the right-hand side by neglecting the terms with $i<n$, to obtain the inequality

$$|V_L(n, 3q_1-1, q_1q_2)| \geqq 2^n \binom{3q_1-1}{n}.$$

If $n<10$ and q_1 sufficiently large, then

$$\frac{2^n}{(2n+1)n!}\,\frac{(3q_1-1)(3q_1-2)\dots(3q_1-n)}{q_1^n} =$$

$$= \frac{6^n}{(2n+1)n!}\left(1-\frac{1}{3q_1}\right)\dots\left(1-\frac{n}{3q_1}\right) > 1,$$

yielding the inequality $2^n\binom{3q_1-1}{n} \geqq (2n+1)q_1^n$.

This means

$$|V_L(n, 3q_1-1, q_1q_2)| \geqq (2n+1)q_1^n,$$

which is equivalent with (2).

Remark 2. Our construction allows us to extend perfect codes to larger alphabet. Let us consider the special case, when C_2 consists of all vectors of $Z_{q_2}^n$. Then the combination of a code $C_1 \in \mathscr{C}_L(n, d, q_1)$ and C_2 yields a code $C \in \mathscr{C}_L(n, d, q)$. C consists of those vectors $\mathbf{c} = \mathbf{c}_1 + q_1\mathbf{x}$ for which $\mathbf{c} \in C_1$ and $\mathbf{x} \in Z_{q_2}^n$. We shall prove that if C_1 is perfect, so does C. If $C_1 \in \mathscr{C}_L(n, d, q_1)$, then

$$|C_1| = \frac{q_1^n}{|V_L(n, e, q_1)|}.$$

The constructed code $C \in \mathscr{C}_L(n, d, q_2q_1)$ is perfect if and only if

$$|C| = \frac{(q_2q_1)^n}{|V_L(n, e, q_2q_1)|}.$$

It follows from the construction that $|C| = (q_2)^n \cdot |C_1|$, whence

$$|C| = \frac{(q_2q_1)^n}{|V_L(n, e, q_1)|}.$$

On the other hand we know that for $q_1 \geqq 2e+1$ the number of the points contained in a sphere does not depend on the alphabet: $|V_L(n, e, q_1)| = |V_L(n, e, q_2 q_1)|$. This proves the assertion.

Decoding. If there is a decoding algorithm for $C_1 \in \mathscr{C}_L(n, d, q_1)$, then the same is true for the code $C \in \mathscr{C}_L(n, d, q)$ constructed in Remark 2. This means that there is a procedure which associates with each $\mathbf{y} \in Z_q^n$ a codeword $\mathbf{c} \in C$ with minimal Lee-distance from $\mathbf{y}$. If the number of errors is less than $\frac{d-1}{2}$, then we obtain the true input word. Let $\mathbf{y} = \mathbf{y}_1 + q_1 \mathbf{y}_2$ be an arbitrary output word, where $\mathbf{y}_1 \in Z_{q_1}^n$ and $\mathbf{y}_2 \in Z_{q_2}^n$. Let $\mathbf{c}_1^*$ denote the codeword of minimal distance from $\mathbf{y}_1$ obtained by the decoding algorithm of C_1. Let $\mathbf{c} = \mathbf{c}_1 + q_1 \mathbf{x}$ be an arbitrary codeword. It is easy to see that

$$d_{Lq}(\mathbf{y}, \mathbf{c}) \geqq d_{Lq_1}(\mathbf{y}, \mathbf{c}) \geqq d_{Lq_1}(\mathbf{y}_1, \mathbf{c}_1) \geqq d_{Lq_1}(\mathbf{y}_1, \mathbf{c}_1^*).$$

It follows from the definition of d_{Lq_1} that

$$d_{Lq_1}(\mathbf{y}, \mathbf{c}_1^*) = \sum_{i=1}^{n} |\tilde{d}_{Lq_1}(y_{1i}, c_{1i}^*)| = \sum_{i=1}^{n} |y_{1i} - c_{1i}^* + k_i q_1| \tag{3}$$

where the value of k_i is uniquely given by the condition

$$|y_{1i} - c_{1i}^* + k_i q_1| \leqq \frac{q_1}{2}.$$

Thus we have

$$|\tilde{d}_{Lq_1}(y_{1i}, c_{1i}^*)| = |\tilde{d}_{Lq_1}(y_{1i}, c_{1i}^* - k_i q_1)| = |\tilde{d}_{Lq_1 q_2}(y_{1i}, c_{1i}^* - k_i q_1)|$$

and consequently $d_{Lq_1 q_2}(\mathbf{y}, \mathbf{c}) \geqq d_{Lq_1 q_2}(\mathbf{y}, \mathbf{c}_1^* + q_1 \mathbf{k})$, where $\mathbf{k} = (k_1, k_2, \ldots, k_n)$.

The decoding algorithm may be summarized as follows: We write the output word in the form $\mathbf{y} = \mathbf{y}_1 + q_1 \mathbf{y}_2$, determine the vector $\mathbf{c}^* \in C_1$ of minimal distance from $\mathbf{y}_1$, evaluate by (3) the k_i-s and obtain as input word $\mathbf{c}^* = \mathbf{c}_1^* + q_1(\mathbf{y}_2 - \mathbf{k})$.

Remark 3. We may construct codes in the Lee-metric by combination of several codes. In Remark 1 we have constructed a new code $C \in \mathscr{C}_L(n, q_1 d, q_1 q_2)$ by combination of the codes $C_1 \in \mathscr{C}_L(n, q_1 d, q_1)$ and $C_2 \in \mathscr{C}_L(n, d, q_2)$. We may well obtain by the combination of two arbitrary codes $C_1 \in \mathscr{C}_L(n, d_1, q_1)$ and $C_2 \in C_L(n, d_2, q_2)$ a new n-dimensional code over an alphabet of $q_1 q_2$ symbols which consists of the codewords $\mathbf{c} = q_1 \mathbf{c}_2 + \mathbf{c}_1$, $\mathbf{c}_1 \in C_1$ and $\mathbf{c}_2 \in C_2$, whose code-distance is equal to $\min(q_1 d_2, d_1)$. This procedure can be repeated. Starting with the codes

$$C_i \in \mathscr{C}_L(n, d_i, q_i), \qquad i = 1, \ldots, s,$$

we obtain a code

$$C \in \mathscr{C}_L\left(n, \min_{1 \leqq i \leqq s} (q_1 q_2 \cdots q_{i-1} d_i), q_1 q_2 \cdots q_s\right)$$

which consists of the codewords

$$\mathbf{c}=\sum_{i=1}^{s} q_1 q_2 \ldots q_{i-1}\mathbf{c}_i, \qquad \mathbf{c}_i \in C_i,\ q_0=1.$$

Combining first $C_1 \in \mathscr{C}_L(n, d_1, q_1)$ and $C_2 \in \mathscr{C}_L(n, d_2, q_2)$ we obtain the code

$$C' \in \mathscr{C}_L(n, \min(q_1 d_2, d_1), q_1 q_2)$$

with codewords $\mathbf{c}'=q_1\mathbf{c}_2+\mathbf{c}_1$. From the codes

$$C' \in \mathscr{C}_L(n, \min(q_1 d_2, d_1), q_1 q_2)$$

and $C_3 \in \mathscr{C}_L(n, d_3, q_3)$ we obtain

$$C'' \in \mathscr{C}_L(n, \min(q_1 q_2 d_3, q_1 d_2, d_1), q_1 q_2 q_3)$$

with codewords

$$\mathbf{c}''=q_1 q_2 \mathbf{c}_3+\mathbf{c}'=q_1 q_2 \mathbf{c}_3+q_1\mathbf{c}_2+\mathbf{c}_1,$$

and so on.

The case is of special interest when $q_1=q_2=\ldots=q_s=q$. Now the obtained code can be described also in the following way: We write $x \in Z_{q^s}$ in the form

$$x=x_0 q^0+x_1 q^1+\ldots+x_{s-1}q^{s+1},$$

and associate with x the column-vector

$$\begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{s-1} \end{pmatrix}.$$

The column-vectors belonging to the coordinates of $\mathbf{x} \in Z_{q^s}^n$ constitute an $s \times n$-matrix. Let $C_i \in Z_q^n$, $(i=1, \ldots, s)$, be codes. We consider the code $C \subset Z_{q^s}^n$ consisting of the vectors

$$\mathbf{c}=\mathbf{c}_1+q\mathbf{c}_2+\ldots+q^{s-1}\mathbf{c}_s \qquad (\mathbf{c}_i \in C_i).$$

In the above matrix representation of $\mathbf{c}$ the i-th row is an element of C_i.

In the case under consideration our construction is nothing else but the construction of the densest sphere-packing of Leech and Sloane adapted to the Lee-metric. This construction results in reasonably good codes provided the values $q^{i-1}d_i$ are approximately equal. The applicability of this construction is bounded by the fact that we know only relatively few codes which can play the role of the codes C_i. In the case when $q=2$, the Lee- and Hamming-distances coincide. Thus we can obtain codes in the Lee-metric over an alphabet of 2^s symbols by combination of codes in the Hamming-metric.

First we apply the above construction to the Reed–Müller codes. It is known that to arbitrary integers m and r $(0 \leqq r \leqq m)$ a so-called Reed–Müller code C_r of order r exists with the following parameters:

codeword length:	$n=2^m$,
number of elements of C_r:	$\lvert C_r \rvert = 2^{1+\binom{m}{1}+\ldots+\binom{m}{r}}$,
minimum code-distance:	$d_r = 2^{m-r}$.

The choice of m, of course, also specifies the length of the codewords. Varying r we obtain codes with different minimal code-distances. If we choose these codes in a suitable way then we can make the values $2^{i-1}d_i$, $i=1, \ldots, s$, to be equal.

Let e.g. $C_i \in \mathscr{C}_H(n, 2^{m-(r_1+i)}, 2)$ be Reed–Müller codes of order r_1+i, $i=1, \ldots, r_2-r_1$, and let us consider the matrix

$$\begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{r_2-r_1\,1} & c_{r_2-r_1\,n} & \ldots & c_{r_2-r_1\,n} \end{pmatrix} = (c_1, \ldots, c_j, \ldots, c_n),$$

where $(c_{i1}, c_{i2}, \ldots, c_{in}) \in C_i$.

Let C consist of the vectors $\mathbf{c}=(c_1, \ldots, c_j, \ldots, c_n)$, where

$$c_j = c_{1j} + 2c_{2j} + \ldots + 2^{r_2-r_1-1} c_{r_2-r_1\,j}.$$

Obviously, C is defined over an alphabet of $2^{r_2-r_1}$ symbols. The code-distance of C in the Lee-metric is

$$d = \min\left(2^{r_2-r_1-1} \cdot 2^{m-r_2}, 2^{r_2-r_1-2} \cdot 2^{m-(r_2-1)}, \ldots, 2 \cdot 2^{m-(r_1+2)}, 2^{m-(r_1+1)}\right) =$$
$$= 2^{m-(r_1+1)}.$$

The number of elements of C is equal to

$$\lvert C \rvert = 2^{1+\binom{m}{1}+\ldots+\binom{m}{r_1+1}+1+\binom{m}{1}+\ldots+\binom{m}{r_1+2}+\ldots+1+\binom{m}{1}+\ldots+\binom{m}{r_2}}$$

References

1. *Astola, J.*, The theory of Lee-codes, Lappeenranta University of Technology, Research Report 1, 1982.
2. *Astola, J.*, On the non-existence of certain perfect Lee-error-correcting codes, Ann. Univ. Turku. Ser. *A1*, 167, 1975.
3. *Leech, J.*, *Sloane, N. J. A.*, Sphere packing and error-correcting codes, Canad. J. Math. **23**, 718–745.
4. *MacWilliams, F. J.*, *Sloane, J. A.*, The theory of error-correcting codes, North-Holland Publ. Co., 1978.
5. *Post, K. A.*, Nonexistence theorems on perfect Lee-codes over large alphabets, Inform. Contr. **29**, 369–380, 1975.
6. *Racsmány, A.*, Perfect-Single-Lee-error-correcting Codes, Studia Sci. Math. Hungar. **9** (1974), 73–75.
7. *Racsmány, A.*, Correction to my paper "Perfect-Single-Lee-error-correcting Codes", Studia Sci. Math. Hungar., to appear.
8. *Sloane, N. J. A.*, Sphere packings constructed from BCH and Justesen codes, Mathematika **19**, 183–190.

О конструкции кодов с заданным расстоянием в Ли-метрик

А. РАЧМАНЬ

(Будапешт)

Пусть $\mathscr{C}_L(n, d, q)$ обозначает класс размерности кодов n над алфавитом размерности q с минимальным кодовым расстоянием d в Ли-метрик. В этой работе мы предлагаем метод для конструкции кода $C \in \mathscr{C}_L(n, \min\limits_{1 \leqq i \leqq s} (q_1 q_2 \dots q_{i-1} d_i), q_1 q_2 \dots q_s)$ при комбинации кодов $C_i \in \mathscr{C}_L(n, d_i, q_i)$ $i = 1, \dots, s$. Это позволяет нам найти хорошие коды над большим алфавитом при использовании хороших кодов над малыми алфавитами.

A. Racsmány
Karl Marx University of Economics
H-1093 Budapest, IX. Dimitrov tér 8.
Budapest, Hungary

Problems of Control and Information Theory, Vol. 15 (5), pp. 385—397 (1986)

ON OPTIMAL CONTROL PROBLEM FOR PARABOLIC–HYPERBOLIC SYSTEM

A. KOWALEWSKI

(*Kraków*)

(Received July 20, 1985)

The purpose of this paper is to show Milutin–Dubovicki's method for solving optimal control problems for distributed parameter systems.

As an example an optimization problem for parabolic–hyperbolic system is considered.

Making use of the Milutin–Dubovicki's method necessary and sufficient conditions of optimality for Dirichlet problem with a quadratic performance functional and constrained control are derived.

The flow chart of the algorithm which can be used in the numerical solving of certain optimization problems for parabolic–hyperbolic systems is also presented.

1. Preliminaries

At the beginning we consider some evolution equation.

We have to find a function $y(x_0, t)$ which fulfils the following conditions (Theorem 9.1 [5]):

$$y \in L^2(\Theta; V) \cap D(\Lambda, L^2(\Theta; V')) \tag{1}$$

and is the solution of the equation

$$\Lambda y + A(x_0, t)y = f, \qquad \{x_0, t\} \in \Theta \tag{2}$$

where

$$\Theta = (0, a_0) \times (0, T), \qquad a_0 < \infty, T < \infty$$

$$V \subset H \subset V', \quad f \in L^2(\Theta; V'), \quad A(x_0, t) \in \mathscr{L}(V, V').$$

The operator is defined by

$$\Lambda y = \frac{\partial y}{\partial t} + \frac{\partial y}{\partial x_0} \tag{3}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

with

$$D(\Lambda; L^2(\Theta; V')) = \left\{ y \,|\, y \in L^2(\Theta; V'), \right.$$

$$\left. \frac{\partial y}{\partial t} + \frac{\partial y}{\partial x_0} \in L^2(\Theta; V'), \; y(x_0, 0) = 0, \; y(0, t) = 0 \right\}. \tag{4}$$

Let $a(x_0, t, \varphi, \psi)$ be a family of bi-linear forms on V with

$$\left. \begin{array}{ll} x_0, t \to a(x_0, t, \varphi, \psi) & \text{bounded measurable on } \Theta \\ a(x_0, t; \varphi, \psi) \geqq \alpha \|\varphi\|_V^2 & \text{a.e. in } \Theta, \; \forall \varphi \in V, \; \alpha > 0 \end{array} \right\}. \tag{5}$$

Then if $\psi \in L^2(\Theta; V)$, the operator "$x_0, t \to A(x_0, t)\psi(x_0, t)$" is defined by

$$(A(x_0, t)\varphi_1, \varphi_2) = a(x_0, t; \varphi_1, \varphi_2) \qquad \forall \varphi_1, \varphi_2 \in V. \tag{6}$$

We assume that (5) holds. Then exists a unique function $y(x_0, t)$ (Theorem 9.2 [5]) having the following properties:

$$y \in L^2(\Theta; V), \qquad \frac{\partial y}{\partial t} + \frac{\partial y}{\partial x_0} \in L^2(\Theta; V') \tag{7}$$

which satisfies (2) with the initial conditions

$$y(x_0, 0) = 0 \qquad x_0 \in (0, a_0) \tag{8}$$

$$y(0, t) = 0 \qquad t \in (0, T). \tag{9}$$

We shall now extend our results to the case where the initial condition (8) is non-zero.

We can prove (Lemma 9.1 [5], Theorem 9.3 [5]):

Let $f, y_p(x_0)$ be given with $f \in L^2(\Theta; V')$ and $y_p(x_0) \in L^2(0, a_0, H)$. Then there exists a function $y = y(x_0, t)$ which is unique having the following properties: y satisfies (2), (7), (9) and

$$y(x_0, 0) = y_p(x_0). \tag{10}$$

2. Statement of Optimal Control Problem. Optimization Theorem

Now we shall formulate optimal control problem in the context of Theorem 9.3 [5]. Let us take

$$V = H_0^1(\Omega), \quad H = L^2(\Omega), \quad V' = H^{-1}(\Omega)$$

where Ω 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.

We consider the parabolic–hyperbolic equation describing the dynamics of controlled system

$$\frac{\partial y}{\partial t}+\frac{\partial y}{\partial x_0}+A\left(x,x_0,t,\frac{\partial}{\partial x}\right)y=f+u \qquad x\in\Omega,\ \{x_0,t\}\in\Theta \tag{11}$$

$$y(x,x_0,t)=0 \qquad x\in\Gamma,\ \{x_0,t\}\in\Theta \tag{12}$$

$$y(x,x_0,0)=y_p(x,x_0) \qquad x_0\in(0,a_0),\ x\in\Omega \tag{13}$$

$$y(x,0,t)=0 \qquad t\in(0,T),\ x\in\Omega \tag{14}$$

where

$y\equiv y(x,x_0,t),\ u\equiv u(x,x_0,t),\ f\equiv f(x,x_0,t)$
$y(\cdot,x_0,0)\in L^2(\Omega),\ y(\cdot,0,t)\in L^2(\Omega)$
$f\in L^2(\Theta;H^{-1}(\Omega))$.

The operator $A\left(x,x_0,t,\frac{\partial}{\partial x}\right)$ in (11) has the form

$$A\left(x,x_0,t,\frac{\partial}{\partial x}\right)y=-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x,x_0,t)\frac{\partial}{\partial x_j}\right)$$

and the functions $a_{ij}(x,x_0,t)$ satisfy conditions in $\Omega\times\Theta$

$$a_{ij}(x,x_0,t)\in L^\infty(\Omega\times\Theta)$$

$$\sum_{i,j=1}^{n}a_{ij}(x,x_0,t)\xi_i\xi_j\geqq\alpha\sum_{i=1}^{n}\xi_i^2,\qquad \alpha>0,\ \forall\xi_i\in R.$$

We must remark that the operator $\frac{\partial}{\partial t}+\frac{\partial}{\partial x_0}+A\left(x,x_0,t,\frac{\partial}{\partial x}\right)$ in equation (11) is the "combination" of the hyperbolic operator $\frac{\partial}{\partial t}+\frac{\partial}{\partial x_0}$ and the parabolic operator $\frac{\partial}{\partial t}+A\left(x,x_0,t,\frac{\partial}{\partial x}\right)$. Applications of these systems will be considered elsewhere. Also it is easy to notice that equations (11)–(14) can be written in the following form

$$P(y,u)=0. \tag{15}$$

Let us denote by $Y=L^2(\Theta;H_0^1(\Omega))$ the space of states, and by $U=L^2(\Omega\times\Theta)$ the space of controls.

The control time T is fixed and the observation is distributed in the set $\Omega\times\Theta$ in our problem.

The performance functional is given by

$$I(y,u)=\langle y-z_d,y-z_d\rangle_{L^2(\Omega\times\Theta)}+\langle Nu,u\rangle_{L^2(\Omega\times\Theta)} \tag{16}$$

where

z_d is a given element in $L^2(\Omega \times \Theta)$,

$N \in \mathscr{L}(U, U)$ is a hermitian operator which performed the coercive condition

$$\langle Nu, u\rangle_U \geqq \nu \|u\|_U^2 \qquad \forall\, u \in U, \quad \nu > 0.$$

We assume the following constraints on controls:

$$u \in U_{ad} \text{ is a closed, convex set with non-empty enterior, a subset of } U. \tag{17}$$

Making use of Milutin–Dubovicki's Theorem we shall derive the necessary and sufficient conditions of optimality for the optimization problem (15)–(17).

The idea of Milutin–Dubovicki's method was particularly described in [3, 4], therefore we shall not present this method here.

The solving of the stated optimal control problem is equivalent to seeking a couple $(y_0, u_0) \in Y \times U$ which satisfies equation (15) and minimizing the performance functional (16) with constraints on controls (17).

We formulate the necessary and sufficient conditions of the optimality in the following form:

Theorem 1. The solution of the optimization problem (15)–(17) exists and it is unique by the assumptions mentioned above and the necessary and sufficient conditions of the optimality are characterized by the following system of partial differential equations and inequalities: equation of the system control

$$P(y_0, u_0) = 0 \tag{18}$$

adjoint equation

$$-\frac{\partial p}{\partial t} \div \frac{\partial p}{\partial x_0} + A^*\left(x, x_0, t, \frac{\partial}{\partial x}\right) p = y_0 - z_d \qquad x \in \Omega, \{x_0, t\} \in \Theta \tag{19}$$

$$p(x, x_0, t) = 0 \qquad x \in \Omega, \{x_0, t\} \in \Theta \tag{20}$$

$$p(x, x_0, T) = 0 \qquad x_0 \in (0, a_0), x \in \Omega \tag{21}$$

$$p(x, a_0, t) = 0 \qquad t \in (0, T), x \in \Omega \tag{22}$$

maximum condition

$$\langle p + Nu_0, u - u_0\rangle_{L^2(\Omega \times \Theta)} \geqq 0 \qquad \forall\, u \in U_{ad}. \tag{23}$$

Proof. According to the theorem of Milutin–Dubovicki 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

$$Q_1 = \{(y, u) \in Y \times U;\, P(y, u) = 0\}. \tag{24}$$

We construct the regular tangent cone of the set Q_1 using the Lusternik theorem (Theorem 9.1 [3]).

For this purpose we define the operator P in the form

$$P(y, u) = \left\{ \frac{\partial y}{\partial t} + \frac{\partial y}{\partial x_0} + A\left(x, x_0, t, \frac{\partial}{\partial x}\right) y - f - u, \right.$$

$$\left. y(x, x_0, t),\, y(x, x_0, 0) - y_p(x, x_0),\, y(x, 0, t) \right\}. \tag{25}$$

The operator P is the mapping from the space $L^2(\Theta;\, H_0^1(\Omega)) \times L^2(\Omega \times \Theta)$ into the space $L^2(\Theta;\, H^{-1}(\Omega)) \times L^2(\Theta;\, H_0^1(\Omega)) \times L^2(\Omega) \times L^2(\Omega)$.

We write down the Fréchet differential of the operator P in the following form

$$P'(y_0, u_0)(\bar{y}, \bar{u}) = \left(\frac{\partial \bar{y}}{\partial t} + \frac{\partial \bar{y}}{\partial x_0} + A\left(x, x_0, t, \frac{\partial}{\partial x}\right) \bar{y} - \bar{u}, \right.$$

$$\left. \bar{y}(x, x_0, t),\, \bar{y}(x, x_0, 0),\, \bar{y}(x, 0, t) \right). \tag{26}$$

Really, $\frac{\partial}{\partial t} + \frac{\partial}{\partial x_0}$ is the linear bounded mapping (Theorem 2.8 [7]) and the operator $A\left(x, x_0, t, \frac{\partial}{\partial x}\right)$ is bounded too (Theorem 9.1 [5]).

Using Theorem 9.3 [5] we can prove that P' is the operator "one to one" from the space $L^2(\Theta;\, H_0^1(\Omega)) \times L^2(\Omega \times \Theta)$ onto $L^2(\Theta;\, H^{-1}(\Omega)) \times L^2(\Theta;\, H_0^1(\Omega)) \times L^2(\Omega) \times L^2(\Omega)$.

Considering that the assumptions of Lusternik's theorem are fulfilled, we can write down the regular tangent cone for the set Q_1 in a point (y_0, u_0) in the form

$$RTC(Q_1, (y_0, u_0)) = \{(\bar{y}, \bar{u}) \in Y \times U;\, P'(y_0, u_0)(\bar{y}, \bar{u}) = 0\}. \tag{27}$$

It is easy to notice that it is a subspace. Therefore, using Theorem 10.1 [3], we know the form of the functional belonging to the adjoint cone

$$f_1(\bar{y}, \bar{u}) = 0 \qquad \forall (\bar{y}, \bar{u}) \in RTC(Q_1, (y_0, u_0)).$$

b) Analysis of the Constraints on Controls

Using Theorem 10.5 [3] we approximate the set $Q_2 = Y \times U_{ad}$ which represents inequality constraints by the regular admissible cone $RAC(Q_2, (y_0, u_0))$: The functional f_2 which belongs to the adjoint cone $[RAC(Q_2, (y_0, u_0))]^*$ has the form

$$f_2(\bar{y}, \bar{u}) = (f'_1, f'_2)$$

where

$f'_1(\bar{y}) = 0 \qquad \forall\, \bar{y} \in Y$ (Theorem 10.1 [3]),

$f'_2(\bar{u})$ is a support functional to the set U_{ad} in a point u_0 (Theorem 10.5 [3]).

c) Analysis of the Performance Functional

On the basis of Theorem 7.5 [3] we construct the regular improvement cone $RFC(I, (y_0, u_0))$ for the performace functional (16).

The Fréchet differential $I'(y_0, u_0)(\bar{y}, \bar{u})$ has the form

$$I'(y_0, u_0)(\bar{y}, \bar{u}) = 2\langle y_0 - z_d, \bar{y}\rangle_{L^2(\Omega\times\Theta)} + 2\langle Nu_0, \bar{u}\rangle_{L^2(\Omega\times\Theta)}.$$

By virtue of Theorem 7.5 [3], we get

$$RFC(I, (y_0, u_0)) = \{(\bar{y}, \bar{u}) \in Y \times U;\, I'(y_0, u_0)(\bar{y}, \bar{u}) < 0\} \tag{28}$$

and $f_3(\bar{y}, \bar{u}) \in [RFC(I, (y_0, u_0))]^*$ is

$$f_3(\bar{y}, \bar{u}) = -\lambda_0 \langle y_0 - z_d, \bar{y}\rangle_{L^2(\Omega\times\Theta)} - \lambda_0 \langle Nu_0, \bar{u}\rangle_{L^2(\Omega\times\Theta)} \tag{29}$$

where $\lambda_0 \geqq 0$.

It is easy to notice that $\lambda_0 \neq 0$ if Slater's condition is fulfilled (Theorem 11.3 [3]). Then $\lambda_0 > 0$.

d) Analysis of the Euler–Lagrange Equation

The Euler–Langrange equation for our optimization problem has the form

$$f'_2(\bar{u}) = \lambda_0 \int_{\Omega\times\Theta} (y_0 - z_d)\bar{y}\, dx\, dx_0\, dt + \lambda_0 \langle Nu_0, \bar{u}\rangle_{L^2(\Omega\times\Theta)}$$

$$\forall (\bar{y}, \bar{u}) \in RTC(Q_1, (y_0, u_0)). \tag{30}$$

We transform the first component of the right-hand side of (30) introducing the adjoint variable by equation (19) and using the formulas

$$\bar{y}(x, x_0, t) = 0, \quad \bar{y}(x, x_0, 0) = 0, \quad \bar{y}(x, 0, t) = 0,$$

$$\frac{\partial \bar{y}}{\partial t} + \frac{\partial \bar{y}}{\partial x_0} + A\left(x, x_0, t, \frac{\partial}{\partial x}\right)\bar{y} = \bar{u}.$$

Then using Lemma 9.2 [5] we get

$$\lambda_0 \int\limits_{\Omega\times\Theta} (y_0 - z_d)\bar{y}\, dx\, dx_0\, dt = \lambda_0 \int\limits_{\Theta} \langle y_0 - z_d, \bar{y}\rangle_{L^2(\Omega)}\, dx_0\, dt =$$

$$= \lambda_0 \int\limits_{\Theta} \left\langle -\frac{\partial p}{\partial t} - \frac{\partial p}{\partial x_0} + A^*\left(x, x_0, t, \frac{\partial}{\partial x}\right) p, \bar{y} \right\rangle_{L^2(\Omega)} dx_0\, dt =$$

$$= \lambda_0 \int\limits_{\Theta} \left\langle p, \frac{\partial \bar{y}}{\partial t} + \frac{\partial \bar{y}}{\partial x_0} + A\left(x, x_0, t, \frac{\partial}{\partial x}\right) \bar{y} \right\rangle_{L^2(\Omega)} dx_0\, dt =$$

$$= \lambda_0 \int\limits_{\Theta} \langle p, \bar{u}\rangle_{L^2(\Omega)}\, dx_0\, dt = \lambda_0 \langle p, \bar{u}\rangle_{L^2(\Omega\times\Theta)}. \tag{31}$$

Taking into account equality (31) we can write down (30) in the form

$$f'_2(\bar{u}) = \lambda_0 \langle p + Nu_0, \bar{u}\rangle_{L^2(\Omega\times\Theta)}. \tag{32}$$

Using the definition of the support functional [3] we get

$$\lambda_0 \langle p + Nu_0, u - u_0\rangle_{L^2(\Omega\times\Theta)} \geqq 0 \qquad \forall\, u \in U_{ad}. \tag{33}$$

Dividing both members of the last inequality by λ_0, we finally get the maximum condition

$$\langle p + Nu_0, u - u_0\rangle_{L^2(\Omega\times\Theta)} \geqq 0 \qquad \forall\, u \in U_{ad}. \tag{34}$$

In order to prove the sufficiency of the derived conditions of the optimality, we use the fact that constraints and the performace functional are convex and exists a point $(\tilde{y}, \tilde{u}) \in \operatorname{int} Q_2$ such that $(\tilde{y}, \tilde{u}) \in Q_1$ (Theorem 15.2 [3]).

This fact follows immediately from the existence of non-empty interior in the set Q_2 and from the existence of the solution of equations (11)–(14) as well.

It is easy to notice that:

— the existence of the optimal control u_0 follows from the form of the set U_{ad} (17) and the maximum condition (34),

— the uniqueness of u_0 follows from the strict convexity of the performance functional (16).

The remarks mentioned above complete the proof of Theorem 1.

Also, one may consider a similar problem with local constraints on controls $u \in U_{ad}$, where U_{ad} is a bounded set in $L^\infty(\Omega\times\Theta)$. This problem will be discussed elsewhere.

3. Conclusions and Certain Example of Optimal Control Problem

We must remark that the optimal conditions we have derived above, allow us to obtain an analytical formula for the optimal control in particular cases (e.g. there are no constraints on controls).

For the special case where $U_{ad}=U$, (34) is satisfied when $u^0=-N^{-1}p(u^0)$.

It results from the following, determining the function $p(x, x_0, t)$ in the maximum condition is possible from the adjoint equation, if and only if we know $y_0(x, x_0, t)$ will suit the control $u_0(x, x_0, t)$.

These mutual connections make the practical usage of the derived optimization formulas difficult. So that we resign from the exact determining of the optimal control and we use approximation methods.

For instance, using the penalty function method we can solve the optimal control problem with quadratic performance functional (16) (if we assume that $N\neq 0$ and $U_{ad}=U$).

In the case of non-coercive performance functional (in formula (16) $N=0$) the optimal control problem (15)–(17) will be reduced to the minimizing of the functional on a closed and convex subset in a Hilbert space.

Then the optimal control problem is equivalent to a quadratic programming one which can be solved by the use of the well-known algorithms, e.g. Gilbert's [2], Barr's [1] or Nahi–Wheeler's [8] ones.

To illustrate the remarks mentioned above we shall formulate the following control problem as an example

equation of the system control

$$\frac{\partial y}{\partial t}+\frac{\partial y}{\partial x_0}+Ay=u \qquad x\in\Omega,\ \{x_0, t\}\in\Theta \tag{35}$$

$$y(x, x_0, 0)=y_p(x, x_0) \qquad x\in\Omega,\ x_0\in(0, a_0) \tag{36}$$

$$y(x, 0, t)=0 \qquad x\in\Omega,\ t\in(0,T) \tag{37}$$

$$y(x, x_0, t)=0 \qquad x\in\Gamma,\ \{x_0, t\}\in\Theta \tag{38}$$

performance functional

$$I(y, u)=\|y-z_d\|^2_{L^2(\Omega\times\Theta)} \tag{39}$$

constraint on controls

$$U_{ad}=\{u\in L^2(\Theta; L^2(\Omega)), \|u(x, x_0, t)\|_{L^2(\Omega\times\Theta)}\leqq 1\}. \tag{40}$$

We shall define the attainable set Y_{ad}

$$Y_{ad}=\left\{y\in Y;\frac{\partial y}{\partial t}+\frac{\partial y}{\partial x_0}+Ay=u,\ y(x,x_0,0)=y_p(x,x_0),\right.$$
$$\left.y(x,0,t)=0,\ u\in U_{ad}\right\}. \tag{41}$$

It can be shown that the set Y_{ad} is a closed, convex and a bounded one in the space $Y=L^2(\Theta;H_0^1(\Omega))$. The proof of this fact is obtained in a similar way as in the case of parabolic equation which is given in [6].

Control problem (35)–(40) can be considered as a one for seeking an element y_0, more precisely the corresponding u_0 belonging to a closed, convex and a bounded set Y_{ad} in a certain Hilbert space whose distance from a given element z_d is minimal. Thus it is a quadratic programming problem in a Hilbert space.

Now, we shall describe a certain iteration procedure for solving a quadratic programming problem.

Let $\{Y_{ad}^i\}$ be a system of closed and convex subsets of the set Y_{ad}. We denote by $Y^i\in Y_{ad}^i$ an element whose distance from the element z_d is minimal, i.e. the condition

$$\|y^i-z_d\|=\min_{y\in Y_{ad}^i}\|y-z_d\| \tag{42}$$

is fulfilled.

By $\bar{y}^{i+1}$ we denote the element such that

$$\langle y^i-z_d,y-\bar{y}^{i+1}\rangle\geqq 0\qquad\forall\, y\in Y_{ad}. \tag{43}$$

The point $\bar{y}^{i+1}$ is a support one of the set Y_{ad} determining by the hyperplane M^i which is orthogonal to the vector (z_d-y^i).

In [6] it is shown that if the system of sets $\{Y_{ad}^i\}$ has the following structure

$$Y_{ad}^{i+1}\supset y^i\cup y^{i+1} \tag{44}$$

then the sequence $\{y^i\}$ is strongly convergent to y_0 in the space Y.

One-by-one algorithms for finding the sequence $\{y^i\}$ convergent to y_0 differ from each other by the constructing of the sets Y_{ad}^i, only. The simplest one of them has been proposed by Gilbert in [2].

In Gilbert's method in the $(i+1)$-th iteration the set Y_{ad}^{i+1} is a section joining the points y^i and $\bar{y}^{i+1}$, and on it we are looking for a point y^{i+1} lying in the minimal distance from the point z_d. So, determining y^{i+1} is reduced to looking for a number $\alpha\in[0,1]$ such as to satisfy

$$\|y^{i+1}-z_d\|=\min_{\alpha\in[0,1]}\|y^\alpha-z_d\| \tag{45}$$

where $y^{\alpha}=(1-\alpha)y^i+\alpha\bar{y}^{i+1}$.

It can be proved that

$$y^{i+1}=(1-\alpha^{i+1})y^i+\alpha^{i+1}\bar{y}^{i+1} \tag{46}$$

where

$$\alpha^{i+1}=\min\left\{\frac{\langle y^i-z_d, y^i-\bar{y}^{i+1}\rangle}{\langle y^i-\bar{y}^{i+1}, y^i-\bar{y}^{i+1}\rangle}, 1\right\}. \tag{47}$$

In the first iteration an arbitrary point of the set Y_{ad} is taken as point y^0. The error δ in the i-th iteration can be evaluated by the formula

$$\delta=\|y^i-z_d\|-\Delta \tag{48}$$

where

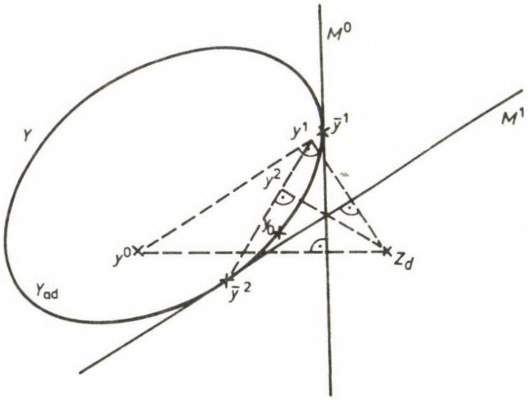
$$\Delta=\frac{\langle\bar{y}^{i+1}-z_d, y^i-z_d\rangle}{\|y^i-z_d\|}$$

is the distance between the point z_d and the hyperplane M^i.

Computing is over when the error δ does not exceed a fixed number ε.

The shortcoming of the described method is a very slow speed of convergence. In this respect Barr's [1] and Nahi–Wheeler's [8] algorithms are better.

The operation of Gilbert's algorithm is shown in Fig. 1.



y^0 — initial point

Y_0 — optimal point

$\bar{y}^1, \bar{y}^2$ — support points of the set Y_{ad} determined by the hyperplanes M^0 and M^1 suitable

Fig. 1

To determine the point y^{i+1} it is necessary to know the element $\bar{y}^{i+1}$ satisfying condition (43).

At present we describe the method of determining the element $\bar{y}^{i+1}$ for the optimal control problem (35)–(40).

We shall introduce the following notation

$$y^i = y(u^i), \quad \bar{y}^{i+1} = \bar{y}(u^{i+1}), \quad p^i = p(u^i). \tag{49}$$

The adjoint variable p^i is given by the equation

$$-\frac{\partial p^i}{\partial t} - \frac{\partial p^i}{\partial x_0} + A^* p^i = y^i - z_d \qquad x \in \Omega, \{x_0, t\} \in \Theta \tag{50}$$

$$p^i(x, x_0, T) = 0 \qquad x \in \Omega, x_0 \in (0, a_0) \tag{51}$$

$$p^i(x, a_0, t) = 0 \qquad x \in \Omega, t \in (0, T) \tag{52}$$

$$p^i(x, x_0, t) = 0 \qquad x \in \Gamma, \{x_0, t\} \in \Theta. \tag{53}$$

Proceeding in a similar way as in deriving formula (34), condition (43) is written in the following way

$$\langle p^i, u - \bar{u}^{i+1} \rangle_{L^2(\Omega \times \Theta)} \geqq 0 \qquad \forall\, u \in U_{ad}. \tag{54}$$

Taking into account the form of the set U_{ad} from formula (54), we get

$$\bar{u}^{i+1} = -\frac{p^i}{\|p^i\|_{L^2(\Omega \times \Theta)}}. \tag{55}$$

Now, it is easy to notice that there are no mutual connections between the equation of the system control, the adjoint equation and the maximum condition which made impossible the determination of the optimal control, earlier.

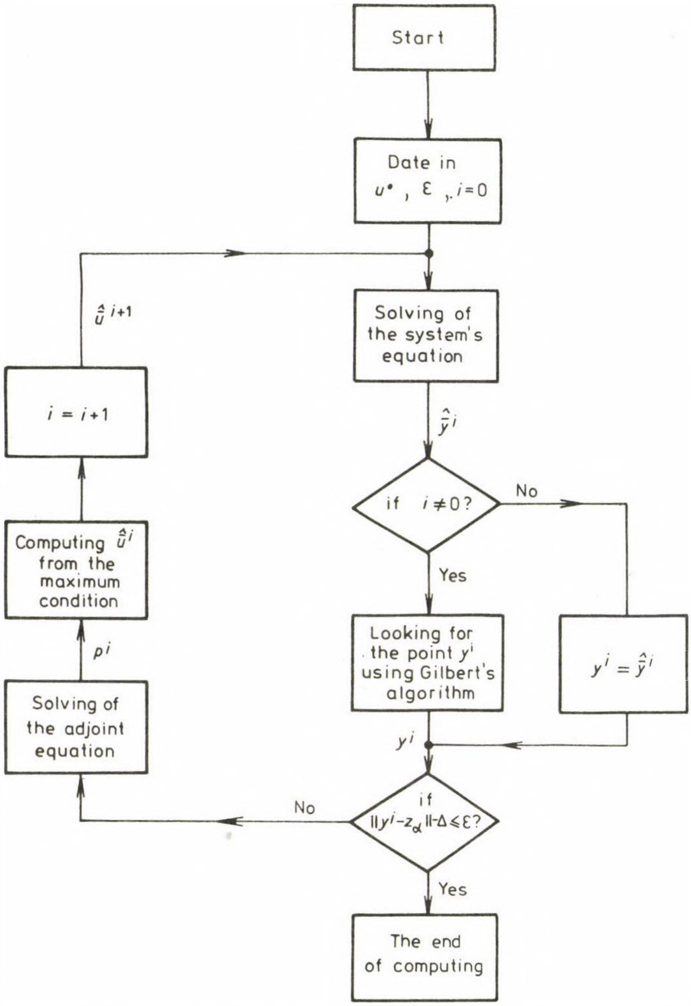
So that from formula (55) we find out $\bar{u}^{i+1}$ for p^i which we determine from equations (50)–(53) knowing y^i from the previous iteration.

Then, having $\bar{u}^{i+1}$ we compute $\bar{y}^{i+1}$ from equations (35)–(38).

In this case we can show the course of calculations on the flow chart (Fig. 2).

We see that in every iteration the partial differential equation must be solved twice.

Because, it is impossible to find the solution of these equations analytically, usúally we use some approximation methods e.g. the Fourier one, as it is shown in [6].


Start
Date in
u°, ε, $i=0$
Solving of the system's equation
$\hat{\hat{y}}^i$
if $i \neq 0$?
No
Yes
Looking for the point y^i using Gilbert's algorithm
$y^i = \hat{\hat{y}}^i$
y^i
if $\|y^i - z_\alpha\| - \Delta \leqslant \varepsilon$?
No
Yes
The end of computing
Solving of the adjoint equation
p^i
Computing $\hat{\hat{u}}^i$ from the maximum condition
$i = i+1$
$\hat{\hat{u}}^{i+1}$

u° is an arbitrary initial control ($u^\circ \in U_{a,d}$)
$\| y^i - z_\alpha \| - \Delta$ is an error in the i-th iteration

Fig. 2

References

1. *Barr, R. O.*, On efficient computational procedure for a generalized quadratic programming problem. SIAM Journal on Control, **7** (1969), pp. 415–429.
2. *Gilbert, E. S.*, An iterative procedure for computing the minimum of a quadratic form on a convex set. SIAM Journal on Control, **4** (1966), pp. 61–80.
3. *Girsanov, I. V.*, Lectures on the Mathematical Theory of extremal problems. Publication University of Moscow 1970 (in Russian).
4. *Kowalewski, A.*, *Kotarski, W.*: On the application of conical approximations to an optimal control problem for systems described by partial differential equations of parabolic type with time delay. Problems of Control and Information Theory, vol. **10** (*5*), pp. 341–351 (1981).
5. *Lions, J. L.*, Optimal Control of Systems Governed by Partial Differential Equations. Springer–Verlag, Berlin, Heidelberg, New York 1971.
6. *Malanowski, K.*, Some applications of Fourier method to optimal control problems for systems described by partial differential equations. Papers of the Institute of Organization and Management. Series B, Bulletin 3, Warsaw 1974 (in Polish).
7. *Maslov, V. P.*, Operators Methods. Publication "Science", Moscow 1973 (in Russian).
8. *Nahi, N. E.*, *Wheeler, L. A.*, Optimal terminal control of continuous systems via successive approximation of the reachable set. IEEE Transaction on Automatic Control, vol. **AC–12** (1967), pp. 515–521.

Об одной задаче оптимального управления для параболического–гиперболического объекта

А. КОВАЛЕВСКИ
(Краков)

В работе доказывается пригодность схемы Милютина–Дубовицкого для решения задач оптимального управления для объектов с распределенными параметрами.

В качестве примера рассматривается проблема оптимизации для параболическо–гиперболического объекта.

Исследуя метод Милютина–Дубовицкого для решения задачи Дирихлета, выводятся необходимые и достаточные условия оптимальности при квадратной функции стоимости и ограничении управления.

Приводится также блок-схема алгоритма, который можно использовать при численном решении некоторых вопросов оптимизации для параболическо–гиперболических объектов.

A. Kowalewski
Institute of Automatic Control, Systems Engineering and Telecommunications,
Stanisław Staszic University of Mining and Metallurgy,
Al. Mickiewicza 30, 30–059 Kraków, Poland

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda, Budapest

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

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). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = \|a_{ij}\|$. Write diagonal matrices as diag $(d_1, d_2, \dots d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 129090 Москва И-90, ул. Щепкина, 8. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 208-60-19).

Объём 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 языке в трех экземплярах, 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 и 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 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

HU ISSN 0370-2529

Index: 26 660

VOL. / ТОМ 15 · NUMBER / НОМЕР 6

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

1986

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
— theory of adaptive systems
— theory of estimation and identification
— theory of controlling robot-technologic and flexible manufacturing systems
— information theory
— information-theoretic aspects of multiple access networks.

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;
— теория адаптивных систем;
— теория оценивания и идентификации;
— теория управления робототехническими и гибкими производственными системами;
— теория информации;
— теория информации в области сетей с множественным доступом.

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 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1986 Subscription Rate US Dollars 175.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

1828

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

Problems of Control and Information Theory, Vol. 15 (6), pp. 399—412 (1986)

STACKELBERG STRATEGIES FOR HIERARCHICAL DIFFERENTIAL TWO-PERSON GAMES

A. F. Kleimenov

(Sverdlovsk)

(Received August 20, 1985)

This paper presents a general approach to an investigation of hierarchical differential two-person games. The proposed formalization of players' actions for these games is the same as in the theory of positional antagonistic differential games [1, 2]. Stackelberg solutions [3, 4] are considered. The problem of finding Stackelberg strategies is reduced to the solution of a nonstandard optimal problem. A structure of Stackelberg strategies is shown. It is established that there exists a pair of Stackelberg strategies of players which is not improvable with respect to the leader Pareto optimal point in the set of equilibrium coalitional strategies introduced in [13, 15]. A sufficient condition for Bellman's principle of optimality for Stackelberg trajectories is obtained. The paper adjoins the investigations [1, 2, 6, 13–16].

1. Introduction

The Stackelberg solution concept, first introduced in [3] and [4] for static games, and then extended and applied to dynamic games in papers [5–11], has recently attracted considerable attention in the literature. The following assumptions are typical for these games [3, 4].

Assumption A. Player 1 (P1), called the leader, announces his strategy ahead of time to player 2.

Assumption B. Player 2 (P2), called the follower, in view of P1's strategy, decides his rational strategy.

It is necessary to define more exactly the concepts "strategy", "rational strategy" in Assumptions A and B. Depending on the chosen definition different problem formulations can be proposed [5–11].

In this paper players' actions are formalized in the class of positional strategies. This formalization is the same as in the general theory of positional antagonistic differential games [1, 2]. The fact of the existence of an universal saddle point in an antagonistic differential game [2] is essentially used. The results in Section 3 are based on [14].

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. Preliminaries

Consider a dynamic system

$$\dot{x}=f(t,x,u,v),\ u\in P,\ v\in Q,\ x(t_0)=x_0,\ t_0\leqq t\leqq \vartheta \tag{1}$$

where $x\in R^n$ is the state variable, $u\in P\subset R^p$ and $v\in Q\subset R^q$ are the control variables handled by P1 and P2, respectively; the sets P and Q are compacts; ϑ is the fixed final time. The function $f:G\times P\times Q\to R^n$ 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) beginning at any position $(t_*,x_*)\in G$ remain in G by all $t\in(t_*,\vartheta]$.

P1 chooses his control u to minimize the cost functional $\sigma_1(x[\vartheta])$ and P2 chooses his control v to minimize the cost functional $\sigma_2(x[\vartheta])$. Here $\sigma_i:R^n\to R^1$ are given continuous functions.

In an antagonistic game interests of players are opposite, therefore any of their joint actions are excepted. Unlike an antagonistic game, in a nonantagonistic game players' decision making should be realized on a contractual basis. In case of system (1) joint actions of players are characterized with the help of a set of trajectories $x(\cdot)=\{x(t),\ t_0\leqq t\leqq\vartheta\}$ which are generated by all possible measurable functions

$$u(\cdot)=\{u(t),\ t_0\leqq t\leqq\vartheta,\ u(t)\in P\}$$

and

$$v(\cdot)=\{v(t),\ t_0\leqq t\leqq\vartheta,\ v(t)\in Q\}$$

from the initial position (t_0,x_0). We denote this set by D. The points $x(\vartheta)$, where $x(\cdot)\in D$, form an attainable set $K(\vartheta)$ at time ϑ for the class of measurable controls. If the set

$$F(t,x)=\{q\in R^n:q=f(t,x,u,v),\ u\in P,\ v\in Q\} \tag{2}$$

$F(t,x)$ is convex, then $K(\vartheta)$ is closed. If $F(t,x)$ is not convex, then $K(\vartheta)$ is, in general, not closed. However, if we extend the class of measurable controls to the class of functions with values in the set of probability measures normed on the product set $P\times Q$, then an attainable set $\bar{K}(\vartheta)$ for this extended class of controls is closed, and also $\bar{K}(\vartheta)$ is the closure of $K(\vartheta)$. For the sake of simplicity we shall assume that $F(t,x)$ is convex in the sequel.

A trajectory $x(\cdot)\in D$ is the best one for P1 if $x(\vartheta)$ is a point of minimum of the function $\sigma_1(\cdot)$ over the set $K(\vartheta)$. Naturally P1 is interested in following this trajectory jointly with P2. Now we analyze how much P2 is interested in following a trajectory $x(\cdot)\in D$ jointly with P1. To this end, consider an antagonistic differential game Γ_2

whose dynamics is described by (1); P2 minimizes the cost $\sigma_2(x[\vartheta])$ and P1 is opposed to him. In the sequel for simplicity we shall assume that the function $f(\cdot,\cdot,\cdot,\cdot)$ (1) satisfies the following condition of saddle point in the minor game ([1], p. 56)

$$\min_{u\in P}\max_{v\in Q} s'f(t,x,u,v)=\max_{v\in Q}\min_{u\in P} s'f(t,x,u,v) \tag{3}$$

for all $s\in R^n$, $(t,x)\in G$. (The general case, when condition (3) is not fulfilled, will be considered in the next paper of the author). It follows from the results of [2] that the game Γ_2 has a continuous value function $\gamma_2(t,x)$, $(t,x)\in G$ and an universal saddle point for the class of pure positional strategies

$$\{u^{(2)}(t,x,\varepsilon),\ v^{(2)}(t,x,\varepsilon)\}. \tag{4}$$

The property of strategies (4) to be universal means that they are optimal not only for the fixed initial position $(t_0,x_0)\in G$ but also for any position $(t_*,x_*)\in G$ assumed as initial one. By using the strategy $v^{(2)}(\cdot,\cdot,\cdot)$ (4) P2 in a current position $(t,x)\in G$ ensures himself the result $\sigma_2(x[\vartheta])$ which is not greater than the value $\gamma_2(t,x)$. On the other hand, by using the strategy $u^{(2)}(\cdot,\cdot,\cdot)$ (4) P1 in a current position $(t,x)\in G$ ensures the result of P2 $\sigma_2(x[\vartheta])$ which is not smaller than the value $\gamma_2(t,x)$.

In the game Γ_2 the class of pure positional strategies of players consists of arbitrary functions $u(t,x,\varepsilon)$ and $v(t,x,\varepsilon)$, $(t,x)\in G$, $\varepsilon>0$, $u(t,x,\varepsilon)\in P$, $v(t,x,\varepsilon)\in Q$, where ε is called a precision parameter [2]. If as pure positional strategies we take arbitrary functions $u(t,x)$ and $v(t,x)$ depending only on the position (t,x) then optimal strategies, in general, are not universal in the game Γ_2 (for a corresponding example, see [12]). The use of the precision parameter permits to ensure the property of universality for optimal strategies in a very large class of antagonistic positional differential games [2].

Now we separate the set D into the union of three disjoint sets D_1, D_2 and D_3. A trajectory $x(\cdot)$ belongs to D_1 if it satisfies

$$\gamma_2(t,x(t))>\gamma_2(\vartheta,x(\vartheta))=\sigma_2(x(\vartheta)),\ t\in[t_0,\vartheta). \tag{5}$$

Inequality (5) means that P2 receives on the trajectory $x(\cdot)$ a result which is better than the value $\gamma_2(t,x(t))$ ensured to him in the game Γ_2. Let us assume that P1 offers to follow a trajectory $x(\cdot)\in D_1$ jointly with P2, and if P2 gives up following $x(\cdot)$, then P1 threats to use the universal strategy $u^{(2)}(\cdot,\cdot,\cdot)$ (4) beginning with a time of the refusal. Obviously, it is more profitable for P2 to accept this offer of P1.

A trajectory $x(\cdot)$ belongs to D_2 if it satisfies

$$\min_{t\in[t_0,\vartheta)}\gamma_2(t,x(t))<\gamma_2(\vartheta,x(\vartheta))=\sigma_2(x(\vartheta)). \tag{6}$$

Let t^* denote the first point of minimum in (6). Now in answer to the offer of P1 to follow a trajectory $x(\cdot)\in D_2$, P2 follows it only till time t^*. Beginning with time t^* it is more profitable for P2 to use the universal strategy $v^{(2)}(\cdot,\cdot,\cdot)$ (4) which ensures him the result $\gamma_2(t^*, x(t^*))$ better than the value $\sigma_2(x(\vartheta))$.

Finally, a trajectory $x(\cdot)$ belongs to D_3 if it satisfies

$$\min_{t\in[t_0,\vartheta)} \gamma_2(t, x(t)) = \gamma_2(\vartheta, x(\vartheta)) = \sigma_2(x(\vartheta)). \tag{7}$$

Let t^{**} denote the first point of minimum in (7). P2 is interested to follow a trajectory $x(\cdot)\in D_3$ only till time t^{**}. Beginning with time t^{**} it is indifferent for P2 whether he continues to follow a trajectory $x(\cdot)$ or he uses the strategy $v^{(2)}(\cdot,\cdot,\cdot)$ (4).

3. Problem formulation. Main result

As it follows from Section 2, it would be convenient to formalize strategies of P1 so that they would contain, firstly, the offer to follow a trajectory $x(\cdot)\in D$ jointly with P2 and, secondly, "the penalty" in case of the refusal of P2 to follow $x(\cdot)$.

We identify a pure strategy (or strategy for short) of P1 (of the leader) with a pair $U=\{u(t, x, \varepsilon), \beta_1(\varepsilon)\}$, where

$$u(t, x, \varepsilon)=\begin{cases} u^*(t, \varepsilon), & \text{if } \|x-x^*(t, \varepsilon)\|\leqq\varepsilon, t_0\leqq t\leqq\vartheta, \varepsilon>0 \\ u^{(2)}(t, x, \varepsilon), & \text{if } \|x-x^*(t, \varepsilon)\|>\varepsilon, t_0\leqq t\leqq\vartheta, \varepsilon>0. \end{cases} \tag{8}$$

Here the functions $u^*(\cdot,\cdot)$ and $x^*(\cdot,\cdot)$ are piecewise continuous and piecewise differentiable with respect to the first argument; $u^*(t,\varepsilon)\in P, t\in[t_0,\vartheta], \varepsilon>0$; the function $u^{(2)}(\cdot,\cdot,\cdot)$ is defined in (4). The function $x^*(\cdot,\cdot)$ is coordinated with the function $u^*(\cdot,\cdot)$ in the following sense: (i) there exists a function $v^*(\cdot,\cdot)$ piecewise continuous with respect to the first argument, $v^*(t,\varepsilon)\in Q$, $t\in[t_0,\vartheta]$, $\varepsilon>0$ such that the controls $u^*(\cdot,\cdot)$ and $v^*(\cdot,\cdot)$ generate the trajectory $x^*(\cdot,\cdot)$; (ii) $\lim_{\varepsilon\to 0} x^*(t,\varepsilon)=x^{**}(t)$, where $x^{**}(\cdot)\in D$. We shall say also that the trajectory $x^{**}(\cdot)\in D$ is coordinated with the function $u(\cdot,\cdot,\cdot)$ (8), and that the function $v^*(\cdot,\cdot)$ is coordinated with $u(\cdot,\cdot,\cdot)$ (8) and with $u^*(\cdot,\cdot)$. (The symbol $\|\cdot\|$ denotes the Euclidean norm).

The function β_1: $(0,\infty)\to(0,\infty)$ is continuous, monotonous and satisfies the condition $\beta_1(\varepsilon)\to 0$ if $\varepsilon\to 0$. A class of such functions is denoted by S. If P1 chooses a value ε_1 of his precision parameter, then a value $\beta_1(\varepsilon_1)$ is the upper bound for the step of subdivisions chosen by P1 in the construction of Euler splines (see below).

For any trajectory $x^{**}(\cdot)\in D$ we can find a function $u(\cdot,\cdot,\cdot)$ (8) with which $x^{**}(\cdot)$ is coordinated. In fact, by the Lusin's theorem for measurable functions $u^{**}(\cdot)$ and $v^{**}(\cdot)$ generating $x^{**}(\cdot)$ we can find functions $u^*(t,\varepsilon)$ and $v^*(t,\varepsilon)$ continuous with respect to t such that for the motion $x^*(t,\varepsilon)$ generated by them the following inequality $\|x^*(t,\varepsilon)-x^{**}(t)\|<\varepsilon$ is true for all $t\in[t_0,\vartheta]$, $\varepsilon>0$. These functions $u^*(\cdot,\cdot)$ and $x^*(\cdot,\cdot)$

are substituted in (8). In particular, if $x^{**}(\cdot)$ is generated by piecewise continuous functions $u^{**}(\cdot)$ and $v^{**}(\cdot)$ then the function $u(\cdot,\cdot,\cdot)$ (8) with which $x^{**}(\cdot)$ is coordinated can be written in a more simple form

$$u(t,x,\varepsilon)=\begin{cases} u^{**}(t), & \text{if} \quad \|x-x^{**}(t)\|\leqq\varepsilon, \ t_0\leqq t\leqq\vartheta, \ \varepsilon>0 \\ u^{(2)}(t,x,\varepsilon), & \text{if} \quad \|x-x^{**}(t)\|>\varepsilon, \ t_0\leqq t\leqq\vartheta, \ \varepsilon>0. \end{cases} \tag{9}$$

We identify a pure strategy (a strategy for short) of P2 (of the follower) with a pair $V=\{v(t,x,\varepsilon),\beta_2(\varepsilon)\}$ where $v(\cdot,\cdot,\cdot)$ is an arbitrary function, $v(t,x,\varepsilon)\in Q$, $(t,x)\in G$, $\varepsilon>0$ and $\beta_2(\cdot)\in S$.

Let be given: strategies U and V, ε_1 and ε_2 which are values of a precision parameter chosen by P1 and P2. Let $\Delta_1=\{\tau_i^{(1)}\}$ and $\Delta_2=\{\tau_j^{(2)}\}$ be subdivisions of the interval $[t_0,\vartheta]$ with finite families of disjoint semi-intervals $[\tau_i^{(1)},\tau_{i+1}^{(1)})$ and $[\tau_j^{(2)},\tau_{j+1}^{(2)})$ chosen by P1 and P2 provided that $\delta(\Delta_S)\leqq\beta_S(\varepsilon_S)$, $s=1,2$ where the step of the subdivision Δ_S is denoted by $\delta(\Delta_S)=\max\limits_i(\tau_{i+1}^{(S)}-\tau_i^{(S)})$. An Euler spline generated by the strategies U and V, by fixed ε_1 and ε_2, by the subdivisions Δ_1 and Δ_2 from the initial position (t_0,x_0) is a piecewise differentiable function

$$x_\Delta^\varepsilon[t]=x_{\Delta_1,\Delta_2}^{\varepsilon_1,\varepsilon_2}[t,t_0,x_0,U,V]$$

satisfying the multistage differential equation

$$\begin{aligned} \dot{x}_\Delta^\varepsilon[t]&=f(t,x_\Delta^\varepsilon[t],u_{\Delta_1}^{\varepsilon_1}[t],v_{\Delta_2}^{\varepsilon_2}[t]),\\ u_{\Delta_1}^{\varepsilon_1}[t]&=u(\tau_i^{(1)},x_\Delta^\varepsilon[\tau_i^{(1)}],\varepsilon_1), \quad \tau_i^{(1)}\leqq t<\tau_{i+1}^{(1)},\\ v_{\Delta_2}^{\varepsilon_2}[t]&=v(\tau_j^{(2)},x_\Delta^\varepsilon[\tau_j^{(2)}],\varepsilon_2), \quad \tau_j^{(2)}\leqq t<\tau_{j+1}^{(2)}. \end{aligned}$$

A limiting motion (a motion for short) generated by the strategies U and V from the initial position (t_0,x_0) is a continuous function $x[t]=x[t,t_0,x_0,U,V]$ for which there exists a sequence of Euler splines $x_{\Delta_1^k,\Delta_2^k}^{\varepsilon_1^k,\varepsilon_2^k}[t,t_0^k,x_0^k,U,V]$ uniformly converging to $x[t]$ on $[t_0,\vartheta]$ if

$$k\to\infty, \quad \varepsilon_1^k\to 0, \quad \varepsilon_2^k\to 0, \quad t_0^k\to t_0, x_0^k\to x_0, \quad \delta(\Delta_S^k)\leqq\beta_S(\varepsilon_S^k), \qquad s=1,2.$$

Without loss of generality we assume that $\varepsilon_2^k\leqq\varepsilon_1^k$. In fact, the possibilities of P2 are not restricted by this since he can change a "scale" of his precision parameter in a reasonable way.

In general, strategies U and V generate a set of motions $x[t,t_0,x_0,U,V]$. We shall denote this compact subset of $C[t_0,\vartheta]$ by $X(t_0,x_0,U,V)$. However, $X(t_0,x_0,U,V)$ consists of single motion, if the strategy $V=\{v(t,x,\varepsilon),\beta_2(\varepsilon)\}$ has the following first component

$$v(t,x,\varepsilon)=\begin{cases} v^*(t,\varepsilon), & \text{if} \quad \|x-x^*(t,\varepsilon)\|\leqq\varepsilon, t_0\leqq t\leqq\vartheta, \varepsilon>0 \\ \text{arbitrary}, & \text{if} \quad \|x-x^*(t,\varepsilon)\|>\varepsilon, t_0\leqq t\leqq\vartheta, \varepsilon>0 \end{cases} \tag{10}$$

where the function $v^*(\cdot,\cdot)$ is coordinated with $u^*(\cdot,\cdot)$ and $x^*(\cdot,\cdot)$; the function $\beta_2(\cdot)$ is chosen jointly with the function $\beta_1(\cdot)$ so that the following inequality

$$\|x^{\varepsilon,\varepsilon_2}_{\Delta_1,\Delta_2}[t, t_0, x_0, U, V]-x^*(t,\varepsilon)\|\leqq\varepsilon, \quad t\in[t_0,\vartheta] \tag{11}$$

holds for any $\varepsilon>0$, $\varepsilon_2\leqq\varepsilon$, $\Delta_1,\Delta_2:\delta(\Delta_1)\leqq\beta_1(\varepsilon)$, $\delta(\Delta_2)\leqq\beta_2(\varepsilon_2)$. Then we have $x[t, t_0, x_0, U, V]=x^{**}(t)$, where $x^{**}(\cdot)$ is coordinated with $u(\cdot,\cdot,\cdot)$ (8).

If a trajectory $x^{**}(\cdot)\in D$ is generated by piecewise continuous functions $u^{**}(\cdot)$ and $v^{**}(\cdot)$, then the function $v(\cdot,\cdot,\cdot)$ (10) can be written in a simpler form

$$v(t,x,\varepsilon)=\begin{cases}v^{**}(t), & \text{if } \|x-x^{**}(t)\|\leqq\varepsilon,\ t_0\leqq t\leqq\vartheta,\ \varepsilon>0\\ \text{arbitrary}, & \text{if } \|x-x^{**}(t)\|>\varepsilon,\ t_0\leqq t\leqq\vartheta,\ \varepsilon>0.\end{cases} \tag{12}$$

Now we shall define more exactly Assumptions A and B, introduced in Section 1.

Assumption A_1. P1 announces his strategy $U=\{u(t,x,\varepsilon),\beta_1(\varepsilon)\}$ to P2 prior to the start of the game. Here $u(\cdot,\cdot,\cdot)$ is of the form (8) and $\beta_1(\cdot)\in S$.

Assumption B_1. When the strategy U is known to P2 his action is assumed to be rational. This means the following. If a trajectory $x^{**}(\cdot)$ coordinated with $u(\cdot,\cdot,\cdot)$ (8) belongs to D_1, then a rational strategy $V$is chosen to follow $x^{**}(\cdot)$. (In particular, a strategy $V$the first component of which is $v(\cdot,\cdot,\cdot)$ (10) (or (12)) is rational.) If $x^{**}(\cdot)$ belongs to D_2, then a rational strategy $V$is chosen to follow $x^{**}(\cdot)$ till time t^* (6). Beginning with time t^* (or with any next points of minimum in (6)) the first component of $V$is the function $v^{(2)}(\cdot,\cdot,\cdot)$ (4). Finally, if $x^{**}(\cdot)$ belongs to D_3, then a rational strategy $V$is chosen to follow $x^{**}(\cdot)$ till time t^{**} (7). Beginning with time t^{**} (or with any next points of minimum in (7)) it is either a strategy following $x^{**}(\cdot)$ till end or a strategy whose first component is the function $v^{(2)}(\cdot,\cdot,\cdot)$ (4).

A set of rational strategies $V$of P2 corresponding to the announced strategy U of P1 will be denoted by $K(t_0, x_0, U)$.

We note that since P1 tends to follow a trajectory $x^{**}(\cdot)\in D$ jointly with P2, he must choose a function $\beta_1(\cdot)\in S$ so that P2 knowing this choice would have a chance to choose a function $\beta_2(\cdot)$ such that inequality (11) will hold. It is natural to assume that P1 realizes such a choice. To guarantee the fulfilment of the inequality $\varepsilon_2\leqq\varepsilon_1$ for the construction of Euler splines we make one more assumption.

Assumption C_1. P1 chooses a value ε_1 of his precision parameter and informs P2 about it simultaneously with the beginning of the game.

Since P2 gets ε_1 only simultaneously with beginning of the game, he cannot use this information to make his rational strategy more precise.

Any strategy belonging to the set $K(t_0, x_0, U)$ ensures P2 one and the same result. However, the result of P1 depends, in general, on what strategy belonging to $K(t_0, x_0, U)$ is chosen by P2. Therefore we distinguish two cases.

Case a). P2 chooses an arbitrary rational strategy from $K(t_0, x_0, U)$.

Obviously, if P1 announces a strategy U, his guaranteed result in Case a) is

$$\rho_a(U) = \sup_{V \in K(t_0, x_0, U)} \max_{x[\cdot] \in X(t_0, x_0, U, V)} \sigma_1(x[\vartheta]). \tag{13}$$

Case b). P2 shows the goodwill to P2 choosing a rational strategy V from the following condition

$$\rho_b(U) = \min_{V^* \in K(t_0, x_0, U)} \max_{x[\cdot] \in X(t_0, x_0, U, V^*)} \sigma_1(x[\vartheta]) = \\ = \max_{x[\cdot] \in X(t_0, x_0, U, V)} \sigma_1(x[\vartheta]). \tag{14}$$

The value $\rho_b(U)$ (14) is a guaranteed result of P1 in Case b).

In the sequel the considered game will be called a hierarchical differential game (HDG) in Case a) and a hierarchical differential game with the well-wishing follower (HDGWF) in Case b). Sometimes it will be convenient not to distinguish Case a) and Case b), then the considered game will be denoted by Γ.

Problem 1. Find the strategy of P1 $U^0 = \{u^0(t, x, \varepsilon), \beta_1^0(\varepsilon)\}$ such that

$$\rho_a(U^0) = \min_U \rho_a(U). \tag{15}$$

Problem 2. Find the strategy of P1 $U_0 = \{u_0(t, x, \varepsilon), \beta_{10}(\varepsilon)\}$ such that

$$\rho_b(U_0) = \min_U \rho_b(U). \tag{16}$$

Definition 1. A solution U^0 of Problem 1 is called a Stackelberg strategy of P1 in a HDG. A strategy from the set $K(t_0, x_0, U^0)$ is called a Stackelberg strategy of P2 in a HDG.

Definition 2. A solution U_0 of Problem 2 is called a Stackelberg strategy of P1 in a HDGWF. A strategy from the set $K(t_0, x_0, U_0)$ satisfying condition (14) is called a Stackelberg strategy of P2 in a HDGWF.

Definition 3. Any trajectory belonging to the set of motions generated by Stackelberg strategies of P1 and P2 is called a Stackelberg trajectory in the game Γ.

Now we formulate the following auxiliary nonstandard optimal problem.

Problem 3. Let the dynamics of the control system be described by (1). Find the measurable functions $u(t)$ and $v(t)$, $t_0 \leqq t \leqq \vartheta$ minimizing the cost $\sigma_1(x(\vartheta))$ provided that

$$\sigma_2(x(\vartheta)) = \gamma_2(\vartheta, x(\vartheta)) \leqq \gamma_2(t, x(t)), \quad t \in [t_0, \vartheta] \tag{17}$$

where $\gamma_2(\cdot, \cdot)$ is the value function of the game Γ_2 and $x(t)$, $t_0 \leqq t \leqq \vartheta$ is a trajectory of (1) generated by controls $u(\cdot)$ and $v(\cdot)$ from the initial position (t_0, x_0).

The set of admissible trajectories in Problem 3 is not empty and, since the set $F(t, x)$ (2) is convex, is compact in $C[t_0, \vartheta]$. Hence, Problem 3 has a solution.

Let measurable functions $u^{00}(\cdot)$ and $v^{00}(\cdot)$ be a solution of Problem 3 and $x^{00}(\cdot)$ be a corresponding trajectory. It follows from (17) that $x^{00}(\cdot)$ belongs either to D_1 or to D_3. As it was established above, for $x^{00}(\cdot)$ one can find a function $u^0(\cdot,\cdot,\cdot)$ of type (8) with which $x^{00}(\cdot)$ is coordinated, i.e.

$$u^0(t, x, \varepsilon) = \begin{cases} u^0(t, \varepsilon), & \text{if} \quad \|x - x^0(t, \varepsilon)\| \leqq \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0 \\ u^{(2)}(t, x, \varepsilon), & \text{if} \quad \|x - x^0(t, \varepsilon)\| > \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0. \end{cases} \tag{18}$$

Consider the strategies $U^0 = \{u^0(t, x, \varepsilon), \beta^0(\varepsilon)\}$ and $V^0 = \{v^0(t, x, \varepsilon), \beta^0(\varepsilon)\}$ where $u^0(\cdot,\cdot,\cdot)$ is defined in (18) and $v^0(\cdot,\cdot,\cdot)$ is of type (10), i.e.

$$v^0(t, x, \varepsilon) = \begin{cases} v^0(t, \varepsilon), & \text{if} \quad \|x - x^0(t, \varepsilon)\| \leqq \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0 \\ \text{arbitrary}, & \text{if} \quad \|x - x^0(t, \varepsilon)\| > \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0. \end{cases} \tag{19}$$

The function $\beta^0(\cdot) \in S$, being the same for both strategies is chosen so as to ensure the fulfilment of an inequality of type (11), i.e.

$$\|x^{\varepsilon, \varepsilon_2}_{\Delta_1, \Delta_2}[t, t_0, x_0, U^0, V^0] - x^0(t, \varepsilon)\| \leqq \varepsilon, \quad t \in [t_0, \vartheta] \tag{20}$$

provided that $\varepsilon_2 \leqq \varepsilon$, $\delta(\Delta_1) \leqq \beta^0(\varepsilon)$, $\delta(\Delta_2) \leqq \beta^0(\varepsilon_2)$.

We recall that if the functions $u^{00}(\cdot)$ and $v^{00}(\cdot)$ are piecewise continuous, then the functions $u^0(\cdot,\cdot,\cdot)$ (18) and $v^0(\cdot,\cdot,\cdot)$ (19) can be written in more simple form (see (9) and (12)).

Obviously, if P1 and P2 choose the strategies U^0 and V^0 (18)–(20) then the set $X(t_0, x_0, U^0, V^0)$ consists of single motion $x^{00}(\cdot)$.

The main result is as follows (see [14]).

Theorem 1. Let $x^{00}(\cdot)$ belong to D_1. Then the strategies U^0 and V^0 (18)–(20) are Stackelberg strategies of P1 and P2 in HDG.

Theorem 2. Let $x^{00}(\cdot)$ belong to D_3. Then the strategies U^0 and V^0 (18)–(20) are Stackelberg strategies of P1 and P2 in HDGWF.

Hence we can draw the following conclusion. If controls $u^{00}(\cdot)$ and $v^{00}(\cdot)$ are a solution of Problem 3, then the trajectory $x^{00}(\cdot)$ generated by them is a Stackelberg trajectory in the game Γ. If $x^{00}(\cdot)$ belongs to D_1, then the strategy U^0 (18), (20) is a solution of Problem 1 and of Problem 2 all the more. If $x^{00}(\cdot)$ belongs to D_3 then the strategy U^0 is a solution of Problem 2 while Problem 1 has not a solution. However, we can construct a minimizing sequence of strategies in Problem 1 on the basis of the strategy U^0.

We summarize this section by noting some other advantages of the proposed formalization of strategies. Firstly, following a trajectory $x^{00}(\cdot)$ can be effected in physically realized Euler splines with any preassigned precision. P1 assigns a precision by means of choice of a value ε_1 of his precision parameter. Secondly, a function

$\beta^0(\cdot) \in S$ can be chosen so that inequality (20) is held with some "reserve". This reserve can be made sufficiently large so as inequality (20) remains time even if P1 and P2 have made some unintentional errors choosing their controls. That is, following a trajectory $x^{00}(\cdot)$ is stable with respect to unintentional errors of players.

4. A Stackelberg strategy pair as the unimprovable for the leader Pareto optimal point of the set of equilibrium coalitional strategies

The concept of equilibrium coalitional strategy (ECS) introduced in [13, 15] generalizes the concept of Nash strategy set. In the beginning of this section we find ECSs in our special case when the dynamics is described by (1), condition (3) is fulfilled and the number of players equals 2.

In parallel with the game Γ_2 introduced in Section 2 we consider an antagonistic differential game Γ_1 whose dynamics is described by (1), P1 minimizes the cost $\sigma_1(x[\vartheta])$ and P2 is opposite to him. The game Γ_1 has a continuous value function $\gamma_1(t, x)$, $(t, x) \in G$ and an universal saddle point

$$\{u^{(1)}(t, x, \varepsilon), \quad v^{(1)}(t, x, \varepsilon)\}. \tag{21}$$

We select now a subset $D_* \subset D$ (the set D is defined in Section 2) including in D_* those trajectories $x(\cdot) \in D$ for which the following inequalities are held

$$\begin{aligned} &\sigma_1(x(\vartheta)) = \gamma_1(\vartheta, x(\vartheta)) \leqq \gamma_1(t, x(t)), \\ &\sigma_2(x(\vartheta)) = \gamma_2(\vartheta, x(\vartheta)) \leqq \gamma_2(t, x(t)), \quad t \in [t_0, \vartheta]. \end{aligned} \tag{22}$$

The set D_* is not empty. It follows from (22) that P1 and P2 obtain on a trajectory $x^*(\cdot) \in D_*$ those results which are not worse than their guaranteed results in the games Γ_1 and Γ_2, respectively. Therefore in the presence of the partner's threat to use the universal strategy $v^{(1)}(\cdot, \cdot, \cdot)$ (21) or $u^{(2)}(\cdot, \cdot, \cdot)$ (4) in case of the refusal to follow a trajectory $x^*(\cdot) \in D_*$, P1 and P2 have to agree to follow this trajectory.

Let controls $u^*(\cdot)$ and $v^*(\cdot)$ generate a trajectory $x^*(\cdot) \in D_*$. For simplicity, we assume that they are piecewise continuous. It follows from [13, 15] that the strategies with first components

$$u^*(t, x, \varepsilon) = \begin{cases} u^*(t), & \text{if} \quad \|x - x^*(t)\| \leqq \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0 \\ u^{(2)}(t, x, \varepsilon), & \text{if} \quad \|x - x^*(t)\| > \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0 \end{cases}$$

and

$$v^*(t, x, \varepsilon) = \begin{cases} v^*(t), & \text{if} \quad \|x - x^*(t)\| \leqq \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0 \\ v^{(1)}(t, x, \varepsilon), & \text{if} \quad \|x - x^*(t)\| > \varepsilon,\ t_0 \leqq t \leqq \vartheta,\ \varepsilon > 0 \end{cases}$$

form an ECS. For details, concerning with construction of Euler splines generated by ECS, the choice of subdivisions of $[t_0, \vartheta]$ etc. see [13, 15].

We discuss now an interrelation between Stackelberg strategies and ECSs. Obviously, if U and V are Stackelberg strategies of P1 and P2 in HDG or HDGWF, then a pair (U, V) forms an ECS.

Returning now to Problem 3 we find its solution for which $\sigma_2(x(\vartheta))$ is minimal. Such a solution denoted by $u^P(\cdot)$ and $v^P(\cdot)$ exists under our assumptions. Denote by U^P and V^P strategies of P1 and P2 which are constructed from U^0 and V^0 (18)–(20) with the substitution of $u^P(\cdot)$, $v^P(\cdot)$ and $x^P(\cdot)$ for $u^{00}(\cdot)$, $v^{00}(\cdot)$ and $x^{00}(\cdot)$ respectively; here $x^P(\cdot)$ is a trajectory generated by $u^P(\cdot)$ and $v^P(\cdot)$.

Theorem 3. A pair (U^P, V^P) is the unimprovable Pareto optimal point for P1 in the set of ECSs.

The proof of Theorem 3 is obvious.

5. Bellman's principle of optimality for Stackelberg trajectories

As it is shown in Section 3, a Stackelberg trajectory in the game Γ is the one generated by controls $u^{00}(\cdot)$ and $v^{00}(\cdot)$ which solve Problem 3. It is a well-known fact (see, e.g. [5, 7–10]), that a Stackelberg trajectory, in general, does not satisfy Bellman's principle of optimality. In other words, if a trajectory $x^{00}(t)$, $t_0 \leqq t \leqq \vartheta$ is a Stackelberg one, then its part $x^{00}(t)$, $\tilde{t} \leqq t \leqq \vartheta$, where $\tilde{t} \in (t_0, \vartheta)$, is not, in general, Stackelberg trajectory in the game Γ for initial position $(\tilde{t}, x^{00}(\tilde{t}))$.

We say that a trajectory $x(\cdot) \in D$ satisfies the condition E, if a value function $\gamma_2(\cdot,\cdot)$ in the game Γ_2 is not increasing along $x(\cdot)$, i.e.

$$\gamma_2(t'', x(t'')) \leqq \gamma_2(t', x(t')), \quad t', t'' \in [t_0, \vartheta], \quad t' < t''. \tag{23}$$

Theorem 4. Let a Stackelberg trajectory in the game Γ satisfy condition E. Then it satisfies Bellman's principle of optimality.

Proof. Let controls $u^{00}(\cdot)$ and $v^{00}(\cdot)$ generate a Stackelberg trajectory $x^{00}(\cdot)$. We fix $\tilde{t} \in (t_0, \vartheta)$ and formulate Problems 1 and 2 for the initial position $(\tilde{t}, x^{00}(\tilde{t}))$. By Theorems 1 and 2, a solution of these problems is reduced to solving of Problem 3 formulated for the initial position $(\tilde{t}, x^{00}(\tilde{t}))$. Now we show that a restriction of functions $u^{00}(\cdot)$ and $v^{00}(\cdot)$ on the interval $[\tilde{t}, \vartheta]$ is a solution of Problem 3. Suppose, to the contrary, that there exist controls $\tilde{u}(t)$ and $\tilde{v}(t)$, $\tilde{t} \leqq t \leqq \vartheta$ generating a trajectory $\tilde{x}(t)$, $\tilde{t} \leqq t \leqq \vartheta$ from the initial position $(\tilde{t}, x^{00}(\tilde{t}))$ such that the following inequalities hold

$$\sigma_1(\tilde{x}(\vartheta)) < \sigma_1(x^{00}(\vartheta)), \tag{24}$$

$$\gamma_2(\vartheta, \tilde{x}(\vartheta)) \leqq \gamma_2(t, \tilde{x}(t)), \quad t \in [\tilde{t}, \vartheta]. \tag{25}$$

Consider controls $u^+(t)$ and $v^+(t)$, $t_0 \leqq t \leqq \vartheta$ which coincide with $u^{00}(\cdot)$ and $v^{00}(\cdot)$ on $[t_0, \tilde{t})$ and with $\tilde{u}(\cdot)$ and $\tilde{v}(\cdot)$ on $[\tilde{t}, \vartheta]$. From condition E holding for the trajectory $x^{00}(\cdot)$ and from (25) it follows that controls $u^+(\cdot)$ and $v^+(\cdot)$ are admissible in Problem 3 formulated for the interval $[t_0, \vartheta]$. Further, it follows from (24) that on the trajectory $x^+(\cdot)$ generated by $u^+(\cdot)$ and $v^+(\cdot)$ P1 receives a result that is better than that obtained on $x^{00}(\cdot)$. This contradicts to the Stackelberg optimality of $x^{00}(\cdot)$.

Theorem 4 gives a sufficient test for the fulfilment of Bellman's principle of optimality for Stackelberg trajectories. This test is not necessary, in general (see Example 2 in Section 6).

6. Examples

Example 1. Consider a system described by

$$\dot{x} = u + v, \quad x = (x_1, x_2), \quad u = (u_1, u_2), \quad v = (v_1, v_2),$$
$$\|u\| \leqq 1, \quad \|v\| \leqq 1.$$

The cost functionals of the players are

$$\sigma_1(x[\vartheta]) = \|x[\vartheta] - a^{(1)}\|, \quad a^{(1)} \in R^2$$
$$\sigma_2(x[\vartheta]) = x_2[\vartheta] - |x_1[\vartheta]|.$$

Obviously, a value function in the game Γ_2 is

$$\gamma_2(t, x) \equiv \gamma_2(x) = x_2 - |x_1|.$$

Let us assume that

$$t_0 = 0, \quad \vartheta = 1, \quad x(0) = (-1, 1), \quad a^{(1)} = (2, 2).$$

Problem 3 is formulated as follows. Find controls

$$u(t), \quad v(t), \quad 0 \leqq t \leqq 1, \quad \|u(t)\| \leqq 1, \quad \|v(t)\| \leqq 1$$

minimizing the cost $\|x(1) - a^{(1)}\|$ provided that

$$x_2(1) - |x_1(1)| \leqq x_2(t) - |x_1(t)|$$

for all $t \in [0, 1]$. It can be verified that the functions

$$u_1^{00}(t) = v_1^{00}(t) = 1, \quad u_2^{00}(t) = v_2^{00}(t) = 0, \quad 0 \leqq t \leqq 1$$

give the unique solution of Problem 3. The Stackelberg trajectory

$$x_1^{00}(t) = -1 + 2t, \quad x_2^{00}(t) = 1, \quad 0 \leqq t \leqq 1$$

is represented on the plane (x_1, x_2) in Fig. 1. It starts in the point $A(-1, 1)$ at time $t=0$ and ends in the point $B(1, 1)$ at time $t=1$. The rays OC and OD form a line $\gamma_2(x)=0$. The initial point $x(0)$ and the target point $a^{(1)}$ of P1 lie on this line. The function $\gamma_2(t, x)$ calculated along the trajectory $x^{00}(\cdot)$ at first increases from 0 in A to 1 in the point K and then decreases to 0 in B. Hence, condition E is not fulfilled for $x^{00}(\cdot)$. It can be easily shown that $x^{00}(\cdot)$ does not satisfy Bellman's principle of optimality.

Since $x^{00}(\cdot)\in D_3$ we have that the strategies U^0 and V^0 (18)–(20) constructed on the basis of the functions $u^{00}(\cdot)$ and $v^{00}(\cdot)$ are Stackelberg strategies in HDGWF.

In Fig. 1 Stackelberg trajectories are represented by dashed lines if we take $\vartheta=1.1$ (the segment AL) and $\vartheta=0.9$ (the segment AM). The corresponding strategies U^0 and V^0 are Stackelberg ones in HDGWF also.

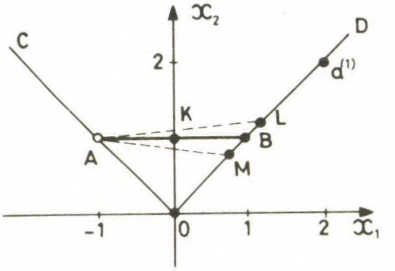


Fig. 1

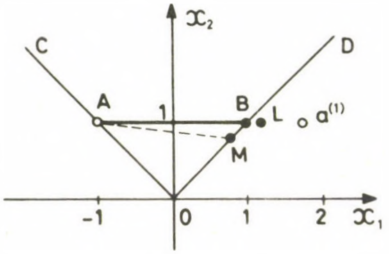


Fig. 2

Example 2 differs from the previous one only in the choice of the target point $a^{(1)}=(\sqrt{3}, 1)$ (see Fig. 2). The solution of Problem 3 and the Stackelberg trajectory are the same as in Example 1. However, here the solution of Problem 3 is also a solution of the problem of finding of unconditional minimum of the cost $\sigma_1(x(\vartheta))$ (i.e. team problem for P1). And although the trajectory $x^{00}(\cdot)$ does not satisfy the condition E as before, it satisfies Bellman's principle of optimality. As in Example 1, we have $x^{00}(\cdot)\in D_3$ and, hence, the strategies U^0 and V^0 (18)–(20) are Stackelberg ones in HDGWF.

If we take $\vartheta=1.1$, then the Stackelberg trajectory is the segment AL and, since $x^{00}(\cdot)\in D_1$, the strategies U^0 and V^0 are Stackelberg ones in HDG. The Stackelberg trajectory does not satisfy condition E but it satisfies Bellman's principle of optimality.

If we take $\vartheta=0.9$, then the Stackelberg trajectory is the segment AM. It does not satisfy Bellman's principle of optimality.

Example 3. The only difference from Example 1 here is that $\sigma_2(x[\vartheta])=\|x[\vartheta]\|$. Then a value function in the game Γ_2 is

$$\gamma_2(t, x)\equiv\gamma_2(x)=\|x\|.$$

It can be verified that the functions

$$u_1^{00}(t)=v_1^{00}(t)=\cos\left(\frac{\pi}{4}-\sqrt{2}t\right),\quad u_2^{00}(t)=v_2^{00}(t)=\sin\left(\frac{\pi}{4}-\sqrt{2}t\right),\ 0\leqq t\leqq 1$$

give the unique solution of Problem 3. The Stackelberg trajectory on the plane (x_1, x_2) is the arc AB of the circle $\|x\|=\sqrt{2}$ (see Fig. 3). The function $\gamma_2(t, x)$ calculated along the trajectory $x^{00}(\cdot)$ is constant and, hence, $x^{00}(\cdot)$ satisfies condition E. By Theorem 4, $x^{00}(\cdot)$ satisfies Bellman's principle of optimality.

In conclusion we note that a more complex example of a Stackelberg trajectory satisfying Bellman's principle of optimality has been considered in [16].

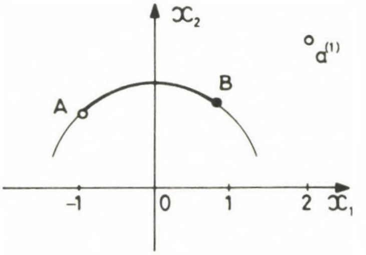


Fig. 3

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. London, Hodge, 1952.
4. *Germeier, Iu. B.*, On two-person games with fixed sequence of moves. Dokl. Akad. 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. *Kononenko, A. F.*, On multi-step conflict with information exchange. Ž. Vyčisl. Mat. i Mat. Fiz., **17,** *4* (1977).
7. *Cruz, J. B., Jr.*, Leader-follower strategies for multilevel systems. IEEE Trans. Automat. Contr., **AC-23,** *2* (1978).
8. *Başar, T., Selbuz, H.*, Closed-loop Stackelberg strategies with applications in the optimal control of multilevel systems. IEEE Trans. Automat. Contr., **AC-24,** *2* (1979).
9. *Başar, T.*, Equilibrium strategies in dynamic games with multilevels of hierarchy. Automatica, **17,** *5* (1981).
10. *Tolwinski, B.*, A Stackelberg solution of dynamic games. IEEE Trans. Automat. Contr., **AC-28,** *1* (1983).
11. *Ishida, T., Shimemura, E.*, Three-levels incentive strategies in differential games. Int. J. Control, **38,** *6* (1983).

12. *Subbotina, N. N.*, Universal optimal strategies in positional differential games. Differ. uravnenija, **19,** *11* (1983).
13. *Kleimenov, A. F.*, Equilibrium coalitional counterstrategies in many-person differential games. Priklad. mat. i meh., **46,** *5* (1982).
14. *Kleimenov, A. F.*, Analysis of a hierarchical differential two-person game. Priklad. mat. i meh., **48,** *4* (1984).
15. *Kleimenov, A. F.*, Equilibrium coalitional mixed strategies in differential games of *m* players. Problems of Control and Information Theory, **11,** *2* (1982).
16. *Kleimenov, A. F.*, Optimal strategies in a hierarchical differential game. Problems of Control and Information Theory, **12,** *6* (1983).

Стратегии Штакельберга в иерархических дифференциальных играх двух лиц

А. Ф. КЛЕЙМЕНОВ

(Свердловск)

В работе дается общий подход к исследованию иерархической дифференциальной игры двух лиц. Динамика игры описывается нелинейным дифференциальным уравнением достаточно общего вида. Показатели игроков суть функции от конечного состояния системы.

Предлагаемая формализация стратегий игроков основывается на формализации в общей теории позиционных антагонистических дифференциальных игр [1, 2]. Существенно используется факт существования универсальных седловых точек в таких играх [2]. Рассматриваются решения Штакельберга [3, 4]. Основной результат состоит в том, что задача нахождения стратегий Штакельберга сводится к решению нестандартной задачи оптимального управления [14]. Выявлена структура стратегий Штакельберга.

А. Ф. Клейменов
Институт математики и механики УНЦ АН СССР
СССР 620066, ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 15 (6), pp. 413—423 (1986)

ANISOCHRONIC MODELLING AND STABILITY CRITERION OF HEREDITARY SYSTEMS

P. ZÍTEK

(*Prague*)

(Received January 5, 1986)

Delays, aftereffects and other hereditary properties in general represent an important class of problems in control systems investigation. Since the classical system state theory is based on the assumption that the system to be modelled is governed by a set of ordinary differential equations, it cannot ensure a faithful means of investigation in case of hereditary phenomena modelling. The behaviour of systems characterized by any type of aftereffects is to be described by differential equations with concentrated or distributed delays in argument.

The article presents an original approach to composing hereditary system models based on the concept of the anisochronic system state [6]. A method is suggested enabling us to express hereditary interactions in the system as a combination of concentrated delays and simultaneous integrations. Further a characteristic complex function of this type of system is introduced and a stability criterion based on it is derived.

1. Anisochronic model formulation

For describing dynamical properties of controlled processes the state space models are widely used in control theory. The state equation of these models

$$\frac{d\mathbf{x}(t)}{dt}=\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{u}(t) \tag{1}$$

(state vector **x**, input vector **u**) represents an interpretation of the investigated process, which can be characterized as a network of n interacting capacities (actual or fictive). The state variables $x_1, x_2, \dots x_n$ represent the appropriate levels of accumulation in these capacities and each of the coefficients A_{ij} characterizes the rate of influence between the level x_j and the rate of x_i change in time. The internal mutual interactions between state variables are assumed to be determined by the coefficients A_{ij} only,

$$q_{ij}(t)=A_{ij}x_j(t), \tag{2}$$

i.e. without any inner dynamics of them. This assumption is well satisfied in case of a process with concentrated parameters. However, if the interactions between capacities are delayed or if they have any other inner dynamics the description by equation (1) is

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

hardly to be applied. The inner dynamics of such a kind of interaction may always be characterized as a continuously distributed delay. Therefore it can be expressed by Stieltjes integral

$$q_{ij}(t) = A_{ij} \int_0^\infty x_j(t-\tau)dh_{ij}(\tau) \tag{3}$$

where τ denotes the time shift variable and h_{ij}, satisfying the condition

$$\lim_{\tau \to \infty} h_{ij}(\tau) = 1, \tag{4}$$

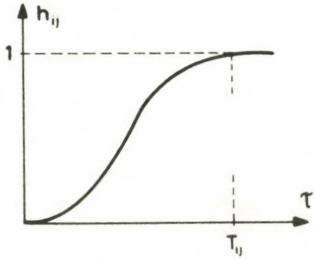


Fig. 1

is the delay distribution function. The aim of condition (4) is to express by h_{ij} the delays only and no static sensitivity relations.

The infinite interval of integration in (3) is an important disadvantage regarding the practical use of this formula. In most practical cases, however, the delays are limited on a finite interval of τ only, so that it can be found such a T_{ij} that for all $\tau > T_{ij}$ it is satisfied that $h_{ij}(\tau) \equiv 1$ (Fig. 1). Then it is possible to substitute the infinite interval in (3) by $\langle 0, T_{ij} \rangle$, i.e.

$$q_{ij}(t) = A_{ij} \int_0^{T_{ij}} x_j(t-\tau)dh_{ij}(\tau). \tag{5}$$

If each of such T_{ij} are determined and if

$$T = \max_{i,j} T_{ij} \tag{6}$$

then all the margins T_{ij} may be replaced by a common value T. The state vector change in time can then be described by the following vector functional differential equation

$$\frac{d\mathbf{x}(t)}{dt} = \int_0^T d\mathbf{A}(\tau)\mathbf{x}(t-\tau) + \mathbf{Bu}(t) \tag{7}$$

where

$$\mathbf{A}(\tau)=\begin{bmatrix} A_{11}h_{11}(\tau),\ldots,A_{1n}h_{1n}(\tau) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ A_{n1}h_{n1}(\tau),\ldots,A_{nn}h_{nn}(\tau) \end{bmatrix}, \tag{8}$$

u is the system input vector and **B** the matrix of input coefficients.

The properties of equation (7) solutions were investigated by many authors [1], [2], [3]. It was shown in [5], [6] that equation (7) may be interpreted as a state space model of process dynamics if the instantaneous system state at t is defined as a segment of vector $\mathbf{x}$ over the time interval $\langle t-T, t\rangle$, i.e.

$$ST(t)=\hat{\mathbf{x}}\langle t-T, t\rangle. \tag{9}$$

In this alternative interpretation $ST(t)$ is introduced as the anisochronic system state with all the basic properties required in state space theory in general (consistency, separability etc.).

The functional character of the state equation (7) allows us to describe a process dynamics by means of a considerably lower number of state variables x_i in comparison with equation (1). However, the main advantage of formulation (8) is its ability to express by the functions h_{ij} even such dynamical properties which are impossible to be described by ordinary differential equations at all. It is due to the fact that in effect the functions h_{ij} in (7), (8) specify the delay distribution of state variables x_i in the system feedback. An example of the function h_{ij} defining a dynamical phenomenon, inexpressible by the framework of equation (1), is mentioned in Fig. 2. This example also emphasizes that the functions h_{ij} need not necessarily be continuous everywhere. The jump at $\tau=T_d$ signifies that not only a continuously distributed delay $\tau\in\langle T_1, T_2\rangle$ but also a concentrated transport delay $\tau=T_d$ is present.

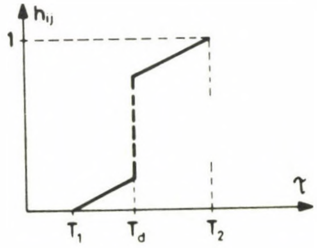


Fig. 2

2. Transforming continuously distributed delays to concentrated ones

The functional form of equation (7) is not the most convenient one for computer models realization. Since the concentrated delays are considerably easier to be simulated on computer, the following transformation was developed to replace continuously distributed delays by an equivalent combination of concentrated ones and simultaneous integrations.

Let the distribution function h_{ij} be given in the form

$$h(\tau)= \begin{cases} h_1 , & \tau \leqq T_1 \\ (\tau - T_1)(h_2 - h_1)/(T_2 - T_1), & \tau \in \langle T_1, T_2 \rangle \\ h_2 , & \tau \geqq T_2 \end{cases} \tag{10}$$

where h_1, h_2 are given and $T_2 > T_1$. Within the interval $\langle T_1, T_2 \rangle$ the derivative of h is constant while beyond it, it is zero. With respect to this fact the integral in (5) may be considered in the form

$$z(t)= \int_{T_1}^{T_2} x(t-\tau)dh(\tau)= \frac{h_2-h_1}{T_2-T_1} \int_{T_1}^{T_2} x(t-\tau)d\tau. \tag{11}$$

By the help of simultaneous integration of x from zero to t it is possible to express $z(t)$ by the formula

$$z(t)= \frac{h_2-h_1}{T_2-T_1} \int_{0}^{t} [x(\vartheta - T_1) - x(\vartheta - T_2)]d\vartheta + z(0) \tag{12}$$

where

$$z(0)= \frac{h_2-h_1}{T_2-T_1} \int_{T_1}^{T_2} x(-\tau)d\tau. \tag{13}$$

This result represents an additional state differential equation

$$\frac{dz(t)}{dt} = \frac{h_2-h_1}{T_2-T_1} [x(t-T_1) - x(t-T_2)] \tag{14}$$

with two concentrated delays on the right-hand side. It defines a relation being equivalent to the distributed delay (11).

The delay transformation introduced above can be applied to a generally defined delay distribution function $h(\tau)$. Let this function be given by N straightline parts continuously linking one to another (Fig. 3). All its $N+1$ break points are given by their coordinates $\tau=0, T_1, \ldots, T_N$ and $h=h_0, h_1, \ldots, h_N$. For such a case of delay integral $z(t)$ obtains the form

$$z(t)=\int_0^{T_N} x(t-\tau)dh(\tau)=\sum_{i=1}^{N}\int_{T_{i-1}}^{T_i}\frac{h_i-h_{i-1}}{T_i-T_{i-1}}x(t-\tau)d\tau. \tag{15}$$

Applying now repeatedly the transformation (11)→(12) and reversing the sequence of summation and integration executing, the following equation is received

$$\frac{dz(t)}{dt}=\sum_{k=0}^{N}C_k x(t-T_k) \tag{16}$$

where $T_0=0$, $C_0=-h_1/T_1$ and

$$C_k=\frac{h_{k+1}-h_k}{T_{k+1}-T_k}-\frac{h_k-h_{k-1}}{T_k-T_{k-1}},\quad C_N=-\frac{1-h_{N-1}}{T_N-T_{N-1}}. \tag{17}$$

With respect to (15), the initial condition $z(0)$ is given by the integral

$$z(0)=\int_0^{T_N} x(-\tau)dh(\tau). \tag{18}$$

In agreement with relations (17) and with the considered shape of $h(\tau)$ the coefficients C_k must satisfy the condition that

$$\sum_{k=0}^{N}C_k=0. \tag{19}$$

Equation (16) was derived on the assumption that the function $h(\tau)$ has the piecewise linear shape sketched in Fig. 3. However, this shape may also be considered as an approximation of an arbitrary distributed delay or step response function, too. In this way the above-mentioned delay transformation represents an advantageous approach to anisochronic model identification.

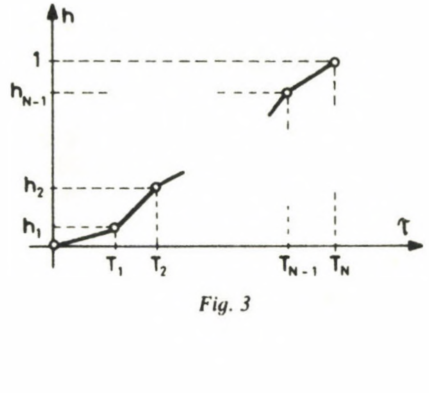


Fig. 3

So, by means of the presented transformation it is possible to substitute all the distributed delays in the original system by equivalent combinations of only concentrated delays and simultaneous integrations. Let it be stressed that the result of this transformation may also be interpreted as the following Stieltjes integral

$$\frac{dz(t)}{dt} = \int_0^{T_N} x_j(t-\tau)dg(\tau) \tag{20}$$

with discontinuous function

$$g(\tau) = \begin{cases} 0 \quad , & \tau < 0, \\ \sum\limits_{v=0}^{k} C_v, & \tau \in \langle T_k, T_{k+1} \rangle, \\ 0 \quad , & \tau > T_N, \end{cases} \tag{21}$$

so that it may be included into the general state equation form (7) as well. The only substantial difference in this analogy is the fact that the functions $g(\tau)$ have to end at zero for $\tau > T_N$, (not 1).

3. Characteristic function of anisochronic model

Similarly as for ordinary linear system (1), the basic dynamical properties of anisochronic system (7) may be indicated by its characteristic equation [1]

$$\det\left[\lambda \mathbf{1} - \int_0^T e^{-\lambda\tau} d\mathbf{A}(\tau)\right] = 0 \tag{22}$$

where λ is a complex-valued variable and $\mathbf{1}$ denotes the (n, n) unit matrix. This equation is difficult to be solved since it has an unlimited set of solutions in the complex domain of λ. Instead of this approach the left-hand side determinant may be interpreted as a so-called characteristic function M of system (7). After executing the determinant operation the function $M(\lambda)$ obtains the form

$$M(\lambda) = \lambda^n + \sum_{i=0}^{n-1} K_i \lambda^i \prod_s \left(\int_0^T e^{-\lambda\tau} dh_{ij}(\tau) \right)_s \tag{23}$$

where s is the counting index of multiplied functions in the determinant terms respectively.

There are two ways in evaluating the integrals in this formula. Firstly the integral functions can be analytically solved if a simple form of distribution h_{ij} is given. The second way, always possible to be applied, can be derived from the above introduced transformation of distributed delays.

Consider a given delay integral of the state equation (7) and the corresponding integral in the characteristic function M:

$$A_{ij}\int_0^T x(t-\tau)dh_{ij}(\tau)\rightarrow A_{ij}\int_0^T e^{-\lambda\tau}dh_{ij}(\tau). \tag{24}$$

Since the distributed delay defined by h_{ij} may be expressed as a combination of concentrated delays, $(x(t-T_k),\ k=0,1,2,\ldots)$, using formula (15), the appropriate integral in the characteristic function $M(\lambda)$ may also be replaced by an equivalent sum of exponential functions $e^{-\lambda T_k}$ divided by λ:

$$\int_0^T e^{-\lambda\tau}dh_{ij}(\tau)=\frac{1}{\lambda}\sum_{k=0}^{N} C_{ijk}\,e^{-\lambda T}{}_{k}. \tag{25}$$

By means of this substitution it is possible to express function (23) in the form

$$M(\lambda)=\lambda^n+\sum_{i=0}^{n-1}K_i\lambda^i\prod_s\left(\frac{1}{\lambda}\sum_{k=0}^{N}C_{ijk}\,e^{-\lambda T_k}\right)_s \tag{26}$$

where K_i and C_{ijk} are the coefficients and s is the counting index of functions (25) to be multiplied in the determinant terms.

4. Stability criterion for anisochronic models

The characteristic function $M(\lambda)$ is an effective mean for dynamic system analysis and synthesis. With respect to the argument increment rule of complex-valued functions, function (23) or (26) is able to indicate the asymptotic stability of the state equation (7). Since equation (22) belonging to a stable system (7) has no roots with non-negative real part, the argument of $M(\lambda)$ cannot achieve any non-zero increment in case, the variable λ has run round any arbitrary closed curve ψ lying completely inside the right half of the complex plane

$$\Delta_\psi \arg M(\lambda)=0. \tag{27}$$

Suppose a dynamic system with delayed inner interactions given by state equation (7). All its continuously distributed delays let be transformed to the concentrated ones substituting integrals (11) by the simultaneous ones (12). For this system formulation the characteristic function $M(\lambda)$ is given by formula (26).

The dynamic system defined in this manner is asymptotically stable if and only if the characteristic function (26) evaluated for positive imaginary $\lambda=j\omega$, $\omega\in\langle 0,\infty)$

represents a characteristic curve $M(j\omega)$ in the complex plane with the following two properties:

a) the starting point $M(0)$, for $\omega=0$, lies on the positive real axis,
b) the argument of $M(j\omega)$ approaches for $\omega\to\infty$ the value

$$\lim \arg M(j\omega)=n\frac{\pi}{2}. \tag{28}$$

Proof. With regard to the symmetry of $\lambda\to M$ mapping, all the test curves ψ for proving the argument increment may be situated in the first quadrant only. Consider a test curve ψ (Fig. 4), consisting of following three parts:

$$\begin{aligned} &\psi_1\colon \lambda=\beta+j0\,, && 0\leqq\beta\leqq R\\ &\psi_2\colon \lambda=Re^{j\varphi}\,, && 0\leqq\varphi\leqq\pi/2\\ &\psi_3\colon \lambda=0+j\omega\,, && 0\leqq\omega\leqq R. \end{aligned} \tag{29}$$

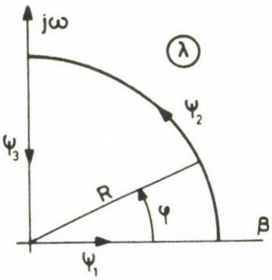


Fig. 4

Regarding that the function $M(\lambda)$ belonging to an asymptotically stable system must satisfy condition (27) for any arbitrarily large area defined by (29), its fulfilling must be proved for unlimited radius of ψ, i.e. for $R\to\infty$. The total increment of arg M is to be obtained as a sum of partial increments corresponding to the parts ψ_1, ψ_2, ψ_3, respectively. Because of real values of M the first one of them is zero

$$\Delta_{\psi 1} \arg M(\lambda)=0. \tag{30}$$

For the second part ψ_2 it is useful to consider $M(\lambda)$ as a product

$$M(\lambda)=\lambda^n[1+m(\lambda)] \tag{31}$$

where

$$m(\lambda)=\sum_{i=0}^{n-1} K_i\lambda^{i-n}\prod_{s}\left(\frac{1}{\lambda}\sum_{k=0}^{N} C_{ijk}e^{-\lambda T_k}\right)_s \tag{32}$$

and its total argument increment therefore to determine as a sum of the partial increments appertaining to λ^n and $1+m(\lambda)$, respectively. The first of these increments attains the well-known value

$$\Delta_{\psi 2}\arg(\lambda^n)=n\pi/2 \tag{33}$$

independently on the radius R. The remaining part of $M(\lambda)$, i.e. $1+m(\lambda)$ is a rather complicated expression, but in case of $\lambda=Re^{j\varphi}$, for $R\to\infty$ all the terms of $m(\lambda)$ approach zero

$$\lim_{R\to\infty}(e^{-\lambda T_k}/\lambda^i)_{\lambda=Rej^\varphi}=0 \tag{34}$$

because all the exponents i are positive. Consequently the function $m(\lambda)$ always approaches zero if $R\to\infty$ and therefore the total argument increment of $M(\lambda)$ along the path ψ_2 attains the value (33) for $R\to\infty$ as well:

$$\lim_{R\to\infty}\Delta_{\psi 2}\arg[\lambda^n(1+m(\lambda)]=n\frac{\pi}{2}. \tag{35}$$

It remains to consider the argument increment along ψ_3, i.e. along the positive imaginary axis. In case of no roots of $M(\lambda)$ with non-negative real part the total argument increment along the closed curve ψ must be zero. It is possible only if

$$\lim_{R\to\infty}\Delta_{\psi 3}\arg M(\lambda)=-n\frac{\pi}{2} \tag{36}$$

which is an equivalent property with the requirement (28).

As it follows from the principle of argument, the mentioned requirements of the $M(j\omega)$ criterion represent not only the sufficient condition of stability but also the necessary one. Let be emphasized that no additional properties of the curve $M(j\omega)$ besides the above-mentioned ones are to be required. Particularly the formulations claiming the number or the sequence of quadrants to be crossed by the curve $M(j\omega)$ must be avoided. Influenced by the transcendental terms namely, the shape of this curve can be considerably more complicated in case of system (7) than it is possible in any system of the type (1). Therefore the presented criterion cannot be changed for the usual formulation of Mikhaylov–Leonhard criterion.

5. Conclusions

The presented method of hereditary system modelling can be applied to the analytically mastered theoretical problems as well as to such ones needing to work with empirically obtained data. By the use of analytical solutions of appropriate partial differential equations it is even possible to describe some systems with distributed parameters by the above-mentioned anisochronic model. It is demonstrated in [5] on a model of a long-distance belt conveyor transport line. In a similar way some dynamical properties of long pipelines (e.g. the resonance effects) may be described by the presented method. Since the necessity of the distributed delay modelling is avoided due to the above introduced delay transformation, originally functional state equations (7) are consistently replaced by equations with concentrated delays only. This approach enables us to achieve models relatively easily applicable on computers.

As to the above presented method of characteristic function, it is necessary to emphasize that the opportunities of its application are considerably larger than it was demonstrated. By means of $M(\lambda)$ also the natural oscillations of a hereditary system can be easily identified, [7]. Using this function the control synthesis of a hereditary system or its model identification can be performed [8], [9] etc. The importance of the application of $M(\lambda)$ lies particularly in the fact that other usual methods of stability proof and system analysis fail in case of hereditary phenomena investigation.

References

1. *Myshkis, A. D.*, Lineynye differentsialnye uravneniya s zapazdyvayushchim argumentom. Nauka, Moskva, 1972.
2. *Kolmanovskij, V. B.*, *Nosov, V. R.*, Ustoychivost i perioditcheskie rezhimy sistem s posledeystviem. Nauka, Moskva, 1981.
3. *Hale, J.*, Theory of Functional Differential Equations. Springer, New York, 1977.
4. *Volterra, V.*, Theory of Functionals and of Integro-Differential Equations. Dover Publ. Inc., New York.
5. *Zítek, P.*, Anisochronic Generalization of Dynamic System State Theory, In: IFAC 4. Symp. on Autom. in Mining, Min. and Metal Proc., Helsinki, 1983.
6. *Zítek, P.*, Anisochronic State Theory of Dynamic Systems, Acta Technica ČSAV, **4**, Academia, Praha, 1983.
7. *Zítek, P.*, Stability Criterion for Anisochronic Dynamic Systems, Acta Technica ČSAV, **4**, Praha, 1984.
8. *Zítek, P.*, Parameter Identification of System with Delays Using the Charact. Function (in Czech), Automatizace **25**, No. *11*, 1982.
9. *Zítek, P.*, Synthesis of a Control with Transport Delays (in Czech), Automatizace **25**, No. *6*, 1982.

Анизохронные модели и критерий устойчивости эредитарных систем

П. ЗИТЕК

(Прага)

Запаздывания, последействия и другие эредитарные свойства представляют собой важные вопросы в задачах исследования объектов управления. Для этого рода объектов и явлений классическая теория состояния системы не является удобным и эффективным средством их анализа. Ввиду этого в статье предлагается применение так называемого анизохронного состояния динамической системы к выражению эредитарных свойств объекта. В связи с этим введено преобразование исходной системы, позволяющее выразить любое последействие объекта при помощи всего лишь сосредоточенных запаздываний. Для таким образом сформулированной динамической системы выведен критерий устойчивости, основанный на принципе аргумента её характеристической функции.

P. Zítek
Faculty of Mechanical Engineering
Czech Technical University
Suchbátarova Str. 4
166 07 Prague 6
Czechoslovakia

Problems of Control and Information Theory, Vol. 15 (6), pp. 425—438 (1986)

A DRIFT ALGORITHM IN CONTROL OF UNCERTAIN PROCESSES

S. V. EMELYANOV, S. K. KOROVIN, L. V. LEVANTOVSKIY

(*Moscow*)

(Received December 17, 1985)

Control of an uncertain process (process with the mathematical model known only to be one of a class of models) described by a smooth dynamic equation system nonlinearly dependent on control is reduced to making a certain scalar function σ of the process vanish. This is done by control algorithms referred to as drift algorithms which use the values of σ taken at discrete times and generate control which is limited, continuous, and piecewise-smooth in time. Conditions are specified under which the magnitudes of σ do not exceed, after some time, a value which is proportional to squared time quantization step. In effect, drift algorithms are second-order real slide algorithms.

1. Introduction

One approach to designing control of dynamic systems

$$\dot{x}=f(t,x,u), \quad t\in I_0=[t_0,+\infty) \tag{1.1}$$

is to impose on the system state an artificial constraint

$$\sigma(t,x)=0\,. \tag{1.2}$$

In equations (1.1) and (1.2) t and t_0 are the current and initial times: x is a state variable with values in X; X is $\mathbf{R}^n$, a smooth real Riemann n-dimensional manifold or Banach space; u is control from a set of feasible controls $\mathscr{U}$; f and σ are mappings defined on associated sets. Relation (1.2) is made to hold by proper choice of control in the form $u(t,x)$, $u(t,\sigma)$, or $u(t,x,\sigma)$. The constraint (1.2) is specified arbitrarily so that when it holds the closed-loop control system

$$\dot{x}=f(t,x,u(t,x,\sigma)), \quad \sigma=0$$

possesses the desired set of qualities from a certain time.

Some of the conditions and ways to choose the constraints have been reported elsewhere [1–6]. When the constraint is specified the control problem is reducible to a simpler problem of making the constraint hold. The specifics of solving this latter problem is completely defined by the set $\mathscr{U}$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

This approach has two advantages. The order of motion equations reduces when relation (1.2) holds, which simplified the analysis and design of the control system. In addition, the motion equations (1.1), (1.2) are independent of some (not any) variations of the right-hand side of (1.1). The latter fact makes this approach effective in control of uncertain dynamic systems.

There are two different ways to implement this approach. In one the control is usually chosen in the form $u = k\sigma$ where $|k|$ is a large, in the limit an infinite gain [6]. In the other the control is discontinuous and can be represented as $u = k\varphi(t, x) \operatorname{sign} \sigma$ where k is a limited gain, $\varphi(t, x)$ is a piecewise-smooth function, strictly positive at every admissible t and $x \neq 0$, and sign is a sign function ($\operatorname{sign} \sigma = \sigma/|\sigma|$, $\sigma \neq 0$) [1, 5].

When the control $u = k\sigma$ the solution is accurate only in the limit with $|k| = \infty$. When the control is discontinuous, $u = k\varphi \operatorname{sign} \sigma$, relation (1.2) is maintained in sliding mode. In effect, in one case the solution is obtained with continuous but infinite control and in the other, with discontinuous control limited in the finite part of X. Therefore the question arises whether there are sufficiently smooth limited controls which solve this problem. A positive answer has been given in [7] where a newly introduced concept of high order sliding mode on the manifold $\sigma = 0$ was used.

An m-th order sliding mode ($m > 1$) is obtained by artificially expanding the state space of dynamic system (1.1) through introduction of several new state variables, including u. Unlike the dynamics of process (1.1) that of the additional state variables (probably discontinuous) is specified arbitrarily, the closed-loop system being understood in Filippov's sense [8]. The m-th order slide is said to occur when under any initial conditions the process goes, within finite time, to the mode $\sigma(t, x(t)) \equiv 0$, all time derivatives by virtue of the system $x, \dot{x}, \ldots, x^{(m-1)}$ being continuous as vector functions of the closed-loop system state variables and time while the next derivative is discontinuous. In this sense a conventional sliding mode [1, 5] is of the first order ($m = 1$).

Several scalar control algorithms which ensure accurate or approximate second-order sliding are mentioned in [7]. Such sliding modes are important in applications as well as in theory. Indeed, in applications the sliding modes are called real. They are featured by a finite "amplitude" of the state point deviation from the manifold $\sigma = 0$ and a limited switching "frequency". Let $\tau = \text{const} > 0$ be a parameter of a real sliding mode and such that with $\tau = 0$ the real sliding becomes ideal. This parameter may be delay in switching. For a usual sliding mode and real sliding the relation $\sup |\sigma| = O(\tau)$ is known to hold. For second-order sliding modes in real sliding the relations $\sup |\sigma| = O(\tau^2)$ may hold [7]. This implies that the effect of small imperfections on sliding modes may decrease as the slide order increases. Since the accuracy of maintaining equality (1.2) directly affect the control performance an increase of the sliding order may prove significant in applications.

This article will thoroughly analyze the properties of various control algorithms mentioned in [7] and ensuring real second-order sliding. All the algorithms below are drift algorithms. In addition to the above properties they are different from other second-order slide algorithms in that the sign of σ is constant until this quantity reaches a certain vicinity of zero, which is interpreted as absence of overshoot for σ. Furthermore, drift algorithms have no sense other than real sliding. When the switching delay goes to zero and the sliding becomes ideal they lose their properties and do not perform their function. This illustrates the fact that an ideal sliding mode of any order may, under certain circumstances, hinder solution and so only real sliding can do the job.

The article will discuss application of drift algorithms to the control of nonlinear uncertain dynamic systems linearly dependent on control and discrete simulation of closed-loop systems which illustrate the main theoretical findings.

2. Statement of the problem, basic assumptions, and drift algorithm equations

Assumptions. The smooth dynamic system (1.1) is considered whose state space can be anything from among those specified in Introduction, even the dimension of X being unimportant. The range of admissible controls $\mathcal{U}$ is assumed to consist of continuous and limited scalar controls

$$|u| \leqq 1+\kappa \tag{2.1}$$

where $\kappa \in (0, 1)$ is a number. Solutions of (1.1) are assumed infinitely extendible in time with any admissible control.

Relation (1.2) is assumed to be specified by a smooth functional $\sigma : I_0 \times X \to \mathbf{R}$. With these assumptions the derivative $\frac{d\sigma}{dt} = \dot{\sigma}(t, x, u)$ is, by virtue of system (1.1), certain to be smooth.

Let $\bar{\sigma}, u_0, K_m$, and K_M be some positive constants, $u_0 \in (0, 1)$. At every t and x such that $|\sigma| \leqq \bar{\sigma}$ let the derivative $\dot{\sigma}$ be negative if $u \leqq -u_0$ and positive if $u \geqq u_0$. Let also the inequalities

$$K_m \leqq \frac{\partial \dot{\sigma}}{\partial u} \leqq K_M \tag{2.2}$$

be true. These assumptions ensure existence of a unique smooth function $u_{eq}(t, x)$ which is the solution of the equation $\dot{\sigma}(t, x, u_{eq}) = 0$ when $|\sigma| \leqq \bar{\sigma}$, and $|u_{eq}| < u_0$ at every specified t and x.

Let the resultant function u_{eq} satisfy the condition that at every t and x such that $|\sigma(t, x)| \leqq \bar{\sigma}$ for a certain number $\alpha_0 > 0$

$$\sup \left| \frac{du_{\mathrm{eq}}}{dt} \right| < \alpha_0 \, . \tag{2.3}$$

Inequality (2.3) limits the rate of variation of the right-hand side of (1.1). In meaningful terms it can be interpreted as bound on the rate of variations of the process "parameters" and "exogenous disturbances".

For simplicity, let us consider the behavior of those closed-loop systems with only such initial states $x(t_0)$ for which $|\sigma(t_0, x(t_0))| < \bar{\sigma} - \delta_0$ where δ_0 is a constant, $\delta_0 \in (0, \bar{\sigma})$, or only in a certain vicinity of the manifold specified by relation (1.2). Then there is no need to solve the well-explored problem which has no bearing on the main subject of the research of making the state point go from an arbitrary initial position into the "strip" $|\sigma| < \bar{\sigma} - \delta_0$.

Statement of the problem. Under the above conditions and bounds it is required to specify an algorithm for which in steady state the oscillation amplitude of σ around zero is proportional to squared delay of switching. This algorithm will be referred to as a second-order real sliding algorithm.

If the control is not required to be continuous, then under the above conditions a possible first-order real slide algorithm is

$$u = -\operatorname{sign} \sigma(t - \tau, x(t - \tau))$$

where $\tau = \mathrm{const} > 0$ is the delay time constant in computing the functional σ. As $\tau \to 0$ this algorithm becomes a first-order ideal sliding algorithm. At $\tau > 0$ the relations

$$\sup |\dot{\sigma}| = O(\tau) \quad \text{and} \quad \sup |\dot{\sigma}| = O(1)$$

hold. It is required to specify a control algorithm for which relations of the form

$$\sup |\sigma| = O(\tau^2) \quad \text{and} \quad \sup |\dot{\sigma}| = O(\tau)$$

are true. To solve this problem, let us introduce a smooth slide manifold

$$L = \{(t, x, u) \,|\, \sigma(t, x) = \dot{\sigma}(t, x, u) = 0\} \, . \tag{2.4}$$

Under these conditions it can also obviously be specified as

$$L = \{(t, x, u) \,|\, \sigma(t, x) = 0, u = u_{\mathrm{eq}}(t, x)\} \, . \tag{2.5}$$

It is also obvious that the motion of a closed-loop system in a small vicinity of the manifold $L$is approximately described by the equation

$$\dot{x} = f(t, x, u_{\mathrm{eq}}(t, x)) \, . \tag{2.6}$$

Note that equations such as (2.6) also occur in the theory of systems whose controls are discontinuous with first-order sliding but are not provable unless the right-hand side of (1.1) linearly depends on control.

The drift algorithms. Let us specify the imperfection which leads to a real second-order sliding. It will be assumed to be caused by discontinuity in measuring the functional $\sigma(t)=\sigma(t, x(t))$.* Let $t_0, t_1, t_2, \ldots, t_i, \ldots$ be a sequence of times such that the differences $\tau_i=t_i-t_{i-1}$ satisfy with every $i=1, 2, \ldots$ the inequalities

$$0<\tau_m \leqq \tau_i \leqq \tau_M \tag{2.7}$$

with some constants τ_m and τ_M. Using this sequence let us define the function $\delta\sigma(t)$; for every $i=1, 2, \ldots$ and $t \in [t_i, t_{i+1})$

$$\delta\sigma(t)=\sigma(t_i)-\sigma(t_{i-1}) \tag{2.8}$$

and at $t \in [t_0, t_1)$ let $\delta\sigma(t)=\sigma(t_0)$.

The drift algorithms have the form

$$\dot{u}=\begin{cases} -u & \text{with} \quad |u| \geqq 1\,, \\ -\alpha_M \operatorname{sign} \delta\sigma & \text{with} \quad \sigma \cdot \delta\sigma>0, |u|<1\,, \\ -\alpha_m \operatorname{sign} \delta\sigma & \text{with} \quad \sigma \cdot \delta\sigma \leqq 0, |u|<1\,, \end{cases} \tag{2.9}$$

where $\alpha_M>\alpha_m>0$ are drift algorithm parameters. Particular cases of the algorithms occur with constant and variable τ_i $(i=0, 1, 2, \ldots)$ which will be referred to as constant and variable step drift algorithms.

The right-hand side of (2.9) is discontinuous and so its solution is understood to be a totally continuous function which satisfies it almost everywhere.

In meaningful terms the drift algorithms can be explained as follows. When $u=u_{\text{eq}}(t, x)$ is substituted in (1.1) the functional σ has a constant value. The drift algorithm generates oscillations of the control signal around $u=u_{\text{eq}}(t, x)$ and associated oscillations of $\dot{\sigma}$ around zero. These oscillations can be made such that the functional σ change in the desired way, for instance, σ drift to zero. Algorithm (2.9) is such an algorithm; other varieties are obtained when α_m and α_M are functions of t, x, and u or $\delta\sigma(t)=\sigma(t)-\sigma(t-\tau)$ where τ is a positive function of t and σ.

* Wherever ambiguity cannot occur all arguments other that time t of the functional computed on solution of system (1.1) with a certain chosen control $u(t)$ are omitted.

3. The drift algorithm mechanism. Properties of a constant step algorithm

Let us consider processes in a closed-loop system (1.1), (2.9). Within the region $|\sigma|<\bar{\sigma}$ the identity holds

$$\dot{\sigma}(t, x, u)=\int\limits_{u_{\mathrm{eq}}(t, x)}^{u} \frac{\partial \dot{\sigma}}{\partial u}(t, x, \xi)\, d\xi . \tag{3.1}$$

Denote $\varepsilon=u-u_{\mathrm{eq}}$ and $K=\dot{\sigma}/\varepsilon$. Then the equalities hold

$$\begin{cases}\dot{\sigma}=K(t, x, u)\varepsilon ,\\ \dot{\varepsilon}=\dot{u}-\dot{u}_{\mathrm{eq}}(t, x, u) ,\end{cases} \tag{3.2}$$

where the continuous function K satisfies, by virtue of (2.2) and (3.1), the inequalities

$$0<K_m \leqq K(t, x, u) \leqq K_M .$$

System (3.2) is of key importance in proving the results below. Let the inequality hold

$$\alpha_m>\alpha_0 \tag{3.3}$$

which ensures that the signs of $\dot{\varepsilon}$ and $\dot{u}$ coincide. Let us take up the dynamics of the variables σ and ε with $|\sigma|<\bar{\sigma}$ (Fig. 1).

The axis $\varepsilon=0$ on the plane (σ, ε) represents the manifold $\dot{\sigma}=0$ and the origin of coordinates, the slide manifold L. In the limit as $\tau_i \to 0$ in (2.9) sign $\delta\sigma(t)$ should be replaced by sign $\dot{\sigma}(t)$ and it follows from (3.2) and (3.3) that a sliding mode occurs on the manifold $\dot{\sigma}=0$ with which the value of the functional σ is constant. When (2.7) holds, a switching mode occurs which results in a drift of the value of σ.

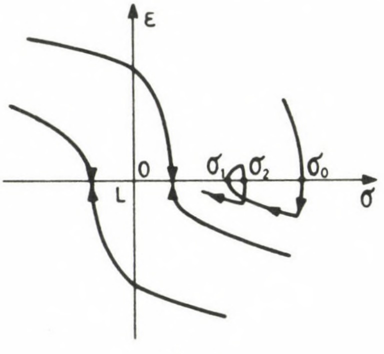


Fig. 1

The switching mode in the region $|\sigma|<\bar{\sigma}$ which follows the first vanishing of $\dot{\sigma}$ will be referred to as the *drift mode*.

A *cycle* will be a stretch of the path $(\sigma(t), \varepsilon(t))$ which does not intersect the axis $\sigma=0$ and stays between two successive transitions from the region $\sigma\dot{\sigma}>0$ into the region $\sigma\dot{\sigma}<0$.

Figure 1 shows a cycle which leaves point σ_0 on the axis σ, intersects the axis σ in point σ_1 and ends in point σ_2. The direction of the drift is dictated by the sign of the difference $\sigma_2-\sigma_0$.

The following inequality may be shown sufficient for the convergence of the functional σ to a vicinity of zero in drift mode when a constant step algorithm is used

$$4\frac{K_M}{K_m}\left(1+\sqrt{\frac{K_M}{K_m}\frac{(\alpha_M+\alpha_0)}{(\alpha_M-\alpha_0)}}\right)^2\frac{(\alpha_M+\alpha_m-2\alpha_0)}{(\alpha_M+\alpha_m+2\alpha_0)}\frac{(\alpha_M+\alpha_0)}{(\alpha_M-\alpha_0)}\frac{(\alpha_m+\alpha_0)^3}{(\alpha_M-\alpha_0)^3}<1. \tag{3.4}$$

The inequalities below ensure invariance of the region $|\sigma|<\bar{\sigma}$ with an initial condition $|\sigma(t_0)|<\bar{\sigma}-\delta_0$

$$\begin{gathered}\alpha_M>\alpha_0+K_M(1+u_0)^2/(2\delta_0)\,,\\ \frac{1}{2}K_M\frac{(\alpha_M+\alpha_m-2\alpha_0)}{(\alpha_M-\alpha_0)(\alpha_m-\alpha_0)}(\alpha_m+\alpha_0)^2\tau_M^2<\delta_0\,.\end{gathered} \tag{3.5}$$

The following theorem expresses the main properties of a constant step drift algorithm.

Theorem 1. If all assumptions of Section 2 and conditions (3.3)–(3.5) hold, then there are constants $a>0$ and $\tilde{a}>0$ such that any solution of system (1.1), (2.9) at $\tau_i=\tau=\text{const}$, $\tau_m\leqq\tau\leqq\tau_M$ reaches within finite time the set

$$\{(t,x,u)\,|\,|\sigma(t,x)|\leqq a\tau^2, |\dot{\sigma}(t,x,u)|\leqq\tilde{a}\tau\}$$

to stay there.

The constants a and $\tilde{a}$ are provided by the expressions

$$a=9K_M(\alpha_M+\alpha_0)^2,\qquad \tilde{a}=3(\alpha_M+\alpha_0)K_M\,. \tag{3.6}$$

The proof is made possible by the following

Lemma. Under the conditions of Theorem 1 there are positive constants r_1 and r_2 that depend on α_m, α_M, α_0, K_m, and K_M and are such that for any cycle that intersects the axis σ successively in the points σ_0, σ_1, and σ_2 with $\sigma_0>0$ the inequalities hold

$$r_2\tau^2<\sigma_0-\sigma_2<\sigma_0-\sigma_1<r_1\tau^2\,.$$

The formulae for some constant are rather cumbersome and so are omitted. Note that the inequality $r_2>0$ is equivalent to (3.4). The attraction area is easily shown to be defined by inequalities of the form $|\sigma|<a\tau^2$ and $|\dot{\sigma}|<\tilde{a}\tau$.

The Lemma also shows that the duration of the transient process at $\tau_i = \tau = \text{const}$ is uniformly limited by a quantity about τ^{-1}. A variable τ_i will be shown to lead to a duration which may be as short as desired with the attraction area being defined by the inequalities $|\sigma| < a\tau_m^2$ and $|\dot{\sigma}| < \tilde{a}\tau_m$.

Theorem 1 remains true in another statement of the problem where the sign of σ in (2.9) is determined at the same times t_i, $i = 0, 1, 2, \ldots$ (The constants $\tilde{a}$ and a in (3.6) are chosen fairly large.)

4. A variable step of the constraint function increment measurement

The relations which specify a drift algorithm where τ_i is variable are

$$\dot{x} = f(t, x, u) \tag{4.1}$$

$$\dot{u} = \begin{cases} -u & \text{with } |u| \geqq 1, \\ -\alpha_m \operatorname{sign} \delta\sigma(t) & \text{with } \sigma \cdot \delta\sigma < 0, |u| < 1, \\ -\alpha_M \operatorname{sign} \delta\sigma(t) & \text{with } \sigma \cdot \delta\sigma > 0, |u| < 1; \end{cases} \tag{4.2}$$

$$\delta\sigma(t) = \begin{cases} \sigma(t_i) - \sigma(t_{i-1}) & \text{with } t_i \leqq t < t_{i+1}, \quad i = 1, 2, \ldots, \\ \sigma(t_0) & \text{with } t_0 \leqq t < t_1; \end{cases} \tag{4.3}$$

$$t_{i+1} - t_i = \tau_{i+1} = \tau(\sigma(t_i)), \quad i = 0, 1, 2, \ldots,$$

$$\tau(s) = \begin{cases} \tau_M & \text{with } \tilde{\tau}(s) \geqq \tau_M, \\ \tilde{\tau}(s) & \text{with } \tau_m < \tilde{\tau}(s) < \tau_M, \\ \tau_m & \text{with } \tilde{\tau}(s) \leqq \tau_m. \end{cases} \tag{4.4}$$

Here $\tilde{\tau}$, $\tilde{\tau}(s) = \tilde{\tau}(-s)$ is an even function which does not decrease with $s > 0$. The variable step drift algorithm (4.2)–(4.4) is said to *converge with utmost accuracy* if with any initial values t_0, $x(t_0)$, and $u(t_0)$ the solution of system (4.1)–(4.4) reaches within finite time the region

$$\{(t, x, u) \,|\, |\sigma(t, x)| < a\tau_m^2, \ |\dot{\sigma}(t, x, u)| < \tilde{a}\tau_m\}$$

where the constants a and $\tilde{a}$ are taken from (3.5) to stay there.

Overshoot is said to occur if in drift mode, Section 3, the process state leaves the region $\{(t, x) \,|\, |\sigma(t, x)| < a\tau_m^2\}$. Under the assumptions of Section 2 the duration of switching to drift mode does not exceed $2/\alpha_M + \tau(\sigma(t_0))$.

Theorem 2. If the assumptions of Section 2 and inequalities (3.3)–(3.5) hold and if with $\tau(\sigma) \geqq \tau_m$ the strict inequality $r_1 \tau^2(\sigma) < \sigma$ where r_1 is a constant from the Lemma of

Section 3 is true, then algorithm (4.2)–(4.4) converges without overshoot with utmost accuracy.

The proof follows from the truth of the inequality $0<\sigma_0-\sigma_2<r_1\tau^2(\sigma_0)$ with $\sigma_0>0$ and $\sigma_1>0$ (see the Lemma).

Theorem 3. If $\tilde{\tau}(\sigma)=\nu|\sigma|^\rho$ where $\nu>0$ and $0.5\leqq\rho<1$ and the assumptions of Section 2 and inequalities (3.3)–(3.5) hold, then with ν fairly small algorithm (4.2)–(4.4) converges without overshoot with utmost accuracy, the convergence time being limited by a constant $T(\nu)$ which does not depend on the initial conditions or on τ_m.

Theorem 2 provides the most important constraint on ν. Another restriction is produced by evaluation of the convergence time proportional to $\int|d\varepsilon(t)|$. The technique is similar to that of the lemma.

With $\rho=1$ the convergence is exponential-like in that the convergence time is proportional to $\ln\tau_m^{-1}$. With $\rho>1$ the convergence is still slower and with $0<\rho<0.5$ it is necessary to introduce the constraint $\nu=\nu(\tau_m)$.

5. Application of a drift algorithm to dynamic systems linearly dependent on control

The process is described by a dynamic system

$$\dot{x}=g(t,x)+b(t)u$$

where $x\in\mathbf{R}^n$ is the process state, g, $b\in\mathbf{R}^n$ are smooth vector functions, and $u\in\mathbf{R}$ is scalar control. It is required that the function $\sigma(t,x)=\sigma_1(t)+(c(t),x)$ be vanished where $\sigma_1(t)\in\mathbf{R}$ and $c(t)\in\mathbf{R}^n$ are smooth functions. With the following assumptions holding, this problem is reducible to that described above.

A) Let $\Phi(x)=\sqrt{(x,x)+h}$ where $h>0$. Let the quantities

$$g/\Phi,\quad \frac{\partial g}{\partial t}\bigg/\Phi,\quad \frac{\partial g}{\partial x},\quad b,\quad \frac{\partial b}{\partial t},$$

$$\frac{\partial\sigma_1}{\partial t}\bigg/\Phi,\quad \frac{\partial^2\sigma_1}{\partial t^2}\bigg/\Phi,\quad \frac{\partial c}{\partial t},\quad c\quad\text{and}\quad \frac{\partial^2 c}{\partial t^2}$$

be bounded.

B) For some $\delta_1>0$ at every t the inequality $(b(t),c(t))>\delta_1$ holds.

C) At every time the values of the functions $\Phi(x)$ and the functional σ are known.

The control signal is generated as

$$u=\mu k\,\Phi(x)$$

and the quantity $\varphi=\sigma/\Phi$ is introduced. With k fairly large it is easy to see that the system

$$\dot{x}=g(t,x)+\mu b(t)k\Phi(x) \tag{5.1}$$

where μ, $|\mu|\leqq 1+\kappa$ is regarded as a new control and the function φ as a new constraint function satisfies the assumptions of Section 2 on system (1.1). The explicitly numerical bounds in assumptions A) and B) make it possible to compute the parameters $\alpha_0, \bar{\sigma}, K_m$, and K_M. The new control μ and constraint function φ transform (4.2)–(4.4). Theorems equivalent to Theorems 1–3 could be easily formulated. In steady state the maximal deviation of σ from zero is proportional to $k\Phi(x)\tau_m^2$.

6. Simulation results

A system such as (5.1), $x\in\mathbf{R}^3$ that satisfies assumptions A) and B) is chosen for simulation. The functions $g(t, x)$ and $b(t)$ are represented in the form

$$g=\begin{pmatrix} -5x_1\cdot\sin t+10x_2+4x_3 \\ 6x_1-3x_2-2x_3+3(x_1+x_2+x_3)\cos t \\ x_1+4x_2\cdot\cos 5t+3x_3+4\sin 5t \end{pmatrix},$$

$$b=\begin{pmatrix} 0 \\ 0 \\ 1+0.5\cos 10t \end{pmatrix}.$$

To demonstrate how a dynamic system which nonlinearly depends on control is handled, assume that

$$\mu=u^3-u^2/2+3u-\sin u\cdot\cos 30t\,.$$

The functions Φ and σ are taken in the form

$$\Phi(x)=\sqrt{(x,x)}+1, \quad \sigma=x_3/\Phi(x)\,.$$

In this case the loss of smoothness in the right-hand side of (5.1) in the point $x=0$ may be shown to have no effect on the above results. The associated system (5.1), (4.2)–(4.4) was integrated by the Euler method with an integration step of $2\cdot 10^{-4}$. The algorithm parameters and initial conditions had the following values

$$k=10, \quad \alpha_M=40, \quad \alpha_m=8, \quad \tau_M=0.015, \quad \tau_m=5\cdot 10^{-4},$$

$$x_1=2, \quad x_2=-2, \quad x_3=10, \quad u=0, \quad t_0=0\,.$$

Figures 2 and 3 plot $\sigma(t)$ with $\tilde{\tau}(\sigma)=0.05\,|\sigma|$ and $\tilde{\tau}(\sigma)=0.02\sqrt{|\sigma|}$, respectively. In both cases the accuracy is $|\sigma|\leqq 7\cdot 10^{-4}$. Figure 4 shows the function $u(t)$ in the latter case; the function is seen to follow a certain time function starting from a certain instant. Figure 5 plots the functions $\mu(u(t))$ and

$$\mu_{\mathrm{eq}}(t, x(t)) = -g_3(t, x(t))/(k\Phi(x(t))b_3(t))$$

in the same case. As was to be expected, from a certain time the function $\mu(u(t))$ follows the function $\mu_{\mathrm{eq}}(t, x(t))$. When τ_m is reduced to its one tenth, the amplitude of σ was found to drop to one percent of its value.

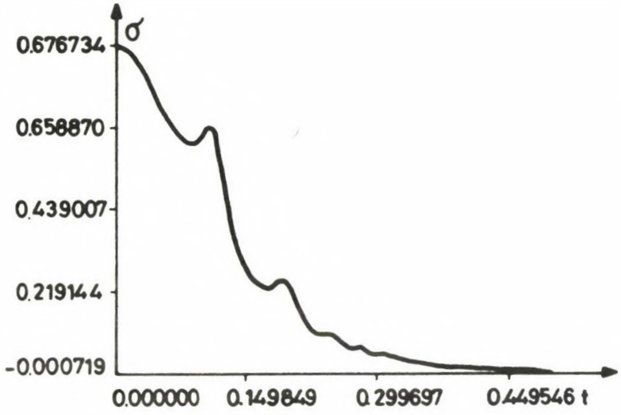


Fig. 2

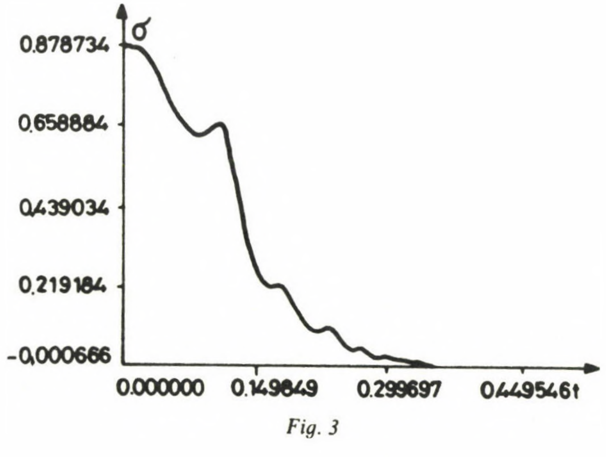


Fig. 3

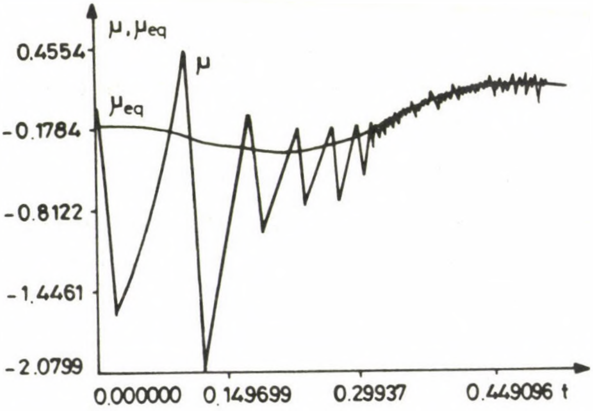


Fig. 4

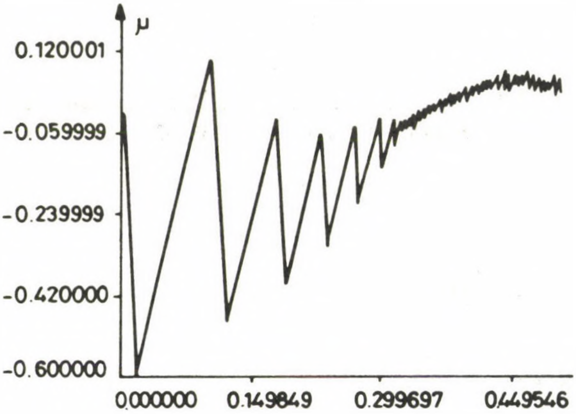


Fig. 5

The initial values of t and x were taken from outside the linear zone $|\sigma| < \bar{\sigma}$. Some reasoning and simulation reveal that the convergence of the algorithm is a rule rather than an exception in this case. To make sure that the convergence time is short enough, however, some techniques have to be employed that ensure reaching the linear zone.

7. Additional remarks

A) A drift algorithm can be specified in a more general way by making $\dot{u}$ equal to $\alpha_1, \alpha_2, \alpha_3$, or α_4 depending on the combination of signs of σ and $\delta\sigma$. Here α_j are limited measurable functions of t, x, and u and the equation is understood in Filippov's sense. Conditions are formulated that are similar to (3.3)–(3.6) for upper and lower bounds of α_j.

B) If in (2.9) $\delta\sigma$ is replaced by the output of some real differentiator, a similar drift mode may probably be expected.

C) The variable step $\tau(\sigma)$ introduced in Section 4 may be shown to make the drift algorithm and other algorithms of [7] stable with small high frequency noises Δf and $\Delta\sigma$.

References

1. *Emelyanov, S. V.* (ed.), Teoriya sistem s peremennoy strukturoi. Moscow: Nauka, 1970, 592 pp.
2. *Emelyanov, S. V.*, Binarnye sistemy avtomaticheskogo regulirovaniya. Ser. "Binarnye dinamicheskie sistemy". Issue 1, Moscow: IRIMS, 1984, 313 pp.
3. *Emelyanov, S. V., Korovin, S. K., Sizikov, V. I.*, Ser. "Binarnye dinamicheskie sistemy", Issues 2–3, Moscow: IRIMS, 1983.
4. *Emelyanov, S. V., Korovin, S. K., Sizikov, V. I.*, Ser. "Binarnye dinamicheskie sistemy", Issues 4–5, Moscow, IRIMS, 1983.
5. *Utkin, V. I.*, Skol'zyashchie rezhimy v zadachakh optimizatsii i upravleniya. Moscow: Nauka, 1981, 368 pp.
6. *Meerov, M. V.*, Sistemy monogosvyaznogo regulirovaniya. Moscow. Nauka, 1967.
7. *Emelyanov, S. V., Korovin, S. K., Levantovskiy, L. V.*, Skol'zyashchie rezhimy vysshikh poriadkov v binarnykh sistemakh upravleniya. DAN SSSP, 1985, Vol. **287**, No. *6*.
8. *Filippov, A. F.*, Differentsyal'nye uravneniya s razryvmoy pravoy chast'yu. Moscow: Nauka, 1985, 224 pp.

Алгоритм дрейфа при управлении неопределенными объектами

С. В. ЕМЕЛЬЯНОВ, С. К. КОРОВИН, Л. В. ЛЕВАНТОВСКИЙ

(Москва)

Проблема управления гладкой динамической системой $\dot{x} = f(t, x, u)$, где $x \in X$ — пространство состояний системы; u — скалярное управление, $|u| \leqq \text{const}$, сводится к задаче возможно более точного выполнения связи $\sigma(t, x) = 0$, здесь σ — гладкая скалярная функция, при условии измерения значений σ в дискретные моменты времени $t_0, t_1, \ldots, 0 < \tau_m \leqq \tau_i = t_i - t_{i-1} \leqq \tau_M$, $i = 1, 2, \ldots$. Для известных методов решения этой задачи в установившемся по σ режиме выполняются соотношения $\sup|\sigma| = O(\tau_m)$, $\sup|\dot{\sigma}| = O(1)$. Предлагаются алгоритмы управления, называемые алгоритмами дрейфа, которые формируют непрерывное кусочногладкое управляющее воздействие, обеспечивают переходный процесс по σ без перерегулирования и выполнение в установившемся режиме соотношений $\sup|\sigma| = O(\tau_m^2)$, $\sup|\dot{\sigma}| = O(\tau_m)$.

При выполнении указанных соотношений алгоритм управления относится к классу алгоритмов реального скольжения второго порядка. Системы управления с подобными алгоритмами управления менее чувствительны к запаздываниям в переключениях, чем системы управления с обычным скользящим режимом или скользящим режимом первого порядка. В статье при естественных предположениях указываются соотношения, при которых алгоритмы дрейфа являются алгоритмами реального скольжения второго порядка. Рассматриваются два типа алгоритмов дрейфа: с постоянным ($\tau_i = \text{const}$) и с переменным ($\tau_i = \tau_i(\sigma(t_{i-1}, x(t_{i-1}))))$ шагом дискретизации по времени. Использование переменного шага дискретизации позволяет время переходного процесса по σ задавать по произволу. Все упомянутые выше факты справедливы в случаях, когда X — пространство $\mathbf{R}^n$, гладкое вещественное риманово многообразие или банахово пространство. Размерность не играет роли. В статье описан пример использования алгоритмов дрейфа для управления неопределенной системой с линейным вхождением управления, приведены результаты численного моделирования замкнутых систем с нелинейным вхождением управления.

С. В. Емельянов
Всесоюзный научно-исследовательский институт
системных исследований Госплан СССР, АН СССР
СССР Москва 117312,
просп. 60 летия Октября

Problems of Control and Information Theory, Vol. 15 (6), pp. 439—450 (1986)

PARETO-OPTIMALITY IN STOCHASTIC DIFFERENTIAL GAMES

S. D. Gaidov

(*Rousse*)

(Received November 15, 1985)

The paper deals with many-player nonzero-sum games where the dynamics is described by Ito stochastic differential equations. Sufficient conditions for Pareto-optimality of the strategies of the players are established. Linear-quadratic games without and with a control dependent noise are studied.

1. Introduction

There is a variety of different approaches to the theory of stochastic differential games. Let us mention especially the martingale methods of M. H. A. Davis, R. J. Elliott and P. P. Varaiya. In this paper we shall follow the approach of Fleming and Rishel [4] to optimal control of stochastic dynamic systems, but we have some ideas to use the above one [2] as well. The approach chosen here gives the opportunity further to make use of the approximation and numerical methods developed by Chernousko and Kolmanovskii [1].

In [9] Varaiya studied Sleiter-optimality in stochastic differential games. In this paper we consider a slightly more precise concept of a cooperative solution, i.e. Pareto-optimal strategies. This aspect of the problem for deterministic differential games is described in [8]. The linear regulator problems [1], [4], [7] are the background for considering linear-quadratic games.

2. Model of the game

Consider the game

$$\Gamma=\langle\{1, \dots, N\}, \Sigma, \{\mathcal{U}_1, \dots, \mathcal{U}_N\}, \{J_1, \dots, J_N\}\rangle . \tag{2.1}$$

Here $\{1, \dots, N\}$ is the set of the players participating in the game Γ. The evolution of the dynamic system Σ is described by a stochastic differential equation of the type

$$d\xi(t)=f(t, \xi, u_1, \dots, u_N)dt+\sigma(t, \xi, u_1, \dots, u_N)dw(t), t\in[t_0, T] \tag{2.2}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

with an initial condition $\xi(t_0)=\xi_0 \in \mathbf{R}^n$ where $T>t_0 \geqq 0$. The process $w=\{w(t), t \in [t_0, T]\}$ is a standard m-dimensional Wiener process defined on some complete probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and is adapted to a family $\mathbf{F}=\{\mathscr{F}_t, t \in [t_0, T]\}$ of nondecreasing sub-σ-algebras of $\mathscr{F}$. $\xi \in \mathbf{R}^n$ is the state vector process and $u_i \in U_i \subset \mathbf{R}^{v_i}$ is the control of the ith player, $i=1, \ldots, N$. Now let us make the following assumptions about the functions $f(t, x, u_1, \ldots, u_N)$ and $\sigma(t, x, u_1, \ldots, u_N)$. Suppose

$$f: [t_0, T] \times \mathbf{R}^n \times U_1 \times \ldots \times U_N \to \mathbf{R}^n$$

and

$$\sigma: [t_0, T] \times \mathbf{R}^n \times U_1 \times \ldots \times U_N \to \mathbf{R}^n \times \mathbf{R}^m$$

have continuous partial derivatives and let $C>0$ be a constant such that

$$|f(t, 0, \ldots, 0)| + |\sigma(t, 0, \ldots, 0)| \leqq C,$$

$$|f_x| + |\sigma_x| + \sum_{i=1}^{N} (|f_{u_i}| + |\sigma_{u_i}|) \leqq C.$$

Each player has perfect information about the state vector $\xi(t)$ at every moment $t \in [t_0, T]$ and constructs his strategy in game (2.1) as an admissible feedback control, i.e.

$$u_i(t)=u_i(t, \xi(t))$$

where

$$u_i(\cdot, \cdot): [t_0, T] \times \mathbf{R}^n \to U_i$$

is a Borel function satisfying the following conditions:

(i) There exists a constant $M_i>0$ such that

$$|u_i(t, x)| \leqq M_i(1+|x|) \quad \text{for all} \quad (t, x) \in [t_0, T] \times \mathbf{R}^n .$$

(ii) For each bounded set $B \subset \mathbf{R}^n$ and $T^* \in (t_0, T)$ there exists a constant $K_i>0$ such that for arbitrary $x, y \in B$ and $t \in [t_0, T^*]$

$$|u_i(t, x)-u_i(t, y)| \leqq K_i |x-y| .$$

Denote by $\mathscr{U}_i$ the set of strategies of the ith player, $i=1, \ldots, N$ and

$$\mathscr{U}=\prod_{i=1}^{N} \mathscr{U}_i, \quad U=\prod_{i=1}^{N} U_i .$$

Let a vector of strategies $u=(u_1, \ldots, u_N) \in \mathscr{U}$ be called for brevity simply a strategy.

The assumptions mentioned above imply the existence and sample path uniqueness of the solution $\xi=\{\xi(t), t \in [t_0, T]\}$ of (2.2) corresponding to the control $u \in \mathscr{U}$. The infinitesimal operator $\mathscr{A}(u)$ of the Markov process ξ has the form, see [3]:

$$\mathscr{A}(u) V(t, x)=f'(t, x, u) V_x(t, x)+\frac{1}{2} \operatorname{tr} [a(t, x, u) V_{xx}(t, x)]$$

where $a = \sigma\sigma'$ and prime denotes matrix transpose. Here $V(t, x)$ is a real-valued function with continuous partial derivatives up to second order for all $(t, x) \in [t_0, T] \times \mathbf{R}^n$.

Let L_i, Ψ_i be continuous functions satisfying the polynomial growth conditions:

$$| L_i(t, x, u_1, \ldots, u_N) | \leqq C_i \left(1 + |x| + \sum_{i=1}^{N} |u_i| \right)^k$$

$$| \Psi_i(t, x) | \leqq C_i (1 + |x|)^k$$

where C_i, k are positive constants. Introduce now the cost-function $J_i(u)$ of the ith player:

$$J_i(u) = \mathbf{E}_{t_0, \xi_0} \left\{ \Psi_i(T, \xi(T) + \int_{t_0}^{T} L_i(t, \xi, u_1, \ldots, u_N) dt \right\}, \qquad i = 1, \ldots, N.$$

Denote by $J(u)$ the vector cost-function of all players:

$$J(u) = (J_1(u), \ldots, J_N(u)).$$

The object of each player in game (2.1) is to minimize his own cost-function.

3. Main results

Definition. The strategy $u^P \in \mathscr{U}$ is called Pareto-optimal (efficient) for game (2.1) if the relations

$$J_i(u) \leqq J_i(u^P), \quad i = 1, \ldots, N$$

for some strategy $u \in \mathscr{U}$ imply the equalities

$$J_i(u) = J_i(u^P), \quad i = 1, \ldots, N.$$

Denote by $\mathscr{P}$ the set of all Pareto-optimal (efficient) strategies for game (2.1). The former definition shows the basic property of the vector cost-function $J(u^P), u^P \in \mathscr{U}$, whose essence is that the further decreasing of J_i is impossible without the increasing J_j for at least one $j \neq i$.

Lemma. Let the vector $\lambda = (\lambda_1, \ldots, \lambda_N) \in \mathbf{R}^N$ be such that

$$\lambda_i > 0, \quad i = 1, \ldots, N, \quad \lambda_1 + \ldots + \lambda_N = 1$$

and

$$\min_{u \in \mathscr{U}} \sum_{i=1}^{N} \lambda_i J_i(u) = \sum_{i=1}^{N} \lambda_i J_i(u^P). \tag{3.1}$$

Then u^P is a Pareto-optimal strategy for game (2.1).

Proof. Let condition (3.1) hold for a strategy u^P. Suppose u^P is not Pareto-optimal, i.e.

$$J_i(u) \leqq J_i(u^P), \quad i=1, \ldots, N$$

does not imply

$$J_i(u) = J_i(u^P), \quad i=1, \ldots, N .$$

Hence there exists a strategy u^0 and at least one i_0, $1 \leqq i_0 \leqq N$, such that

$$J_{i_0}(u^0) < J_{i_0}(u^P) .$$

Then

$$\sum_{i=1}^{N} \lambda_i J_i(u^0) < \sum_{i=1}^{N} \lambda_i J_i(u^P) = \min_{u \in \mathscr{U}} \sum_{i=1}^{N} \lambda_i J_i(u) \leqq \sum_{i=1}^{N} \lambda_i J_i(u^0) ,$$

which is wrong. Therefore, u^P is a Pareto-optimal strategy for game (2.1).

The above lemma makes possible the use of some techniques from stochastic optimal control [4; Ch. VI]. Indeed, the strategies providing the minimum of

$$I_\lambda(u) = \sum_{i=1}^{N} \lambda_i J_i(u) , \quad \lambda_i > 0 , \quad i=1, \ldots, N , \quad \lambda_1 + \ldots + \lambda_N = 1$$

are Pareto-optimal. This remark is the background for establishing sufficient conditions for Pareto-optimality of strategies in game (2.1). Denote

$$H_\lambda(t, x, u) = V_t(t, x) + \mathscr{A}(u) V(t, x) + \sum_{i=1}^{N} \lambda_i L_i(t, x, u)$$

for all $t \in [t_0 , T]$, $x \in \mathbf{R}^n$, $u \in U$. The vector $\lambda = (\lambda_1 , \ldots, \lambda_N) \in \mathbf{R}^N$ is given a priori. The constant λ_i characterizes the roie of the cost-function J_i of the ith player in game (2.1), $i=1, \ldots, N$.

Now we shall present a result analogous to that of Varaiya [9].

Theorem. The strategy u^P is Pareto-optimal (efficient) for game (2.1) if there exist a vector $\lambda = (\lambda_1 , \ldots, \lambda_N) \in \mathbf{R}^N$, $\lambda_i > 0$, $i=1, \ldots, N$, $\lambda_1 + \ldots + \lambda_N = 1$ and a real-valued function $V(t, x)$ such that for all $t \in [tt_0 , T]$, $x \in \mathbf{R}^n$ the following conditions jointly hold:

(a) V, V_t , V_x , V_{xx} are continuous;

(b) $H_\lambda(t, x, u^P) = 0$;

(c) $H_\lambda(t, x, u) \geqq 0$ for each strategy $u \in \mathscr{U}$;

(d) $V(T, x) = \sum_{i=1}^{N} \lambda_i \Psi_i(T, x)$.

Proof. Let $\xi(t)$ and $\xi^P(t)$, $t \in [t_0, T]$ be the sample paths of the solutions of (2.2) corresponding to the strategies u and u^P respectively. Next write Ito–Dynkin's formula with an arbitrary u and $\xi(t)$

$$V(t, x) = \mathbf{E}_{t,x}\left\{V(T, \xi(T)) - \int_t^T [V_\tau(\tau, \xi(\tau)) + \mathscr{A}(u)V(\tau, \xi(\tau))]\,d\tau\right\}.$$

Then, together with conditions (c) and (d) it impliees

$$V(t, x) \leqq \mathbf{E}_{t,x}\left\{\sum_{i=1}^N \lambda_i \Psi_i(T, \xi(T)) + \int_t^T \sum_{i=1}^N \lambda_i L_i(\tau, \xi, u)\,d\tau\right\}$$

and hence

$$V(t_0, \xi_0) \leqq \sum_{i=1}^N \lambda_i \mathbf{E}_{t_0, \xi_0}\left\{\Psi_i(T, \xi(T)) + \int_{t_0}^T L_i(t, \xi, u)\,dt\right\}.$$

Once more write Ito–Dynkin's formula, but with u^P and $\xi^P(t)$ respectively

$$V(t, x) = \mathbf{E}_{t,x}\left\{V(T, \xi^P(T)) - \int_t^T [V_\tau(\tau, \xi^P(\tau)) + \mathscr{A}(u^P)V(\tau, \xi^P(\tau))]\,d\tau\right\}.$$

Now taking into account conditions (b) and (d), we get

$$V(t, x) = \mathbf{E}_{t,x}\left\{\sum_{i=1}^N \lambda_i \Psi_i(T, \xi^P(T)) + \int_t^T \sum_{i=1}^N \lambda_i L_i(\tau, \xi^P, u^P)\,d\tau\right\},$$

which leads to

$$V(t_0, \xi_0) = \sum_{i=1}^N \lambda_i \mathbf{E}_{t_0, \xi_0}\left\{\Psi_i(T_0 \xi^P(T)) + \int_{t_0}^T L_i(t, \xi^P, u^P)\,dt\right\}.$$

Finally we obtain

$$V(t_0, \xi_0) = \sum_{i=1}^N \lambda_i J_i(u^P) \leqq \sum_{i=1}^N \lambda_i J_i(u) \quad \text{for each} \quad u \in \mathscr{U}.$$

This means that u^P is a Pareto-optimal strategy for game (2.1).

Note that $V(t, x)$ is a solution of a dynamic programming equation of the type

$$\min_u H_\lambda(t, x, u) = 0$$

with a boundary condition

$$V(T, x) = \sum_{i=1}^N \lambda_i \Psi_i(T, x)$$

for all $t \in [t_0, T]$, $x \in \mathbf{R}^n$.

4. Linear-quadratic games with state and control independent noise

Consider game (2.1) where the evolution of the dynamic system Σ is described by the linear stochastic differential equation

$$d\xi(t)=\left[A(t)\xi+\sum_{i=1}^{N}B_i(t)u_i\right]dt+\sigma(t)dw(t),\quad t\in[t_0,T]$$

with an innitial condition $\xi(t_0)=\xi_0$. Here ξ_0, w, u_i, $i=1,\dots,N$ are the same as in Section 2. $A(t)$ is an $n\times n$-matrix, $\sigma(t)$ is an $n\times m$-matrix and $B_i(t)$ are $n\times v_i$-matrices, $i=1,\dots,N$. The cost-function $J_i(u)$ of the ith player is the following quadratic functional:

$$J_i(u)=\mathbf{E}_{t_0,\xi_0}\left\{\xi'(T)D_i\xi(T)+\int_{t_0}^{T}\left[\xi'(t)M_i(t)\xi(t)+\sum_{i=1}^{N}u_j'(t)N_j^{(i)}(t)u_j(t)\right]dt\right\}.$$

Here D_i, $M_i(t)$ and $N_j^{(i)}(t), j=1,\dots,N$, are symmetric matrices with dimensions $n\times n$, $n\times n$ and $v_i\times v_i$ respectively, $i=1,\dots,N$.

Next consider the function

$$H_\lambda(t,x,u)=V_t(t,x)+\mathscr{A}(u)V(t,x)+\sum_{i=1}^{N}\lambda_iL_i(t,x,u)=$$

$$=V_t(t,x)+\left[A(t)x+\sum_{i=1}^{N}B_i(t)u_i\right]'V_x(t,x)+\frac{1}{2}\operatorname{tr}\left[a(t)V_{xx}(t,x)\right]+$$

$$+\sum_{i=1}^{N}\lambda_i\left[\sum_{j=1}^{N}u_j'N_j^{(i)}(t)u_j+x'M_i(t)x\right].$$

Conditions (b) and (c) imply the relation

$$\min_u H_\lambda(t,x,u)=H_\lambda(t,x,u^P)=0\,. \tag{4.1}$$

Hence the Pareto-optimal strategies u_i^P are solutions of the equations

$$\left.\frac{\partial H_\lambda}{\partial u_i}\right|_{u^P}=B_i'(t)V_x(t,x)+2\sum_{j=1}^{N}\lambda_jN_i^{(j)}(t)u_i^P=0\,,$$

i.e.

$$u_i^P=-\frac{1}{2}\left[\sum_{j=1}^{N}\lambda_jN_i^{(j)}(t)\right]^{-1}B_i'(t)V_x(t,x)\,,\quad i=1,\dots,N\,.$$

Further substitute u^P in (b) and search $V(t, x)$ in the following special form:

$$V(t, x) = x'K_\lambda(t)x + q_\lambda(t) \tag{4.2}$$

where $K_\lambda(t)$ is a symmetric $n \times n$-matrix with

$$K_\lambda(T) = \sum_{i=1}^{N} \lambda_i D_i$$

and $q_\lambda(t)$ is a scalar function. Evidently

$$V_x(t, x) = 2K_\lambda(t)x\,, \quad V_{xx}(t, x) = 2K_\lambda(t)$$

and then

$$u_i^P = -\left[\sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\right]^{-1} B_i'(t)K_\lambda(t)x\,, \quad i = 1, \ldots, N\,.$$

Thus, for equation (b) we obtain

$$0 = x' \frac{dK_\lambda(t)}{dt} x + \frac{dq_\lambda(t)}{dt} +$$

$$+\left\{x'A'(t) - \sum_{i=1}^{N} x'K_\lambda(t)B_i(t)\left[\sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\right]^{-1} B_i'(t)\right\} 2K_\lambda(t)x +$$

$$+ \operatorname{tr}\,[a(t)K_\lambda(t)] + \sum_{i=1}^{N} \lambda_i x'M_i(t)x +$$

$$+ \sum_{i=1}^{N} \lambda_i \left\{\sum_{j=1}^{N} x'K_\lambda(t)B_j(t)\left[\sum_{k=1}^{N} \lambda_k N_j^{(k)}(t)\right]^{-1} N_j^{(i)}(t)\left[\sum_{k=1}^{N} \lambda_k N_j^{(k)}(t)\right]^{-1} B_j'(t)K_\lambda(t)x\right\}$$

or

$$0 = x'\left\{\frac{dK_\lambda(t)}{dt} + A'(t)K_\lambda(t) + K_\lambda(t)A(t) + \sum_{i=1}^{N} \lambda_i M_i(t) -\right.$$

$$\left. - \sum_{i=1}^{N} K_\lambda(t)B_i(t)\left[\sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\right]^{-1} B_i'(t)K_\lambda(t)\right\} x + \frac{dq_\lambda(t)}{dt} + \operatorname{tr}\,[a(t)K_\lambda(t)]\,.$$

Hence for $K_\lambda(t)$ we get Riccati's matrix differential equation

$$\frac{dK_\lambda(t)}{dt} + A'(t)K_\lambda(t) + K_\lambda(t)A(t) + \sum_{i=1}^{N} \lambda_i M_i(t) -$$

$$- K_\lambda(t)\left\{\sum_{i=1}^{N} B_i(t)\left[\sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\right]^{-1} B_i'(t)\right\} K_\lambda(t) = 0$$

with the boundary condition

$$K_\lambda(T) = \sum_{i=1}^{N} \lambda_i D_i$$

and

$$q_\lambda(t) = \int_t^T \operatorname{tr}\,[a(\tau) K_\lambda(\tau)]\,d\tau\,.$$

Introduce the matrices

$$D_\lambda = \sum_{i=1}^{N} \lambda_i D_i\,, \quad M_\lambda(t) = \sum_{i=1}^{N} \lambda_i M_i(t)$$

and

$$N_\lambda^{(i)}(t) = \sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\,, \quad i = 1, \ldots, N\,.$$

Thus we obtain the following

Proposition 1. Let in the linear-quadratic game (2.1) the matrices $A(t)$, $B_i(t)$, $\sigma(t)$, $M_i(t)$, $N_j^{(i)}(t)$ be continuous and D_i be constant. Let the vector $\lambda = (\lambda_1, \ldots, N) \in \mathbf{R}^N$ be such that $\lambda_i > 0, i = 1, \ldots, N, \lambda_i + \ldots + \lambda_N = 1$, the matrices D_λ, $M_\lambda(t)$ are positive semi-definite and the matrices $N_\lambda^{(i)}(t)$ are positive definite. Then:

(i) $K_\lambda(t)$ is the solution of Riccati's matrix differential equation

$$\frac{dK_\lambda(t)}{dt} + A'(t)K_\lambda(t) + K_\lambda(t)A(t) + M_\lambda(t) - K_\lambda(t)\left\{\sum_{i=1}^{N} B_i(t)\,[N_\lambda^{(i)}(t)]^{-1} B_i'(t)\right\}K_\lambda(t) = 0$$

with the boundary condition

$$K_\lambda(T) = D_\lambda$$

and

$$q_\lambda(t) = \int_t^T \operatorname{tr}\,[a(\tau) K_\lambda(\tau)]\,d\tau\,;$$

(ii) $V(t, x)$ given by (4.2) is the solution of the dynamic programming equation (4.1) and

$$u_i^P = -[N_\lambda^{(i)}(t)]^{-1} B_i'(t) K_\lambda(t) x\,, \quad i = 1, \ldots, N$$

are Pareto-optimal strategies for the linear-quadratic game (2.1) considered in this section.

5. Linear-quadratic games with controlled drift and diffusion coefficients

Let the evolution of the dynamic system Σ of game (2.1) be described by the linear stochastic differential equation

$$d\xi(t)=\left[A(t)\xi+\sum_{i=1}^{N} B_i(t)u_i\right]dt+\sigma(t,\xi,u_1,\dots,u_N)dw(t), \quad t\in[t_0,T]$$

with initial condition $\xi(t_0)=\xi_0\in\mathbf{R}$. Here $\xi\in\mathbf{R}$ is the state process, $w=\{w(t), t\in[t_0,T]\}$ is an $(N+2)$-dimensional standard Wiener process and $u_i\in U_i\subset\mathbf{R}$ is the control of the ith player, $i=1,\dots,N$. $\sigma(t,\xi,u_1,\dots,u_N)$ is an $1\times(N+2)$-matrix of the form

$$\sigma=(\sigma_0(t)\xi\ \sigma_1(t)u_1\ \dots\ \sigma_N(t)u_N\ \sigma_{N+1}(t)).$$

Henceforth $A(t)$, $B_i(t)$, $i=1,\dots,N$, $\sigma_j(t)$, $j=0,1,\dots,N+1$ are functions taking values in $\mathbf{R}$. The cost-function $J_i(u)$ of the ith player is the functional

$$J_i(u)=\mathbf{E}_{t_0,\xi_0}\left\{D_i\xi^2(T)+\int_{t_0}^{T}\left[M_i(t)\xi^2(t)+\sum_{j=1}^{N}N_j^{(i)}(t)u_j^2(t)\right]dt\right\}, \quad i=1,\dots,N.$$

Here D_i are constants and $M_i(t)$, $i=1,\dots,N$, $N_j^{(i)}(t)$, $i,j=1,\dots,N$ are real-valuedd functions.

Next consider the function

$$H_\lambda(t,x,u)=V_t(t,x)+\mathscr{A}(u)V(t,x)+\sum_{i=1}^{N}\lambda_iL_i(t,x,u)=$$

$$=\frac{\partial V(t,x)}{\partial t}+\left[A(t)x+\sum_{i=1}^{N}B_i(t)u_i\right]\frac{\partial V(t,x)}{\partial x}+\frac{1}{2}\left[\sigma_0^2(t)x^2+\sum_{i=1}^{N}\sigma_i^2(t)u_i^2+\right.$$

$$\left.+\sigma_{N+1}^2(t)\right]\frac{\partial^2V(t,x)}{\partial x^2}+\sum_{i=1}^{N}\lambda_i\left[M_i(t)x^2+\sum_{j=1}^{N}N_j^{(i)}(t)u_j^2\right].$$

Conditions (b) and (c) are equivalent to the relation

$$\min_u H_\lambda(t,x,u)=H_\lambda(t,x,u^P)=0. \tag{5.1}$$

Hence the Pareto-optimal strategies u_i^P are solutions of the equations

$$\left.\frac{\partial H_\lambda}{\partial u_i}\right|_{u^P}=B_i(t)\frac{\partial V(t,x)}{\partial x}+\sigma_i^2(t)\frac{\partial^2V(t,x)}{\partial x^2}u_i^P+2\sum_{j=1}^{N}\lambda_jN_i^{(j)}(t)u_i^P=0,$$

i.e.

$$u_i^P = -\left[\sigma_i^2(t)\frac{\partial^2 V(t,x)}{\partial x^2} + 2\sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\right]^{-1} B_i(t)\frac{\partial V(t,x)}{\partial x}, \quad i=1,\ldots,N\,.$$

Next substitute u^P in (b) and search $V(t, x)$ in the following special form

$$V(t,x) = K_\lambda(t)x^2 + q_\lambda(t) \tag{5.2}$$

where $K_\lambda(t)$ and $q_\lambda(t)$ are real-valued functions. Evidently

$$\frac{\partial V(t,x)}{\partial x} = 2K_\lambda(t)x\,, \quad \frac{\partial^2 V(t,x)}{\partial x^2} = 2K_\lambda(t)$$

lead to

$$u_i^P = -\left[\sigma_i^2(t)K_\lambda(t) + \sum_{j=1}^{N} \lambda_j N_i^{(j)}(t)\right]^{-1} B_i(t)K_\lambda(t)x\,, \quad i=1,\ldots,N\,.$$

Then for equation (b) we have

$$\begin{aligned} 0 = {} & \frac{dK_\lambda(t)}{dt}x^2 + \frac{dq_\lambda(t)}{dt} + \Bigg\{A(t)x - \sum_{i=1}^{N}\Bigg[\sigma_i^2(t)K_\lambda(t) + \\ & + \sum_{j=1}^{N}\lambda_j N_i^{(j)}(t)\Bigg]^{-1} B_i^2(t)K_\lambda(t)x\Bigg\}2K_\lambda(t)x + \Bigg\{\sigma_0^2(t)x^2 + \sum_{i=1}^{N}\sigma_i^2(t)\Bigg[\sigma_i^2(t)K_\lambda(t) + \\ & + \sum_{j=1}^{N}\lambda_j N_i^{(j)}(t)\Bigg]^{-2} B_i^2(t)K_\lambda^2(t)x^2 + \sigma_{N+1}^2(t)\Bigg\}K_\lambda(t) + \sum_{i=1}^{N}\lambda_i M_i(t)x^2 + \\ & + \sum_{i=1}^{N}\lambda_j\sum_{j=1}^{N}N_j^{(i)}(t)\Bigg[\sigma_j^2(t)K_\lambda(t) + \sum_{K=1}^{N}\lambda_K N_j^{(K)}(t)\Bigg]^{-2} B_j^2(t)K_\lambda^2(t)x^2 \end{aligned}$$

or

$$\begin{aligned} 0 = x^2\Bigg\{ & \frac{dK_\lambda(t)}{dt} + 2A(t)K_\lambda(t) + \sum_{i=1}^{N}\lambda_i M_i(t) + \sigma_0^2(t)K_\lambda(t) - \\ & - \sum_{i=1}^{N}\Bigg[\sigma_i^2(t)K_\lambda(t) + \sum_{j=1}^{N}\lambda_j N_i^{(j)}(t)\Bigg]^{-1} B_i^2(t)K_\lambda^2(t)\Bigg\} + \frac{dq_\lambda(t)}{dt} + \sigma_{N+1}^2(t)K_\lambda(t)\,. \end{aligned}$$

Hence for $K_\lambda(t)$ we get the nonlinear differential equation

$$\begin{aligned} & \frac{dK_\lambda(t)}{dt} + 2A(t)K_\lambda(t) + \sum_{i=1}^{N}\lambda_i M_i(t) + \sigma_0^2(t)K_\lambda(t) - \\ & - K_\lambda^2(t)\sum_{i=1}^{N}\Bigg[\sigma_i^2(t)K_\lambda(t) + \sum_{j=1}^{N}\lambda_j N_i^{(j)}(t)\Bigg]^{-1} B_i^2(t) = 0 \end{aligned}$$

with the boundary condition

$$K_\lambda(T)=\sum_{i=1}^{N}\lambda_i D_i$$

and

$$q_\lambda(t)=\int_t^T \sigma_{N+1}^2(\tau)K_\lambda(\tau)d\tau\,.$$

Denote

$$D_\lambda=\sum_{i=1}^{N}\lambda_i D_i\,,\quad M_\lambda(t)=\sum_{i=1}^{N}\lambda_i M_i(t)\quad\text{and}\quad N_\lambda^{(i)}(t)=\sum_{j=1}^{N}\lambda_j N_i^{(j)}(t)\,,\quad i=1,\ldots,N\,.$$

Summarizing the above reasoning, we come to the following.

Proposition 2. Let in the linear-quadratic game (2.1) the functions $A(t)$, $B_i(t)$, $\sigma_j(t)$, $M_i(t)$, $N_j^{(i)}(t)$ be continuous. Let the vector $\lambda=(\lambda_1\,,\ldots,\lambda_N)\in\mathbf{R}^N$ be such that $\lambda_i>0$, $i=1,\ldots,N$, $\lambda_1+\ldots+\lambda_N=1$, D_λ is a non-negative constant, $M_\lambda(t)$ is a non-negative function and $N_\lambda^{(i)}(t)$ is a positive function, for each $t\in[t_0\,,T]$. Then:

(i) $K_\lambda(t)$ is the solution of the nonlinear differential equation

$$\frac{dK_\lambda(t)}{dt}+2A(t)K_\lambda(t)+M_\lambda(t)+\sigma_0^2(t)K_\lambda(t)-K_\lambda^2(t)\sum_{i=1}^{N}[\sigma_i^2(t)K_\lambda(t)+N_\lambda^{(i)}(t)]B_i^2\quad(t)=0$$

with the boundary condition

$$K_\lambda(T)=D_\lambda$$

and

$$q_\lambda(t)=\int_t^T \sigma_{N+1}^2(\tau t)K_\lambda(\tau)d\tau\,;$$

(ii) $V(t,x)$ given by (5.2) is the solution of the dynamic programming equation (5.1) and

$$u_i^P=-[\sigma_i^2(t)K_\lambda(t)+N_\lambda^{(i)}(t)]^{-1}B_i(t)K_\lambda(t)x\,,\quad i=1,\ldots,N$$

are Pareto-optimal strategies for the linear-quadratic game (2.1) considered in this section.

6. Concluding remarks

In this paper we study Pareto-optimality using the approach of Fleming and Rishel [4] to the stochastic optimal control. These results are a good starting point for the further consideration of other concepts to solve a stochastic differential game. As far as the approach is concerned we intend also to make use of the methods developed by Davis and Varaiya [2] and Krylov [6]. The results presented here have been announced without any proof in our paper [5].

Acknowledgement. The author expresses his gratitude to Dr. J. Stoyanov for the helpful discussions.

References

1. *Chernousko, F. L., Kolmanovskii, V. B.*, Optimal Stochastic Control. Nauka, Moscow 1978 (in Russian).
2. *Davis, M. H. A., VVaraiya, P. P.*, Dynamic programming conditions for partially observable stochastic systems. SIAM J. Control, 1973, **11**, pp. 226–261.
3. *Dynkin, E. B.*, Markov Processes. Fizmatgiz, Moscow 1963 (In Russian; Engl. transl.: Springer-Verlag, Berlin 1965).
4. *Fleming, W. H., Rishel, R. W.*, Deterministic and Stochastic Optimal Control. Springer-Verlag, Berlin–Heidelberg–New York 1975.
5. *Gaidov, S. D.*, Basic optimal strategies in stochastic differential games. Compt. Rend. Acad. Bulg. Sci., 1984, **37**, pp. 457–460.
6. *Krylov, N. V.*, Controlled Diffusion Processes. Nauka, Moscow 1977 (In Russian; Engl. transl.: Springer-Verlag, Berlin 1980).
7. *Roitenberg, Y. N.*, Automatic Control. Nauka, Moscow 1978 (In Russian; French transl.: Mir, Moscow 1974).
8. *Vaisbord, E. M., Zhukovskii, V. I.*, Introduction to Many-Player Differential Games and Their Applications. Sovetskoe Radio, Moscow 1980 (In Russian).
9. *Varaiya P. P.*, *N*-player stochastic differential games. SIAM J. Control and Optimiz., 1976, **14**, pp. 538–545.

Оптимальность по Парето в стохастических дифференциальных играх

С. Д. ГАЙДОВ

(Русе)

В работе рассматриваются игры с ненулевой суммой и многими игроками, когда динамика описывается стохастическими дифференциальными уравнениями Ито. Установлены достаточные условия для стратегий игроков быть оптимальными по Парето. Изучены также и линейные квадратические игры, когда коэффициент шума зависит или не зависит от управления.

S. D. Gaidov
Center of Mathematics
P.O.Box 325
BG-7000 Rousse
Bulgaria

Problems of Control and Information Theory, Vol. 15 (6), pp. 451—463 (1986)

STABILITY PROPERTIES OF THE VALUE FUNCTION OF A DIFFERENTIAL GAME AND VISCOSITY SOLUTIONS OF HAMILTON–JACOBI EQUATIONS

A. I. SUBBOTIN, A. M. TARASYEV

(Sverdlovsk)

(Received August 7, 1985)

Differential inequalities being used in the differential games theory for determination of stability properties of the value function in infinitesimal form [9–11] are examined in the present paper. Another group of inequalities being considered below is the group of inequalities, which has been introduced in [2, 3] for definition of viscosity solutions of Hamilton–Jacobi equations. The main result of the paper is the direct proof of the equivalence of the mentioned inequalities.

1. Introduction

As it is known in the differential games theory (see, [5, 6]) the value function satisfies the so-called stability properties, which are related to the optimality principle of dynamic programming. In [9–11] it has been demonstrated that the stability properties may be expressed by a pair of inequalities for directional derivatives. These inequalities together with the boundary conditions form a group of necessary and sufficient conditions which the value function must satisfy. In a domain, where the value function is a differentiable one, the mentioned inequalities turn into the partial differential equation, which is called Bellman–Isaacs equation (BIE). Therefore, one can say that the value function is a generalized solution of BIE. By this definition (with the help of differential inequalities) a generalized solution of BIE satisfying given boundary conditions exists and is unique.

Necessity of introduction of generalized solutions is caused by non-existence, as a rule, of classical global solutions of BIE. It should be noted that BIE is a first-order partial differential equation of Hamilton–Jacobi type (HJE). In recent papers [2, 3] the notion of viscosity solutions of HJE has been introduced and it has been proved that a viscosity solution exists and is unique. It is shown (see e.g. [1, 7]) that the viscosity solution of BIE coincides with the value function. In definition of viscosity solutions a partial differential equation is substituted by a pair of differential inequalities.

However, these inequalities differ formally from the inequalities which have been obtained in [9–11]. Below inequalities used in [2, 3, 9–11] for definition of generalized solutions of BIE and HJE are cited and the direct proof of the equivalence of these inequalities is given.

2. Statement of the problem

Let the dynamics of a control system be described by the equation

$$\dot{x}=f(t, x, u, v)\,, \tag{2.1}$$

where

$$t\in[0, \vartheta], \quad x\in R^n, \quad u\in P, \quad v\in Q,$$

P and Q are compact sets in R^p and R^q respectively. The function

$$f\colon [0, \vartheta]\times R^n\times P\times Q\mapsto R^n$$

is a continuous one and satisfies the Lipschitz condition

$$\sup_{(t,\, x^{(i)},\, u,\, v)} \|f(t, x^{(1)}, u, v)-f(t, x^{(2)}, u, v)\|\cdot\|x^{(1)}-x^{(2)}\|^{-1}<\infty\,,$$

$$(t, x^{(i)}, u, v)\in[0, \vartheta]\times G\times P\times Q(i=1, 2), \quad x^{(1)}\neq x^{(2)}\,,$$

where G is an arbitrary restricted domain in R^n. Here and below $\|x\|=\langle x, x\rangle^{1/2}$, the symbol $\langle\cdot,\cdot\rangle$ denotes the inner product.

It is also assumed that the motions of the system (2.1) are extendable to the moment $t=\vartheta$.

Let us consider the upper differential game. The first player (minimizing pay-off) selects positional strategies $U:[0, \vartheta]\times R^n\mapsto P$ and the second player (maximizing pay-off) selects counterstrategies

$$V\colon[0, \vartheta]\times R^n\times P\mapsto Q$$

in this game. Suppose that a pay-off functional has the following form $\gamma(x(\cdot))=\sigma(x(\vartheta))$, where the function $\sigma:R^n\mapsto R^1$ satisfies the Lipschitz condition.

The formalization of the described differential game is contained, for example, in [5, 6].

Let $\omega^0:[0, \vartheta]\times R^n\mapsto R^1$ be the value function, i.e. this function transform an initial position (t_0, x_0) into the value $\omega^0(t_0, x_0)$ of the considered differential game. It is known that $\omega^0\in\mathrm{Lip}$. Here the symbol Lip denotes the class of functions $\omega:[0, \vartheta]\times\times R^n\mapsto R^1$ satisfying the Lipschitz condition

$$\sup_{(t^{(i)},\, x^{(i)})} |\omega(t^{(1)}, x^{(1)})-\omega(t^{(2)}, x^{(2)})|\cdot(|t^{(1)}-t^{(2)}|+\|x^{(1)}-x^{(2)}\|)^{-1}<\infty\,,$$

$$(t^{(i)}, x^{(i)}) \in G (i=1,2), \quad |t^{(1)}-t^{(2)}| + \|x^{(1)}-x^{(2)}\| > 0 ,$$

where G is an arbitrary restricted domain in $[0, \vartheta] \times R^n$.

From the definition of the value function it follows immediately that

$$\omega^0(\vartheta, x) = \sigma(x) \, (x \in R^n) . \tag{2.2}$$

Since $\omega^0 \in$ Lip, then according to Rademacher's theorem [8] the value function is differentiable almost everywhere. It is known (see e.g. [4]) that at each position (t, x), where the value function is differentiable, it satisfies the following BIE

$$\frac{\partial \omega^0}{\partial t}(t, x) + H\left(t, x, \frac{\partial \omega^0}{\partial x}(t, x)\right) = 0 . \tag{2.3}$$

Here $\partial\omega/\partial x = (\partial\omega/\partial x_1, \ldots, \partial\omega/\partial x_n)$ is a column vector, the quantity $H(t, x, l)$ is called the Hamiltonian or the unificator of the considered differential game and defined by the equality

$$\begin{gathered} H(t, x, l) = \min_{u \in P} \max_{v \in Q} \langle l, f(t, x, u, v) \rangle \\ ((t, x, l) \in [0, \vartheta] \times R^n \times R^n) . \end{gathered} \tag{2.4}$$

Let us consider the terminal value problem for BIE (2.3) with boundary condition (2.2). As it is noted above this problem has a classical global solution in exceptional cases only. There are known various definitions of generalized solutions of terminal value problem (2.2), (2.3). In a number of papers the following definition of generalized solutions has been used. A function $\omega \in$ Lip is called a generalized solution of terminal value problem (2.2), (2.3) if it satisfies boundary condition (2.2) and satisfies BIE (2.3) almost everywhere. In particular the value function satisfies this definition. The trouble with the given definition is that the set of solutions may be infinite.

Below we consider four definitions of generalized solutions of terminal value problem (2.2), (2.3). For each of these definitions a generalized solution exists and is unique. Furthermore, it coincides with the value function.

3. Main definitions and results

Let $\omega \in$ Lip, $(t, x) \in [0, \vartheta) \times R^n$, $h \in R^n$, $l \in R^n$. Let us introduce the following notations

$$\begin{gathered} \partial_+ \omega(t, x)|(h) = \limsup_{\delta \downarrow 0} (\omega(t+\delta, x+\delta h) - \omega(t, x))\delta^{-1} \\ \partial_- \omega(t, x)|(h) = \liminf_{\delta \downarrow 0} (\omega(t+\delta, x+\delta h) - \omega(t, x))\delta^{-1} , \end{gathered} \tag{3.1}$$

$$D^*\omega(t,x)|(l) = \sup_{h\in R^n}(\langle l,h\rangle - \partial_-\omega(t,x)|(h))$$
$$D_*\omega(t,x)|(l) = \inf_{h\in R^n}(\langle l,h\rangle - \partial_+\omega(t,x)|(h))\,. \tag{3.2}$$

Equalities (3.1) define upper and lower derivatives of the function ω by the direction $(1, h)$ at the point (t, x).

The quantities D^* and D_*, defined according to (3.2), we shall call the conjugate derivatives of the function ω.

Let $u\in P$ and $v(\cdot)\in \mathbf{V}$, where $\mathbf{V}$ is the set of mappings from P to Q. Let us assume

$$F_1(t,x,u) = \mathrm{co}\,\{f(t,x,u,v): v\in Q\}\,,$$
$$F_2(t,x,v(\cdot)) = \mathrm{co}\,\mathrm{cl}\,\{f(t,x,u,v(u)): u\in P\}\,.$$

Here the symbols $\mathrm{co}\,\{F\}$ and $\mathrm{cl}\,\{F\}$ denote the convex and closed hulls, respectively, of the set F.

By the symbol $\mathbf{C}^1$ we denote the space of continuously differentiable functions $\varphi:(0,\vartheta)\times R^n \mapsto R^1$. For a position $(t,x)\in(0,\vartheta)\times R^n$ and a number $\varepsilon>0$ the symbol $N_\varepsilon(t,x)$ denotes the open ε-neighbourhood of (t,x), i.e.

$$N_\varepsilon(t,x) = \{(\tau,y)\in(0,\vartheta)\times R^n : (t-\tau)^2 + \|x-y\|^2 < \varepsilon^2\}\,.$$

Following to [2] let us introduce the notations

$$\Phi_{\min}(\omega) = \{(t,x,\varphi)\in(0,\vartheta)\times R^n\times \mathbf{C}^1 : \exists\,\varepsilon>0 \quad \text{such that}$$
$$\omega(\tau,y)-\varphi(\tau,y) \geqq \omega(t,x)-\varphi(t,x), \quad \forall\,(\tau,y)\in N_\varepsilon(t,x)\} \tag{3.3}$$
$$\Phi_{\max}(\omega) = \{(t,x,\varphi)\in(0,\vartheta)\times R^n\times \mathbf{C}^1 : \exists\,\varepsilon>0 \quad \text{such that}$$
$$\omega(\tau,y)-\varphi(\tau,y) \leqq \omega(t,x)-\varphi(t,x), \quad \forall\,(\tau,y)\in N_\varepsilon(t,x)\}\,.$$

Thus $(t,x,\varphi)\in\Phi_{\min}(\omega)$ if and only if the difference $\omega(\tau,y)-\varphi(\tau,y)$ attains a local minimum at the point (t,x).

According to [3] we shall introduce also the following definitions.

The super differential of $\omega\in\mathrm{Lip}$ at $(t,x)\in(0,\vartheta)\times R^n$, denoted by $D^+\omega(t,x)$, is the set defined by:

$$D^+\omega(t,x) = \{(\alpha,a)\in R^1\times R^n : \limsup_{(\tau,y)\to(t,x)} (\omega(\tau,y)-\omega(t,x)-$$
$$-(\alpha(\tau-t)+\langle a, y-x\rangle))/(|\tau-t|+\|y-x\|) \leqq 0\}\,.$$

The sub differential of $\omega\in\mathrm{Lip}$ at $(t,x)\in(0,\vartheta)\times R^n$, denoted by $D^-\omega(t,x)$, is the set given by:

$$D^-\omega(t,x)=\{(\alpha,a)\in R^1\times R^n:\ \liminf_{(\tau,y)\to(t,x)}\ (\omega(\tau,y)-\omega(t,x)-$$

$$-(\alpha(\tau-t)+\langle a,y-x\rangle))/(|\tau-t|+\|y-x\|)\geqq 0\}\,.$$

In order to define generalized solutions of the terminal value problem (2.2), (2.3), we replace BIE (2.3) by a pair of differential inequalities. Let us consider four pairs of inequalities of such kind

$$\max_{v(\cdot)\in \mathbf{V}}\ \min_{h\in F_2(t,x,v(\cdot))}\ \partial_-\omega(t,x)|(h)\leqq 0,\quad \forall\,(t,x)\in(0,\vartheta)\times R^n \tag{3.4a}$$

$$\min_{u\in P}\ \max_{h\in F_1(t,x,u)}\ \partial_+\omega(t,x)|(h)\geqq 0,\quad \forall\,(t,x)\in(0,\vartheta)\times R^n\,, \tag{3.5a}$$

$$D^*\omega(t,x)|(l)\geqq H(t,x,l),\quad \forall\,(t,x,l)\in(0,\vartheta)\times R^n\times R^n \tag{3.4b}$$

$$D_*\omega(t,x)|(l)\leqq H(t,x,l),\quad \forall\,(t,x,l)\in(0,\vartheta)\times R^n\times R^n\,, \tag{3.5b}$$

$$\frac{\partial\varphi}{\partial t}(t,x)+H\left(t,x,\frac{\partial\varphi}{\partial x}(t,x)\right)\leqq 0,\quad \forall\,(t,x,\varphi)\in\Phi_{\min}(\omega) \tag{3.4c}$$

$$\frac{\partial\varphi}{\partial t}(t,x)+H\left(t,x,\frac{\partial\varphi}{\partial x}(t,x)\right)\geqq 0,\quad \forall\,(t,x,\varphi)\in\Phi_{\max}(\omega)\,, \tag{3.5c}$$

$$\alpha+H(t,x,a)\leqq 0,\quad \forall(t,x)\in(0,\vartheta)\times R^n,\quad (\alpha,a)\in D^-\omega(t,x) \tag{3.4d}$$

$$\alpha+H(t,x,a)\geqq 0,\quad \forall\,(t,x)\in(0,\vartheta)\times R^n,\quad (\alpha,a)\in D^+\omega(t,x)\,. \tag{3.5d}$$

Definition *A* (*B*, *C* or *D*). A function $\omega\in\mathrm{Lip}$ satisfying equality (2.2) and inequalities (3.4a), (3.5a) ((3.4b), (3.5b); (3.4c), (3.5c) or (3.4d), (3.5d)) is called the generalized solution of the terminal value problem (2.2), (2.3).

Remark 3.1. Definitions *A*, *B*, *C*, *D* have been introduced in [9, 11, 2, 3] respectively.

Remark 3.2. It is not difficult to show that in a domain, where a function $\omega\in\mathrm{Lip}$ is a differentiable one, each of the mentioned above four pairs of inequalities turns into BIE (2.3).

It has been proved in [9–11] that a generalized solution in the sense of the definition *A* or *B* exists, is unique and coincides with the value function ω^0. Definitions *C*, *D* coincide in fact with the definitions of a viscosity solution (see [2, 3]). A distinction without a difference here from the definitions, which are contained in [2, 3], consists in the following. In the mentioned papers uniformly continuous functions have been considered, while here we operate with locally Lipschitz continuous functions. As it has been proved in [1–3, 7] a viscosity solution of the terminal value problem (2.2), (2.3) exists, is unique and coincides with the value function. Hence, definitions *A*, *B*, *C*, *D* are equivalent.

For the analysis of the structure of the value function, it is useful to make more precisely the noted fact about the equivalence of definitions A, B, C, D. This sharpening consists in the equivalence of inequalities (3.4a), (3.4b), (3.4c), (3.4d) and also in the equivalence of inequalities (3.5a), (3.5b), (3.5c), (3.5d). As it has been shown in [9–11] inequalities (3.4a) and (3.4b) (resp. (3.5a) and (3.5b)) express the so-called u-stability property (resp. v-stability property). Therefore inequalities (3.4a) and (3.4b) ((3.5a) and (3.5b)) are equivalent. It has also been shown in [3] that inequalities (3.4c) and (3.4d) ((3.5c) and (3.5d)) are equivalent.

The following statement is the main result of the present paper.

Theorem 3.1. Inequalities (3.4b) and (3.4c) ((3.5b) and (3.5c)) are equivalent.

4. Preliminary results

We shall prove now two auxiliary statements.

Lemma 4.1. Let $\psi \in \mathrm{Lip}$, $(t_0, x_0) \in [0, \vartheta) \times R^n$,

$$a \in (0, \vartheta - t_0], b > 0 \quad \text{and} \quad \psi(t, x) > \psi(t_0, x_0),$$

when

$$t_0 \leqq t \leqq t_0 + a, \quad \|x - x_0\| \leqq b, \quad t - t_0 + \|x - x_0\| > 0.$$

Then for any $\alpha > 0$ there exists $(t_\alpha, x_\alpha, \varphi_\alpha) \in \Phi_{\min}(\psi)$ such that

$$t_0 < t_\alpha < t_0 + \alpha, \quad \|x_\alpha - x_0\| \leqq \alpha, \tag{4.1}$$

$$\frac{\partial \varphi_\alpha}{\partial t}(t_\alpha, x_\alpha) \geqq -\alpha, \quad \frac{\partial \varphi_\alpha}{\partial x}(t_\alpha, x_\alpha) = 0. \tag{4.2}$$

Proof of Lemma 4.1. Without restriction on the generality of arguments we shall assume that

$$t_0 = 0, \quad x_0 = 0, \quad \psi(t_0, x_0) = 0. \tag{4.3}$$

Let us introduce the following notations

$$B = \{x \in R^n : \|x\| \leqq b\},$$

$$R(\varepsilon) = \{r \geqq 0 : \psi(t, x) \geqq rt, \quad \forall\, t \in [0, \varepsilon], x \in B\},$$

$$r(\varepsilon) = \sup R(\varepsilon), \quad r_0 = \lim_{\varepsilon \downarrow 0} r(\varepsilon).$$

It should be noted that $0 \in R(\varepsilon)$, when $\varepsilon \in (0, a]$. Therefore

$$r_0 \geqq 0. \tag{4.4}$$

From the definition of the number r_0 it follows that for any $r_* < r_0$ there exists $\varepsilon(r_*) \in (0, a]$ such that

$$\psi(t, x) > r_* t, \quad \text{when} \quad t \in (0, \varepsilon(r_*)], x \in B, \tag{4.5}$$

and for any $r^* > r_0$ there exists a sequence $(t_k, x_k) \in (0, a] \times B$ such that

$$\psi(t_k, x_k) < r^* t_k (k = 1, 2, \ldots), \quad t_k \to 0, \quad \text{when} \quad k \to \infty. \tag{4.6}$$

We may assume in estimate (4.5) that

$$\varepsilon(r) \to 0, \quad \text{when} \quad r \uparrow r_0. \tag{4.7}$$

Let $r < r_0$. Assume

$$\rho(t, r) = rt - (2t - \varepsilon(r))^2 (r_0 - r)/\varepsilon(r), \tag{4.8}$$

$$c(r) = \min_{(t, x)} (\psi(t, x) - \rho(t, r)), \quad \text{when} \quad t \in [0, \varepsilon(r)], x \in B, \tag{4.9}$$

$$M(r) = \{(\tau, y) \in [0, \varepsilon(r)] \times B : \psi(\tau, y) - \rho(\tau, r) = c(r)\}. \tag{4.10}$$

Let us prove that

$$0 < c(r) < (r_0 - r)\varepsilon(r). \tag{4.11}$$

From (4.5) and (4.8) it follows that

$$\psi(t, x) - \rho(t, r) > (2t - \varepsilon(r))^2 (r_0 - r)/\varepsilon(r) \geqq 0, \tag{4.12}$$

when $t \in (0, \varepsilon(r)]$, $x \in B$. Therefore $c(r) > 0$. Let $r^* = 2r_0 - r$. Since $r^* > r_0$, then from (4.6) it follows that there exists a point (t^*, x^*) satisfying the conditions

$$0 < t^* < \varepsilon(r)/2, x^* \in B, \quad \psi(t^*, x^*) - (2r_0 - r)t^* < 0. \tag{4.13}$$

From (4.8) and (4.9) we have

$$c(r) < \psi(t^*, x^*) - rt^* + 4t^*(r_0 - r)(t^*/\varepsilon(r) - 1) + (r_0 - r)\varepsilon(r).$$

From the estimate $t^* < \varepsilon(\tau)/2$ it follows that $t^*/\varepsilon(r)/2$ it follows that $t^*/\varepsilon(r) - 1 \leq$ $\leq -1/2$. Taking into account (4.13) we obtain $c(r) < (r_0 - r)\varepsilon(r)$. Thus estimates (4.11) are proved.

Let us show that

$$0 < \tau < \varepsilon(r) \quad \text{for all} \quad (\tau, y) \in M(r). \tag{4.14}$$

Really, according to (4.11) and (4.12) we have

$$\psi(0, \chi) - \rho(0, r) \geq (r_0 - r)\varepsilon(r) > c(r),$$

$$\psi(\varepsilon(r), x) - \rho(\varepsilon(r), r) \geq (r_0 - r)\varepsilon(r) > c(r).$$

Then (4.14) is valid.

Let us prove the following implication

$$((\tau_k, y_k) \in M(r_k) \quad (k=1, 2, \ldots), \quad r_k \uparrow r_0, k \to \infty) \Rightarrow$$
$$\Rightarrow (\|y_k\| \to 0, \quad k \to \infty). \tag{4.15}$$

Let us suppose the contrary. Then without restriction on the generality of arguments we may assume that there exists the limit $y_k \to y_*$, when $k \to \infty$, $\|y_*\| > 0$. Note that from (4.7) and (4.14) it follows that $\tau_k \to 0$, when $k \to \infty$. From the condition $(\tau_k, y_k) \in M(r_k)$ we have

$$\psi(\tau_k, y_k) - \rho(\tau_k, r_k) = c(r_k).$$

According to (4.8) and (4.14) we obtain $\rho(\tau_k, r_k) \to 0$, when $k \to \infty$. From (4.11) we have $c(r_k) \to 0$, when $k \to \infty$. From (4.11) we have $c(r_k) \to 0$, when $k \to \infty$. Therefore, taking into account (4.3) we conclude that

$$\psi(0, y_*) = \psi(t_0, x_0) = 0, \quad \|y_*\| > 0.$$

But these relations contradict conditions of Lemma 4.1. So, implication (4.15) is proved.

Let $0 < \alpha < \min\{a, b\}$. Choose a parameter r and a point $(t_\alpha, x_\alpha) \in M(r)$ so that the following conditions will be valid

$$r_0 - \alpha/5 < r < r_0, \quad 0 < t_\alpha \leq \alpha, \quad \|x_\alpha\| \leq \alpha. \tag{4.16}$$

As it is shown, the above selection of such a kind is possible. Assume

$$\varphi_\alpha(t, x) = \rho(t, r), \quad t \in [0, \vartheta], \quad x \in R^n. \tag{4.17}$$

Note that $\varphi_\alpha \in \mathbf{C}^1$, $(t_\alpha, x_\alpha) \in \operatorname{int}([0, \varepsilon(r)] \times B)$. Therefore, from (4.10) and the definition of sets $\Phi_{\min}$ it follows that $(t, x_\alpha, \varphi_\alpha) \in \Phi_{\min}(\psi)$. It is obvious that $\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial x = 0$. It remains to convince ourselves that $\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial t \geq -\alpha$. From (4.8) and (4.17) we have

$$\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial t = r - 4(2t_\alpha - \varepsilon(r))(r_0 - r)/\varepsilon(r).$$

According to (4.14) we obtain

$$\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial t \geq r - 4(r_0 - r) = 5r - 4r_0.$$

Finally taking into account (4.4) and (4.16) we conclude that

$$\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial t \geq 5r - 4r_0 \geq 5(r - r_0) \geq -\alpha.$$

Lemma 4.1 is proved.

Lemma 4.2. Let

$$\mu \in Lip, \ (t_0, x_0) \in (0, \vartheta) \times R^n$$

and

$$\partial_- \mu(t_0, x_0)|(h) \geq 0, \quad \forall h \in R^n. \tag{4.18}$$

Then for any $\beta > 0$ there exist $a \in (0, \vartheta - t_0]$ and $b > 0$ such that the function

$$\Psi(t, x) = \mu(t, x) + \beta((t - t_0)^2 + \|x - x_0\|^2)^{1/2}. \tag{4.19}$$

satisfies the conditions of Lemma 4.1.

Proof of Lemma 4.2. For

$$\mu \in Lip, \quad (t_0, x_0) \in (0,\vartheta) \times R^n, \quad \alpha \in R^1 \quad \text{and} \quad h \in R^n$$

we introduce the quantity

$$\partial_- \mu(t_0, x_0)|(\alpha, h) = \liminf_{\delta \downarrow 0} (\mu(t_0 + \alpha\delta, x_0 + \delta h) - \mu(t_0, x_0))\delta^{-1}.$$

From this definition,

$$\partial_- \mu(t_0, x_0)|(\lambda\alpha, \lambda h) = \lambda \partial_- \mu(t_0, x_0)|(\alpha, h)$$

for $\lambda \geq 0$. It is possible to check that the mapping $(\alpha, h) \mapsto \partial_- \mu(t_0, x_0)|(\alpha, h)$ satisfies the Lipschitz condition. Therefore from the conditions of Lemma 4.2 it follows that

$$\partial_- \mu(t_0, x_0)|(\alpha, h) \geq 0, \quad \forall \alpha \geq 0, \quad h \in R^n. \tag{4.20}$$

Let us suppose by contradiction that the function Ψ mentioned in Lemma 4.2 does not satisfy the requirements of Lemma 4.1. Then there exists a sequence (t_k, x_k) $(k = 1, 2, \ldots)$ such that

$$\begin{gathered} t_k \geq t_0, \quad (t_k, x_k) \to (t_0, x_0), \quad k \to \infty, \\ \Psi(t_k, x_k) \leq \Psi(t_0, x_0), \quad \delta_k = t_k - t_0 + \|x_k - x_0\| > 0. \end{gathered} \tag{4.21}$$

Assume $\alpha_k = (t_k - t_0)\delta_k^{-1}$. Note that $\alpha_k \in [0, 1]$. Without restriction on the generality of arguments we may assume that $\alpha_k \to \alpha_*$, $k \to \infty$. The following equalities are valid

$$t_k = t_0 + \alpha_k \delta_k, \quad x_k = x_0 + h_k \delta_k,$$

$$h_k = (x_k - x_0)(1 - \alpha_k)\|x_k - x_0\|^{-1}, \quad \text{when} \quad x_k \neq x_0,$$

$$h_k = 0, \quad \text{when} \quad x_k = x_0.$$

Here we may also assume that $h_k \to h_*$, $k \to \infty$. It should be noted that $\alpha_k + \|h_k\| = 1$, therefore

$$\alpha_* + \|h_*\| = 1. \tag{4.22}$$

So we have

$$t_k = t_0 + \alpha_* \delta_k + (\alpha_k - \alpha_*)\delta_k,$$

$$x_k = x_0 + h_* \delta_k + (h_k - h_*)\delta_k.$$

Since $\Psi \in \mathrm{Lip}$, then

$$\Psi(t_k, x_k) \geq \Psi(t_0 + \alpha_* \delta_k, x_0 + h_* \delta_k) - \lambda(|\alpha_k - \alpha_*| + \|h_k - h_*\|)\delta_k.$$

Hence

$$\partial_- \Psi(t_0, x_0)|(\alpha_*, h_*) \leq$$

$$\leq \liminf_{k \to \infty} (\Phi(t_0 + \alpha_* \delta_k, x_0 + h_* \delta_k) - \Psi(t_0, x_0))\delta_k^{-1} =$$

$$= \lim_{i \to \infty} (\Psi(t_0 + \alpha_* \delta_{k_i}, x_0 + h_* \delta_{k_i}) - \Psi(t_0, x_0))\delta_{k_i}^{-1} \leq$$

$$\leq \limsup_{i \to \infty} ((\Psi(t_{k_i}, x_{k_i}) - \Psi(t_0, x_0))\delta_{k_i}^{-1} +$$

$$+ \lambda(|\alpha_{k_i} - \alpha_*| + \|h_{k_i} - h_*\|)) \leq 0. \tag{4.23}$$

According to the definition of the function Ψ the following equality is valid

$$\partial_- \Psi(t_0, x_0)|(\alpha_*, h_*) = \partial_- \mu(t_0, x_0)|(\alpha_*, h_*) +$$

$$+ \beta(\alpha_*^2 + \|h_*\|^2)^{1/2}.$$

Therefore from (4.22) and (4.23) we obtain the inequality

$$\partial_- \mu(t_0, x_0)|(\alpha_*, h_*) < 0,$$

which contradicts (4.20). The obtained contradiction proves Lemma 4.2.

5. Proof of Theorem 3.1

Let us prove the implication (3.4b)⇒(3.4c). Let a function $\omega \in \mathrm{Lip}$ satisfy condition (3.4b) and $(t, x, \varphi) \in \Phi_{\min}(\omega)$. From (3.1), (3.3) it follows

$$-\frac{\partial \varphi}{\partial t}(t, x) \geq \left\langle \frac{\partial \varphi}{\partial x}(t, x), h \right\rangle - \partial_- \omega(t, x)|(h), \quad \forall h \in R^n.$$

According to (3.2) we have

$$-\frac{\partial \varphi}{\partial t}(t, x) \geq D^* \omega(t, x) \left| \left(\frac{\partial \varphi}{\partial x}(t, x) \right) \right..$$

Then we obtain (3.4c) from (3.4b).

The implication (3.5b)⇒(3.5c) is proved similarly.

We shall prove now the implication (3.4c)⇒(3.4b). Let a function $\omega \in \mathrm{Lip}$ satisfies condition (3.4c). Let us fix a point

$$(t_0, x_0, l_0) \in (0, \vartheta) \times R^n \times R^n.$$

If $D^*\omega(t_0, x_0)|(l_0) = +\infty$, then (3.4b) is valid. Let

$$D^*\omega(t_0, x_0)|(l_0) = c < +\infty. \tag{5.1}$$

According to (3.2) we have

$$\langle l_0, h\rangle - \partial_-\omega(t_0, x_0)|(h) \leq c, \quad \forall\, h \in R^n. \tag{5.2}$$

Let

$$\mu(t, x) = \omega(t, x) - \langle l_0, x\rangle + ct. \tag{5.3}$$

From (5.2) it follows that the function μ satisfies condition (4.18). According to Lemma 4.2, a function Ψ of the form (4.19) satisfies the conditions of Lemma 4.1. Let $(t_\alpha, x_\alpha, \varphi_\alpha)$ be an element of the set $\Phi_{\min}(\Psi)$, for which the estimates (4.1) and (4.2) are valid.

From (3.3), (4.19), (5.3) we have

$$\begin{aligned}\Psi(t, x) - \varphi_\alpha(t, x) = \omega(t, x) - v_*(t, x) \geq \\ \geq \Psi(t_\alpha, x_\alpha) - \varphi_\alpha(t_\alpha, x_\alpha) = \omega(t_\alpha, x_\alpha) - v_*(t_\alpha, x_\alpha),\end{aligned} \tag{5.4}$$

when

$$(t, x) \in N_\varepsilon(t_\alpha, x_\alpha).$$

Here

$$v_*(t, x) = \langle l_0, x\rangle - ct - \beta((t - t_0)^2 + \|x - x_0\|^2)^{1/2} + \varphi_\alpha(t, x). \tag{5.5}$$

It should be noted that the function v_* is non-differentiable at the point (t_0, x_0). We may assume that

$$(t_0 - t_\alpha)^2 + \|x_0 - x_\alpha\|^2 > \varepsilon^2.$$

Then we may define a function

$$v \in \mathbf{C}^1, \quad v(t, x) = v_*(t, x),$$

when

$$(t, x) \in N_\varepsilon(t_\alpha, x_\alpha).$$

From (3.3), (5.4) it follows that

$$(t_\alpha, x_\alpha, v) \in \Phi_{\min}(\omega).$$

According to (3.4c) we have

$$-\partial v(t_\alpha, x_\alpha)/\partial t \geq H(t_\alpha, x_\alpha, \partial v(t_\alpha, x_\alpha)/\partial x).$$

From (4.2) and (5.5) we obtain

$$-\partial v(t_\alpha, x_\alpha)/\partial t < c + \alpha + \beta,$$

$$\partial v(t_\alpha, x_\alpha)/\partial x = l_0 - s,$$

where

$$s = \beta(x - x_0)/((t - t_0)^2 + \|x - x_0\|^2)^{1/2}, \quad \|s\| \leq \beta.$$

Therefore $c \geq H(t_\alpha, x_\alpha, l_0 - s) - \alpha - \beta$. The function $(t, x, l) \mapsto H(t, x, l)$ is a continuous one, we may choose the parameters α and β as small as we want, a point $((t_\alpha, x_\alpha)$ satisfies condition (4.1). Hence $c \geq H(t_0, x_0, l_0)$. Taking into account notation (5.1) we obtain (3.4b).

The implication (3.4c)⇒(3.4*b*) is proved.

The implication (3.5c)⇒(3.5b) is proved analogously.

Theorem 3.1 is proved.

6. Concluding remark

Theorem 3.1 is valid for any continuous Hamiltonian $H(t, x, l)$. In the given proof we have not used the fact that the Hamiltonian is defined by equality (2.4).

References

1. *Barron, E. N., Evans, L C., Jensen, R.*, Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls, J. Different. Equat., 1984, **53,** *2*, pp. 213–233.
2. *Crandall, M. G., Lions, P. I.*, Viscosity solutions of Hamilton—Jacobi equations, Trans. Amer. Math. Soc., 1983, **277,** *1*, pp. 1–42.
3. *Crandall, M. G., Lions, P. L*, Some properties of viscosity solutions of Hamilton–Jacobi equations, Trans. Amer. Math. Soc., 1984, **282,** *2*, pp. 487–502.
4. *Isaacs, R.*, Differential Games, Wiley, New York, 1965.
5. *Krasovskii, N. N.*, An approach-evasion differential game, I, II, Izv. Akad. Nauk SSSR, Tekhn. Kibernet., 1973, *2*, pp. 3–18, *3*, pp. 22–42 (in Russian).
6. *Krasovskii, N. N., Subbotin, A. I.*, Positional Differential Games, Nauka, Moscow, 1974 (in Russian).
7. *Lions, P. L, Souganidis, P. E.*, Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations, SIAM J. Control Optimiz., 1985, **23,** *4*, pp. 566–583.
8. *Rademacher, H.*, Über Partielle und Totale Differenzierbarkeit von Funktionen Mehrer Variablen unter die Transformation der Doppelintegrable, Math. Ann., 1918, **79,** pp. 340–354.

9. *Subbotin, A. I.*, A generalization of the basic equation of the theory of differential games, Dokl. Akad. Nauk SSSR, 1980, **254,** *2*, pp. 293–297 (in Russian).
10. *Subbotin, A. I.*, Generalization of the main equation of differential game theory, J. Optimiz. Theory and Appl., 1984, **43,** *1*, pp. 103–133.
11. *Subbotin, A. I., Tarasyev, A. M.*, Conjugate derivatives of the value function of a differential game, Dokl. Akad. Nauk SSSR, 1985, **283,** *3*, pp. 559–564 (in Russian).

Свойства стабильности функции цены дифференциальной игры и вязкие решения уравнений Гамильтона–Якоби

А. И. СУББОТИН, А. М. ТАРАСЬЕВ

(Свердловск)

В работе доказана эквивалентность дифференциальных неравенств, которые используются в теории дифференциальных игр для определения свойства стабильности функции цены и в теории уравнений в частных производных для определения вязких решений уравнений Гамильтона–Якоби. Тем самым показано, что любое из рассматриваемых неравенств можно использовать для определения обобщенного решения уравнения Гамильтона–Якоби. При таком определении обобщенное решение существует и единственно. Более того, для уравнения Беллмана–Айзекса это решение совпадает с функцией цены дифференциальной игры.

А. И. Субботин, А. М. Тарасьев
Институт математики и механики УНЦ АН СССР
СССР, 620066, г. Свердловск
ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 15 (6), pp. 465—474 (1986)

ON MULTISTAGE QUEUING SYSTEM WITH LOSSES AND ARRIVING FLOW OF PHASE TYPE

P. P. BOCHAROV, S. S. SPESIVOV

(*Moscow*)

(Received July 8, 1985)

An $(n+1)$-stage queuing system is considered. Zero stage is a single server queue with buffer of finite capacity. All the other stages are single server queues without buffers. The interarrival times to zero stage and customers' service times have probability distribution function of phase type. There is an additional Poisson flow to server i, $i=\overline{1,n}$. The service times of customers of main and additional flows at server i are assumed to be exponentially distributed.

If a customer arrives at a time when the buffer of zero stage is full or the server i, $i=\overline{1,n}$, is busy, it will be lost.

Recursive formulas for calculating output rate, probability of losses, probability of queue idleness and some other performance characteristics are given.

1. Introduction

An $(n+1)$-stage queuing system is considered. Stage 0 is a single server queue with buffer of finite capacity r, $0\leqq r<\infty$. Customers arrive to this stage according to a renewal process with probability distribution function (P.D.F.) $A(t)$. The service times of customers are i.i.d. random variables with P.D.F. $B(t)$. $A(t)$ and $B(t)$ are assumed to be distributions of phase type (PH-distributions)

$$A(t)=1-\vec{\alpha}^T e^{\Lambda t}\vec{1},\quad t\geqq 0,\quad \vec{\alpha}^T\vec{1}=1\,,$$

$$B(t)=1-\vec{\beta}^T e^{Mt}\vec{1},\quad t\geqq 0,\quad \vec{\beta}^T\vec{1}=1$$

with PH-representations $(\vec{\alpha}, \Lambda)$ and $(\vec{\beta}, M)$ of order 1 and m respectively. $\vec{1}$ is a column vector with all elements equal to one.

Upon service completion at the stage i a customer proceeds to stage $i+1$ with probability q_i, $i=\overline{0,n-1}$, and with additional probability it returns to stage i, $i=\overline{0,n}$. With probability q_n a customer departs from the queue. The case when $q_0=1$, $0<q_i\leqq 1$, $i=\overline{1,n}$, is considered.

There is an additional Poisson flow of rate γ_i to the server i, $i=\overline{1,n}$. The service times of customers of main and additional flows at server i, $i=\overline{1,n}$, are independent and exponentially distributed with rate η_i.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

If a customer arrives at a time when the buffer of zero stage is full or the server i, $i=\overline{1,n}$, is busy, it will be lost. In Kendall's notation such a system may be classified as $PH/PH/1/r\to(M/M/1/0)^n$ (Fig. 1).

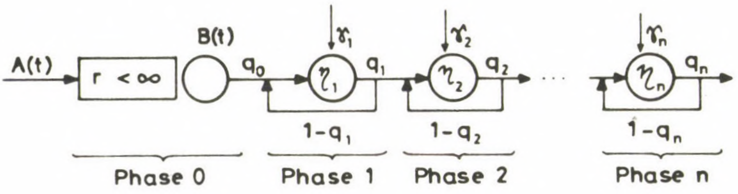


Fig. 1

The closed-form formula of probability of losses for the queue similar to the above-mentioned under the conditions that the flow arriving to stage 0 is stationary and ordinary, $r=0$ and there are no additional flows to stage i, $i=\overline{1,n}$, was reported in [3]. In [4] the results of paper [3] are extended for the case of nonhomogeneous groups of servers. In [5] for the case when interarrival times have Coxian distribution the efficient recursive procedure for calculation of probability of losses and some other queue characteristics which gives stable results for arbitrary values of parameters η_i was obtained.

In this paper which generalizes the results of [5] the recursive algorithm is obtained for calculating queue idleness and some other performance characteristics.

2. Main relations

Using probabilistic interpretation of PH-distribution the studied queue may be considered as a homogeneous Markov process $X(t)$, $t\geqq 0$, on the state space

$$\mathscr{X}=\bigcup_{k=0}^{r+1} x_k\,,$$

where

$$\mathscr{X}_0=\{(i,0,i_1,\ldots,i_n)/i=\overline{1,l},\, i_s=0,1,\, s=\overline{1,n}\}\,,$$

$$\mathscr{X}_0=\{(i,k,j,i_1,\ldots,i_n)/i=\overline{1,l},\, j=\overline{1,m},\, i_s=0,1,\, s=\overline{1,n}\},\ k=\overline{1,r+1}\,.$$

Here for some moment of time t:

1) $X(t)=(i,0,i_1,\ldots,i_n)$, if stage 0 is empty, the customer being generated is at the phase i [2], component i_s, $s=\overline{1,n}$, corresponds to the state of server i: $i_s=0$—the server is idle, $i_s=1$—the server is busy.

2) $X(t)=(i, k, j, i_1, \ldots, i_n)$, if the generated customer is at phase i, there are k customers at stage 0, the customer being served at zero stage is at phase j; component i_s, as previously, corresponds to the state of server i.

We shall use the notation i^l for the series $i, \ldots, i$ of length l, $l=\overline{1, n}$. If $l=0$ we agree on $(i_1, \ldots, i_j, i_{j_1}^0, i_{j+1}, \ldots, i_n)=(i_1, \ldots, i_j, i_{j+1}, \ldots, i_n)$. We shall also denote by dot the parameter with respect to all possible values of which the summing is made.

The irreducibility of the PH-representations $(\vec{\alpha}, \Lambda)$ and $(\vec{\beta}, M)$ [1, 2] entails that the process $X(t)$ is irreducible and hence, there exists stationary distribution of process $X(t)$

$$p(x)=\lim_{t\to\infty} P\{X(t)=x\}, \quad x\in\mathscr{X}.$$

For further account we introduce the following notations:

$$\nu_i=\eta_i q_i,\ i=\overline{1, n}; \quad g_i=\sum_{i=1}^{j}\gamma_i, \quad j=\overline{1, n};$$

$$\vec{\lambda}=-\Lambda\vec{1}; \quad \vec{\mu}=-M\vec{1}.$$

We also agree to denote by $A\mathbf{v}B$ the tensor product of matrices A and B [6], by I the identity matrix of appropriate order. Besides that, define the following vectors:

$$\vec{x}^T=(i_1, \ldots, i_n), \quad \vec{\gamma}^T=(\gamma_1, \ldots, \gamma_n), \quad \vec{\nu}^T=(\nu_1', \ldots, \nu_n),$$

$$\vec{p}_{0,n}^T(\vec{x})=(p(1, 0, \vec{x}), \ldots, p(l, 0, \vec{x})),$$

$$\vec{p}_{k,n}(\vec{x})=(p(1, k, 1, \vec{x}), \ldots, p(1, k, m, \vec{x}), \ldots, p(l, k, m, \vec{x})),$$

$$\vec{e}_j=(0, \ldots, 0, 1, 0, \ldots, 0),$$

where all n elements are zero exept the j-the element being one.

The stationary distribution $p(x)$, $x\in\mathscr{X}$, satisfies the following system of steady state equilibrium equations:

$$\vec{p}_{0,n}^T(\vec{x})\,[\Lambda-(g_n+\vec{x}^T(\vec{\nu}-\vec{\gamma}))I]+$$
$$+[\vec{p}_{1,n}^T(\vec{x})+\vec{p}_{1,n}^T(\vec{x}-\vec{e}_1)]\,(I\otimes\vec{\mu})u(i_1)+\vec{C}_{0,n}^T(\vec{x})=\vec{0}^T, \tag{1}$$
$$n=1, 2, \ldots$$

$$\vec{p}_{0,n}^T(\vec{x})\,(\vec{\lambda}\vec{\alpha}^T\otimes\vec{\beta}^T)+\vec{p}_{1,n}^T(\vec{x})\,[\Lambda\otimes I+I\otimes M-(g_n+\vec{x}^T(\vec{\nu}-\vec{\gamma}))I]+$$
$$+[\vec{p}_{2,n}^T(\vec{x})+\vec{p}_{2,n}^T(\vec{x}-\vec{e})]\,(I\otimes\vec{\mu}\vec{\beta}^T)u(i_1)+\vec{C}_{1,n}^T(\vec{x})=\vec{0}^T, \tag{2}$$
$$n=1, 2, \ldots$$

$$\vec{p}_{k-1,n}^T(\vec{x})\,(\vec{\lambda}\vec{\alpha}^T\otimes I)+\vec{p}_{k,n}^T(\vec{x})\,[\Lambda\otimes I+I\otimes M-(g_n+\vec{x}^T(\vec{\nu}-\vec{\gamma}))I]+$$
$$+[\vec{p}_{k+1,n}^T(\vec{x})+\vec{p}_{k+1,n}^T(\vec{x}-\vec{e}_1)]\,(I\otimes\vec{\mu}\vec{\beta}^T)u(i_1)+\vec{C}_{k,n}^T(\vec{x})=\vec{0}^T, \tag{3}$$
$$k=\overline{2, r}, \quad n=1, 2, \ldots$$

$$\vec{p}^{\,T}_{r,n}(\vec{x})\,(\vec{\lambda}\vec{\alpha}^T \otimes I) + \vec{p}^{\,T}_{r+1,n}(\vec{x})\,[(\Lambda + \vec{\lambda}\vec{\alpha}^T) \otimes I + I \otimes M - \tag{4}$$

$$-(g_n + \vec{x}^T(\vec{v} - \vec{\gamma}))I] + \vec{C}^{\,T}_{r+1,n}(\vec{x}) = \vec{0}^T\,, \quad n = 1, 2, \ldots \tag{5}$$

with the normalizing condition

$$\sum_{k=0}^{r+1} \vec{p}^{\,T}_{k,n}(\cdot^n)\vec{1} = 1\,.$$

Here

$$u(x) = \begin{cases} 1 & \text{for} \quad x > 0, \\ 0 & \text{for} \quad x \leqq 0\,. \end{cases}$$

$$\vec{C}^{\,T}_{k,n}(\vec{x}) = \sum_{j=1}^{n-1} (1 - i_j) i_{j+1} v_j \vec{p}^{\,T}_{k,n-1}(\vec{x} + \vec{e}_j) +$$

$$+ v_n(1 - i_n)\vec{p}^{\,T}_{k,n}(\vec{x} + \vec{e}_n) + \sum_{j=1}^{n} i_j \gamma_j \vec{p}^{\,T}_{k,n}(\vec{x} - \vec{e}_j)\,,$$

$$k = \overline{0, r+1}, \quad n = 1, 2, \ldots$$

Equations (1)–(4) together with the normalizing condition (5) define stationary distribution uniquely. Here the order of a system of equations to be solved is equal to

$$|\mathscr{X}| = [l + lm(r+1)]\,2^n\,.$$

Obviously, the order of the system becomes greater while the parameter n which defines the number of stages increases. It may be very big even for small values of buffer capacity r. Therefore, the solution of such a system of equations even with the help of computer is rather difficult. A procedure which enables us to calculate some characteristics of a multistage queue recursively with respect to the parameter n not solving the steady state equilibrium system of equations was reported in [5]. Analysis of equations (1)–(4) which was done by using the above-mentioned procedure enabled to establish a recursive dependence similar to [5] with respect to n of some characteristics of the considered queue. By this procedure the maximum order of the system of equations to be solved is lm.

Let us consider the relations for probabilities $\vec{p}_{k,n}(i_1, \ldots, i_{n-1}, \cdot)$ which are obtained from equations (1)–(4) by summing with respect to all possible values of i_n. Comparing these relations and the steady state system of equilibrium equations for n-stage system we get

$$\vec{p}_{k,n}(i_1, \ldots, i_{n-1}, \cdot) = \vec{p}_{k,n-1}(i_1, \ldots, i_{n-1})\,. \tag{6}$$

From this it follows that $\vec{p}_k = \vec{p}_{k,n}(\cdot^n)$ define the stationary distribution of stage 0 which is PH/PH/1/r queue.

For short account we define the matrices:

$$H_n = \Lambda - (v_n + g_n)I\,,$$

$$Q_n = (\vec{\lambda}\vec{\alpha}^T + \Lambda)\otimes I + I\otimes M - (v_n + g_n)I\,,$$

$$R_{n,i} = \vec{\lambda}\vec{\alpha}^T + \Lambda - (v_i + v_n + g_n - g_i)I\,,$$

$$W_n = \vec{\lambda}\vec{\alpha}^T + \Lambda - (v_n + g_n)I\,,$$

$$Y_i = u(2-i)\,(I\otimes\vec{\mu}) + u(i-1)\,v_{i-1}I\,,$$

$$K_n = u(2-n)\,(I\otimes\vec{\mu}\vec{\beta}^T) + u(i-1)\,v_{n-1}I\,.$$

From expression (1) for fixed $n=1, 2, \ldots, i_1 = \ldots = i_{n-1} = 0$ and $i_n = 0, 1$ using (6) we get the following recursive relations for quantities $\vec{p}_{0,n}(0^{n-1}, i_n)$. Let for convenience

$$\vec{S}_{n,i} = \vec{p}_{0,n}(0^{n-1}, i_n).$$

Then

$$\begin{gathered} \vec{S}^T_{n,0} H_n + v_n \vec{S}^T_{n-1,0} = \vec{0}^T\,, \\ \vec{S}^T_{n,1} H_n + \vec{S}^T_{n-1,1}\,Y_n + \gamma_n \vec{S}^T_{n-1,0} = \vec{0}^T\,, \end{gathered} \tag{7}$$

where $\vec{S}_{0,0} = \vec{p}_{0,n}(\cdot^n)$, $\vec{S}_{0,1} = \vec{p}_{1,n}(\cdot^n)$.

Now, fix $n=1, 2, \ldots$ and assume that $i_1 = \ldots = i_{n-1} = 0$, $i_n = 0, 1$, in equations (2)–(4). Summing these equations with respect to k from 1 to $r+1$ and using (6) we get the following recursive expression for quantities

$$\vec{Z}_{n,0,j} = \sum_{k=1}^{r+1} \vec{p}_{k,n}(0^{n-1}, j):$$

$$\begin{gathered} \vec{Z}^T_{n,0,0} Q_n + \vec{S}^T_{n,0}(\vec{\lambda}\vec{\alpha}^T\otimes\vec{\beta}^T) + v_n \vec{Z}^T_{n-1,0,0} = \vec{0}^T\,, \\ \vec{Z}^T_{n,0,1} Q_n + \vec{S}^T_{n,1}(\vec{\lambda}\vec{\alpha}^T\otimes\vec{\beta}^T) + \vec{Z}^T_{n-1,0,1} K_n + \gamma_n \vec{Z}_{n-1,0,0} = \vec{0}^T\,, \end{gathered} \tag{8}$$

where

$$\vec{Z}_{0,0,0} = \sum_{k=1}^{r+1} \vec{p}_{k,n}(\cdot^n),\ \vec{Z}_{0,0,1} = \sum_{k=2}^{r+1} \vec{p}_{k,n}(\cdot^n)\,.$$

Now, consider the quantities

$$\vec{Z}_{n,i,j} = \sum_{k=1}^{r+1} \vec{p}_{k,n}(\cdot^{i-1}, 1, 0^{n-i-1}, j)\,(I\otimes\vec{1}) + \vec{p}_{0,n}(\cdot^{i-1}, 1, 0^{n-i-1}, j)\,,$$

$$i = \overline{1, n-1}, \quad j = 0, 1, \quad n = 2, 3, \ldots\,,$$

for which recursive relations with respect to n may be obtained. For this in expressions (2)–(4) sum those equations in which the element i_i of vector $\vec{x}$ is equal to 1, the following $n-i-1$ elements are zeros and $i=0, 1$. Multiplying both sides of the result on the right by the matrix $(I \otimes 1)$ and summing up with those equations from expression (1) for which vector $\vec{x}$ has the above described elements we get

$$\begin{gathered} Z^T_{n,i,0} R_{n,i} + \vec{Z}^T_{n,i-1,0} Y_i + u(n-2) v_n \vec{Z}^T_{n-1,i,0} + \gamma_i \vec{V}^T_{n,i,0} = \vec{0}^T , \\ \vec{Z}^T_{n,i,1} R_{n,i} + \vec{Z}^T_{n,i-1,1} Y_i + u(n-i-1) v_{n-1} \vec{Z}^T_{n-1,i,1} + \\ + \gamma_i \vec{V}^T_{n,i,1} + \gamma_n \vec{Z}^T_{n-1,i,0} = \vec{0}^T . \end{gathered} \tag{9}$$

Here

$$\vec{V}_{n,i,j} = \sum_{s=0}^{i-1} \vec{Z}_{n,s,j} + \vec{S}_{n,j} .$$

Similarly, from expressions (2)–(4) the relations are obtained for the quantities

$$\begin{gathered} \vec{Z}_{n,n,0} = \sum_{k=1}^{r+1} \vec{p}_{k,n}(\cdot^{n-1}, 1)(I \otimes 1) + \vec{p}_{0,n}(\cdot^{n-1}, 1), \quad n = 1, 2, \ldots : \\ \vec{Z}^T_{n,n,0} W_n + (\vec{Z}^T_{n-1,n-1,0} + \vec{Z}^T_{n,n-1,1}) Y_n + \vec{1}^T \gamma_n = \vec{0}^T . \end{gathered} \tag{10}$$

For the convenience of further account we shall say that a matrix is stable if all its eigenvalues h_i satisfy $\mathrm{Re}\, h_i < 0$ and substable if $\mathrm{Re}\, h_i \leqq 0$.

Theorem. For fixed $n = 1, 2, \ldots$ the quantities $\vec{S}_{n,i}$, $i = 0, 1$, and $\vec{Z}_{n,i,j}$, $i = \overline{0, n}$, $j = 0, 1$ are defined uniquely by relations (7)–(10).

For proof it is sufficient to show the invertibility of matrices H_n, Q_n, $R_{n,i}$, W_n from relations (7)–(10). Let us prove that matrix Q_n is invertible. Note that the substable matrix $\Lambda + \vec{\lambda}\vec{\alpha}^T$ is the generator of a homogeneous Markov process. From the properties of tensor product of matrices [6] it follows that the eigenvalues of matrix $(\Lambda + \vec{\lambda}\vec{\alpha}^T) \otimes I + I \otimes M$ are the sum of respective eigenvalues of matrix $\Lambda + \vec{\lambda}\vec{\alpha}^T$ and stable matrix M [1]. Therefore, this matrix is stable. From this fact and positivity of the sum $v_n + g_n$ follows that matrix Q_n is stable and hence, invertible. The invertibility of other matrices is proved in a similar way.

3. Calculation of steady state characteristics

The above obtained relations enable to define such queue characteristics as probability of queue idleness, mean number of busy servers, output rate and probability of losses. Let us give the algorithm of their calculation.

First of all, using a procedure developed in [2, 7] the stationary distribution of stage 0 is calculated. From this distribution the values of $\vec{S}_{0,0}$, $\vec{S}_{0,1}$, $\vec{Z}_{0,0,0}$, $\vec{Z}_{0,0,1}$ are defined.

Let us denote by $p_n(0)$ the probability that the $(n+1)$-stage queuing system is idle. From (7) it follows that if $\vec{S}_{0,0}$ is known the quantities $\vec{S}_{k,i}$, $k=\overline{1,n}$, $i=0$, 1, and consequently the probability of queue idleness can be recursively calculated:

$$p_n(0)=\vec{S}^T_{n,0}\vec{1}\,. \tag{11}$$

We note that if the probability of zero stage idleness is known the closed-form formula for queue idleness is readily obtained:

$$p_n(0)=(-1)^n(\vec{S}^T_{0,0}\prod_{i=1}^{n} v_i H_i)\vec{1}\,. \tag{12}$$

Further, for given values $\vec{Z}_{0,0,j}$, $j=0$, 1, and previously calculated $\vec{S}_{k,i}$, $k=\overline{1,n}$, $i=0$, 1, from relations (8) the values of $\vec{Z}_{k,0,j}$, $k=\overline{1,n}$, $j=0$, 1, are determined recursively. By means of them in its turn the values of $\vec{Z}_{n,i,j}$, $i=\overline{1,n}$, $j=0$, 1, are defined from expressions (9)–(10).

Let E_n be the mean number of busy servers in an $(n+1)$-stage queuing system. Also, let $\lambda_{D,n}$ be the output rate and π_n the probability of losses of a customer having arrived at the queue.

Clearly, the following equality holds

$$E_n=\sum_{j=0}^{n}\vec{Z}^T_{j,\cdot,1}\vec{1},\quad n=1,2,\ldots \tag{13}$$

From ergodicity of the process $X(t)$ it follows that

$$\lambda_{D,n}=\mu_n\vec{Z}^T_{n,\cdot,1}\vec{1},\quad n=1,2,\ldots \tag{14}$$

Note that the rate λ of main flow to stage 0 is [2]

$$\lambda=(-\vec{\alpha}^T\Lambda^{-1}\vec{1})^{-1}\,.$$

Then rate λ_n of total flow to the queue is $\lambda_n=\lambda+g_n$. Hence, for steady state the probability of losses is defined by

$$\pi_n=1-\frac{\lambda_{D,n}}{\lambda_n},\quad n=1,2,\ldots$$

4. Examples

To illustrate the above obtained algorithm, we consider some examples of PH-representations for P.D.F. $A(t)$ and $B(t)$ which define stage 0.

1°. Queue $M/M/1/r\to(M/M/1/0)^n$. In this case $\vec{\alpha}=(1)$, $\Lambda=(-\lambda)$ and $\vec{\beta}=(1)$, $M=(-\mu)$. The matrices from relations (7)–(10) are given by

$$H_n = -(\lambda + v_n + g_n),\ Q_n = -(\mu + v_n + g_n),$$

$$R_{n,i} = -(v_i + v_n + g_n - g_i),\ W_n = -(v_n + g_n)\,,$$

$$Y_i = u(2-i)\mu + u(i-1)v_{i-1}\,,\ K_n = Y_n\,.$$

For this queuing system under the assumptions that $\gamma_1 = \ldots = \gamma_n = 0$ and $r=0$ the results were reported in [3].

2°. Queue $C_1/HM_m/1/r \to (M/M/1/0)^n$. Here, C_1 denotes Coxian distribution. The Laplace–Stieltjes transform (L.S.T.) of P.D.F. of interarrival times to stage 0 is the following:

$$\alpha(s) = \sum_{i=1}^{l} \alpha_1 \ldots \alpha_{i-1}\beta_i \prod_{j=1}^{i} \frac{\lambda_j}{\lambda_j + s}$$

where

$$0 \leqq \alpha_i \leqq 1, \quad i = \overline{1, l-1}, \quad \alpha_0 = 1, \quad \alpha_k = 0, \quad \beta_k = 1, \quad 0 < \lambda_i < \infty, \quad i = \overline{1, l}\,.$$

P.D.F. of service times are hyperexponential with L.S.T.

$$\beta(s) = \sum_{j=1}^{m} \frac{\beta_j \mu_j}{\mu_j + s}$$

where $0 \leqq \beta_j \leqq 1$, $0 < \mu_j < \infty$, $j = \overline{1, m}$. Then PH-representations $(\vec{\alpha}, \Lambda)$ and $(\vec{\beta}, M)$ are given by

$$\Lambda = \begin{pmatrix} -\lambda_1 & \alpha_1\lambda_1 & & 0 \\ & -\lambda_2 & \alpha_2\lambda_2 & \\ & & \vdots & \\ 0 & & \alpha_{l-1}\lambda_{l-1} & \\ & & & -\lambda_l \end{pmatrix} \qquad \vec{\alpha}^T = (1.0, \ldots 0)\,, \tag{16}$$

$$M = \operatorname{diag}(-\mu_1, \ldots, -\mu_m), \quad \vec{\beta}^T = (\beta_1, \ldots, \beta_m)\,. \tag{17}$$

In the case when $\gamma_1 = \ldots = \gamma_n = 0$, $r=0$ and $m=1$ we get an algorithm derived in [5].

3°. For the queuing system $HM_l/C_m/1/r \to (M/M/1/0)^n$ the PH-representations $(\vec{\alpha}, \Lambda)$ and $(\vec{\beta}, M)$ within renotations are defined by (16)–(17). Similarly, the queuing systems $C_l/C_m/1/r \to (M/M/1/0)^n$ and $HM_l/HM_m/1/r \to (M/M/1/0)^n$ may be considered.

5. Conclusion

The accuracy of the developed algorithm depends on the number of stages but does not depend on values of parameters η_i while calculations by closed-form formulae of paper [3] give uncorrect results for near values of η_i.

We note that the considered queuing system being of selfimportance is a part of arbitrary configuration queuing networks of finite capacity with internal losses. That is why the results obtained for initial multistage queue may be used to validate the approximate procedures for analyzing queuing networks which are intensively derived lately [8, 9].

Note also that the obtained results may be generalized for the case when parameters of PH-distribution which define stage 0 depend on the number of customers in this stage. In this case for calculation of stationary distribution of phase 0 the matrix algorithm developed in [7] is used. The description of zero stage by a PH-distribution whose parameters depend on queue length enables us to consider the question of queue control and optimization [10, 11].

References

1. *Neuts, M. F.*, Matrix-geometric solutions in stochastic models. The Johns Hopkins Univ. Press, Baltimore and London, 1981.
2. *Bocharov, P. P., Naumov, V. A.*, Matrix-geometric stationary distribution for the $PH/PH/1/r$ queue. Rapports de Recherche, *304*, INRIA, France, 1984.
3. *Bromberg, M. A., Kokotushkin, V. A., Naumov, V. A.*, Obslugivanie posledovatelnoi tsepochkoi priborov. Avtomatika i telemekhanika, *3*, 1977, pp. 60–64.
4. *Bromberg, M. A.*, Mnogophaznie sistemi s poteryami pri exponentsialnom oblugivanii. Avtomatika i telemekhanika, *10*, 1979, pp. 27–31.
5. *Bocharov, P. P.*, O mnogophaznoi sisteme s poteryami. Avtomatika i telemekhanika, *10*, 1984, pp. 40–65.
6. *Voevodin, V. V., Kuznetsov, U. A.*, Matritsi i vichisleniya. M. Nauka, 1984.
7. *Bocharov, P. P.* O sisteme massovogo obslugivaniya ogranichennoi emkosti s raspredeleniyami phazovogo tipa, zavisyaschimi ot sostoyaniya ocheredi. Avtomatika i telemekhanika, *10*, 1985.
8. *Bocharov, P. P.*, Pribligenii metod rascheta razomknutikh neexponentsialnikh setei ogranichennoi emkosti s poteryami. Tez. dokl. IX Vsesouznoi shkoli-seminara po vichislitelnim setyam, chast 1.1, pp. 88–94.
9. *Bronshtein, O., Gertsbakh, I.*, An open exponential queuing network with limited waiting spaces and losses: a method of approximate analysis. Perf. Eval., **4**, *1*, 1984, *pp.* 31–43.
10. *Nazarov, A. A.*, Upravlenie sistemami obslugivaniya i ikh optimizatsiya. Izd. Tomskogo universiteta, 1984.
11. *Laugen, H. J.*, Applying the method of phases in the optimization of queuing systems. Adv. Appl. Prob., **14**, 1982, pp. 122–142.

О многофазной системе с потерями с входящим потоком фазового типа

П. П. БОЧАРОВ, С. С. СПЕСИВОВ

(Москва)

Рассматривается $(n+1)$-фазная система массового обслуживания (СМО). Нулевая фаза представляет собой однолинейную СМО с накопителем ограниченной емкости, а остальные фазы — однолинейные системы без мест для ожидания. Длительности интервалов между заявками, поступающими на нулевую фазу, и длительности их облуживания имеют функции распределения фазового типа. На прибор i, $i=\overline{1,n}$, поступает дополнительный пуассоновский поток. Длительности обслуживания заявок как основного, так и дополнительного потоков на приборе i имеют экспоненциальный закон распределения.

Заявка, заставшая накопитель нулевой фазы заполненным, либо прибор фазы i, $i=\overline{1,n}$, занятым, теряется.

Выводятся рекуррентные формулы, позволяющие вычислить интенсивность выходящего из системы потока, вероятность потери заявки, поступившей в систему, вероятность простоя и некоторые другие характеристики качества функционирования системы.

Представляя самостоятельный интерес, рассматриваемая система является также фрагментом общего случая сети массового обслуживания конечной емкости с внутренними потерями. Поэтому результаты, полученные для исходной многофазной системы, можно использовать для тестирования приближенных алгоритмов расчета сетей, разработка которых интенсивно развивается в последнее время.

П. П. Бочаров, С. С. Спесивов
Университет дружбы народов им. П. Лумумбы
СССР, 117923 Москва, ГСП,
ул. Орджоникидзе, д. 3

РУССКИЙ ПЕРЕВОД

Проблемы Управления и Теории Информации, том **15,** *номер 6 (1986)*

СВОЙСТВА СТАБИЛЬНОСТИ ФУНКЦИИ ЦЕНЫ ДИФФЕРЕНЦИАЛЬНОЙ ИГРЫ И ВЯЗКИЕ РЕШЕНИЯ УРАВНЕНИЙ ГАМИЛЬТОНА—ЯКОБИ

А. И. Субботин, А. М. Тарасьев

(*Свердловск*)

В данной работе исследуются дифференциальные неравенства, которые используются в теории дифференциальных игр для определения в инфинитезимальной форме свойства стабильности функции цены [4, 5, 11]. Другую группу неравенств, которые рассматриваются ниже, составляют неравенства, введенные в работах [7, 8] для определения вязких решений уравнения Гамильтона–Якоби. Основным результатом статьи является прямое доказательство эквивалентности упомянутых неравенств.

1. Введение

В теории дифференциальных игр известно (см. [2, 3]), что функция цены удовлетворяет так называемым свойствам стабильности, которые можно трактовать как принцип оптимальности динамического программирования. В работах [4, 5, 11] было показано, что свойства стабильности можно выразить парой неравенств для производных по направлениям. Эти неравенства, дополненные краевым условием, образуют группу необходимых и достаточных условий, которым должна удовлетворять функция цены. В области, где функция цены дифференцируема, упомянутые неравенства обращаются в уравнение в частных производных первого порядка, которое называется уравнением Беллмана–Айзекса (УБА). Поэтому можно сказать, что функция цены является обобщенным решением УБА. При таком определении (с помощью дифференциальных неравенств) существует и единственно обобщенное решение УБА, удовлетворяющее заданному краевому условию.

Необходимость введения обощенных решений вызвана тем, что УБА, как правило, не имеет классических глобальных решений. Отметим, что УБА являются уравнениями в частных производных первого порядка и относятся к известным уравнениям типа Гамильтона–Якоби (УГЯ). В недавних работах [7, 8] было введено понятие вязкого решения УГЯ и доказано, что решение существует и единственно. В работах [6, 9] показано, что вязкое решение УГЯ совпадает с функцией цены. При определении вязкого решения уравнение в частных производных заменяется парой дифференциальных неравенств. Однако формально эти неравенства отличаются от неравенств из работ [4, 5, 11]. Ниже приведены неравенства, используемые в работах [4, 5, 7, 8, 11] при определении обобщенных решений УГЯ, и дано прямое доказательство эквивалентности этих неравенств.

2. Постановка задачи

Пусть движение управляемой системы описывается уравнением

$$\dot{x} = f(t, x, u, v)\,, \tag{2.1}$$

где $t \in [0, \vartheta]$, $x \in R^n$, $u \in P$, $v \in Q$, P и Q-компакты в R^p и R^q соответственно. Функция

$$f \colon [0, \vartheta] \times R^n \times P \times Q \mapsto R^n$$

непрерывна и удовлетворяет условию Липшица

$$\sup_{(t,\, x^{(i)},\, u,\, v)} \|f(t, x^{(1)}, u, v) - f(t, x^{(2)}, u, v)\| \cdot \|x^{(1)} - x^{(2)}\|^{-1} < \infty\,,$$

$$(t, x^{(i)}, u, v) \in [0, \vartheta] \times G \times P \times Q (i = 1, 2)\,, \quad x^{(1)} \neq x^{(2)}\,,$$

где G — произвольная ограниченная область в R^n. Здесь и ниже $\|x\| = \langle x, x \rangle^{1/2}$, символом $\langle \cdot , \cdot \rangle$ обозначено скалярное произведение.

Предполагается также, что движения системы (2.1) продолжимы до момента $t = \vartheta$.

Рассмотрим верхнюю дифференциальную игру. В этой игре первый игрок (минимизирующий плату) выбирает позиционные стратегии $U : [0, \vartheta] \times R^n \mapsto P$, а второй игрок (максимизирующий плату) выбирает контрстратегии

$$V \colon [0, \vartheta] \times R^n \times P \mapsto Q\,.$$

Полагаем, что функционал платы имеет вид $\gamma(x(\cdot))) = \sigma(x(\vartheta))$, где функция $\sigma : R^n \mapsto R^1$ удовлетворяет условию Липшица.

Формализация рассматриваемой дифференциальной игры приведена, например, в [2, 3].

Пусть $\omega^0 : [0, \vartheta] \times R^n \mapsto R^1$ — функция цены, т.е. эта функция начальной позиции $(t_0, x_0) \in [0, \vartheta] \times R^n$ ставит в соответствие цену $\omega^0(t_0, x_0)$ рассматриваемой дифференциальной игры. Известно, что $\omega^0 \in \mathrm{Lip}$. Символом

Lip здесь обозначен класс функций $\omega:[0,\ \vartheta]\times R^n\mapsto R^1$, удовлетворяющих условию Липшица

$$\sup_{(t^{(i)},\, x^{(i)})} |\omega(t^{(1)}, x^{(1)})-\omega(t^{(2)}, x^{(2)})|\cdot(|t^{(1)}-t^{(2)}|+\|x^{(1)}-x^{(2)}\|)^{-1}<\infty\,,$$

$$(t^{(i)}, x^{(i)})\in G(i=1, 2), |t^{(1)}-t^{(2)}|+\|x^{(1)}-x^{(2)}\|>0\,,$$

где G — произвольная ограниченная область в $[0,\ \vartheta]\times R^n$.

Из определения функции цены сразу следует, что

$$\omega^0(\vartheta, x)=\sigma(x)\,(x\in R^n)\,. \tag{2.2}$$

Поскольку $\omega^0\in$ Lip, то согласно теореме Радемахера [10] функция цены почти всюду дифференцируема. Известно (см., например, [1]), что в каждой точке (t, x), где функция цены дифференцируема, она удовлетворяет следующему УБА

$$\frac{\partial\omega^0}{\partial t}(t, x)+H\left(t, x, \frac{\partial\omega^0}{\partial x}(t, x)\right)=0\,. \tag{2.3}$$

Здесь $\partial\omega/\partial x=(\partial\omega/\partial x_1,\ \ldots,\ \partial\omega/\partial x_n)$ — вектор-столбец, величина $H(t, x, l)$ называется гамильтонианом или унификатором рассматриваемой дифференциальной игры и определяется равенством

$$\begin{gathered} H(t, x, l)=\min_{u\in P}\max_{v\in Q}\langle l, f(t, x, u, v)\rangle \\ ((t, x, l)\in[0,\ \vartheta]\times R^n\times R^n)\,. \end{gathered} \tag{2.4}$$

Рассмотрим краевую задачу (2.2), (2.3). Как отмечалось выше, эта задача имеет классическое глобальное решение лишь в отдельных исключительных случаях. Известны различные определения обобщенных решений краевой задачи (2.2), (2.3). В ряде работ приведено следующее определение. Обобщенным решением краевой задачи (2.2), (2.3) называется функция $\omega\in$ Lip, которая удовлетворяет граничному условию (2.2) и почти всюду удовлетворяет УБА (2.3). Этому определению удовлетворяет, в частности, функция цены. Недостатком данного определения является то, что множество решений может оказаться бесконечным.

Ниже приведены четыре определения обобщенного решения краевой задачи (2.2), (2.3). Для каждого из этих определений обощенное решение существует и единственно. Более того, оно совпадает с функцией цены.

3. Основные определения и результаты

Пусть $\omega \in \mathrm{Lip}$, $(t, x) \in [0, \vartheta) \times R^n$, $h \in R^n$, $l \in R^n$. Введем следующие обозначения

$$\partial_+ \omega(t, x)|(h) = \limsup_{\delta \downarrow 0} (\omega(t+\delta, x+\delta h) - \omega(t, x))\delta^{-1}$$

$$\partial_- \omega(t, x)|(h) = \liminf_{\delta \downarrow 0} (\omega(t+\delta, x+\delta h) - \omega(t, x))\delta^{-1}, \tag{3.1}$$

$$D^* \omega(t, x)|(l) = \sup_{h \in R^n} (\langle l, h\rangle - \partial_- \omega(t, x)|(h))$$

$$D_* \omega(t, x)|(l) = \inf_{h \in R^n} (\langle l, h\rangle - \partial_+ \omega(t, x)|(h)). \tag{3.2}$$

Равенства (3.1) определяют верхнюю и нижнюю производные функции ω по направлению $(1, h)$ в точке (t, x).

Величины D^* и D_*, определенные согласно (3.2), будем называть сопряженными производными функции ω.

Пусть $u \in P$ и $v(\cdot) \in \mathbf{V}$, где $\mathbf{V}$ — множество отображений из P в Q. Полагаем

$$F_1(t, x, u) = \mathrm{co}\, \{f(t, x, u, v) : v \in Q\},$$

$$F_2(t, x, v(\cdot)) = \mathrm{co}\, \mathrm{cl}\, \{f(t, x, u, v(u)) : u \in P\}.$$

Здесь $\mathrm{co}\,\{F\}$ и $\mathrm{cl}\,\{F\}$ — выпуклая и замкнутая оболочки множества F.

Символом $\mathbf{C}^1$ обозначим пространство непрерывно дифференцируемых функций $\varphi : (0, \vartheta) \times R^n \mapsto R^1$. Для точки $(t, x) \in (0, \vartheta) \times R^n$ и числа $\varepsilon > 0$ символ $N_\varepsilon(t, x)$ обозначает открытую ε-окрестность точки (t, x), т.е.

$$N_\varepsilon(t, x) = \{(\tau, y) \in (0, \vartheta) \times R^n : (t-\tau)^2 + \|x-y\|^2 < \varepsilon^2\}.$$

Следуя [7], введем следующие обозначения

$$\Phi_{\min}(\omega) = \{(t, x, \varphi) \in (0, \vartheta) \times R^n \times \mathbf{C}^1 : \exists\, \varepsilon > 0 \quad \text{такое, что}$$

$$\omega(\tau, y) - \varphi(\tau, y) \geqq \omega(t, x) - \varphi(t, x),\ \forall\, (\tau, y) \in N_\varepsilon(t, x)\}. \tag{3.3}$$

$$\Phi_{\max}(\omega) = \{(t, x, \varphi) \in (0, \vartheta) \times R^n \times \mathbf{C}^1 : \exists\, \varepsilon > 0 \quad \text{такое, что}$$

$$\omega(\tau, y) - \varphi(\tau, y) \leqq \omega(t, x) - \varphi(t, x), \quad \forall\, (\tau, y) \in N_\varepsilon(t, x)\}.$$

Таким образом, $(t, x, \varphi) \in \Phi_{\min}(\omega)$ тогда и только тогда, когда разность $\omega(\tau, y) - \varphi(\tau, y)$ достигает локального минимума в точке (t, x).

Согласно [8] введем также следующие определения.

Множество, определяемое соотношением

$$D^+\omega(t,x)=\{(\alpha,a)\in R^1\times R^n:\ \limsup_{(\tau,y)\to(t,x)}\ (\omega(\tau,y)-\omega(t,x)-$$

$$-(\alpha(\tau-t)+\langle a,y-x\rangle))/(|\tau-t|+\|y-x\|)\leqq 0\},$$

называется супер дифференциалом функции $\omega\in\mathrm{Lip}$ в точке (t,x).

Множество, определяемое соотношением

$$D^-\omega(t,x)=\{(\alpha,a)\in R^1\times R^n:\ \liminf_{(\tau,y)\to(t,x)}\ (\omega(\tau,y)-\omega(t,x)-$$

$$-(\alpha(\tau-t)+\langle a,y-x\rangle)/(|\tau-t|+\|y-x\|)\leqq 0\},$$

называется суб дифференциалом функции $\omega\in\mathrm{Lip}$ в точке $(t,\ x)$.

Чтобы определить обобщенное решение, заменим УБА парой диффренциальных неравенств. Рассмотрим четыре пары таких неравенств

$$\max_{v(\cdot)\in V}\ \min_{h\in F_2(t,x,v(\cdot))}\ \partial_-\omega(t,x)|(h)\leqq 0,\quad \forall\,(t,x)\in(0,\vartheta)\times R^n \tag{3.4a}$$

$$\min_{u\in P}\ \max_{h\in F_1(t,x,u)}\ \partial_+\omega(t,x)|(h)\geqq 0,\quad \forall\,(t,x)\in(0,\vartheta)\times R^n, \tag{3.5a}$$

$$D^*\omega(t,x)|(l)\geqq H(t,x,l),\quad \forall\,(t,x,l)\in(0,\vartheta)\times R^n\times R^n \tag{3.4в}$$

$$D_*\omega(t,x)|(l)\leqq H(t,x,l),\quad \forall\,(t,x,l)\in(0,\vartheta)\times R^n\times R^n, \tag{3.5в}$$

$$\frac{\partial\varphi}{\partial t}(t,x)+H\left(t,x,\frac{\partial\varphi}{\partial x}(t,x)\right)\leqq 0,\quad \forall\,(t,x,\varphi)\in\Phi_{\min}(\omega) \tag{3.4c}$$

$$\frac{\partial\varphi}{\partial t}\,t,x)+H\left(t,x,\frac{\partial\varphi}{\partial x}(t,x)\right)\geqq 0,\quad \forall\,(t,x,\varphi)\in\Phi_{\max}(\omega), \tag{3.5c}$$

$$\alpha+H(t,x,a)\leqq 0,\quad \forall\,(t,x)\in(0,\vartheta)\times R^n,(\alpha,a)\in D^-\omega(t,x) \tag{3.4d}$$

$$\alpha+H(t,x,a)\geqq 0,\ \forall\,(t,x)\in(0,\vartheta)\times R^n,\ (\alpha,a)\in D^+\omega(t,x). \tag{3.5d}$$

Определение $A(B,C$ или $D)$. Обобщенным решением задачи (2.2), (2.3) называется функция $\omega\in\mathrm{Lip}$, удовлетворяющая равенству (2.2) и неравенствам (3.4a), (3.5a) ((3.4b), (3.4c), (3.5b); (3.5c) или (3.4d), (3.5d)).

Замечание 3.1. Определения A, B, C, D были предложены в работах [4, 5, 7, 8] соответственно.

Замечание 3.2. Нетрудно показать, что там, где функция ω дифференцируема, любая из указанных выше четырех пар неравенств обращается в УБА (2.3).

В работах [4, 5, 11] показано, что обобщенное решение в смысле определения A или B существует, единственно и совпадает с функцией цены ω^0. Определения C, D фактически совпадают с определением вязкого решения (см. [7, 8]). Несущественное отличие от формулировки из работ [7, 8] состоит в том, что в упомянутых работах рассматривались равномерно непрерывные функции, а здесь обобщенное решение содержится в классе Lip. Как установлено в [6–9], вязкое решение краевой задачи (2.2), (2.3) существует, единственно и совпадает с функцией цены. Следовательно, определения A, B, C, D эквивалентны.

Для изучения структуры функции цены полезно уточнить отмеченный факт эквивалентности определений A, B, C, D. Это уточнение состоит в том, что эквивалентны неравенства (3.4a), (3.4b), (3.4c), (3.4d), а также эквивалентны неравенства (3.5a), (3.5b), (3.5c), (3.5d). В работах [4, 5] показано, что неравенства (3.4a) и (3.4b) (соответственно (3.5a) и (3.5b)) выражают так называемое свойство u-стабильности функции ω (соответственно свойство v-стабильности). Поэтому неравенства (3.4a) и (3.4b) ((3.5a) и (3.5b)) эквивалентны.

В работе [8] было также показано, что неравенства (3.4c) и (3.4d) ((3.5c) и (3.5d)) эквивалентны.

Основным результатом данной работы является следующее утверждение.

Теорема 3.1. Неравенства (3.4b) и (3.4c) ((3.5b) и (3.5c)) эквивалентны.

4. Предварительные результаты

Мы докажем сейчас два вспомогательных утверждения.

Лемма 4.1. Пусть $\psi \in \mathrm{Lip}$, $(t_0, x_0) \in [0, \vartheta) \times R^n$,

$$a \in (0, \vartheta - t_0],\ b > 0 \quad \text{и} \quad \psi(t, x) > \psi(t_0, x_0)$$

при

$$t_0 \leqq t \leqq t_0 + a,\ \|x - x_0\| \leqq b,\ t - t_0 + \|x - x_0\| > 0\,.$$

Тогда для любого $\alpha > 0$ существует $(t_\alpha, x_\alpha, \varphi_\alpha) \in \Phi_{\min}(\psi)$ такой, что

$$t_0 < t_\alpha < t_0 + \alpha, \quad \|x_\alpha - x_0\| \leqq \alpha\,, \tag{4.1}$$

$$\frac{\partial \varphi_\alpha}{\partial t}(t_\alpha, x_\alpha) \geqq -\alpha, \quad \frac{\partial \varphi_\alpha}{\partial x}(t_\alpha, x_\alpha) = 0\,. \tag{4.2}$$

Доказательство леммы 4.1. Без ограничения общности можно принять, что

$$t_0 = 0, \quad x_0 = 0, \quad \psi(t_0, x_0) = 0\,. \tag{4.3}$$

Полагаем

$$B=\{x\in R^n:\|x\|\leqq b\},$$

$$R(\varepsilon)=\{r\geqq 0:\psi(t,x)\geqq rt,\quad \forall\, t\in[0,\varepsilon],\ x\in B\},$$

$$r(\varepsilon)=\sup R(\varepsilon),\quad r_0=\lim_{\varepsilon\downarrow 0} r(\varepsilon).$$

Отметим, что $0\in R(\varepsilon)$ при $\varepsilon\in(0,a]$. Поэтому

$$r_0\geqq 0. \tag{4.4}$$

Из определения числа r_0 следует, что для любого $r_*<r_0$ существует $\varepsilon(r_*)\in(0,a]$ такое, что

$$\psi(t,x)>r_*t\quad \text{при}\quad t\in(0,\varepsilon(r_*)],\ x\in B, \tag{4.5}$$

и для любого $r^*>r_0$ существует последовательность $(t_k,x_k)\in(0,a]\times B$ такая, что

$$\psi(t_k,x_k)<r^*t_k(k=1,2,\ldots),\quad t_k\to 0\quad \text{при}\quad k\to\infty. \tag{4.6}$$

В оценке (4.5) можно полагать, что

$$\varepsilon(r)\to 0\quad \text{при}\quad r\uparrow r_0. \tag{4.7}$$

Пусть $r<r_0$. Полагаем

$$\rho(t,r)=rt-(2t-\varepsilon(r))^2\,(r_0-r)/\varepsilon(r), \tag{4.8}$$

$$\text{при}\quad t\in[0,\varepsilon(r)],\ x\in B,$$

$$c(r)=\min_{(t,x)}(\psi(t,x)-\rho(t,r))\quad \text{при}\quad t\in[0,\varepsilon(r)],\ x\in B, \tag{4.9}$$

$$M(r)=\{(\tau,y)\in[0,\varepsilon(r)]\times B:\psi(\tau,y)-\rho(\tau,r)=c(r)\}. \tag{4.10}$$

Покажем, что

$$0<c(r)<(r_0-r)\varepsilon(r). \tag{4.11}$$

Из (4.5) и (4.8) следует, что при $t\in(0,\varepsilon(r)]$, $x\in B$

$$\psi(t,x)-\rho(t,r)>(2t-\varepsilon(r))^2(r_0-r)/\varepsilon(r)\geqq 0. \tag{4.12}$$

Поэтому $c(r)>0$. Пусть $r^*=2r_0-r$. Поскольку $r^*>r_0$, то из (4.6) видно, что существует точка (t^*,x^*), удовлетворяющая условиям

$$0<t^*<\varepsilon(r)/2,\ x^*\in B,\ \psi(t^*,x^*)-(2r_0-r)t^*<0. \tag{4.13}$$

Из (4.8) и (4.9) имеем

$$c(r)<\psi(t^*,x^*)-rt^*+4t^*(r_0-r)\,(t^*/\varepsilon(r)-1)+$$

$$+(r_0-r)\varepsilon(r).$$

Из оценки $t^*<\varepsilon(r)/2$ следует, что $t^*/\varepsilon(r)-1\leqq-1/2$. Учитывая (4.13), получаем $c(r)<(r_0-r)\varepsilon(r)$. Итак, оценки (4.11) доказаны.

Покажем, что

$$0<\tau<\varepsilon(r) \quad \text{для всех} \quad (\tau, y)\in M(r)\,. \tag{4.14}$$

Действительно, согласно (4.11) и (4.12) имеем

$$\psi(0, x)-\rho(0, r)\geqq(r_0-r)\varepsilon(r)>c(r)\,,$$

$$\psi(\varepsilon(r), x)-\rho(\varepsilon(r), r)\geqq(r_0-r)\varepsilon(r)>c(r)\,.$$

Отсюда следует (4.14).

Докажем импликацию

$$((\tau_k\,, y_k)\in M(r_k)\,(k=1, 2, \ldots)\,, \quad r_k\uparrow r_0 \quad \text{при} \quad k\to\infty)\Rightarrow$$
$$\Rightarrow(\,\|\, y_k\|\to 0 \quad \text{при} \quad k\to\infty)\,. \tag{4.15}$$

Допустим противное. Тогда без ограничения общности можно принять, что существует предел $y_k\to y_*$ при $k\to\infty$, $\|\, y_*\,\|>0$. Отметим, что из (4.7) и (4.14) следует, что $\tau_k\to 0$ при $k\to\infty$. Из условия $(\tau_k\,,\, y_k)\in M(r_k)$ имеем

$$\psi(\tau_k\,, y_k)-\rho\,(\tau_k\,, r_k)=c(r_k)\,.$$

Согласно (4.8) и (4.14) получаем $\rho(\tau_k\,, r_k)\to 0$ при $k\to\infty$. Из (4.11) имеем $c(r_k)\to 0$ при $k\to\infty$. Поэтому, учитывая (4.3), заключаем, что

$$\psi(0, y_*)=\psi(t_0\,, x_0)=0\,, \quad \|\, y_*\,\|>0\,.$$

Пришли к противоречию с условиями леммы 4.1. Итак, импликация (4.15) доказана.

Пусть $0<\alpha<\min\{a, b\}$. Выберем параметр r и точку $(t_\alpha\,, x_\alpha)\in M(r)$ так, чтобы выполнялись условия

$$r_0-\alpha/5<r<r_0\,, \quad 0<t\leqq\alpha\,, \quad \|\, x_\alpha\,\|\leqq\alpha\,. \tag{4.16}$$

Как показано выше, такой выбор возможен. Полагаем

$$\varphi_\alpha(t, x)=\rho(t, r)\,, \quad t\in[0, \vartheta]\,, \quad x\in R^n\,. \tag{4.17}$$

Заметим, что $\varphi_\alpha\in\mathbf{C}^1$, $(t_\alpha\,, x_\alpha)\in\operatorname{int}([0, \varepsilon\,(r)]\times B)$. Поэтому из (4.10) и определения множества $\Phi_{\min}$ следует, что $(t_\alpha\,, x_\alpha\,, \varphi_\alpha)\in\Phi_{\min}\,(\psi)$.

Очевидно, что $\partial\varphi_\alpha(t_\alpha\,, x_\alpha)/\partial x=0$. Остается проверить, что $\partial\varphi_\alpha(t_\alpha\,, x_\alpha)/\partial t\geqq -\alpha$. Из (4.8) и (4.17) имеем

$$\partial\varphi_\alpha(t_\alpha\,, x_\alpha)/\partial t=r-4(2t_\alpha-\varepsilon(r))\,(r_0-r)/\varepsilon(r)\,.$$

Согласно (4.14) получаем

$$\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial t \geqq r - 4(r_0 - r) = 5r - 4r_0 .$$

Наконец, учитывая (4.4) и (4.16), заключаем

$$\partial\varphi_\alpha(t_\alpha, x_\alpha)/\partial t \geqq 5r - 4r_0 \geqq 5(r - r_0) \geqq -\alpha .$$

Лемма 4.1 доказана.

Лемма 4.2. Пусть

$$\mu \in \mathrm{Lip}\, , (t_0, x_0) \in (0, \vartheta) \times R^n$$

и

$$\partial_-\mu(t_0, x_0) \mid (h) \geqq 0 , \quad \forall h \in R^n . \tag{4.18}$$

Тогда для любого $\beta > 0$ существуют $a \in (0, \vartheta - t_0]$, $b > 0$ такие, что функция

$$\psi(t, x) = \mu(t, x) + \beta((t - t_0)^2 + \| x - x_0 \|^2)^{1/2} \tag{4.19}$$

будет удовлетворять требованиям, указанным в лемме 4.1.

Доказательство леммы 4.2. Для

$$\mu \in \mathrm{Lip}\, , (t_0, x_0) \in (0, \vartheta) \times R^n , \quad \alpha \in R^1 \quad \text{и} \quad h \in R^n$$

полагаем

$$\partial_-\mu(t_0, x_0) \mid (\alpha, h) = \liminf_{\delta \downarrow 0} (\mu(t_0 + \delta\alpha, x_0 + \delta h) - \mu(t_0, x_0))\delta^{-1} .$$

Из этого определения следует равенство

$$\partial_-\mu(t_0, x_0) \mid (\lambda\alpha, \lambda h) = \lambda\partial_-\mu(t_0, x_0) \mid (\alpha, h)$$

для $\lambda \geqq 0$. Можно проверить, что отображение $(\alpha, h) \to \partial_-\mu(t_0, x_0) \mid (\alpha, h)$ удовлетворяет условию Липшица. Поэтому из условия леммы 4.2 следует, что

$$\partial_-\mu(t_0, x_0) \mid (\alpha, h) \geqq 0 , \quad \forall \alpha \geqq 0 , \quad h \in R^n . \tag{4.20}$$

Предположим от противного, что указанная в лемме 4.2 функция ψ не удовлетворяет требованиям, сформулированным в лемме 4.1. Тогда существует последовательность (t_k, x_k) $(k = 1, 2, \dots)$ такая, что

$$\begin{gathered} t_k \geqq t_0 , (t_k, x_k) \to (t_0, x_0) \quad \text{при} \quad k \to \infty , \\ \psi(t_k, x_k) \leqq \psi(t_0, x_0) , \delta_k = t_k - t_0 + \| x_k - x_0 \| > 0 . \end{gathered} \tag{4.21}$$

Полагаем $\alpha_k = (t_k - t_0)\delta_k^{-1}$. Заметим, что $\alpha_k \in [0, 1]$. Без ограничения общности можно принять, что $\alpha_k \to \alpha_*$ при $k \to \infty$. Справедливы равенства

$$t_k = t_0 + \alpha_k\delta_k , \quad x_k = x_0 + h_k\delta_k ,$$

$$h_k = (x_k - x_0)(1 - \alpha_k) \| x_k - x_0 \|^{-1} \quad \text{при} \quad x_k \neq x_0 ,$$

$$h_k = 0 \quad \text{при} \quad x_k = x_0 .$$

Здесь тоже можно полагать, что $h_k \to h_*$ при $k \to \infty$. Заметим, что $\alpha_k + \| h_k \| = 1$, поэтому

$$\alpha_* + \| h_* \| = 1 . \tag{4.22}$$

Итак, получаем

$$t_k = t_0 + \alpha_* \delta_k + (\alpha_k - \alpha_*) \delta_k ,$$

$$x_k = x_0 + h_* \delta_k + (h_k - h_*) \delta_k .$$

Поскольку $\psi \in \mathrm{Lip}$, то

$$\psi(t_k , x_k) \geqq \psi(t_0 + \alpha_* \delta_k , x_0 + h_* \delta_k) - \lambda(| \alpha_k - \alpha_* | + \| h_k - h_* \|) \delta_k .$$

Следовательно,

$$\begin{aligned} &\partial_- \psi(t_0 , x_0) \,|\, (\alpha_* , h_*) \leqq \\ &\leqq \liminf_{k \to \infty} (\psi(t_0 + \alpha_* \delta_k , x_0 + h_* \delta_k) - \psi(t_0 , x_0)) \delta_k^{-1} = \\ &= \lim_{i \to \infty} (\psi(t_0 + \alpha_* \delta_{k_i} , x_0 + h_* \delta_{k_i}) - \psi(t_0 , x_0)) \delta_{k_i}^{-1} \leqq \\ &\leqq \limsup_{i \to \infty} ((\psi(t_{k_i} , x_{k_i}) - \psi(t_0 , x_0)) \delta_{k_i}^{-1} + \\ &+ \lambda(| \alpha_{k_i} - \alpha_* | + \| h_{k_i} - h_* \|)) \leqq 0 . \end{aligned} \tag{4.23}$$

По определению функции ψ справедливо равенство

$$\begin{aligned} \partial_- \psi(t_0 , x_0) \,|\, (\alpha_* , h_*) = \partial_- \mu(t_0 , x_0) \,|\, (\alpha_* , h_*) + \\ + \beta(\alpha_*^2 + \| h_* \|^2)^{1/2} . \end{aligned}$$

Поэтому из (4.22) и (4.23) получаем неравенство

$$\partial_- \mu(t_0 , x_0) \,|\, (\alpha_* , h_*) < 0 ,$$

которое противоречит (4.20). Полученное противоречие доказывает лемму 4.2.

5. Доказательство теоремы 3.1

Докажем импликацию (3.4в)⇒(3.4с).

Пусть функция $\omega \in \mathrm{Lip}$ удовлетворяет условию (3.4в) и $(t, x, \varphi) \in \Phi_{\min}(\omega)$. Из (3.1), (3.3) следует

$$\frac{-\partial\varphi}{\partial t}(t, x) \geqq \left\langle \frac{\partial\varphi}{\partial x}(t, x), h \right\rangle - \partial_{-}\omega(t, x) \mid (h), \quad \forall\, h \in R^n .$$

Согласно (3.2) имеем

$$\frac{-\partial\varphi}{\partial x}(t, x) \geqq D^*\omega(t, x) \left| \left(\frac{\partial\varphi}{\partial x}(t, x) \right) \right. .$$

Тогда из (3.4в) получаем (3.4с). Аналогично доказывается импликация (3.5в)⇒(3.5с).

Докажем теперь импликацию (3.4с)⇒(3.4в).

Пусть функция $\omega \in \mathrm{Lip}$ удовлетворяет условию (3.4с). Выберем точку

$$(t_0, x_0, l_0) \in (0, \vartheta) \times R^n \times R^n .$$

Если $D^*\omega(t_0, x_0) \mid (l_0) = +\infty$, то (3.4в) выполнено. Пусть

$$D^*\omega(t_0, x_0) \mid (l_0) = c < +\infty . \tag{5.1}$$

Согласно (3.2) имеем

$$\langle l_0, h \rangle - \partial_{-}\omega(t_0, x_0) \mid (h) \leqq c, \quad \forall\, h \in R^n . \tag{5.2}$$

Пусть

$$\mu(t, x) = \omega(t, x) - \langle l_0, x \rangle + ct . \tag{5.3}$$

Из (5.2) следует, что функция μ удовлетворяет условию (4.18). Согласно лемме 4.2 функция ψ вида (4.19) удовлетворяет условиям леммы 4.1. Пусть $(t_\alpha, x_\alpha, \varphi_\alpha)$ — элемент множества $\Phi_{\min}(\psi)$, для которого выполнены оценки (4.1) и (4.2).

Из (3.3), (4.19), (5.3) имеем

$$\begin{gathered} \psi(t, x) - \varphi_\alpha(t, x) = \omega(t, x) - v_*(t, x) \geqq \\ \geqq \psi(t_\alpha, x_\alpha) - \varphi_\alpha(t_\alpha, x_\alpha) = \omega(t_\alpha, x_\alpha) - v_*(t_\alpha, x_\alpha) \end{gathered} \tag{5.4}$$

при

$$(t, x) \in N_\varepsilon(t_\alpha, x_\alpha) .$$

Здесь

$$v_*(t, x) = \langle l_0, x \rangle - ct - \beta((t - t_0)^2 + \| x - x_0 \|^2)^{1/2} + \varphi_\alpha(t, x) . \tag{5.5}$$

Отметим, что функция v_* недифференцируема в точке (t_0, x_0).

Можно полагать, что

$$(t_0-t_\alpha)^2+\|x_0-x_\alpha\|^2>\varepsilon^2 .$$

Тогда можно определить функцию

$$v\in \mathbf{C}^1 , \quad v(t,x)=v_*(t,x)$$

при

$$(t,x)\in N_\varepsilon(t_\alpha, x_\alpha) .$$

Из (3.3), (5.4) следует, что

$$(t_\alpha, x_\alpha, v)\in \Phi_{\min}(\omega) .$$

Согласно (3.4с) имеем

$$-\partial v(t_\alpha, x_\alpha)/\partial t \geqq H(t_\alpha, x_\alpha, \partial v(t_\alpha, x_\alpha)/\partial x) .$$

Из (4.2) и (5.5) получаем

$$-\partial v(t_\alpha, x_\alpha)/\partial t \leqq c+\alpha+\beta ,$$

$$\partial v(t_\alpha, x_\alpha)/\partial x=l_0-s ,$$

где

$$s=\beta(x-x_0)/((t-t_0)^2+\|x-x_0\|^2)^{1/2} , \quad \|s\|\leqq\beta .$$

Поэтому $c\geqq H(t_\alpha, x_\alpha, l_0-s)-\alpha-\beta$. Функция $(t,x,l)\mapsto H(t,x,l)$ непрерывна, параметры α и β можно взять сколь угодно малыми, точка (t_α, x_α) удовлетворяет соотношениям (4.1). Следовательно, $c\geqq H(t_0, x_0, l_0)$. Учитывая обозначение (5.1), получаем (3.4в).

Импликация (3.4с)⇒(3.4в) доказана.

Аналогичным образом можно доказать импликацию (3.5с)⇒(3.5в).

Теорема 3.1 доказана.

6. Заключительно.е примечание

Теорема 3.1 справедлива для любого непрерывного гамильтониана $H(t,x,l)$. В приведенном доказательстве не используется тот факт, что гамильтониан определен равенством (2.4).

Литература

1. *Айзекс Р.* Дифференциальные игры. М., Мир, 1967, 479 с.
2. *Красовский Н. Н.* Дифференциальная игра сближения — уклонения, I, II Изв. АН СССР, Техн. кибернетика, 1973, № *2*, с. 3–18, № *3*, с. 22–42.
3. *Красовский Н. Н., Субботии А. И.* Позиционные дифференциальные игры. М., Наука, 1974. 456 с.
4. *Субботин А. И.* Обобщение основного уравнения теории дифференциальных игр. Докл. АН СССР, 1980, т. **254,** № *2,* с. 293–297.
5. *Субботин А. И., Тарасьев А. М.* Сопряженные производные функции цены дифференциальной игры. Докл. АН СССР, 1985, т. **283,** № *3,* с. 559–564.
6. *Barron, E. N., Evans, L. C., Jensen, R.*, Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls, J. Different. Equat., 1984, vol. **53,** No. *2*, pp. 213–233.
7. *Crandall, M. G., Lions, P. L.*, Viscosity solutions of Hamilton–Jacobi equations, Trans. Amer. Math. Soc., 1983, vol. **277,** No. *1*, pp. 1–42.
8. *Crandall, M. G., Linons, P. L.*, Some properties of viscosity solutions of Hamilton–Jacobi equations, Trans. Amer. Math. Soc., 1984, vol. **282,** No. *2*, pp. 487–502.
9. *Lions, P. L., Souganidis, P. E.*, Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations, SIAM J. Control. Optimiz., 1985, vol. **23,** No. *41*, pp. 566–583.
10. *Rademacher, H.*, Über Partielle und Totale Differenzierbarkeit von Funktionen Mehrer Variablen under die Transformation der Doppelintegrale, Mathematishe Annalen, 1918, vol. **79,** pp. 340–354.
11. *Subbotin, A. I.*, Generalization of the main equation of differential game theory, J. Optimiz. Theory and Appl., 1984, vol. **43,** No. *1*, pp. 103–133.

PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda, Budapest

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

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). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = \|a_{ij}\|$. Write diagonal matrices as diag (d_1, d_2, . . . d_n).

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 129090 Москва И-90, ул. Щепкина, 8. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 208-60-19).

Объём 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 языке в трех экземплярах, 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 и 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 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

HU ISSN 0370-2529 Index: 26 660