VOL. TOM 18 · NUMBER HOMEP 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

1989

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 PLC Headington Hill Hall, Oxford OX3 OBW, England

or

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

1989 Subscription Rate DM 535,— per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 18 (1989)

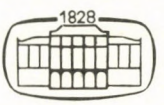
SUBJECT INDEX

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 18 (1989)

AUTHOR INDEX

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. 18 (1), pp. 3—16 (1989)

MULTIINFORMATION AND THE PROBLEM OF CHARACTERIZATION OF CONDITIONAL INDEPENDENCE RELATIONS

M. STUDENÝ

(Prague)

(Received January 10, 1988)

Certain algebraic relation between multiinformation and conditional mutual information is established. It is shown to be applicable to the problem of characterization of conditional independence relations arising in connection with probabilistic expert systems. More concretely, a new axiom of these relations is derived. Some auxiliary results have their own significance: the characterization of marginally continuous measures in Proposition 1 and the information theoretical significance of the conditional product of measures mentioned in Consequence 3.

Introduction

The main concept of this paper is a certain generalization of the concept of mutual information, namely the so-called *multiinformation*. Simply, multiinformation is the relative entropy of the simultaneous distribution of a finite collection of random variables with respect to the product of the distributions of individual random variables. It is nonnegative and vanishes iff the corresponding random variables are independent. So, similarly as the mutual information which can serve as a measure of dependence of two random variables (see [13]), multiinformation enables us to characterize the level of dependence of more than two random variables. From this point of view it was studied by Perez in [8].

There are several papers belonging to information theory which indirectly handle multiinformation. For example, in [1] the studied algorithm IPFP converges to such probability measure which minimizes multiinformation in some given family of measures having prescribed marginals. As statistical properties of multiinformation are concerned, they are investigated in [12].

In this paper we want to show that multiinformation is also useful in apparently remote spheres. Namely, its certain "algebraic" properties can be applied to the problem of characterization of *conditional independence relations* (we shall use the abbreviation CIR here). This problem arises in connection with probabilistic expert systems, i.e. expert systems based on principles of probability theory.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

The first section contains the definitions of the basic concepts and recalls some facts used later. Note that we take multiinformation as a characteristic of a probability measure; those who prefer to speak about random variables can regard the probability measure as the distribution of the corresponding variables. Moreover, we subjoin a proposition which establish an interesting equivalence connection between marginally continuous measures and measures that can be formed by a dominated kernel.

In the second section the conditional product of measures is defined and some facts about it are mentioned.

The third section deals with the concept of conditional mutual information which is defined by means of the concept of conditional product of measures. In Consequence 1 the fundamental formula for the conditional mutual information is given.

The fourth section considers both multiinformation and conditional mutual information as a set function on subsets of the index set. An important algebraic connection between them is established there.

Finally, the mentioned connection is applied in the last section. The problem of characterization of conditional independence relations (CIR's) is formulated there and it is shown how it is possible to utilize multiinformation.

1. Basic definitions, auxiliary concepts and results

Given measurable spaces $(X, \mathcal{X})$, $(Y, \mathcal{Y})$ and a probability measure R on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ *the marginal measure* (or simply the *marginal*) *of* R *on* $(X, \mathcal{X})$ is defined by

$$R^X(A) = R(A \times Y), \qquad A \in \mathcal{X}.$$

We denote it by the symbol of the original measure having as upper index the symbol of the corresponding space.

Let us suppose that two probability measures P and Q on a measurable space $(X, \mathcal{X})$ are given. In case $P \ll Q$ we take some function $f: X \to \langle 0, \infty)$ (it means that f is defined everywhere on X and has all values finite and nonnegative) which is a version of the Radon–Nikodym derivative dP/dQ and define the *relative entropy of* P *w. r. to* Q (we use the abbreviation w. r. instead of "with respect") as the integral:

$$H(P, Q) = \int_{x \in X} \ln f(x) dP(x).$$

Since $P\{x \in X;\ f(x) = 0\} = 0$, it is not essential what is $\ln(0)$. Evidently, the value of $H(P, Q)$ does not depend on the choice of a version of dP/dQ. In case $P \not\ll Q$ we put $H(P, Q) = \infty$. In this paper we denote relative entropy by the letter H.

Relative entropy is always nonnegative and vanishes iff $P=Q$. Moreover, if $\mathscr{Y}$ is a sub-σ-algebra of $\mathscr{X}$ and $\tilde{P}$ or $\tilde{Q}$ is the restriction of P or Q on $\mathscr{Y}$ (respectively), then $H(\tilde{P}, \tilde{Q}) \leqq H(P, Q)$. Especially, it follows for every pair of probability measures P, Q on a product $(X \times Y, \mathscr{X} \times \mathscr{Y})$:

$$H(P^X, Q^X) \leqq H(P, Q). \tag{1}$$

These basic properties are well known, see e.g. [10] or [9].

If P is a probability measure on a product $(X \times Y, \mathscr{X} \times \mathscr{Y})$ then the *mutual information between X and Y* is defined as the relative entropy of P w. r. to $P^X \times P^Y$.

Analogously, given a finite nonempty collection of measurable spaces $(X_i, \mathscr{X}_i)$, $i \in N$ and a probability measure P on $\left(\prod_{i \in N} X_i, \prod_{i \in N} \mathscr{X}_i \right)$ we define the *multiinformation of P* as the relative entropy of P w. r. to the product of its one-dimensional marginals:

$$M(P) = H\left(P, \prod_{i \in N} P^{X_i} \right).$$

In this paper multiinformation is denoted by the letter M.

In this paragraph $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ are measurable spaces and R is a probability measure on $(X \times Y, \mathscr{X} \times \mathscr{Y})$. By a *representative of conditional probability on* $(Y, \mathscr{Y})$ *w. r. to* $(X, \mathscr{X})$ we shall understand every mapping $K: \mathscr{Y} \times X \to \langle 0, 1 \rangle$ such that for each $B \in \mathscr{Y}$ the function $x \mapsto K(B|x)$ is a variant of conditional probability of the set B given the σ-algebra $\mathscr{X}$, i.e. $\mathscr{X}$-measurable function satisfying:

$$\int_{x \in A} K(B|x) dR^X(x) = R(A \times B) \qquad \text{for each} \quad A \in \mathscr{X}. \tag{2}$$

We shall use the abbreviation c. p. instead of "conditional probability". Note that (2) can be formulated equivalently as follows:

$$\int_{x \in X} g(x) \cdot K(B|x) dR^X(x) = \int_{(x,y) \in X \times B} g(x) dR(x, y) \qquad \text{for each} \quad g: X \to \langle 0, 1 \rangle \quad \mathscr{X}\text{-measurable}. \tag{3}$$

The existence of a representative of c. p. is a trivial consequence of the Radon–Nikodym theorem. Indeed, for each $B \in \mathscr{Y}$ the function $A \mapsto R(A \times B)$ is a measure on $(X, \mathscr{X})$ which is absolutely continuous w. r. to R^X. Evidently, representatives of c. p. are determined uniquely in the framework of this equivalence:

$$K \simeq K' \quad \text{iff} \quad K(B|x) = K'(B|x) \quad \text{for} \quad R^X\text{-a.e.} \quad x \in X \quad \text{for every} \quad B \in \mathscr{Y}.$$

We shall use the symbol $R_{Y|X}$ to denote an arbitrary representative of c. p., i.e. the symbol of the original measure having as lower index the separate symbols of the respective spaces.

A representative K of c. p. on $(Y, \mathcal{Y})$ w. r. to $(X, \mathcal{X})$ is called *regular* iff for each $x \in X$ the function $B \mapsto K(B|x)$ is a probability measure on $(Y, \mathcal{Y})$. In case there exists a regular representative of c. p. on $(Y, \mathcal{Y})$ w. r. to $(X, \mathcal{X})$ we shall say that c. p. on $(Y, \mathcal{Y})$ w. r. to $(X, \mathcal{X})$ is regular.

In this paragraph we suppose measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ are given. By stochastic (or Markov) *kernel from* $(X, \mathcal{X})$ *to* $(Y, \mathcal{Y})$ we understand a collection $\mathcal{P} = \{P_x; x \in X\}$ of probability measures on $(Y, \mathcal{Y})$ such that for each $B \in \mathcal{Y}$ the function $x \mapsto P_x(B)$ is $\mathcal{X}$-measurable. This concept is also known as a crossing probability or as a channel (in information theory, especially).

We shall say that a kernel $\mathcal{P} = \{P_x; x \in X\}$ is *dominated* iff there exists a probability measure τ on $(Y, \mathcal{Y})$ such that for each $x \in X$ $P_x \ll \tau$.

Given a kernel $\mathcal{P} = \{P_x; x \in X\}$ from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ and a probability measure Q on $(X, \mathcal{X})$ we can define a probability measure $Q * \mathcal{P}$ on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ by:

$$Q * \mathcal{P}(A \times B) = \int_{x \in A} P_x(B) dQ(x) \qquad A \in \mathcal{X},\ B \in \mathcal{Y} \tag{4}$$

and by the standard extension argument (see [4], III.2.1). We shall say that Q *and* $\mathcal{P}$ *form the measure* $Q * \mathcal{P}$. Note that (4) can be extended as follows:

$$Q * \mathcal{P}(C) = \int_{x \in X} P_x(C_x) dQ(x) \qquad C \in \mathcal{X} \times \mathcal{Y}$$

where $C_x = \{y \in Y;\ (x, y) \in C\}$. Especially:

$$Q * \mathcal{P}(C) = 0 \quad \text{iff} \quad P_x(C_x) = 0 \quad \text{for} \quad Q\text{-a.e.} \quad x \in X. \tag{5}$$

Remark 1. a) Let us point out an interesting connection. Supposing that R is a probability measure on $(X \times Y, \mathcal{X} \times \mathcal{Y})$, we can easily derive that $R = Q * \mathcal{P}$ for some probability measure Q on $(X, \mathcal{X})$ and some kernel $\mathcal{P}$ from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ iff the c. p. on $(Y, \mathcal{Y})$ w. r. to $(X, \mathcal{X})$ is regular.

b) Further, we note the known fact that the assumption saying that Y is a separable complete metric space and $\mathcal{Y}$ is the σ-algebra of its Borel subsets suffices for regularity of c. p. on $(Y, \mathcal{Y})$ w. r. to $(X, \mathcal{X})$ (see [4], it follows from the consequence of V.4.4). Especially, it holds for finite Y with $\mathcal{Y} = \exp Y$.

If a probability measure P on a product $(X \times Y, \mathcal{X} \times \mathcal{Y})$ satisfies $P \ll P^X \times P^Y$, then it is called *marginally continuous*. Evidently, this condition is necessary for finiteness of the mutual information between X and Y. The following lemma leads to some characterization of marginally continuous measures in Proposition 1.

Lemma 1. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ be measurable spaces, λ a probability measure on $(X, \mathscr{X})$, τ on $(Y, \mathscr{Y})$ and μ on $(X \times Y, \mathscr{X} \times \mathscr{Y})$. Then $\mu \ll \lambda \times \tau$ iff $\mu^X \ll \lambda$ and there exists a kernel $\mathscr{P} = \{P_x; x \in X\}$ from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$ such that $\mu = (\mu^X) * \mathscr{P}$ and for each $x \in X$ it holds $P_x \ll \tau$.

Proof. a) In case $\mu \ll \lambda \times \tau$ we can take such a version $L: X \times Y \to \langle 0, \infty)$ of $d\mu/d(\lambda \times \tau)$ that

$$l(x) = \int_{y \in Y} L(x, y)\, d\tau(y) < \infty \qquad x \in X.$$

Evidently, l is a version of $d(\mu^X)/d\lambda$. We define $k(x, y) = l^{-1}(x) \cdot L(x, y)$ if $l(x) > 0$ and $k(x, y) = 1$, otherwise. Finally, we put:

$$P_x(B) = \int_{y \in B} k(x, y) d\tau(y) \qquad x \in X, B \in \mathscr{Y}.$$

It makes no problem to verify that $\mathscr{P} = \{P_x; x \in X\}$ is the desired kernel.

b) The sufficiency can be seen using (5). For $C \in \mathscr{X} \times \mathscr{Y}$ the relation $(\lambda \times \tau)(C) = 0$ implies $\tau(C_x) = 0$ for λ-a.e. $x \in X$. So $P_x(C_x) = 0$ for μ^X-a.e. $x \in X$, i.e. $(\mu^X) * \mathscr{P}(C) = 0$. ■

Proposition 1. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ be measurable spaces. Then the following conditions on a probability measure μ on $(X \times Y, \mathscr{X} \times \mathscr{Y})$ are equivalent:

(a) μ is marginally continuous

(b) there exist a probability measure λ on $(X, \mathscr{X})$ and a probability measure τ on $(Y, \mathscr{Y})$ such that $\mu \ll \lambda \times \tau$

(c) μ can be formed by a dominated kernel from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$.

Proof. Directly from Lemma 1 we conclude that (b) implies (c). Conversely, if $\mu = Q * \mathscr{P}$ where Q is a measure on $(X, \mathscr{X})$ and $\mathscr{P}$ is the mentioned kernel, then necessarily $Q = \mu^X$. So, we can take $\lambda = \mu^X$ in Lemma 1 to show that (c) implies (b). In fact we have just proved that $\mu \ll \lambda \times \tau$ implies $\mu \ll \mu^X \times \tau$. Replacing of $(X, \mathscr{X})$ by $(Y, \mathscr{Y})$ we get that $\mu \ll \lambda \times \tau$ implies $\mu \ll \lambda \times \mu^Y$. So, let us take $\lambda = \mu^X$ here and see that (b) implies (a). The converse is trivial. ■

Note that Proposition 1 yields a sufficient condition for regularity of c. p., which is not of topological nature (see Remark 1).

2. Conditional product of measures

Definition 1. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$, $(Z, \mathscr{Z})$ be measurable spaces and P a probability measure on $(X \times Y \times Z, \mathscr{X} \times \mathscr{Y} \times \mathscr{Z})$. We shall say that P is a *conditional product on* $X \times Y$ *under condition* Z iff it holds

$$\left.\begin{aligned} P_{X \times Y | Z}(A \times B | z) = P_{X|Z}(A|z) \cdot P_{Y|Z}(B|z) \quad &\text{for} \quad P^Z\text{-a.e.} \quad z \in Z \\ &\text{for each} \quad A \in \mathscr{X}, \quad B \in \mathscr{Y}. \end{aligned}\right\} \tag{6}$$

Naturally, we write $P_{X|Z}$ instead of $(P^{X\times Z})_{X|Z}$. Evidently, the validity of (6) does not depend on the choice of representatives of c. p. Further, it is easy to see that (6) is equivalent to:

$$\left.\begin{aligned} P(A\times B\times C)=\int_{z\in C} P_{X|Z}(A|z)\cdot P_{Y|Z}(B|z)\,dP^Z(z) \\ \text{for each}\quad A\in\mathcal{X},\ B\in\mathcal{Y},\ C\in\mathcal{Z}. \end{aligned}\right\} \tag{7}$$

We use this terminology in order not to impair analogy with the "unconditional" case: a probability measure R on $(X\times Y, \mathcal{X}\times\mathcal{Y})$ is the product of its marginals R^X and R^Y iff in the probability space $(X\times Y, \mathcal{X}\times\mathcal{Y}, R)$ the σ-algebras $\mathcal{X}\times\mathcal{Y}'$ and $\mathcal{X}'\times\mathcal{Y}$ are independent ($\mathcal{X}'$, $\mathcal{Y}'$, $\mathcal{Z}'$ are respectively trivial σ-algebras on $X\times Y\times Z$). Analogously, (6) means that in the probability space $(X\times Y\times Z, \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}, P)$ the σ-algebras $\mathcal{X}\times\mathcal{Y}'\times\mathcal{Z}'$ and $\mathcal{X}'\times\mathcal{Y}\times\mathcal{Z}'$ are conditionally independent given the σ-algebra $\mathcal{X}'\times\mathcal{Y}'\times\mathcal{Z}$ (see [5], chapter VII, § 25.3).

Remark 2. The usual "unconditional" product of measures can be viewed as a special case of the conditional product. Indeed, supposing that $\mathcal{Z}$ is the trivial σ-algebra on Z, a measure P on $(X\times Y\times Z, \mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ is a conditional product on $X\times Y$ under condition Z iff $P^{X\times Y}=P^X\times P^Y$.

Definition 2. Let $(X, \mathcal{X})$, $(Y, \mathcal{Y})$, $(Z, \mathcal{Z})$ be measurable spaces, $Q_{X\times Z}$ and $Q_{Y\times Z}$ be consonant probability measures, respectively, on $X\times Z$ and on $Y\times Z$, i.e. $(Q_{X\times Z})^Z=(Q_{Y\times Z})^Z$.

In case there exists a measure P on $(X\times Y\times Z, \mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ having $Q_{X\times Z}$ and $Q_{Y\times Z}$ as marginals which is moreover a conditional product on $X\times Y$ under condition Z, we shall call it the *conditional product of* $Q_{X\times Z}$ *and* $Q_{Y\times Z}$.

Proposition 2. Under assumptions of Definition 2 it holds:

a) The conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$ is determined uniquely.

b) Supposing that $Q_{X\times Z}$ has regular c. p. on (X, x) w. r. to (Z, z) or that $Q_{Y\times Z}$ has regular c. p. on (Y, y) w. r. to (Z, z)) there exists the conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$.

Since measures having the same marginals on $X\times Z$ have the same set of representatives of c. p. on $(X, \mathcal{X})$ w. r. to $(Z, \mathcal{Z})$, we can show the first part of Proposition 2 using (7). For the proof of the second part we refer to the translator's remarks to chapter 3 of [10].

Combining Propositions 1 and 2b we see the known fact mentioned in [10] (p. 56), namely: supposing that $Q_{X\times Z}$ (or $Q_{Y\times Z}$) in Definition 2 is marginally continuous, there exists the conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$.

Nevertheless, under assumptions of Definition 2 the conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$ may not exist, moreover it holds:

Proposition 3. There exist measurable spaces $(X, \mathcal{X})$, $(Y, \mathcal{Y})$, $(Z, \mathcal{Z})$ and a probability measure P on $(X\times Y\times Z, \mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ such that the conditional product of $P^{X\times Z}$ and $P^{Y\times Z}$ does not exist.

For the proof we refer to [11], where the desired example is constructed.

3. Conditional mutual information

Definition 3. Let $(X, \mathcal{X})$, $Y, \mathcal{Y})$, $(Z, \mathcal{Z})$ be measurable spaces and P a probability measure on $(X\times Y\times Z, \mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$. In case there exists the conditional product of $P^{X\times Z}$ and $P^{Y\times Z}$ (denoted by $\bar{P}$), we put:

$$C(X;\, Y|Z)=H(P,\bar{P}).$$

In the opposite case we put $C(X;\, Y|Z)=\infty$. The number $C(X;\, Y|Z)$ we shall call the *conditional mutual information between X and Y under condition Z.*

The following lemma is a trivial consequence of the basic properties of the relative entropy:

Lemma 2. Under assumptions of Definition 3 it holds $C(X;\, Y|Z)\geqq 0$. Moreover, $C(X;\, Y|Z)=0$ iff P is a conditional product on $X\times Y$ under condition Z.

The well-known notion of mutual information can be viewed as a special case of conditional mutual information, if we take $\mathcal{Z}$ as the trivial σ-algebra on Z (cf. Remark 2). Indeed, it must hold

$$P=P^{X\times Y}\times P^{Z} \quad\text{and}\quad \bar{P}=P^{X}\times P^{Y}\times P^{Z} \quad\text{and}\quad H(P,\bar{P})=H(P^{X\times Y},\, P^{X}\times P^{Y}).$$

The following lemma we need for the proof of the fundamental formula (10) in Consequence 1:

Lemma 3. Under assumptions of Definition 3 we denote $R=P^{Y}\times P^{X\times Z}$.

a) If $P^{Y\times Z}\ll P^{Y}\times P^{Z}$, then there exists the conditional product $\bar{P}$ of $P^{X\times Z}$ and $P^{Y\times Z}$. Moreover, $\bar{P}\ll R$ and there exists a function k: $Y\times Z\to\langle 0,\infty)$ which is a version of $d(P^{Y\times Z})/d(P^{Y}\times P^{Z})$ and viewed as a function on $X\times Y\times Z$ a version of $d\bar{P}/dR$.

b) The following two conditions are equivalent:

$$P\ll R, \tag{8}$$

$$\left.\begin{array}{l} P^{Y\times Z}\ll P^{Y}\times P^{Z} \text{ and there exists the conditional product} \\ \bar{P} \text{ of } P^{X\times Z} \text{ and } P^{Y\times Z} \text{ which, moreover, satisfies } P\ll\bar{P}. \end{array}\right\} \tag{9}$$

We shall not prove this lemma. The proof can be found in [2] (pp. 42–44), but with the proviso that one must be careful whether the conditional product of measures exists. Namely, in the mentioned paper there is an erroneous consideration leading to the conclusion that the existence of P suffices for the existence of the conditional product of $P^{X\times Z}$ and $P^{Y\times Z}$ (more exactly, the set function (2.7.7) is not countably additive). It was said in Proposition 3 that the mentioned conclusion is wrong.

In this paper we extended the definition of conditional mutual information in order to preserve the general validity of relation (16) mentioned below.

Consequence 1. Under assumption of Definition 3 it holds

$$H(P, P^Y \times P^{X\times Z}) = C(X; Y|Z) + H(P^{Y\times Z}, P^Y \times P^Z). \tag{10}$$

Proof. If (8) does not hold, then according to Lemma 3b both sides of (10) are infinite. In case (8) holds, we use Lemma 3a and fix the function k: $Y\times Z\to\langle 0, \infty)$ mentioned there. Further, according to Lemma 3b we may consider some version l: $X\times Y\times Z\to\langle 0, \infty)$ of $dP/d\bar{P}$. So, k being considered as a function on $X\times Y\times Z$, the product $k\cdot l$ is a version of dP/dR. Finally, integrating the identity

$$\ln(k\cdot l) = \ln(k) + \ln(l) \qquad (\text{where } \ln 0 = -\infty)$$

with respect to P, we get (10). ■

4. Multiinformation viewed as a set function

(S) In the remaining two sections we shall consider the following situation. A finite nonempty collection of measurable spaces $(X_i, \mathscr{X}_i)$, $i\in N$ is given. If $A\subset N$ is nonempty, we shall write $(X_A, \mathscr{X}_A)$ instead of $\left(\prod_{i\in A} X_i, \prod_{i\in A}\mathscr{X}_i\right)$. Further, a probability measure P on $(X_N, \mathscr{X}_N)$ is given. For the sake of brevity, the marginal of P on $(X_A, \mathscr{X}_A)$ will be denoted by P^A.

Definition 4. Assuming (S), we define for nonempty $A\subset N$:

$$I_m[A] = M(P^A).$$

Moreover, for empty A we put $I_m[\emptyset] = 0$.

From basic properties of relative entropy we easily conclude that assuming (S) the function I_m: exp $N \to \langle 0, \infty \rangle$ satisfies:

$$A \subset B \quad \text{implies} \quad I_m[A] \leqq I_m[B] \tag{11}$$

$$\text{if card } A \leqq 1 \qquad \text{then} \quad I_m[A] = 0. \tag{12}$$

Definition 5. Assuming (S), we define for every ordered triplet $\langle A, B, C \rangle$ of disjoint subsets of N the number $I_c[A; B|C] \in \langle 0, \infty \rangle$. If all the sets A, B, C are nonempty, then we define it as the conditional mutual information between X_A and X_B under condition X_C (logically it is computed from $P^{A \cup B \cup C}$), i.e.

$$I_c[A; B|C] = C(X_A; X_B|X_C).$$

For empty C and nonempty A, B we define $I_c[A; B|\emptyset]$ as the mutual information between X_A and X_B, i.e.

$$I_c[A; B|\emptyset] = H(P^{A \cup B}, P^A \times P^B).$$

Finally, in case that A or B is empty we put:

$$I_c[\emptyset; B|C] = 0 \qquad \text{and} \quad I_c[A; \emptyset|C] = 0.$$

Lemma 4. Assuming (S), the function I_c satisfies (A, B, C are supposed to be disjoint):

$$I_c[A; B|C] = I_c[B; A|C] \tag{13}$$

$$0 \leqq I_c[A; B|C] \tag{14}$$

$$I_c[A; B \cup C|\emptyset] = I_c[A; B|C] + I_c[A; C|\emptyset] \tag{15}$$

$$\text{if} \quad A' \subset A, \ B' \subset B, \quad \text{then} \quad I_c[A', B'|C] \leqq I_c[A; B|C]. \tag{16}$$

Proof. (13) and (14) are easy consequences of the definition; (15) follows directly from (10) and (13). (16) is trivial in case $I_c[A; B|C] = \infty$. In the opposite case there exists the conditional product of $P^{A \cup C}$ and $P^{B \cup C}$. It makes no problem to verify that its restriction onto $X_{A' \cup B' \cup C}$ is the conditional product of $P^{A' \cup C}$ and $P^{B' \cup C}$. So (16) follows from (1). ■

The substantial relation between I_m and I_c is established by the following statement.

Proposition 4. Assuming (S), it holds for every $D, E \subset N$ (not necessarily disjoint):

$$I_m[D \cup E] + I_m[D \cap E] = I_m[D] + I_m[E] + I_c[E \setminus D; D \setminus E | D \cap E]. \tag{17}$$

Proof. a) First we prove (17) for disjoint D and E. So, if D and E are nonempty (otherwise trivial), then we denote $Q_A = \prod_{i \in A} P^{(i)}$ for nonempty $A \subset N$. In case $P^{D \cup E} \not\ll P^D \times P^E$ it is, according to Proposition 1, $P^{D \cup E} \not\ll Q_D \times Q_E = Q_{D \cup E}$. So, both $I_m[D \cup E]$ and $I_c[E \setminus D; D \setminus E | D \cap E]$ are infinite and (17) holds. Analogously we proceed in case $P^D \not\ll Q_D$ or $P^E \not\ll Q_E$ (using (11)). So, we can suppose $P^D \ll Q_D$, $P^E \ll Q_E$ and $P^{D \cup E} \ll P^D \times P^E$. We take a version $f: X_{D \cup E} \to \langle 0, \infty)$ of $d(P^{D \cup E})/d(P^D \times P^E)$, a version $h: X_D \to \langle 0, \infty)$ of dP^D/dQ_D and a version $g: X_E \to \langle 0, \infty)$ of dP^E/dQ_E. The proof we conclude similarly as the proof of Consequence 1.

b) Now we suppose arbitrary D, E. According to part a) we see:

$$I_m[D \cup E] = I_m[E \setminus D] + I_m[D] + I_c[E \setminus D; D | \emptyset]$$

$$I_m[E] = I_m[E \setminus D] + I_m[D \cap E] + I_c[E \setminus D; D \cap E | \emptyset].$$

So, for the proof of (17) it suffices to prove the identity:

$$I_c[E \setminus D; D | \emptyset] = I_c[E \setminus D; D \setminus E | D \cap E] + I_c[E \setminus D; D \cap E | \emptyset].$$

We simply put $A = E \setminus D$, $B = D \setminus E$, $C = D \cap E$ in (15). ■

Consequence 2. Assuming (S), the function I_m: $\exp N \to \langle 0, \infty \rangle$ is convex (or supermodular), i.e. it holds:

$$I_m[D \cup E] + I_m[D \cap E] \geqq I_m[D] + I_m[E] \qquad \text{for each} \quad D, E \subset N. \tag{18}$$

Proof. (14) implies $I_c[E \setminus D; D \setminus E | D \cap E] \geqq 0$. We add $I_m[D] + I_m(E)$ to both sides and use (17). ■

So, Consequence 2 leads to the following question.

Problem 1. We know that, assuming (S), function I_m satisfies (12) and (18) ((11) follows from them). Can it be conversed? More precisely, whether these conditions on a function I: $\exp N \to \langle 0, \infty \rangle$ suffice for the existence of measurable spaces and probability measure described in (S) such that $I = I_m$.

The last consequence shows some information-theoretical significance of the conditional product of measures.

Consequence 3. Let $(X_i, \mathcal{X}_i)$, $i \in N$ be measurable spaces and $\{A, B, C\}$ some decomposition of N (finite, nonempty sets). Let μ and τ be consonant probability measures, μ on $(X_{A \cup C}, \mathcal{X}_{A \cup C})$, τ on $(X_{B \cup C}, \mathcal{X}_{B \cup C})$. Further, we denote

$$\Phi = \{P;\ P \text{ is a probability measure on } X_N,\ M(P) < \infty,\ P^{A \cup C} = \mu,\ P^{B \cup C} = \tau\}.$$

Then a) $\Phi \neq \emptyset$ iff $M(\mu) < \infty$ and $M(\tau) < \infty$.

b) Supposing $\Phi \neq \emptyset$, there exists the conditional product of μ and τ and minimizes the multiinformation on Φ.

Proof. If $\Phi \neq \emptyset$, then $M(\mu)$ and $M(\tau)$ are finite according to (11). Conversely, let $M(\mu)$ and $M(\tau)$ be finite. We deduce that $\mu \ll \prod_{i \in A} \mu^{(i)} \times \prod_{i \in C} \mu^{(i)}$ and by Proposition 1 and Remark 1a we see that c. p. on $(X_A, \mathscr{X}_A)$ w. r. to $(X_C, \mathscr{X}_C)$ is regular. So, Proposition 2 yields the existence of the conditional product P of μ and τ. Evidently, for this measure $I_c[A; B|C] = 0$ and, according to Proposition 4, it is $M(P) = M(\mu) + M(\tau) - M(\mu^C) < \infty$, so $P \in \Phi$ and $\Phi \neq \emptyset$. Moreover, by (17) and (14) applied to another $Q \in \Phi$ we deduce $M(Q) \geqq M(\mu) + M(\tau) - M(\mu^C) = M(P)$. ■

5. Application to the problem of characterization of CIR's

Definition 6. Assuming (S), we define a ternary relation $I(.,.|.)$ having as the domain all ordered triplets $\langle A, B, C \rangle$ of mutually disjoint subsets of N. If both A and B is nonempty, then $I(A; B|C)$ holds iff $P^{A \cup B \cup C}$ is the conditional product of $P^{A \cup C}$ and $P^{B \cup C}$ (for empty C it means $P^{A \cup B} = P^A \times P^B$). If A or B is empty, then we postulate that $I(A; B|C)$ holds. We shall call this relation the *conditional independence relation* corresponding to P and shall use the abbreviation CIR.

Note that CIR determines the conditional dependence relation D as its complementary relation (i.e. $D(A; B|C)$ holds iff $I(A; B|C)$ does not hold). Now, what is the problem of characterization of CIR's?

Problem 2. Let N be nonempty finite set. The problem is to find all independent properties (axioms) of a ternary relation I (defined on all ordered triplets of disjoint subsets of N) which together yield a necessary and sufficient condition for the existence of finite spaces X_i, $i \in N$ and of a probability measure P on $\prod_{i \in N} X_i$ such that I coincides with the CIR corresponding to P.

In this form the CIR was introduced by Pearl in [6] and his previous papers. But restricts to strictly positive measures. In the mentioned paper five properties of CIR's are formulated. The first one is the axiom of symmetry:

$$I(A; B|C) \Leftrightarrow I(B; A|C). \tag{A.1}$$

Three other axioms can be integrated into the following one:

$$I(A; B \cup C|D) \Leftrightarrow [I(A; B|C \cup D) \wedge I(A; C|D)]. \tag{A.2}$$

These two axioms hold without the assumption of strict positivity of the measure, while the last property:

$$[I(A; B|C\cup D) \wedge I(A; C|B\cup D)] \Rightarrow I(A; B\cup C|D) \tag{19}$$

does not so (i.e. it is not relevant to Problem 2).

Pearl expressed the completeness conjecture, i.e. (A.1), (A.2), (19) is the solution of Problem 2 modified by the demand that P must be strictly positive. The rest of Pearl's paper is concerned with graphical representations of probabilistic knowledge that are possible owing to (A.1), (A.2), (19).

The desired solution of Problem 2 seems to be significant in the theory of probabilistic expert systems. Let us mention the intensional expert system INES (see [7]). According to this approach, the knowledge base of an expert system is modelled by a multidimensional probability measure, while partial knowledges obtained from experts are described by means of less-dimensional probability measures which should be marginals of the mentioned multidimensional one. For capacity reasons it is usually impossible to store the multidimensional measure in the memory of a computer. This imperfection is solved by the help of so-called DSS's (dependence structure simplifications). These multidimensional measures are "formed successively as conditional products of given less-dimensional measures". So, we have to store only those in the memory. The choice of the DSS (i.e. of the order of making conditional products) is made from a certain information-theoretical point of view.

The solution of Problem 2 would make possible some improvement. Since the notion of conditional independence (or dependence) is easy to interpret we would be able to determine the proper structure of dependences and independences directly by asking experts. By means of the solution of Problem 2 we would be able to decide whether the statements of various experts are contradictory or whether there exists a probabilistic model having the requisite dependences and independences (i.e. there exists a CIR having prescribed dependences and independences).

Now, how to use the multiinformation? From Lemma 2 and Definitions 5, 6 it is easy to see:

Proposition 5. Assuming (S), it holds for disjoint A, B, $C \subset N$:

$$I(A; B|C) \text{ holds} \Leftrightarrow I_c[A; B|C] = 0. \tag{20}$$

Further, according to (17) we can express $I_c[A; B|C]$ by means of the function I_m (in Problem 2 X_i are finite, so I_m is finite). So, by this procedure we verify for disjoint A, B, C, D:

$$\left.\begin{aligned} &I_c[A; B|C\cup D] + I_c[C; D|A] + I_c[C; D|B] + I_c[A; B|\emptyset] = \\ &= I_c[C; D|A\cup B] + I_c[A; B|C] + I_c[A; B|D] + I_c[C; D|\emptyset]. \end{aligned}\right\} \tag{21}$$

Finally, from Proposition 5 we easily derive using (14):

$$\left.\begin{aligned}&[I(A;B|C\cup D]\wedge I(C;D|A)\wedge I(C;D|B)\wedge I(A;B|\emptyset)]\Leftrightarrow\\&\Leftrightarrow[I(C;D|A\cup B)\wedge I(A;B|C)\wedge I(A;B|D)\wedge I(C;D|\emptyset)].\end{aligned}\right\}\tag{A.3}$$

Example. We can take $N=\{a, b, c, d\}$ and construct a certain ternary relation as follows:

1. $I(a, b|cd)$, $I(c, d|a)$, $I(c, d|b)$, $I(a, b|\emptyset)$ and symmetric independences hold
2. $I(A; B|C)$ for empty A or B holds
3. no other independence holds.

The desired relation satisfies (A.1), (A.2), (19) but not (A.3).

So, using algebraic properties of multiinformation a new axiom (A.3) of CIR's was derived and Pearl's completeness conjecture was disproved. Note that (A.1), (A.2) can be derived similarly. Perhaps, it is possible to derive further axioms of CIR's analogously.

Nevertheless, I do not know the complete solution of Problem 2. I would like to ask readers for help. If somebody knows something relevant to this problem (maybe the solution is known since, for example, the theory of Markov fields meets with similar problems), I would like him (or her) to send me a reference or a reprint or any information. The similar wish concerns Problem 1.

References

1. *Csiszár, I.*, I-divergence geometry of probability distributions and minimization problems. Ann. Probab. **3** (1975), pp. 146–158.
2. *Добрушин, Р. Л.*, Общая формулировка основной теоремы Шеннона в теории информации. Успехи матем. наук. **14** (1959), *6*, с. 3–104. (In Russian, translation: Ann. Math. Soc. Translation **33**, *2*, pp. 323–438).
3. *Halmos, P. R.*, Measure Theory. D. v. Nostrand Comp. Inc., New York 1950.
4. *Neveu, J.*, Bases mathematiques du calcul des probabilités. Masson et Cie, Paris 1964 (in French, Russian translation: Мир, Москва 1969).
5. *Loéve, M.*, Probability Theory. D. v. Nostrand Comp. Inc., New York 1960.
6. *Pearl, J.*, Markov and Bayes networks. Technical Report CSD 860024, R-46-I, October 1986, University of California, Los Angeles.
7. *Perez, A., Jiroušek, R.*, Constructing an intensional expert system (INES). Medical Decision Making: diagnostic strategies and expert systems, North Holland 1985, pp. 307–315.
8. *Perez, A.*, ε-admissible simplifications of the dependence structure of a set random variables. Kybernetika **13** (1977), pp. 439–449.
9. *Perez, A.*, Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie des martingales. Trans. of the 1-st Prague Conference on Information Theory, Statistical Decision Function, Random Processes (1956), Prague 1957, pp. 183–208 (in French).

10. *Pinsker, M. S.*, Information and Information Stability of Random Variables and Processes. Translation from Russian by A. Feinstein, Holden-Day, San Francisco 1964. (The original: Информация и информационная устройчивость случайных величин и процессов, Академия наук СССР, Москва 1960).
11. *Сазонов, В. В.*, Ответ на один вопрос Р. Л. Добрушина. Теория вероятностей и ее применение **9** (1964), *1*, с. 180–181.
12. *Studený, M.*, Asymptotic behaviour of empirical multiinformation. Kybernetika **23** (1987), *2*, pp. 124–135.
13. *Zvárová, J.*, Informační míry statistické závislosti a výběrové vlastnosti zobecněné entropie řádu α. Thesis, Prague 1973 (in Czech).

Мультиинформация и проблема характеризации отношений условной независимости

М. СТУДЕНЫ

(Прага)

Установлено некоторое алгебраическое соотношение между мультиинформацией и условной информацией. Показано, что это соотношение применимо к проблеме характеризации отношений условной независимости, которая возникает в связи с вероятностными экспертными системами. Более конкретно выведена новая аксиома для этих отношений. Некоторые подготовительные результаты имеют самостоятельное значение: характеризация маргинально-непрерывных мер в теореме 1 и информационно-теоретическое значение условного произведения мер, упомянутое в следствии 3.

M. Studený
Ústav teorie informace a automatizace ČSAV
(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. 18 (1), pp. 17—26 (1989)

REMARKS ON A CLASS OF NONLINEAR MATRIX EQUATIONS AND ASSOCIATED STABLE TRANSFORMATIONS

PHAM T. NHU

(Hanoi)

(Received November 5, 1987)

This note is intended to explain the relationship between the following three results (Morozan 1979, Nhu 1981, Paksin 1982) for a class of nonlinear matrix equations arising from optimal control problems of discrete-time systems under Markov disturbances. Stable transformations associated with these equations will be also discussed and their applications are shown.

1. Introduction

The equations under consideration are of the form

$$P(i)=Q(i)+A^*(i)\sum_{j=1}^{S}P(j)p_{ij}A(i)-$$

$$-\left\{A^*(i)\sum_{j=1}^{S}P(j)p_{ij}B(i)\right\}\left\{R(i)+B^*(i)\sum_{j=1}^{S}P(j)p_{ij}B(i)\right\}^{-1}\cdot$$

$$\cdot\left\{B^*(i)\sum_{j=1}^{S}P(j)p_{ij}A(i)\right\},\qquad i=1,2,\ldots,S \tag{1}$$

where $P=(P(i),\ 1\leqq i\leqq S)$ is unknown, $P(i)$, $1\leqq i\leqq S$, are nonnegative definite $n\times n$-matrices, $Q(i)$ — nonnegative $n\times n$-matrices, $R(i)$ — positive $m\times m$-matrices, $A(i)$ — $n\times n$-matrices, $B(i)$ — $n\times m$-matrices and p_{ij}, $1\leqq i,j\leqq S$, are numbers such that $0\leqq p_{ij}\leqq 1$, $\sum_{j=1}^{S}p_{ij}=1$.

The class of such equations arises from the infinite horizon quadratic cost optimal control problem of the discrete-time linear system (see [1]–[4])

$$x_{t+1}=A(\eta_t)x_t+B(\eta_t)u_t,\quad x_0\in R^n,\qquad t=0,1,2,\ldots \tag{2}$$

where feedback controls are of the form $u_t=u_t(x_t,\ \eta_t)$ and $\{\eta_t,\ t=0,\ 1,\ 2,\ \ldots\}$ is a Markov chain with the set of states $I=1,\ 2,\ \ldots,\ S$ and the transition probabilities $\{p_{ij},\ 1\leqq i,j\leqq S\}$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Entering of the Markov disturbance in coefficients of dynamic equations (2), they describe abrupt changes of this system in a finite number of parameter values perhaps because of component failures or sudden shifts in environment. The continuous-time version of the problem with such sudden changes was first considered by N. N. Krasovskii and E. A. Lidskii, later by W. M. Wonham, D. D. Sworder, R. Rishel and others. References given in [5, 6] were detailed enough to enable necessary informations to the direction.

Since nonnegative definite solutions of (1) link with the design of admissible controllers yielding finite expected costs, and the existence and uniqueness of the positive definite solution link with an optimal controller, many forms of sufficient conditions seeming different were given (see [2]–[4]). The present paper discusses the relationship between them from the unified point of view of stable transformations (see [3]) associated with (1). Moreover, we show some classes of optimal control and stability problems to which stable transformations can advantageously be applied in order to derive some sufficient conditions for solving them.

We begin by a review of the results of the existence and uniqueness of the solution of (1) and then show relations between them. In the rest of the paper some comments on different forms of associated transformations are given and their applications are discussed.

2. Review

In the sequel, the following notations are used (see [3]).

Denote by $\mathscr{W}$ the set of all sequences of S $m \times n$-matrices $w=(W(i), 1 \leqq i \leqq S)$ and by $\mathscr{E}$ the Banach space of all sequences of S symmetric $n \times n$-matrices $P=(P(i), 1 \leqq i \leqq S)$ with the norm

$$|P| = \max_{1 \leqq i \leqq S} |P(i)| = \max_{1 \leqq i \leqq S} \left\{ \max_{\|x\| \leqq 1} \|P(i)x\| \right\}$$

where $|P(i)|$ denotes the spectral norm of the matrix $P(i)$, and $\|\cdot\|$ the Euclidean norm in R^n.

Definition 1. A linear transformation $\mathscr{B}: \mathscr{E} \mapsto \mathscr{E}$ is said to be stable if there exists $N>0$ such that $|\mathscr{B}^N|<1$, where $\mathscr{B}^N=\mathscr{B}(\mathscr{B}^{N-1})$, $N \geqq 2$, and $|\cdot|$ denotes the norm of a linear transformation from $\mathscr{E}$ into itself.

Recall that the norm of a linear transformation $\mathscr{L}$ is given by

$$|\mathscr{L}| = \sup_{|P| \leqq 1} |\mathscr{L}(P)|.$$

For stable transformations, the next lemma is used (see [3]).

Lemma 1. (a) If $\mathscr{B}$: $\mathscr{E} \mapsto \mathscr{E}$ is a linear and nondecreasing transformation, then $|\mathscr{B}| = |\mathscr{B}(I)|$ where $I = (I(i), 1 \leqq i \leqq S)$ and $I(i)$ is the identity $n \times n$-matrix.

(b) If $\mathscr{B}$: $\mathscr{E} \mapsto \mathscr{E}$ is a linear and nondecreasing transformation such that $P \geqq \mathscr{B}(P) + L$ for some $P \in \mathscr{E}$ and $L > 0$, then $\mathscr{B}$ is stable. Here $L = (L(i), 1 \leqq i \leqq S) > 0$ means $L(i) > 0$ for all $1 \leqq i \leqq S$.

(c) If $\mathscr{B}$: $\mathscr{E} \mapsto \mathscr{E}$ is a stable transformation then, for every $L \in \mathscr{E}$, equation $P = \mathscr{B}(P) + L$ has the unique solution given by the formula $P = \sum_{t=0}^{\infty} \mathscr{B}^t(L)$.

For given $w \in \mathscr{W}$, we associate with equations (1) the following transformations

$$G(w, \cdot) = (G_1(w, \cdot), G_2(w, \cdot), \ldots, G_S(w, \cdot)): \mathscr{E} \mapsto \mathscr{E}$$

$$F(w, \cdot) = (F_1(w, \cdot), F_2(w, \cdot), \ldots, F_S(w, \cdot)): \mathscr{E} \mapsto \mathscr{E}$$

$$\mathscr{A}(\cdot) = (\mathscr{A}_1(\cdot), \mathscr{A}_2(\cdot), \ldots, \mathscr{A}_S(\cdot)): \mathscr{K} \mapsto \mathscr{K}$$

where $\mathscr{K}$ is the cone of all sequences of S nonnegative definite $n \times n$-matrices,

$$\begin{aligned} G_i(w, P) &= (A(i) - B(i)W(i))^* \sum_{j=1}^{S} P(j)p_{ij}(A(i) - B(i)W(i)) \\ F_i(w, P) &= G_i(w, P) + W^*(i)R(i)W(i) + Q(i) \qquad (3) \\ \mathscr{A}_i(P) &= F_i(w(P), P) \\ w(P) &= (w_1(P), w_2(P), \ldots, w_S(P)) \end{aligned}$$

and

$$w_i(P) = \left\{ R(i) + B^*(i) \sum_{j=1}^{S} P(j)p_{ij}B(i) \right\}^{-1} \left\{ B^*(i) \sum_{j=1}^{S} P(j)p_{ij}A(i) \right\}. \qquad (4)$$

Then (1) becomes

$$P = \mathscr{A}(P). \qquad (5)$$

Definition 2. The system of random matrices $\{A(i), B(i), 1 \leqq i \leqq S\}$ is called to be observable if there exists $N \geqq 1$ such that $F^N(0) > 0$, where $F^N(w) = F(F^{N-1}(w))$ and $F(o, \cdot)$ is the transformation corresponding to $w = (0, 0, \ldots, 0)$ and $F(o, 0)$ is the image of the element $0 \in \mathscr{K}$ through the transformation $F(o, \cdot)$.

Lemma 2. If $Q(i) > 0$, $1 \leqq i \leqq S$, or $Q(i) \geqq 0$, $p_{ii} > 0$ and all pairs $(A^*(i), Q(i))$, $1 \leqq i \leqq S$, are completely controllable then the system $\{A(i), Q(i), 1 \leqq i \leqq S\}$ is observable (see [3]).

The results of [2, 3, 4] can be summarized as follows.

Theorem 1 (Morozan 1979). If, for every $i = 1, 2, \ldots, S$, the pair $(A(i), B(i))$ is stabilizable, then there exists $\gamma > 0$ such that if for every $i = 1, 2, \ldots, S$, $1 - p_{ii} < \gamma$ then system (1) has a solution $P = (P(i), 1 \leqq i \leqq S$ with $P(i) \geqq 0$, $1 \leqq i \leqq S$.

Theorem 2 (Nhu 1981). If there exists $w \in \mathscr{W}$ such that $G(w, \cdot)$ is stable then there exists a nonnegative definite solution of (1).

Moreover, equations (1) has a nonnegative definite solution if and only if there exists $w \in \mathscr{W}$ such that the equation $P = F(w, P)$ has a nonnegative definite solution.

Specially, if the system $\{A(i), Q(i), 1 \leqq i \leqq S\}$ is observable then the above assertions are equivalent and in this case, the solution of (1) is unique and positive definite.

Theorem 3 (Paksin 1982). If all pairs $(A(i), B(i))$, $1 \leqq i \leqq S$, are stabilizable, $(\sqrt{Q(i)}, A(i))$, $1 \leqq i \leqq S$, are observable and for all $i = 1, 2, \ldots, S$

$$\inf_{W(i)} \left| (1-p_{ii}) \sum_{k=1}^{\infty} (A(i)-B(i)W(i))^* k_{p_{ii}}^{k-1} (A(i)-B(i)W(i))^k \right| < 1 \tag{6}$$

then there exists the unique positive definite solution of (1).

Remark 1. In Theorem 3, the conditions $p_{ii} > 0$, $1 \leqq i \leqq S$, should be assumed. The following example demonstrates that all assumptions of Theorem 3 are not sufficient for the existence of the unique positive definite solution.

Example. Consider a Markov chain with the set of states $I = \{1, 2\}$ and transition probabilities $p_{11} = p_{12} = \frac{1}{2}$, $p_{21} = 1$, $p_{22} = 0$. The equations of (1) for $P(1)$ and $P(2)$ are

$$P(1) = \frac{1}{2} A^*(i)P(1)A(1) + \frac{1}{2} A^*(1)P(2)A(1) + Q(1)$$

$$P(2) = \quad A^*(2)P(1)A(2) + \quad Q(2)$$

where

$$A(1) = A(2) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad Q(1) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad Q(2) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

All assumptions of Theorem 3 hold, e.g. for $i = 1$, $A^2(1) = 0$, $p_{11} = \frac{1}{2}$,

$$\operatorname{Inf}_{W(1)} \left| \frac{1}{2} \sum_{k=1}^{\infty} (A^*(1))^k \left(\frac{1}{2}\right)^{k-1} A^k(1) \right| = \frac{1}{2} \left| \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right| = \frac{1}{2} < 1$$

but the solution is

$$P(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix} > 0, \qquad P(2) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} \geqq 0.$$

3. Relations

Lemma 3. (a) Theorem 1 can be regarded as a special case of Theorem 2 and it concerns with the sequence $\{\mathscr{A}^N(0), N \geqq 1\}$ converging from below to the minimal solution of (1) (see [3]).

(b) Theorem 3 can be regarded as a special case of Theorem 2 and it concerns with the sequence $\{P_N,\ N \geqq 1\}$ converging from above to the unique solution of (1), where $P_{N+1} = F(w(P_N), P_N)$ and $w(P_N)$ is given by (4) (see [3]).

Furthermore, in this case Theorem 2 gives us a stronger result: for any $P_0 \geqq 0$, the sequence $\{\mathscr{A}^N(P_0),\ N \geqq 1\}$ converges geometrically fast to the unique solution of (1) (see [3]).

Proof. Assertion (a) has been shown in [3].

We now give a sketch of the proof of Assertion (b) with $p_{ii} > 0$, $1 \leqq i \leqq S$ (see Remark 1).

Denote $I_1 = \{i \in I \mid p_{ii} = 1\}$.

From the stabilizability assumption of $(A(i), B(i))$ for $i \in I_1$ and (6) for $i \in I \backslash I_1$ it follows that there exists $\bar{w} = (\bar{W}(i),\ 1 \leqq i \leqq S)$ such that the linear and nondecreasing transformation

$$\mathscr{G} = \{\mathscr{G}_i(\bar{w}, \cdot),\ 1 \leqq i \leqq S\},$$

$$\mathscr{G}_i(\bar{w}, P) = \sum_{k=1}^{\infty} (A(i) - B(i)\bar{W}(i))^{*k} p_{ii}^{k-1} \sum_{j=1, j \neq i}^{S} P(j) p_{ij} (A(i) - B(i)\bar{W}(i))^k \tag{7}$$

is well defined and $|\mathscr{G}_i(\bar{w}, \cdot)| = |\mathscr{G}_i(\bar{w}, I)| < 1$ (see Lemma 1).

Thus, $\mathscr{G}$ is stable and the equation

$$P(i) = \mathscr{G}_i(\bar{w}, P) + \sum_{k=0}^{\infty} (\bar{A}^*(i))^k p_{ii}^{k-1} \bar{Q}(i) \bar{A}^k(i) \tag{8}$$

where

$$\bar{Q}(i) = \bar{W}^*(i) R(i) \bar{W}(i) + Q(i), \quad \bar{A}(i) = A(i) - B(i)\bar{W}(i), \quad 1 \leqq i \leqq S,$$

has a nonnegative definite solution P.

We show that $(\bar{w}, \bar{P})$ is also a solution of the equation $P = F(w, P)$ in Theorem 2.

$$\begin{aligned}
F_i(\bar{w}, \bar{P}) = G_i(\bar{w}, \bar{P}) + \bar{Q}(i) &= \bar{A}^*(i) p_{ii} \bar{P}(i) \bar{A}(i) + \\
&+ \bar{A}^*(i) \sum_{j=1, j \neq i}^{S} \bar{P}(j) p_{ij} \bar{A}(i) + \bar{Q}(i) = \\
&= \bar{A}^*(i) p_{ii} \left[\sum_{k=1}^{\infty} (\bar{A}^*(i))^k p_{ii}^{k-1} \sum_{j=1, j \neq i}^{S} P(j) p_{ij} \bar{A}^k(i) \right] \bar{A}(i) + \\
&+ \bar{A}^*(i) p_{ii} \left[\sum_{k=0}^{\infty} \bar{A}^*(i))^k p_{ii}^{k-1} \bar{Q}(i) \bar{A}^k(i) \right] \bar{A}(i) + \\
&+ \bar{A}^*(i) \sum_{j=1, j \neq i}^{S} P(j) p_{ij} \bar{A}(i) + \bar{Q}(i) = \\
&= \mathscr{G}_i(\bar{w}, \bar{P}) + \sum_{k=0}^{\infty} (\bar{A}^*(i))^k p_{ii}^{k-1} \bar{Q}(i) \bar{A}^k(i) = \bar{P}(i).
\end{aligned}$$

From the second assertion of Theorem 2 it follows that (1) has a nonnegative definite solution. Then Lemma 2, Remark 1, and the third assertion of Theorem 2 imply the assertion of Theorem 3. The proof is complete.

Remark 2. From the above proof (see (7)–(8)) we note that a sufficient condition for the existence of a nonnegative definite solution of (8) is that $(A(i), B(i))$, $i \in I_1$, are stabilizable and condition (6) holds for all $i \in I \setminus I_1$. Thus the stabilizability assumption and (6) of Theorem 3 can be weakened to this sufficient condition. Moreover, this note suggests the possibility of decomposing of I into $I_1, I_2, \ldots, I_m$ for which forms of transformations G and $\mathscr{G}$ can differ from one another and they act in the corresponding Banach subspaces, e.g. in $\mathscr{E}_m = \{P = (P(i), i \in I_m) \mid P(i) = P^*(i)\}$.

4. Comments

The transformations G and $\mathscr{G}$ (if $\mathscr{G}$ is well defined) arose from the optimal control problem in the class $\mathscr{U}_1$ of all admissible feedback strategies $u_t = u_t(x_t, \eta_t)$ the design of which needs complete observations of η_t until the control moment t. As noted in [3], the problem can be regarded as a special case of the optimal control problem in the class $\mathscr{U}_2$ consisting of all feedback strategies $u_t = u_t(x_t, \eta_{t-1})$. For the latter problem, equations (1) and the transformations G, $\mathscr{G}$ are replaced by

$$P(i) = Q(i) + \sum_{j=1}^{S} A^*(j)P(j)p_{ij}A(j) - \left\{ \sum_{j=1}^{S} A^*(j)P(j)p_{ij}B(j) \right\} \cdot$$

$$\cdot \left\{ R(i) + \sum_{j=1}^{S} B^*(j)P(j)p_{ij}B(j) \right\}^{-1} \left\{ \sum_{j=1}^{S} B^*(j)P(j)p_{ij}A(j) \right\}$$

$$i = 1, 2, \ldots, S, \tag{9}$$

$$\tilde{G}(w, \cdot) = \{\tilde{G}_i(w, \cdot),\ 1 \leqq i \leqq S\},$$

$$\tilde{G}_i(w, P) = \sum_{j=1}^{S} (A(j) - B(j)W(i))^* P(j)p_{ij}(A(j) - B(j)W(i)),$$

$$\tilde{\mathscr{G}} = \{\tilde{\mathscr{G}}_i(w, \cdot),\ 1 \leqq i \leqq S\},$$

$$\tilde{\mathscr{G}}_i(w, P) = \sum_{k=1}^{\infty} (A(i) - B(i)W(i))^{*k} p_{ii}^{k-1} \cdot$$

$$\cdot \sum_{j=1, j \neq i}^{S} (A(j) - B(j)W(i))^* P(j)p_{ij}(A(j) -$$

$$- B(j)W(i))(A(i) - B(i)W(i))^k. \tag{10}$$

Theorem 2 corresponding to this case was described in [3] and by an argument analogous to that used for the proof of Assertion (b) of Lemma 3 we have (see also Remarks 1–2)

Corollary 1. Assume that $0<p_{ii}\leqq 1$, all pairs $(A(i), B(i))$ with $p_{ii}=1$ are stabilizable, and for every i with $p_{ii}<1$

$$\operatorname*{Inf}_{W(i)}\left|\sum_{k=1}^{\infty}(A(i)-B(i)W(i))^k p_{ii}^{k-1}\cdot\right.$$

$$\left.\cdot\sum_{j=1,j\neq i}^{S}(A(j)-B(j)W(i))^* p_{ij}(A(j)-B(j)W(i))(A(i)-B(i)W(i))^k\right|<1.$$

If $(\sqrt{Q(i)}, A(i))$ is observable for every $i=1, 2, \ldots, S$ then there exists the unique positive definite solution of (9).

So far we have concerned ourselves only with transformations G and $\tilde{G}$ associated with the optimal control problems. We now specialize them to the exponential in mean square stability problem of linear system

$$x_{t+1}=A(\eta_t)x_t$$

$$x_0\in R^n, \qquad t=0, 1, 2, \ldots \tag{10}$$

In this case, transformations G and $\tilde{G}$ with $B(i)=0$, $1\leqq i\leqq S$, are denoted by $H(\cdot)=(H_i(\cdot), 1\leqq i\leqq S$ and $K(\cdot)=(K_i(\cdot), 1\leqq i\leqq S$, where

$$H_i(P)=A^*(i)\sum_{j=1}^{S}P(j)p_{ij}A(i), \tag{11}$$

$$K_i(P)=\sum_{j=1}^{S}A^*(j)P(j)p_{ij}A(j). \tag{12}$$

The relation between them and the stability problem is shown by the following theorem.

Theorem 4. System (10) is exponentially in mean square stable if and only if Transformation H is stable, or equivalently, if and only if Transformation K is stable.

Proof. We begin by establishing the first assertion of the theorem.

If H is stable then there exist $\beta>0$ and $0<\alpha<1$ such that

$$|H^k|<\beta\cdot\alpha^k, \qquad k=1, 2, \ldots \tag{13}$$

On the other hand, from (10) it follows by induction

$$E[x_{t+1}^* x_{t+1}|\eta_0]=x_0^* H_{\eta_0}^t(I)x_0 \tag{14}$$

where E denotes the conditional expectation. Hence (13) implies

$$E[x_{t+1}^* x_{t+1}|\eta_0]\leqq|H_{\eta_0}^t(I)|\,\|x_0\|^2\leqq\beta\alpha^t\|x_0\|^2.$$

Conversely, if system (10) is exponentially in mean square stable, then there exist $\beta > 0$ and $0 < \alpha < 1$ such that (14) implies the last inequality. Hence $|H^t(I)| < \beta \cdot \alpha^t$.

Since Transformation H is linear and nondecreasing we have (see Lemma 1) $|H^t| = \max_{1 \leqq i \leqq S} |H_i^t(I)| < \beta \cdot \alpha^t$. Thus H is stable and the first assertion is proved.

The second part of the theorem follows from the next lemma.

Lemma 4. Transformation H is stable if and only if Transformation K is stable.

Proof of Lemma 4. If H is stable then from Lemma 1 follows that there exists a nonnegative solution $\bar{P}_I = (\bar{P}_I(i), 1 \leqq i \leqq S)$ of the equation

$$P = H(P) + I, \tag{15}$$

that means

$$\bar{P}_I(i) = A^*(i) \sum_{j=1}^{S} \bar{P}_I(j) p_{ij}, A(i) + I(i), 1 \leqq i \leqq S,$$

where $I = (I(i), 1 \leqq i \leqq S)$ and $I(i)$ is the identity $n \times n$-matrix.

Hence $P_I(i) = \sum_{j=1}^{S} P_I(j) p_{ij}$, $1 \leqq i \leqq S$, is a nonnegative definite solution of the equation

$$P = K(P) + I. \tag{16}$$

By Lemma 1, K is stable.

Conversely, if $P_I = (P_I(i), 1 \leqq i \leqq S)$ is a nonnegative definite solution of (16) then $\bar{P}_I = (\bar{P}_I(i), 1 \leqq i \leqq S)$ with

$$\bar{P}_I(i) = I(i) + A^*(i) P_I(i) A(i)$$

is a positive definite solution of (15). Again by Lemma 1, Transformation H is stable. The proof of Lemma 4 is complete and this concludes the proof of Theorem 4.

Remark that the final assertion of Theorem 4 is a special case ($B(i) = 0$) of Theorem 5 in [3] and comparing it with the corresponding results in [7], Theorem 4 gives lightly weakened conditions for the exponential stability of (10).

For making the usefulness of Theorem 4 more apparent, some sufficient conditions for the stability of (10) will be given.

Corollary 2. System (10) is exponentially in mean square stable if one of the following conditions is fulfilled

(i) $|A^*(i)A(i)| < 1, \quad 1 \leqq i \leqq S,$

(ii) $\left| \sum_{j=1}^{S} A^*(j)A(j)p_{ij} \right| < 1, \quad 1 \leqq i \leqq S,$

(iii) $\left| A^*(i) \sum_{j=1}^{S} A^*(j)A(j)p_{ij}A(i) \right| < 1, \quad 1 \leqq i \leqq S,$

(iv) $\left| \sum_{j=1}^{S} A^*(k) \left[\sum_{j=1}^{S} A^*(j)A(j)p_{kj} \right] A(k)p_{ik} \right| < 1, \quad 1 \leqq i \leqq S,$

(v) $|H_i^N(I)| < 1, \quad 1 \leqq i \leqq S,$

(vi) $|K_i^N(I)| < 1, \quad 1 \leqq i \leqq S,$

where $I = (I(i), 1 \leqq i \leqq S)$, $I(i)$ is the identity $n \times n$-matrix, N is some natural number, $H_i^N(I) = H_i(H^{N-1}(I))$ and $K_i^N(I) = K_i(K^{N-1}(I))$.

Proof. Since Transformations H and K are linear and nondecreasing, Lemma 1 gives $|H^N| = |H^N(I)|$, $|K^N| = |K^N(I)|$ for every $N \geqq 1$. Hence if one of the above conditions is satisfied, then Transformations H and K are stable and the corollary follows from applying Theorem 4. The proof is complete.

In the conclusion, we emphasize that introduced transformations give us an efficient tool for the analysis of the stability and control of linear systems under Markov disturbances.

Acknowledgement

The author wishes to thank Prof. L. Györfi and the referees for the benefit of their advices.

References

1. *Blair, W. P., Sworder, D. D.*, Feedback control of class of linear discrete time systems with jump parameters and quadratic cost criteria, Int. J. Control, **21** (1975).
2. *Morozan, T.*, Stochastic stability and control for discrete time systems with jump Markov disturbances, Rev. Roum. Math. pures et Appl., **24,** *1* (1979).
3. *Nhu, Pham T.*, On a system of matrix Riccati equations with applications to optimal stabilization under stochastic disturbances, Rev. Roum. Math. pures et Appl., **26,** *7* (1981).
4. *Paksin, P. V.*, Optimal linear control of discrete time systems with random jump parameters, Problems of Control and Information Theory, **11,** *3* (1982).
5. *Wonham, W. M.*, Random differential equations in control theory, Probabilistic methods in applied mathematics, A. T. Bharucha-Reid Ed., Chapt. 2, Academic Press, New York, 1970.
6. *Rishel, R.*, Dynamic programming and minimum principles for systems with jump Markov disturbances, SIAM J. Control, **13,** *2* (1975).
7. *Morozan, T.*, Stability and control of some linear discrete time systems with jump Markov disturbances, Rev. Roum. Math. pures et Appl., **26,** *1* (1981).

Замечание о классе нелинейных алгебраических уравнений матриц и устойчивых отображениях, связанных с ними

ПХАМ Т. НХУ
(Ханой)

Заметка является попыткой объяснения взаимосвязи между теоремами существования Пакшина, Морозана и Нью, полученными для класса нелинейных уравнений матриц, возникающих в проблемах оптимального управления дискретными системами при возмущениях в виде Марковских цепей. Приведены устойчивые отображения, связанные с этими нелинейными уравнениями, и их применения.

Pham T. Nhu
Institute of Computer Science and Cybernetics
Hanoi, Badinh Lieugiai
Vietnam

Problems of Control and Information Theory, Vol. 18 (1), pp. 27—41 (1989)

ON MINIMAX STATE ESTIMATES FOR MULTISTAGE STATISTICALLY UNCERTAIN SYSTEMS

B. I. ANAN'EV

(Sverdlovsk)

(Received November 9, 1987)

A minimax state estimation problem for nonlinear multistage systems both being subjected to random disturbances and containing uncertain parameters is considered. A class of admissible nonlinear estimates is defined, and the optimal minimax item is determined among them. A necessary optimality condition being also sufficient under some assumption is obtained. With sufficiency the optimal estimate is uniquely defined in the class of admissible ones. A finite-dimensional approximation of the solution is examined. Some special cases are noted, and model examples are given.

1. Introduction

The minimax (or game-theoretic, guaranteed) approach got the wide spreading when state estimation problems for dynamic systems with incomplete information about parameters and disturbance distributions were solved [1–11]. In most of papers of the mentioned direction, linear systems are considered. Among the investigations dealing with nonlinear multistage systems we indicate, for example [6, 8].

It is known [1, 2] that the problem of estimation (or observation) of coordinates of the state vector of the dynamic system under incomplete or inexact measurements occupy one of the central place in the theory of control. The minimax approach [10] to the estimation problem for systems with uncertain values got started almost simultaneously with the appearance of the well-known work [12] on stochastic filtering. The importance of minimax estimation methods can be motivated by the fact that in many practical problems the detailed information touching upon the dynamics of the process and its statistics turns out to be inaccessible. Furthermore, it is necessary to examine processes being under the action of unknown for an observer parameters, controls or disturbances. In these cases, it is meaningful to assume that the unknown parameters can be formed on the base of any accessible information including the knowledge of the observation structure with the aim to maximize the estimation error. In this connection the observer's information of uncertain values becomes frequently exhausted with the knowledge of limitations on its possible magnitudes.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

The statistically uncertain situation in the estimating may arise in the case when, for example, the information vector contains both fast changing (random) errors and slowly changing ones [13]. Sometimes, the last errors can be simulated by more or less complicated forming filters and after that one may apply the statistical methods. However in that case, the minimax approach will be the less artificial one in which the slowly changing errors belonging to the given constraints are considered to be uncertain. Note that, when solving the real estimation problems by means of a computer, it is necessary to simulate differential equations with the help of multistage systems. Therefore, on the stage of the information treating and the integration of the differential equations some additional errors of the electronic apparatus, approximation method errors and so on will take place, as a rule, along with the statistically uncertain disturbances having the mechanical nature (measurement devices errors, uncertain controls). Furthermore, the parameters of statistical distributions of that additional errors may be both known and unknown. The above considerations lead us to the necessity to solve the estimation problem under the joint examination of uncertain and random disturbances in multistage systems.

In this work, the methods of [6, 8] are developed for general nonlinear multistage systems being subjected to both random disturbances and uncertain ones. One defines an admissible nonlinear estimation class in which the optimal minimax estimates are specified, and defining relations for them are given. Problems of finite-dimensional approximation of the solution are considered and two particular cases are pointed out: when the uncertain parameters are absent and when the random ones do as well. In the first case the solution is reduced to known results of the stochastic theory of filtering [14] and in the second one it is reduced to known results of the theory of guaranteed estimation [2]. The results are illustrated by model examples.

2. Preliminary problem formulation

Consider the multistage system

$$x^t = f^t(x^{t-1}, \theta^t, \xi^t), \qquad t \geqq 1,$$

$$x^0 = f^0(\theta^0, \xi^0), \tag{2.1}$$

and the equation of observation:

$$y^t = g^t(x^t, \theta^t, \eta^t), \qquad t \geqq 1, \tag{2.2}$$

where $x^t \in R^n$, $y^t \in R^m$. The elements of the sequences $\{\xi^t\}$, $\{\eta^t\}$ represent independent random vector values with known distributions. The uncertain vectorial parameter

θ^t satisfy the a priori inclusion

$$\theta^t \in \Theta_t, \qquad t \geqq 0, \tag{2.3}$$

where Θ_t is a compact set infinite-dimensional Euclidean space. The union of all uncertain parameters up to the instant t, denoted by $\theta_{0t}=(\theta^0, \ldots, \theta^t)$, will comply with the inclusion

$$\theta_{0t} \in \Theta_{0t} \tag{2.4}$$

where Θ_{0t} is a corresponding bounded set (the Cartesian product of compact sets from (2.3)). The vector-functions $f^t(\cdot,\cdot,\cdot)$, $g^t(\cdot,\cdot,\cdot)$ in equations (2.1), (2.2) are assumed to be continuous.

The multistage system (2.1), (2.2) is transformed to the one-step system for further purposes. To this end, the following notations are introduced:

$$\begin{aligned} x&=[x^{1'} \vdots \ldots \vdots x^{t'}]' \in R^{nt}, \qquad \xi=[\xi^{1'} \vdots \ldots \vdots \xi^{t'}]', \\ y&=[y^{1'} \vdots \ldots \vdots y^{t'}]' \in R^{mt}, \qquad \eta=[\eta^{1'} \vdots \ldots \vdots \eta^{t'}]', \end{aligned} \tag{2.5}$$

where sign $'$ means the transposition. Then one has

$$\begin{aligned} x&=F(x^0, \theta_{1t}, \xi), \\ y&=G(x, \theta_{1t}, \eta), \end{aligned} \tag{2.6}$$

and also

$$x^t=\pi_t x, \qquad \pi_t=[0 \vdots \ldots \vdots I_n] \in R^{n \times nt}. \tag{2.7}$$

From now on, I_n is the unit $n \times n$-matrix, $R^{n \times nt}$ is the space of $n \times nt$-matrices.

In equations (2.6), $F(\cdot,\cdot,\cdot)$ and $G(\cdot,\cdot,\cdot)$ are the continuous vector-functions constructed according to (2.1), (2.2).

Consider the class Δ of Borelian mappings $\delta(\cdot): R^{mt} \to R^n$ for which

$$\sup_{\theta_{0t}} E\|\delta(y)\|^2 < \infty. \tag{2.8}$$

From now on, $\|\cdot\|$ is the Euclidean norm, E is the mathematical expectation. Note that the class Δ represents the linear space.

The problem under consideration in this paper has a form

$$I(\delta(\cdot)) = \sup_{\theta_{0t}} E\|\pi_t x - \delta(y)\|^2 \to \min_{\delta(\cdot)}. \tag{2.9}$$

Besides of the papers listed above, problems of that type are studied in a monograph [15], in [16], and elsewhere. One difference of this work from the preceding papers consists in using of a non-compact class of admissible decision functions [17]. This has entailed a new proof of the existence of (2.7).

The examination of the class Δ determined by inequality (2.8) may be inconvenient from technical point of view. Therefore, some subclass $\tilde{\Delta} \subset \Delta$ in which problem (2.9) with random parameters will be solved is defined below. The additional conditions on the probability distributions of the values x, y (see (2.6)) will be also given below.

3. A class of admissible nonlinear estimates

Fix the instant t, and omit the subscripts for the parameters θ_{0t} from the inclusion (2.4). Let P and Q be measures both defined on one and the same space. We will write $P \ll Q$ if the measure P is absolutely continuous with respect to Q [14]. The notation $P \sim Q$ will mean that $P \ll Q$ and $Q \ll P$ simultaneously.

The following assumptions are taken.

Assumptions 3.1. 1) Let $P_\theta \ll \nu$ for some σ-finite measure ν on R^{mt} for $\forall \theta \in \Theta$ where P_θ is the distribution of the random vector y from (2.6). In addition, for the density function we require that

$$dP_\theta(y)/d\nu(y) = p(y, \theta) \leqq \bar{p}(y) \qquad (\nu \text{ — a.e.}) \tag{3.1}$$

where $\bar{p}(\cdot) \in L_1(\nu)$.

2) The value $\max_\theta \|x^t\|^2$ has the finite mathematical expectation. The density function $p(y, \theta)$ is continuous in θ for vectors y, $\bar{P}$-almost everywhere where $d\bar{P}(y) = \bar{p}(y)d\nu(y)$.

3) There exists the variant* of conditional mean $E[x^t | y, \theta]$ satisfying the inequality

$$\|E[x^t | y, \theta]\| \leqq \bar{f}(y) \qquad (\bar{P} \text{ — a.e.}) \tag{3.2}$$

where $\bar{f}(\cdot) \in L_2(\bar{P})$. The variant $E[x^t | y, \theta]$ is continuous in θ on the set $\{\theta \in \Theta : p(y, \theta) > 0\}$ for vectors y $\bar{P}$ — a.e.

Remark 3.1. Assumptions 3.1 are automatically fulfilled in case $E\|x^t\|^2 < \infty$ when the indeterminacy in θ is absent for equations (2.1), (2.2).

Problem (2.9) will be solved in the class $\tilde{\Delta}$ of Borelian mappings belonging to the space

$$L_2^n(\bar{P}) \subset \Delta \tag{3.3}$$

where the measure $\bar{P}$ is defined in point 2) of assumptions 3.1. Estimates $\delta(\cdot)$ from the class $\tilde{\Delta}$ will be called admissible ones.

* The definition of the variant of the conditional mean is given in [18, ch. 2, def. 44, 46].

Let us introduce the notation

$$K(\delta(\cdot), \theta) = E_\theta \|x^t - \delta(y)\|^2. \tag{3.4}$$

Here and further, E_θ is the symbol of the mathematical expectation for the known value θ.

The following assertion is valid.

Lemma 3.1. Let assumptions 3.1 be fulfilled. Then function (3.4) is continuous in θ for the arbitrary fixed $\delta(\cdot) \in \tilde{\Delta}$ and is convex in $\delta(\cdot)$ for the arbitrary fixed $\theta \in \Theta$.

Proof. One needs to justify only the continuity of the functional (3.4) in θ.

By formula (3.4), applying the basic properties of conditional mathematical expectations, we get

$$K(\delta(\cdot), \theta) = E_\theta \|x^t\|^2 - 2 \int_{R^{mt}} \delta'(y) E[x^t | y, \theta] \cdot$$

$$\cdot p(y, \theta) dv(y) + E_\theta \|\delta(y)\|^2. \tag{3.5}$$

By virtue of assumptions 3.1, the functions continuous in θ are situated under the integral sign in expression (3.5). Consequently, by the Lebesque theorem on majorized convergence, functional (3.4) will be continuous in θ.

4. A problem solution in the class of admissible estimates

From now on, we suppose assumptions 3.1 to be fulfilled. Let Λ be the set of all probability measures on the compact set Θ. Denote the averaging of function (3.4) over a measure λ by a symbol

$$\tilde{K}(\delta, \lambda) = \int_\Theta K(\delta(\cdot), \theta) d\lambda(\theta). \tag{4.1}$$

For the functional $I(\cdot)$ of problem (2.9), the equality

$$I(\delta(\cdot)) = \max_{\theta \in \Theta} K(\delta(\cdot), \theta) = \max_{\lambda \in \Lambda} \tilde{K}(\delta, \lambda) \tag{4.2}$$

is valid. Hence, a solution to problem (2.9) in the class $\tilde{\Delta}$ (see (3.3)) is equivalent to the determination of a saddle point for functional (4.1). Indeed, providing the set with the topology of the weak convergence of measures converts that set into the metric compact one [15]. Then the minimax theorem [16] gives the equality

$$\inf_{\delta(\cdot)} I(\delta(\cdot)) = \inf_\delta \max_\lambda K(\delta, \lambda) = \max_\lambda \inf_\delta \tilde{K}(\delta, \lambda). \tag{4.3}$$

Further it is shown that the minimum of functional (4.1) over δ for a fixed λ is equal to

$$\min_{\delta} \tilde{K}(\delta, \lambda) = \tilde{K}(\delta_\lambda, \lambda) = \int_{\Theta} (E_\theta \| x^t \|^2 - E_\theta \| \delta_\lambda(y) \|^2) d\lambda(\theta) \tag{4.4}$$

where the symbol

$$\delta_\lambda(y) = \int_{\Theta} E[x^t | y, \theta] p(y, \theta) d\lambda(\theta) / \int_{\Theta} p(y, \theta) d\lambda(\theta) \tag{4.5}$$

stands for the estimate ensuring that minimum. For vectors $y \notin Y_\lambda$, where

$$Y_\lambda = \{y : \int_{\Theta} p(y, \theta) d\lambda(\theta) > 0\}, \tag{4.6}$$

estimate (4.5) is determined arbitrarily. We shall assume that $\delta_\lambda(y) = 0$ in that case.

The following lemma is true.

Lemma 4.1. Estimate (4.5) is admissible for any measure $\lambda \in \Lambda$.

The proof of that assertion follows from the easily verified inclusion

$$\delta_\lambda(\cdot) \in S (\forall \lambda \in \Lambda)$$

where

$$S = \{\delta(\cdot) : \| \delta(y) \| \leqq \bar{f}(y) \qquad (\bar{P} \text{ — a.e.})\} \tag{4.7}$$

is a weak compact set in the space (3.3).

For further considerations, let us define the probability measure P_λ:

$$dP_\lambda(y) = (\int_{\Theta} p(y, \theta) d\lambda(\theta)) dv(y). \tag{4.8}$$

Note that $P_\lambda(Y_\lambda) = 1 (\forall \lambda \in \Lambda)$ and also $P_\lambda \sim \bar{P}$ on the set Y_λ.

Summarize the above considerations into a theorem.

Theorem 4.1. A solution to problem (2.9) in the class of admissible estimates is equivalent to the determination of a saddle point for functional (4.1). The saddle point denoted as $(\bar{\delta}, \bar{\lambda})$ exists, and also $\bar{\delta}(y) = \delta_{\bar{\lambda}}(y)$ ($P_{\bar{\lambda}}$-almost sure), $\| \bar{\delta}(y) \| \leqq \bar{f}(y)$ ($\bar{P}$ — a.e.). Here $\delta_{\bar{\lambda}}(\cdot)$ is the element (4.5) for $\lambda = \bar{\lambda}$, and $\bar{f}(\cdot)$ is the function from the inequality (3.2). The measure $\lambda \in \Lambda$ characterizing the worst possible distribution of uncertain parameters maximizes the concave in λ functional (4.4) and has a property: $\bar{\lambda}(\bar{\Theta}) = 1$, where the set

$$\bar{\Theta} = \{\theta \in \Theta : K(\bar{\delta}(\cdot), \theta) = I(\bar{\delta}(\cdot))\}. \tag{4.9}$$

Proof. Functional (4.1) is convex in $\delta(\cdot) \in \tilde{\Delta}$ and linear in $\lambda \in \Lambda$. In addition, it is continuous in λ and weak lower semi-continuous in $\delta(\cdot)$ in the space (3.3). Equality (4.3) follows from the minimax theorem [16] since Λ is compact. The greatest lower

bound of weak lower semi-continuous functional (4.2) is reached on the set (4.7) in view of the inclusion $\delta_\lambda(\cdot) \in S(\forall \lambda \in \Lambda)$. Thus, the saddle point for functional (4.1) does exist. Then we integrate quantity (3.5) over the measure λ and change the position of integrals according to Fubini's theorem. As the result, the equality

$$\tilde{K}((\delta, \lambda) = \int\limits_{R^{mt}} \|\delta(y) - \delta_\lambda(y)\|^2 dP_\lambda(y) + K(\delta_\lambda, \lambda)$$

is obtained in which the corresponding values are defined by relations (4.4), (4.5), and (4.6). From the last equality the other assertions of the theorem are derived except for the latter one. The assumption $\bar{\lambda}(\bar{\Theta}) < 1$ leads us to contradiction, because $\tilde{K}(\bar{\delta}, \bar{\lambda}) < I(\bar{\delta}(\cdot))$ in that case. But the last inequality is impossible. Q.E.D.

Definition 4.1. A measure $\lambda \in \Lambda$ is called to be nondegenerate if $\bar{P} \sim P_\lambda$ (see (4.8)) (equivalently: $\bar{P}(Y_\lambda) = \bar{P}(R^{mt})$ where Y_λ is the set (4.6)).

Corollary of Theorem 4.1. Let exist the non-degenerate measure $\bar{\lambda} \in \Lambda$ maximizing functional (4.4). Then the equality

$$\bar{\delta}(y) = \delta_{\bar{\lambda}}(y) \qquad (P_{\bar{\lambda}} \text{ — a.e.}) \tag{4.10}$$

represents the necessary and sufficient condition for the case that the estimate $\bar{\delta}(\cdot) \in \tilde{\Delta}$ provides the minimum for problem (2.9). The solution to that problem is unique (mod $\bar{P}$) under the condition of the given corollary.

A computation of the maximum of functional (4.4) over the measures is a rather complicated mathematical problem. Therefore, questions concerning a finite-dimensional approximation of the problem are considered below.

5. A finite-dimensional approximation of the problem solution

First, let us establish properties of continuity for functional (4.4) and estimate (4.5).

Lemma 5.1. The concave in λ functional (4.4) is continuous in the sense of the weak measure convergence in the set Λ. If $\lambda_k \to \lambda$ in Λ, then the sequence of the estimates $\delta_{\lambda_k}(y)$ converge to $\delta_\lambda(y)$ on the set (4.6) $\bar{P}$ — a.e.

Note that functional (4.1) is continuous in the joint variables if the convergence of estimates δ is regarded in the strong topology of the space (3.3) and the convergence of measures λ is regarded as weak convergence.

Let $\lambda_k \to \lambda$ in the set Λ. The numerator and the denominator of expression (4.5) are continuous in λ on a subset of the full $\bar{P}$-measure. Besides, the inclusion $\delta_{\lambda_k}(\cdot) \in S$ (see (4.7)) holds. Consequently, the assertions of Lemma 5.1 are valid.

Consider the set $\Lambda_n \subset \Lambda$ of all probability measures concentrated on the finite n^{-1}-net of the compact set Θ, $n=1, 2, \ldots$. A sequence of n^{-1}-nets is selected in this way that the previous n^{-1}-net is contained in the next $(n+1)^{-1}$-net. Then one has the inclusion $\Lambda_1 \subset \Lambda_2 \subset \ldots$. Further, a sequence of n^{-1}-nets constructed as above will be mentioned as being an increasing one.

Lemma 5.2. The set $\bigcup_{n=1}^{\infty} \Lambda_n$ is weak dense in Λ for any increasing sequence of n^{-1}-nets.

The proof of that assertion is omitted.

Let $\{\bar{\lambda}_n\}$ be a sequence of measures maximizing functional (4.4) on the set Λ_k. At least, one such sequence does exist, because Λ_n is the finite-dimensional compact set for $\forall n$. Then it follows from Lemmas 5.1 and 5.2 that

$$d_n = \tilde{K}(\delta_{\bar{\lambda}_n}, \bar{\lambda}_n) \uparrow \max \{\tilde{K}(\delta_\lambda, \lambda) : \lambda \in \Lambda\} \tag{5.1}$$

when $n \to \infty$, i.e. the sequence of measures $\{\bar{\lambda}_n\}$ is maximizing for functional (4.4) on the set Λ.

The mentioned arguments lead us to the following assertion.

Theorem 5.1. Let $(\bar{\delta}_n, \bar{\lambda}_n)$ be a sequence $(\bar{\lambda}_n \in \Lambda_n)$ of a saddle point being computed for $\delta \in \tilde{\Delta}$, $\lambda \in \Lambda_n$, for functional (4.4). Suppose that there exists an increasing sequence of n^{-1}-nets for which the corresponding maximizing measures sequences $\{\bar{\lambda}_n\}$ having at least one non-degenerate limit point $\bar{\lambda} \in \Lambda$ may be found. Then, if $\bar{\lambda}_{n(k)} \to \bar{\lambda}$ when $k \to \infty$, we have

$$\lim_{k \to \infty} \bar{\delta}_{n(k)}(y) = \delta(y) \qquad (\bar{P} \text{ — a.e.}) \tag{5.2}$$

where $\delta(\cdot)$ is the estimate (4.10) providing the unique minimum in problem (2.9). In addition, for $\forall \varepsilon > 0$, $\exists N$ such that for $\forall k > N$ the pair $(\bar{\delta}_{n(k)}, \bar{\lambda}_{n(k)})$ forms an ε-saddle point for functional (4.4), i.e.

$$\begin{aligned} -\varepsilon + \tilde{K}(\bar{\delta}_{n(k)}, \lambda) \leqq \tilde{K}(\bar{\delta}_{n(k)}, \bar{\lambda}_{n(k)}) \leqq \\ \leqq \tilde{K}(\delta, \bar{\lambda}_{n(k)}), \qquad \forall \delta \in \tilde{\Delta}, \quad \forall \lambda \in \Lambda. \end{aligned} \tag{5.3}$$

Proof. Let $\bar{\lambda}_{n(k)} \to \bar{\lambda}$ where $\bar{\lambda}$ is a non-degenerate measure. Then $\bar{\lambda}$ is a maximizing measure by virtue of relation (5.1). It follows from Lemma 5.1 that the sequence of estimates $\delta_{\bar{\lambda}_{n(k)}}(\cdot)$ converges to $\delta_{\bar{\lambda}}(\cdot) = \bar{\delta}(\cdot)$ $\bar{P}$ — a.e. From here, equality (5.2) is deduced, since the estimate $\bar{\delta}_{n(k)}(\cdot)$ coincides (mod $\bar{P}$) with $\delta_{\bar{\lambda}_{n(k)}}(\cdot)$ on the set $Y_{\bar{\lambda}_{n(k)}}$. Assuming that (5.3) does not hold, we arrive to the contradiction by virtue of the continuity of corresponding functions, Lemma 5.2, and the compactness of the set Λ. Q.E.D.

Corollary of Theorem 5.1. Under conditions of the theorem the equality

$$\lim_{k\to\infty} I(\bar{\delta}_{n(k)}(\cdot)) = I(\bar{\delta}(\cdot)) = \min_{\delta(\cdot)\in\bar{\Delta}} I(\delta(\cdot)) \tag{5.4}$$

holds.

Remark 5.1. If the density function $p(y,\theta)>0$ for y $\bar{P}$-almost everywhere and for $\forall\,\theta\in\bar{\Theta}$ (see (4.9)), then any maximizing measure $\bar{\lambda}\in\Lambda$ will be non-degenerate according to definition 4.1. Therefore, the conditions of Theorem 5.1 and that of the corollary of the theorem will be fulfilled. Besides, in Theorem 5.1, an arbitrary increasing sequence of n^{-1}-nets can be taken as an approximating one. In this connection, equality (5.2) and relation (5.3) will hold for any subsequences, in particular, for the sequence $\{\bar{\delta}_n(\cdot)\}$.

6. The case of linear-Gaussian systems with uncertain parameters

Let equations (2.1), (2.2) be the following form

$$x^t = A_t(\theta^t)x^{t-1} + b^t(\theta^t) + \sigma_{1t}(\theta^t)\xi^t, \quad t\geq 0, \tag{6.1}$$

$$x^{-1} = 0,$$

$$y^t = C_t(\theta^t)x^t + d^t(\theta^t) + \sigma_t(\theta^t)\eta^t, \quad t\geq 1, \tag{6.2}$$

where the matrix continuous functions $A_t(\cdot)$, $C_t(\cdot)$, $\sigma_{1t}(\cdot)$, $\sigma_t(\cdot)$, have the appropriate dimensions. Vectorial functions $b^t(\cdot)$, $d^t(\cdot)$ are also supposed to be continuous. Elements of the sequences $\{\xi^t\}$, $\{\eta^t\}$ are independent Gaussian vectorial values with zero means and unit covariance matrices.

Further the value $E[x^t|y,\theta]$ will be denoted by $\hat{x}^t(\theta)$. It is known from relations of the Kalman filter [14] that the last value is recurrently defined under the non-degeneracy condition

$$\sigma_t(\theta)\sigma_t'(\theta)\geq \varkappa I_m, \quad \forall\, t\geq 1, \quad \forall\,\theta\in\Theta_t, \quad \varkappa>0. \tag{6.3}$$

Thus, the value $\hat{x}^t(\theta)$ may be assumed to be linear in y and to be continuous in θ on the compact set Θ. Besides, one may assert that the inequality

$$\|\hat{x}^t(\theta)\|\leq a + b\|y\| = \bar{f}(y), \tag{6.4}$$

where $a, b>0$, is fulfilled. The vectorial value y has in the given case a non-degenerate Gaussian distribution with the density function

$$p(y,\theta) = N(y-\bar{y},F) = (2\pi)^{-mt/2}(\det F)^{-1/2}\cdot$$

$$\cdot\exp\left(-(y-\bar{y})'F^{-1}(y-\bar{y})/2\right) \tag{6.5}$$

where

$$\bar{y} = E_\theta y \in R^{mt}, \quad F = E_\theta[(y-\bar{y})(y-\bar{y})'] \in R^{mt \times mt}.$$

The inequality

$$N(y-\bar{y}, F) \leqq kN(y, 4dI_{mt}) = \bar{p}(y),$$

$$d = \max_\theta \|F\|, \tag{6.6}$$

is valid for the density function (6.5).

From relations (6.4–(6.6), one may conclude that assumptions 3.1 (where $dv(y) = dy$ now) are fulfilled in the given case.

Note that the class $\tilde{\Delta}$ (see (3.3)) of admissible estimates contains here, in particular, all Borelian functions admitting a polynomial growth at infinity. The estimate (4.5) will be here continuous in y and λ function having the form

$$\delta_\lambda(y) = \int_\Theta \hat{x}^t(\theta) N(y-\bar{y}, F) d\lambda(\theta) / \int_\Theta N(y-\bar{y}, F) d\lambda(\theta). \tag{6.7}$$

In view of inequality (6.4) function (6.7) has in y a linear growth at infinity.

An approximation of solution (4.10) can be realized according to Remark 5.1 and Theorem 5.1 with the help of any increasing sequence of n^{-1}-nets in the compact set Θ.

7. The case of determinate systems with uncertain parameters

Let the functions f^t, g^t in equations (2.1), (2.2) do not depend on the random values ξ^t, η^t. Then the system becomes determinate, and the information on uncertain parameters is, as before, exhausted by the inclusion (2.3), (2.4). In the given case, the space Δ of bounded Borelian mappings $\delta(\cdot)$: $R^{mt} \to R^n$ with the norm

$$\|\delta(\cdot)\|_\infty = \sup_{\theta \in \Theta} \|\delta(y)\| \tag{7.1}$$

is chosen as being a class of admissible non-linear estimates according to inequality (2.8).

In the indicated class, it is necessary to solve problem (2.9) that takes now the form

$$I(\delta(\cdot)) = \sup_{\theta \in \Theta} \|x^t - \delta(y)\|^2 \to \min_{\delta(\cdot) \in \Delta}. \tag{7.2}$$

Below the Chebyshev center, i.e. the vector

$$\bar{\delta}(y)=\arg\min\{\max\{\|z-x\|:x\in X_t(y)\}:z\in R^n\},\tag{7.3}$$

of the informational set $X_t(y)$ in the *a posteriori* estimation problem for the determinate system (2.1), (2.2) is denoted by the symbol $\bar{\delta}(y)$ as in (7.3). The corresponding notions are defined in monograph [2] and in [6].

The following assertion is true.

Theorem 7.1. Function (7.3) belongs to the space Δ and provides the minimum for problem (7.2). Besides, for any function $\delta_1(\cdot)\in\Delta$ with the property $I(\delta_1(\cdot))=I(\bar{\delta}(\cdot))$, the equality $\delta_1(y(\theta^*))=\bar{\delta}(y(\theta^*))$ is fulfilled at any point $\theta^*\in\Theta$ for which

$$\|x^t(\theta^*)-\bar{\delta}(y(\theta^*))\|^2=I(\bar{\delta}(\cdot)).$$

Note that, if the sign of max in (4.2) is substituted by the sign of sup, then equality (4.2) is valid yet. But equality (4.3) is invalid, generally speaking. However, the following special result takes place.

Theorem 7.2. Let the compact set Θ be a finite set of points. Then vector (7.3) is represented by the formula $\bar{\delta}(y)=E_{\bar{\lambda}}[x^t|y]$ where $\bar{\lambda}$ is a probability measure on Θ providing the maximum over Λ for the concave in λ functional $E_\lambda\|x^t-E_\lambda[x^t|y]\|^2$. Here E_λ is the mathematical expectation corresponding to the distribution λ on Θ. A saddle point of functional (4.1) for examined problem (7.2) is formed by the pair $(\bar{\delta},\bar{\lambda})$.

A proof of Theorem 7.1 is developed by contradiction, and Theorem 7.2 follows from the above-stated considerations.

8. Examples

1. Let the one-dimensional equations

$$\begin{aligned}x&=v+\sigma_0\xi,\\ y&=x+\sigma\eta\end{aligned}\tag{8.1}$$

be given where v is an uncertain value constrained by the inequality $|v|\leqq 1$. Numbers σ_0, σ are fixed, and ξ, η are standard independent Gaussian values with zero means and unit dispersion.

In order to determine the optimal non-linear estimate (4.10), by means of formulae (6.4)–(6.7) one finds $\bar{y}=v$, $F=\sigma^2+\sigma_0^2$,

$$\delta_\lambda(y)=F^{-1}(\sigma_0^2 y+\sigma^2\int_{-1}^{1}v\exp(-(y-v)^2/(2F))\cdot$$

$$\cdot\,d\lambda(v)/\int_{-1}^{1}\exp(-(y-v)^2/(2F))d\lambda(v).\tag{8.2}$$

Further, it is necessary to find the maximum over measures for the functional of type (4.4) and to substitute the optimal measure into expression (8.2). The problem of maximization of functional (4.4) can be numerically solved with the help of Theorem 5.1. Note that the optimal minimax estimate is non-linear in y here.

2. Consider equations (8.1) where an uncertain parameter v can take only two values: $+1$ or -1. The values $\sigma_0\xi$ and $\sigma\eta$ will be assumed to be uniformly distributed and independent with the same density function

$$p(x)=\begin{cases}1/2, & |X|\leqq 1,\\ 0, & |X|>1.\end{cases}$$

Then the density function of value y with respect to Lebesgue measure is equal to

$$p(y,v)=\begin{cases}(2-|y-v|)/4, & |y-v|\leqq 2,\\ 0, & |y-v|>2,\end{cases}$$

under given v. The conditional mathematical expectation $E[x|y,v]=(y+v)/2$. Formula (4.5) takes the form

$$\delta_\lambda(y)=\begin{cases}(y+1)/2, & 1\leqq y\leqq 3;\\ y+(2\lambda-1+y)/(1+(2\lambda-1)y)/2, & |y|\leqq 1;\\ (y-1)/2, & -3\leqq y\leqq -1.\end{cases}$$

Here the number $\lambda\in(0,\,1)$ is the weight of the measure $\lambda(dv)$ at the point 1. The measure $\bar{\lambda}(dv)$ having the same weights at the points $+1$ and -1 (i.e. $\lambda=1/2$) provides the maximum for functional (4.4). Such measure $\bar{\lambda}$ is non-degenerate in the sense of definition 3.1. Consequently, equality (4.10) for the given example takes the form

$$\bar{\delta}(y)=\delta_{\bar{\lambda}}(y)=\begin{cases}(y+1)/2, & 1\leqq y\leqq 3;\\ y, & |y|\leqq 1;\\ (y-1)/2, & -3\leqq y\leqq -1.\end{cases}\tag{8.3}$$

In addition, the value of functional $I(\cdot)$ is equal to 1/4. Note that the optimal in the class of linear operations estimate

$$\bar{\delta}_1(y)=4y/S$$

ensures the value of functional $I(\bar{\delta}_1(\cdot))=4/15>1/4$.

If it is assumed for this example that the uncertain parameter v fills the entire segment $[-1,\,1]$, then the form of the optimal linear estimate and the value of its

functional are not changed. However, estimate (8.3) gives already a better result that, in turn, may be improve yet if functional (4.4) is optimized over all probability measures on $[-1, 1]$.

9. Discussion of results

First, it should be noted that the optimal minimax estimate of type (4.10) where $\bar{\lambda}$ is a measure maximizing functional (4.4) depends essentially on the known distributions of values ξ^t, η^t in equations (2.1), (2.2). This estimate is robust being giving the guaranteed result over a class of distributions known to within the parameter θ, of values x, y (see (2.6)). In principle, the above stated considerations turn out to be correct for an arbitrary metric compact set Θ.

Second, we shall indicate that expression (4.5) contains the conditional mean $E[x^t|y, \theta]$ representing the best estimate of vector x^t with respect to observations $y^1, \ldots, y^t$ under known parameters θ. The computation of the noted estimate is a rather complicated problem constituting the subject of the mathematical theory of stochastic filtering [14]. The proposed approach permits to take into account developed approximate methods for the determination of the conditional mean when optimal minimax estimates are evaluated.

Third, note that procedures of the recurrent recalculation which are not discussed in this paper are considered for linear systems in [3–6, 9, 11]. Finally, it should be said that the stated results are similar to a certain extent to the general theory of statistical decisions [17].

References

1. *Krasovskii, N. N.*, Theory of control of the motion. Moscow, Nauka, 1968 (in Russian).
2. *Kurzhanskii, A. B.*, Control and observation under uncertainty. Moscow, Nauka, 1977 (in Russian).
3. *Kats, I. Ya., Kurzhanskii, A. B.*, Minimax estimation for multistage systems. Dokl. Akad. Nauk SSSR, 1975, **221,** *3*, pp. 535–538 (in Russian).
4. *Bublik, B. N., Kirichenko, N. F., Nakonechnyi, A. G.*, Minimax estimates and regulators for dynamic systems. Prepr. Kiev Univ., *31*, 1978, pp. 1–50 (in Russian).
5. *Pshenichnyi, B. N., Pokotilo, V. G.*, On problems of observation for discrete systems. Prikl. Mat. i Mekh., 1981, **45,** *1*, pp. 3–10 (in Russian).
6. *Koshcheev, A. S., Kurzhanskii, A. B.*, Adaptive estimation of evolution of multistage systems under uncertainty. Izv. Akad. Nauk SSSR, Tekh. Kibernetika, 1983, **2,** pp. 72–93 (in Russian).
7. *Bakhshiyan, B. Ts., Nazirov, R. R., El'yasberg, P. Ye.*, Determination and correction of the motion. Moscow, Nauka, 1980 (in Russian).
8. *Anan'ev, B. I., Kurzhanskii, A. B.*, The nonlinear filtering problem for a multistage system with statistical uncertainty. In: Prepr. of Second IFAC symp. on stochast. control, Vilnius, USSR, 1986, Pt. 1, pp. 205–210.

9. *Martin, C. J., Mintz, M.*, Robust filtering and prediction for linear systems with uncertain dynamics: A game-theoretic approach. IEEE Trans. Automat. Control, 1983, **AC-28,** *9*, pp. 888–896.
10. *Krasovskii, N. N.*, On the theory of control lability and observability for linear dynamic systems. Prikl. Mat. i Mekh., 1964, **28,** *1*, pp. 3–14 (in Russian).
11. *Anan'ev, B. I.*, Minimax mean-square estimates for statistically uncertain systems. Different. Uravn., 1984, **20,** *8*, pp. 1291–1297 (in Russian).
12. *Kalman, R. E.*, A new approach to linear filtering and prediction problems. J. Basic Eng. ASME Trans., 1960, **82D,** pp. 35–45.
13. *Boguslavskii, I. A.*, Applied problems of filtering and control. Moscow, Nauka, 1983 (in Russian).
14. *Liptser, R. Sh., Shiryayev, A. N.*, Statistics of random processes. I, II. New York, Springer-Verlag, 1978.
15. *Repin, V. G., Tartakovskii, G. P.*, Statistical design under a priori uncertainty. M.: Sovetskoye Radio, 1977 (in Russian).
16. *Beloglazov, I. N.*, Recurrent searching algorithms of estimation. Dokl. Akad. Nauk SSSR, 1977, **236,** *2*, pp. 292–295 (in Russian).
17. *Wald, A.*, Statistical decision functions. New York, J. Wiley & Sons, 1950.
18. *Meyer, P.*, Probability and potentials. Blaisdell, 1966.
19. *Billingsley, P.*, Convergence of probability measures. New York, J. Wiley & Sons, 1968.
20. *Fan Ky*, Minimax theorems. Proc. Nat. Acad. Sci. USA, 1953, **39,** *1*, pp. 42–47.

О минимаксных оценках состояния многошаговых статистически неопределенных систем

Б. И. АНАНЬЕВ

(Свердловск)

Рассматривается минимаксная задача оценивания фазового состояния для нелинейных многошаговых систем вида (2.1), подверженных воздействию как случайных, так и неопределенных возмущений. Предполагается, что процесс наблюдения описывается уравнением (2.2), в котором также могут содержаться неопределенные параметры. Задачи подобного рода возникают в различных областях техники при рассмотрении процессов, находящихся под воздействием неизвестных наблюдателю параметров, управлений или возмущений.

Для выбора оптимальных минимаксных оценок определяется класс допустимых нелинейных оценок, удовлетворяющих неравенству (2.8) при дополнительных предположениях 3.1. После этого максимум функционала по неопределенным параметрам в выражении (2.9) заменяется максимумом по вероятностным мерам, сосредоточенным на априорно заданном компакте. В результате из теоремы о минимаксе с учетом сделанных предположений выводится теорема 4.1, в которой установлено существование седловой точки для функционала (4.1) и получены необходимые условия оптимальности (4.5), (4.9). Если максимизирующая мера в задаче является невырожденной в смысле определения 4.1, то равенство (4.10) представляет собой необходимое и достаточное условие оптимальности, а оптимальная минимаксная оценка в этом случае определяется единственным образом с точностью до эквивалентности.

Далее рассматриваются вопросы конечномерной аппроксимации задачи. А именно, установлена теорема 5.1, в которой при условиях невырожденности доказана возможность аппроксимации оптимального решения при помощи элементов (4.5) с мерой, сосредоточенной в конечном множестве точек. Кроме того, установлены соотношения (5.2)–(5.4).

Разобраны два частных случая общего решения: для линейно-гауссовских систем вида (6.1), (6.2) и для детерминированных систем с неопределенными параметрами. В первом случае оптимальная оценка вида (6.7) имеет линейый по измерениям рост на бесконечности и совпадает при отсутствии неопределенных параметров с известной оценкой условного среднего. Во втором

случае полученное решение сводится к известным результатам теории гарантированного оценивания. Результаты иллюстрируются на модельных примерах.

Полученные оптимальные оценки являются робастными, дающими гарантированный результат на классе известных с точностью до параметров распределений векторных величин (2.6). Предлагаемый подход позволяет учитывать разработанные приближенные методы определения условного среднего при нахождении оптимаьных минимаксных оценок.

Б. И. Ананьев
Институт математики и механики
Уральского отделения АН СССР,
СССР 620219, г. Свердловск, ГСП-384, ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 18 (1), pp. 43—54 (1989)

SYNTHESIS OF FAULT-TOLERANT DYNAMIC CONTROL SYSTEMS WITH FAULT IDENTIFICATION

V. I. SALYGA, I. B. SIRODGA, A. S. KULIK, V. L. OBRUCHEV

(*Moscow, Kharkov*)

(Received September 24, 1987)

The problem of construction of fault-tolerant control systems is examined here. An approach to the synthesis of fault-tolerant control systems based on hardware algorithm method of diagnosis criterion realization in relation to predetermined set of faults, and on application of the method of data classification processing for fault identification and for the control of algorithm and hardware redundancy with the aim of the system serviceability restoration are proposed.

1. Introduction

Continuous increase of the functional capabilities of automatic systems and the rise of demands for the quality of their operation require the development of new conceptions and methods of fault-tolerant control system designing. Familiar methods in the theory of self-organizing and self-adjusting control systems [1, 2] are based on the principle of compensation of the resulting disturbances via estimation of some resulting quality factor of control (consequence) rather than the disturbance factor (reason) itself. This results in a set of essential limitations connected with undesirable construction of the system's adaptive capability depending on the place, size and character of the disturbances, with the impossibility of disturbances compensation resulting in structural changes of the control object, and also with unjustified energy consumption of the system supply. The adaptive system construction is based upon direct or indirect use of methods of model identification describing the current technical condition of the control object [3], with the purpose of developing evaluation values of the chosen model class coefficient. Here, as the information about the place, value and character of disturbances is hidden in the coefficients of the identification models and can very seldom be produced in explicit form. The principle of direct compensation (counteraction) of disturbing fault action which resulted in disruption of the correct system operation is proposed for construction of fault-tolerant control system with the help of timely fault identification. The problem of the fault identification lies in the synthesis of diagnostic procedure giving the possibility to determine timely the place, value and character of the fault on the basis of data

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

measured. Its counteraction is performed according to the results of fault identification. The problem of counteraction lies in the synthesis for various types of faults of compensation algorithms consisting in signal and parametric retraining for compensable fault and in connecting a sound element instead of a failed one for uncompensable fault.

As a result of structuring of the problem of the fault-tolerant control the following generalized structure has been obtained (Fig. 1). The traditional task of control of the first level is carried out by combination of the control object and the controller

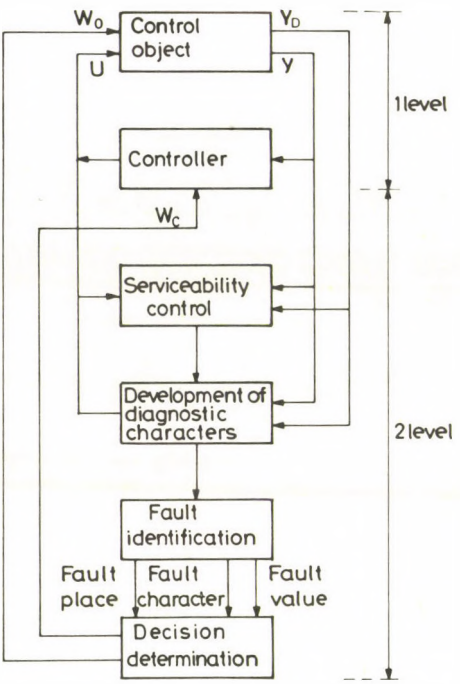


Fig. 1. Generalized two-level control structure

(regulator). The aim of the second level control lies in detection of all the offsets of the control process in relation to calculated regime, the fault identification, and in their compensation and counteraction. At the second level the following problems are being solved. 1. Serviceability supervision of the original control system. 2. Development of diagnostic character. Hereat the characters which are unambiguously related to the fault characters in all on-line modes of the control system are being developed. 3. Timely discrimination of fault location, value and type on the basis of

the characters developed. 4. Making decision on the fault compensation and counteraction. Hereat the control actions for changing structure, regulator parameters — W_c and the control actions — W_0 for disconnection of the defective equipment and connection of the redundant equipment are being developed depending on the fault.

In this paper an approach to the treatment of the second level problem providing fault-tolerant control of the object is described.

2. Problem statement

A linearized control system dynamics of the first level is described by the equations

$$\dot{x}(t) = A(\alpha)x(t) + B(\alpha)U(t), \qquad y(t) = x(t) + \xi(t), \tag{1}$$

where $x(t)$ is n-dimensional state vector; $U(t)$ is r-dimensional vector of control actions; $y(t)$ is n-dimensional vector of measurements; $\xi(t)$ is n-dimensional vector of normal interference measurements; $A(\alpha)$ and $B(\alpha)$, respectively, $(n \times n)$ and $(n \times r)$, are coefficient-dimensional matrices depending on the diagnostic parameter vector

$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_q], \qquad \alpha \in \Omega, \tag{2}$$

where Ω is a set of tolerable values of diagnostic parameters. The moment of inertia, the moment of resistance, capacitance, inductance and other moments characterizing possible physical failures of the control system are used as diagnostic parameters.

Poisson arrival of gradual failures which result in a relatively slow change of diagnostic parameters in the period of the fault identification is supposed to be present in the system.

Control system faults result in occurrence of measurement vector deviation in relation to nominal: $\Delta y(t) = y(t) - y_N(t)$. The task lies in the detection of deviation, identification of the fault which caused this deviation, compensation or counteraction of the fault in order to maintain the minimum of the following function

$$I[\Delta y(t), \sigma(t)] = \Delta y^T(t) Q \Delta y(t) + \int_0^{\sigma(t)} f(\lambda) d\lambda, \tag{3}$$

where Q is a symmetrical positive matrix; $f(\lambda)$ is the characteristics of the regulator serviceability recovery; $\sigma(t)$ is the regulator control signal.

3. System diagnosis provision

The necessary condition for fault-intolerance is the original control system diagnostic which provides the principal possibility of the fault identification on the results of the system signal measurements. The application of the diagnosis criterion [6, 10] gives the possibility to perform the directional control system decomposition into diagnostic components and to formulate the measurement tractable technical condition vector $y^T_{TC}(t)=[y(t), y_D(t)]$, where $y_D(t)$ is the additional measurement vector. The basis of the diagnosis criterion is the diagnostic model of the first approximation representing linearized mathematical description unambiguously relating the system diagnostic character $\Delta y(t)$ to the fault character $\Delta\alpha_i$ ($\Delta\alpha_i=\alpha_i-\alpha_{iN}$, α_{iN} is the parameter nominal value, $i=\overline{1, q}$). The diagnostic model is formally formulated by application of the sensitivity function of the system model. Thus for i-fault the diagnostic model can be written as

$$\Delta y_i(t)=S_i(t)\Delta\alpha_i+\xi(t), \tag{4}$$

$$\dot{S}_i(t)=A(\alpha_N)S_i(t)+[\mathscr{A}_i\mathscr{B}_i]\begin{bmatrix}x(t)\\U(t)\end{bmatrix}=A(\alpha_N)S_i(t)+C_iv(t), \tag{5}$$

where $S_i(t)$ is the sensitivity function for i-fault;

$$\mathscr{A}_i=\partial A(\alpha)/\partial\alpha_i|_{\alpha=\alpha_N}, \quad \mathscr{B}_i=\partial B(\alpha)/\partial\alpha_i|_{\alpha=\alpha_N},$$

are the sensitivity coefficient matrices for i-fault.

The control system will be fully the subject to diagnosis at the interval $[t_0, t_1]$ in relation to Poisson arrival faults if the location, value and type of the fault at the time moment t_0 can be unambiguously identified from the values of the system diagnostic characters $\Delta y(t)$ at the interval $[t_0, t_1]$. According to the criterion for the complete system diagnostic it is necessary and sufficient for matrices C_i and C_j in all pairwise combinations to be linearly independent. We assume that there is always a control action vector providing complete system diagnosis. If among the pairwise combinations occurs a linearly dependent one then the system will not be the subject to diagnosis in relation to the faults corresponding to the matrices. In this case it is necessary to take special measures including organization of additional direct and indirect measurements $y_D(t)$ or the control system structure change.

Example 1. Let us analyse the system of a space apparatus stabilization from the point of view of its roll channel and estimate its diagnostic. The system consists of the apparatus itself with the inertia moment I, jet nozzles with transmission coefficient K_I, angle and angular velocity transducers correspondingly to the transmission coefficient K_φ and $K_{\dot{\varphi}}$ and the governor with the transmission coefficient K_{p_1} and K_{p_2}. In order to simplify the calculations the stabilization system will be

examined when accurate measurements are available, then it can be written as

$$\begin{bmatrix} \dot{U}_\varphi \\ \dot{U}_{\dot\varphi} \end{bmatrix} = \begin{bmatrix} 0 & \frac{K_\varphi}{K_{\dot\varphi}} \\ \frac{K_{\dot\varphi} K_{p_1} K_I}{I} & \frac{K_{\dot\varphi} K_{p_2} K_I}{I} \end{bmatrix} \begin{bmatrix} U_\varphi \\ U_{\dot\varphi} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{K_{\dot\varphi} K_{p_1} K_I}{I} \end{bmatrix} U_3,$$

where U_φ and $U_{\dot\varphi}$ are the voltages taken off the transducers correspondingly to the angle and angular velocity; U_3 is the voltage of the unit. The system assumes five faults characterizing the following diagnostic parameters: K_φ, $K_{\dot\varphi}$, K_I, K_{p_1} and K_{p_2}. Compound matrices of the sensitivity coefficients will be shown as

$$C_{K_\varphi} = \begin{bmatrix} 0 & \frac{1}{K_{\dot\varphi}} & 0 \\ 0 & 0 & 0 \end{bmatrix}; \quad C_{K_{\dot\varphi}} = \begin{bmatrix} 0 & -\frac{K_\varphi}{K_{\dot\varphi}} & 0 \\ \frac{K_{p_1} K_I}{I} & \frac{K_{p_2} K_I}{I} & \frac{K_{p_1} K_I}{I} \end{bmatrix},$$

$$C_{K_I} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{K_{\dot\varphi} K_{p_1}}{I} & \frac{K_{\dot\varphi} K_{p_2}}{I} & \frac{K_{\dot\varphi} K_{p_1}}{I} \end{bmatrix}, \quad C_{K_{p_1}} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{K_{\dot\varphi} K_I}{I} & 0 & \frac{K_{\dot\varphi} K_I}{I} \end{bmatrix},$$

$$C_{K_{p_2}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{K_{\dot\varphi} K_I}{I} & 0 \end{bmatrix}.$$

The number of pairwise combinations is equal to 10. As among the matrices there is not a single pair with identical structure all the combinations are linearly independent and the system is fully the subject of diagnosis in relation to predetermined faults.

4. Serviceability supervision

Serviceability supervision of the control system is a familiar problem which has effective solutions [4, 8]. The supervision method with the standard model is widely used. System equation (1) is used as a standard model at nominal parameter characters, then

$$\dot{y}_N(t) = A(\alpha_N) y_N(t) + B(\alpha_N) U(t). \tag{6}$$

With no-fault system (1) the result of the comparison is

$$\Delta y(t) = y(t) - y_N(t) = \xi(t).$$

As a condition for serviceability evaluation a two-valued predicate equation is used

$$(c - |\hat{M}[\Delta y(t)]|) \geqq 0 = \begin{cases} 1, & \text{if the system is serviceable,} \\ 0, & \text{if the system is unserviceable} \end{cases} \tag{7}$$

where c is the vector of acceptable accuracy of the control system operation; $\hat{M}[\cdot]$ is the mathematical expectation estimation.

When various malfunctions result in the situation where $|\hat{M}[\Delta y(t)]| > c$ then the predicate equation takes 0-value which corresponds to the unserviceable system.

5. Diagnostic character development

Control system decomposition gives the possibility to proceed with the solution of the problem of diagnostic character development. All the familiar parametric methods of functional diagnosis of dynamic systems [4, 5] are connected with explicit estimation of diagnostic parameter variations. Hereat closed and open identification procedures [3] are used, their practical application requiring a large size of storage and time interval. The application of diagnosis models gives the possibility to develop rather simple algorithms of character formation for further classification processing. For the development of estimating value $\Delta\hat{\alpha}_i$ from equation (4) it is possible to use the method of the least squares, then

$$\Delta\hat{\alpha}_i = [S_i^T S_i]^{-1} S_i^T \Delta y_i, \tag{8}$$

where S_i is the arrays of the selective values of the system diagnosis character; T is the symbol of transposition. Equation (8) allows to develop the estimation values of the fault character by small-scale arrays.

The above estimation is developed from the conditions of application of diagnosis character Δy_i corresponding to i-fault. As the moment of the fault appearance and its value are random events, only diagnostic character $\Delta y(t)$ showing the presence of some malfunction from the predetermined set Ω can be used in the estimating equation. In order to develop estimating values of all fault characters from the predetermined set it is necessary to use the following equation

$$\Delta\hat{\alpha}_i = [S_i^T S_i]^{-1} S_i^T \Delta y, \qquad i = \overline{1, q}, \tag{9}$$

which develops the system from q equations.

For the fault identification it is possible to use the following estimate of the system diagnostic characters

$$\Delta\hat{y}(t)=\hat{M}[\Delta y(t)], \tag{10}$$

which are applied for the system serviceability supervision.

The space of diagnostic characters is developed on the basis of sensitivity functions $S_i(t)$, $i=\overline{1, q}$, the choice of the diagnostic characters is made with the purpose of unambiguous fault identification from the predetermined set by the analysis of the vector pairwise combination structures $S_i(t)$. Thus, if among the vectors there are not a couple with the same structure, then for the given system $\Delta y(t)$ they must be selected as diagnostic characters. The selection of such diagnostic characters provides unambiguous fault identification. If among vectors $S_i(t)$ occurs a couple with the same structure then $\Delta\alpha_i$ must be selected as diagnostic character. The choice of different diagnostic characters gives the possibility to carry out the operational procedure of fault detection.

6. Fault identification

The problem of fault identification lies in identifying the fault location, calculation of estimating value of its size, and definition of the fault type.

The structural-analytic method of the identification of images with different characters [7] is used for the synthesis of the procedure of the fault location identification. We assume that the local attribute of the fault location image $K_i \in K$, $i=\overline{1, q}$ defines the "likeness" between the corresponding realizations in the character space $z^j \in Z^n$, $j=\overline{1, n}$ in a sense of their submission to a single law. Then the local attribute can be described with a two-valued predicate equation

$$(\varphi_i(z, \beta)>0)=\begin{cases}1, & \text{if } z \in K_i,\\ 0, & \text{if } z \bar{\in} K_i,\end{cases} \tag{11}$$

where $\varphi_i(z, \beta)$ is an analytical dependence determining the law φ_i which governs the realization z of the image $K_i \in K$; β is the parameter vector of the law φ_i which is developed from the analytical relation of characters. Any local attribute φ_i estimated by expression (11) is called attribute-predicate $\varphi_i(A-P)$. Then object z has the attribute φ_i if for corresponding $A-P\varphi_i$ the predicate equation is developed in the point $z \in Z^j$. The type of subimages and corresponding $A-P$'s depends on the type of the scale of the character measurements and is defined by the type of dependence $\varphi_i(z, \beta)$, vector dimension z, β, and the method of parameter calculation β.

Example 2. Let the character be $z_j \in \{0, 1\}$; this occurs for the character $\Delta y(t)$ which takes values in the logical scale, then Heavyside function with constant threshold

$\beta = 0.5$ is used to develop the corresponding $A - P\varphi_i$

$$\varphi_i(z, \beta) \equiv (z_j - 0.5) \geqq 0) = 1.$$

After identification of the fault location the estimation value of its size is defined. If $\Delta y(t)$ is used as the diagnostic character then in order to develop the estimation value of the fault character $\Delta\hat{\alpha}_i$ it is necessary to use analytical dependence relating $\Delta\bar{\alpha}_i$ to $\Delta y(t)$ or relationship (8), which allows to estimate $\Delta\alpha_i$ via selective values of sensitivity function $S_i(t)$ and diagnostic character $\Delta y_i(t)$. If $\Delta\alpha_i$ is used as a diagnostic character, then after identification of i-fault location the estimation value of its size will be equal to the character $\Delta\hat{\alpha}_i$.

Classification of the fault type is performed on the basis of estimation value of the fault size $\Delta\hat{\alpha}_i$. Two classes of fault types are used: 1) compensable faults — ω_i^1; 2) uncompensable faults — ω_i^2. Two-valued predicate equation is used for classification

$$(\psi(\Delta\hat{\alpha}_i, \beta) \geqq 0) = \begin{cases} 1, & \text{if} \quad \Delta\alpha_i \in \omega_i^1, \\ 0, & \text{if} \quad \Delta\hat{\alpha}_i \in \omega_i^2. \end{cases}$$

A set of terminal $A - P$'s gives the possibility to develop a treelike procedure of timely fault identification.

7. Serviceability restoration

The results of fault identification provide the logical basis for making decision aimed at the compensation and counteraction of a fault occurred. The solution will be satisfactory if the functional minimum is available (3).

Asymptotic stability of the control system behaviour $\lim y(t) = y_N(t)$ with $t \to \infty$ may be provided with the synthesis of restoration algorithm structure on the basis of the second Lyapunov's method.

Serviceability restoration is connected with the solution of two problems. The first is the parametric compensation of control system faults. The second is the standby control when the failure of the system hardware occurs.

In solving the tasks enumerated sufficient conditions of asymptotic stability of the control system are obtained by the development of negative sign of the functional derivative (3).

The solution of the problem of serviceability restoration of control systems is given in [8, 9].

Example 3. Let us consider some results of digital simulation of fault-tolerant system of stabilization of the electric motor angular speed. Figure 2 shows the functional scheme of the regulation system. The adder is built on the basis of

operational amplifier. Input resistors R_1, R_2 and the resistor R_0 in the feedback circuit characterize the technical condition of the adder. Power amplifier is an inertialess element with coefficient K_A. The electric motor is described by the first order differential equation with the following parameters: K_M is the transmission coefficient; I is the inertia moment of armature; r is a resistance of armature winding.

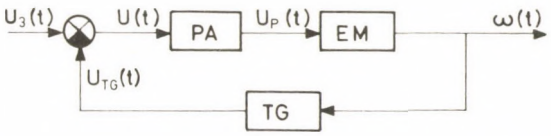


Fig. 2. Functional scheme of stabilization system of electromotor angular speed. *A* — adder; *PA* — power amplifier; *EM* — electromotor; *TG* — tachogenerator

Tachogenerator is inertialess element with transmission coefficient K_{TG}. Seven different faults characterized by the following vector of diagnostic parameters $\alpha=[R_0, R_1, R_2, K_A, K_M, r, K_{TG}]$ were considered in this investigation. Linearized system model has been obtained and its decomposition has been developed to provide the diagnosis of the original system in relation to the predetermined set of faults subjected to Poisson arrival. As a result the system has been divided into three subsystems which can be described by the following equations

$$\begin{cases} U(t)=\dfrac{R_1}{R_0}U_3(t)-\dfrac{R_2}{R_0}U_{TG}(t), \\ U_p(t)=K_A U(t), \\ \dot{U}_{TG}(t)=-\dfrac{1}{IrK_M^2}U_{TG}(t)+\dfrac{K_M K_{TG}}{IrK_M^2}U_p(t). \end{cases}$$

The measurement vector will consist of three components

$$\begin{bmatrix} \tilde{U}(t) \\ \tilde{U}(t) \\ \tilde{U}_{TG}(t) \end{bmatrix} = \begin{bmatrix} U(t) \\ U(t) \\ U_{TG}(t) \end{bmatrix} + \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \xi_3(t) \end{bmatrix},$$

where $\xi_i(t)$ is the measurement interferences, $i=\overline{1,3}$.

For the supervision of stabilization system serviceability we used a standard model of the following architecture

$$U_{TGS}(t)=aU_{TGS}(t)+bU_3(t),$$

where a and b are determined via functional element parameters.

As a result of the analysis of diagnostic subsystem models the diagnostic characters have been formulated. For the adder it is $\Delta R_1, \Delta R_2, \Delta R_0$. For the amplifier it is $\Delta U_p(t)$, for the electromotor-tachogenerator system it is $\Delta \hat{K}_M$, $\Delta \hat{r}$, $\Delta \hat{K}_{TG}$.

Treelike procedures of fault location identification and its type have been developed by the local characteristics of predicate equations.

The algorithm structures of parametric compensation and standby control for each subsystem have been developed for serviceability restoration of stabilization system.

Investigation of fault-tolerant system serviceability was carried out on the basis of digital simulation. Hereat different types of faults from the predetermined set have been introduced and the quality of functioning has been evaluated proceeding from the reaction of the tachogenerator voltage $U_{TG}(t)$ to the step input $U_3(t) = A \cdot 1(t)$. Here the parameter values were used: $R_0 = R_1 = R_2 = 1\text{M}\Omega$; $K_A = 1.67$; $K_M = 27.76$ 1/vs; $I = 2.2 \times 10^{-6}$ kg m^2; $r = 117\ \Omega$; $K_{TG} = 0.04$ vs; $A = 7$ V.

Figure 3 indicates the results of the digital simulation of the condition when in the stabilization system resistor R_1 fault which lies in decreasing by a minimal 20% that corresponds to compensable type of the fault has occured.

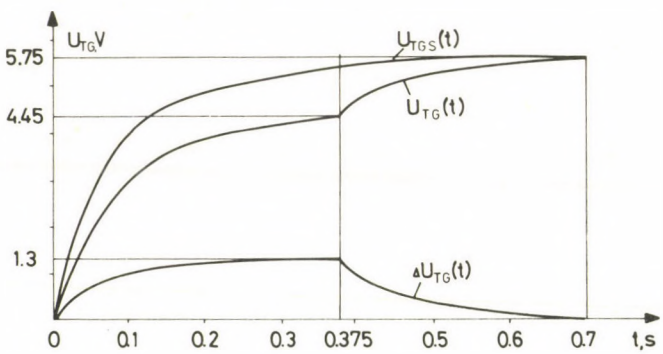


Fig. 3. Transition processes in the stabilization of fault-tolerant system with compensable fault R_1

As it follows from the plots represented the system serviceability control has taken place when resistor R_1 compensable fault reached $t = 0.375$ s and the fact of stabilization system non-serviceability has been established, then the fault location, resistor R_1, its value and type have been determined, and the decision as to parametric compensation of this fault has been taken. "Drawing up" the system to the standard behaviour $U_{TGS}(t)$ which ends at $t = 0.8$ s, the fault being $\Delta U_{TG}(t) \leqq 0.14$ V, begins on the plot of transmission process of the stabilization system $U_{TG}(t)$ where $t = 0.375$ s.

Figure 4 shows the plots of the transmission processes when in fault-tolerant system non-compensable tachogenerator fault has occurred. It is evident that before

the time moment $t=0.375$ s a blunder $\Delta U_{TG}(t)$ had occurred in the system, as the real behaviour of the system $U_{TG}(t)$ is considerably different from the standard $U_{TGS}(t)$. Tachogenerator fault identification had been carried out and the decision to disconnect the failed tachogenerator had been made by the time moment $t=0.375$ s. Connected standby tachogenerator began functioning since time moment $t=0.44$ s and the blunder streamed to 0. By the time moment $t=0.8$ s the stabilization system serviceability restoration had been completed.

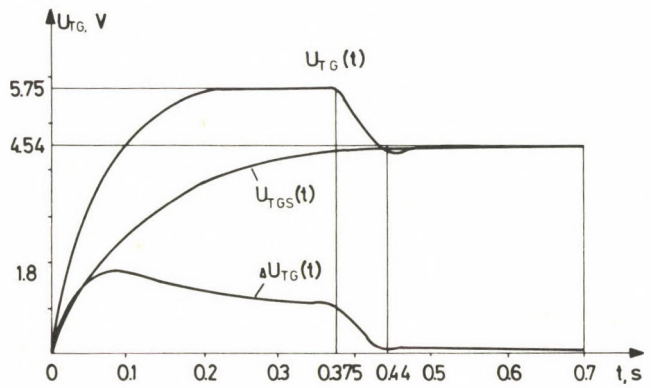


Fig. 4. Transition processes in the stabilization of fault-tolerant system with non-compensable fault K_{TG}

8. Conclusions

The proposed approach to the synthesis of fault-tolerant control systems has been presented for application with linearized conditions systems. Investigations carried out for the class of discrete systems as well as systems with substantial nonlinearities showed serviceability and effectiveness of the approach with digital simulation.

References

1. *Saridis, J.*, Self-organizing stochastic control systems. Moscow, Nauka, 1980, 400 pp.
2. *Petrov, V. N., Portnov-Sokolov, Y. P., Andriyenko, A. Y., Ivanov, V. P.*, Airborne terminal control systems: Construction principles and theory elements. Moscow, Mashinostroenie 1983, 200 pp.
3. *Eyckhoff, P.*, Foundations of control system identification. Moscow, Mir, 1975, 684 pp.
4. *Isermann, R.*, Process Fault Diagnosis with Parameter Estimation Methods. 7th IFAC Conference "Digital Computer Application to Process Control", Sept. 17–20. 1985, Vienna, pp. 361–375.
5. *Frank, P. M.*, Erhöhung der Robustheit und Zuverlässigkeit automatischen Prozesse-Automatisierungstechnische Praxis. 1985, No. 2, S. 64–72.

6. *Salyga, V. N., Sirodga, I. V., Kulik, A. S.*, System diagnosis in the problem of versatile and reliable control. In the book: X All-Union conference on control problems. Theses of reports. Moscow, IPU, 1986, pp. 513–514.
7. *Sirogda, I. B.*, Structural-analytic method of computer identification of objects with different characters. The Theory of *R*-functions and Actual Problems of Applied Mathematics. Kiev, Naukova Dumka, 1986, pp. 212–243.
8. *Kulik, A. S., Rubanov, V. G., Sokolov, Y. N.*, Synthesis of systems adapting to variations of element parameters and to their failures. Avtomatika i Telemechanika, No. *1*, 1978, pp. 96–107.
9. *Kulik, A. S.*, Algorithm of standby control with system hardware failures. Radioelektronika Letatelnych Apparatov. Kharkov, 1976, *8*, pp. 62–69.
10. *Kulik, A. S.*, Diagnosability of linear continuous systems. Avtomatika i Telemechanika, *6*, 1987, pp. 148–155.

Синтез отказоустойчивых динамических систем управления с идентификацией дефектов

В. И. САЛЫГА, И. Б. СИРОДЖА, А. С. КУЛИК, В. Л. ОБРУЧЕВ

(Харьков, Москва)

Рассматривается проблема построения отказоустойчивых систем управления на основе использования нетрадиционных принципов идентификации возмущающих воздействий (дефектов). Формулируется и решается задача синтеза отказоустойчивых линейных систем управления, а также указываются пути построения нелинейных систем в рамках предложенной общей концепции диагностической модели, которая описывает связи диагностических признаков системы с признаками дефектов и обеспечивает идентификацию последних. Вопрос принципиальной возможности распознавания заданного множества дефектов системы по результатам измерений решается с помощью введенного конструктивного критерия диагностируемости. Этот критерий обеспечивает обоснованное разбиение (декомпозицию) системы управления на идентифицируемые компоненты по отношению к заданному множеству дефектов и построение процедуры оперативного оценивания состояния системы с определением места дефекта, его величины и характера. Для определения места и (характера) вида дефекта используется структурно-аналитический метод распознавания образов с разнотипными признаками.

Парирование дефектов системы управления осуществляется с помощью алгоритмов восстановления работоспособности, синтезированных на основе второго метода Ляпунова.

Приведены результаты цифрового моделирования, иллюстрирующие и подтверждающие эффективность предлагаемого подхода к синтезу отказоустойчивых динамических систем.

В. И. Салыга
МинВУЗ СССР, Главное управление научно-исследовательских работ
СССР, 113833 Москва, М-230, ГСП, ул. Люсиновская, 51

И. Б. Сироджа, А. С. Кулик
Харьковский авиационный институт
СССР, 310191 Харьков, ул. Чкалова, 17

В. Л. Обручев
Московский институт стали и сплавов
СССР, 117936 Москва, Ленинский пр., 4

Problems of Control and Information Theory, Vol. 18 (1), pp. 55—63 (1989)

DECENTRALIZED CONTROL OF LINEAR SYSTEMS

I. HEJDA, J. MURGAŠ

(*Bratislava*)

(Received January 5, 1988)

Three methods of decentralized *LQ* control design are given. They differ in the solution of the initial conditions dependence problem. One of them is the new concept of decentralized dominant feedback following from the centralized one introduced by Allwright in 1982.

1. Introduction

Large scale systems, such as power networks, vast technological processes, fast complex servomechanisms, urban traffic networks, socioeconomic systems, etc. present a great challenge to the control system designers. The classical control design presupposition based on the centrality fails because of geographical separation, elevated cost of communication, reliability issues, great complexity of the control system and limited capability of centralized computing. Thus, a decentralized control scheme is imposed. The system is then acted upon several controllers which observe only local system outputs, control local inputs and do not communicate among themselves. Great capability and low price of microprocessors are another reasons favouring the distributed computation.

In the last 20 years, the greatest effort was concentrated on a so-called linear quadratic problem (linear dynamics, quadratic criterion). It was shown that the linear quadratic design methods have desirable sensitivity and robustness properties. Such optimal control, though the system is linear, is characterized by the dependence on initial conditions. In this paper, we discuss three methods of decentralized *LQ* control design differing in the solution of the initial conditions dependence problem. One of them is the new concept of decentralized dominant feedback. It follows from the centralized output feedback introduced by Allwright [1].

2. Problem statement

Consider the system consisting of s subsystems:

$$dx(t)/dt = Ax(t) + \sum_{i=1}^{s} B_i u_i(t) \tag{1a}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

$$y_i(t) = C_i x(t), \qquad i = 1, \ldots, s \tag{1b}$$

where $x(t) \in R^n$ is the state vector, $u_i(t) \in R^{m_i}$ and $y_i(t) \in R^{r_i}$ are the input and output vector, respectively, of the ith subsystem. A_i, B_i, C_i are real matrices of appropriate size.

The objective of the optimal decentralized control design is to find the control laws of the form

$$u_i(t) = -K_i y_i(t), \qquad i = 1, \ldots, s \tag{2}$$

minimizing the quadratic criterion

$$J = \int_0^\infty \{x^T(t) Q x(t) + u^T(t) R u(t)\} dt \tag{3}$$

where $K_i \in R^{m_i \times r_i}$, $Q \in R^{n \times n}$ is positive semidefinite and $R \in R^{m \times m}$ is positive definite,

$$m = m_1 + \ldots + m_s, \quad u(t) = [u_1^T(t), \ldots, u_s^T(t)]^T.$$

Let

$$B = (B_1, \ldots, B_s), \quad C = (C_1^T, \ldots, C_s^T)^T.$$

Then the system is given by

$$dx(t)/dt = Ax(t) + Bu(t) \tag{4a}$$

$$y(t) = Cx(t) \tag{4b}$$

where $y(t) = [y_1^T(t), \ldots, y_s^T(t)]^T$. The design problem can be formulated in the following form:

$$\min_{K \in \mathcal{K}} J, \tag{5}$$

where

$$\mathcal{K} = \{K: K = \text{block diag}\,(K_1, \ldots, K_s); \quad \sigma[A - BKC] \subset C^-\}. \tag{6}$$

$\sigma[X]$ denotes the set of the eigenvalues of X and C^- is the subset of complex numbers with negative real parts. The criterion considered is an infinite horizon. The finite horizon problem is treated in [2] and infinite in [2], [6], [7], [9], [10]. Synthesis methods based on Lyapunov theory of stability (e.g. [17]) can be employed when some additional constraints are imposed (connective stability, etc.). The feedback matrix $K \in R^{m \times r}$, $r = r_1 + \ldots + r_s$, in (6) has not to be strictly block diagonal. It can have only a similar form, e.g. due to the overlapping decomposition or information constraints in the subsystems. Let Ω denote the set of $K \in R^{m \times r}$ having this structure. In the following, we will consider the optimal control design problem (5) with $\mathcal{K}$ given by

$$\mathcal{K} = \{K: K \in \Omega; \sigma[A - BKC] \subset C^-\}. \tag{7}$$

As mentioned above, this problem depends on initial conditions. The optimal feedback K minimizing (5), (7) is, therefore, a function of $x(0)$.

3. Decentralized control

A. Statistical method

The most widely used procedure to eliminate the above dependence was introduced in [3]. It can be done by assuming that the initial state $x(0)$ is a random variable with known or estimated covariance matrix $Z_0 = E[x(0)x^T(0)]$, where $E[\cdot]$ is the mean value. Then the design problem can be transformed to the form:

$$\min_{K \in \mathscr{K}} J_A = \min_{K \in \mathscr{K}} E[J], \quad E[J] = \operatorname{tr}(PZ_0) \tag{8}$$

$$(A - BKC)^T P + P(A - BKC) + Q + C^T K^T RKC = 0\,. \tag{9}$$

The criterion (8) is differentiable with respect to K with the gradient:

$$\partial J_A / \partial K = 2(RKC - B^T P)VC^T \tag{10}$$

where V is the solution of the Lyapunov equation

$$(A - BKC)V + V(A - BKC)^T + Z_0 = 0. \tag{11}$$

The overall system is stable only if the solution V of (11) is positive definite. Then, the following iterative gradient algorithm can be pointed out. Let D be the projection of $\partial J_A / \partial K$ on the set Ω. It means that D is obtained from $\partial J_A / \partial K$ by setting some of its elements to zero, in order that the structure prescribed by the set Ω (e.g. block diagonal) is reached. Choose $K_0 \in \mathscr{K}$. Iterate $K_{p+1} = K_p - \alpha_p D_p$, α_p is the step size such that $J_A(K_{p+1}) < J_A(K_p)$ and $K_{p+1} \in \mathscr{K}$. Stop when $|D_p| < \varepsilon$. ε is a small positive number and $|D_p|$ denotes some norm of D_p.

Such techniques of optimal control design can be found in [4]–[6] for the centralized case and in [6]–[9] for the decentralized case. Hierarchical iterative method solving (8) was given in [10].

B. Minimizing the worst performance

Another idea avoiding the initial conditions dependence is taken from [2]. The criterion used is

$$J_B = \max_{Z_0} [J : \|Z_0\|^2 < \varphi^2] \tag{12}$$

$$\|Z_0\| = \{\operatorname{tr}(Z_0^T Z_0)\}^{1/2} \tag{13}$$

where $\varphi > 0$. Thus, the optimal decentralized feedback minimizes

$$\min_{K \in \mathcal{K}} J_B. \tag{14}$$

The criterion J_B is the worst performance for every initial conditions satisfying $\|Z_0\|^2 < \varphi^2$. The idea is to minimize this worst performance. J_B is differentiable, its gradient is given by (9)–(11) with $Z_0 = \varphi P / \|P\|$. The proof [2] consists on the fact that $Z_0 = \varphi P / \|P\|$ is the solution of problem (12). The same gradient iterative algorithm as in section A can be used for the solution of problem (14).

C. Dominant decentralized feedback

The above two methods use the same iterative algorithm. Suppose that the algorithm starts with the initial feedback K_0 and ends with K^* satisfying the necessary conditions of optimality. Because of statistical simplification, there may be initial conditions $x(0)$ such that $J[K^*, x(0)] > J[K_0, x(0)]$. It is possible that the optimized feedback K^* for some initial conditions gives worse behaviour and more elevated cost than $K = 0$, i.e. the open loop system (see [1]).

It will be said that $\bar{K}$ dominates K if $J[\bar{K}, x(0)] < J[K, x(0)]$ for every $x(0) \neq 0$. In [1] Allwright derived a method giving a sequence of output feedbacks K_p for the centralized system satisfying for every $x(0) \neq 0$

$$J[K_{p+1}, x(0)] < J[K_p, x(0)]. \tag{15}$$

The same principle can be used for the decentralized control, as it will be shown. The idea is to determine the feedback dominating arbitrary $K_0 \in \mathcal{K}$ (e.g. $K_0 = 0$), when it is possible. The used algorithm is gradient and iterative. The value of the cost (3) can be expressed as $J[K, x(0)] = x^T(0) P x(0)$, where P is the solution of (9). P is differentiable at K ($K = (k_{ab})$) with the differential $\partial P / \partial k_{ab} = G_{ab}$; G_{ab} is the solution of the equation ([1]):

$$(A - BKC)^T G_{ab} + G_{ab}(A - BKC) + C^T E_{ab}^T (B^T P - RKC) + \\ + (B^T P - RKC)^T E_{ab} C = 0. \tag{16}$$

E_{ab} is an $m \times r$ matrix having all zero elements except of element $[a, b]$ with value one. Let $dP(K, S)$ be the first order approximation of $P(K + S) - P(K)$, $S = (s_{ab}) \in \Omega$. Then

$$dP(K, S) = \sum_{a=1}^{m} \sum_{b=1}^{r} G_{ab} s_{ab}. \tag{17}$$

Lemma 1. Let

$$\Xi = \{S:: S \in \Omega \subset R^{m \times r};\ \sum_{a=1}^{m} \sum_{b=1}^{r} s_{ab}^2 = 1\},\ \pi(K) =$$

$$= \min\,[\lambda_{\max}\{dP(K, S)\} : S \in \Xi],\ \lambda_{\max}[\,\cdot\,]$$

be the maximum eigenvalue and $\bar{S}$ minimizes $\lambda_{\max}\{dP(K, S)\}$ on Ξ. Then, for $K \in \mathscr{K}$ if $\pi(K) < 0$, there exists $\delta > 0$ such that $J[K + \beta\bar{S},\ x(0)] < J[K,\ x(0)]$ for every $x(0) \neq 0$ and all $\beta \in (0,\ \delta)$.

The proof of this lemma is equivalent to the proof of lemma 3.2 in [1].

The decentralized dominant feedback can be determined by the next algorithm.

1) Let $p = 0$. Choose $K_0 \in \mathscr{K}$.
2) If $\pi(K_p) \geqq 0$ then stop, else continue.
3) Find S_p minimizing $\lambda_{\max}[dP(K_p, S)]$; $S \in \Xi$.
4) Find β_p minimizing $\lambda_{\max}[P(K_p + \beta S_p) - P(K_p)]$.
5) Update $K_{p+1} = K_p + \beta_p S_p$, let $p = p + 1$, go to step 2).

Theorem 1. If the above algorithm generates a sequence K_p, then relation (15) is satisfied for every p and $x(0) \neq 0$.

Proof. Since $\pi(K_p) < 0$, then β_p obtained in step 4) satisfies

$$\lambda_{\max}[P(K_p + \beta_p S_p) - P(K_p)] < 0,\ \text{i.e.}\ \lambda_{\max}[P(K_{p+1}) - P(K_p)] < 0.$$

Also

$$J[K_{p+1}, x(0)] - J[K_p, x(0)] = x^T(0)\,\{P(K_{p+1}) - P(K_p)\}x(0). \tag{18}$$

Hence $\lambda_{\max}[P(K_{p+1}) - P(K_p)] < 0$, $P(K_{p+1}) - P(K_p)$ is negative definite. Then the right-hand side of (18) is negative for every $x(0) \neq 0$. This proves (15). Hence $K_p \in \Omega$ and $S_p \in \Omega$, then $K_{p+1} \in \Omega$, too.

The implementation of steps 2) and 3) is not very simple. The problem from step 3) has a nondifferentiable convex cost and its feasible set is not convex. Though our problem is more general than the problem stated in [1], the same method based on convex analysis can be used. The algorithm for the solution of step 3) derived there must submit some modifications. The criterion from step 3) can be expressed as

$$\lambda_{\max}\left[\sum_{a=1}^{m} \sum_{b=1}^{r} G_{ab}(K_p) s_{ab}\right] = \lambda_{\max}\left[\sum_{k=1}^{q} H_k s_k\right] \tag{19}$$

where q is the number of nonzero elements of $S = (s_{ab})$ defined by the control constraints Ω, $q \leqq m \times r$, $s = (s_1, \ldots, s_q)^T \in R^q$ is the vector formed from these elements and H_k is equal to $G_{ab}(K_p)$ associated with $s_k = s_{ab}$. Then, for the solution of step 3), one has to determine the vector $\bar{s} \in \Phi = \{s \in R^q : \|s\| = 1\}$ minimizing the right-hand side of (19).

For this reason, the following sequence of vectors $z^j \in R^q$ is used

$$z^{j+1} = z^j + \operatorname{sat} \psi(z^j)\,[g(-z^j) - z^j]$$

$$\psi(z) = \begin{cases} 0 & \text{if } z - g(-z) = 0 \\ z^T[z - g(-z)]/\|z - g(-z)\|^2, & \text{otherwise} \end{cases}$$

$$g(z) = \begin{cases} z^0 & \text{if } z = 0 \\ [v_z^T H_1 v_z, \ldots, v_z^T H_q v_z]^T, & \text{otherwise} \end{cases}$$

where v_z is a normalized eigenvector of $\sum_{k=1}^{q} z_k H_k$ corresponding to the maximum eigenvalue. The sequence starts with $z^0 = [v_h^T H_1 v_h, \ldots, v_h^T H_q v_h]^T$, $h \in \Phi$. Let $s^j = -z^j/\|z^j\|$. Then, the following two statements can be proven by the same way as in [1]:

a) $\pi(K_p) < 0$ if $\lambda_{\max}\left[\sum_{k=1}^{q} s_k^j H_k\right] < 0$ for some finite j.

b) If $\pi(K_p) < 0$, then $s^j \to \bar{s}$ and for $\varepsilon > 0$, if the iteration stops in the step w with

$$\lambda_{\max}\left[\sum_{k=1}^{q} s_k^w H_k\right] + \|z^j\| < \varepsilon,$$

then

$$\lambda_{\max}\left[\sum_{k=1}^{q} s_k^w H_k\right] - \pi(K_p) < \varepsilon.$$

These two statements allow us to determine whether $\pi(K_p) < 0$ (step 2) and to obtain arbitrarily good estimate s^w of $\bar{s}$, i.e. of S_p too (step 3).

4. Example

Consider a decentralized system with $n=4$, $s=2$, $m_1 = m_2 = 1$, $r_1 = 2$, $r_2 = 1$ described by (4) with

$$A = \begin{bmatrix} -0.5 & 0 & 0 & 0.05 \\ 0 & -0.25 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0.1 & 0 & 0 & -0.25 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & 0 \\ 0.5 & 0 \\ 0 & 1 \\ 0 & 0.5 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}; \quad K = \begin{bmatrix} k_{11} & k_{12} & 0 \\ 0 & 0 & k_{23} \end{bmatrix}.$$

Here Ω is the set of matrices K of the above form. Weight matrices in (3) are $Q=\text{diag}\,(1; 10; 1; 1)$, $R=\text{diag}\,(0.5; 0.5)$. The optimization for the methods in sections 3A, 3B was given by conjugate gradients, with quadratic interpolation for determining the step size α_p ([9]) and with $K_0=0$. The convergence test was

$$|D_p|<\varepsilon=0.001;$$

where

$$|D_p|=\max|d_{ab}|;\ a=1, \ldots, m;\ b=1, \ldots, r;\ D_p=(d_{ab}).$$

The first, with $Z_0=I$, stopped in the sixteenth iteration,

$$K_{16}=\begin{bmatrix} 0.522 & 3.716 & 0 \\ 0 & 0 & 0.888 \end{bmatrix}.$$

The cost $J_A(K_{16})=8.901$. The second method stopped in the seventeenth iteration,

$$K_{17}=\begin{bmatrix} 0.522 & 3.711 & 0 \\ 0 & 0 & 1.268 \end{bmatrix}$$

with the cost $J_B(K_{17})=6.765$. For the method from section C the open loop was chosen as K_0, too. The algorithm stopped after the first iteration, the decentralized feedback dominating $K_0=0$ is

$$K_1=\begin{bmatrix} 0.212 & 1.850 & 0 \\ 0 & 0 & 0.613 \end{bmatrix}.$$

The eigenvalues of $P(K_1)-P(K_0)$ are -13.25; -0.837; -0.026; -0.008. For comparison, $J_B(K_{16})=6.711$; $J_A(K_{17})=8.933$; $J_A(K_1)=9.523$; $J_B(K_1)=7.288$. Here K_{16} dominates $K_0=0$, too. However, it is possible that the feedback received by the first two methods does not dominate the open loop (an example for the centralized output feedback can be found in [1]). For this concrete system, K_{16} is very similar to K_{17}. The experience with other systems showed us that it is not usual.

5. Remarks

All three algorithms need a starting matrix $K_0 \in \mathcal{K}$. Such a matrix can be obtained by the method of Armentano and Singh [11]. Their gradient method minimizes the real part of the dominant eigenvalue of the system. It can be applied to a wider set of control constraints as it was proposed originally. Constraints considered in this paper can also be covered (see [9]). Such initialization seems us to be more efficient as those based on Lyapunov theory.

The first method (section A) for the initial condition dependence elimination will be successfully used when the statistical properties of $x(0)$ are exactly or

approximately known. The second (section B) will be favoured when such information is reduced. The decentralized dominant feedback (section C) can be chosen as optimal when this information does not exist at all. We must note that the control obtained by the first two methods satisfy only necessary conditions of optimality. The feedback obtained can be only a local minimum of J_A or J_B. For its verification we can restart the optimization with various K_0. For the decentralized dominant feedback, the situation, when there is no dominating K near K_0 is not very unlikely. It is possible that it does not exist at all. In this case, the third method is unable to find the optimal control. All algorithms can also be used in the centralized case and/or with arbitrary control constraints given by the feasible set Ω. Control with dynamic compensators can be obtained by the state extention as in [12].

6. Conclusion

Three possibilities of *LQ* initial conditions dependence elimination were considered. There exists a lot of other ways [13]–[15]. The optimal feedback computation is not very simple (especially in the third case). However, it is applied off-line and requires the solution of linear matrix equations only. Because of applicability to a wide class of constrained control problems, the introduced algorithms are very efficient. Some experiments with on-line application of similar methods can be found in [16].

References

1. *Allwright, J. C.*, LQP: Dominant Output Feedbacks. IEEE Trans. on Aut. Control, **AC-27,** 1982, pp. 915–921.
2. *Geromel, J. C., Bernussou, J.*, Optimal Decentralized Control of Dynamic Systems. Automatica, **18,** 1982, pp. 545–557.
3. *Levine, W. S., Athans, M.*, On the Determination of the Optimal Constant Output Feedback Gains for Linear Multivariable Systems. IEEE Trans. on Aut. Control, **AC-15,** 1970, pp. 44–48.
4. *Horisberger, P., Bélanger, H. R.*, Solution of the Optimal Constant Output Feedback Problem by Conjugate Gradients. IEEE Trans. on Aut. Control, **AC-19,** 1974, pp. 434–435.
5. *Choi, S., Sirisena, H. R.*, Computation of Optimal Output Feedback Gains for Linear Multivariable Systems. IEEE Trans. on Aut. Control, **AC-19,** 1974, pp. 257–258.
6. *Wenk, C. J., Knapp, Ch. H.*, Parameter Optimization in Linear Systems with Arbitrarily Constrained Controller Structure. IEEE Trans. on Aut. Control, **AC-25,** 1980, pp. 496–500.
7. *Geromel, J. C., Bernussou, J.*, An Algorithm for Optimal Decentralized Regulation of Linear Quadratic Interconnected Systems. Automatica, **15,** 1979, pp. 489–491.
8. *Travé, L., Tarras, A., Titli, A.*, Some Problems in Decentralized Control in Presence of Fixed Modes. In: Proc. 1984 IFAC World Congress, Budapest.
9. *Hejda, I., Murgaš, J.*, Solution of the *LQ* Problem with Given Structure of Control. Automatizace, **30,** 1987, pp. 214–216 (in Slovak).

10. *Xinogalas, T. P., Mahmoud, M. S., Singh, M. G.*, Hierarchical Computation of Decentralized Gains for Interconnected Systems. In.: Proc. 1981 IFAC World Congress, Kyoto.
11. *Armentano, W. A., Singh, M. G.*, A New Approach to the Decentralized Controller Initialisation Problem. In: Proc. 1981 IFAC World Congress, Kyoto.
12. *Johnson, T. L., Athans, M.*, On the Design of Optimal Constrained Dynamic Compensators for Linear Constant Systems. IEEE Trans. on Aut. Control, **AC-15,** 1970, pp. 658–660.
13. *Dabke, K. P.*, Suboptimal Regulators with Incomplete State Feedback. IEEE Trans. on Aut. Control, **AC-15,** 1970, pp. 120–122.
14. *Man, F. T.*, Suboptimal Control of Linear Time-Invariant Systems with Incomplete Feedback. IEEE Trans. on Aut. Control, **AC-15,** 1970, pp. 112–114.
15. *Allwright, J. C., Mao, J. Q.*, Optimal Output Feedback by Minimizing $\|K(F)\|_2$. IEEE Trans. on Aut. Control, **AC-27,** 1982, pp. 729–731.
16. *Friedlander, B., Porat, B.*, Adaptive Design of Decentralized Controllers. Proc. of 21st IEEE Conference on Decision and Control, 1982.
17. *Veselý, V., Murgaš, J., Bízik, J.*, Decentralized Control of Dynamic Linear Systems. In: Proc. 1981 IFAC World Congress, Kyoto.

Децентрализованное управление линейными системами

И. ГЕЙДА, Я. МУРГАШ

(Братислава)

В статьи представлены три методы синтеза децентрализованного управления. Они отличаются способам решения проблемы зависимости оптимального управления от начального вектора состояния. Одним из них является метод децентрализованной домминирующей оборотной связи основан на принципах метода Олрайта (1982) для централизованных систем.

I. Hejda and J. Murgaš
Faculty of Electrical Engineering,
Slovak Technical University, Mlynská dolina,
812 19 Bratislava
Czechoslovakia

РУССКИЙ ПЕРЕВОД

Проблемы управления и теории информации, том **18,** *номер 1 (1989)*

О МИНИМАКСНЫХ ОЦЕНКАХ СОСТОЯНИЯ МНОГОШАГОВЫХ СТАТИСТИЧЕСКИ НЕОПРЕДЕЛЕННЫХ СИСТЕМ

Б. И. АНАНЬЕВ

(*Свердловск*)

Для нелинейных многошаговых систем, подверженных случайным возмущениям и содержащих неопределенные параметры, рассмотрена минимаксная задача оценивания фазового состояния. Определен класс допустимых нелинейных оценок, в котором отыскиваются оптимальные минимаксные оценки. Получены необходимые условия оптимальности, которые при определенном условии невырожденности являются также и достаточными. В последнем случае оптимальная оценка определяется единственным образом в классе допустимых. Рассмотрена конечномерная аппроксимация решения. Отмечены частные случаи и приведены модельные примеры.

1. Введение

При решении задач оценивания состояния динамических систем с неполной информацией о параметрах и распределениях помех широкое распространение получил минимаксный (или игровой, гарантирующий) подход [1–11]. В большинстве работ указанного направления рассматриваются линейные системы. Среди исследований, посвященных нелинейным многошаговым системам, укажем, например [6, 8].

Известно [1, 2], что задачи оценивания (или наблюдения) координат вектора состояния динамической системы при неполных или неточных измерениях занимают одно из центральных мест в теории управления. Минимаксный подход [10] к задачам оценивания для систем с неопределенными величинами начал развиваться почти одновременно с появлением известной работы [12] по стохастической фильтрации. Важность методов минимаксного оценивания можно мотивировать тем обстоятельством, что во многих практических задачах детальная информация, касающаяся динамики процесса и его статистики, оказывается недоступной. Более того, приходится рассматривать процессы, находящиеся под воздействием неизвестных наблюдателю параметров, управлений или возмущений. В этих случаях имеет смысл предположение о

том, что неизвестные параметры могут формироваться на основе любой доступной информации, в том числе на основе знания структуры наблюдения, с целью максимизации ошибки оценивания. При этом информация наблюдателя о неопределенных величинах часто исчерпывается лишь знанием ограничений на их возможные значения.

Статистически неопределенная ситуация при оценивании может возникнуть в случае, если, например, в состав измерений вектора информации кроме быстро меняющихся (случайных) ошибок входят медленно меняющиеся ошибки [13]. Иногда модели этих последних ошибок удается описать формирующими фильтрами той или иной степени сложности, и после этого можно применять статистические методы. Однако менее искусственным подходом в данном случае будет минимаксный подход, при котором медленно меняющиеся ошибки считаются неопределенными, принадлежащими заданным ограничениям. Отметим, что при решении реальных задач оценивания на ЭВМ приходится моделировать дифференциальные уравнения с помощью многошаговых систем. В связи с этим на этапе обработки информации и дискретизации уравнений к статистически неопределенным возмущениям механической природы (ошибки измерительных приборов, неизвестные управления), добавляются, как правило, дополнительные ошибки электронные устройств, ошибки метода аппроксимации и т. д. Причем параметры вероятностных распределений дополнительных возмущений могут быть как известными, так и не известными. Отмеченные обстоятельства приводят к необходимости решать задачу при совместном рассмотрении неопределенных и случайных возмущений в многошаговых системах.

В настоящей работе развиваются изложенные в [6, 8] методы для нелинейных многошаговых систем общего вида, подверженных воздействию как случайных, так и неопределенных возмущений. Выделен класс допустимых нелинейных оценок, в котором найдены оптимальные минимаксные оценки и определяющие соотношения для них. Рассмотрены вопросы конечномерной аппроксимации решения и отмечены 2 частных случая, когда отсутствуют неопределенные параметры и когда отсутствуют случайные параметры. В первом случае решение сводится к известным результатам теории стохастической фильтрации [14], а во втором — к известным результатам теории гарантированного оценивания [2]. Результаты иллюстрируются на модельных примерах.

2. Предварительная постановка задачи

Рассматривается многошаговая система

$$x^t = f^t(x^{t-1}, \theta^t, \xi^t), \qquad t \geqq 1,$$
$$x^0 = f^0(\theta^0, \xi^0), \tag{2.1}$$

и уравнение измерения

$$y^t = g^t(x^t, \theta^t, \eta^t), \qquad t \geqq 1. \tag{2.2}$$

Здесь $x^t \in R^n$, $y^t \in R^m$. Элементы последовательностей $\{\xi^t\}$, $\{\eta^t\}$ представляют собой независимые случайные векторные величины с известными распределениями. Неопределенный векторный параметр θ^t удовлетворяет априорному включению

$$\theta^t \in \Theta_t, \qquad t \geqq 0, \tag{2.3}$$

где Θ_t — компакт в конечномерном евклидовом пространстве. Совокупность всех неопределенных параметров вплоть до момента t будем обозначать символом $\theta_{0t} = (\theta^0, \ldots, \theta^t)$, причем будет справедливо включение

$$\theta_{0t} \in \Theta_{0t} \tag{2.4}$$

где Θ_{0t} — соответствующее ограничивающее множество (декартово произведение компактов из (2.3)). Вектор-функции $f^t(\cdot, \cdot, \cdot)$, $g^t(\cdot, \cdot, \cdot)$ в уравнениях (2.1), (2.2) предполагаются непрерывными.

Для дальнейшего преобразуем многошаговую систему (2.1), (2.2) к одношаговой. Для этого введем обозначения

$$x = [x^{1\prime} \vdots \ldots \vdots x^{t\prime}]' \in R^{nt}, \qquad \xi = [\xi^{1\prime} \vdots \ldots \vdots \xi^{t\prime}]',$$
$$y = [y^{1\prime} \vdots \ldots \vdots y^{t\prime}]' \in R^{mt}, \qquad \eta = [\eta^{1\prime} \vdots \ldots \vdots \eta^{t\prime}]', \tag{2.5}$$

где знак $'$ означает транспонирование. Тогда будем иметь

$$x = F(x^0, \theta_{1t}, \xi),$$
$$y = G(x, \theta_{1t}, \eta), \tag{2.6}$$

причем

$$x^t = \pi_t x, \qquad \pi_t = [0 \vdots \ldots \vdots I_n] \in R^{n \times nt}. \tag{2.7}$$

Здесь и ниже I_n — единичная $n \times n$-матрица, $R^{n \times nt}$ — пространство матриц размера $n \times nt$.

В уравнениях (2.6) $F(\cdot, \cdot, \cdot)$ и $G(\cdot, \cdot, \cdot)$ суть непрерывные вектор-функции, которые строятся согласно (2.1), (2.2).

Рассмотрим класс Δ борелевских отображений $\delta(\cdot): R^{mt} \to R^n$, для которых

$$\sup_{\theta_{0t}} E\|\delta(y)\|^2 < \infty. \tag{2.8}$$

Здесь и далее $\|\cdot\|$ — евклидова норма, E — знак математического ожидания. Заметим, что класс Δ представляет собой линейное пространство.

Задача, изучаемая в работе, имеет вид

$$I(\delta(\cdot)) = \sup_{\theta_{0t}} E\|\pi_t x - \delta(y)\|^2 \to \min_{\delta(\cdot)}. \tag{2.9}$$

Помимо перечисленных выше работ задачи подобного типа изучались в монографии [15], статье [16] и некоторых других. Одно из отличий данной работы от предшествующих состоит в использовании некомпактного класса допустимых решающих функций [17]. Это повлекло за собой новое доказательство существования цены игры и ее аппроксимации.

По соображениям технического характера рассмотрение класса Δ, определяемого неравенством (2.8), может оказаться неудобным. Поэтому ниже определен некоторый подкласс $\tilde{\Delta} \subset \Delta$, в котором будет решаться задача (2.9) при наличии случайных параметров. Приведены также некоторые дополнительные условия на вероятностные распределения величин x, y (см. (2.6)).

3. Класс допустимых нелинейных оценок

Фиксируем момент времени t и будем опускать нижние индексы в обозначении параметров θ_{0t} из включения (2.4). Пусть P и Q — меры, заданные на одном и том же пространстве. Будем писать $P \ll Q$, если мера P абсолютно непрерывна относительно Q [14]. Запись $P \sim Q$ будет означать, что $P \ll Q$ и $Q \ll P$.

Примем следующие предположения.

Предположения 3.1. 1) Пусть для некоторой σ — конечной меры ν на R^{mt} при $\forall\, \theta \in \Theta$ имеем $P_\theta \ll \nu$, где P_θ — распределение случайного вектора y из (2.6). При этом для функции плотности требуем выполнения неравенства

$$dP_\theta(y)/d\nu(y) = p(y, \theta) \leqq \bar{p}(y) \qquad (\nu \text{ — п. в.}), \tag{3.1}$$

где $\bar{p}(\cdot) \in L_1(\nu)$.

2) Величина $\max_\theta \|x^t\|^2$ обладает конечным математическим ожиданием. Для $\bar{P}$ — почти всех y, где $d\bar{P}(y) = \bar{p}(y)d\nu(y)$, функция плотности $p(y, \theta)$ непрерывна по θ.

3) Существует вариант* условного среднего $E[x^t|y, \theta]$, удовлетворяющий неравенству

$$\|E[x^t|y, \theta]\| \leqq \bar{f}(y) \qquad (\bar{P} \text{ — п. в.}), \tag{3.2}$$

где $\bar{f}(\cdot) \in L_2(\bar{P})$. Для $\bar{P}$ — почти всех y вариант $E[x^t|y, \theta]$ непрерывен по θ на множестве $\{\theta \in \Theta : p(y, \theta) > 0\}$.

Замечание 3.1. При отсутствии неопределенности по θ в уравнениях (2.1), (2.2) предположения 3.1 автоматически выполняются, если только $E\|x^t\|^2 < \infty$.

Будем решать задачу (2.9) в классе $\tilde{\Delta}$ борелевских отображений $\delta(\cdot)$, принадлежащих пространству

$$L_2^n(\bar{P}) \subset \Delta, \tag{3.3}$$

где мера $\bar{P}$ определена в пункте 2) предположения 3.1. Оценки $\delta(\cdot)$ из класса $\tilde{\Delta}$ будем называть допустимыми.

Введем обозначение

$$K(\delta(\cdot), \theta) = E_\theta \|x^t - \delta(y)\|^2. \tag{3.4}$$

Здесь и далее E_θ — знак математического ожидания при известном значении θ.

Справедливо следующее утверждение.

Лемма 3.1. Пусть выполнены предположения 3.1. Тогда функция (3.4) непрерывна по θ при любом фиксированном $\delta(\cdot) \in \tilde{\Delta}$ и выпукла по $\delta(\cdot)$ при каждом фиксированном $\theta \in \Theta$.

Доказательство. В обосновании нуждается лишь непрерывность функционала (3.4) по θ. По формуле (3.4), применяя основные свойства условных математических ожиданий, получаем

$$K(\delta(\cdot), \theta) = E_\theta \|x^t\|^2 - 2 \int_{R^{mt}} \delta(y) E[x^t|y, \theta] \cdot$$

$$\cdot\, p(y, \theta) dv(y) + E_\theta \|\delta(y)\|^2. \tag{3.5}$$

В силу предположений 3.1 под знаками интегралов в выражении (3.5) стоят непрерывные по θ функции. Следовательно, по теореме Лебега о мажорированной сходимости функционал (3.4) будет непрерывным по θ.

* Определение варианта условного среднего дано в книге [18, гл. 2, 044,46].

4. Решение задачи в классе допустимых оценок

Далее будем считать предположения 3.1 выполненными. Пусть Λ — множество всех вероятностных мер на компакте Θ. Обозначим усреднение функции (3.4) по мере λ символом

$$\tilde{K}(\delta, \lambda) = \int_{\Theta} K(\delta(\cdot), \theta) d\lambda(\theta). \tag{4.1}$$

Для функционала $I(\cdot)$ задачи (2.9) справедливо равенство

$$I(\delta(\cdot)) = \max_{\theta \in \Theta} K(\delta(\cdot), \theta) = \max_{\lambda \in \Lambda} \tilde{K}(\delta, \lambda). \tag{4.2}$$

Отсюда следует, что решение задачи (2.9) в классе $\tilde{\Delta}$ (см. (3.3)) эквивалентно нахождению седловой точки для функционала (4.1). Действительно, наделяя множество Λ топологией слабой сходимости мер, превратим его в метрический компакт [19]. Тогда из теоремы о минимаксе [20] получим равенство

$$\inf_{\delta(\cdot)} I(\delta(\cdot)) = \inf_{\delta} \max_{\lambda} \tilde{K}(\delta, \lambda) = \max_{\lambda} \inf_{\delta} \tilde{K}(\delta, \lambda). \tag{4.3}$$

Ниже показано, что минимум по δ функционала (4.1) при фиксированном λ равен

$$\min_{\delta} \tilde{K}(\delta, \lambda) = \tilde{K}(\delta_\lambda, \lambda) = \int_{\Theta} (E_\theta \| x^t \|^2 - E_\theta \| \delta_\lambda(y) \|^2) d\lambda(\theta), \tag{4.4}$$

где символом

$$\delta_\lambda(y) = \int_{\Theta} E[x^t | y, \theta] p(y, \theta) d\lambda(\theta) / \int_{\Theta} p(y, \theta) d\lambda(\theta) \tag{4.5}$$

обозначена оценка, доставляющая этот минимум. Для векторов $y \notin Y_\lambda$, где

$$Y_\lambda = \{ y : \int_{\Theta} p(y, \theta) d\lambda(\theta) > 0 \}, \tag{4.6}$$

оценка (4.5) определяется произвольно. Далее будем считать, что $\delta_\lambda(y) = 0$ для отмеченного случая.

Справедлива лемма.

Лемма 4.1. Оценка (4.5) является допустимой для всякой меры $\lambda \in \Lambda$.

Доказательство этого утверждения вытекает из легко проверяемого включения $\delta_\lambda(\cdot) \in S (\forall \lambda \in \Lambda)$, где

$$S = \{ \delta(\cdot) : \| \delta(y) \| \leqq \bar{f}(y) \qquad (\bar{P} \text{ — п. в.}) \} \tag{4.7}$$

есть слабо компактное в пространстве (3.3) множество.

Для дальнейшего определим вероятностную меру P_λ:

$$dP_\lambda(y) = (\int_{\Theta} p(y, \theta) d\lambda(\theta) dv(y). \tag{4.8}$$

Отметим, что $P_\lambda(Y_\lambda) = 1$ $(\forall \lambda \in \Lambda)$, причем $P_\lambda \sim \bar{P}$ на множестве Y_λ.

Суммируем проведенные рассуждения в виде теоремы.

Теорема 4.1. Решение задачи (2.9) в классе допустимых оценок эквивалентно нахождению седловой точки для функционала (4.1). Упомянутая седловая точка $(\bar{\delta}, \bar{\lambda})$ существует, причем $\bar{\delta}(y) = \delta_{\bar{\lambda}}(y)$ $(P_{\bar{\lambda}}$ — почти наверное) и $\|\bar{\delta}(y)\| \leqq \bar{f}(y)$ $(\bar{P}$ — п. в.). Здесь $\delta_{\bar{\lambda}}(\cdot)$ — элемент вида (4.5) при $\lambda = \bar{\lambda}$, а $\bar{f}(\cdot)$ — функция из неравенства (3.2). Мера $\bar{\lambda} \in \Lambda$, характеризующая наихудшее распределение параметров, максимизирует вогнутый по λ функционал (4.4) и обладает тем свойством, что $\bar{\lambda}(\bar{\Theta}) = 1$, где

$$\bar{\Theta} = \{\theta \in \Theta : K(\bar{\delta}(\cdot), \theta) = I(\bar{\delta}(\cdot))\}. \tag{4.9}$$

Доказательство. Функционал (4.1) является выпуклым по $\delta(\cdot) \in \tilde{\Delta}$ и линейным по $\lambda \in \Lambda$. Кроме того, он непрерывен по λ и слабо полунепрерывен снизу по $\delta(\cdot)$ в пространстве (3.3). Так как Λ — компакт, то из теоремы о минимаксе [20] следует равенство (4.3). Ввиду включения $\delta_\lambda(\cdot) \in S(\forall \lambda \in \Lambda)$ нижняя грань слабо полунепрерывного снизу функционала (4.2) достигается на множестве (4.7). Таким образом, седловая точка для функционала (4.1) существует. Далее проинтегрируем равенство (3.5) по мере λ и переставим интегралы согласно теореме Фубини. В итоге приходим к равенству

$$\tilde{K}(\delta, \lambda) = \int_{R^{mt}} \|\delta(y) - \delta_\lambda(y)\|^2 dP_\lambda(y) + \tilde{K}(\delta_\lambda, \lambda),$$

в котором соответствующие величины определены соотношениями (4.4), (4.5) и (4.8). Из этого равенства вытекают утверждения теоремы за исключением последнего. Предположение $\bar{\lambda}(\bar{\Theta}) < 1$ приводит к противоречию, ибо тогда $\tilde{K}(\bar{\delta}, \bar{\lambda}) < I(\bar{\delta}(\cdot))$, что невозможно. Теорема доказана.

Определение 4.1. Меру $\lambda \in \Lambda$ назовем невырожденной, если $\bar{P} \sim P_\lambda$ (см. (4.8)) (эквивалентно: $\bar{P}(Y_\lambda) = \bar{P}(R^{mt})$, где Y_λ — множество (4.6)).

Следствие теоремы 4.1. Пусть существует невырожденная мера $\bar{\lambda} \in \Lambda$, максимизирующая функционал (4.4). Тогда равенство

$$\bar{\delta}(y) = \delta_{\bar{\lambda}}(y) \qquad (P_{\bar{\lambda}} \text{— п. н.}) \tag{4.10}$$

является необходимым и достаточным условием того, что оценка $\bar{\delta}(\cdot) \in \tilde{\Delta}$ доставляет минимум в задаче (2.9). Решение последней задачи единственно (mod $\bar{P}$) при выполнении условия данного следствия.

Вычисление максимума функционала (4.4) по мерам является достаточно сложной математической задачей. Поэтому ниже обсуждаются вопросы конечномерной аппроксимации этой задачи.

5. Конечномерная аппроксимация решения

Установим вначале свойства непрерывности функционала (4.4) и оценки (4.5).

Лемма 5.1. Вогнутый по λ функционал (4.4) непрерывен в смысле слабой сходимости мер в множестве Λ. Если $\lambda_k \to \lambda$ в Λ, то последовательность оценок $\delta_{\lambda_k}(y)\bar{P}$ — п. в. сходится к $\delta_\lambda(y)$ на множестве (4.6).

Отметим, что функционал (4.1) непрерывен по совокупности переменных, если сходимость оценок δ понимать в сильной топологии пространства (3.3), а сходимость мер λ — как слабую сходимость.

Пусть $\lambda_k \to \lambda$ в множестве Λ. На множестве полной $\bar{P}$-меры числитель и знаменатель выражения (4.5) непрерывны по λ. Кроме того, имеют место включения $\delta_{\lambda_k}(\cdot) \in S$ (см. (4.7)). Следовательно, заключения леммы справедливы.

Рассмотрим множество $\Lambda_n \subset \Lambda$ всех вероятностных мер с носителем, содержащимся в конечной n^{-1}-сети компакта Θ, $n = 1, 2, \ldots$. Последовательность n^{-1}-сетей выбираем так, чтобы предыдущая n^{-1}-сеть содержалась в последующей $(n+1)^{-1}$-сети. Тогда имеет включения $\Lambda_1 \subset \Lambda_2 \subset \ldots$. В дальнейшем построенную указанным выше способом последовательность n^{-1}-сетей будем называть возрастающей.

Лемма 5.2. Для любой возрастающей последовательности n^{-1}-сетей множество $\bigcup_{n=1}^{\infty} \Lambda_n$ слабо плотно в Λ.

Доказательство этого утверждения опускается.

Пусть $\{\bar{\lambda}_n\}$ — последовательность мер, каждая из которых максимизирует функционал (4.4) на множестве Λ_n. Хотя бы одна такая последовательность существует, т. к. Λ_n — конечномерный компакт при $\forall n$. Из лемм 5.1, 5.2 следует тогда соотношение

$$d_n = \tilde{K}(\delta_{\bar{\lambda}_n}, \bar{\lambda}_n) \uparrow \max\{\tilde{K}(\delta_\lambda, \lambda) : \lambda \in \Lambda\} \tag{5.1}$$

при $n \to \infty$, т. е. последовательность мер $\{\bar{\lambda}_n\}$ является максимизирующей для функционала (4.4).

Проведенные рассуждения приводят к утверждению

Теорема 5.1. Пусть $(\bar{\delta}_n, \bar{\lambda}_n)$ — последовательность $(\bar{\lambda}_n \in \Lambda_n)$ седловых точек функционала (4.4), подсчитанных при $\delta \in \bar{\Delta}$, $\lambda \in \Lambda_n$. Предположим, что существует возрастающая последовательность n^{-1}-сетей, для которой найдется соответствующая последовательность $\{\bar{\lambda}_n\}$ максимизирующих мер, имеющая хотя бы одну невырожденную предельную точку $\bar{\lambda} \in \Lambda$. Тогда, если $\bar{\lambda}_{n(k)} \to \bar{\lambda}$ при $k \to \infty$, то

$$\lim_{k \to \infty} \delta_{n(k)}(y) = \bar{\delta}(y) \qquad (\bar{P} \text{ — п. в.}), \tag{5.2}$$

где $\bar{\delta}(y)$ — оценка (4.10), доставляющая единственный минимум в задаче (2.9). Кроме того, для $\forall \varepsilon > 0$, $\exists N$ такое, что при $\forall k > N$ пара $(\bar{\delta}_{n(k)}, \bar{\lambda}_{n(k)})$ образует ε — седловую точку функционала (4.4), т. е.

$$-\varepsilon + \tilde{K}(\bar{\delta}_{n(k)}, \lambda) \leqq \tilde{K}(\bar{\delta}_{n(k)}, \bar{\lambda}_{n(k)}) \leqq \tilde{K}(\delta, \bar{\lambda}_{n(k)}),$$

$$\forall \delta \in \tilde{\Delta}, \qquad \forall \lambda \in \Lambda. \tag{5.3}$$

Доказательство. Пусть $\bar{\lambda}_{n(k)} \to \bar{\lambda}$ и $\bar{\lambda}$ — невырожденная мера. Тогда в силу соотношения (5.1) $\bar{\lambda}$ — максимизирующая мера. Из леммы 5.1 следует, что последовательность оценок $\delta_{\bar{\lambda}_{n(k)}}(\cdot)$ $\bar{P}$ — п. в. сходится к $\delta_{\bar{\lambda}}(\cdot) = \bar{\delta}(\cdot)$. Отсюда выводим равенство (5.2), ибо оценка $\bar{\delta}_{n(k)}(\cdot)$ совпадает (mod $\bar{P}$) с $\delta_{\bar{\lambda}_{n(k)}}(\cdot)$ на множестве $Y_{\bar{\lambda}_{n(k)}}$. Предполагая, что (5.3) не выполняется, приходим к противоречию в силу непрерывности соответствующих функций, леммы 5.2 и компактности множества Λ.

Следствие теоремы 5.1. При условии теоремы имеет место равенство

$$\lim_{k \to \infty} I(\bar{\delta}_{n(k)}(\cdot)) = I(\bar{\delta}(\cdot)) = \min_{\delta(\cdot) \in \tilde{\Delta}} I(\delta(\cdot)). \tag{5.4}$$

Замечание 5.1. Если для $\bar{P}$ — почти всех y плотность $p(y, \theta) > 0$ для $\forall \theta \in \bar{\Theta}$ (см. (4.9)), то любая максимизирующая мера $\bar{\lambda} \in \Lambda$ в соответствии с определением 4.1 будет невырожденной. Следовательно, условия теоремы 5.1 и условия следствиа теоремы 4.1 будут выполнены. Более того, в теореме 5.1 в качестве аппроксимирующей может быть взята произвольная возрастающая последовательность n^{-1}-сетей, при этом равенство (5.2) и соотношение (5.3) будут выполняться для всех подпоследовательностей, в частности, для самой последовательности $\{\bar{\delta}_n(\cdot)\}$.

6. Случай линейно-гауссовских систем с неопределенными параметрами

Пусть уравнения (2.1), (2.2) имеют вид

$$x^t = A_t(\theta^t) x^{t-1} + b^t(\theta^t) + \sigma_{1t}(\theta^t) \xi^t, \quad t \geqq 0, \tag{6.1}$$

$$x^{-1} = 0,$$

$$y^t = C_t(\theta^t) x^t + d^t(\theta^t) + \sigma_t(\theta^t) \eta^t, \quad t \geqq 1, \tag{6.2}$$

где матричные непрерывные функции $A_t(\cdot)$, $C_t(\cdot)$, $\sigma_{1t}(\cdot)$, $\sigma_t(\cdot)$ имеют подходящие размерности. Векторные функции $b^t(\cdot)$, $d^t(\cdot)$ также предполагаются непрерывными. Элементы последовательностей $\{\xi^t\}$, $\{\eta^t\}$ суть независимые гауссовские

векторные величины с нулевыми средними и единичными матрицами ковариаций.

Далее будем обозначать величину $E[x^t|y,\theta]$ через $\hat{x}^t(\theta)$. Известно, что эта величина при условии невырожденности

$$\sigma_t(\theta)\sigma_t'(\theta)\geqq\kappa I_m,\quad \forall\, t\geqq 1,\quad \forall\,\theta\in\Theta_t,\quad \kappa>0, \tag{6.3}$$

рекуррентным образом определяется из соотношений фильтра Калмана [12].

Таким образом, величина $\hat{x}^t(\theta)$ линейно зависит от y и непрерывна по θ на компакте Θ. Более того, можно утверждать, что выполняется неравенство

$$\|\hat{x}^t(\theta)\|\leqq a+b\|y\|=\bar{f}(y);\quad a,b>0. \tag{6.4}$$

Векторная величина y имеет в данном случае невырожденное гауссовское распределение с плотностью

$$p(y,\theta)=N(y-\bar{y},F)=(2\pi)^{-mt/2}(\det F)^{-1/2}\cdot$$
$$\cdot\exp\left(-(y-\bar{y})'F^{-1}(y-\bar{y})/2\right), \tag{6.5}$$

где

$$\bar{y}=E_\theta y\in R^{mt},\quad F=E_\theta[(y-\bar{y})(y-\bar{y})']\in R^{mt\times mt}.$$

Для плотности (6.5) справедливо неравенство

$$N(y-\bar{y},F)\leqq kN(y,4dI_{mt})=\bar{p}(y),$$

$$d=\max_\theta\|F\|. \tag{6.6}$$

Соотношения (6.4)–(6.6) показывают, что предположения 3.1, где полагаем $dv(y)=dy$, в данном случае выполняются.

Отметим, что в класс $\bar{\Delta}$ (см. (3.3)) допустимых оценок здесь входят, в частности, все борелевские функции, допускающие полиномиальный рост на бесконечности. Оценка (4.5) будет здесь непрерывной по y и λ функцией, имеющей вид

$$\delta_\lambda(y)=\int_\Theta\hat{x}^t(\theta)N(y-\bar{y},F)d\lambda(\theta)/\int_\Theta N(y-\bar{y},F)d\lambda(\theta). \tag{6.7}$$

Ввиду неравенства (6.4) данная функция имеет линейный по y рост на бесконечности.

По замечанию 5.1 аппроксимация решения (4.10) может быть осуществлена согласно теореме 5.1 с помощью произвольной возрастающей последовательности n^{-1}-сетей компакта Θ.

7. Случай детерминированных систем с неопределенными параметрами

Пусть функции f^t, g^t в уравнениях (2.1), (2.2) не зависят от случайных величин ξ^t, η^t. Тогда система становится детерминированной, а информация о неопределенных параметрах по-прежнему исчерпывается заданием включений (2.3), (2.4). В данном случае в качестве класса допустимых нелинейных оценок в соответствии с неравенством (2.8) выберем пространство Δ ограниченных борелевских отображений $\delta(\cdot): R^{mt} \to R^n$ с нормой

$$\|\delta(\cdot)\|_\infty = \sup_{\theta \in \Theta} \|\delta(y)\|. \tag{7.1}$$

В указанном классе требуется решить задачу (2.9), которая примет вид

$$I(\delta(\cdot)) = \sup_{\theta \in \Theta} \|x^t - \delta(y)\|^2 \to \min_{\delta(\cdot) \in \Delta}. \tag{7.2}$$

Ниже символом $\bar{\delta}(y)$ обозначим чебышевский центр информационного множества $\mathcal{X}_t(y)$, т. е. вектор

$$\bar{\delta}(y) = \arg\min \{\max \{\|z - x\| : x \in \mathcal{X}_t(y)\} : z \in R^n\}, \tag{7.3}$$

в задаче апостериорного оценивания для детерминированной системы (2.1), (2.2). Соответствующие понятия определены в монографии [2] и работе [6].

Справедливо утверждение.

Теорема 7.1. Функция (7.3) принадлежит пространству Δ и доставляет минимум в задаче (7.2). Более того, для любой функции $\delta_1(\cdot) \in \Delta$ со свойством $I(\delta_1(\cdot)) = I(\bar{\delta}(\cdot))$ выполняется равенство $\delta_1(y(\theta^*)) = \bar{\delta}(y(\theta^*))$ для всех точек $\theta^* \in \Theta$, для которых $\|x^t(\theta^*) - \bar{\delta}(y(\theta^*))\|^2 = I(\bar{\delta}(\cdot))$.

Отметим, что для функционала $I(\cdot)$ задачи (7.2) равенство (4.2) также справедливо, если только знаки максимума заменить знаками супремума. Равенство (4.3), вообще говоря, не выполняется. Однако имеет место следующий частный результат.

Теорема 7.2. Пусть компакт Θ состоит из конечного множества точек. Тогда вектор (7.3) представим в виде $\bar{\delta}(y) = E_{\bar{\lambda}}[x^t | y]$, где $\bar{\lambda}$ — вероятностная мера на Θ, доставляющая максимум на Λ вогнутому по λ функционалу $E_\lambda \|x^t - E_\lambda[x^t | y]\|^2$. Здесь E_λ — математическое ожидание, соответствующее распределению λ на Θ. Пара $(\bar{\delta}, \bar{\lambda})$ образует седловую точки функционала (4.1) для рассматриваемой задачи (7.2).

Доказательство теоремы 7.1 проводится методом от противного, а теорема 7.2 следует из изложенного выше.

8. Примеры

1. Пусть заданы одномерные уравнения

$$x = v + \sigma_0 \xi,$$

$$y = x + \sigma\eta, \tag{8.1}$$

где v — неопределенная величина, стесненная ограничением $|v| \leqq 1$. Числа σ_0, σ считаем фиксированными; ξ, η — стандартные независимые гауссовские величины с нулевыми средними и единичными дисперсиями.

Для определения оптимальной нелинейной оценки (4.10) по формулам (6.4)–(6.7) находим $\bar{y} = v$, $F = \sigma^2 + \sigma_0^2$,

$$\delta_\lambda(y) = F^{-1}(\sigma_0^2 y + \sigma^2 \int_{-1}^{1} v \exp(-(y-v)^2/(2F)) \cdot$$

$$d\lambda(v) / \int_{-1}^{1} \exp(-(y-v)^2/(2F)) d\lambda(v)). \tag{8.2}$$

Далее следует найти максимум по мерам функционала типа (4.4) и подставить оптимальную меру в выражение (8.2). Решить задачу максимизации функционала (4.4) можно численно, используя теорему 5.1. Отметим, что оптимальная минимаксная оценка здесь нелинейна по y.

2. Рассмотрим уравнения (8.1), где неопределенный параметр v может принимать лишь два значения $+1$ или -1. Будем считать, что зеличины $\sigma_0\xi$ и $\sigma\eta$ равномерно распределены с одинаковой плотностью распределения:

$$p(x) = \begin{cases} 1/2, & |x| \leqq 1, \\ 0, & |x| > 1, \end{cases}$$

и независимы между собой. Тогда плотность распределения величины y относительно меры Лебега при заданном v будет равна

$$p(y, v) = \begin{cases} (2 - |y-v|)/4, & |y-v| \leqq 2; \\ 0, & |y-v| > 2. \end{cases}$$

Условное математическое ожидание $E[x|y, v] = (y+v)/2$. Формула (4.5) примет вид

$$\delta_\lambda(y) = \begin{cases} (y+1)/2, & 1 \leqq y \leqq 3; \\ y + (2\lambda - 1 + y)/(1 + (2\lambda - 1)y)/2, & |y| \leqq 1; \\ (y-1)/2, & -3 \leqq y \leqq -1. \end{cases}$$

Здесь число $\lambda \in (0, 1)$ есть вес меры $\lambda(\mathrm{dv})$ в точке 1. Максимум функционала (4.4) достигается на мере $\bar{\lambda}(dv)$, имеющей равные веса в точках $+1$ и -1, t. e. $\lambda = 1/2$. Такая мера $\bar{\lambda}$ будет невырожденной в смысле определения 3.1. Следовательно, равенство (4.10) для данного примера примет вид

$$\bar{\delta}(y) = \delta_{\bar{\lambda}}(y) = \begin{cases} (y+1)/2, & 1 \leqq y \leqq 3; \\ y \quad\quad , & |y| \leqq 1; \\ (y-1)/2, & -3 \leqq y \leqq -1. \end{cases} \tag{8.3}$$

Значение функционала $I(\cdot)$ при этом равно 1/4. Отметим, что оптимальная в классе линейных операций оценка

$$\bar{\delta}_1(y) = 4y/5$$

дает значение функционала $I(\bar{\delta}_1(\cdot) = 4/15 > 1/4$.

Если в этом примере считать, что неопределенный параметр заполняет весь отрезок $[-1, 1]$, то вид оптимальной линейной оценки и значение функционала от нее не изменится. Однако уже оценка (8.3) дает более хороший результат, который, в свою очередь, можно еще улучшить, если оптимизировать функционал (4.4) по всем вероятностным мерам на $[-1, 1]$.

9. Обсуждение результатов

Во-первых следует отметить, что оптимальная минимаксная оценка вида (4.10), где $\bar{\lambda}$ — мера, максимизирующая функционал (4.4), существенно зависит от известных распределений величин ξ^t, η^t в уравнениях (2.1), (2.2). Эта оценка будет робастной, дающей гарантированный результат на классе известных с точностью до параметра θ распределений величин x, y (см. (2.6)). В принципе, изложенное выше остается верным и для произвольного метрического компакта Θ. Во-вторых, укажем, что выражение (4.5) содержит условное среднее $E[x^t | y, \theta]$, представляющее собой наилучшую оценку вектора x^t по наблюдениям $y^1, \ldots, y^t$ при известных параметрах θ. Вычисление указанной оценки является достаточно сложной задачей, составляющей предмет математической теории стохастической фильтрации [14]. Предлагаемый подход позволяет учитывать разработанные приближенные методы определения условного среднего при нахождении оптимальных минимаксных оценок. В-третьих, отметим, что процедуры рекуррентного пересчета, которые в данной статье не обсуждаются, для оптимальных минимаксных оценок рассматривались в случае линейных систем в работах [3–6, 9, 11]. Наконец, следует сказать, что изложенные результаты в определенной мере близки общей теории статистических решений [17].

Литература

1. *Красовский Н. Н.* Теория управления движеннем. М., Наука, 1968.
2. *Куржанский А. Б.* Управление и наблюдение в условиях неопределенности. М., Наука, 1977.
3. *Кац И. Я., Куржанский А. Б.* Минимаксное оценивание в многошаговых системах. Докл. АН СССР, 1975, Т. **221,** № *3,* с. 535 < 538.
4. *Бублик Б. Н., Кириченко Н. Ф., Наконечный А. Г.* Минимаксные оценки и регуляторы в динамических системах. Препринт Киевского государственного университета № *31,* 1978, 50 с.
5. *Пшеничный В. Н., Покотило В. Г.* О задачах наблюдения в дискретных системах. Прикл. матем. и механ., Т. **45,** вып. *1,* с. 3 < 10.
6. *Кощеев А. С., Куржанский А. Б.* Адаптивное оценивание эволюции многошаговых систем в условиях неопределенности. Изв. АН СССР. Техн. кибернетика, 1983. № *2,* с. 72 < 93.
7. *Бахшиян Б. Ц., Назиров Р. Р., Эльясберг П. Е.* Определение и коррекция движения. М., Наука, 1980.
8. *Anan'ev, B. I., Kurzhanskiĭ, A. B.,* The nonlinear filtering problem for a multistage system with statistical uncertainty. Second IFAC symp on stoch. control, Vilnius, USSR, 1986: Preprint-M., 1986, Pt. 1. pp. 205–210.
9. *Martin, C. I., Mintz, M.,* Robust filtering and prediction for linear systems with uncertain dynamics: A game-theoretic approach. IEEE Trans. Auto. Control. 1983. V. **AC-28,** No. *9,* pp. 888–896.
10. *Красовский Н. Н.* К теории управляемости и наблюдаемости линейных динамических систем. Прикл. матем. и мех., 1964, т. **28,** вып. *И,* с. 3–14.
11. *Апапьев Б. И.* Минимаксные среднеквадратичные оценки в статистически неопределенных системах. Дифференц. уравнения. 1984. т. **20,** № 8, с. 1291–1297.
12. *K.álmán, R. E.,* A new approach to linear filtering and prediction problems. J. Basic Engr. ASME Trans. 1960, V. **82,** D. pp. 35–45.
13. *Богуславский И. А.* Прикладные задачи фильтрации и управления. М., Наука, 1983.
14. *Липцер Р. Ш., Ширяев А. Н.* Статистика случайных процессов. М., Наука, 1974.
15. *Репин В. Г., Тартаковский Г. П.* Статистический синтез при априорной неопределенности. М., Советское радио, 1977.
16. *Белоглазов И. Н.* Рекуррентно-поисковые алгоритмы оценивания. Докл. АН СССР, 1977, т. **236,** № *2.* с. 292–295.
17. *Вальд А.* Статистические решающие функции. Позиционные игры. М., Наука, 1967, с. 300–522.
18. *Мейер П.* Вероятность и потенциалы. М., Мир, 1973.
19. *Билингели П.* Сходимость вероятностных мер. М., Наука, 1977.
20. *Фань Цзы.* Некоторые теоремы о минимаксе. Бесконечные антагонистические игры. М., Физматгиз, 1963. с. 31–39.

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda Vállalat, 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

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: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

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. TOM 18 · NUMBER HOMEP 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 С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1989

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 PLC Headington Hill Hall, Oxford OX3 OBW, England

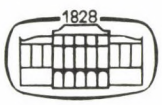
or

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

1989 Subscription Rate DM 535,— 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. 18 (2), pp. 65—77 (1989)

ESTIMATION OF NONLINEAR FUNCTIONALS FROM THE REGRESSION FUNCTION WITH THE POSSIBILITY OF THE REGRESSOR'S DESIGN

YU. I. PASTUCHOVA, R. Z. HASMINSKII

(*Moscow*)

(Received May 10, 1988)

Asymptotical minimax lower bounds of mean square risk for estimators of a differentiable functional on regression with the possibility of experiment design are obtained. The sample design and the estimator, for which this bound is asymptotically tight, are constructed in the case of the known conditional variance of the "noise", if some conditions of smoothness for a functional and an unknown regression function are satisfied.

1. Statement of the problem

Let $P(\cdot\,|t)$, $t \in [0, 1]$ be the family of the conditional distributions on R^1. We assume that this family is determined but unknown to the statistician. Assume that it is possible to observe the values $X_1(t_1), \ldots, X_n(t_n)$ for arbitrary time points $t_1, \ldots, t_n$ and $X_1(t_1), \ldots, X_n(t_n)$ are conditionally independent for a given design $t^{(n)} = (t_1, \ldots, t_n)$, and the conditional (for a given t_i) distribution of $X_i(t_i) - E\{X_i(t_i)|t_i\}$ coincides with $(P(\cdot\,|t_i)$.

The value $X_i = X_i(t_i)$ can be written in the form

$$X_i = R(t_i) + \xi_i(t_i), \qquad R(t) = E\{X_i(t_i)|t_i = t\}.$$

Here $\xi_i(t_i)$ is the "noise", $E\{\xi_i(t_i)|t_i\} = 0$, and $\xi_1(t_1), \ldots, \xi_n(t_n)$ are conditionally independent for a given design $t^{(n)}$, and $R(t)$ is the regression function.

The problem is to estimate the value $F(R)$, F being a known functional defined on $L_2[0, 1] = L_2$, by means of the known values $t_1, X_1, \ldots, t_n, X_n$. The admissible (in the sense of [1]) design $t^{(n)}$ and estimator $\hat{F}_n$, for which the value $\Delta_n = E\{\hat{F}_n - F(R)\}^2$ is minimal in the asymptotically minimax sense, are constructed.

Similarly to [1] we distinguish two cases:

1. Only the conditional variance of the noise

$$\sigma^2(t) = E\{\xi_i^2(t_i)|t_i = t\}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

is known, and

$$0 < \inf_{[0,1]} \sigma(t) \leqq \sup_{[0,1]} \sigma(t) < \infty, \quad \sigma \in L_2, \tag{1.1}$$

$$\sigma \in C[0,1] \tag{1.1'}$$

(where $C[0,1]$ is the space of continuous functions on $[0,1]$).

2. The distribution of $\xi_i(t_i)$ for a given $t_i = t$ is known, and its density function $p(x|t)$ with respect to the Lebesgue measure in R^1 is absolutely continuous in x, has the finite Fisher's information

$$I(t) = \int_{R'} \frac{(p'_x(x|t))^2}{p(x|t)}\, dx$$

and satisfies the following regularity conditions

$$0 < \inf_{[0,1]} I(t) \leqq \sup_{[0,1]} I(t) < \infty, \quad \sup_{[0,1]} \int |x| p(x|t)\, dx < \infty, \tag{1.2}$$

$$\sup_{[0,1]} \int \left| \frac{p'_x(x+s|t)}{p^{1/2}(x+s|t)} - \frac{p'_x(x|t)}{p^{1/2}(x|t)} \right|^2 dx \to 0 \ (s \to 0). \tag{1.3}$$

2. Lower bounds

The following theorem gives a lower bound for quality of an arbitrary estimator of any smooth functional for any admissible design $t^{(n)}$.

Let us remind, that the functional $F(R)$ is Fréchet differentiable in the point R_0 in L_2 if

$$F(R_0 + h) - F(R_0) = \int_0^1 F'(R_0, t) h(t)\, dt + o(\|h\|), \qquad (\|h\| \to 0).$$

Theorem 2.1. Let conditions (1.2), (1.3) be satisfied, the functional $F(R)$ is Fréchet differentiable on the compact $\mathscr{P} \subset L_2$, and the derivative $F'(R, t)$ satisfies the Hölder condition in L_2 with some index $\alpha > 0$

$$\|F'(R_2, \cdot) - F'(R_1, \cdot)\| \leqq C \|R_2 - R_1\|^{\alpha}, \qquad (\alpha \leqq 1).$$

Let $R_0 \in \mathscr{P}$, and there exist $\delta_k > 0$ and $\varphi_k(t)$, $t \in [0, 1]$ $k = 1, 2, \ldots$ such that $R_0(t) + s\varphi_k(t) \in \mathscr{P}$ for all k with $|s| < \delta_k$, and

$$|\varphi_k(t)| \leqq I^{-1/2}(t),$$

$$\int_0^1 \varphi_k(t) F'(R_0, t)\, dt \to \int_0^1 I^{-1/2}(t) |F'(R_0, t)|\, dt, \qquad (k \to \infty). \tag{2.1}$$

Then, for any estimator $\tilde{F}_n$ of $F(R)$, the inequality

$$\varliminf_{n\to\infty}\left[n \sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2 \right] \geqq \left(\int_0^1 I^{-1/2}(t)|F'(R_0, t)|dt \right)^2 \tag{2.2}$$

is true.

Proof. Let us consider the parametric family

$$R_h^k(t) = R_0(t) + (h-\theta)g_k(t), \quad \theta = F(R_0),$$

$$g_k(t) = \varphi_k(t)\lambda_k^{-1}, \quad \lambda_k = \int_0^1 \varphi_k(t)F'(R_0, t)dt, \qquad k = 1, 2, \ldots \tag{2.3}$$

It is clear that $(g_k(\cdot), F'(R_0, \cdot)) = 1$ (where $(\cdot, \cdot)$ is the inner product in L_2). Condition (2.1) implies that $R_h^k \in \mathscr{P}$ for $|h-\theta| < \delta_k\lambda_k$.

Therefore, the inequality

$$\sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2 \geqq \sup_{|h-\theta|<\delta_k\lambda_k} E\{\tilde{F}_n - F(R_h^k)\}^2$$

is true for any estimator $\tilde{F}_n$.

Put $\varepsilon = n^{-1/2}$. Let $\delta(\varepsilon)$, $\varepsilon > 0$ be a function such that $\delta(\varepsilon)\to\infty$, $\varepsilon\delta(\varepsilon)\to 0$ if $\varepsilon\to 0$. Then the relation

$$\sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2 \geqq \sup_{|h-\theta|<\varepsilon\delta(\varepsilon)} \{\tilde{F}_n - F(R_h^k)\}^2 \tag{2.4}$$

is true for sufficiently small ε.

Lagrange theorem implies equality $(0 < \bar{\lambda} < 1)$

$$F(R_h^*) = F(R_0 + (h-\theta)g_k) = \theta + (h-\theta)(g_k(t), F'(R_0 + \bar{\lambda}(h-\theta)g_k, t)).$$

From this equality, the Hölder condition and (2.3) follows that for $|h-\theta| < \varepsilon\delta(\varepsilon)$ we have

$$(h - F(R_h^k))^2 = (h-\theta)^2 (g_k(t), F'(R_0, t) - F'(R_0 + \bar{\lambda}(h-\theta)g_k, t))^2 \leqq$$

$$\leqq C(h-\theta)^{2(\alpha+1)}\|g_k\|^{2(\alpha+1)} \leqq C(\varepsilon\delta(\varepsilon))^{2(\alpha+1)}. \tag{2.5}$$

Let us consider now for every $k = 1, 2, \ldots$ the estimation problem of parameter h, by the known sample $t^{(n)}$, $X^{(n)}$ with the joint distribution

$$dP_{k,h}^{(n)}(t^{(n)}, X^{(n)}) = Q_1(dt_1)p(x_1 - R_0(t_1) -$$

$$-(h-\theta)g_k(t_1)|t_1)Q_2(dt_2|t_1, x_1)\ldots$$

$$\ldots Q_n(dt_n|t^{(n-1)}, x^{(n-1)})p(x_n - R_0(t_n) -$$

$$-(h-\theta)g_k(t_n)|t_n)dx_1, \ldots, dx_n$$

(here $Q_i(dt_i|\dots)$ is the probability distribution for the choice t_i for a given $t^{(i-1)}$, $X^{(i-1)}$ (see [1] for details). Fisher's information about h in $t^{(n)}$, $X^{(n)}$ can be written in the form

$$I_k^{(n)}(h)=\int\dots\int\left(\frac{\partial}{\partial h}\ln\frac{P_{k,\theta+h}^{(n)}(t^{(n)},X^{(n)})}{P_{k,\theta}^{(n)}(t^{(n)},X^{(n)})}\right)^2 dP_{k,\theta+h}^{(n)}(t^{(n)},X^{(n)})=$$

$$=\sum_{i=1}^{n}\int_0^1\varphi^2(t_i)I(t_i)\tilde{Q}_i(dt_i).$$

Here

$$\tilde{Q}_1(dt_1)=Q_1(dt_1),\ \tilde{Q}_2(dt_2)=\iint p(x_1-R_0(t_1)-hg_k(t_1)|t_1)Q_2(dt_2|t_1,x_1)\times$$

$$\times Q_1(dt_1)dx_1,\dots,\tilde{Q}_n(dt_n)=\int\dots\int p(x_{n-1}-R_0(t_{n-1})-hg_k(t_{n-1})|t_{n-1})dx_n\times$$

$$\times Q_n(dt_n|t^{(n-1)},x^{(n-1)})\dots p(x_1-R_0(t_1)-hg_k(t_1)|t_1)Q_1(dt_1)dx_1.$$

Relation $I_k^{(n)}(h)\leqq n\lambda_k^{-2}$ is clear from (2.1). The Rao–Kramer inequality is true for any estimator of h in the parametric family $P_{k,h}^{(n)}$, if conditions (1.2)–(1.3) are satisfied (see [2], p. 104). Therefore, the inequality

$$E_h\{\tilde{F}_n-h\}^2\geqq(I_k^{(n)}(h))^{-1}(1+b_k'(h))^2+b_k^2(h)\geqq\frac{\lambda_k^2}{n}(1+b_k'(h))^2+b_k^2(h)$$

is also true. Here $b_k(h)$ is the bias of estimator $\tilde{F}_n$. Consequently

$$\sup_{|h-\theta|<\varepsilon\delta(\varepsilon)}E_h\{\tilde{F}_n-h\}^2\geqq(2\varepsilon\delta(\varepsilon))^{-1}\int_{\theta-\varepsilon\delta(\varepsilon)}^{\theta+\varepsilon\delta(\varepsilon)}E_h\{\tilde{F}_n-h\}^2dh\geqq$$

$$\geqq(2\varepsilon\delta(\varepsilon))^{-1}\int_{\theta-\varepsilon\delta(\varepsilon)}^{\theta+\varepsilon\delta(\varepsilon)}\left[\frac{\lambda_k^2}{n}(1+b_k'(h))^2+b_k^2(h)\right]dh.$$

Let us denote $\varepsilon_k=\lambda_k\varepsilon$. The application of Euler's equation in the calculus of variations leads to the relation

$$\inf_{y(h)}\frac{1}{b-a}\int_a^b[\varepsilon_k^2(1+y'(h))^2+y^2(h)]dh\geqq\varepsilon_k^2-C_1\frac{\varepsilon^3}{b-a}.$$

Here C_1 is absolute constant.

So, we have established for all estimators $\tilde{F}_n$ the inequality

$$\sup_{|h-\theta|<\varepsilon\delta(\varepsilon)}E_h\{\tilde{F}_n-h\}^2\geqq\varepsilon^2\lambda_k^2+o(\varepsilon^2).$$

Notice further that the inequality

$$(\tilde{F}_n - F(R_h^k))^2 \geqq (\tilde{F}_n - h)^2(1-\gamma^{-1}) - (h - F(R_h^k))^2\gamma$$

is true for any $\gamma > 0$. The two latter inequalities, and (2.5) finally imply the relation

$$\sup_{|h-\theta|<\varepsilon\delta(\varepsilon)} E_h\{\tilde{F}_n - F(R_h^k)\}^2 \geqq (1-\gamma^{-1})(\varepsilon^2\lambda_k^2 + o(\varepsilon^2)) - C(\varepsilon\delta(\varepsilon)^{2(\alpha+1)}\gamma.$$

Choose now $\delta(\varepsilon) = \varepsilon^{-\alpha/4}$, $\gamma = \gamma(\varepsilon) = \varepsilon^{-\alpha/2}$. Then, using (2.4) and remembering the relation $\varepsilon^2 = n^{-1}$, we obtain the assertion of the theorem.

Remark. The Hölder condition for $F'(R, \cdot)$ can be weakened, more precisely, it can be changed to the condition

$$\|(F'(R_2, \cdot) - F'(R_1, \cdot)\| \leqq \beta(\|R_2 - R_1\|), \quad R_i \in \mathscr{P}, \quad i = 1, 2$$

for some function $\beta(x)$ satisfying $\beta(x) < C \ln^{-1}(1/x)$. The proof is similar, it is necessary only to choose functions $\gamma(x)$, $\delta(x)$ by another way, putting, for example,

$$\gamma(x) = \delta^2(x) = \ln\ln(1/x).$$

Corollary. Let the conditions of Theorem 2.1 be satisfied and

$$p(x|t) = (2\pi\sigma^2(t))^{-1/2} \exp\left\{-\frac{x^2}{2\sigma^2(t)}\right\}. \tag{2.6}$$

Then $I(t) = \sigma^{-2}(t)$ and Theorem 2.1 implies the relation

$$\underline{\lim}_{n\to\infty}\left[n \sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2\right] \geqq \left(\int_0^1 \sigma(t)|F'(R_0, t)|dt\right)^2.$$

Let us denote by $\mathscr{P}_\sigma$ the set of $p(x|t)$, for which

$$E\{\xi_i^2(t_i)|t_i = t\} = \sigma^2(t).$$

The following theorem is the consequence of Theorem 2.1 and the Corollary.

Theorem 2.2. Let the conditions of Theorem 2.1 with the substitute ((1.2), (1.3))⇒(1.1) be satisfied. Then, for any estimator $\tilde{F}_n$, we have the inequality

$$\underline{\lim}_{n\to\infty}\left[n \sup_{R\in\mathscr{P},\, p(x|t)\in\mathscr{P}_\sigma} E\{\tilde{F}_n - F(R)\}^2\right] \geqq \left(\int_0^1 \sigma(t)|F'(R_0, t)|dt\right)^2. \tag{2.7}$$

It follows from (2.2) and (2.7), that the mean square error of any estimator of $F(R)$, multiplied by n, is not less (in the asymptotically minimax sense) than the value $\Phi_1(R, I) = (\int I^{-1/2}(t)|F'(R, t)|dt)^2$, if the conditional distribution of the noise is known, and the value $\Phi_2(R, \sigma) = (\int \sigma(t)|F'(R, t)|dt)^2$ if the conditional variance of the noise is only known. (More precisely, this conclusion is true, if Φ_i $(i = 1, 2)$ are continuous in L_2, relative to R.) Therefore, the following definitions are natural.

Definition 1. The estimator $\hat{F}_n$ is asymptotically effective in K nonparametrical estimator (AENE) of $F(R)$ for the noise conditional density $p(x|t)$, if

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 I^{-1/2}(t)|F'(R,t)|dt\right)^2\right]\to 0 \qquad (n\to\infty).$$

Definition 2. The estimator $\hat{F}_n$ is AENE of $F(R)$ for the noise conditional variance $\sigma^2(t)$, if

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R,t)|dt\right)^2\right]\to 0 \qquad (n\to\infty).$$

AENE in the sense of Definition 2 is constructed in Section 3. This estimator is AENE in the sense of Definition 1, if $p(x|t)$ is determined from (2.6).

3. Asymptotically effective nonparametrical estimator

Denote by $W_2^\beta(L)$ the set of $\tau=\lceil\beta\rceil$ times differentiable functions on $[0, 1]$, satisfying the relation

$$\|f^{(\tau)}(t+h)-f^{(\tau)}(t)\|\leqq L|h|^\rho, \qquad \beta=\tau+\rho, \quad 0<\rho<1.$$

Denote also by $W_2^\beta(L_1, T)$ the set of periodical with period $T>1$ functions, having absolutely continuous derivative of order $\lceil\beta\rceil-1$ for which ($\tau=\lceil\beta\rceil$)

$$\int_a^{a+T}(f^{(\tau)}(t+h)-f^{(\tau)}(t))^2dt\leqq L_1|h|^{2\rho}, \qquad \beta=\tau+\rho, \quad 0<\rho\leqq 1.$$

It is supposed further that

$$R\in K\subset W_2^\beta(L), \qquad \beta>1/2, \tag{3.1}$$

here K is a known compact in $L_2[0, 1]$. Then

$$\sup_{R\in K}|R(0)|<\infty. \tag{3.2}$$

Conditions (1.1′), (3.1) guarantee the uniform continuity of $R(t)/\sigma(t)$ and the boundedness of R on K, i.e.

$$\sup_{R\in K}\max_{[0,1]}|R(t)|<\infty. \tag{3.1′}$$

Remark. It is known (see [5]) that condition (3.1) guarantees the possibility of the extension of the function $R(t)$ in such a way that this extension R_1 would coincide with R for $t\in[0, 1]$, and would be periodic with period $T>1$, and finally would

belong to $W_2^\beta(L_1, T)$ for some $L_1 > 0$. We shall use this extension later and denote it by the same letter R.

The following smoothness conditions of $F(R)$ will be used.

(F1) The functional F: $L_2 \to R^1$ is Frechet differentiable in L_2 for $R \in K$.

(F2) The derivative $F'(R, \cdot)$ satisfies the Hölder condition in L_2

$$\|F'(R_1, \cdot) - F'(R_2, \cdot)\| < C_1 \|R_1 - R_2\|^\alpha, \qquad R_i \in K$$

with some α satisfying the inequalities

$$(2\beta)^{-1} < \alpha \leqq 1.$$

(F3) The derivative $F'(R, t)$ satisfies the Hölder condition in t with some index $\gamma > 0$ uniformly in $R \in K$:

$$|F'(R, t_1) - F'(R, t_2)| < C_2 |t_1 - t_2|^\gamma, \qquad R \in K, \quad t_i \in [0, 1].$$

Now we describe the construction of the asymptotically effective estimator.

Let us divide the sample $t_1, X_1, \ldots, t_n, X_n$ into two parts. The first one has the volume $n_0 = [n^\varkappa]$ with $\varkappa$ satisfying the inequalities

$$\frac{2\beta + 1}{2\beta(\alpha + 1)} < \varkappa < 1. \tag{3.3}$$

(Such a choise is possible, see condition (F2)).

The second part of the sample has the volume $n_1 = n - n_0$.

It is known (see [3]) that for $R \in W_2^\beta(L_1, T)$ there exist a sample design $t^{(n_0)}$ and based on $t^{(n_0)}$, $X^{(n_0)}$ an estimator $\hat{R}_{n_0}(t)$, for which $\hat{R}_{n_0} \in K$ almost sure and

$$\sup_{R \in K} E\|\hat{R}_{n_0} - R\|^2 \leqq M n_0^{-\frac{2\beta}{2\beta+1}} \leqq M n^{-\frac{2\beta\varkappa}{2\beta+1}}. \tag{3.4}$$

The remaining part of the sample is used for the linear functional

$$L_{n_0}(R) = \int_0^1 R(t) F'(\hat{R}_{n_0}, t) dt$$

estimation. Our estimator is similar to the one from [4]. Let $\mathcal{F}_{n_0}$ be a σ-algebra, generated by $t^{(n_0)}$, $X^{(n_0)}$. The sample design $(t_{n_0+1}, \ldots, t_n) = t^{(n_1)}$ is chosen as conditionally independent for given $\mathcal{F}_{n_0}$, and the corresponding conditional density for t_i, $i = n_0 + 1, \ldots, n$ is

$$p_{n_0}(t) = \sigma(t) |F'(\hat{R}_{n_0}, t)| \left(\int_0^1 \sigma(t) |F'(\hat{R}_{n_0}, t)| dt \right)^{-1}.$$

Let

$$N = [c n_1 / \ln n_1], \qquad 0 < c < 1.$$

Let us divide [0, 1] into N mutually disjoint parts $\Delta_1,\ldots,\Delta_N$ by points $a_1,\ldots,a_{N-1}$ so that

$$\int_{\Delta_k} \sigma(t)|F'(\hat{R}_{n_0}, t)|dt = \frac{1}{N}\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|dt. \tag{3.5}$$

Then

$$P\{t_i \in \Delta_k | \mathscr{F}_{n_0}\} = \int_{\Delta_k} p_{n_0}(t)dt = N^{-1}. \tag{3.6}$$

Let us denote by ν_k the number of elements from $t^{(n_1)}$, belonging to Δ_k and consider the estimator

$$\hat{L}_{n_1} = N^{-1}\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|dt \sum_{k=1}^{N} \nu_k^{-1} \sum_{t_i \in \Delta_k} \frac{X_i \operatorname{sign} F'(\hat{R}_{n_0}, t_i)}{\sigma(t_i)}. \tag{3.7}$$

(We assume the agreement here and later that $\nu_k^{-l}\sum_{t_i\in\Delta_k}(\ldots)=0$ for $\nu_k=0$, $l=1, 2$.)

The final estimator for $F(R)$ has the form

$$\hat{F}_n = F(\hat{R}_{n_0}) + \hat{L}_{n_1} - \int_0^1 \hat{R}_{n_0}(t)F'(\hat{R}_{n_0}, t)dt. \tag{3.8}$$

Our aim now is to prove the inequality ($n\to\infty$)

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R, t)|dt\right)^2\right] \leqq o(1), \tag{3.9}$$

i.e. the AENE-property for $\hat{F}_n$ (see Section 2).

The equality

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R, t)|dt\right)^2\right] =$$

$$= \sup_{R\in K}\left[nE\{F(\hat{R}_{n_0}) - F(R) + \int_0^1 F'(\hat{R}_{n_0}, t)\,(R(t) - \hat{R}_{n_0}(t))dt - \right.$$

$$\left. - \int_0^1 F'(\hat{R}_{n_0}, t)R(t)dt + \hat{L}_{n_1}\}^2 - \left(\int_0^1 \sigma(t)|F'(R, t)|dt\right)^2\right]$$

is trivial. The Lagrange theorem gives (see e.g. [6])

$$F(\hat{R}_{n_0}) - F(R) = -\int_0^1 F'(\hat{R}_{n_0} + \theta(R - R_{n_0}), t)\,(R(t) - \hat{R}_{n_0}(t))dt, \qquad 0<\theta<1.$$

These two equalities imply the relation

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R, t)|dt\right)^2\right] \leqq$$

$$\leqq \sup_{R\in K}\left[nE\left\{\int_0^1 (F'(\hat{R}_{n_0},t)-F'(\hat{R}_{n_0}+\hbar(R-\hat{R}_{n_0}),t))\,(R(t)-\hat{R}_{n_0}(t))dt\right\}^2\right]+$$

$$+2\sup_{R\in K}\left[nE\left|\left\{\int_0^1 (F'(\hat{R}_{n_0},t)-F'(\hat{R}_{n_0}+\hbar(R-\hat{R}_{n_0}),t))\,(R(t)-\hat{R}_{n_0}(t))dt\times\right.\right.$$

$$\left.\left.\times(\hat{L}_{n_1}-L_{n_0}(R))\right\}\right|\Bigg]+\sup_{R\in K}\left[nE\{\hat{L}_{n_1}-L_{n_0}(R)\}^2-\left(\int_0^1\sigma(t)|F'(R,t)|dt\right)^2\right].$$

Conditions (F2), (3.4) and (3.3) give the following relation

$$\sup_{R\in K}\left[nE\left\{\int_0^1 (F'(\hat{R}_{n_0},t)-F'(\hat{R}_{n_0}+\theta(R-\hat{R}_{n_0}),t))\,(R(t)-\hat{R}_{n_0}(t))dt\right\}^2\right]\leqq$$

$$\leqq C_1\sup_{R\in K} nE\|R-R_{n_0}\|^{2(\alpha+1)}\leqq C_1 Mn^{1-\frac{2\varkappa\beta(\alpha+1)}{2\beta+1}}=o(1),\qquad (n\to\infty).\tag{3.10}$$

Let us denote

$$A_n=\sup_{R\in K}\left[nE\left\{\hat{L}_{n_1}-L_{n_0}(R)\right\}^2-\left(\int_0^1\sigma(t)|F'(R,t)|dt\right)^2\right].\tag{3.11}$$

It follows from (3.10), that the inequality

$$\overline{\lim_{n\to\infty}}\,A_n=A_0\leqq 0$$

implies (3.9). It is easy to verify (similarly to (3.10)) that

$$\sup_{R\in K}E\left\{\int_0^1 F'(\hat{R}_{n_0},t)\,R(t)dt\right\}^2\leqq\sup_{R\in K}E\|R(\cdot)F'(\hat{R}_{n_0},\cdot)\|^2\leqq$$

$$\leqq\sup_{R\in K}\|R(\cdot)F'(R,\cdot)\|^2+o(1)<\infty,$$

$$\sup_{R\in K}E\left\{\int_0^1\sigma(t)|F'(\hat{R}_{n_0},t)|dt\right\}^2\leqq\sup_{R\in K}\left\{\int_0^1\sigma(t)|F'(R,t)|dt\right\}^2+$$

$$+o(1)<\infty.\tag{3.12}$$

Let us denote, similarly to [4],

$$\zeta_k=v_k^{-1}\sum_{t_i\in\Delta_k}\frac{R(t_i)\operatorname{sign}F'(\hat{R}_{n_0},t_i)}{\sigma(t_i)}\int_0^1\sigma(t)|F'(\hat{R}_{n_0},t)|dt,$$

$$R_k=N\int_{\Delta_k}R(t)F'(\hat{R}_{n_0},t)dt.$$

It is clear that

$$A_n \leqq \sup_{R\in K} \frac{n}{N^2} E\left\{ E\left[\left(\sum_{k=1}^{N} (\zeta_k - R_k) \right)^2 \middle| \mathscr{F}_{n_0} \right] \right\} + \sup_{R\in K} \left[\frac{n}{N^2} \times \right.$$

$$\times E\left\{ \left(\int_0^1 \sigma(t) |F'(\hat{R}_{n_0}, t)| dt \right)^2 \sum_{k=1}^{N} E\left[\nu_k^{-2} \sum_{i=n_0+1}^{n} \chi(t_i \in \Delta_k) | \mathscr{F}_{n_0} \right] \right\} -$$

$$\left. -\left(\int_0^1 \sigma(t) |F'(R, t)| dt \right)^2 \right].$$

The first term in the right-hand part of (3.13) can be estimated using the arguments of Theorem 3.1 from [4] and (3.6), (3.12). The new point here is the proof of the smallness of the expression ($N \to \infty$)

$$\sup_{R\in K} N \sum_{k=1}^{N} E\left\{ \int_{\Delta_k} \frac{R^2(t)}{\sigma(t)} |F'(\hat{R}_{n_0}, t)| dt \int_{\Delta_k} \sigma(t) |F'(\hat{R}_{n_0}, t)| dt - \right.$$

$$\left. - \left(\int_{\Delta_k} R(t) F'(\hat{R}_{n_0}, t) dt \right)^2 \right\}.$$

Let us prove this. The first step is the equality

$$\sup_{R\in K} N \sum_{k=1}^{N} E\left\{ \int_{\Delta_k} \frac{R(t)}{\sigma(t)} |F'(\hat{R}_{n_0}, t)| dt \int_{\Delta_k} \sigma(t) |F'(\hat{R}_{n_0}, t)| dt - \right.$$

$$\left. - \left(\int_{\Delta_k} R(t) F'(\hat{R}_{n_0}, t) dt \right)^2 \right\} = \sup_{R\in K} N \sum_{k=1}^{N} E\left\{ \int_{\Delta_k} \int_{\Delta_k} \left(\frac{R(t)}{\sigma(t)} \operatorname{sign} F'(\hat{R}_{n_0}, t) - \right. \right.$$

$$\left. \left. - \frac{R(s)}{\sigma(s)} \operatorname{sign} F'(\hat{R}_{n_0}, s) \right)^2 |F'(\hat{R}_{n_0}, t)| |F'(\hat{R}_{n_0}, s)| \sigma(t)\sigma(s) dt\, ds \right\}.$$

The latter sum can be divided into two parts:

$$\sum\nolimits_1 = N \sum_{k:|\Delta_k| > N^{-\varepsilon}} (\ldots) \qquad \text{and} \qquad \sum\nolimits_2 = N \sum_{k:|\Delta_k| \leqq N^{-\varepsilon}} (\ldots).$$

Here $\varepsilon \in]\,(1+\gamma)^{-1}, 1[$. Obviously, the number of terms in $\sum_1$ is less than N^{ε}. Recalling also (3.5) we obtain

$$\sup_{R\in K} \sum\nolimits_1 \leqq G N^{\varepsilon+1} \sup_{R\in K} E\left\{ \int_{\Delta_k} \sigma(t) |F'(\hat{R}_{n_0}, t)| dt \right\}^2 =$$

$$= G\frac{N^{1+\varepsilon}}{N^2}\sup_{R\in K} E\left\{\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\,dt\right\}^2 \leqq GN^{\varepsilon-1}\to 0, \qquad (N\to\infty).$$

(Here and further we denote by G a positive, not necessary the same constant.)

The sum $\sum_2$ can be divided into two parts, too: $\sum_{21}$ includes sum over such k, for which $|\Delta_k|\leqq N^{-\varepsilon}$, and the function $|F'(\hat{R}_{n_0}, t)|$ do not change the sign in Δ_k, and $\sum_{22}=\sum_2-\sum_{21}$. We have the following inequality for $\sum_{21}$

$$N\sup_{R\in K}\sum_{21}\leqq\rho(N^{-\varepsilon})\sup_{R\in K} E\left\{\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\,dt\right\}^2\leqq$$

$$\leqq G\rho(N^{-\varepsilon})\to 0, \qquad (N\to\infty).$$

Here $\rho(t)$ is the continuity module of the function R/σ. Finally, for Δ_k belonging to $\sum_{22}$ the inequality

$$|F'(\hat{R}_{n_0}, t)|<\tilde{\rho}(N^{-\varepsilon})$$

is true. Here $\tilde{\rho}(t)$ is the continuity module of the function $F'(\hat{R}_{n_0}, t)$. It follows from (F3) that $\tilde{\rho}(t)\leqq C_2 t^{\gamma}$. Therefore, recalling that $\varepsilon>(\gamma+1)^{-1}$, we obtain the relation

$$\sup_{R\in K} N\sum_{22}\leqq G\frac{N^{\varepsilon}}{N^{2\gamma\varepsilon}}\cdot\frac{1}{N^{2\varepsilon}}\sup_{R\in K} E\left\{\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\,dt\right\}^2\leqq$$

$$\leqq GN^{2(1-\varepsilon(\gamma+1))}\to 0, \quad (N\to\infty).$$

To prove (3.9) it remains to show that

$$\sup_{R\in K}\left[\frac{n}{N^2} E\left\{\left(\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\,dt\right)^2\sum_{k=1}^{N} E\left[v_k^{-2}\sum_{i=n_0+1}^{n}\chi(t_i\in\Delta_k)|\mathscr{F}_{n_0}\right]\right\}-\right.$$

$$\left.-\left(\int_0^1 \sigma(t)|F'(R, t)|\,dt\right)^2\right]=o(1), \qquad (n\to\infty).$$

In fact, similarly to Theorem 3.1, we establish by Lemma 3.1 from [4] that

$$\sum_{k=1}^{N} E\left[v_k^{-2}\sum_{i=n_0+1}^{n}\chi(t_i\in\Delta_k)|\mathscr{F}_{n_0}\right]=\sum_{k=1}^{N}\sum_{i=n_0+1}^{n} E\left[\frac{\chi(t_i\in\Delta_k)}{(1+\sum_{j\neq i}\chi(t_j\in\Delta_k))^2}\middle|\mathscr{F}_{n_0}\right]=$$

$$=\sum_{k=1}^{N}\sum_{i=n_0+1}^{n} E\left\{\left[(\chi(t_i\in\Delta_k)|\mathscr{F}_{n_0}\right]E\left[\left(1+\sum_{\substack{j=n_0+1\\ j\neq i}}^{n}\chi(t_j\in\Delta_k)\right)^{-2}|\mathscr{F}_{n_0}\right]\right\}=\frac{N^2}{n_1}+o(1).$$

Finally, using (3.12) we have

$$\sup_{R\in K}\left[\frac{n}{N^2} E\left\{\left(\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\,dt\right)^2\sum_{k=1}^{N} E\left[v_k^{-2}\sum_{i=n_0+1}^{n}\chi(t_i\in\Delta_k)|\mathscr{F}_{n_0}\right]\right\}-\right.$$

$$-\left(\int_0^1 \sigma(t)|F'(R,t)|\,dt\right)^2\Bigg] = \sup_{R\in K}\left[\frac{n}{n_1}E\left\{\int_0^1 \sigma(t)|F'(\hat{R}_{n_0},t)|\,dt\right\}^2 - \right.$$

$$\left. -\left(\int_0^1 \sigma(t)|F'(R,t)|\,dt\right)^2\right] + o(1) = o(1).$$

Thus, the following theorem is proved.

Theorem. Let conditions (1.1)–(1.1′) be satisfied, $R \in K \subset W_2^{\beta}(L)$ with $\beta > 1/2$, functional $F(R)$ safisfies conditions (F1)–(F3). Let also the estimator F_n be defined by equality (3.8) with $n_0 = [n^{\varkappa}]$, $N = [n_1 c/\ln n_1]$ where $\varkappa$ and c are any constants satisfying

$$\frac{2\beta+1}{2\beta(\alpha+1)} < \varkappa < 1, \qquad 0 < c < 1.$$

Then this estimator $\hat{F}_n$ is asymptotically efficient in K nonparametric estimator of the functional $F(R)$ of regression function for square loss function, i.e.

$$\lim_{n\to\infty} \sup_{R\in K} [nE\{\hat{F}_n - F(R)\}^2 - (\int_0^1 \sigma(t)|F'(R,t)|\,dt)^2] = 0.$$

References

1. *Has'minskii, R. Z.*, On nonparametric estimation of the linear functional from regression under observation design. Problemy peredachi informatsii, 1986, **22,** 3, pp. 43–61.
2. *Ibragimov, I. A., Has'minskii, R. Z.*, Statistical estimation. Asymptotic theory. Springer, 1981.
3. *Chentsov, N. N.*, On estimation of unknown mean in multidimensional Gaussian distribution. Probability theory and its applications, 1967, **12,** *4*, pp. 619–633.
4. *Nikolskii, S. M.*, On the continuation of functions from several variables with the preservation of the differential properties. Matematicheskii Sbornik, 1956, **40** (82), pp. 243–268.
5. *Ibragimov, I. A., Has'minskii, R. Z.*, The asymptotic bound of the quality of nonparametric regression estimation in L_p. Zapiski nauchnych seminarov LOMI, 1981, **97,** pp. 88–101.
6. *Pastuhova, Ju. I., Has'minskii, R. Z.*, Asymptotically effective estimators of the linear functional from the regression function under the fixed observation design. Problemy peredachi informatsii, 1988, **24,** *3*, pp. 42–51.
7. *Cartan, H.*, Calcul differential formes differentielles. Hermann, Paris, 1967.

Оценивание нелинейных функционалов от регрессии при возможности планирования

Ю. И. ПАСТУХОВА, Р. З. ХАСЬМИНСКИЙ

(Москва)

Рассмотрена задача оценивания некоторого гладкого функционала $F(R)$ от функции регрессии $R(t)$ на отрезке $[0, 1]$ при возможности планирования наблюдений $X_1(t_1), \ldots, X_n(t_n)$. Как для случая известного условного распределения «шумов» наблюдений, так и для случая их известной условной дисперсии $\sigma(t)$ получены асимптотически минимаксные нижние границы среднеквадратического риска оценок функционала $F(R)$, дифференцируемого по Фреше на некотором компакте $\mathscr{P}$ из $L_2[0, 1]$, производная которого $F'(R, t)$ удовлетворяет условию Гельдера в $L_2[0, 1]$ на том же компакте $\mathscr{P}$.

Допустим, что функция регрессии принадлежит некоторому компакту $K \subset W_2^\beta(L)$, $\beta > 1/2$ и ограничена в норме $\|\cdot\|_\infty$ на этом компакте, а производная Фреше $F'(R, t)$ функционала $F(R)$ удовлетворяет на компакте K условию Гельдера в $L_2[0, 1]$ с показателем $1/\beta < \alpha \leqq 1$ и условию Гельдера по t на отрезке $[0, 1]$ с некоторым показателем $\gamma > 0$. Показано, что в этом случае можно указать такой план оценивания и такую оценку $\hat{F}_n$ функционала $F(R)$, построенную по наблюдениям $X_1(t_1), \ldots, X_n(t_n)$, что в случае известной условной дисперсии на указанном компакте K нижняя минимаксная граница среднеквадратического риска оценки $\hat{F}_n$ асимптотически достигаются, т. е. выполнено неравенство

$$\lim_{n\to\infty} \sup_{R\in K} \left[nE\left\{ \hat{F}_n - F(R) \right\}^2 - \left(\int_0^1 \sigma(t) |F'(R, t)|\, dt \right)^2 \right] = 0.$$

Р. З. Хасьминский
Ю. И. Пастухова
Институт проблем передачи информации АН СССР
СССР, 101447, Москва ГСП4, ул. Ермоловой 19

Problems of Control and Information Theory, Vol. 18 (2), pp. 79—89 (1989)

COMPARISON OF ASYMPTOTIC VARIANCES FOR SEVERAL ESTIMATORS OF LOCATION

I. VAJDA

(*Prague*)

(Received April 2, 1988)

This is a comparative study presenting asymptotic variances of several consistent asymptotically normal estimators of location for sources of independent data defined by mixtures of standard normal and some other, normal or non-normal, distributions. An estimator introduced in Vajda [6, 7] is compared with a group of estimators considered by Huber [3] and with a group of estimators proposed by Kovanic [5]. Variances of all estimators are compared with the reversed value of Fisher information. The result are presented in a graphical form.

1. Estimators of location

Let us consider a family $\mathcal{Q}=(Q_\theta|\theta\in\mathbf{R})$ of absolutely continuous probability measures on $\mathbf{R}$ with densities $q(x-\theta)$ and a parameter $0\leqq\alpha\leqq 1$. By the symbol T^α we denote the so-called α-estimator for location with projection density $q(x)$. This means that T^α denotes a sequence $(T_n^\alpha|n\in\mathbf{N})$ of measurable mappings $T_n^\alpha:\mathbf{R}^n\to\mathbf{R}$ such that $T_n^\alpha(x_1,\ldots,x_n)$ belongs to the set

$$\operatorname{argmin}\frac{1}{\alpha}\left(1-\sum_{i=1}^{n}q(x_i-\theta)^\alpha\right)\qquad\text{if}\quad 0<\alpha\leqq 1$$

and to the set

$$\operatorname{argmin}-\sum_{i=1}^{n}\ln q(x_i-\theta)\qquad\text{if}\quad \alpha=0.$$

T^α is a particular case of an α-estimator of more general parameter introduced in our paper [6]. Hereafter we are interested in the asymptotic properties of the just defined T^α under the assumption that the data in sample vectors $\mathbf{x}=(x_1,\ldots,x_n)$ are distributed by a member of a family $\mathcal{P}=(P_\theta|\theta\in\mathbf{R})$ of absolutely continuous probability measures on $\mathbf{R}$ with densities $p(x-\theta)$. Let us note that the data generating density $p(x)$ may differ from the projection density $q(x)$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

In [7] we recommended for practical use the estimator $T^{1/5}$ with the projection density of standard normal probability distribution $N(0, 1)$, i.e. with

$$q(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

in all cases where the data generating density is even and unimodal. This estimator is denoted in the sequel by the symbol $T^{1/5,N}$, i.e.

$$T_n^{1/5,N}(\mathbf{x}) \in \operatorname{argmax} \sum_{i=1}^{n} e^{-(x_i-\theta)^2/10}. \tag{1}$$

Throughout this paper we denote by H the probability distribution with the hyperbolic-secant type density

$$h(x) = \frac{1}{2\operatorname{ch}^2 x} = \frac{2}{(e^x + e^{-x})^2}.$$

The mean of this distribution is 0 and the variance is 3/4.

Let us denote by $T^{\alpha,H}$ the estimator T^{α} with the projection density $q(x) = h(x)$, i.e.

$$T^{\alpha,H}(\mathbf{x}) \in \operatorname{argmax} \sum_{i=1}^{n} (e^{x_i-\theta} + e^{-x_i+\theta})^{-2\alpha} \qquad \text{if} \quad 0 < \alpha \leqq 1, \tag{2}$$

$$T_n^{0,H}(\mathbf{x}) \in \operatorname{argmax} \sum_{i=1}^{n} \ln (e^{x_i-\theta} + e^{-x_i+\theta}) \qquad \text{if} \quad \alpha = 0. \tag{3}$$

As shown by Fabian [1] (for further details see Vajda [9]), the estimators $T^{\alpha,H}$ for $\alpha \in \{0, 1/2, 1\}$ are equivalent with the estimators introduced on the lines 5, 7, 8 of Table 2 on p. 311 in Kovanic [5].

We shall compare the asymptotic properties of the estimator $T^{1/5,N}$ with the asymptotic properties of estimators $T^{0,H}$, $T^{1/2,H}$, $T^{1,H}$. Numerical comparisons are also extended to some of the estimators of location considered in Table 1 of Huber [3].

2. Existence and asymptotic properties

For details about the definition of the estimators considered in Huber's paper [3], as well as about their existence and asymptotic properties, we refer to Huber's book [4].

Our estimators $T^{1/5,N}$ and $T^{\alpha,H}$, $0 \leqq \alpha \leqq 1$, are close to special versions of the M-estimators of Huber, but the definition given above formally differs from that in [4] (or in [2], where the book [4] refers to as to the general theory of M-estimators). Thus, as to the existence and asymptotic properties of the estimators $T^{1/5,N}$ and $T^{\alpha,H}$, $\alpha \in \{0, 1/2, 1\}$, we have the possibility to refer ourselves either to the book [10], where

one can find a general theory of α-estimators covering the particular cases considered here, or to the research report [8] and the paper [9], where the general theory of [10] was specified to the α-estimators of location $T^{\alpha,N}$ and $T^{\alpha,H}$ for $0 \leqq \alpha \leqq 1$. We have chosen the second option.

Proposition 1a. The estimator $T^{1/5,N}$ exists in the sense specified above. If p is even, bounded, unimodal and almost everywhere differentiable with a derivative $p'(x)$ and if $p'(x) \neq 0$ almost everywhere, then $T^{1/5,N}$ is consistent in the sense that, for every $\theta \in \mathbf{R}$, $T_n^{1/5,N}(\mathbf{x})$ tends in the P_θ-probability to θ as $n \to \infty$. If p satisfies the just stated conditions and F is the distribution function of p then the influence curve $IC(x, T^{1/5,N}, F)$ exists and satisfies the relation

$$IC(x, T^{1/5,N}, F) = \frac{xe^{-x^2/10}}{C(F)}$$

where

$$C(F) = -\int_{\mathbf{R}} xe^{-x^2/10} p'(x)dx.$$

Proof. The first assertion follows from Theorem A on p. 179 of [8]. The second assertion follows from the Corollary on p. 181 ibid, and the third assertion from Theorem D on p. 182 ibid, and from the identity

$$\int_{\mathbf{R}} \left(1 - \frac{x^2}{5}\right) \exp\left\{-\frac{x^2}{10}\right\} p(x)dx = C(F)$$

which follows from the per partes integration rule. □

Proposition 1b. Let p and F satisfy all conditions considered in Proposition 1a. Then $T^{1/5,N}$ is asymptotically normal in the sense that, for every $\theta \in \mathbf{R}$, $\sqrt{n}(T_n^{1/5,N}(\mathbf{x}) - \theta)$ tends in the law induced by the probability distribution P_θ to the random variable $N(0, \sigma^2)$ with

$$\sigma^2 = \int_{\mathbf{R}} IC(x, T^{1/5,N}, T)^2 p(x)dx.$$

It holds

$$\sigma^2 = \frac{\int_{\mathbf{R}} x^2 e^{-x^2/5} p(x)dx}{C(F)^2} > \frac{1}{I(p)},$$

where $I(p)$ denotes the Fisher information of the data generating density, i.e.

$$I(p) = \int_{\mathbf{R}} \left[\frac{p'(x)}{p(x)}\right]^2 p(x)dx > 0.$$

Proof. The first assertion follows from Theorem E on p. 184 in [8]. The formula for σ^2 in the second assertion follows from the formula for the influence curve in Proposition 1a. The positivity of $I(p)$ follows from the assumptions about p. The inequality

$$\sigma^2 \geqq \frac{1}{I(p)}$$

follows from the formula for σ^2 and from the Cauchy–Schwartz inequality. The equality takes place iff there exists $c \in \mathbf{R}$ such that

$$xe^{-x^2/10} = c\frac{d}{dx}\ln p(x) \qquad \text{a.e.}$$

This, however, means that there is a constant $\tilde{c} \in \mathbf{R}$ for which

$$e^{-x^2/10} = c \ln p(x) + \tilde{c}, \qquad x \in \mathbf{R}.$$

This is obviously never satisfied by p under consideration. □

Proposition 2a. For every $0 \leqq \alpha \leqq 1$ the estimator $T^{\alpha,H}$ exists in the sense specified in Section 1. If p and F satisfy the conditions considered in Proposition 1a then $T^{\alpha,H}$ is consistent in a similar sense as $T^{1/5,N}$ in Proposition 1a and its influence curve $IC(x, T^{\alpha,H}, F)$ exists and satisfies the relation

$$IC(x, T^{\alpha,H}, F) = \frac{e^x - e^{-x}}{C(F)(e^x + e^{-x})^{2\alpha+1}}$$

where

$$C(F) = -\int_{\mathbf{R}} \frac{e^x - e^{-x}}{(e^x + e^{-x})^{2\alpha+1}} p'(x)dx.$$

Proof. The first assertion follows from Theorem 1 in [9]. The second assertion follows from Theorems 2 and 3 ibid. □

Proposition 2b. Let p and F satisfy all conditions considered in Proposition 2a. Then, for every $0 \leqq \alpha \leqq 1$, $T^{\alpha,H}$ is asymptotically normal in a similar sense as $T^{1/5,N}$ in Proposition 1b and

$$\sigma^2 = \int_{\mathbf{R}} IC(x, T^{\alpha,H}F)^2 p(x)dx =$$

$$= \frac{\displaystyle\int_{\mathbf{R}} \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^{4\alpha+2}} p(x)dx}{C(F)^2} \geqq \frac{1}{I(p)},$$

where $I(p)$ is defined as in Proposition 1b and where the sign of equality takes place iff there is $c \in \mathbf{R}$ such that

$$\frac{e^x - e^{-x}}{(e^x + e^{-x})^{2\alpha+1}} = c\frac{p'(x)}{p(x)} \qquad \text{a.e.}$$

Proof. This assertion follows from Theorems 4 and 5 in [9]. □

Example. It follows from Examples 1 and 2 in [9] that if $p(x) = h(x)$, then the constant $C(F)$ considered in Propositions 2a and 2b is given as follows

$$C(F) = \frac{2\Gamma(1+\alpha)^2}{(3+2\alpha)\Gamma(2+2\alpha)}$$

where Γ denotes the gamma function. In the same way it follows that the asymptotic variance σ^2 considered in Proposition 2b is in this case given by

$$\sigma^2 = \frac{[(3+2\alpha)(1+2\alpha)]^2}{4(3+4\alpha)(1+4\alpha)\Gamma(1+4\alpha)}\left(\frac{\Gamma(1+2\alpha)}{\Gamma(1+\alpha)}\right)^4.$$

Let us now consider $\alpha = 0$. From the last formula we obtain

$$\sigma^2 = \frac{3^2}{4.3\Gamma(1)}\left(\frac{\Gamma(1)}{\Gamma(1)}\right)^4 = \frac{3}{4}$$

which is the variance of the distribution H with density h. This is not yet an argument that $T^{0,H}$ is asymptotically efficient when $p(x) = h(x)$. But it suffices to refer to the last assertion of Proposition 2b, since in this case it holds for every $x \in \mathbf{R}$

$$p'(x) = \frac{d}{dx}h(x) = \frac{d}{dx}\frac{1}{2\operatorname{ch}^2 x} = \frac{\operatorname{sh} x}{\operatorname{ch}^3 x},$$

so that

$$\frac{p'(x)}{p(x)} = 2\frac{\operatorname{sh} x}{\operatorname{ch} x} = 2\operatorname{th} x = 2\frac{e^x - e^{-x}}{e^x + e^x}.$$

Hence, by the cited assertion, the asymptotic variance of $T^{0,H}$ attains under $p(x) = h(x)$ the lower bound

$$\frac{1}{I(p)} = \frac{3}{4}.$$ □

The influence curves $IC(x, T, N(0, 1))$ for $T \in \{T^{1/5,N}, T^{0,H}, T^{1/2,H}, T^{1,H}\}$ are shown in Fig. 1.

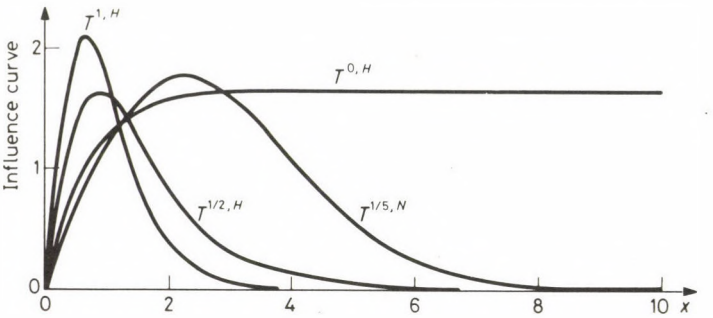


Fig. 1

3. Asymptotic variances

The asymptotic variances σ^2 specified for estimators $T^{1/5,N}$, $T^{1/2,H}$, $T^{1,H}$ in Propositions 1b and 2b are measures of quality of these estimators in data sources described by a family $\mathscr{P}$ with an absolutely continuous parent probability measure $P=P_0$, the density of which is $p(x)$. In this section we shall consider parent probability measures P of the form

$$P=(1-\varepsilon)P_1+\varepsilon P_2, \qquad 0\leqq\varepsilon\leqq 1, \tag{4}$$

where $P_1\equiv N(0,1)$ and where P_2 describes one of the following probability distributions:

— H with the density $p_2(x)=h(x)$ defined above,
— $N(0, 9)$ with the density

$$p_2(x)=\frac{1}{3\sqrt{2\pi}}e^{-x^2/18},$$

— $N(0, 100)$with the density

$$p_2(x)=\frac{1}{10\sqrt{2\pi}}e^{-x^2/200},$$

— Cauchy with the density

$$p_2(x)=\frac{1}{\pi(1+x^2)}.$$

Each of the listed densities $p_2(x)$ satisfies the assumptions of propositions of the last section so that so does the density

$$p(x)=(1-\varepsilon)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}+\varepsilon p_2(x)$$

of the mixture (4). It follows from here and from Propositions 1b and 2b that, if F_1 and F_2 are the distribution functions of the standard normal distribution and of the distribution with a density p_2 considered above and σ_1^2 and σ_2^2 the corresponding asymptotic variances of an estimator $T\in\{T^{1/5,N},\ T^{0,H},\ T^{1/2,H},\ T^{1,H}\}$, then the asymptotic variance σ^2 of T for the mixed density p is given by

$$\sigma^2=\frac{(1-\varepsilon)C(F_1)^2\sigma_1^2+\varepsilon C(F_2)^2\sigma_2^2}{((1-\varepsilon)C(F_1)+\varepsilon C(F_2))^2}. \tag{5}$$

Using this formula and the values $C(F_i)$, σ_i^2 for $i=1, 2$, calculated with the help of Mr. R. Hartl and stated in Table 1, we evaluated in Figs 2–5 the asymptotic variances σ^2 of estimators $T^{1/5,N}$ (interrupted line), $T^{0,H}$, $T^{1/2,H}$, $T^{1,H}$ (solid lines), and of the following estimators considered by Huber [3]

Used notation	Estimator
○	sample mean
△	trimmed mean for $\alpha=0.1$
□	Huber for $k=1$
+	Hodges–Lehman
✳	Takeuchi
×	Hampel 25A

as functions of the mixture variable $0\leqq\varepsilon\leqq 1$. For comparison the minimum attainable asymptotic variance represented by the inverse Fisher information is shown by the dotted line in each figure.

Table 1

Estimator / F	$T^{1/5,N}$		$T^{0,H}$		$T^{1/2,H}$		$T^{1,H}$	
	$C(F)$	σ^2	$C(F)$	σ^2	$C(F)$	σ^2	$C(F)$	σ^2
$N(0, 1)$	0.760726	1.043162	0.605705	1.074726	0.151279	1.543585	0.045449	2.270803
$N(0, 9)$	0.213434	20.02533	0.254833	11.47470	0.016878	70.73459	0.003255	199.9985
$N(0, 100)$	0.010391	3527.623	0.079463	145.7807	0.000605	18026.81	0.000099	68219.00
H	0.806163	0.757298	0.666666	0.750000	0.196350	0.864607	0.066666	1.071429
Cauchy	0.511469	2.361529	0.449953	2.642520	0.109080	2.139826	0.036129	2.538425

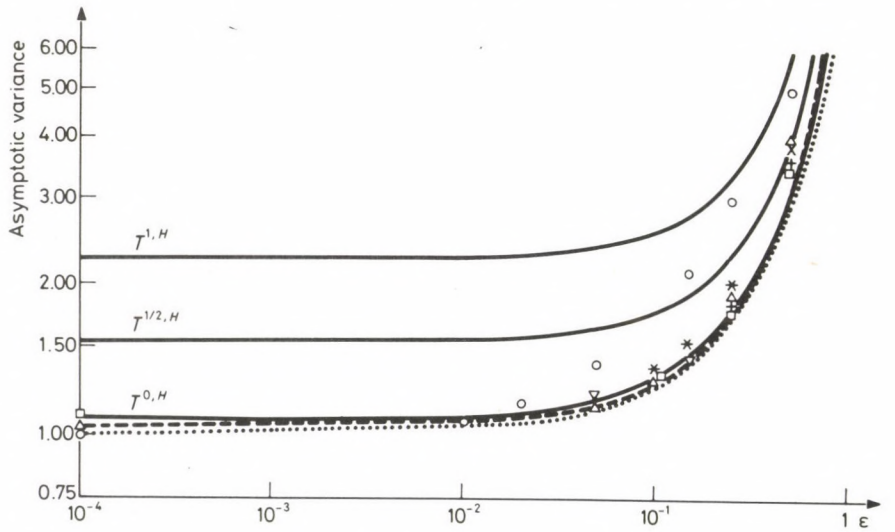


Fig. 2. $P_2 \equiv N(0, 9)$

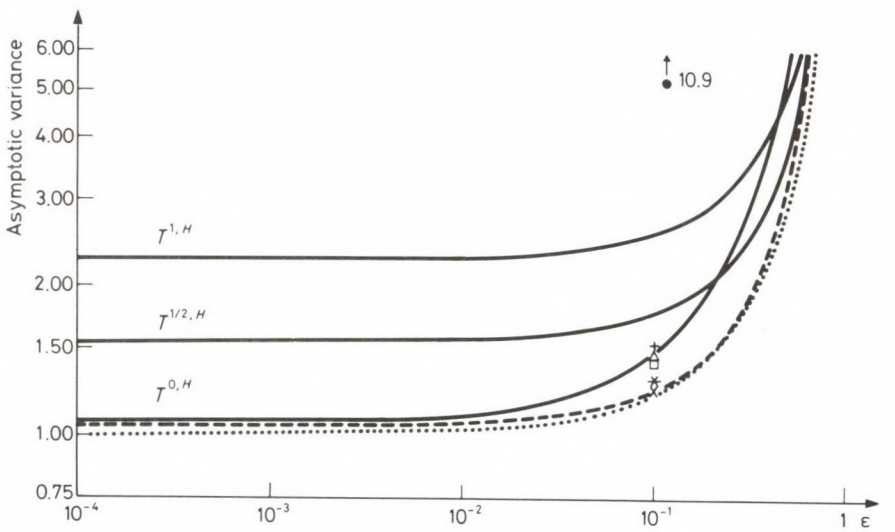


Fig. 3. $P_2 \equiv N(0, 100)$

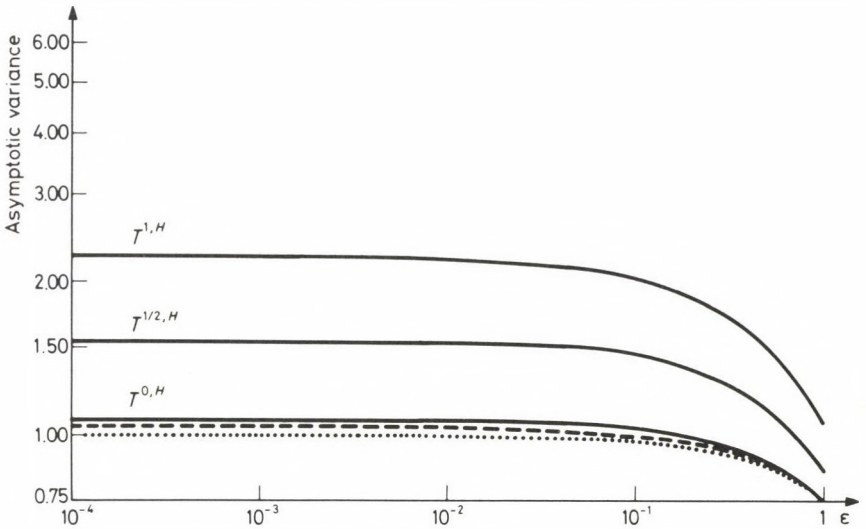


Fig. 4. $P_2 \equiv H$

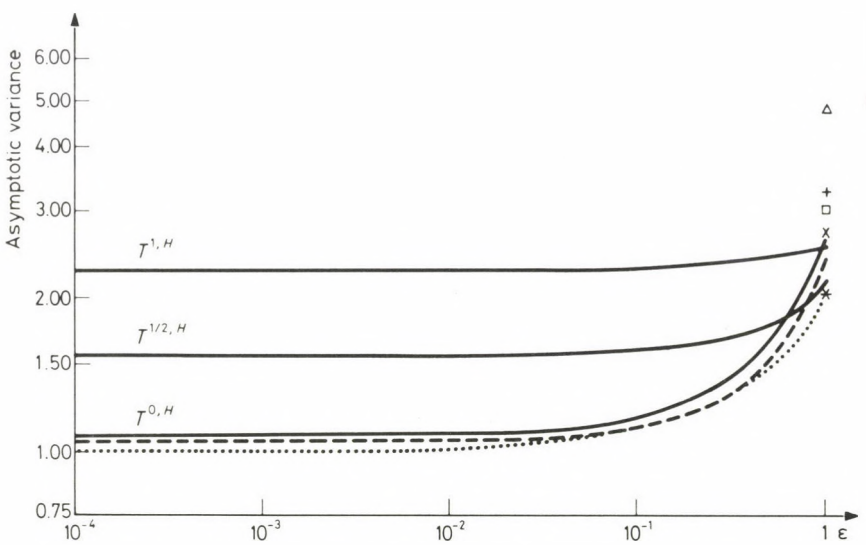


Fig. 5. $P_2 \equiv$ Cauchy

Let us mention here a special historical link between the Hampel estimator 25A and the α-estimators as such. In 1982, when the author jointly with his colleague Dr. J. A. Víšek numerically evaluated the influence curves $IC(x, T^{\alpha,N}, N(0, 1))$ for $\alpha \in \{1/10, 1/5, 1/2, 1\}$, it has become clear immediately that these curves for $\alpha = 1/10$ or $\alpha = 1/5$ quite well fit the piecewise linear influence curve of the estimator 25A known from the Princeton robustness study. This greatly helped to understand the properties of the new estimators (for details about their derivation see [9]) and stimulated further research on them.

4. Conclusions

Let us now turn to some conclusions which can be drawn from Figs 2–5. At first, maximum inefficiency observed in all mixture models considered in these figures is minimized by $T^{1/5,N}$ — the observed value for this estimator is about 10%. Let us notice that the asymptotic variance of this estimator in mixture families under consideration is more than comparable with the best estimators of location considered in Huber [3]. Even for the hyperbolic — cosecant alternative H, for which $T^{0,H}$ is the maximum likelihood estimator, the asymptotic variance of $T^{1/5,N}$ is below that of $T^{0,H}$ until the data are distributed practically purely by H, and then the inefficiency of $T^{1/5,N}$ is practically zero. The only noticeable cases where $T^{1/5,N}$ is outperformed is the contamination ε about 0.1–0.2 where "Hodges–Lehman", trimmed mean and "Huber" are very slightly better and the contaminations above 0.4, where $T^{0,H}$ is very slightly better — all for $P_2 \equiv N(0, 9)$. Moreover, $T^{1/2,H}$ is slightly better for $P_2 \equiv$ Cauchy and for the contaminations $\varepsilon > 0.8$.

It seems to us that the first conclusion which can be made from these figures is that $T^{1/5,N}$ is a good estimator of location in cases where the data distribution density is symmetric and unimodal, in particular if the density is close to the standard normal one.

The second conclusion which can be made is that the estimators $T^{1/2,H}$ and $T^{1,H}$ can hardly compete, and least in data sources under consideration, with the remaining estimators. The only observed exceptions are very pure Cauchy sources for which $T^{1/2,H}$ performs well.

References

1. *Fabian, Z.*, Point estimation in case of small data sets. Trans. 10th Prague Conf. on Information Theory, ..., Academia, Praha 1987.
2. *Huber, P. J.*, The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. on Math. Statist. and Probab., Vol. **1**, Calif. Univ. Press, Berkeley 1967.

3. *Huber, P. J.*, Robust statistics: a review. Ann. Math. Statist. **43** (1972), 1041–1067.
4. *Huber, P. J.*, Robust Statistics. J. Wiley, New York, 1981.
5. *Kovanic, P.*, Gnostical theory of small samples of real data. Probl. of Contr. and Inform. Theory **13** (1984), 309–319.
6. *Vajda, I.*, A new general approach to minimum distance estimation. Trans. 9th Prague Conf. on Information Theory, . . ., 103–112. Academia, Praha 1983.
7. *Vajda, I.*, Minimum divergence principle in statistical estimation. Statistics & Decisions, Supplement Issue No. *1*, 239–261. Oldenburg Verlag, München 1984.
8. *Vajda, I.*, Information in a Statistical Experiment and its Use in a Statistical Decision (in Czech). Research report No. 1500. Institute of Information Theory and Automation, Prague 1987.
9. *Vajda, I.*, Minimum-distance and gnostical estimators. Probl. of Control and Inform. Theory **17** (1988), No. *5*.
10. *Vajda, I.*, Theory of Statistical Inference and Information. D. Reidel, Dordrecht.

Сравнение асимптотических дисперсий некотоых оценок сдвига

И. ВАЙДА
(Праха)

Сравниваются дисперсии оценки автора [6, 7], Кованица [5] и Губера [3]. Приводится также обратное значение информации Фишера.

I. Vajda
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věži 4
182 08 Prague 8
Czechoslovakia

Problems of Control and Information Theory, Vol. 18 (2), pp. 91—106 (1989)

ASYMPTOTIC EXPANSIONS FOR THE MUTUAL INFORMATION AND FOR THE CAPACITY OF CONTINUOUS MEMORYLESS CHANNELS WITH WEAK INPUT SIGNAL

V. V. PRELOV

(Received May 10, 1988)

(*Moscow*)

Asymptotic expansions for the mutual information in a continuous memoryless channel with independent Gaussian noise are obtained under the assumption that the signal X_ε at the input of the channel depending on a small parameter ε satisfies either the condition of the absolute moment of the order $n+a$

$$\mathbf{E}|X_\varepsilon/\varepsilon|^{n+a} \leqq K < \infty$$

for some integer $n \geqq 2$ and some α, $0 < \alpha < 1$, or a stronger condition to the peak power $|X_\varepsilon| \leqq K\varepsilon$. Using this result asymptotic expansions for the capacity of the considered channel under various restrictions to the input signal are obtained.

1. Introduction

The asymptotic behaviour of the mutual information and of the capacity for continuous memoryless channels with small input signal was studied in a large number of papers (see, e.g. [1] where references to this topic are contained, or [2] where one can find a short review of some of the results). However, up to now only the main term of the asymptotic expansion of the mutual information and of the capacity under some assumptions about the noise distribution and certain restrictions on the input signal were obtained.

In the present paper for the first time the conditions at the additive noise distribution and at the input noise are obtained which enable to get not only the main terms, but also subsequent terms of similar asymptotic expansions. Earlier a similar problem was solved by the author [3] for channels with almost Gaussian noise.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. Statement of the main results

The continuous memoryless communication channel with additive noise which we will consider in the present paper is given by the following formula:

$$Y = X + Z \tag{2.1}$$

where X is the signal at the input of the channel, Y is the signal at the output and Z is the additive noise in the channel. We will assume that $X=(X_1, \ldots, X_N)$, $Y=(Y_1, \ldots, Y_N)$, $Z=(Z_1, \ldots, Z_N)$ are random vectors taking values in the N-dimensional Euclidean space $\mathbf{R}^N$ and X and Z are independent. Let us assume further that the input signal $X = X_\varepsilon = (X_{\varepsilon 1}, \ldots, X_{\varepsilon N})$ depends on a small parameter ε in such a way that one of the following condition is satisfied: either

$$\mathbf{E}|X_\varepsilon/\varepsilon|^{n+\alpha} \leqq K < \infty \tag{2.2}$$

for some integer $n \geqq 2$ and some α, $0 < \alpha < 1$, or

$$|X_\varepsilon| \leqq K\varepsilon, \quad K < \infty, \quad \text{with probability 1} \tag{2.3}$$

(here $|\cdot|$ is the usual norm in $\mathbf{R}^N$).

Let us state now the main assumptions on the distribution of the additive noise Z under which the results proved below are true.

(A) There exists a probability distribution $p_Z(x) = p(x)$, $x=(x_1, \ldots, x_N) \in \mathbf{R}^N$ which is bounded on $\mathbf{R}^N$, has bounded continuous partial derivatives up to the order $n+1$ and satisfies the condition

$$\lim_{|x| \to \infty} p(x) = \lim_{|x| \to \infty} \frac{\partial^k p(x)}{\partial x_{i_1} \ldots \partial x_{i_k}} = 0 \tag{2.4}$$

for $k = 1, \ldots, n-1$ and for all $i_1, \ldots, i_k$.

(B) There exist sufficiently large $M > 0$ and $\delta_0 > 0$ such that

$$\left| \frac{\partial^k p(x)}{\partial x_{i_1} \ldots \partial x_{i_k}} \right| \leqq \delta^M p(x) \tag{2.5}$$

for all $x \in G_\delta$, $\delta \geqq \delta_0$, $k = 1, 2, \ldots, n$ and all $i_1, \ldots, i_k$ where

$$G_\delta = \{x \in \mathbf{R}^N \mid p(x) > e^{-\delta^2}\}. \tag{2.6}$$

(C) There exist $\lambda > 0$, $M > 0$ and $\delta_0 > 0$ such that

(C1) $$\int_{\bar{G}_\delta} \left| \frac{\partial^j p(x)}{\partial x_{i_1} \ldots \partial x_{i_j}} \log p(x) \right| dx \leqq \delta^M e^{-\lambda\delta^2}$$

for all $\delta \geqq \delta_0$, $j=0, 1, \ldots, n$ (here, as usual, we agree that the derivative of order 0 of the function $p(x)$ is $p(x)$ itself)

(C2) $$\int\limits_{\bar{G}_\delta} \left| \frac{\partial^j p(x)}{\partial x_{i_1} \ldots \partial x_{i_j}} \Big/ p(x) \right|^k p(x)dx \leqq \delta^M e^{-\lambda\delta^2}$$

for all $\delta \geqq \delta_0$, $j=0, 1, \ldots, n$, $k=2, \ldots, [\nu/2]$,

(C3) $$\sup_{y:|y|<\theta} \int\limits_{\bar{G}_\delta} |p(x-y) \log p(x)|\, dx \leqq \delta^M e^{-\lambda\delta^2(1-\beta)}$$

for all $\delta \geqq \delta_0$, all sufficiently small $\beta > 0$ and some $\theta = \theta(\beta) > 0$.

Here we used the following notations:

$$\bar{G}_\delta = \mathbf{R}^N \backslash G_\delta = \{x \in \mathbf{R}^N \mid p(x) \leqq e^{-\delta^2}\}, \tag{2.7}$$

$$\nu = \frac{\lambda}{\lambda+\mu} n \tag{2.8}$$

and μ is the smallest nonnegative number satisfying the conditions

$$\lambda + \mu \geqq 1 \tag{2.9}$$

and

$$\text{mes } G_\delta \leqq \delta^M e^{\mu\delta^2} \qquad \text{for all} \quad \delta \geqq \delta_0 \tag{2.10}$$

(mes G is the Lebesgue measure of a set G, $[z]$ is the integral part of the number z);

(D) $$\limsup_{|y| \to \infty} (H(p, p^y)/|y|^{n+\alpha}) < \infty$$

where

$$H(p, p^y) = \int\limits_{\mathbf{R}^N} p(x) \log \frac{p(x)}{p(x-y)}\, dx. \tag{2.11}$$

The following theorem is proved in Section 3:

Theorem 1. Let the distribution density $p(x)$ of the noise Z in channel (2.1) satisfy conditions (A), (B), (C). Then, under restriction (2.3) on the input signal, the following asymptotic expansion of the mutual information between input and output signals holds:

$$I(X_\varepsilon, Y) = -\int\limits_{\mathbf{R}^N} a_{[\nu]}(x) \log p(x)dx -$$

$$-\log e \sum_{k=2}^{[v/2]} \frac{(-1)^k}{(k-1)k} \int_{\mathbf{R}^N} \frac{a^k_{[v]-2}(x)}{p^{k-1}(x)}\, dx + o(\varepsilon^v), \qquad \varepsilon \to 0 \tag{2.12}$$

where

$$a_m(x) = \sum_{k=2}^{m} \frac{(-1)^k}{k!} \mathbf{E}\left[\left(\frac{\partial}{\partial x_1} X'_{\varepsilon 1} + \ldots + \frac{\partial}{\partial x_N} X'_{\varepsilon N}\right)^k p(x)\right], \tag{2.13}$$

$$X'_{\varepsilon k} = X_{\varepsilon k} - \mathbf{E} X_{\varepsilon k}, \qquad k = 1, \ldots, N. \tag{2.14}$$

If, moreover, condition (D) is satisfied, then formula (2.12) is true also under restriction (2.2) to the input signal. The expression $o(\varepsilon^v)$ in the right-hand side of (2.12) (as well as everywhere later) stands for some function in ε such that $o(\varepsilon^v)/\varepsilon^v \to 0$ as $\varepsilon \to 0$ uniformly in all distributions X_ε satisfying (2.3) or (2.2), respectively.

Corollary. If conditions (A), (B), (C) are satisfied for all n (with $\lambda = \lambda(n) > 0$, $M = M(n) > 0$, $\delta_0 = \delta_0(M) > 0$) depending, in general, on n then, under restriction (2.3), the representation of $I(X_\varepsilon, Y)$ by an asymptotic series based on (2.12) holds. Similar representation holds if, for any positive integer n, the input signal satisfies the following condition

$$\mathbf{E}|X_\varepsilon/\varepsilon|^n \leqq K_n < \infty$$

where K_n might depend on n.

Let us discuss the conditions of Theorem 1 in more details. Condition (A) is the smoothness and regularity condition for the density $p(x)$ as $x \to 0$, and the existence of partial derivatives of the function $p(x)$ is necessary for the statement of Theorem 1. Conditions (B) and (C) can be easily checked for a wide class of densities (in particular, for densities decreasing as a power or as an exponent in x as $|x| \to \infty$ and satisfying weak additional regularity conditions). Let us note, in particular, that for exponentially decreasing densities of the form

$$p(x) \approx c \exp\{-|x|^\beta\}, \qquad x \in \mathbf{R}^1, \quad \beta > 0, \quad x \to 0$$

the parameters λ and μ in the statement of Theorem 1 are equal to 1 and 0, respectively, and for densities with the power decrease

$$p(x) \approx c|x|^{-\beta}, \qquad x \in \mathbf{R}^1, \quad \beta > 1, \quad x \to \infty$$

one can easily see that $\lambda = (\beta - 1)/\beta$, $\mu = 1/\beta$ (let us remark that in each case $\lambda + \mu = 1$).

One can also check that under some additional regularity conditions condition (D) is satisfied only in the case when $p(x)$, $x \in \mathbf{R}^1$, decreases for $x \to \infty$ not faster than $\exp\{-|x|^\beta\}$ with $\beta \leqq n + a$. The question whether or not Theorem 1 is true under restriction (2.2) but without condition (D) remains open.

Let us introduce now the following notations. Let $C^{(h,K)}(\varepsilon)$ be the capacity of channel (2.1) under the condition that the following two restrictions on the input signal distribution hold:

$$\mathbf{E}|X_\varepsilon|^h \leqq \varepsilon^h, \quad h>0 \tag{2.15}$$

$$\mathbf{E}|X_\varepsilon| \leqq K\varepsilon, \quad 1 \leqq K < \infty. \tag{2.16}$$

Denote also by

$$C_1^{(h)}(\varepsilon) = C^{(h,\infty)}(\varepsilon) \tag{2.17}$$

the capacity of the considered channel under the only restriction (2.15) and by

$$C_2(\varepsilon) = C^{(h,1)}(\varepsilon) \tag{2.18}$$

the capacity of this channel under the only restriction (2.16) with $K=1$.

The following result for channel (2.1) in the one-dimensional case is proved in section 4:

Theorem 2. Let the probability density $p(x)$, $x \in \mathbf{R}^1$, satisfy conditions (A), (B), (C) for some n and v. Let also $p(x)$ be an odd function. Then

1°. A random variable X_ε^0 such that

$$\mathbf{P}\{X_\varepsilon^0 = -\varepsilon\} = \mathbf{P}\{X_\varepsilon^0 = \varepsilon\} = 1/2$$

is the asymptotically optimal signal at the input of channel (2.1) up to terms of order $O(\varepsilon^{2m})$, $\varepsilon \to 0$ provided either:

a) the input signal restriction is given in the form (2.15) with $h = n+\alpha$, $\alpha > 0$, where n is such that $[v] = 2m$, $m \geqq 1$, and condition (D) is satisfied; or

b) the input signal restriction is given in form (2.16) with $K=1$ and $[v]=2m$, $m \geqq 1$.

In other words, under the above conditions the asymptotic expansions for $C_1^h(\varepsilon)$ and $C_2(\varepsilon)$ can be found by formula (2.12) with $N=1$ and $X_\varepsilon = X_\varepsilon^0$.

2°. For $v \geqq 6$ the following asymptotic expansion holds:

$$C^{(2,K)}(\varepsilon) = \log e \left[\frac{1}{2} \int_{\mathbf{R}^1} \frac{(p'(x))^2}{p(x)}\, dx\, \varepsilon^2 + \right.$$

$$\left. + \left(\frac{L^2}{24} \int_{\mathbf{R}^1} \frac{p'(x)p'''(x)}{p(x)}\, dx - \frac{1}{8} \int_{\mathbf{R}^1} \frac{(p''(x))^2}{p(x)}\, dx \right) \varepsilon^4 \right] + O(\varepsilon^6)$$

where

$$L = \begin{cases} K & \text{if } I = \int\limits_{\mathbf{R}^1} \dfrac{p'(x)p'''(x)}{p(x)}\,dx \geqq 0 \\ 1 & \text{if } I < 0. \end{cases}$$

3. Asymptotic expansion of information

In this section we prove the above stated Theorem 1. Let us note first of all that we can assume $\mathbf{E}X_\varepsilon = 0$; everywhere in this section this property of the noise will be assumed satisfied. Theorem 1 would follow from Lemmas 1–4 stated and proved below. By c, c_1, c_2 we will denote various constants, possibly different in different inequalities.

Lemma 1. Let the probability density $p(x)$ of the distribution of the noise Z satisfy condition (A) (we do not require the fulfilment of (2.4)). The density $p_Y(x)$ of the signal Y at the channel output admits the following representation:

$$p_Y(x) = p(x) + a_n(x, \varepsilon) + r_n(x, \varepsilon) \tag{3.1}$$

where

$$a_n(x, \varepsilon) = \sum_{k=2}^{n} \frac{(-1)^k}{k!} \mathbf{E}\left[\left(\frac{\partial}{\partial x_1} X_{\varepsilon 1} + \ldots + \frac{\partial}{\partial x_N} X_{\varepsilon N}\right)^k p(x)\right] \tag{3.2}$$

and

$$|r(x, \varepsilon)| \leqq c\varepsilon^{n+\gamma}, \quad \gamma = \alpha/(1+\alpha) \qquad \text{for all} \quad x \in \mathbf{R}^N \tag{3.3}$$

if condition (2.2) is satisfied, and

$$|r(x, \varepsilon)| \leqq c\varepsilon^{n+1} \qquad \text{for all} \quad x \in \mathbf{R}^N \tag{3.4}$$

if condition (2.3) is satisfied.

Proof. Using the formula

$$p_Y(x) = \int\limits_{\mathbf{R}^N} p(x-y)\,dF_{X_\varepsilon}(y)$$

where $F_{X_\varepsilon}(x)$ is the distribution function of X_ε and the Taylor decomposition for $p(x-y)$ at the neighbourhood of the point x we get the required equalities (3.1), (3.2)

with

$$r_n(x,\varepsilon)=\frac{(-1)^n}{n!}\int\limits_{\mathbf{R}^N}\left(\frac{\partial}{\partial x_1}y_1+\ldots+\frac{\partial}{\partial x_N}y_N\right)^n\Big[\,p(x-\theta y)-p(x)\,\Big]dF_{X_\varepsilon}(y) \quad (3.5)$$

where $0<\theta=\theta(y)<1$ and $y=(y_1,\ldots,y_N)$. We have to show that $r_n(x,\varepsilon)$ satisfies (3.3) and (3.4).

Let us assume first that condition (2.2) is satisfied. For convenience we introduce the notations

$$\Delta^n_{i_1,\ldots,i_n}(x,\theta y)=\frac{\partial^n p(x-\theta y)}{\partial x_{i_1},\ldots,\partial x_{i_n}}-\frac{\partial^n p(x)}{\partial x_{i_1},\ldots,\partial x_{i_n}}$$

and estimate the integrals

$$I_m=\int\limits_{\mathbf{R}^N}\Big|\Delta^n_{i_1\ldots i_n}(x,\theta y)\Big|\,|y_m|^n\,dF_{X_\varepsilon}(y),\qquad m=1,\ldots,N.$$

It is clear that for any fixed $\beta>0$ we have

$$I_m\leqq I_m^{(0)}+\sum_{k=1}^{N}I_m^{(k)}$$

where

$$I_m^{(0)}=\int\limits_{\{|y|\leqq\varepsilon^\beta,\,k=1,\ldots,N\}}|\Delta^n_{i_1\ldots i_n}(x,\theta y)|\,|y_m|^n\,dF_{X_\varepsilon}(y)\leqq$$

$$\leqq c\varepsilon^\beta\mathbf{E}|X_\varepsilon|^n\leqq c_1\varepsilon^{n+\beta}.$$

In proving this estimate we used conditions (A) and (2.2), together with inequalities

$$|\Delta^n_{i_1\ldots i_n}(x,\theta y)|\leqq|y_m||p^{(n+1)}(\,\cdot\,)|\leqq c\varepsilon^\beta,$$

$$\mathbf{E}|X_\varepsilon|^n\leqq(\mathbf{E}|X_\varepsilon|^{n+\alpha})^{n/(n+\alpha)}.$$

Now,

$$I_m^{(k)}=\int\limits_{\{|y_k|>\varepsilon^\beta,\,|y_m|\leqq\varepsilon^\beta\}}|\Delta^n_{i_1\ldots i_n}(x,\theta y)|\,|y_m|^n\,dF_{X_\varepsilon}(y)\leqq$$

$$\leqq c\varepsilon^{n\beta}\mathbf{P}\{|X_\varepsilon|>\varepsilon^\beta\}\leqq c_1\varepsilon^{n+\alpha-\alpha\beta},\qquad k\neq m$$

$$I_m^{(m)}=\int\limits_{\{|y_m|>\varepsilon^\beta\}}|\Delta^n_{i_1\ldots i_n}(x,\theta y)|\,|y_m|^n\,dF_{X_\varepsilon}(y)\leqq$$

$$\leqq c\varepsilon^{n\beta}\int\limits_{\{|y_m|>\varepsilon^\beta\}}|y_m/\varepsilon^\beta|^n\,dF_{X_\varepsilon}(y)\leqq$$

$$\leqq c_1\varepsilon^{n\beta-(n+\alpha)\beta}\int\limits_{\mathbf{R}^N}|y_m|^{n+\alpha}\,dF_{X_\varepsilon}(y)<c_2\varepsilon^{n+\alpha-\alpha\beta}.$$

The above bounds for $I_m^{(0)}$, $I_m^{(k)}$ imply that

$$I \leqq c \min_{\beta} \max \{\varepsilon^{n+\beta}, \varepsilon^{n+\alpha-\alpha\beta}\} = c\varepsilon^{n+\gamma}, \qquad \gamma = \alpha/(1+\alpha).$$

Finally, the last estimate for I_m imply similar (up to some unessential multiplicative constant) estimate for $r_n(x, \varepsilon)$. The estimate (3.3) for $r_n(x, \varepsilon)$ is proved.

In order to prove (3.4) under the assumption that condition (2.3) is satisfied we have only to note that in this case the integral I_m satisfies $I_m \leqq cI_m^{(0)}$ for $\beta = 1$. This immediately implies (3.4). Lemma 1 is proved.

Lemma 2. Under the conditions of Theorem 1 the differential entropy $h(Y)$ can be written in the following form:

$$h(Y) = h(Z) - \int_{\mathbf{R}^N} a_{[v]}(x, \varepsilon) \log p(x) dx -$$

$$- \log e \sum_{k=2}^{\infty} \frac{(-1)^k}{(k-1)k} \int_{G_\delta} \frac{(a_n^k(x, \varepsilon) + r_n(x, \varepsilon))^k}{p^{k-1}(x)} dx -$$

$$- \int_{\bar{G}_\delta} p_Y(x) \log p_Y(x) dx + o(\varepsilon^v), \qquad \varepsilon \to 0 \tag{3.6}$$

where

$$\delta^2 = \delta^2(\varepsilon) = \frac{n+\varkappa}{\lambda+\mu} \ln \frac{1}{\varepsilon} \tag{3.7}$$

and $\varkappa$ is sufficiently small.

Proof. We will use the formula

$$h(Y) = - \int_{\bar{G}_\delta} p \log p \, dx - \int_{G_\delta} a_n \log p \, dx - \int_{G_\delta} r_n \log p \, dx -$$

$$- \log e \int_{G_\delta} (a_n + r_n) dx -$$

$$- \log e \sum_{k=2}^{\infty} \frac{(-1)^k}{(k-1)k} \int_{G_\delta} \frac{(a_n + r_n)^k}{p^{k-1}} dx - \int_{\bar{G}_\delta} p_Y \log p_Y dx \tag{3.8}$$

(here and below we will skip for brevity arguments in functions $p(x)$, $p_Y(x)$, $a_n(x, \varepsilon)$, $r_n(x, \varepsilon)$). Formula (3.8) easily follows from the equality

$$h(Y) = - \int_{\bar{G}_\delta} p_Y \log p_Y dx - \int_{G_\delta} (p + a_n + r_n) \left[\log p + \log \left(1 + \frac{a_n + r_n}{p} \right) \right] dx$$

(the convergence of the integral $\int_{\bar{G}_\delta} p_Y \log p_Y dx$ follows, for example, from Lemma 4 below) if we expand $\log\left(1+\frac{a_n+r_n}{p}\right)$ into power series, integrate it term by term and perform some rather natural algebraic transformations. The possibility of the term by term integration and the convergence of the corresponding integrals can be easily proved using conditions (A), (B), (2.7) together with estimates (3.3) or (3.4).

Now, condition (C1) and relations (3.7) and (2.8) immediately imply

$$\left| -\int_{G_\delta} p \log p\, dx - h(Z) \right| = \left| \int_{\bar{G}_\delta} p \log p\, dx \right| \leqq \delta^M e^{-\lambda\delta^2} = o(\varepsilon^\nu), \tag{3.9}$$

$$\left| \int_{G_\delta} a_n \log p\, dx - \int_{\mathbf{R}^N} a_{[\nu]} \log p\, dx \right| \leqq \left| \int_{\bar{G}_\delta} a_n \log p\, dx \right| +$$

$$+ \left| \int_{\mathbf{R}^N} (a_n - a_{[\nu]}) \log p\, dx \right| = o(\varepsilon^\nu). \tag{3.10}$$

By (2.10), together with (3.9) or (3.4) and definition (2.6) one easily gets the estimates

$$\int_{G_\delta} r_n\, dx = o(\varepsilon^\nu) \tag{3.11}$$

$$\int_{G^\delta} r_n \log p\, dx = o(\varepsilon^\nu). \tag{3.12}$$

Finally, to get the required estimate of $\int_{G_\delta} a_n\, dx$ we have to note first that condition (2.4) implies the equality $\int_{\mathbf{R}^N} a_n\, dx = 0$ so that

$$\left| \int_{G_\delta} a_n\, dx \right| = \left| \int_{\bar{G}_\delta} a_n\, dx \right| = o(\varepsilon^\nu). \tag{3.13}$$

Lemma 2 immediately follows from (3.9)–(3.13).

Lemma 3. In the assumptions of Theorem 1 the following equality holds:

$$\sum_{k=2}^{\infty} \frac{(-1)^k}{(k-1)k} \int_{G_\delta} \frac{(a_n+r_n)^k}{p^{k-1}}\, dx =$$

$$= \sum_{k=2}^{[\nu/2]} \frac{(-1)^k}{(k-1)k} \int_{\mathbf{R}^N} \frac{a_{[\nu]-2}^k}{p^{k-1}}\, dx + o(\varepsilon^\nu) \tag{3.14}$$

where $\delta = \delta(\varepsilon)$ is chosen according to (3.7).

Proof. Let us show first that

$$\left| \int\limits_{G_\delta} \frac{a_n^j r_n^{k-j}}{p^{k-1}} dx \right| \leqq \varepsilon^{\nu_1} \varepsilon_1^k, \qquad 0 \leqq j \leqq k-1, \quad k \geqq 2 \tag{3.15}$$

where $\nu_1 > \nu$ does not depend on k and $\varepsilon_1 = \varepsilon_1(\varepsilon) \to 0$ as $\varepsilon \to 0$. In fact, using estimates (3.3) (or (3.4)), (2.10), definition (2.6) and taking into account (3.7) together with the obvious estimate

$$|a(x, \varepsilon)| \leqq c\delta^M \varepsilon^2 p(x), \qquad x \in G_\delta$$

following from assumption (2.5) and equality $\mathbf{E}X_\varepsilon = 0$ we get

$$\left| \int\limits_{G_\delta} \frac{a_n^j r_n^{k-j}}{p^{k-1}} dx \right| \leqq c\delta^j \varepsilon^{2j} \varepsilon^{(n+\gamma)(k-j)} e^{(k-j-1)\delta^2} \operatorname{mes} G_\delta \leqq$$

$$\leqq c_1 \delta^{M+j} \varepsilon^{2j+(n+\gamma)(k-j)-(k-j-1)-\frac{n+\varkappa}{\lambda+\mu}-\frac{\mu(n+\varkappa)}{\lambda+\mu}}.$$

Performing rather simple algebraic transformations and using inequalities $k-j \geqq 1$ and $\lambda+\mu \geqq 1$ (see (2.9)) one can easily get that ε that the degree exponent for ε in the right-hand side of the last formula is not less that

$$\nu + \frac{\varkappa}{\lambda+\mu}\lambda + (\gamma-\varkappa)k + j(2-(\gamma-\varkappa)).$$

Therefore, choosing $\varkappa > 0$ in such a way that $\gamma - \varkappa > 0$ and denoting $\nu_1 = \nu + \frac{\varkappa}{\lambda+\mu}\lambda$, $\varepsilon_1 = \varepsilon^{\gamma-\varkappa}$ we get the required estimate (3.15).

Similar but even more simple arguments show that

$$\left| \int\limits_{G_\delta} \frac{a_n^k}{p^{k-1}} dx \right| \leqq (\delta^{M_1})^k \varepsilon^{2k}, \qquad M_1 > 0, \quad k \geqq 2. \tag{3.16}$$

Let us prove now the estimate

$$\left| \int\limits_{\bar{G}_\delta} \frac{a_n^k}{p^{k-1}} dx \right| \leqq c\varepsilon^2 \delta^M e^{-\lambda\delta^2}, \qquad k = 2, 3, \ldots, [\nu/2]. \tag{3.17}$$

To do this we estimate the integrals

$$I = \int\limits_{\bar{G}_\delta} \left[\left(\frac{\partial^{j_1} p}{\partial x_{i_1} \ldots \partial x_{i_{j_1}}} \right)^{k_1} \ldots \left(\frac{\partial^{j_s} p}{\partial x_{l_1} \ldots \partial x_{l_{j_s}}} \right)^{k_s} \Big/ p^k \right] p \, dx$$

for all $1 \leqq j_1, \ldots, j_s \leqq n$, $k_1 + \ldots + k_s = k$, $k = 2, 3, \ldots, [\nu/2]$.

To do this we use the generalized Hölder inequality

$$\int_E |b_1 \dots b_n| p\,dx \leqq \left(\int_E |b_1|^{q_1} p\,dx\right)^{1/q_1} \dots \left(\int_E |b_n|^{q_n} p\,dx\right)^{1/q_n}$$

where $q_i > 1$, $(1/q_1) + \dots + (1/q_n) = 1$. Applying this inequality and using condition (C2) we get immediately

$$|I| \leqq \delta^M e^{-\lambda\delta^2}.$$

This inequality immediately implies (3.17), because $\int_{\bar{G}_\delta} \frac{a_n^k}{p^{k-1}}\,dx$ can be expressed as the finite sum of integrals of type I, each multiplied by some mixed central moment of X_ε of order $\geqq 2$.

Estimates (3.15), (3.16) and (3.17) imply that

$$\sum_{k=2}^{\infty} \frac{(-1)^k}{(k-1)k} \int_{G_\delta} \frac{(a_n + r_n)^k}{p^{k-1}}\,dx = \sum_{k=2}^{[\nu/2]} \frac{(-1)^k}{(k-1)k} \int_{\mathbf{R}^N} \frac{a_n^k}{p^{k-1}}\,dx + o(\varepsilon^\nu).$$

In fact, first

$$\sum_{k=2}^{\infty} \frac{(-1)^k}{(k-1)k} \int_{G_\delta} \frac{(a_n + r_n)^k - a_n^k}{p^{k-1}}\,dx = o(\varepsilon^\nu)$$

(this series can be majorized by the series

$$\sum_{k=2}^{\infty} \frac{1}{(k-1)k} 2^k \varepsilon^{\nu_1} \varepsilon_1^k = o(\varepsilon^\nu)$$

because each term of this series consists of the sum of not more than 2^k integrals of the form

$$\int_{G_\delta} \frac{a_n^j r_n^{k-j}}{p^{k-1}}\,dx$$

for which estimate (3.15) is satisfied). Second, using estimate (3.16) one can easily see that

$$\sum_{k=2}^{\infty} \int_{G_\delta} \frac{a_n^k}{p^{k-1}}\,dx = \sum_{k=2}^{[\nu/2]} \int_{G_\delta} \frac{a_n^k}{p^{k-1}}\,dx + o(\varepsilon^\nu).$$

Finally, choosing δ according to (3.7), one gets from (3.7) that

$$\sum_{k=2}^{[v/2]} \frac{(-1)^k}{(k-1)k} \int_{\bar{G}_\delta} \frac{a_n^k}{p^{k-1}}\, dx = o(\varepsilon^v)$$

so that

$$\sum_{k=2}^{[v/2]} \frac{(-1)^k}{(k-1)k} \int_{G_\delta} \frac{a_n^k}{p^{k-1}}\, dx = \sum_{k=2}^{[v/2]} \frac{(-1)^k}{(k-1)k} \int_{\mathbf{R}^N} \frac{a_n^k}{p^{k-1}}\, dx + o(\varepsilon^v).$$

To complete the proof of Lemma 3 one has only to note that for any integral in the right-hand side of (3.18) one has an easily verified formula

$$\int_{\mathbf{R}^N} \frac{a_n^k}{p^{k-1}}\, dx = \int_{\mathbf{R}^N} \frac{a_{[v]-2}^k}{p^{k-1}}\, dx + o(\varepsilon^v).$$

Lemma 4. Let conditions (A) and (C3) be satisfied. Then, under restriction (2.3), we have

$$\int_{\bar{G}_\delta} p_Y(x) \log p_Y(x)\, dx = o(\varepsilon^v), \qquad \varepsilon \to 0$$

where δ is chosen according to (3.7); if, moreover, condition (D) is satisfied then (3.19) holds under restriction (2.2) as well.

Proof. We will use the following inequality

$$\exp\{(\int q \ln f\, dx)/\int q\, dx\} \leqq (\int q\, f^r dx/\int q\, dx)^{1/r}$$

(see [4], pp. 163–167, formulas (6.6.1), (6.7.1), (6.7.6)). Here the integration is performed over some set E, functions $f = f(x) \geqq 0$ and $q = q(x) \geqq 0$ are finite almost everywhere on E, $f(x) \neq \text{const}$, and $r \neq 0$ is a real number. Taking in this inequality $E = \bar{G}_\delta$, $r = 1$, $q = p_Y$, $f = p/p_Y$ we get after some transformations (using, in particular, the inequality $\left(\int_{\bar{G}_\delta} p_Y dx\right) \log\left(\int_{\bar{G}_\delta} p_Y dx\right) \leqq 0$), we get

$$0 \leqq -\int_{\bar{G}_\delta} p_Y \log p_Y\, dx \leqq -\int_{\bar{G}_\delta} p_Y \log p\, dx - \left(\int_{\bar{G}_\delta} p_Y dx\right) \log\left(\int_{\bar{G}_\delta} p_Y dx\right) \tag{3.20}$$

(the non-negativity of the integral $\int_{\bar{G}_\delta} p_Y \log p_Y\, dx$ for sufficiently small ε and sufficiently large δ follows, for example, from representation (3.1) showing that $p_Y(x) < 1$ for $x \in \bar{G}_\delta$).

Let us estimate now both summands in the right-hand side of (3.20) assuming that restriction (2.2) is satisfied:

$$\left| -\int_{\bar{G}_\delta} p_Y \log p_Y dx \right| \leqq \int_{|y|<\theta} \left(\sup_{|y|\leqq\theta} \left| \int_{\bar{G}_\delta} p(x-y) \log p(x) dx \right| \right) dF_{X_\varepsilon}(y) +$$

$$+ \int_{|y|\geqq\theta} \left(\int_{\bar{G}_\delta} |p(x-y) \log p(x)| dx \right) dF_{X_\varepsilon}(y) \leqq$$

$$\leqq \delta^M e^{-\lambda\delta^2(1-\beta)} + \int_{|y|\geqq\theta} c|y|^{n+\alpha} dF_{X_\varepsilon}(y) \leqq c_1 \varepsilon^{\nu+\tilde{\beta}} \tag{3.21}$$

where $\tilde{\beta}>0$. In proving (3.21) we have used (2.2), equality (2.7) and conditions (C3) and (D). If, on the other hand, restriction (2.3) is satisfied, then the second integral in the right-hand side of the first inequality in (3.21) vanishes and (3.21) follows from condition (C3).

As $p_Y(x) \leqq |p_Y(x) \log p(x)|$ on $\bar{G}_\delta$ for sufficiently small δ, we have

$$\left| \int_{\bar{G}_\delta} p_Y dx \right| \leqq c_1 \varepsilon^{\nu+\tilde{\beta}}, \qquad \tilde{\beta}>0. \tag{3.22}$$

Formula (3.19) follows from (3.21), (3.22) and (3.20). Lemma 4 is proved.

4. Asymptotic expansions for the capacity

Theorem 1 proved in Section 3 is the basis of all the results about asymptotic expansions of the capacity of the considered memoryless channel under various restrictions on the input signal. To simplify the exposition we will restrict ourselves in this section to the one-dimensional case.

Let us note first that, by Theorem 1, the mutual information $I(X_\varepsilon, Y)$ considered up to terms of order $o(\varepsilon^m)$ for some integer m is a polynomial of order $\leqq m$ on moments of the distribution of the input signal X_ε with coefficients depending on the density of the noise. Therefore, by the well-known Caratheodory theorem (see, e.g. [5], Theorem 3.6, p. 31) to find the asymptotic expansion of the capacity up to terms of order $o(\varepsilon^m)$ it is sufficient to consider discrete distributions of the input signal supported in $\leqq m+1$ points. Therefore, to find the asymptotic expansion for the capacity up to $o(\varepsilon^m)$ we have to solve a finite-dimensional problem: to find the maximum point of a function of $2(m+1)$ variables under conditions (2.15) and (2.16) (or one of them) and the conditions $\sum_{i=1}^{m+1} p_i = 1$ and $p_i \geqq 0$, $i=1, \ldots, m+1$ (here $\{p_i\}$, $i=1, \ldots, m+1$ is the probability distribution of the input signal X_ε supported at points $x_1, \ldots, x_{m+1}$). Although to obtain the complete solution of this problem seems to be impossible, in

several special cases one can get explicit expressions for the first few terms of the asymptotic expansion of the capacity, and in some cases, as for example, Theorem 2 shows from Section 2, one can get the complete asymptotic expansion.

Proof of Theorem 2. Let us prove the first statement of Theorem 2. The convexity property of the mutual information $I(X_\varepsilon, Y)$ with respect to the distribution of X_ε and the symmetry of the noise probability density $p(x)$ imply that the upper bound of $I(X_\varepsilon, Y)$ under conditions (2.15) and (2.16) is achieved on signals X_ε with the symmetric (with respect to the origin) distribution. Under this assumption, together with the assumptions of part 1 of Theorem 2, the mutual information $I(X_\varepsilon, Y)$ can, according to Theorem 1, be written as follows:

$$I(X_\varepsilon, Y) = \sum_{l=1}^{m} \sum_{k_1 + \ldots k_l = 2l} b_{k_1 \ldots k_l} \mathbf{E} X_\varepsilon^{k_1} \mathbf{E} X_\varepsilon^{k_2} \ldots \mathbf{E} X_\varepsilon^{k_l} + o(\varepsilon^{2m}) \tag{4.1}$$

where the second summation is taken over all partitions of the number $2l$ into l even summands $k_1 \geqq k_2 \ldots \geqq k_l \geqq 0$; the coefficients $b_{k_1 \ldots k_l}$ depend on the noise probability density $p(x)$ and can be computed using (2.12). In particular, one can easily check that

$$b_2 = \frac{\log e}{2} \int_{\mathbf{R}^1} \frac{(p')^2}{p}\, dx, \qquad b_{4,0} = \frac{\log e}{24} \int_{\mathbf{R}^1} \frac{p' p'''}{p}\, dx,$$

$$b_{2,2} = -\frac{\log e}{8} \int_{\mathbf{R}^1} \frac{(p'')^2}{p}\, dx$$

and so on.

Let X_ε^* be the optimal input signal, i.e. the one maximizing $I(X_\varepsilon, Y)$ under restriction (2.15), and let X_ε^0 be the signal taking values $-\varepsilon$ and ε with equal probabilities. Then, using (4.1) and the equalities $\mathbf{E}(X_\varepsilon^0)^{2k} = \varepsilon^{2k}$, $k = 1, \ldots, m$, one can express the difference $I(X_\varepsilon^*, Y) - I(X_\varepsilon^0, Y)$ in the following form:

$$I(X_\varepsilon^*, Y) - I(X_\varepsilon^0, Y) =$$

$$= \sum_{l=2}^{m} \sum_{\substack{k_1 + \ldots k_l = 2l \\ k_1 \geqq \ldots \geqq k_l \geqq 0 \\ k_i \text{ even}}} b_{k_1 \ldots k_l} (\mathbf{E}(X_\varepsilon^*)^{k_1} \ldots \mathbf{E}(X_\varepsilon^*)^{k_l} - \varepsilon^{2l}) + o(\varepsilon^{2m}). \tag{4.2}$$

Setting

$$\mathbf{E}(X_\varepsilon^*)^2 = \varepsilon^2 (1 - f(\varepsilon)), \qquad f(\varepsilon) \geqq 0,$$

and using (2.15) and the known relations among moments of the distributions we get

$$\varepsilon^{2j} \geqq \mathbf{E}(X_\varepsilon^*)^{2j} \geqq \varepsilon^{2j} (1 - f(\varepsilon))^{2j}. \tag{4.3}$$

Let us note that, obviously, these bounds remain true under condition (2.16) as well.

As the difference $I(X_\varepsilon^*, Y)-I(X_\varepsilon^0, Y)$ should be non-negative, (4.3) implies immediately that, first, $f(\varepsilon)\to 0$ for $\varepsilon\to 0$, and, second, the above difference can be represented in the following form:

$$I(X_\varepsilon^*, Y)-I(X_\varepsilon^0, Y)=-b_2\varepsilon^2 f(\varepsilon)+$$

$$+\sum_{l=2}^{m}\sum_{\substack{k_1+\ldots k_l=2l\\ k_1\geq\ldots\geq k_l\geq 0\\ k_i \text{ even}}} b_{k_1\ldots k_l} o(\varepsilon^{2l} f(\varepsilon^{2l}))+o(\varepsilon^{2m}).$$

As $b_2>0$, we immediately get that $\varepsilon^2 f(\varepsilon)=o(\varepsilon^{2m})$ and this implies the first statement of Theorem 2.

Let us prove now the second statement. Let us take for X_ε the random variable with the distribution

$$\mathbf{P}\{X_\varepsilon=-K\varepsilon\}=\mathbf{P}\{X_\varepsilon=K\varepsilon\}=1/2K^2, \qquad \mathbf{P}\{X_\varepsilon=0\}=1-1/K^2$$

if

$$\int\frac{p'p'''}{p}\,dx\geqq 0$$

and with the distribution

$$\mathbf{P}\{X_\varepsilon=-\varepsilon\}=\mathbf{P}\{X_\varepsilon=\varepsilon\}=1/2$$

if

$$\int\frac{p'p'''}{p}\,dx<0$$

and substituting into (2.12) we see that the right-hand side of (2.19) is a lower bound for $C^{(2,K)}(\varepsilon)$.

In order to prove that the right-hand side of (2.19) is also an upper bound for $C^{(2,K)}(\varepsilon)$ one has to apply arguments similar to ones used in the proof of the first part of Theorem 2 taking into account the sign of the summand

$$\frac{1}{24}\int_{\mathbf{R}^1}\frac{p'p'''}{p}\,dx\,\mathbf{E}X_\varepsilon^4$$

entering into the asymptotic expansion (4.1) and the inequality $\mathbf{E}X_\varepsilon^4\leqq K^2\varepsilon^4$; this last inequality follows from (2.15) with $h=2$ and from (2.16) because

$$\mathbf{E}X_\varepsilon^4\leqq(K\varepsilon)^2\int_{-K\varepsilon}^{K\varepsilon}x^2 dP_{X_\varepsilon}(x)\leqq K^2\varepsilon^4.$$

Thus, the proof of Theorem 2 is complete.

Let us remark in conclusion that in a general case when the density of the noise distribution is not assumed to be an even function, to get explicit expressions for nonprincipal terms of the asymptotic expansion for the capacity appears to be a much more complicated problem. This problem, as well as the corresponding multi-dimensional problems, will be studied in a separate paper.

References

1. *Prelov, V. V.*, Asymptotic behavior of the capacity of a continuous channel with large smooth noise, Probl. Peredachi Inform., 1980, **16,** *2*, 3–17.
2. *Prelov, V. V.*, The asymptotic behavior of the capacity and the zero-error capacity of a continuous channel with large noise, Technical Report no. *179*, K. U. Leuven, 1984, 27 pp.
3. *Prelov, V. V.*, Asymptotic expansion for the capacity of almost Gaussian channels, Proc. 3rd Soviet–Swedish Workshop on Inform. Theory, Sochi, 1987, pp. 138–141.
4. *Hardy, G. H., Littlewood, J. E., Pólya, G.*, Inequalities. 1934.
5. *Krein, M. G., Nudelman, A. A.*, Markov moment problem and extremum problems. Moscow, Nauka, 1973, 551 pp.

Асимптотические раложения информации и пропускной способности непрерывных каналов без памяти при слабом входном сигнале

В. В. ПРЕЛОВ

(Москва)

Приведены явные условия на плотность распределения $p(x)$ шума Z в непрерывном канале без памяти с независимым аддитивным шумом $Y=X_\varepsilon+Z$ в предположении, что сигнал на входе X_ε, зависящий от малого параметра $\varepsilon>0$, удовлетворяет либо ограничению на абсолютный момент порядка $n+\alpha$ вида $\mathbf{E}|X_\varepsilon/\varepsilon|^{n+\alpha}\leqq K<\infty$ для некоторого натурального $n\geqq 2$ и $0<\alpha<1$, либо более сильному ограничению на пиковую мощность $|X_\varepsilon|\leqq K\varepsilon$, при выполнении которых информация $I(X_\varepsilon; Y)$ допускает следующее асимптотическое разложение

$$I(X_\varepsilon; Y)=-\int_{\mathbf{R}^N} a_{[\nu]}(x)\log p(x)dx-$$

$$-\log e\sum_{k=2}^{[\nu/2]}\frac{(-1)^k}{(k-1)^k}\int_{\mathbf{R}^N}\frac{a^k_{[\nu]-2}(x)}{p^{k-1}(x)}dx+o(\varepsilon^\nu),\qquad \varepsilon\to 0,$$

где

$$a_m(x)=\sum_{k=2}^{m}\frac{(-1)^k}{k!}\mathbf{E}\left\{\left(\frac{\partial}{\partial x_1}(X_{\varepsilon 1}-\mathbf{E}X_{\varepsilon 1})+\ \dots\ +\frac{\partial}{\partial x_N}(X_{\varepsilon N}-\mathbf{E}X_{\varepsilon N})\right)^k p(x)\right\},$$

а для константы ν, $\nu\leqq n$ указано явное выражение.

Этот результат используется затем для вывода асимптотических разложений пропускной способности рассматриваемого канала при различного типа ограничениях на входной сигнал.

В. В. Прелов
Институт проблем передачи информации АН СССР
СССР 101447 Москва ГСП-4, ул. Ермоловой, 19

Problems of Control and Information Theory, Vol. 18 (2), pp. 107—123 (1989)

TRACKING OF TIME-VARYING PARAMETERS IN DELTA MODELS

R. KULHAVÝ, E. KLIOKYS

(Prague, Kaunas)

(Received April 10, 1988)

A difference operator description of system input-output behaviour by delta models is promising to achieve a higher numerical robustness mainly in case of fast sampling. Recent results in state and parameter estimation for delta models are extended in this paper to the case of time-varying parameters. Tracking of time-varying parameters is achieved by modelling their variations through a simple one-factor random walk. The model is adapted so that just information modified by the latest data may be forgotten. A reliable parameter tracking is ensured in such a way even when the identified system is poorly excited.

1. Introduction

Recently Peterka [7] described a new algorithmic solution of LQG self-tuning control based on so-called delta models which express the input-output relation through difference operators. The concept of delta models has been introduced in [2]. The delta representation has been found less sensitive to numerical errors produced by a finite word-length of the computer, especially in case when the system to be controlled is fast sampled.

The tasks of state and parameter estimation, output prediction as well as control synthesis were solved in [7] in an excellent tutorial way for the special case of constant parameters of the delta model. However, this assumption forms a serious resctriction for practical applications. For this reason, we have studied various possibilities of extending the results of [7] to time-varying processes, too.

The solution given below represents a natural extension of the approach chosen by Peterka in [7] to obtain joint estimates of both state and constant parameters. We propose an extremely simple model of parameter variations motivated by experience from investigation of rational forgetting in parameter estimation (cf. [4]). Using this model we derive (in a consistent Bayesian way) an algorithm for estimation of both state and time-varying parameters.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Owing to the special (nontraditional) form of the model of parameter variations, the algorithm is extraordinarily robust to noninformative data caused by improper excitation of the identified system. Thus, reliable operation of the algorithm is ensured even in case of linear feedback, overparametrization, rare changes of external disturbance, or input saturation, to mention only a few of the possible causes of poor system excitation.

The derivation as well as motivation of the results adopted from other papers are reduced or omitted here because of the limited scope. The emphasis is laid on the interpretation of the resulting algorithm and elucidating differences between the cases of constant and time-varying parameters.

2. Delta/ARMA models

1. We consider a stochastic process (system output) $y(t)$ which can be influenced by a sequence of previous system inputs $u(\tau)$, $\tau = 1, 2, \ldots, t$. The data items $u(t)$, $y(t)$ are observed at discrete time instants labelled $t = 1, 2, \ldots$. Both input and output are supposed univariate. Extension to multivariate models is touched briefly in Section 5.

2. Throughout the paper we assume the dependence of the system output on previous data described by means of the canonical state model introduced in [7]

$$A(t)\begin{Vmatrix} y(t) \\ s(t) \end{Vmatrix} = Hs(t-1) + b(t)u(t) + k_x(t) + ce(t). \tag{2.1}$$

Here $s(t)$ denotes the n-dimensional state vector and $e(t)$ stands for the univariate white-noise component which is normally distributed with zero mean and variance

$$\mathrm{Var}\,[e(t)] = \rho. \tag{2.2}$$

Matrices $A(t)$, H and vectors $b(t)$, $k_x(t)$, c are composed as follows:

$$A(t) = \begin{Vmatrix} 1 & 0 \\ \bar{a}(t) & I_n \end{Vmatrix}$$

with $\bar{a}'(t) = \|a_1(t), a_2(t), \ldots, a_n(t)\|$,

$$H = \mu \begin{Vmatrix} 0 \\ I_n \end{Vmatrix} + \begin{Vmatrix} I_n \\ 0 \end{Vmatrix},$$

$$b'(t) = \|b_0(t), b_1(t), \ldots, b_n(t)\|,$$

$$k_x'(t) = \|0, 0, \ldots, k_c(t)\|,$$

$$c' = \|1, c_1, \ldots, c_n\| \tag{2.3}$$

where the symbol ′ denotes transposition and I_n stands for the square unit matrix of order n. Equation (2.1) covers in a unified form both delta (for $\mu=1$) and ARMA (for $\mu=0$) models. The reader interested in the way how this representation has been obtained should consult [7] for a detailed discussion.

3. The parameters $a(t)$, $b(t)$, $k_x(t)$ of the delta/ARMA model (2.1) are taken unknown and time-varying. It is useful to treat them as an n-dimensional vector

$$\theta'(t)=\|a_1(t), \ldots, a_n(t), b_0(t), \ldots, b_n(t), k_c(t)\|. \tag{2.4}$$

The results presented below can be extended in a straightforward way to the case of unknown and even time-varying variance ρ, too (in the vein of [6] and [4]). However, as knowledge of ρ is not required for standard adaptive control strategies, we prefer here to follow the line of explanation adopted in [7] and suppose ρ known.

The reason for avoiding the case of unknown parameters c is more essential. We are not able to give a satisfactory recursively feasible Bayesian solution of the problem of estimating the parameters c. Therefore, we rely on the possibility to specify the values of c a priori.

4. In addition to the problem formulation in [7] we admit that the parameters θ may vary in time. Their variations will be modelled by means of the random walk

$$\theta(t+1)=\theta(t)+v(t), \quad v(t)\sim N(0, \rho C_{\Delta\theta}(t)) \tag{2.5}$$

with the normally distributed discrete white noise component $v(t)$ having the zero mean and a covariance matrix $\rho C_{\Delta\theta}(t)$ depending in general on observed data.

The relationship between modelling parameter variations and forgetting obsolete information is well known (see e.g. [5]) and has been also utilized in solving some problems of parameter estimation (cf. [1]). The distinguishing feature in comparison with the cited papers is a special structure of $C_{\Delta\theta}$ which will be introduced in Section 4 with the aim to increase uncertainty of only those parameters about which the latest data have brought some information.

3. Joint state and parameter estimation

1. From the Bayesian point of view full information about the system state and parameters is represented by the joint probability density function $p(\theta(t), s(t-1)|\,t-1)$ conditioned on all data available up to the time $t-1$. Using a new measurement $u(t)$, $y(t)$ and taking into account the assumed model of parameter variations, we can update the density as follows

$$p(\theta(t+1), s(t)|\,t)\propto$$

$$\propto\int\int p(y(t), s(t), \theta(t+1)|\,t-1; u(t), s(t-1), \theta(t))\times$$

$$\times p(\theta(t), s(t-1)|t-1)\, ds(t-1)\, d\theta(t) \tag{3.1}$$

where the symbol $\propto$ stands for the proportionality (i.e. equality up to a normalizing factor). The above relation follows directly from basic formulae of probability theory as given e.g. in [7].

2. For further explanation it is useful to factorize operation (3.1) into two steps.

(a) First, we carry out the measurement update and state time update, i.e. using the canonical state model (2.1) we evaluate

$$p(\theta(t), s(t)|t) \propto$$

$$\propto \int p(y(t), s(t)|t-1; u(t), s(t-1), \theta(t)) \times$$

$$\times p(\theta(t), s(t-1)|t-1) ds(t-1). \tag{3.2}$$

(b) Second, we perform the parameter time update employing the model of parameter variations (2.5)

$$p(\theta(t+1), s(t)|t) \propto$$

$$\propto \int p(\theta(t+1)|t; s(t), s(t-1), \theta(t)) \times$$

$$\times p(\theta(t), s(t)|t)\, d\theta(t). \tag{3.3}$$

Supposing the transition probability due to (2.5) depends only on previous data and not on state values

$$p(\theta(t+1)|t; s(t), s(t-1), \theta(t)) = p(\theta(t+1)|t; \theta(t)) \tag{3.4}$$

relation (3.3) simplifies to

$$p(\theta(t+1), s(t)|t) \propto$$

$$\propto \int p(\theta(t+1)|t; \theta(t))\, p(\theta(t), s(t)|t) d\theta(t). \tag{3.5}$$

3. The just performed factorization makes it possible to separate the cases of estimation with constant and time-varying parameters. We employ the fact that step (3.2) has been completely solved by Peterka in [7] and focus our attention on solving step (3.5).

4. Algorithmic solution

1. Let the prior joint probability density function of unknown parameters $\theta(t)$, $s(t-1)$ at the time t (before observing $u(t)$, $y(t)$) be normal

$$p(\theta(t), s(t-1)|t-1) \propto \exp\left\{ -(1/2\rho) \times \right.$$

$$\times \left[(\theta(t)-\hat{\theta}(t|t-1))' C_\theta^{-1}(t|t-1)\,(\theta(t)-\hat{\theta}(t|t-1)) + \right.$$

$$+ \left(s(t-1)-X(t|t-1) \left\| \begin{matrix} 1 \\ \theta(t) \end{matrix} \right\| \right)' C_{s|\theta}^{-1}(t|t-1) \times$$

$$\left. \times \left(s(t-1)-X(t|t-1) \left\| \begin{matrix} 1 \\ \theta(t) \end{matrix} \right\| \right) \right] \Bigg\}. \tag{4.1}$$

Notice that the density is fully specified by the statistics $\hat{\theta}(t|t-1)$, $C_\theta(t|t-1)$, $X(t|t-1)$, $C_{s|\theta}(t|t-1)$.

2. After observing new data $u(t)$, $y(t)$, we compose the vector of filtered data $\|\bar{y}(t),\ \bar{z}'(t)\|'$ using the statistics $X(t|t-1)$:

$$\begin{aligned} \bar{y}(t) &= y(t) - X_{1,1}(t|t-1) \\ \bar{z}_{n+1}(t) &= u(t) + X_{1,n}(t|t-1) \\ \bar{z}_i(t) &= X_{1,i-1}(t|t-1) \qquad \text{for} \quad i \neq n+1. \end{aligned} \tag{4.2}$$

The role of this vector will be cleared up at once.

3. The following theorem summarizes the solution of the data update and state time update according to (3.2).

Theorem 1. If the joint probability density function $p(\theta(t), s(t-1)|t-1)$ is of the form (4.1), then $p(\theta(t), s(t)|t)$ is of the same form but with updated statistics $\hat{\theta}(t|t)$, $C_\theta(t|t)$, $X(t|t)$, $C_{s|\theta}(t|t)$. Introducing the following auxiliary statistics

$$\hat{\varepsilon}(t|t-1) = \bar{y}(t) - \hat{\theta}'(t|t-1)\bar{z}(t) \tag{4.3}$$

$$\zeta(t|t-1) = \bar{z}'(t) C(t|t-1) \bar{z}(t) \tag{4.4}$$

and denoting by I_k^j the j-th column of I_k (the unit matrix of order k), the corresponding recursion is done by

(a) update of the conditional state covariance:

$$HC_{s|\theta}(t|t-1)H' + cc' = \left\| \begin{matrix} d_y(t) & d_y(t)\tilde{c}'(t) \\ \tilde{c}(t)d_y(t) & C_{s|\theta}(t|t) \end{matrix} \right\|, \tag{4.5}$$

(b) update of the parameter statistics:

$$\hat{\theta}(t|t) = \hat{\theta}(t|t-1) + \frac{C_\theta(t|t-1)\bar{z}(t)\hat{\varepsilon}(t|t-1)}{d_y(t) + \zeta(t|t-1)} \tag{4.6}$$

$$C_\theta^{-1}(t|t) = C_\theta^{-1}(t|t-1) + \frac{\bar{z}(t)\bar{z}'(t)}{d_y(t)}, \tag{4.7}$$

(c) update of the data statistics:

$$X(t|t) = [\mu I_n + \|\tilde{c}(t), I_n^1, \ldots, I_n^{n-1}\|]X(t|t-1) + \\ + \|-\tilde{c}(t), I_n\| \cdot \|y(t)I_{n+1}, u(t)I_{n+1}, I_{n+1}^{n+1}\|. \tag{4.8}$$

Proof. The above equations have been proved in [7] by substituting system model (2.1) and joint density (4.1) into the relation (3.2). ∎

Notice the relationships among particular steps. The scalar $d_y(t)$ evaluated in (a) is used in the parameter estimation (b). Similarly, the vector $\tilde{c}(t)$ computed as a by-product in (a) is employed for generating $X(t|t)$ in (c).

Note that there is a close link between parameter estimation of regression-type models and delta/ARMA models. Describing the input-output system behaviour by a regression-type model on filtered data

$$\bar{y}(t) = \theta'(t)\bar{z}(t) + \bar{e}(t) \qquad \text{with} \quad \bar{e}(t) = \sqrt{d_y(t)}\, e(t) \tag{4.9}$$

then using the standard Bayesian estimation scheme (cf. [6]), we derive immediately the recursive formulae (4.6) and (4.7).

4. Before performing the parameter time update step, we return to the specification of the model of parameter variations. First, let us analyse the effect of data on the uncertainty of unknown parameters $\theta(t)$. It is clear from (4.9) that taking into account just the last observation we are able to distinguish only among the values of $\theta(t)$ giving different values of $\theta'(t)\bar{z}(t)$ (the projection onto the the straight line with the direction of $\bar{z}(t)$).

If $\zeta(t|t-1) > 0$, the marginal probability density function of $\theta'(t)\bar{z}(t)$ can be evaluated using the basic properties of the Gaussian distribution

$$p(\theta'(t)\bar{z}(t)|t) \propto \\ \propto \exp\left\{ -(1/2\rho)\frac{(\theta'(t)\bar{z}(t) - \hat{\theta}'(t|t)\bar{z}(t))^2}{\zeta(t|t)} \right\}. \tag{4.10}$$

The variable $\zeta(t|t)$ is defined by

$$\zeta(t|t) = \bar{z}'(t) C_\theta(t|t) \bar{z}(t) \tag{4.11}$$

or, in terms of statistics $\zeta(t|t-1)$, $d_y(t)$,

$$\frac{1}{\zeta(t|t)} = \frac{1}{\zeta(t|t-1)} + \frac{1}{d_y(t)}. \tag{4.12}$$

Now we can express $p(\theta(t)|t)$ in the factorized form

$$p(\theta(t)|t) \propto \exp\left\{ -(1/2\rho) \times \right.$$

$$\times (\theta(t)-\hat{\theta}(t|t))' \frac{\bar{z}(t)\bar{z}'(t)}{\zeta(t|t)} (\theta(t)-\hat{\theta}(t|t)) \Bigg\} \times$$

$$\times \exp \Bigg\{ -(1/2\rho)(\theta(t)-\hat{\theta}(t|t))' \times$$

$$\times \left(C_\theta^{-1}(t|t) - \frac{\bar{z}(t)\bar{z}'(t)}{\zeta(t|t)} \right) (\theta(t)-\hat{\theta}(t|t)) \Bigg\}. \tag{4.13}$$

Notice that all new information contained in the latest data has been incorporated into the first (marginal probability) factor. Thus, if the identified system is poorly excited due to a linear dependency of filtered observations $\bar{z}_i(t)$, we do not obtain information about changes of some parameters and the decision whether the corresponding parameters vary in time actually or not may be based only on prior information. As the evolution of parameters is rarely well known in practice, we prefer to model merely time variations of the linear projection $\theta'(t)\bar{z}(t)$ about which we have obtained new information. To increase the variance of $\theta'(t)\bar{z}(t)$, equal to the scalar $\rho\zeta(t|t)$, as simply as possible, we multiply it by the factor $1/\varphi$, $\varphi \in (0, 1)$. In such a way we derive

$$p(\theta(t+1)|t) \propto \exp \{ -(1/2\rho)(\theta(t+1)-\hat{\theta}(t|t))' \times$$

$$\times C_\theta^{-1}(t+1|t)(\theta(t+1)-\hat{\theta}(t|t)) \} \tag{4.14}$$

with

$$C_\theta^{-1}(t+1|t) = C_\theta^{-1}(t|t) - (1-\varphi)\frac{\bar{z}(t)\bar{z}'(t)}{\zeta(t|t)}. \tag{4.15}$$

Inversion of (4.15) gives

$$C_\theta(t+1|t) = C_\theta(t|t) + \frac{1-\varphi}{\varphi} \frac{C_\theta(t|t)\bar{z}(t)\bar{z}'(t)C_\theta(t|t)}{\zeta(t|t)}. \tag{4.16}$$

However, the result (4.14) can be obtained immediately when taking the second term of (4.16) as the matrix $C_{\Delta\theta}(t)$ in the model of parameter variations (2.5).

Another situation arises if $\zeta(t|t-1)=0$ either because of the zero regressor $\bar{z}(t)=0$, or because of the zero Kalman gain $C(t|t-1)\bar{z}(t)=0$ (if $C(t|t-1)$ becomes singular). In this case the data update step has no effect on the uncertainty of unknown parameters, thus, it is reasonable to omit the time update step, too.

To sum up, we suggest to take a special structure of the covariance matrix of parameter increments in time through

$$C_{\Delta\theta}(t) = \begin{cases} \dfrac{1-\varphi}{\varphi} \dfrac{C_\theta(t|t)\bar{z}(t)\bar{z}'(t)C_\theta(t|t)}{\zeta(t|t)} & \text{if} \quad \zeta(t|t-1)>0 \\ C_{\Delta\theta}(t)=0 & \text{if} \quad \zeta(t|t-1)=0. \end{cases} \tag{4.17}$$

Using the auxiliary scalar

$$\alpha(t)=\frac{\zeta(t|t)}{\zeta(t|t-1)} \tag{4.18}$$

we can express (4.17) in terms of prior statistics

$$C_{\Delta\theta}(t)=\frac{1-\varphi}{\varphi}\,\alpha(t)\,\frac{C_\theta(t|t-1)\bar{z}(t)\bar{z}'(t)C_\theta(t|t-1)}{\zeta(t|t-1}\qquad \text{if}\quad \zeta(t|t-1)>0$$

$$C_{\Delta\theta}(t)=0\qquad \text{if}\quad \zeta(t|t-1)=0. \tag{4.19}$$

5. The solution of the parameter time update step (3.5) for the model (2.5) with (4.19) is described by the following theorem. This theorem represents the main result of the paper.

Theorem 2. If the joint probability density function $p(\theta(t), s(t)|t)$ is of the form (4.1) with the statistics $\hat{\theta}(\mu t|t)$, $C_\theta(t|t)$, $X(t|t)$, $C_{s|\theta}(t|t)$,then $p(\theta(t+1), s(t|t)$ is of the same form but with updated statistics $\hat{\theta}(t+1|t)$, $C_\theta(t+1|t)$, $X(t+1|t)$, $C_{s|\theta}(t|t)$. To evaluate them, let us introduce the following auxiliary variables

$$\kappa(t)=(1-\varphi)/\zeta(t|t) \tag{4.20}$$

$$h(t)=\alpha(t)Z(t|t)C_\theta(t|t-1)\bar{z}(t) \tag{4.21}$$

$$\hat{y}(t|t)=\bar{y}(t)-\alpha(t)\hat{\varepsilon}(t|t-1) \tag{4.22}$$

where quantities (4.12) and (4.18) are referred to and $Z(t|t)$ denotes the submatrix of $X(t|t)$ arisen by omitting its first column. If $\zeta(t|t-1)>0$, the recursion of the statistics due to the parameter evolution is done by

(a) update of the conditional state covariance:

$$C_{s|\theta}(t+1|t)=C_{s|\theta}(t|t)+\kappa(t)h(t)h'(t) \tag{4.23}$$

(b) update of the parameter statistics:

$$\hat{\theta}(t+1|t)=\hat{\theta}(t|t) \tag{4.24}$$

$$C_\theta^{-1}(t+1|t)=C_\theta^{-1}(t|t)-\kappa(t)\bar{z}(t)\bar{z}'(t) \tag{4.25}$$

(c) update of the data statistics:

$$X(t+1|t)=X(t|t)-\kappa(t)h(t)\left\|\begin{matrix}-\hat{y}(t|t)\\ \bar{z}(t)\end{matrix}\right\|'. \tag{4.26}$$

If $\zeta(t|t-1)=0$, the statistics $\hat{\theta}(t|t)$, $C_\theta(t|t)$, $X(t|t)$, $C_{s|\theta}(t|t)$ remain without change.

Proof. First, we rewrite the joint normal probability density function $p(\theta(t), s(t)|t)$ into the form

$$p(\theta(t), s(t)|t) \propto \exp\left\{ -(1/2\rho) \left\| \begin{matrix} \theta(t)-\hat{\theta}(t|t) \\ s(t)-\hat{s}(t|t) \end{matrix} \right\|' \times \right.$$

$$\left. \times \left\| \begin{matrix} C_{\theta\theta}(t|t)\ C'_{s\theta}(t|t) \\ C_{s\theta}(t|t)\ C_{ss}(t|t) \end{matrix} \right\|^{-1} \left\| \begin{matrix} \theta(t)-\hat{\theta}(t|t) \\ s(t)-\hat{s}(t|t) \end{matrix} \right\| \right\}.$$

Then we perform integration according to (3.5) (assuming the case $\zeta(t|t-1)>0$). To avoid pseudoinversion of the singular matrix $C_{\Delta\theta}(t)$, it is suitable first to perform the integration for a positive definite matrix $C_{\Delta\theta}$ and then find the limit of the resulting solution for $C_{\Delta\theta}\to C_{\Delta\theta}(t)$ (4.19). After straightforward but cumbersome arrangements which are based on repetitive application of the matrix inversion lemma (see Rao [9]), we derive

$$p(\theta(t+1), s(t)|t) \propto \exp\left\{ -(1/2\rho) \left\| \begin{matrix} \theta(t+1)-\hat{\theta}(t|t) \\ s(t)-\hat{s}(t|t) \end{matrix} \right\|' \times \right.$$

$$\left. \times \left\| \begin{matrix} C_{\theta\theta}(t+1|t)\ C'_{s\theta}(t+1|t) \\ C_{s\theta}(t+1|t)\ C_{ss}(t+1|t) \end{matrix} \right\|^{-1} \left\| \begin{matrix} \theta(t+1)-\hat{\theta}(t|t) \\ s(t)-\hat{s}(t|t) \end{matrix} \right\| \right\}$$

with the statistics

$$C_{\theta\theta}^{-1}(t+1|t) = C_{\theta\theta}^{-1}(t|t) - \kappa(t)\bar{z}(t)\bar{z}'(t)$$

$$C_{s\theta}(t+1|t) = C_{s\theta}(t|t)$$

$$C_{ss}(t+1|t) = C_{ss}(t|t)$$

where $\kappa(t)$ is defined by (4.20).

Now we arrange $p(\theta(t+1), s(t)|t)$ into the form of (4.1) again. After necessary manipulations we get

$$C_{\theta}^{-1}(t+1|t) = C_{\theta}^{-1}(t|t) - \kappa(t)\bar{z}(t)\bar{z}'(t)$$

$$C_{s|\theta}(t+1|t) = C_{s|\theta}(t|t) + \kappa(t)h(t)h'(t)$$

$$X(t+1|t) = X(t|t) - \kappa(t)h(t) \left\| \begin{matrix} -\hat{y}(t|t) \\ \bar{z}(t) \end{matrix} \right\|'$$

with

$$h(t) = Z(t|t)C_{\theta}(t|t)\bar{z}(t)$$

$$\hat{y}(t|t) = \hat{\theta}'(t|t)\bar{z}(t).$$

From these formulae the relations to be proved follow immediately after substitution for the posterior statistics where needed. ∎

5. Extensions to incremental and multi-output models

1. The above results can be easily extended to the case of incremental models [7], too. The incremental delta model ($\mu = 1$) needs to substitute increments for all input, output and state variables

$$u(t) \to \Delta u(t) = u(t) - u(t-1),$$

$$y(t) \to \Delta y(t) = y(t) - y(t-1),$$

$$s(t) \to \Delta s(t) = s(t) - s(t-1),$$

while the incremental ARMA model ($\mu = 0$) needs to substitute increments just for the input and output $u(t) \to \Delta u(t)$, $y(t) \to \Delta y(t)$. In both cases, the absolute term $k_x(t)$ disappears.

2. Generalization to models with a ∂u-dimensional input ($\partial u > 1$) is easy when treating particular input items $u_1, \ldots, u_{\partial u}$ as independent variables

$$A(t) \left\| \begin{matrix} y(t) \\ s(t) \end{matrix} \right\| = Hs(t-1) + b^1(t)u_1(t) + \ldots +$$

$$+ b^{\partial u}(t)u_{\partial u}(t) + k_x(t) + ce(t). \tag{5.1}$$

Then the vector of unknown parameters takes the form

$$\theta'(t) = \| a_1(t), \ldots, a_n(t), b_0^1(t), \ldots, b_n^1, \ldots,$$

$$b_0^{\partial u}(t), \ldots, b_n^{\partial u}(t), k_c(t) \| \tag{5.2}$$

the regressor $\bar{z}(t)$ is composed as follows

$$\bar{z}_{k(n+1)}(t) = u_k(t) + X_{1, kn+k-1}(t|t-1) \qquad \text{for} \quad k = 1, \ldots, \partial u$$

$$\bar{z}_i(t) = X_{1, i-1}(t|t-1) \qquad \text{for} \quad i \neq k(n+1) \tag{5.3}$$

and the statistics $X(t|t-1)$ is updated according to the formula

$$X(t|t) = [\mu I_n + \| \tilde{c}(t), I_1^1, \ldots, I_n^{n-1} \|] X(t|t-1) + \| -\tilde{c}(t), I_n \| \times$$

$$\times \| y(t) I_{n+1}, u_1(t) I_{n+1}, \ldots, u_{\partial u}(t) I_{n+1}, I_{n+1}^{n+1} \|. \tag{5.4}$$

We can proceed fully analogously if a measurable external disturbance (possibly multivariate again) is available.

Extension to models with a ∂y-dimensional output ($\partial y > 1$) is less transparent. Peterka has proved in [7] that under rather natural assumptions the joint state and parameter estimation for the multivariate model can be realized by ∂y separate single-output filters when the lower triangular factor of the noise covariance matrix is estimated in addition to θ. This fact enables the user to handle particular outputs

independently and to apply to each of them the results of both Theorems 1 and 2. A detailed description is outside the scope of this paper and the interested reader is referred to [7].

6. Simulation example

1. To check that the derived estimation algorithm is able to track time-varying parameters in closed-loop conditions, we tested the performance of the adaptive controller based on the LQG control synthesis described in [7] and employing the novel parameter tracking algorithm.

The controlled plant was simulated by the ARMA model

$$A(\zeta)y(t)=B(\zeta)u(t)+C(\zeta)e(t).$$

A, B, C denote polynomials in the backward shift operator $\zeta(\zeta x(t)=x(t-1))$. Within the period of 300 samples the polynomials A and B changed abruptly according to the following "time-table"

$$\begin{aligned}
t=1\div 110: \quad & A(\zeta)=1.-1.7\zeta+0.72\zeta^2 \\
& B(\zeta)=0.4\zeta+0.8\zeta^2 \\
t=111\div 220: \quad & A(\zeta)=1.-1.5\zeta+0.7\zeta^2 \\
& B(\zeta)=\zeta+0.5\zeta^2 \\
t=221\div 300: \quad & A(\zeta)=1.-1.5\zeta+0.7\zeta^2 \\
& B(\zeta)=0.2\zeta+0.3\zeta^2.
\end{aligned}$$

The polynomial $C(\zeta)=1.+0.1\zeta-0.2\zeta^2$ did not change during simulation. The variance $\rho=1$ of the noise $e(t)$ was chosen.

We applied the LQG controller based on the second-order incremental delta model and using the receding horizon control strategy with 10 iterations of the Riccati equation at any time instant and with the penalty 1.0 on input increments. The initial value of the parameters θ was selected to ensure the unit static gain of the model. The vector of parameters c was set to $||1, 0.1, 0.9||'$ and the factor $\varphi=0.9$. The set-point signal changed between two levels each 50 samples.

2. The results of the simulation obtained without and with parameter tracking are illustrated in Figs 1 and 2. Experience from a series of other simulation runs has indicated that the suggested parameter tracking algorithm makes it possible to extend significantly the range of conditions under which the adaptive controller is still applicable. It is worth mentioning that the use of different c-parameters or even time-varying c-parameters in the above example has not degraded substantially the control accuracy.

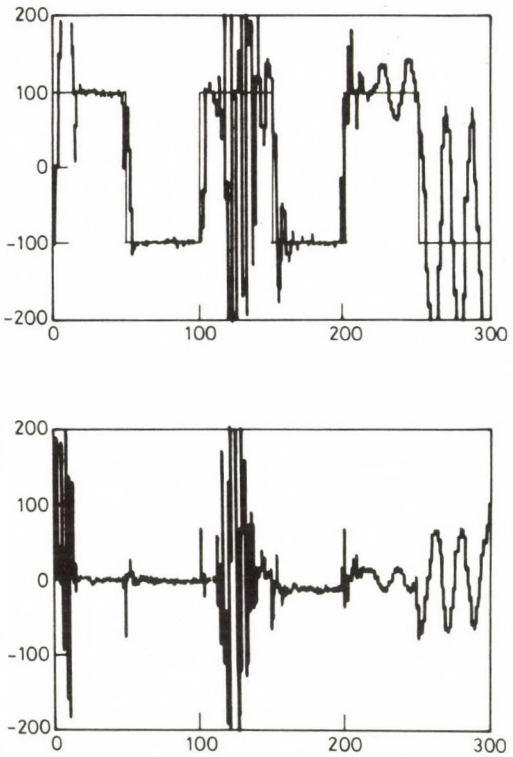


Fig. 1. The controlled output and input action in case that no parameter tracking has been used

7. Remarks to implementation

1. The crucial operations of the joint state and parameter estimation are (4.5), (4.7), (4.23) and (4.25) where a symmetric positive (semi)definite matrix is updated by a matrix dyad of the rank 1. At this step the symmetry and possibly positive (semi)definiteness (when the dyad is subtracted) of the original matrix may be lossed. A reliable measure to ensure the theoretical properties of the covariance matrices is to use the *LD*-factorization of

$$C_\theta(t|t-1) = L_\theta(t|t-1)D_\theta(t|t-1)L'_\theta(t|t-1) \tag{7.1}$$

$$C_{s|\theta}(t|t-1) = L_s(t|t-1)D_s(t|t-1)L'_s(t|t-1) \tag{7.2}$$

and to update the L and D factors directly by a data dyad.

2. The numerical implementation of the recursive formulae of Theorem 1 has been discussed in detail in [7]. The algorithms CGEN, LDFIL, and CFIL have been suggested there for solving the steps (a), (b), and (c), respectively.

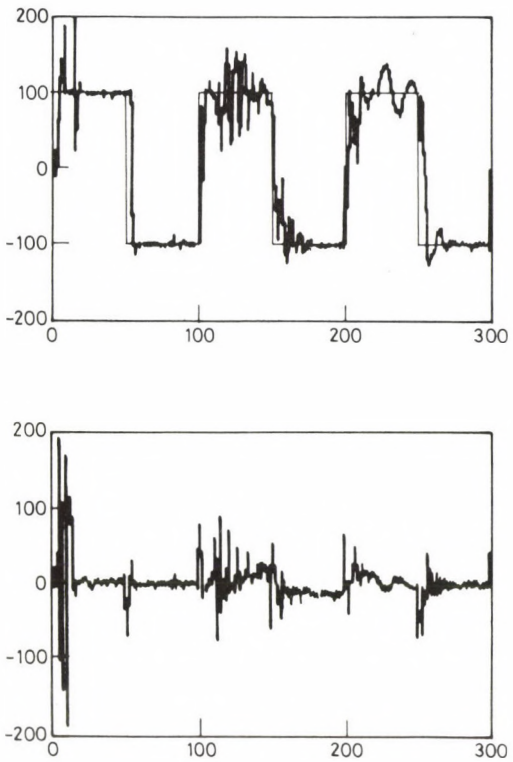


Fig. 2. The controlled output and input action when the novel parameter tracking algorithm has been used ($\varphi = 0.9$)

Provided that the matrices $C_{s|\theta}(t|t)$ and $C_\theta(t|t)$ are in the *LD*-factorized form of (7.2) and (7.1) updated by the algorithms CGEN and LDFIL, steps (4.23) and (4.25) may be performed analogously by applying the algorithm of dyadic reduction DYDR published in [7], too.

3. As the measurement update (4.6)–(4.7) and time update (4.24)–(4.25) are of similar structure, it is advantageous to combine them as follows

$$\hat{\theta}(t|t) = \hat{\theta}(t|t-1) + \frac{C(t|t-1)\bar{z}(t)\hat{\varepsilon}(t|t-1)}{d_y(t) + \zeta(t|t-1)} \tag{7.3}$$

$$C_\theta^{-1}(t|t) = C_\theta^{-1}(t|t-1) + \left[\frac{\varphi}{d_y(t)} - \frac{1-\varphi}{\zeta(t|t-1)}\right]\frac{\bar{z}(t)\bar{z}'(t)}{d_y(t)} \tag{7.4}$$

to minimize the number of the necessary operations. The updating (7.4) can be performed again by applying the dyadic reduction to the *LD*-factorized matrix $C_\theta(t|t)$.

In the Appendix we append the FORTRAN IV subroutine REFIT which carries out the recursions (7.3) and (7.4) in a compact way.

Remark 1. The presented procedure is slightly more general than needed for our task. Statistics $DX(1)$ and DF, internally updated in REFIT, make it possible to return through R the estimate of the variance ρ. The theoretical background for its estimation will be discussed elsewhere.

Remark 2. Subroutine REFIT can be immediately used also for parameter tracking of regression-type models. In such a case the formal parameter D of REFIT is to contain the vector of nonfiltered data and DY is to be equal to 1 for any t.

8. Conclusions

We have achieved the tracking of time-varying parameters $a(t)$, $b(t)$ and $k_x(t)$ of the delta/ARMA model in a consistent Bayesian way by properly modelling changes of the parameters in time. The model has been designed with the aim to increase uncertainty of only evidentially changing parameters. Thus, parameters have been supposed to vary in the direction of the last change of their mean value. In such a way the resulting algorithm has become highly robust with respect to poor system. excitation.

This idea is not new. It forms the basis of the previously suggested technique of restricted exponential (directional) forgetting [4]. An alternative model-based formulation used in this paper may serve as an illustration of a close relationship between the explicit modelling and implicit forgetting.

The novel feature lies in using the concept of a rationally based restriction of parameter variations to solve the joint state and parameter estimation. The results summarized in Theorem 2 demonstrate the link between parameter variations and state uncertainty — owing to the assumed model of parameter variations the increase of the marginal covariance of unknown parameters is inevitably accompanied with the increase of the conditional state covariance.

Recent results indicate that it will be possible to extend the implicit "forgetting" approach used in [4] for parameter tracking to the case of joint state and parameter estimation, too. Within this framework a sensible algorithm for tracking the noise variance $p(t)$ can be developed similarly as in [4] and additional information about parameter variations can be incorporated in the vein of [3].

Appendix. FORTRAN subroutine REFIT

When calling REFIT, the subroutine parameters have the following meaning: LX = an ML-dimensional vector ($ML = N(N+1)/2$) containing in the first N entries the parameter "estimates" $\hat{\theta}(t|t-1)$ and in the remaining entries the strictly lower triangular part of the L-factor of the covariance matrix $C_\theta(t|t-1)$ — stored column-by-column with the negative sign; DX = an MD-dimensional ($MD = N+1$) vector containing in the first entry the sum of the prediction error squares and in the remaining entries the diagonal of the D-factor of the covariance matrix $C_\theta(t|t-1)$; D = the vector of filtered data $\|\bar{y}(t), \bar{z}'(t)\|'$; DY = the scalar $d_y(t)$; DF = the number of observations (weighted by φ); FI = the forgetting factor φ; $MZRO$ = "machine zero" to test a nearly zero $\zeta(t|t-1)$ (estimation of θ is suppressed for $\zeta(t|t-1) \leqq MZRO$). Only the parameters LX, DX, DF are updated by the subroutine.

The remaining parameters get the meaning after exit: DZETA = $\zeta(t|t-1)$; $EP = \hat{\varepsilon}(t|t-1)$; E = an estimate of the noise variance ρ; F, G, H are auxiliary vectors, G contains the vector $C_\theta(t|t-1)\bar{z}(t)$ starting from the 2nd entry.

```
      SUBROUTINE REFIT (LX,DX,D,ML,MD,DY,DF,
     —     DZETA,MZRO,EP,R,FI,F,G,H)
      INTEGER ML,MD
      REAL LX(ML),DX(MD),D(MD),DY,DF,MZRO,EP,
     —     FI,F(MD),G(MD),H(MD)
      DF = DF + 1.
      KD = MD
      KL = ML
      GK = DX(KD) * D(KD)
      G(KD) = GK
      DELTA = GK * D(KD)
      H(KD) = DELTA
    1 KD = KD-1
      EP = D(KD)
      K = MD
    2 EP = EP-LX(KL) * D(K)
      KL = KL-1
      K = K-1
      IF(K-KD)3,3,2
    3 F(KD) = EP
      GK = EP * DX(KD)
      G(KD) = GK
      ETA = EP * GK
      DELTA = DELTA + ETA
      H(KD) = DELTA
      IF(KL)4,4,1
    4 DZETA = H(2)
      DZETA1 = DZETA + DY
      DELTA1 = DELTA + DY
      IF(DZETA-DZETA0)10,10,5
    5 EPS = FI/DY-(1.-FI)/DZETA
      KD = MD
      KL = ML
      A = 1. + EPS * H(KD)
      DX(KD) = DX(KD)/A
    6 KD = KD-1
      IF(KD-1)11,7,8
```

```
 7    EPS=1.
      A=DZETA1
 8    C=DX(KD)*A
      B=EPS*F(KD)/A
      A=1.+EPS*H(KD)
      DX(KD)=C/A
      GK=G(KD)
      K=MD
 9    C=LX(KL)
      LX(KL)=C+B*G(K)
      IF(KD.GT.1)G(K)=G(K)-C*GK
      KL=KL-1
      K=K-1
      IF(K-KD)6,6,9
10    DX(1)=DX(1)*DZETA1/DELTA1
11    DX(1)=DX(1)/FI
      DF=FI*DF
      R=1./((DF-2.)*DX(1))
      RETURN
      END
```

Acknowledgement

The authors would like to thank V. Peterka for stimulating and helpful discussions.

References

1. *Bohlin, T.*, Four Cases of Identification of Changing Systems. In R. K. Mehra and D. G. Lainiotis (eds.), System Identification: Advances and Case Studies. Academic Press, New York, 1976.
2. *Goodwin, G. G.*, Some observations on robust estimation and control. In Preprints of the 7th IFAC/IFORS Symposium on Identification and System Parameter Estimation, York, Vol. **1**, pp. 851–858, 1985.
3. *Kulhavý, R.*, Directional tracking of regression-type model parameters. In Preprints of the 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing, Lund, pp. 97–102, 1986.
4. *Kulhavý, R.*, Restricted exponential forgetting in real time identification. Automatica, **23**, pp. 589–600, 1987.
5. *Ljung, L.*, System Identification — Theory for the User. Prentice-Hall, 1987.
6. *Peterka, V.*, Bayesian approach to system identification. In P. Eykhoff (ed.), Trends and Progress in System Identification, Chap. 8, pp. 239–304, Pergamon Press, Oxford, 1981.
7. *Peterka, V.*, Control of uncertain processes: applied theory and algorithms. Supplement to the journal Kybernetika, **22**, no. *3*, *4*, *5*, *6*, 1986.
8. *Peterka, V.*, Algorithms for *LQG* self-tuning control based on input-output Delta models. In Preprints of the 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing, Lund, pp. 13–18, 1986.
9. *Rao, C. R.*, Linear Statistical Inference and Its Applications. Wiley, New York, 1973.

Слежение за временно-переменными параметрами в дельта-моделях

Р. КУЛГАВЫ, Э. КЛЁКИС
(Прага, Каунас)

Описание входно-выходного поведения систем с помощью разностных операторов, известное также под названием дельта-модель, появилось чтобы добиться высшей численной робастности, прежде всего, в случае высокочастотной дискретизации. В статье рассматривается обобщение последних достижений касающихся оценки состояния и параметров дельта-модели на случай временно-переменных параметров. Слежение за параметрами достигается внедрением простой однофакторной модели параметрических изменений типа случайной прогулки. Модель выбрана так, чтобы забывалась только информация, модифицированная новейшими данными. Таким образом, обеспечивается надежное оценивание параметров даже в случае, когда идентифицируемая система плохо возбуждена.

R. Kulhavý
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Prague 8
Czechoslovakia

Problems of Control and Information Theory, Vol. 18 (2), pp. 125—133 (1989)

ON THE ZERO ERROR FEEDBACK CAPACITY REGION OF THE BINARY ADDER CHANNEL

A. YA. BELOKOPYTOV

(*Moscow*)

(Received November 19, 1987)

The set of all achievable zero error rate triplets for the binary adder channel with feedback is considered. It is shown that the classical capacity region is larger than zero error one.

1. Introduction

Let the message sources 0, 1 and 2 produce random integers $W_0 \in \{1, 2, \ldots, M_0\}$, $W_1 \in \{1, 2, \ldots, M_1\}$ and $W_2 \in \{1, 2, \ldots, M_2\}$ in every N seconds, respectively, so that each triplet (W_0, W_1, W_2) occurs with probability $1/M_0 M_1 M_2$.

The two-user multiple-access channel (MAC) is that which connects two senders with a single receiver. The channel is defined by the input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, output alphabet $\mathcal{Y}$, and the set of transition probabilities $p^0(y|x_1, x_2)$.

The messages from sources 0 and 1 are inputs to the first encoder and those from sources 0 and 2 are inputs to the second one (source 0 is called common information source). It is said that the communication via MAC uses feedback if both encoders know every output signal of the channel.

Let x_t^1, x_t^2 be the signals of the encoders at moment t ($t=1, \ldots, N$) and let y_t be the output at the very instant (coders produce one symbol per second and signal transmission and feedback are instantaneous).

An (M_0, M_1, M_2, N) zero error code for the MAC with feedback is given by a collection of encoding functions f_t^s:

$$x_1^s = f_1^s(W_0, W_s), \qquad x_t^s = f_t^s(W_0, W_s, y_1, \ldots, y_{t-1}) \qquad (t=2, \ldots, N, s=1, 2)$$

and by a decoding function g: $g(y_1, \ldots, y_N) = (W_0^*, W_1^*, W_2^*)$ such that for every triplet (W_0, W_1, W_2):

$$p\{(W_0^*, W_1^*, W_2^*) = (W_0, W_1, W_2) | (W_0, W_1, W_2)\} = 1.$$

A rate triplet (R_0, R_1, R_2) is said to be achievable if for every $\varepsilon > 0$ there exist an (M_0, M_1, M_2, N) code satisfying

$$R_i - \varepsilon \leqq \frac{1}{N} \log_2 M_i \qquad (i=0, 1, 2).$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Dueck [1] considered the class C of MACs for which it was possible to find two functions α: $\mathcal{X}_1 \times \mathcal{Y} \to \mathcal{X}_2$ and β: $\mathcal{X}_2 \times \mathcal{Y} \to \mathcal{X}_1$ such that $p^0(y|x_1, x_2)=0$ if $x_2 \neq \alpha(x_1, y)$ or $x_1 \neq \beta(x_2, y)$, i.e. both of the inputs were functions of the output and the other input. For MACs belonging to C he proved that if $R_{0f} \neq \{(0, 0, 0)\}$ (here R_{0f} denotes the set of all achievable rate triplets) then it is given by the set of all triplets (R_0, R_1, R_2) satisfying:

$$0 \leqq R_1 \leqq H(X_1|U), \tag{1}$$

$$0 \leqq R_2 \leqq H(X_2|U), \tag{2}$$

$$0 \leqq R_0 \leqq \min_{*} \{I(U; Y^*) - H(X_1^*, X_2^*|U, Y^*)\} \tag{3}$$

for some joint distribution $p(u, x_1, x_2, y)$ such that

$$p(u, x_1, x_2, y) = p(u)p(x_1|u)p(x_2|u)p^0(y|x_1, x_2)$$

where $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, $y \in \mathcal{Y}$, $u \in \mathcal{U}$ ($\mathcal{U}$ is a finite set: $|\mathcal{U}| \leqq |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2$) $p^0(y|x_1, x_2)$ are the transition probabilities of the channel, and the minimum in (3) is taken over all distributions of random variables X_1^*, X_2^*, Y^* on $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$ such that $P_{X_1^*|U} = P_{X_1|U}$, $P_{X_2^*|U} = P_{X_2|U}$, $P_{Y^*|X_1^*, X_2^*}(y|x_1, x_2)=0$ whenever $p^0(y|x_1, x_2)=0$. In the present paper we confine ourselves to the binary adder channel ($\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$, $\mathcal{Y} = \{0, 1, 2\}$ and $p^0(2|1, 1) = p^0(1|1, 0) = p^0(1|0, 1) = p^0(0|0, 0) = 1$).

The methods of constructing the zero-error codes for this channel with feedback were discussed in [2] and [3] (Fig. 1 establishes corresponding results ($R_0 = 0$); $H = 0.347$, $A_1 = 0.717$, $A_2 = 0.7532$). However, no upper bound, except of the classical feedback capacity region has been obtained so far.

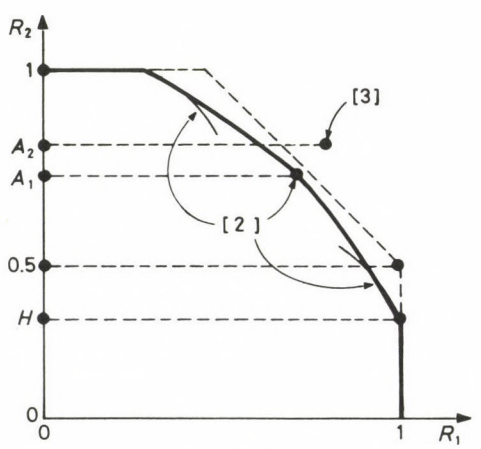


Fig. 1

2. Main results

Let R_{0f}^{0} and R_{0f}^{*} denote the intersections of R_{0f} with planes $R_0=0$ and $R_1=R_2$, respectively. Define, then, $R=R_1+R_2$ and introduce

$$H(x_1, \ldots, x_n)=-\sum_{i=1}^{n} x_i \log_2 x_i, \qquad h(x)=H(x, 1-x),$$

$$L(x)=H\left(\frac{1-x}{2}, x, \frac{1-x}{2}\right)=h(x)+1-x \qquad \left(x_i \geqq 0, \sum_{i=1}^{n} x_i=1, 0 \leqq x \leqq 1\right).$$

Theorem 1. Let $p_1, p_2 \in [0, 1/2]$ satisfy the equations

$$L(p_1+p_2-2x)=H(x, p_1-x, p_2-x, 1-p_1-p_2+x),$$

where x is the unique root of the equation

$$4x(1+x-p_1-p_2)(p_1+p_2-2x)^2=(p_1-x)(p_2-x)(1-p_1-p_2+2x)^2,$$

belonging to $[0, \min(p_1, p_2)]$.

Then the rate pair $(R_1, R_2)=(h(p_1), h(p_2))$ belongs to R_{0f}^{0}.

(See Fig. 3, $E=0.46015$, $F_0=0.78974$.)

Theorem 2. The set of points (R, R_0) established as follows:

$$\left\{R=2h(p), R_0=L(2p-2x)-H(x, p-x, p-x, 1+x-2p), x=\frac{2\sqrt{3}-3}{6}(1-2p),\right.$$

$$\left.p \in \left[\frac{3-\sqrt{3}}{6}, 0.23684\right]\right\} \cup \left\{R=x, R_0=\log_2 3-x, x \in \left[0, 2h\left(\frac{3-\sqrt{3}}{6}\right)\right]\right\}$$

belongs to R_{0f}^{*}.

(See Fig. 2, $B=1.48801$, $C_0=1.57958$, $C=1.58226$, $D=1.58496$.)

The corresponding bounds for classical (nonzero) capacity region are shown in Figs 3 and 2 to compare them with those from Theorems 1 and 2, respectively ($F=0.79113$).

We suspect that the sets of points specified in Theorems 1 and 2 are indeed the bounds of R_{0f}^{0} and R_{0f}^{*}, respectively, but are not able to prove this conjecture.

Theorem 3. If the rate pair $(1, R_2)$ belongs to R_{0f}^{0} then $R_2<1/2$.

Remark 2. The method suggested in Section 4 helped the author to find four decimal digits for the maximum value of R_2 from Theorem 3: $R_2=0.4601\ldots$

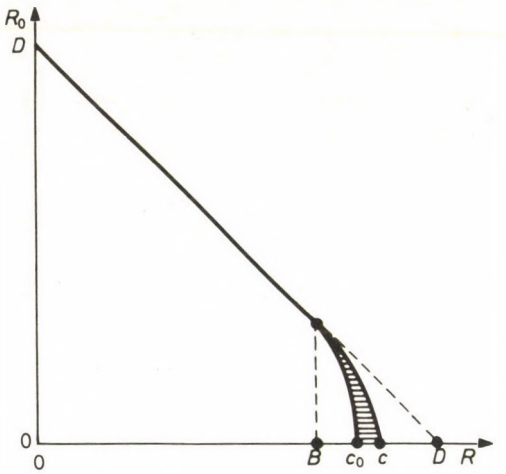


Fig. 2

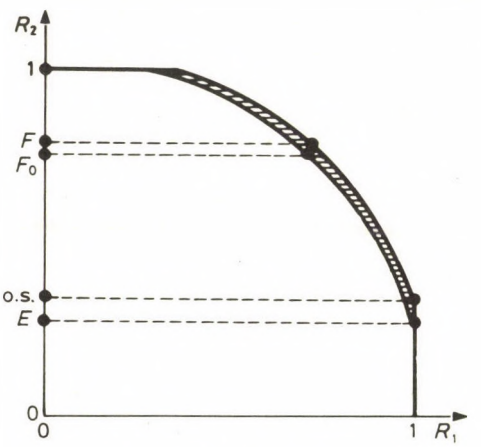


Fig. 3

3. Proof of Theorems 1 and 2

Let $\mathcal{U} = \{1, \ldots, m\}$ where $m = |\mathcal{U}|$, and assume that $p(i) = \alpha_i$ and $p(X_s = 1 | U = i) = p_s^i$ $(i = 1, \ldots, m, s = 1, 2)$. Then to define random variables X_1^*, X_2^* and Y^* it suffices to set $p(X_1^* = 1, X_2^* = 1 | U = i) = q^i$ where $q^i \in D_i = [\max(0, p_1^i + p_2^i - 1), \min(p_1^i, p_2^i)]$ (note that $p(X_1 = 1, X_2 = 1 | U = i) = p_1^i p_2^i$). Conditions (1)–(3) can now be rewritten in

this way

$$0 \leqq R_1 \leqq \sum_{i=1}^{m} \alpha_i h(p_1^i), \tag{4}$$

$$0 \leqq R_2 \leqq \sum_{i=1}^{m} \alpha_i h(p_2^i), \tag{5}$$

$$0 \leqq R_0 \leqq \min_{*} \{H(Y^*) - H(X_1^*, X_2^* | U)\} =$$

$$= \min_{\vec{q}} \left\{ H\left(\sum_{i=1}^{m} \alpha_i q^i, \sum_{i=1}^{m} \alpha_i (p_1^i + p_2^i - 2q^i), \sum_{i=1}^{m} \alpha_i (1 - p_1^i - p_2^i + q^i) \right) - \right.$$

$$\left. - \sum_{i=1}^{m} \alpha_i H(q^i, p_1^i - q^i, p_2^i - q^i, 1 - p_1^i - p_2^i + q^i) \right\} \leqq$$

$$\leqq \min_{\vec{q}} \left\{ L\left(\sum_{i=1}^{m} \alpha_i (p_1^i + p_2^i - 2q^i) \right) - \right.$$

$$\left. - \sum_{i=1}^{m} \alpha_i H(q^i, p_1^i - q^i, p_2^i - q^i, 1 - p_1^i - p_2^i + q_i) \right\} \tag{6}$$

where the last inequality was derived from

$$H(a, b, c) \leqq H\left(\frac{a+c}{2}, b, \frac{a+c}{2} \right) = L(b) \qquad (a + b + c = 1).$$

Consider, then, the set $\mathcal{U}^0 = \{1, \ldots, m, m+1, \ldots, 2m\}$. Define

$$p(i+m) = p(i) = \alpha_i/2, \quad p(X_s = 1 | U = i) = p_s^i, \quad p(X_s = 1 | U = i+m) = 1 - p_s^i,$$

$$p(X_1^* = 1, \quad X_2^* = 1 | U = i) = q_1^i, \quad p(X_1^* = 0, X_2^* = 0 | U = i+m) = q_2^i,$$

$$q_1^i, q_2^i \in D_i \qquad (i = 1, \ldots, m, \ s = 1, 2). \tag{7}$$

Note that the substitution the distributions on $\mathcal{U}^0 \times \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$ for those on $u \times x_1 \times x_2 \times y$ according to (7) does not change conditions (4) and (5), and at the same time, by virtue of convexity of $H(x, p_1^i - x, p_2^i - x, 1 - p_1^i - p_2^i + x)$ on p_i, inequality (6) can be transformed into

$$0 \leqq R_0 \leqq \min_{\vec{x}} \left\{ L\left(\sum_{i=1}^{m} \alpha_i (p_1^i + p_2^i - 2x_i) \right) - \right.$$

$$\left. - \sum_{i=1}^{m} \alpha_i H(x_i, p_1^i - x_i, p_2^i - x_i, 1 - p_1^i - p_2^i + x_i) \right\} \left(x_i = \frac{q_1^i + q_2^i}{2} \right).$$

Therefore, it can be assumed that distributions of the random variables X_1, X_2, U, X_1^*, X_2^* are given by (7) and $q_1^i = q_2^i = x_i$. Setting $m=1$ it follows that R_{0f} contains all triplets (R_0, R_1, R_2) such that $R_1 = h(p_1)$, $R_2 = h(p_2)$,

$$0 \leqq R_0 = \min_x \{L(p_1+p_2-2x) - H(x, p_1-x, p_2-x, 1-p_1-p_2+x)\}$$

$$(x \in [0, \min(p_1, p_2)]) \tag{8}$$

for some $p_1, p_2 \in [0, 1/2]$.

The function in brackets is not only convex down of x on $[0, \min(p_1, p_2)]$, it has strictly one stationary point on the segment (consider its derivative in 0 and $\min(p_1, p_2)$), and if $p_1 = p_2 = p$ then the equation from Theorem 1 has an explicit solution $x = (2\sqrt{3}-3)(1-2p)/6$.

Thus, Theorems 1 and 2 are proved.

4. Investigation of the bound of R_{0f}^0

From the above reasoning it follows that all points (R_1, R_2) of the bound of R_{0f}^0 satisfy

$$R_1 = \sum_{i=1}^{m} \alpha_i h(p_1^i), \qquad R_2 = \sum_{i=1}^{m} \alpha_i h(p_2^i)$$

where

$$0 \leqq \min_{\vec{x}} \{f(\vec{x})\} = \min_{\vec{x}} \left\{ L\left(\sum_{i=1}^{m} \alpha_i (p_1^i + p_2^i - 2x_i) \right) - \right.$$

$$\left. - \sum_{i=1}^{m} \alpha_i H(x_i, p_1^i - x_i, p_2^i - x_i, 1 - p_1^i - p_2^i + x_i) \right\} \tag{9}$$

and

$$\alpha_i, p_1^i, p_2^i \in [0, 1], \quad x_i \in D_i = [\max(0, p_1^i + p_2^i - 1), \quad \min(p_1^i, p_2^i)], \quad \sum_{i=1}^{m} \alpha_i = 1.$$

Let K denote the subset of all $i \in \{1, \ldots, m\}$ such that D_i consists of more than one point. Then for any $i \in K$ we can write:

$$\frac{\partial f}{\partial x_i} = \frac{\alpha_i}{\ln 2} \left(-2 \ln \frac{1-F}{2F} + \ln \frac{x_i(1+x_i-p_1^i-p_2^i)}{(p_2^i - x_i)(p_1^i - x_i)} \right), \tag{10}$$

where $F = \sum_{j=1}^{m} \alpha_j (p_2^j + p_1^j - 2x_j)$.

Let there be some $i \in K$ that $x_1^0 = \max(0,\ p_1^i + p_2^i - 1)$ (here x_i^0 is the i-th coordinate of $\vec{x}^0 = (x_1^0, \ldots, x_m^0)$ that minimizes (9)). If $x_i \to x_i^0 +$ then $\dfrac{\partial f}{\partial x_i} \geqq 0$, and it follows from (10) that if $x_i \to x_i^0 +$, then $F \to 1$ ($L(F) \to 0$). Because of (9), this yields $H(x_j^0,\ p_1^j - x_j^0,\ p_2^j - x_j^0,\ 1 - p_1^j p_2^j + x_j^0) = 0$ for any $j \in \{1, \ldots, m\}$, and then $p_1^j, p_2^j \in \{0, 1\}$ i.e. the expressions for R_1 and R_2 are simply zeros. Let, then, for some $i \in K$, $x_i^0 = \min(p_1^i,\ p_2^i)$. Analogously, if $x_i \to x_i^0 -$, then $\dfrac{\partial f}{\partial x_i} \leqq 0$ and $F \to 0$, i.e. for any $j \in \{1, \ldots, m\}$ $(p_1^j + p_2^j)/2 = x_j^0 \leqq \min(p_1^j, p_2^j)$ and then $x_j^0 = p_1^j = p_2^j = p_j$. Considering (9) with x_i replaced by $p_1^i p_2^i$ $(i = 1, \ldots, m)$ we obtain:

$$L(0) - \sum_{i=1}^{m} \alpha_i h(p_i) \leqq L\left(\sum_{i=1}^{m} \alpha_i (2p_i - 2p_i^2) \right) - 2 \sum_{i=1}^{m} \alpha_i h(p_i)$$

that yields

$$R_1 = \sum_{i=1}^{m} \alpha_i h(p_i) = R_2 \leqq L\left(\sum_{i=1}^{m} \alpha_i (2p_i - 2p_i^2) \right) - 1 \leqq \log_2 3 - 1 < F_0.$$

The above reasoning allows us to assume p_1^j, p_2^j and α_j $(j = 1, \ldots, m)$ be chosen in such a way that for any $i \in K$, x_i^0 is an interior point of D_i.

Consider $R_1 = 1$. It implies that for all $i \in \{1, \ldots, m\}$ $p_1^i = 1/2$.

Let $\vec{x}^* = \left(\dfrac{p_2^1}{2}, \ldots, \dfrac{p_2^m}{2} \right)$, then $f(\vec{x}^*) = L(1/2) - 1 - \sum_{i=1}^{m} \alpha_i h(p_2^i) = 1/2 - R_2$, and for any $i \in K \neq \emptyset$

$$\left. \frac{\partial f}{\partial x_i} \right|_{\vec{x} = \vec{x}^*} = \alpha_i \left(-2 \log_2 \frac{1 - 1/2}{1} + \log_2 1 \right) = 2\alpha_i > 0.$$

It follows that

$$0 \leqq \min_{\vec{x}} \{ f(\vec{x}) \} < f(\vec{x}^*) = \frac{1}{2} - R_2.$$

Hence $R_2 < 1/2$, and Theorem 3 is proved.

It is convenient to introduce $\psi = \left(\dfrac{1 - F}{2F} \right)^2$. Then, according to (10), it follows that for every $i \in \{1, \ldots, m\}$

$$x_i^0 (1/2 + x_i^0 - p_2^i) = \psi (p_2^i - x_i^0)(1/2 - x_i^0). \tag{11}$$

Let x_ψ^i be the unique root of (11) belonging to D_i ($\psi \geqq 0$, $i = 1, \ldots, m$). Inequality (9) can now be rewritten in this way

$$0 \leqq \min_{\psi \geqq 0} \left\{ L\left(\sum_{i=1}^{m} \alpha_i (1/2 + p_2^i - 2x_\psi^i) \right) - \right.$$

$$- \sum_{i=1}^{m} \alpha_i H\left(x_\psi^i, p_2^i - x_\psi^i, \frac{1}{2} - x_\psi^i, \frac{1}{2} + x_\psi^i - p_2^i\right)\Bigg\}. \tag{12}$$

Introduce then

$$p_i = p_2^i, \; l_\psi(p_i) = H\left(x_\psi^i, p_i - x_\psi^i, \frac{1}{2} - x_\psi^i, \frac{1}{2} + x_\psi^i - p_i\right),$$

$$g_\psi^i = g_\psi(p_i) = p_i + \frac{1}{2} - 2x_\psi^i = (1 - \sqrt{\psi + (\psi - 1)^2 (p_i - 1/2)^2})/(1 - \psi) \qquad (g_1(p) \equiv 1/2).$$

Note that

$$g_\psi^i = g_\psi(p_i) = g_\psi(1 - p_i), \qquad l_\psi(p_i) = l_\psi(1 - p_i), \qquad h(p_i) = h(1 - p_i),$$

$l_\psi(p_i) = l_{1/\psi}(p_i)$, $g_\psi^i + g_{1/\psi}^i = 1$ and if $F \in [1/2, 1]$ (i.e. $\psi \in [0, 1]$) it yields

$$L\left(\sum_{i=1}^{m} \alpha_i g_\psi^i\right) = L(F) \leqq L(1 - F) = L\left(\sum_{i=1}^{m} \alpha_i g_{1/\psi}^i\right).$$

Observe now that the desired R_2 is equal to $\max\limits_{\psi \in [0, 1]} R(\psi)$ where $R(\psi)$ is the solution of the following problem

$$R(\psi) = \max \sum_{i=1}^{m} \alpha_i h(p_i) \tag{13}$$

where $p_i \in [0, 1/2]$, $\alpha_i \geqq 0 \left(\sum\limits_{i=1}^{m} \alpha_i = 1\right)$ satisfy the constraints

$$\sum_{i=1}^{m} \alpha_i g_\psi(p_i) = F = \frac{1}{1 + 2\sqrt{\psi}}, \tag{14}$$

$$\sum_{i=1}^{m} \alpha_i l_\psi(p_i) \leqq L(F) = L\left(\frac{1}{1 + 2\sqrt{\psi}}\right). \tag{15}$$

Let then $R(\psi) = 0$ if there are no $\vec{p}$ and $\vec{\alpha}$ to satisfy (14) and (15). It could be easily shown that $R(\psi) > 0$ only if $\psi \in I = (0.11 \ldots, 0.25)$. By virtue of convexity of $h(p(g_\psi))$ for every $\psi \in I$, problem (13)–(15) can be reduced to that with $m = 1$ or to problems (13), (14), (16), where

$$\sum_{i=1}^{m} \alpha_i l_\psi(p_i) = L\left(\frac{1}{1 + 2\sqrt{\psi}}\right). \tag{16}$$

Fix $\vec{p} = (p_1, \ldots, p_m)$ and consider (13)–(16) as a problem of linear programming, so it can be assumed that for every $k \in \{u, \ldots, m\}$, $\alpha_k = 0$ and α_1, α_2, α_3 satisfy

$$\alpha_1+\alpha_2+\alpha_3=1$$

$$\alpha_1 g_\psi(p_1)+\alpha_2 g_\psi(p_2)+\alpha_3 g_\psi(p_3)=\frac{1}{1+2\sqrt{\psi}},$$

$$\alpha_1 l_\psi(p_1)+\alpha_2 l_\psi(p_2)+\alpha_3 l_\psi(p_3)=L\left(\frac{1}{1+2\sqrt{\psi}}\right) \qquad (0\leqq p_1<p_2<p_3\leqq 1/2).$$

Using numerical methods of finding the maximum of functions of 3 variables, it can be shown that $R(\psi)\leqq 0.4602$ where

$$\arg\max_{\psi}\{R(\psi)\}=0.1429\ldots$$

The author is grateful to L. Bassalygo and V. Rykov for fruitful discussions and also to A. Djachkov for supervision.

References

1. *Dueck, G.*, The zero error feedback capacity region of a certain class of multiple-access channels, Problems of Control and Information Theory, **14,** 1985*(2)*, pp. 89–103.
2. *Zhang, Z., Berger, T., Massey, J. L.*, Some families of zero-error block codes for the two-user binary adder channel with feedback, IEEE Trans. on Inform. Theory, 1987, **33,** *5*, pp. 613–619.
3. *Belokopytov, A. Ya., Luzgin, V. N.*, Block information transmission via the binary adder channel with feedback, Problems of information transmission, 1987, **23,** *4*, pp. 114–118 (in Russian).

Замечание о пропускной способности с нулевой вероятностью ошибки двоичного суммирующего канала с обратной связью

А. Я. БЕЛОКОПЫТОВ

(Москва)

В настоящей статье рассматривается задача нахождения области пропускной способности двоичного суммирующего канала с обратной связью при наличии источника общей информации. Получены аналитические описания подмножеств характерных сечений множества достижимых троек скоростей (R_0, R_1, R_2). Определены численные значения координат экстремальных точек полученных областей.

На основании результатов Г. Дьюка о пропускной способности с нулевой ошибкой класса каналов с обратной связью для двоичного суммирующего канала доказана достижимость следующих троек скоростей: (0, 0.78974, 0.78974), (0, 1, 0.46015) и (0.097, 0.744, 0.744).

Вместе с тем, удалось показать, что, если $R_1=1$, то $R_2\leqq 0.4602$. Таким образом, двоичный суммирующий канал, будучи детерминированным $(p(y|x_1, x_2)\in\{0, 1\})$, обладает различной пропускной способностью при передачах с нулевой и стремящейся к нулю ошибкой. Заметим, что для обычных детерминированных каналов $(p(y|x)\in\{0, 1\})$ это невозможно.

А. Я. Белокопытов
Московский государственный университет им. М. В. Ломоносова,
мех-мат факультет, кафедра теории вероятностей,
Ленинские горы, СССР

РУССКИЙ ПЕРЕВОД

Проблемы управления и теории информации, том **18,** *номер 2 (1989)*

ОЦЕНИВАНИЕ НЕЛИНЕЙНЫХ ФУНКЦИОНАЛОВ ОТ РЕГРЕССИИ ПРИ ВОЗМОЖНОСТИ ПЛАНИРОВАНИЯ

Ю. И. Пастухова, Р. З. Хасьминский

(*Москва*)

Получены асимптотически минимаксные нижние границы среднеквадратического риска оценок дифференцируемого функционала от регрессии при возможности планирования эксперимента. В случае, когда известна лишь условная дисперсия шумов, построен план наблюдений и оценка, для которой эти границы асимптотически достигаются при выполнении некоторых условий на гладкость функционала и неизвестную функцию регрессии.

1. Постановка задачи

Пусть $P(\cdot\,|\,t)$, $t\in[0,\,1]$ — семейство условных распределений на прямой R^1, которое в дальнейшем считается заданным, но неизвестным статистику. Предположим, что в произвольных точках $t_1,\ \ldots,\ t_n$ отрезка $[0,\,1]$ можно производить наблюдения $X_1(t_1),\ \ldots,\ X_n(t_n)$, которые условно независимы при фиксированном плане $t^{(n)}=(t_1,\ \ldots,\ t_n)$, причем условное распределение $X_i(t_i)-E(X_i(t_i)|t_i)$ при условии $t_i=t$ совпадает с $P(\cdot\,|\,t)$.

Результат наблюдения X_i в точке t_i можно записать в форме

$$X_i=R(t_i)+\xi_i(t_i),\qquad R(t)=E\{X_i(t_i)|t_i=t\},$$

где $E\{\xi_i(t_i)|t_i\}=0$ и $\xi_i(t_i)$ условно независимы при заданном плане $t^{(n)}$.

Решается задача оценивания некоторого функционала $F(R)$ от функции регрессии $R(t)$, определенного в пространстве $L_2[0,\,1]=L_2$ и построения такого допустимого в смысле [1] плана $t^{(n)}$, что величина $\Delta_n=E\{\hat F_n-F(R)\}^2$ в асимптотически минимаксном смысле оказывается наименьшей ($\hat F_n$ — оценка, основанная на $t_1,\ X_1,\ \ldots,\ t_n,\ X_n$).

Как и в [1], будут рассмотрены два случая.

1. Известна условная дисперсия наблюдений

$$\sigma^2(t)=E\{\xi_i^2(t_i)|t_i=t\},\quad \sigma(t)\in L_2, \tag{1.1}$$

$$0<\inf_{[0,1]}\sigma(t)\leqq\sup_{[0,1]}\sigma(t)<\infty.$$

$\sigma\in C[0,1]$ — пространству непрерывных на $[0,1]$ функций. (1.1′)

2. Известно распределение $\xi_i(t_i)$ при условии $t_i=t$. Его плотность $p(x|t)$ относительно меры Лебега в R^1 абсолютно непрерывна по x, имеет конечное информационное количество Фишера

$$I(t)=\int_{R^1}\frac{(p'_x(x|t))^2}{p(x|t)}\,dx$$

и удовлетворяет следующим условиям регулярности

$$0<\inf_{[0,1]}I(t)\leqq\sup_{[0,1]}I(t)<\infty;\quad \sup_{[0,1]}\int|x|p(x|t)dx<\infty; \tag{1.2}$$

$$\sup_{[0,1]}\int\left|\frac{p'_x(x+s|t)}{p^{1/2}(x+s|t)}-\frac{p'_x(x|t)}{p^{1/2}(x|t)}\right|^2dx\to 0\quad(s\to 0). \tag{1.3}$$

2. Границы снизу

Установим нижние границы качества произвольной оценки $\tilde{F}_n$ функционала $F(R)$ при любом допустимом плане $t^{(n)}$.

Напомним, что функционал $F(R)$ дифференцируем по Фреше в L_2 в точке R_0, если

$$F(R_0+h)-F(R_0)=\int_0^1 F'(R_0,t)h(t)dt+o(\|h\|)$$

$(\|h\|)\to 0)$.

Теорема 2.1. Пусть выполнены условия (1.2)–(1.3), функционал $F(R)$ дифференцируем по Фреше на некотором компакте $\mathscr{P}\subset L_2$ и производная $F'(R,t)$ в L_2 удовлетворяет на этом компакте условию Гельдера с некоторым показателем $\alpha>0$:

$$\|F'(R_2,\cdot)-F'(R_1,\cdot)\|\leqq c\|R_2-R_1\|^\alpha,\qquad \alpha\leqq 1.$$

Предположим, что компакт $\mathscr{P}$ содержит функцию $R_0(t)$ и при каждом $k=1,2,\ldots$ найдется число $\delta_k>0$ и функция $\varphi_k(t)$ такие, что $R_0(t)+s\varphi_k(t)\in\mathscr{P}$ для

всех $|s|<\delta_k$,

$$|\varphi_k(t)| \leqq I^{-1/2}(t),$$

$$\int_0^1 \varphi_k(t)F'(R_0,t)dt \to \int_0^1 I^{-1/2}(t)|F'(R_0,t)|dt (k\to\infty). \tag{2.1}$$

Тогда для любой оценки $\tilde{F}_n$ функционала $F(R)$ справедливо неравенство

$$\varliminf_{n\to\infty}\left[n \sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2 \right] \geqq \left(\int_0^1 I^{-1/2}(t)|F'(R_0,t)|dt \right)^2. \tag{2.2}$$

Доказательство. Рассмотрим параметрическое семейство

$$R_h^k(t) = R_0(t) + (h-\theta)g_k(t); \quad \theta = F(R_0);$$

$$g_k(t) = \varphi_k(t)\lambda_k^{-1}, \quad \lambda_k = \int_0^1 \varphi_k(t)F'(R_0,t)dt. \tag{2.3}$$

Ясно, что $(g_k(t), F'(R_0,t)) = 1((\cdot,\cdot)$ — скалярное произведение в $L_2)$.

В силу условия (2.1) $R_h^k \in \mathscr{P}$ при $|h-\theta|<\delta_k\lambda_k$. Поэтому для любой оценки $\tilde{F}_n$ функционала $F(R)$ выполнено неравенство

$$\sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2 \geqq \sup_{|h-\theta|<\delta_k\lambda_k} E\{\tilde{F}_n - F(R_h^k)\}^2.$$

Положим далее $\varepsilon = n^{-1/2}$. Для любой функции $\gamma(\varepsilon)\to+\infty$ при $\varepsilon\to 0$ очевидно неравенство

$$(\tilde{F}_n - F(R_h^k))^2 \geqq (\tilde{F}_n - h)^2(1-\gamma^{-1}(\varepsilon) - (h-F(R_h^k))^2\gamma(\varepsilon). \tag{2.4}$$

Пусть $\delta(\varepsilon)$ — функция, удовлетворяющая условиям $\delta(\varepsilon)\to\infty$, $\varepsilon\delta(\varepsilon)\to 0 (\varepsilon\to 0)$. Тогда при достаточно малом ε для любого k имеем $\varepsilon\delta(\varepsilon)<\delta_k\lambda_k$. И потому

$$\sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2 \geqq \sup_{|h-\theta|<\varepsilon\delta(\varepsilon)} E\{\tilde{F}_n - F(R_h^k)\}^2. \tag{2.5}$$

По теореме о среднем

$$F(R_h^k) = F(R_0 + (h-\theta)g_k) = \theta + (h-\theta)\,(g_k(t), F'(R_0 + \bar{\lambda}(h-\theta)g_k, t)),$$

$$0<\bar{\lambda}<1.$$

Отсюда с учетом условия Гельдера и соотношения (2.3) получим неравенство

$$(h-F(R_h^k))^2 = (h-\theta)^2(g_k(t), F'(R_0,t) - F'(R_0+\bar{\lambda}(h-\theta)g_k,t))^2 \leqq$$

$$\leqq C(h-\theta)^{2(1+\alpha)}\|g_k\|^{2(1+\alpha)} \leqq C(\varepsilon\delta(\varepsilon))^{2(1+\alpha)}.$$

Рассмотрим теперь задачу оценивания параметра h при каждом k и малых значениях $|h-\theta|$ по наблюдениям $t^{(n)}$, $X^{(n)}$ с совместным распределением (см. [1]).

$$dP^{(n)}_{k,h}(t^{(n)}, x^{(n)}) = Q_1(dt_1)p(x_1-(h-\theta)g_k(t_1)|t_1)Q_2(dt_2|t_1, x_1)\ldots$$

$$\ldots Q_n(dt_n|t_1, \ldots, t_{n-1}, x_1, \ldots, x_{n-1})p(x_n-(h-\theta)g_k(t_n)|t_n)dx_1, \ldots, dx_n.$$

(Здесь Q_i задает выбор очередной точки t_i при условии $t_1, x_1, \ldots, t_{i-1}, x_{i-1}$). В этом случае информационное количество Фишера для параметра h, содержащееся в наблюдениях $t^{(n)}$, $X^{(n)}$ может быть записано в виде

$$I^{(n)}_k(h) = \int \ldots \int \left(\frac{\partial}{\partial h} \ln \frac{dP^{(n)}_{k,\theta+h}(t^{(n)}, x^{(n)})}{dP^{(n)}_{k,\theta}(t^{(n)}, x^{(n)})} \right)^2 dP^{(n)}_{k,\theta+h}(t^{(n)}, x^{(n)}) =$$

$$= \sum_{i=1}^{n} \int_0^1 g_k^2(t_i) I(t_i) \tilde{Q}_i(dt_i),$$

где

$$\tilde{Q}_1(dt_1) = Q_1(dt_1); \quad \tilde{Q}_2(dt_2) = \iint p(x_1 - R_0(t_1) - hg_k(t_1)|t_1)dx_1 Q_1(dt_1) \times$$

$$\times Q_2(dt_2|t_1, x_1); \ldots; \tilde{Q}_n(dt_n) = \int \ldots \int p(x_{n-1} - R_0(t_{n-1}) - hg_k(t_{n-1})|t_{n-1})dx_{n-1} \times$$

$$\times Q_n(dt_n|t_1, \ldots, t_{n-1}, x_1, \ldots, x_{n-1}) \ldots p(x_1 - R_0(t_1) - hg_k(t_1)|t_1)dx_1 Q_1(dt_1)$$

некоторые распределения. Из условий (2.1) ясно, что $I^{(n)}_k(h) \leqq n\lambda_k^{-2}$.

Для оценивания параметра h в параметрическом семействе $P^{(n)}_{k,h}$ справедливо в условиях (1.2)–(1.3) неравенство Крамера–Рао (см. [2], с. 104). Поэтому

$$E_h\{\tilde{F}_n - h\}^2 \geqq (I^{(n)}_k(h))^{-1}(1 + b'_k(h))^2 + b_k^2(h) \geqq \frac{\lambda_k^2}{n}(1 + b'_k(h))^2 + b_k^2(h),$$

$b_k(h)$ — смещение оценки $\tilde{F}_n$ параметра h. Отсюда

$$\sup_{|h-\theta|<\varepsilon\delta(\varepsilon)} E_h\{\tilde{F}_n - h\}^2 \geqq (2\varepsilon\delta(\varepsilon))^{-1} \int_{\theta-\varepsilon\delta(\varepsilon)}^{\theta+\varepsilon\delta(\varepsilon)} E_h\{\tilde{F}_n - h\}^2 dh \geqq$$

$$\geqq (2\varepsilon\delta(\varepsilon))^{-1} \int_{\theta-\varepsilon\delta(\varepsilon)}^{\theta+\varepsilon\delta(\varepsilon)} \left[\frac{\lambda_n^2}{n}(1 + b'_k(h))^2 + b_k^2(h) \right] dh.$$

Обозначим $\varepsilon_k = \lambda_k \varepsilon$ и, используя методы вариационного исчисления (см. [3]), получим неравенство

$$\inf_{y(h)} \frac{1}{b-a} \int_a^b [\varepsilon_k^2(1 + y'(h))^2 + y^2(h)]dh \geqq \varepsilon_k^2 - C_1 \frac{\varepsilon^3}{b-a},$$

где $C_1 > 0$ — абсолютная константа.

Таким образом, установлено, что

$$\sup_{|h-\theta|<\varepsilon\delta(\varepsilon)} E_h\{\tilde{F}_n - h\}^2 \geqq \varepsilon^2\lambda_k^2 + o(\varepsilon^2).$$

Далее из (2.4) находим

$$\sup_{|h-\theta|<\varepsilon\delta(\varepsilon)} E_h\{\tilde{F}_n - F(R_h^k)\}^2 \geqq (1-\gamma^{-1}(\varepsilon))(\varepsilon^2\lambda_k^2 + o(\varepsilon^2)) - C(\varepsilon\delta(\varepsilon))^{2(\alpha+1)}\gamma(\varepsilon).$$

Отсюда и из (2.5), выбирая, например, $\delta(\varepsilon)=\varepsilon^{-\alpha/4}$, $\gamma(\varepsilon)=\varepsilon^{-\alpha/2}$ и вспоминая, что $\varepsilon^2=n^{-1}$, приходим к (2.2).

Замечание. Условие Гельдера на $F'(R, t)$ можно существенно ослабить, заменив его следующим неравенством

$$\|F'(R_1,\cdot)-F'(R_2,\cdot)\| \leqq \beta(\|R_1-R_2\|)$$

для любых R_1, $R_2\in\mathscr{P}$, где $\beta(x)$ удовлетворяет неравенству $\beta(x)<\dfrac{C}{\ln(1/x)}$. Доказательство в этом случае сохраняется, нужно лишь по-другому выбрать функции $\gamma(\varepsilon)$, $\delta(\varepsilon)$, например, положив $\gamma(\varepsilon)=\delta^2(\varepsilon)=\ln\ln\varepsilon$.

Следствие. Пусть выполнены условия теоремы 2.1 и

$$p(x|t)=(2\pi\sigma^2(t))^{-1/2}\exp\left\{-\frac{x^2}{2\sigma^2(t)}\right\}. \tag{2.6}$$

Тогда $I(t)=\sigma^{-2}(t)$ и справедливо соотношение

$$\varliminf_{n\to\infty}\left[n\sup_{R\in\mathscr{P}} E\{\tilde{F}_n - F(R)\}^2\right] \geqq \left(\int_0^1 \sigma(t)|F'(R,t)|dt\right)^2.$$

Обозначим $\mathscr{P}_\sigma$ множество условных плотностей «шума» $p(x|t)$, для которых $E\{\xi_i^2(t_i)|t_i=t\}=\sigma^2(t)$.

Из теоремы 2.1 и следствия из нее вытекает следующая теорема.

Теорема 2.2. Пусть выполнено условие (1.1), функционал $F(R)$ дифференцируем по Фреше на некотором компакте $\mathscr{P}<L_2$ и производная $F'(R, t)$ в L_2 удовлетворяет на этом компакте условию Гельдера с некоторым показателем $\alpha>0$. Предположим, что компакт $\mathscr{P}$ содержит функцию $R_0(t)$ и при каждом $k=1, 2, \ldots$ найдется число $\delta_k>0$, и функция $\varphi_k(t)$ такие, что $R_0(t)+s\varphi_k(t)\in\mathscr{P}$ для всех $|s|<\delta_k$

$$|\varphi_k(t)|\leqq\sigma(t), \quad \int_0^1 \varphi_k(t)F'(R_0,t)dt \to \int_0^1 \sigma(t)|F'(R_0,t)|dt (k\to\infty).$$

Тогда для любой оценки $\tilde{F}_n$ функционала $F(R)$ справедливо неравенство

$$\underline{\lim}_{n\to\infty}\left[n \sup_{\substack{R\in\mathscr{P} \\ p(\cdot,\cdot)\in\mathscr{P}_\delta}} E\{\tilde{F}_n - F(R)\}^2 \right] \geqq \left(\int_0^1 \sigma(t)|F'(R_0, t)|dt \right)^2. \tag{2.7}$$

Из неравенства (2.2) и (2.7) следует, что квадратическое отклонение оценки функционала $F(R)$ не может в асимптотически минимаксном смысле быть меньше, чем величина $(\int I^{-1/2}(t)|F'(R, t)|dt)^2 = \Phi_1(R, I)$ в случае известного распределения «шумов», или величины $(\int \sigma(t)|F'(R, t)||dt)^2 = \Phi_2(R, \sigma)$ в случае, когда известна условная дисперсия «шумов», если функционалы $\Phi_1(R, I)$ и $\Phi_2(R, \sigma)$ непрерывны по $R\in\mathscr{P}$. В связи с этим естественны следующие определения (ср. с [2]).

Определение 1. Оценка $\hat{F}_n$ называется асимптотически эффективной в K непараметрической оценкой (АЭНО) функционала $F(R)$ при условном распределении «шумов» с плотностью $p(x|t)$, если

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 I^{-1/2}(t)|F'(R, t)|dt \right)^2 \right] \to 0 \qquad (n\to\infty).$$

Определение 2. Оценка $\hat{F}_n$ называется асимптотически эффективной в K непараметрической оценкой (АЭНО) функционала $F(R)$ при известной условной дисперсии $\sigma^2(t)$, если

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R, t)|dt \right)^2 \right] \to 0 \qquad (n\to\infty).$$

В этой работе будет построена оценка гладкого функционала от регрессии, асимптотически эффективная в смысле определения 2. Для случая плотности $p(x|t)$ из (2.6) она является АЭНО и в смысле определения 1.

3. Асимптотически эффективная непараметрическая оценка

Как обычно, обозначим $W_2^\beta(L)$ множество τ раз дифференцируемых функций на интервале (0, 1), причем

$$\|f^{(\tau)}(t+h) - f^{(\tau)}(t)\| \leqq L|h|^\rho, \qquad \beta = \tau + \rho, \quad 0 < \rho \leqq 1.$$

(Здесь и далее, $\|\cdot\|$ — норма в пространстве L_2). При этом рассматриваются лишь такие значения h, при которых $x + h \in [0, 1]$ для $x \in [0, 1]$.

Ниже $W_2^\beta(L_1, T, [a, b])$ будет обозначаться класс периодических с периодом T функций, имеющих непрерывную производную порядка $\tau - 1 (\tau > 0)$, удовлетворяющую условию

$$\int_a^b (f^{(\tau)}(t+h)-f^{(\tau)}(t))^2 dt \leqq L_1 |h|^{2\rho}, \quad \beta=\tau+\rho, \quad 0<\rho\leqq 1, \quad T=b-a.$$

Пусть функция регрессии

$$R \in K \subset W_2^{\beta}(L), \qquad \beta > 1/2, \tag{3.1}$$

где K — известный статистику компакт в $L_2[0, 1]$. Тогда

$$\sup_{R\in K} |R(0)| < \infty. \tag{3.2}$$

Условия (1.1′) и (3.1) обеспечивают равномерную непрерывность функции $R(t)/\sigma(t)$ и равномерную ограниченность функции $R(t)$ на компакте K, т.е.

$$\sup_{R\in K} \max_{[0,1]} |R(t)| < \infty; \tag{3.1′}$$

Замечание. Из условия (3.1) и результатов [4] вытекает, что функцию $R(t)$ можно, не изменяя ее на $[0, 1]$, так продолжить вне этого отрезка, что полученная функция $R_1(t)$ принадлежит $W_2^{\beta}(L_1, T, [a, b])$, $a<0<1<b$ с периодом $T>1$. Ниже, там, где это необходимо, вместо R будет рассматриваться ее продолжение R_1, однако обозначаться продолженная функция будет также R.

Предположим, что для функционала $F(R)$ выполнены следующие условия

(F1) Функционал F: $L_2 \to R^1$ имеет производную Фреше $F'(R, t)$ в каждой точке $R \in K$.

(F2) Производная $F(R, t)$ удовлетворяет условию Гёльдера с показателем α равномерно в L_2, т.е. существуют такие положительные постоянные α, C_1, что для любых R_1, $R_2 \in K$

$$\|F'(R_1, \cdot)-F'(R_2, \cdot)\| \leqq C_1 \|R_1-R_2\|^{\alpha},$$

причем выполнено неравенство $(2\beta)^{-1} < \alpha \leqq 1$.

(F3) Производная $F'(R, t)$ на отрезке $[0, 1]$ удовлетворяет условию Гёльдера по t с каким-нибудь показателем $\gamma > 0$ равномерно по $R \in K$, т.е. существует такая положительная постоянная C_2, что для любых t_1, t_2 из отрезка $[0, 1]$ и $R \in K$

$$|F'(R, t_1)-F'(R, t_2)| \leqq C_2 |t_2-t_1|^{\gamma}.$$

Разобьем число наблюдений n на две части n_0 и $n_1 (n=n_0+n_1)$. Выберем $n_0=[n^{\varkappa}]$, где $\varkappa$ — какое-нибудь число из интервала

$$\frac{2\beta+1}{2\beta(\alpha+1)} < \varkappa < 1. \tag{3.3}$$

Очевидно, что $n/n_1 \to 1$ $(n\to\infty)$.

В [5] показано, что если $R(t) \in W_2^{\beta}(L_1, T, [a, b])$, то можно построить такой план $t^{(n_0)}$ и оценку $\hat{R}_{n_0}(t) \in K$, основанную на $t_1, X_1, \ldots, t_{n_0}$ наблюдениях, для которой справедливо неравенство

$$\sup_{R \in K} E\|R - \hat{R}_{n_0}\|^2 \leqq M n_0^{-\frac{2\beta}{2\beta+1}} \leqq M n^{-\frac{2\beta\varkappa}{2\beta+1}}, \quad M = \text{Const}. \tag{3.4}$$

Оставшиеся n_1 наблюдений потратим на оценку линейного функционала от регрессии

$$L_{n_0}(R) = \int_0^1 R(t) F'(\hat{R}_{n_0}, t)\, dt,$$

при этом воспользуемся способом, сходным с описанным в [6].

Пусть $\mathscr{F}_{n_0}$ — минимальная σ-алгебра событий, порожденная наблюдениями $t_1, \mathscr{X}_1, \ldots, t_{n_0}, \mathscr{X}_{n_0}$. Точки $t_{n_0+1}, \ldots, t_n$ плана $t^{(n_1)}$ выберем условно независимыми относительно фиксированной $\mathscr{F}_{n_0}$ с плотностью

$$p_{n_0}(t) = \sigma(t)|F'(\hat{R}_{n_0}, t)| \left(\int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\, dt \right)^{-1}.$$

Определим величину N равенством

$$N = \left[\frac{n_1 c}{\ln n_1} \right].$$

Разделим отрезок [0, 1] точками $a_1, \ldots, a_{N-1}$ на N частей $\Delta_1, \ldots, \Delta_N$ так, чтобы $[0, 1] = \bigcup_k \Delta_k$, $\Delta_i \bigcap \Delta_j = \varnothing$,

$$\int_{\Delta_k} \sigma(t)|F'(\hat{R}_{n_0}, t)|\, dt = N^{-1} \int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\, dt. \tag{3.5}$$

Предложенный выбор разбиения отрезка [0, 1] обеспечивает справедливость равенства

$$P\{t_i \in \Delta_k | \mathscr{F}_{n_0}\} = \int_{\Delta_k} p_{n_0}(t)\, dt = \frac{1}{N}. \tag{3.6}$$

Обозначим ν_k — число точек плана $t^{(n_1)}$, попавших в интервал Δ_k. В качестве оценки $\hat{L}_{n_1}$ функционала $L_{n_0}(R)$ рассмотрим величину

$$\hat{L}_{n_1} = \frac{1}{N} \int_0^1 \sigma(t)|F'(\hat{R}_{n_0}, t)|\, dt \sum_{k=1}^{N} \nu_k^{-1} \sum_{t_i \in \Delta_k} \frac{X_i \operatorname{sign} F'(\hat{R}_{n_0}, t_i)}{\sigma(t_i)}. \tag{3.7}$$

(Здесь и далее считаем $\nu_k^{-l} \sum_{t_i \in \Delta_k} (\ldots) = 0$, если $\nu_k = 0$, $l = 1, 2$).

Покажем, что оценка

$$\hat{F}_n = F(\hat{R}_{n_0}) + \hat{L}_{n_1} - \int_0^1 \hat{R}_{n_0}(t) F'(\hat{R}_{n_0}, t)\,dt \tag{3.8}$$

является АЭНО, т.е. докажем неравенство

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R,t)|\,dt \right)^2 \right] \leqq o(1) \qquad (n\to\infty). \tag{3.9}$$

Очевидно, что

$$\sup_{R\in K} \{nE[\hat{F}_n - F(R)]^2 - (\int_0^1 \sigma(t)|F'(R,t)|\,dt)^2\} =$$

$$= \sup_{R\in K} [nE\{F(\hat{R}_{n_0}) - F(R) + \int_0^1 F'(\hat{R}_{n_0}, t)\,(R(t) - \hat{R}_{n_0}(t))\,dt -$$

$$- \int_0^1 F'(\hat{R}_{n_0}, t) R(t)\,dt + \hat{L}_{n_1}\}^2 - (\int_0^1 \sigma(t)|F'(R,t)|\,dt)^2].$$

С помощью формулы Лагранжа (см. [7], с. 191)

$$F(\hat{R}_{n_0}) - F(R) = -\int_0^1 F'(\hat{R}_{n_0} + \theta(R - \hat{R}_{n_0}, t)\,(R(t) - \hat{R}_{n_0}(t))\,dt, \qquad 0<\theta<1$$

запишем неравенство

$$\sup_{R\in K}\left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R,t)|\,dt \right)^2 \right] \leqq$$

$$\leqq \sup_{R\in K}\left[nE\left\{ \int_0^1 (F'(\hat{R}_{n_0}, t) - F'(\hat{R}_{n_0} + \theta(R - \hat{R}_{n_0}), t))\,(R(t) - \hat{R}_{n_0}(t))\,dt \right\}^2 \right] +$$

$$+ 2 \sup_{R\in K}\left[n \left| E\left\{ \int_0^1 (F'(\hat{R}_{n_0}, t) - F'(\hat{R}_{n_0} + \theta(R - \hat{R}_{n_0}), t))\,(R(t) - \hat{R}_{n_0}(t))\,dt \times \right.\right.\right.$$

$$\left.\left.\left. \times (\hat{L}_{n_1} - L_{n_0}(R)) \right\} \right| \right] + \sup_{R\in K}\left[nE\{\hat{L}_{n_1} - L_{n_0}(R)\}^2 - \right.$$

$$\left. - \left(\int_0^1 \sigma(t)|F'(R,t)|\,dt \right)^2 \right], \qquad 0<\theta<1.$$

Используя предположение (F2), свойство (3.4) оценки $\hat{R}_{n_0}$ и неравенство (3.3), имеем

$$\sup_{R\in K}\left[nE\left\{ \int_0^1 (F'(\hat{R}_{n_0}, t) - F'(\hat{R}_{n_0} + \theta(R - \hat{R}_{n_0}), t))\,(R(t) - \hat{R}_{n_0}(t))\,dt \right\}^2 \right] \leqq$$

$$\leqq C_1 \sup_{R\in K} nE\|R-\hat{R}_{n_0}\|^{2(\alpha+1)} \leqq C_1 M n^{1-\frac{2\alpha\beta(\alpha+1)}{2\beta+1}} = o(1). \tag{3.10}$$

Таким образом, для доказательства (3.9) достаточно показать, что

$$\sup_{R\in K}\left[nE\{\hat{L}_{n_1}-L_{n_0}(R)\}^2 - \left(\int_0^1 \sigma(t)|F'(R,t)|dt\right)^2\right] = A_n \tag{3.11}$$

имеет верхний предел $A_0 \leqq 0$ при $n\to\infty$.

Отметим прежле, что аналогично (3.10) легко установить

$$\sup_{R\in K} E\left\{\int_0^1 F'(\hat{R}_{n_0},t)R(t)dt\right\}^2 \leqq \sup_{R\in K} E\|F'(\hat{R}_{n_0},\cdot)R(\cdot)\|^2 \leqq$$

$$\leqq \sup_{R\in K}\|R(\cdot)F'(R,\cdot)\|^2 + o(1) < \infty; \tag{3.12}$$

$$\sup_{R\in K} E\left\{\int_0^1 \sigma(t)|F'(\hat{R}_{n_0},t)|dt\right\}^2 \leqq$$

$$\leqq \sup_{R\in K}\left(\int_0^1 \sigma(t)|F'(R,t)|dt\right)^2 + o(1) < \infty. \tag{3.12'}$$

Введем аналогичные обозначения [6]

$$\zeta_k = v_k^{-1}\sum_{t_i\in\Delta_k}\frac{R(t_i)\operatorname{sign}F'(\hat{R}_{n_0},t_i)}{\sigma(t_i)}\int_0^1 \sigma(t)|F'(\hat{R}_{n_0},t)|dt,$$

$$R_k = N\int_{\Delta_k} R(t)F'(\hat{R}_{n_0},t)dt.$$

Далее, из (3.11) находим

$$A_n \leqq \sup_{R\in K}\frac{n}{N^2}E\left\{E\left[\left(\sum_{k=1}^N(\zeta_k-R_k)\right)^2\middle|\mathscr{F}_{n_0}\right]\right\} +$$

$$+\sup_{R\in K}\left[\frac{n}{N^2}E\left\{\left(\int_0^1\sigma(t)|F'(\hat{R}_{n_0},t)|dt\right)^2\sum_{k=1}^N E\left[v_k^{-2}\sum_{i=n_0+1}^n \chi(t_i\in\Delta_k)|\mathscr{F}_{n_0}\right]\right\} -\right.$$

$$\left. -\left(\int_0^1\sigma(t)|F'(R,t)|dt\right)^2\right]. \tag{3.13}$$

Учитывая соотношение (3.6) и неравенство (3.12), применим к первому слагаемому в (3.13) рассуждения теоремы 3.1 работы [6]. Новым, по сравнению с указанной работой, здесь является доказательство малости при сделанных выше предположениях выражения

$$\sup_{R\in K} N\sum_{k=1}^{N} E\left\{\int_{\Delta_k}\frac{R^2}{\sigma}(t)|F'(\hat{R}_{n_0},t)|dt\int_{\Delta_k}\sigma(t)|F'(\hat{R}_{n_0},t)|dt-\right.$$

$$\left.-\left(\int_{\Delta_k}R(t)F'(\hat{R}_{n_0},t)dt\right)^2\right\}\qquad(N\to\infty).$$

Легко проверить, что

$$\sup_{R\in K} N\sum_{k=1}^{N} E\left\{\int_{\Delta_k}\frac{R^2}{\sigma}(t)|F'(\hat{R}_{n_0},t)|dt\int_{\Delta_k}\sigma(t)|F'(\hat{R}_{n_0},t)|dt-\right.$$

$$\left.-\left(\int_0^1 R(t)F'(\hat{R}_{n_0},t)dt\right)^2\right\}=\frac{1}{2}\sup_{R\in K} N\sum_{k=1}^{N} E\left\{\int_{\Delta_k}\int_{\Delta_k}\left(\frac{R}{\sigma}(t)\operatorname{sign}F'(\hat{R}_{n_0},t)-\right.\right.$$

$$\left.\left.-\frac{R}{\sigma}(s)\operatorname{sign}F'(\hat{R}_{n_0},s)\right)^2\sigma(t)\sigma(s)|F'(\hat{R}_{n_0},t)|\,|F'(\hat{R}_{n_0},s)|dt\,ds\right\}.$$

Разобьем последнюю сумму на две части: по таким k, что $|\Delta_k|>N^{-\varepsilon}$ (обозначим ее Σ_1) и таким, что $|\Delta_k|\leqq N^{-\varepsilon}$ (обозначим ее Σ_2). Здесь ε какое-либо число, удовлетворяющее неравенствам $(1+\gamma)^{-1}<\varepsilon<1$. Покажем, что Σ_1 и Σ_2 стремятся к нулю при $N\to\infty$. Ясно, что число отрезков длины большей, чем $N^{-\varepsilon}$ не превосходит N^{ε}. Поэтому, с учетом (3.5), имеем

$$\Sigma_1\leqq GN^{\varepsilon+1}\sup_{R\in K}\left\{\int_0^1\sigma(t)|F'(\hat{R}_{n_0},t)|dt\right\}^2=G\frac{N^{\varepsilon+1}}{N^2}.$$

$$\sup_{R\in K} E\left\{\int_0^1\sigma(t)|F'(\hat{R}_{n_0},t)|dt\right\}^2\leqq GN^{\varepsilon-1}\to 0\qquad(N\to\infty).$$

(Здесь и далее символом G обозначены положительные постоянные, не обязательно одинаковые.)

Далее, среди тех интервалов, длина которых не превосходит $N^{-\varepsilon}$, рассмотрим отдельно те, где $F'(\hat{R}_{n_0},t)$ не меняет знак (интервалы Δ_k^+). Для них получаем

$$\sup_{R\in K} N\sum_{k:|\Delta_k^+|\leqq N^{-\varepsilon}} E\left\{\int_{\Delta_k}\int_{\Delta_k}\left(\frac{R}{\sigma}(t)-\frac{R}{\sigma}(s)\right)^2\sigma(t)\sigma(s)|F'(\hat{R}_{n_0},t)|\,|F'(\hat{R}_{n_0},s)|dt\,ds\right\}\leqq$$

$$\leqq\frac{N^2}{N^2}\rho(N^{-\varepsilon})\sup_{R\in K} E\left\{\int_0^1\sigma(t)|F'(\hat{R}_{n_0},t)|dt\right\}^2\leqq$$

$$\leqq G\rho(N^{-\varepsilon})\to 0\qquad(N\to\infty).$$

Здесь $\rho(t)$ — модуль непрерывности функции R/σ (R/σ — непрерывна в силу равенства (3.1′)).

Заметим теперь, что на интервалах, на которых $F'(\hat{R}_{n_0}, t)$ меняет знак (интервалы Δ_k^-), справедливо неравенство

$$|F'(\hat{R}_{n_0}, t)| \leqq C_2 \tilde{\rho}(N^{-\varepsilon}), \quad t \in \Delta_k^-, \quad |\Delta_k^-| \leqq N^{-\varepsilon},$$

где $\tilde{\rho}(t)$ — модуль непрерывности функции $F'(\hat{R}_{n_0}, t)$, не превышающий согласно (F3) величины $C_2 t^{\gamma}$. Поэтому

$$\sup_{R \in K} \sum_{k:|\Delta_k| \leqq N^{-\varepsilon}} NE \left\{ \int_{\Delta_k} \int_{\Delta_k} \left(\frac{R}{\sigma}(t) + \frac{R}{\sigma}(s) \right)^2 \sigma(t)\sigma(s) |F'(\hat{R}_{n_0}, t)| \, |F'(\hat{R}_{n_0}, s)| \, dt \, ds \right\} \leqq$$

$$\leqq G \frac{N^2}{N^{2\gamma\varepsilon}} \frac{1}{N^{2\varepsilon}} \sup_{R \in K} E \left\{ \int_0^1 \sigma(t) |F'(\hat{R}_{n_0}, t)| \, dt \right\}^2 \leqq G N^{2(1 - \varepsilon(\gamma + 1))}$$

в силу выбора ε.

Чтобы доказать (3.9) остается показать, что

$$\sup_{R \in K} \left[\frac{n}{N^2} E \left\{ \left(\int_0^1 \sigma(t) |F'(\hat{R}_{n_0}, t)| \, dt \right)^2 \sum_{k=1}^N E \left[\nu_k^{-2} \sum_{i=n_0+1}^n \chi(t_i \in \Delta_k) | \mathscr{F}_{n_0} \right] \right\} - \right.$$

$$\left. - \left(\int_0^1 \sigma(t) |F'(R, t)| \, dt \right)^2 \right] = o(1).$$

В самом деле, применяя, как и при доказательстве теоремы 3.2 в [6], лемму 3.1 из [6], находим

$$\sum_{k=1}^N E \left[\nu_k^{-2} \sum_{i=n_0+1}^n \chi(t_i \in \Delta_k) | \mathscr{F}_{n_0} \right] = \sum_{k=1}^N \sum_{i=n_0+1}^n E \left[\frac{\chi(t_i \in \Delta_k)}{\left(1 + \sum_{j=1, i \neq j}^n \chi(t_j \in \Delta_k) \right)^2} \middle| \mathscr{F}_{n_0} \right] =$$

$$= \sum_{k=1}^N \sum_{i=n_0+1}^n E[\chi(t_i \in \Delta_k) | \mathscr{F}_{n_0}] E \left[\left(1 + \sum_{j=1, j \neq i}^n \chi(t_j \in \Delta_k) \right)^{-2} \middle| \mathscr{F}_{n_0} \right] = \frac{N^2}{n_1} + o\left(\frac{1}{n_1} \right).$$

Наконец, учитывая (3.12), получаем

$$\sup_{R \in K} \left[\frac{n}{N^2} E \left\{ \left(\int_0^1 \sigma(t) |F'(\hat{R}_{n_0}, t)| \, dt \right)^2 \sum_{k=1}^N E \left[\nu_k^{-2} \sum_{i=n_0+1}^n \chi(t_i \in \Delta_k) | \mathscr{F}_{n_0} \right] \right\} - \right.$$

$$\left. - \left(\int_0^1 \sigma(t) |F'(R, t)| \, dt \right)^2 \right] = \sup_{R \in K} \left[\frac{n}{n_1} E \left\{ \int_0^1 \sigma(t) |F'(\hat{R}_{n_0}, t)| \, dt \right\}^2 - \right.$$

$$\left. - \left(\int_0^1 \sigma(t) |F'(R, t)| \, dt \right)^2 \right] + o(1) = o 1).$$

Итак, нами доказана следующая теорема.

Теорема. Пусть выполнены условия (1.1)–(1.1′), $R \in W_2^{\beta}(L)$ при $\beta > 1/2$, $\sup_{R \in K} |R(0)| < \infty$, функционал $F(R)$ удовлетворяет условиям (F1)–(F3). Тогда оценка $\hat{F}_n$, определяемая равенством (3.8) с $n_0 = [n^{\varkappa}]$, $N = \left[\dfrac{n, c}{\ln n_1} \right]$, где числа $\varkappa$ и c — любые, удовлетворяющие неравенствам

$$\frac{2\beta + 1}{2\beta(\alpha + 1)} < \varkappa < 1, \qquad 0 < c < 1$$

является асимптотически эффективной в K непараметрической оценкой нелинейного функционала $F(R)$ от функции регрессии для квадратической функции потерь, так что

$$\lim_{n \to \infty} \sup_{R \in K} \left[nE\{\hat{F}_n - F(R)\}^2 - \left(\int_0^1 \sigma(t) |F'(R, t)| dt \right)^2 \right] = 0.$$

Литература

1. *Хасьминский Р. З.* О непараметрическом оценивании линейного функционала от регрессии при планировании наблюдений. Проблемы передачи информации, 1986. Т. **22,** Вып. *3*, с. 43–61.
2. *Ибрагимов И. А., Хасьминский Р. З.* Асимптотическая теория оценивания. М.: Наука, 1979.
3. *Ченцов Н. Н.* Об оценке неизвестного среднего многомерного нормального распределения. Теория вероятностей и ее применения, 1967. Т. **12,** № *4*, с. 619–633.
4. *Никольский С. М.* О продолжении функций многих переменных с сохранением дифференциальных свойств. Математический сборник, 1956. Т. **40***(82)*, с. 243–268.
5. *Ибрагимов И. А., Хасьминский Р. З.* Асимптотическая граница качества непараметрического оценивания регрессии в $\mathscr{L}_p$. Записки научных семинаров ЛОМИ, 1981. Т. **97,** с. 88–101.
6. *Пастухова Ю. И., Хасьминский Р. З.* Асимптотические оценки линейного функционала от регрессии при заданном плане наблюдений. Проблемы передачи информации, 1988. Т. **24,** Вып. *3*, с. 42–51.
7. Справочная математическая библиотека. Функциональный анализ. Под ред. С. Г. Крейна. М.: Наука, 1964.

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda Vállalat, 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

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: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

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. TOM 18 · NUMBER HOMEP 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

1989

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 PLC Headington Hill Hall, Oxford OX3 OBW, England

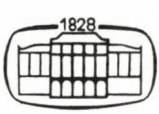
or

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

1989 Subscription Rate DM 535,— 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. 18 (3), pp. 135—149 (1989)

STABILIZATION OF UNCERTAIN DYNAMIC DELAYED PROCESSES BY BINARY CONTROL SYSTEMS

S. V. EMELYANOV, S. K. KOROVIN, I. G. MAMEDOV, A. L. NERSISYAN

(*Moscow*)

(Received April 7, 1988)

Uncertain dynamic delayed systems represented as linear differential systems over convolution rings are stabilized when the uncertainty is uniformly constrained and active in the space where the control is. A discontinuous control algorithm is proposed, conditions for sliding modes to start are indicated. The necessary and sufficient conditions are specified under which such systems are assigned the required spectrum in sliding mode. Discontinuous controls are regularized in the class of continuous functions. Finally, examples are provided.

1. Introduction

Stabilization of dynamic delayed systems is a major subject in control theory and is usually obtained by the Reswick–Smith controller [1]. In recent years functional analysis has yielded the necessary and sufficient conditions under which any desired spectrum the system is assigned by using feedback [2], [4]. The feedback equation is derived from the current and past system states and integrals of past values of the state variables [6]. Algebraic methods of studying delayed systems [5] identify cases where the closed-loop system is assigned have the desired spectrum by using only the current and past values of the state coordinates. In these methods the process parameters are assumed available and independent of time.

Controls of uncertain delayed processes are obtained chiefly by using Lyapunov–Krasovsky functionals [4] and Razumikhin functions [10]. In [9], [10] the quasi-stationarity conditions are assumed to hold when gradient adaptation algorithms are affected and the Lyapunov–Krasovskii functionals are employed to prove stability of certain adaptive control systems. In [11] the Razumikhin functions and in [12] the Lyapunov–Krasovskii functionals stabilize systems with the feedback providing data on the current values of phase variables of essentially non-stationary delayed processes. The algorithm of [11], [12] does not, however, assign the desired spectrum to the system, but this spectrum may be important in, say, optimal control. For this reason control algorithms are needed which would assign the desired

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

spectrum in the face of any perturbations which are known at every time up to compact sets. In other words, such feedbacks make the closed-loop system independent of or not very sensitive to, exogenous perturbations. This paper is chiefly concerned with design of such algorithms by structural and space decomposition so as to simplify the design of the control system and obtain feedbacks with compact operators. The control system performance is understood in the sense of the preassigned spectrum, which is equivalent mathematically to existence of linear manifolds in the system state space the mapping of motions on which has the preassigned spectrum. Therefore, it is initially required to establish the existence of this manifold and derive the functional relation which describes it; then a control is chosen which imparts the properties of a stable attractor to this manifold. The functional relation results from structural transformations which were proposed in [14] for finite-dimensional systems while stability and attraction of the chosen manifold are obtained by using discontinuous controls. The motion on the manifold proceeds in sliding mode. The space decomposition method [14] leads to a general approach to regularization of discontinuous controls by a class of continuous functions. For regularized control algorithms the complete insensitivity of transient processes to perturbations becomes a "weak" dependence whose degree of smallness decreases with the weakening of constraints on the control signal rate of change. The examples will confirm that the proposed methods are useful.

2. Notation and definitions

Let us take up the set $\mathcal{I}$ of formal sums

$$\alpha + \sum_{i=0}^{q} a_i \delta_{b_i},$$

where

$$\alpha \in L_0^{\mathrm{loc}}, \quad a_i \in \mathbf{R}, \quad b_i \in \mathbf{R}, \quad b_0 = 0, \quad b_j > 0, \quad j = 1, \ldots, q, \quad \delta_a = \delta(t-a)$$

is a Dirac distribution in the point a, L_0^{loc} is the space of locally-integrable functions whose supports are constrained and are in $[0, \infty)$. Let L_+^{loc} denote the space of locally integrable functions with supports constrained on the left and C the space of continuous functions. The set $\mathcal{I}$ with the summation and convolution operations forms a commutative ring with an identity δ_0, or $k * \delta_0 = k$, $k \in \mathbf{R}$. Every element θ of the ring $\mathcal{I}$ defines the mapping $L_+^{\mathrm{loc}} \to L_+^{\mathrm{loc}}$. Let p be a generalized derivative of δ_0 and $\mathfrak{N}$ a field of proper fractional rational functions such as

$$\left(p^m + \sum_{i=0}^{m-1} \gamma_i * p^i\right) * \left(p^n + \sum_{j=0}^{n-1} v_i * p^j\right)^{-1}, \qquad n \geqq m,$$

where

$$p^k = \underbrace{p * \ldots * p}_{k}, \quad \gamma_i, \quad v_j \in \mathscr{I}, \quad i=0,1,\ldots,m-1, \quad j=0,1,\ldots,n-1.$$

Every $\mathscr{I}$ obviously determines the mapping. Let $(L_+^{\mathrm{loc}})^{n\times m}$, $\mathscr{I}^{n\times m}$, and $\mathfrak{R}^{n\times m}$ denote spaces of $(n\times m)$ matrices whose components are elements of L_+^{loc}, $\mathscr{I}$, and $\mathfrak{R}$, respectively.

3. Problem statement and basic assumption

Dynamic systems are considered

$$\dot{x} = F * x + g * (u+\varphi), \tag{3.1}$$

where $x(t) \in (L_+^{\mathrm{loc}})^n$ is the process, $u(t) \in L_+^{\mathrm{loc}}$ is the control, $\varphi(t, x) \in L_+^{\mathrm{loc}}$ is the exogenous perturbation, and $F \in \mathscr{I}^{n\times n}$ and $g \in \mathscr{I}^{n\times 1}$ are process parameters. The process is a solution of equation (3.1) with a control $u(t)$ and perturbation $\varphi(t, x)$.

Example. A delayed process

$$\dot{x}_1 = x_1(t-\tau) + x_2 + x_2(t-\tau)$$

$$\dot{x}_2 = x_1 + u(t-\tau) + \varphi(t-\tau)$$

is represented by a system of convolution equations

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \delta_\tau & \delta_0+\delta_\tau \\ \delta_0 & 0 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \delta_\tau \end{bmatrix} * (u+\rho).$$

To formulate the problem, the following definitions will be needed.

Definition 1. System (3.1) is exponentially stable with an order $\lambda > 0$ if any of its solution $x(t)$ at any time $t \geqq t_1$ satisfies the estimate

$$|x(t)| \leqq N \exp[-\lambda t],$$

with $N > 0$ and $t_1 \geqq t_0$, depending on the specific solution.

Definition 2. The pair (F, g), $F \in \mathscr{I}^{n\times n}$, $g \in \mathscr{I}^{n\times 1}$ is spectrally controllable if for the specified elements $\alpha_0, \ldots, \alpha_{m-1}$, $m \geqq n$ of the ring $\mathscr{I}$ there is a vector $h \in \mathfrak{R}^{1\times n}$ such that

$$\det(pE_n - F + g * h) = p^m + \sum_{i=0}^{m-1} \alpha_i * p^i.$$

The following equality [2] is a plain criterion of spectral controllability

$$\operatorname{rank} L(s) = n, \quad \forall s \in \mathbf{C},$$

where $L(s)$ is a superposition of the Laplace transformation and the operator $U=[pE_n-F\vdots g]$ and $\mathbf{C}$ is a set of complex numbers.

The following assumptions which identify the class of systems to be considered are believed to hold.

Assumption I. The pair (F, g) is spectrally controllable.

Assumption II. The perturbation $\varphi(t, x)$ at any $(t, x)\in\mathbf{R}_+\times\mathbf{R}_x^n$ satisfies the inequality

$$|\rho(t, x)|\leqq\Phi(t, x)$$

where $\Phi(t, x)$ is a known function.

Assumption III. There is a vector $h_0\in\mathscr{I}^{1\times n}$ such that

$$(h_0 * g)^{-1}\in\mathscr{I}.$$

Assumption IV. For any $x=[x', x'']\in(L_+^{\mathrm{loc}})^n$, $x''\in L_+^{\mathrm{loc}}$ the equation $h_0 * x=0$ can be solved for x'' in the ring $\mathscr{I}$, or

$$x''=\tilde{h}_0 * x', \quad \tilde{h}_0\in\mathscr{I}^{1\times(n-1)}.$$

Example. The vector $g=[\delta_0+\delta_\tau, \delta_\tau]^T$ satisfies Assumptions III and IV while $h_0=[\delta_0, -\delta_0]$, $(h_0 * g)^{-1}=\delta_0\in\mathscr{I}$, $\tilde{h}_0=\delta_0$. But the vector $g=[0, \delta_\tau]^T$ does not satisfy Assumption III and the vector $g=[\delta_0+\delta_{3\tau}, \delta_{5\tau}]^T$ — Assumption IV.

4. Problem statement

With Assumptions I–IV satisfied, it is required to determine the feedback with the control $u(x(t))\in L_+^{\mathrm{loc}}$ whereby the closed-loop system (3.1) is exponentially stable with any specified order $\lambda>0$ while the solution of the closed-loop system $x(t)\in L_+^{\mathrm{loc}}$ coincides, following a certain time interval, with solutions for a stationary system having the desired spectrum.

The quasi-decoupling method

The problem is solved by the quasi-decoupling method [14] which proceeds in two stages:

— structural transformation; and

— space decomposition.

At the first stage for specified $\alpha_i\in\mathscr{I}$, $i=1, \ldots, m$, $m\geqq n$ the coordinates are replaced

$$x=M * \sigma, \quad M\in\mathfrak{R}^{n\times n}, \quad M^{-1}\in\mathfrak{R}^{n\times n} \tag{4.1}$$

so as to represent system (3.1) as two interconnected subsystems

$$\dot{\sigma}_1 = A_{11} * \sigma_1 + A_{12} * \sigma_2 \tag{4.2a}$$

$$\dot{\sigma}_2 = A_{21} * \sigma_1 + A_{22} * \sigma_2 + b_2 * (u + \varphi), \tag{4.2b}$$

so that

$$\det (pE_{n-1} - A_{11}) = p^m + \sum_{i=0}^{m-1} \alpha_i * p^i \tag{4.3}$$

and

$$b_2^{-1} \in \mathscr{I}. \tag{4.4}$$

The space decomposition specifies the functional dependence of the coordinates σ_1 and σ_2 with which process $x(t)$ has the desired properties. The dependences of the process $x(t)$ on the perturbations are represented by a pair of nonnegative numbers (δ, γ) and the inequality

$$|\sigma_2| \leqq \delta |\sigma_1| + \gamma. \tag{4.5}$$

If $\delta = 0$ and $\gamma = 0$, then

$$\sigma_2 = 0 \tag{4.6}$$

and σ_1 and so the process $x(t)$ are independent of $\varphi(t)$. Functional relations (4.5) identify in $(L_+^{\text{loc}})^n$ the parametric family of the sets

$$G(\delta, \gamma) = \{\sigma \in (L_+^{\text{loc}})^n : |\sigma_2| \leqq \delta |\sigma_1| + \gamma\},$$

which are referred to as decompositional sets because the set of processes $\sigma(t)$ may be divided into three subsets $\mathscr{X} = \bigcup_{i=1}^{3} \mathscr{X}_i$.

$$\mathscr{X}_1 = \{\sigma(t) \in \mathscr{X} : \exists\, t' < \infty, \quad \forall\, t \geqq t', \quad \sigma(t) \in G(\delta, \gamma)\},$$

$$\mathscr{X}_2 = \{\sigma(t) \in \mathscr{X} : \exists\, t' < \infty, \quad \forall\, t \geqq t', \quad \sigma(t) \bar{\in} G(\delta, \gamma)\},$$

$$\mathscr{X}_3 = \mathscr{X} \setminus \bigcup_{i=1}^{2} \mathscr{X}_i.$$

With small $\delta > 0$ and $\gamma > 0$ the asymptotic behavior of processes in the class $\mathscr{X}_1$ is defined by the equation

$$\dot{\sigma}_1 = A_{11} * \sigma_1. \tag{4.7}$$

The properties of processes from the class $\mathscr{X}_2$ are dictated by the system

$$\dot{\sigma}_2 = A_{22} * \sigma_2 + b_2 * (u + \varphi). \tag{4.8}$$

Now, if with some control $u(t)$ the set $G(\delta, \gamma)$ is right-invariant for system (4.2), the space of solutions for a closed-loop system is

$$X \subset \mathscr{X}_1 \bigcup \mathscr{X}_2.$$

But with $\gamma = 0$ processes from $\mathscr{X}_1$ and $\mathscr{X}_2$ are linearly independent and their properties are defined by independent subsystems (4.7) and (4.8). In other words, the solution space is decomposed, although subsystems (4.2a) and (4.2b) remain interdependent. This fact has dictated, in particular, the name of the method. Consequently, in the quasi-decoupling method the feedback is designed so as to make the sets $G(\delta, \gamma)$ invariant and system (4.8) have the desired properties. Let us now proceed to conditions and form of the transformation M with which (3.1) can be represented as (4.2)–(4.4).

5. Structural transformations

Let us assume a vector $h \in \mathfrak{R}^{1\times n}$ which generates the transformation M from (4.1). If $(h * g)^{-1} \in \mathscr{I}$, then the space $(L_+^{\mathrm{loc}})^n$ expands into a direct sum

$$(L_+^{\mathrm{loc}})^n = H(h) + \mathscr{R}(g), \tag{5.1}$$

where

$$H(h) = \{x \in (L_+^{\mathrm{loc}})^n : h * x = 0\}, \quad \mathscr{R}(g) = \{x \in (L_+^{\mathrm{loc}})^n : \exists\, y \in L_+^{\mathrm{loc}}, \quad x = g * y\}.$$

The decoupling $(L_+^{\mathrm{loc}})^n$ in (5.1) is carried out by the mapping operators

$$P_1 : (L_+^{\mathrm{loc}})^n \to H(h), \quad P_2 : (L_+^{\mathrm{loc}})^n \to \mathscr{R}(g) : P_1 = E_n - g * (h * g)^{-1} * h,$$
$$P_2 = g * (h * g)^{-1} * h. \tag{5.2}$$

Represent an arbitrary function $x(t) \in (L_+^{\mathrm{loc}})^n$ in the form

$$x(t) = x_1(t) + x_2(t), \quad x_i(t) = (P_i * x)(t), \quad i = 1, 2;$$

denote

$$\sigma_2 = h * x, \quad x_1 = [\sigma_1^T, \sigma_1'], \quad \sigma_1 \in (L_+^{\mathrm{loc}})^{n-1}$$

and assume that the equation $h * x_1 = 0$ can be solved in H for $\sigma_1' = \in L_+^{\mathrm{loc}}$, or $\sigma_1' \tilde{h} * \sigma_1$, $\tilde{h} \in H^{1\times(n-1)}$. Then, in the coordinates $\sigma = (\sigma_1^T, \sigma_2^T)$ system (3.1) takes the form of (4.2) and

$$A_{11} = F_{11} + F_{12} * \tilde{h}, \quad P_1 * F = \left[\begin{array}{c|c} F_{11} & F_{12} \\ \hline F_{21} & F_{22} \end{array}\right] \begin{matrix} \}n-1 \\ \}1 \end{matrix}, \quad A_{12} = Q_1,$$

$$P_1 * F * g * (h * g)^{-1} = \left[\begin{array}{c} Q_1 \\ \hline Q_2 \end{array}\right] \begin{matrix} \}n-1 \\ \}1 \end{matrix}, \quad A_{22} = h * F * g * (h * g)^{-1},$$

$$A_{21} = h * F * \left[\begin{array}{c|c} E_{n-1} & \\ \hline & \tilde{h} \end{array}\right], \quad b_2 = h * g.$$

Projectors P_1 and P_2 make it possible to determine the form of transformation (4.1)

$$M=\left[\begin{array}{c|c} E_{n-1} & g * (h * g)^{-1} \\ \hline \tilde{h} & \end{array}\right], \quad M \in \mathfrak{R}^{n\times n}. \tag{5.3}$$

It is easy to see that $M^{-1} \in \mathfrak{R}^{n\times n}$.

Equations (4.1) and (5.3) suggest that in general the initial conditions for $\sigma(t)$ cannot be determined in terms of those for $x(t)$. This is obvious in the following example.

The matrix $M=\begin{bmatrix} \delta_0 & 0 \\ -\delta_\tau & \delta_0 \end{bmatrix}$ transforms a system of ordinary differential equations[1]

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = u, \quad x(0)=x^0$$

into a system of equations with delays

$$\dot{\sigma}_1 = -\sigma_1(t-\tau)+\sigma_2$$

$$\dot{\sigma}_2 = -\sigma_1(t-2\tau)+\sigma_2(t-\tau)+u,$$

$$\sigma(\theta)=\rho(\theta) \quad \text{with} \quad \theta\in[-2\tau, 0].$$

For the solution spaces of initial system (3.1) and the transformed one (4.2) to match, the latter has to be considered at $t\geqq t_0$ where t_0 is the time starting with initial conditions which may be specified for (3.1). In the case now discussed $t_0=2\tau$.

The following theorem defines the necessary and sufficient conditions under which subsystem (3.7) may be compelled to have the desired spectrum by a proper choice of $M\in\mathfrak{R}^{n\times n}$ (or of the vector $h\in\mathfrak{R}^{1\times n}$).

Theorem 1. With Assumptions II and IV satisfied the pair (F, g) is spectrally controllable if and only if for arbitrary elements $\alpha_0, \ldots, \alpha_{m-1}$ of the ring $\mathscr{I}$ there is a vector $h=(\tilde{h}, \delta_0)\in\mathfrak{R}^{1\times n}$ such that $(h * g)^{-1}\in\mathscr{I}$ and

$$\det(pE_{n-1}-A_{11}) = p^m + \sum_{i=0}^{m-1} \alpha_i * p^i.$$

Proof of Theorem 1. From Assumptions III and IV follows existence of the transformation

$$M_0=\left[\begin{array}{c|c} E_{n-1} & g * (h_0 * g)^{-1} \\ \hline \tilde{h}_0 & \end{array}\right], \quad M_0\in\mathscr{I}^{n\times n}, \quad M_0^{-1}\in\mathscr{I}^{n\times n}$$

[1] For details see F. Kappel, J. Diff. Eq., vol. **24**, 99–126, 1977.

with which the pair (F, g) takes the form $(\tilde{F}, \tilde{g})$, where

$$\tilde{F} = M_0^{-1} * F * M_0, \quad g = M_0^{-1} * g = [0, \ldots, 0, (h_0 * g)]^T.$$

To prove the Theorem, let us use one finding of control theory.

Lemma. The pair (F, g) is controllable if and only if the pair $(\tilde{F}_{11}, \tilde{F}_{12})$ is also controllable, where

$$\tilde{F} = \left[\begin{array}{c|c} \tilde{F}_{11} & \tilde{F}_{12} \\ \hline \tilde{F}_{21} & \tilde{F}_{22} \end{array}\right].$$

Necessity. Assuming that (F, g) is spectrally controllable so is $(\tilde{F}_{11}, \tilde{F}_{12})$ and for specified $\alpha_1, \ldots, \alpha_{m-1} \in \mathscr{I} (m \geqq n)$ there is [7] a vector $\tilde{h} \in \mathfrak{R}^{1 \times (n-1)}$ such that

$$\det (pE_{n-1} - (\tilde{F}_{11} - \tilde{F}_{12} * \tilde{h})) = p^m + \sum_{i=0}^{m-1} \alpha_i * p^i .$$

Choose $h = (\tilde{h}, \delta_0)$; then $(h * \tilde{g})^{-1} = (h_0 * g)^{-1} \in \mathscr{I}$. The transformation $M = M_0 * \tilde{M}$ where

$$\tilde{M} = \left[\begin{array}{c|c} E_{n-1} & \tilde{g} * (h * \tilde{g})^{-1} \\ \hline \tilde{h} & \end{array}\right], \quad \tilde{M} \in \mathfrak{R}^{n \times n}, \quad \tilde{M}^{-1} \in \mathfrak{R}^{n \times n}$$

makes it possible to represent (3.1) as (4.2)–(4.4) and

$$A_{11} = \tilde{F}_{11} - \tilde{F}_{12} * \tilde{h}, \quad b_2 = h_0 * g .$$

Sufficiency. Immediately follows from the Lemma. Indeed, the condition of the Theorem presumes spectral controllability of the pair $(\tilde{F}_{11}, \tilde{F}_{12})$ and so, according to the Lemma, is also the pair (F, g). This completes the proof of the Theorem.

Remark 1. Reference [6] reports conditions for the case of commensurable delays and spectral controllability of the pair $(\tilde{F}_{11}, \tilde{F}_{12})$ under which system (4.7) is exponentially stable with an order $\lambda > 0$.

For this it is sufficient that

$$\det (pE_{n-1} - A_{11}) = (p + p_1) * \ldots * (p + p_m), \quad m \geqq n-1, \; p_j \in \mathbf{R}, \; p_j < \lambda, \quad j = 1, \ldots, m.$$

Remark 2. If

$$\det{}^{-1} [\tilde{F}_{12} \,|\, \tilde{F}_{11} * \tilde{F}_{12} \,|\, \cdots \,|\, F_{11}^{n-1} * F_{12}] \in \mathscr{I} ,$$

then $h \in \mathscr{I}^{1 \times n}$.

Remark 3. If $n > 2$ and for $\tilde{F}_{12}$ assumptions similar to III and IV hold, then all the propositions of this Section also hold for the pair $(\tilde{F}_{11}, \tilde{F}_{12})$.

6. Discontinuous control in delayed systems

The control is chosen so that solutions of the closed-loop system abide by functional relation (4.5) which ensures independence of the process of exogenous disturbances. The closed-loop system is exponentially stable with an exponent $\lambda>0$ if solutions of such a system from $\mathscr{X}_1$ and $\mathscr{X}_2$ decay with the same exponent $\lambda>0$ while the $G(0, 0)$ is right-invariant for the system. Choose the vector $h\in\mathfrak{R}^{1\times n}$ in compliance with Remark 1. Then the processes from $\mathscr{X}_1$ decay with an order $\lambda>0$. If, however, with $\sigma_2\neq 0$ the inequality

$$\operatorname{sgn}\sigma_2(\dot{\sigma}_2+\lambda\sigma_2)\leqq 0 \tag{6.1}$$

holds, then the processes from $\mathscr{X}_2$ also decay with an order $\lambda>0$. By virtue of (4.2) we have from (6.1)

$$\operatorname{sgn}\sigma_2(A_{21}\ast\sigma_1+(A_{22}+\lambda\delta_0)\ast\sigma_2+(h\ast g)\ast(u+\rho))\leqq 0 \tag{6.2}$$

or, by virtue of the initial system

$$\operatorname{sgn}(h\ast x)(h\ast(\lambda E_n+F)\ast x+(h\ast g)\ast(u+\rho))\leqq 0. \tag{6.3}$$

Because $(h\ast g)^{-1}\in\mathscr{I}$, without limiting the generality $h\ast g=\delta_0$. Choose the control in the form

$$u=u_1+u_2 \tag{6.4}$$

$$u_1=-h\ast(\lambda E_n+F)\ast x \tag{6.5}$$

$$u_2=-\operatorname{sgn}(h\ast x)\cdot u_0+\rho \tag{6.6}$$

$u_0\in L_+^{\mathrm{loc}}$ and at every $t\geqq t_0$, $u_0(t)>\Phi(t, x)$.

Control system (3.1), (6.4)–(6.6) is a variable structure system, VSS [13]. Solutions of the VSS are understood in the Filippov sense [16]. Indeed, the solution of the equation

$$\dot{\sigma}_2=-\lambda\sigma_2-\operatorname{sgn}\sigma_2\cdot u_0+\varphi \tag{6.7}$$

is to be found, by Filippov's definition, in the space L_+^{loc}, or $\sigma_2(t)\in L_+^{\mathrm{loc}}$ while solutions of (4.2a) are understood in the conventional sense [16] because $\sigma_2(t)\in L_+^{\mathrm{loc}}$.

Then holds

Theorem 2. If the vector $h\in\mathfrak{R}^{1\times n}$ is chosen in compliance with Remark 1, the VSS is exponentially stable with an exponent $\lambda>0$ at every $t\geqq t_0$. If, however, $u_0>\Phi(t, x)+\Delta$ where $\Delta>0$, then all solutions of the VSS are in $\mathscr{X}_1$.

Proof. According to Filippov, the set $G(0, 0)$ is right-invariant for the VSS. Then all solutions of the VSS are in the set $\mathscr{X}_1\bigcup\mathscr{X}_2$ and so decay with an order $\lambda>0$. If, however, $u_0(t)>\Phi(t, x)+\Delta$, then $\operatorname{sgn}\sigma_2\cdot\dot{\sigma}_2<-\Delta$ and all solutions of the VSS are in $\mathscr{X}_1$. This completes the proof of Theorem 2.

7. Regularization of discontinuous control

The constraint, which is frequently to be found in applications on the rate of change of the control makes discontinuous controls impossible. This Section will describe a technique to regularize discontinuous controls by a class of continuous functions, the properties of the closed-loop system being very nearly those of the VSS.

The set $G(\delta, \gamma)$ defines with $\delta \geqq 0$ and $\gamma > 0$ a certain vicinity of the surface $\sigma_2 = 0$. With $\delta \geqq 0$ and $\gamma > 0$ fairly small the asymptotic behaviour of processes from $\mathscr{X}_1$ is defined starting from a certain time, by system (4.7) and is independent of the control $u(t)$. This is why at time $t \geqq t_0$ when $\sigma(t) \in G$ the control will be generated so that $u(t) \in C$. For this purpose introduce an indicatrix $\omega(t) = \dfrac{\sigma_2(t)}{\delta|\sigma_2(t)| + \gamma}$ of the set $G(\delta, \gamma)$. If $|\omega| \leqq 1$, then $\sigma(t) \in G(\delta, \gamma)$. The regularized algorithm is specified by the relation

$$u_2 = \Gamma \cdot u_0 ,$$

$$\Gamma(t) = \begin{cases} \Gamma_+(t) & |\omega(t)| \geqq 1 \\ \Gamma_0(t) & |\omega(t)| < 1 \\ \Gamma_-(t) & |\omega(t)| \leqq -1 , \end{cases} \tag{7.1}$$

where $u_0(t), \Gamma_+(t), \Gamma_0(t), \Gamma_-(t) \in L_+^{\text{loc}}$ and $\Gamma(t) \in C$. Control system (3.1), (6.4), (6.5), (7.1) will be referred to as a regularized variable structure system and denoted as RVSS. The following Theorem provides condition for the set $G(\delta, \gamma)$ to be right-invariant for solutions of the RVSS.

Theorem 3. If $\gamma > 0$, $\delta > 0$ and at every $t \geqq t_0$, $\Gamma_+(t) \leqq -1$, $\Gamma_-(t) \geqq 1$,

$$u_0(t) > \Phi(t, x) - \delta|(H * x)(t)| , \tag{7.2}$$

where $H = P_1' * F$, $P_1 = \begin{bmatrix} P_1' \\ \hline P_1'' \end{bmatrix} \begin{matrix} \}n-1 \\ \}1 \end{matrix}$, the set $G(\delta, \gamma)$ is right-invariant for the RVSS.

Proof. Let us first show that the set $G(\delta, \gamma)$ is right-invariant for those solutions of the RVSS for which at every $t \geqq t_0 |\sigma_1(t)| \neq 0$. To make this true, it is sufficient that the condition holds:

$$\text{if} \quad |\omega| = 1, \quad \text{then} \quad \operatorname{sgn} \omega \cdot \dot{\omega} < 0 . \tag{7.3}$$

Because

$$\dot{\omega}\Big|_{|\omega|=1} = \frac{\dot{\sigma}_2 - \delta \dfrac{\langle \sigma_1, \dot{\sigma}_1 \rangle}{|\sigma_2|}}{\delta|\sigma_1| + \gamma}$$

and $\operatorname{sgn} \sigma_2 = \operatorname{sgn} \omega$, condition (7.3) for the RVSS holds whenever $\operatorname{sgn} \sigma_2(\gamma u_0 + \Phi) - -\delta|H * x| < 0$. The truth of this follows from condition (7.2).

Now let us consider in more detail the behaviour of solutions of the RVSS which goes through the points $|\sigma_1|=0$, $|\sigma_2|=\gamma$. Let us show that these paths stay within the set $G(0, \gamma)$. It is easily seen that in this case $\operatorname{sgn}\sigma_2 \cdot \dot{\sigma}_2\Big|_{|\sigma_2|=\gamma} < 0$. But $G(0, \gamma) \subseteq G(\delta, \gamma)$. This proves the Theorem.

Remark 4. Condition (7.1) is satisfied, for instance, by the function

$$\Gamma(t)=\omega(t)\left(\frac{\omega^2(t)+1}{2}\right)_4^{-\frac{1}{2}}.$$

Now it can be shown that with $\gamma>0$ and $\delta>0$ small enough, the asymptotic behaviours of solutions of the RVSS and VSS are similar.

Theorem 4. With the vector $h \in \mathfrak{R}^{1\times n}$ chosen in compliance with Remark 1 and $\delta \geqq 0$ and $\gamma>0$ are small enough, any solution $x(t)$ of the RVSS in $\mathscr{X}_1$ and

$$|x(t)| \leqq N \exp\left[(-\lambda+o(\delta))t\right]+o(\gamma),$$

where N is a constant dictated by the specific solution.

Proof. First let us show that any solution of RVSS belongs to the class $\mathscr{X}_1$. It follow from the inequality $\operatorname{sgn}\sigma_2 \cdot \dot{\sigma}_2 < -\lambda|\sigma_2|+|\rho|$, (7.2) and from the fact that if $\sigma(t) \in G(\delta, \gamma)$ then $|\sigma_2(t)|>\gamma$. Because $\tilde{h} \in \mathfrak{R}^{1\times(n-1)}$, with additional state variables $\sigma_0(t)$ equation (4.2a) transforms into

$$\dot{v}=L * v+R * \sigma_2, \tag{7.4}$$

where

$$v(t)=\begin{bmatrix}\sigma_0(t)\\ \hline \sigma_1(t)\end{bmatrix}\begin{matrix}\}m-n+1\\ \}n-1\end{matrix} \in \left(L_+^{\mathrm{loc}}\right)^m, \quad m \geqq n-1, \quad L \in \mathscr{I}^{m\times n}, \quad R \in \mathscr{I}^{m\times 1}.$$

Without limiting the generality $\operatorname{supp} R \subseteq \operatorname{supp} L$. Let $T(t): C \to C$ be the shift operator [3] of the equation

$$\dot{v}=L * v.$$

Following Remark 1, choose the vector $\tilde{h}$ so that

$$\|T(t)\| \leqq N \exp(-\lambda t). \tag{7.5}$$

The general solution of equation (7.4) is [3]

$$v(t-\theta)=T(t-t_0)v(t-\theta)+\int_{t_0}^{t} T(t-s)X_0(\theta)(R * \sigma_2)(s)ds, \tag{7.6}$$

where

$$X_0(\theta)=\begin{cases}0; & \text{if} \quad 0 \leqq \theta \leqq -\max\{t \in \operatorname{supp} L\}\\ E_m; & \text{if} \quad \theta=0.\end{cases}$$

Let us take up an arbitrary solution of system (7.4) with

$$t \geqq t' + \max \{t \in \operatorname{supp} L\} \triangleq \bar{t}\,.$$

Denote

$$\|v_t\| = \max_{0 < \theta \leqq \max\{t \in \operatorname{supp} L\}} |v(t-\theta|\,.$$

Then, because at $t \geqq \bar{t}$ $\sigma(t) \in G(\delta, \gamma)$, hence

$$\max_{0 < \theta \leqq \max\{t \in \operatorname{supp} L\}} |(R * \sigma_2)(t-\theta)| \leqq q\|v_t\| + c \tag{7.7}$$

where $q > 0$ and $c > 0$ are constant while $q = o(\delta)$ and $c = o(\gamma)$. By virtue of the Gronwall–Bellman inequality it follows from relations (7.5)–(7.7) that at every $t \geqq \bar{t}$

$$\|v_t\| \leqq \bar{N} \exp\left[(-\lambda + o(\delta))t\right] + o(\gamma)\,,$$

where $\bar{N} > 0$. Finally, because

$$|v(t)| \leqq \|v_t\|, \quad |\sigma_1(t)| \leqq |v(t)|\,,$$

$$|(\tilde{h} * \sigma_1)(t)| \leqq |v(t)|, \; |\sigma_2(t)| \leqq \delta|\sigma_1(t)| + \gamma\,,$$

from the relation

$$x(t) = \left(\begin{bmatrix} E_{n-1} \\ \hline \tilde{h} \end{bmatrix} * \sigma_1\right)(t) + (g * (h * g)^{-1} * \sigma_2)(t)$$

follows truth of the Theorem.

8. Examples

1. It is required to stabilize a parametrically uncertain system

$$\dot{x}_1 = x_2(t\text{-}\tau)$$

$$\dot{x}_2 = (1 + n_\tau(t))x_1(t-\tau) + u(t)\,,$$

where at every $t \geqq t_0$ $|n_\tau(t)| \leqq n^0$. According to (3.1) these propositions have the form

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & \delta_\tau \\ \delta_\tau & 0 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \delta_0 \end{bmatrix} * (u + \varphi)\,, \tag{8.1}$$

where at every $t \geqq t_0$, $\varphi(t, x) = n_\tau(t) \cdot x_1(t-\tau)$. Assumptions I–IV are easily seen to hold for system (8.1). Thus, spectral controllability follows from the fact that

$$\operatorname{rank} L(s) = \operatorname{rank} \begin{bmatrix} s & -e^{-s\tau} & 0 \\ -e^{-s\tau} & s & 1 \end{bmatrix} = 2 \quad \forall s \in \mathbf{C}\,.$$

Let us assume the desired polynomial $p^2+\alpha_2 * p+\alpha_1$. By Theorem 1 choose

$$h=[k_1 * (p+k_2)^{-1}, \delta_0],$$

where

$$k_1=\alpha_1, \quad k_2=\alpha_2+\alpha_1 * (\delta_0-\delta_\tau) * p^{-1}.$$

Consequently,

$$\sigma_2=k_1 * (p+k_2)^{-1} * x_1+x_2.$$

By Theorem 2 the VSS is exponentially stable with an exponent $\lambda_0>0$ if the roots of the characteristic polynomial $\lambda^2+\alpha_2\lambda+\alpha_1=0$ meet the condition $\operatorname{Re}\lambda<-\lambda_0$ and

$$u=u_1+u_2$$

where

$$u_1=-(\lambda_0 k_1 * (p+k_2)^{-1}+\delta_\tau) * x_1-(k_1 * (p+k_2)^{-1} * \delta_\tau+\lambda_0\delta_\tau) * x_2,$$

$$u_2=-n^0 \operatorname{sgn}\sigma_2 \cdot |x_1(t-\tau)|.$$

2. In the system

$$\dot{x}_1=x_2(t-\tau)$$

$$\dot{x}_2=x_3(t)$$

$$\dot{x}_3=(1+n(t))x_1(t)+(1+n_\tau(t))x_2(t-\tau)+u(t)$$

at every $t\geqq t_0$ $|n(t)|\leqq n^0$, $|n_\tau(t)|\leqq n_\tau^0$.

In the form of (3.1) this system is

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & \delta_\tau & 0\\ 0 & 0 & \delta_0\\ \delta_0 & \delta_\tau & 0\end{bmatrix} * \begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ \delta_0\end{bmatrix} * (u+\varphi), \tag{8.2}$$

where

$$\varphi(t, x)=n(t)x_1(t)+n_\tau(t)x_2(t-\tau).$$

For system (8.2) Assumptions I–IV are true. Assume the desired polynomial $p^2+\alpha_2 * p+\alpha_1$. Then, in compliance with Theorem 1 choose

$$h=[\alpha_1, \alpha_2+\alpha_1 * (\delta_0-\delta_\tau) * p^{-1}, \delta_0]$$

Then

$$\sigma_2=\alpha_1 * x_1+\alpha_2 * x_2+\alpha_1 * (\delta_0-\delta_\tau) * p^{-1} * x_2+x_3.$$

By virtue of Theorem 2 the VSS is exponentially stable with an exponent $\lambda_0>0$ if

the roots of the characteristic polynomial $\lambda^2+\alpha_2\lambda+\alpha_1=0$ satisfy the condition $\operatorname{Re}\lambda<-\lambda_0$ and

$$u=u_1+u_2$$

$$u_1=(-\alpha_1\lambda_0+\alpha_2\delta_\tau+\alpha_1\delta_\tau * (\delta_0-\delta_\tau) * p^{-1}) * x_1-(\alpha_2\lambda_0+\alpha_1\lambda_0 * (\delta_0-\delta_\tau) *$$

$$* p^{-1}+\delta_0) * x_2-(\alpha_1+\alpha_2\delta_\tau+\alpha_1\delta_\tau * (\delta_0-\delta_\tau) * p^{-1}+\lambda_0) * x_3$$

$$u_2=(n_0|x_1(t)|+n_\tau^0|x_2(t-\tau)|)\operatorname{sgn}\sigma_2\,.$$

References

1. *Янушевский Р. Т.* Управление объектами с запаздыванием. М.: Наука, 1978.
2. *Осипов Ю. С.* О стабилизации управляемых систем с запаздыванием. Дифференциальные уравнения. 1965, **1,** *5,* 605–618.
3. *Hale, J. K.*, Theory of Functional Differential Equations. New York: Springer Verlag, 1977.
4. *Красовский Н. Н.* К аналитическому конструированию оптимального управления в системах с запаздыванием. Прикладная математика и механика, **26,** 1962, 50–69.
5. *Kamen, E. W.*, An Operator Theory of Linear Functional Differential Equations. Journal of Differential Equations, **27**, 1978, 274–297.
6. *Watanabe, K., Ito, M., Kaneko, M.*, Finite Spectrum Assignment Problem of Systems with Multiple Commensurate Delays in States and Control. International Journal of Control, **39,** 1984, *5*, 1073–1082.
7. *Kamen, E. W., Khargonekar, P., Tannenbaum, A.*, Stabilization of Time-Delay Systems Using Finite-Dimensional Compensators. IEEE Transactions on Automatic Control, **30,** 1985, 75–79.
8. *Разумихин Б. С.* Применение метода Ляпунова к задачам устойчивости систем с запаздыванием. Автоматика и телемеханика, **21,** 1960, 740–748.
9. *Колмановский В. Б., Носов, В. Р.* Устойчивость и периодические режимы регулируемых систем с последействием. М.: Наука, 1981.
10. *Цыкунов А. М.* Адаптивное управление объектами с последействием. М.: Наука, 1984.
11. *Hasnul-Bacher, A. M., Mukumdan, R.*, Model Reference Control of Uncertain Systems with Time Delay in Plant State and Control. International Journal of System Science, **18,** 1987, *9*, 1609–1626.
12. *Yu, Y.* On Stabilizing Uncertain Linear Delay Systems. Journal of Optimization Theory and Applications, **41,** 1983, *3*, 503–507.
13. Теория систем с переменной структурой. Под. ред. С. В. Емельянова. М.: Наука, 1970.
14. *Емельянов С. В., Коровин С. К., Мамедов И. Г.* Метод квазирасщепления и его применение для синтеза систем автоматического управления. ДАН СССР, **286,** 1986, *2*, 311–315.
15. *Emelyanov, S. V.*, Binary Automatic Control System. Moscow: Mir Publishers, 1987.
16. *Филиппов А. Ф.* Дифференциальные уравнения с разрывной правой частью. М.: Наука, 1985.

Стабилизация неопределенных динамических объектов с запаздыванием в классе бинарных систем управления

С. В. ЕМЕЛЬЯНОВ, С. К. КОРОВИН, И. Г. МАМЕДОВ, А. Л. НЕРСИСЯН

(Москва)

Решается задача стабилизации неопределенных динамических систем с запаздыванием, представленных линейными дифференциальными уравнениями над сверточными кольцами. Рассмотрен случай, когда неопределенность равномерно ограничена и действует в том же подпространстве, что и управление. Предложены алгоритмы разрывного управления, приведены условия

возникновения скользящих режимов. Указаны необходимые и достаточные условия, при выполнении которых рассматриваемые системы в скользящем режиме имеют наперед заданный спектр. Проведена регуляризация разрывных управлений в классе непрерывных функций. Даются иллюстративные примеры.

С. В. Емельянов
С. К. Коровин
И. Г. Мамедов
А. Л. Нерсисян
Всесоюзный научно-исследовательский институт системных исследований АН СССР, 117312, Москва, просп. 60-летия Октября, 9.

Problems of Control and Information Theory, Vol. 18 (3), pp. 151—160 (1989)

THE MAXIMUM PRINCIPLE AND THE SUPERDIFFERENTIAL OF THE VALUE FUNCTION

N. N. SUBBOTINA

(Sverdlovsk)

(Received February 10, 1988)

In this paper the important connection between the Pontryagin maximum principle and the Bellman dynamic programming is established. It is proved that adjoint variables appearing in the maximum principle conditions are generalized gradients of the value function, which is the "viscosity" solution of the Hamilton–Jacobi–Bellman equation. This relationship completes the maximum principle conditions to the necessary and sufficient optimality conditions. The paper was initiated by the recent results of F. Clarke and R. Vinter [25] who had obtained the similar connection as a necessary optimality condition.

1. Introduction

The Pontryagin maximum principle [1] gives the major approach to derive necessary conditions for the optimal controlled processes. To obtain sufficient optimality conditions one can use the specificity of the problem keeping in mind the structure of the controlled system, the structure of the cost functional and the restrictions, properties of families of the extremals constructed with the help of the maximum principle, etc. (see, e.g. [3, 5, 9, 14, 22]). Another way to obtain sufficient conditions leads to the Bellman dynamic programming and to constructions of global estimates for the value function. This function of initial time and state is defined as the infimum cost of an optimal control problem (see [6, 9, 11, 16, 18, 19, 21–26]). The value function satisfies a partial differential equation of the Hamilton–Jacobi type with the non-linearity of the types "min" or "max" [2]. This so-called Bellman equation has no classical smooth solutions as a rule [1, 8, 11, 18, 22]. The value function is a generalized solution of the Bellman equation. It is the "viscosity" solution [8, 19, 21–23]. The value function belongs to the class of quasi-differentiable functions [10, 12, 15, 21], which have a nonempty set of complete generalized gradients at any interior point of their domain. This set is called the superdifferential [20, 23], and its elements are called the supergradients. The superdifferential coincides with the Clarke subdifferential [15] in the considered problem.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

The aim of this paper is to relate the complete supergradients of the value function and the complete adjoint variables of the maximum principle conditions and thereby to clarify the interconnection between the dynamic programming and the maximum principle. Theorem 4 of the present paper states that this relationship together with the maximum principle conditions gives the necessary and sufficient optimality conditions in the considered Mayer problem. This result was achieved by the further development of ideas of [21, 24, 25]. The mathematical tool of the generalized control theory [4, 7, 13, 17] is applied in the construction below.

2. Statement of the problem

Let us consider the following free final state optimal control problem. Minimize the cost functional

$$\gamma = \gamma_{t_0, x_0}(x(\cdot), u(\cdot)) = \sigma(x(\vartheta)) \tag{1}$$

subject to $(x(\cdot), u(\cdot)) : [t_0, \vartheta] \to R^n \times R^P$,

$$\dot{x}(t) = f(t, x(t), u(t)) \qquad \text{a.e.} \quad t \in [t_0, \vartheta] \tag{2}$$

$$u(t) \in P \qquad \text{a.e.} \quad t \in [t_0, \vartheta] \tag{3}$$

$$x(t_0) = x_0 . \tag{4}$$

Here $x \in R^n$ is a state vector, $u \in R^P$ is a control parameter, whose values belong to the compact $P \in R^P$; ϑ is the fixed final time, $(t_0, x_0) \in (-\infty, \vartheta] \times R^n$ are the given initial time and state.

It is supposed that

1) functions $f(\cdot) : (-\infty, \vartheta) \times R^n \times P \to R^n$, $\sigma(\cdot) : R^n \to R$ are continuous;

2) there exist their derivatives

$$\frac{\partial f}{\partial t}(\cdot) : (-\infty, \vartheta) \times R^n \times P \to R^n ,$$

$$\frac{\partial f}{\partial x}(\cdot) : (-\infty, \vartheta) \times R^n \times P \to R^{n \times n} ,$$

$$\frac{\partial \sigma}{\partial x}(\cdot) : R^n \to R$$

and these functions are continuous;

3) an extendability condition for trajectories $x(\cdot)$ of system (2)–(4) is fulfilled.

We denote by U_{t_0} the set of Borel measurable control functions $u(\cdot) : [t_0, \vartheta] \to R^n$ under restrictions (3). As follows from assumptions 1)–3) there is the unique trajectory

$x(\cdot):[t_0, \vartheta]\to R^n$ of system (2)–(4) for each given control function $u(\cdot)\in U_{t_0}$. The pair $(x(\cdot), u(\cdot))$ is called an admissible controlled process for the initial position (t_0, x_0) if $u(\cdot)\in U_{t_0}$ and $x(\cdot)$ is the trajectory of system (2)–(4) corresponding to this control function. The symbol $D(t_0, x_0)$ denotes the set of all admissible controlled processes.

The number

$$\omega^0=\omega^0(t_0, x_0)=\inf_{(x(\cdot),u(\cdot))\in D(t_0,x_0)}\gamma_{t_0,x_0}(x(\cdot), u(\cdot)) \tag{5}$$

is the optimal result or the value for the given initial position (t_0, x_0) in problem (1)–(4). The function $(t, x)\to\omega^0(t, x):(-\infty, \vartheta]\times R^n\to R$ is called the optimal result function or the value function. The admissible process $(x^0(\cdot), u^0(\cdot))$ is optimal for the given initial position (t_0, x_0) if

$$\gamma_{t_0,x_0}(x^0(\cdot), u^0(\cdot))=\omega^0(t_0, x_0)=\min_{(x(\cdot),u(\cdot))\in D(t_0,x_0)}\gamma_{t_0,x_0}(x(\cdot), u(\cdot)). \tag{6}$$

To provide the existence of the optimal process we impose the following conditions

4) the sets

$$F(t, x)=\{f(t, x, u): u\in P\} \tag{7}$$

are convex at any $(t, x)\in(-\infty, \vartheta]\times R^n$.

3. The maximum principle

Let us formulate the necessary optimality conditions for admissible processes $(x(\cdot), u(\cdot))\in D(t_0, x_0)$ in the considered problem (1)–(4) with the help of the Pontryagin maximum principle [1].

Theorem 1. If $(x^0(\cdot), u^0(\cdot))$ is the optimal process for the initial position (t_0, x_0) and if $(\lambda^0(\cdot), \psi^0(\cdot)):[t_0, \vartheta]\to R\times R^n$ is the absolutely continuous vector function satisfying the adjoint system

$$\frac{d\lambda^0}{dt}(t)=-\left\langle\frac{\partial f}{\partial t}(t, x^0(t), u^0(t)), \psi^0(t)\right\rangle \qquad \text{a.e. } t\in[t_0, \vartheta] \tag{8}$$

$$\frac{d\psi^0}{dt}(t)=-\left(\frac{\partial f}{\partial x}(t, x^0(t), u^0(t))\right)^T\psi^0(t) \qquad \text{a.e. } t\in[t_0, \vartheta] \tag{9}$$

and the boundary conditions

$$\lambda^0(\vartheta)=-\min_{u\in P}\langle f(\vartheta, x^0(\vartheta), u), \psi^0(\vartheta)\rangle \tag{10}$$

$$\psi^0(\vartheta)=\frac{\partial\sigma}{\partial x}(x^0(\vartheta)) \tag{11}$$

then the following equality holds

$$\lambda^0(t) \cdot 1 + \langle \psi^0(t),\ f(t,\ x^0(t),\ u^0(t)) \rangle = 0 \qquad \text{a.e.} \quad t \in [t_0,\ \vartheta]\,. \tag{12}$$

Here symbol T denotes transposition, the symbol $\langle a,\ b \rangle$ denotes the inner product of the vectors a and b.

Note that the solution $\lambda^0(\cdot)$ of equation (8) with boundary condition (10) has the form

$$\lambda^0(t) = -\min_{u \in P} \langle f(t,\ x^0(t),\ u),\ \psi^0(t) \rangle \qquad \text{for all} \quad t \in [t_0,\ \vartheta]\,. \tag{13}$$

Variables λ, ψ are called the adjoint variables. A controlled process $(x^*(\cdot),\ u^*(\cdot)) \in$ $\in D(t_0,\ x_0)$ satisfying conditions (8)–(12) is called the extremal one. As can be seen, the velocity $v^*(t) = (1,\ f(t,\ x^*(t),\ u^*(t)))$ of the point $(t,\ x^*(t))$ with the extremal dynamics is orthogonal to the corresponding adjoint variables vector $s^*(t) = (\lambda^*(t),\ \psi^*(t))$ on the time interval $[t_0,\ \vartheta]$. From the definitions of the value function $\omega^0(t,\ x)$ and the optimal process $(x^0(\cdot),\ u^0(\cdot))$ one can obtain that the value function is constant along an optimal trajectory, i.e. $\omega^0(t,\ x^0(t)) \equiv \text{const}$ for all $t \in [t_0,\ \vartheta]$. It means that the velocity $v^0(t) =$ $= (1,\ f(t,\ x^0(t),\ u^0(t)))$ of the point $(t,\ x^0(t))$ with the optimal dynamics belongs to the corresponding tangent plane of the value function level surface. According to (12) the adjoint variables vector $s^0(t) = (\lambda^0(t),\ \psi^0(t))$ coincides with a generalized gradient of the non-smooth value function. We shall give a rigorous proof of the relationship between the adjoint variables and the generalized gradients in Theorem 4.

4. The superdifferential and the quasi-differentiability of the value function

Let us remind the definition of the superdifferential [20, 23] which will be used in the necessary and sufficient optimality conditions below.

Definition 1. The superdifferential $\partial\omega(t,\ x)$ of a function $\omega(\cdot) : (-\infty,\ \vartheta] \times R^n \to R$ at the point $(t,\ x)$ is the set

$$\partial\omega(t,\ x) = \Big\{ (p^0,\ p) \in R \times R^n : \limsup_{\substack{\tau \to t \\ y \to x}} [(|\tau - t| + \|y - x\|)^{-1} \times$$

$$\times (\omega(\tau,\ y) - \omega(t,\ x) - p^0(\tau - t) - \langle p,\ (y - x) \rangle)] \leqq 0 \Big\}\,. \tag{14}$$

To obtain the expression for the superdifferential of the value function $\omega^0(\cdot)$ (5) one can use the following assertion [21].

Theorem 2. The value function $\omega^0(\cdot)$ (5) in the considered problem (1)–(4) under assumptions 1)–4) belongs to the class Ω of quasi-differentiable [10, 12] functions

$\omega(\cdot):(-\infty, \vartheta]\times R^n\to R$ of the following form

$$\omega^0(t, x)=\min_{\alpha\in A}\varphi(t, x, \alpha) \qquad \text{for} \quad t\leqq\vartheta,\ x\in R^n \tag{15}$$

Here α is a parameter, A is a metric compactum, $\varphi(\cdot):(-\infty, \vartheta]\times R^n\times A\to R$ is a continuous function, which has continous partial derivatives $\partial\varphi(\cdot)/\partial t$, $\partial\varphi(\cdot)/\partial x$: $(-\infty, \vartheta)\times R^n\times A\to R\times R^n$.

Remark 1. One can take admissible control functions $u(\cdot)$ for prototypes of parameters α in (15). To be more precise, one should take generalized control functions, which are measurable on the standard interval [0, 1] with values in the set of regular probability measures on P (see [7, 13, 17]). If the values of a generalized control function α are point-concentrated measures then the linear transformation $[0, 1]\to$ $\to[t, \vartheta]:\tau\to\xi$ of the form $\xi=t+(\vartheta-t)\tau$ turns this control α into an admissible control $u_t(\cdot)\in U_t$. So, for α of this type and corresponding $u_t(\cdot)$ the sense of the function $(t, x)\to\varphi(t, x, \alpha)$ is given by

$$\varphi(t, x, \alpha)=\gamma_{t,x}(x(\cdot), u(\cdot))=u_t(x(\vartheta; t, x, u_t(\cdot))) \tag{16}$$

where $x(\xi)=x(\xi; t, x, u_t(\cdot))$, $t\leqq\xi\leqq\vartheta$ is the solution of the problem

$$\frac{dx}{d\xi}(\xi)=f(\xi, x(\xi), u_t(\xi)) \qquad \text{a.e.} \quad \xi\in[t, \vartheta],$$

$$x(t)=x, \qquad u_t(\cdot)\in U_t. \tag{17}$$

From Theorem 2 it follows that the function $\omega^0(\cdot)$ (5) is locally Lipschitz continuous. It is differentiable in any direction $(\tau, f)\in R\times R^n$. As follows from [10, 12, 15] the expression for directional derivative has the form

$$D\omega^0(t, x)|(\tau, f)=\lim_{\lambda\downarrow 0}\lambda^{-1}\cdot[\omega^0(t+\lambda\tau, x+\lambda f)-\omega^0(t, x)]=$$

$$=\min_{(p^0, p)\in d\omega^0(t,x)}[\tau\cdot p^0+\langle f, p\rangle], \qquad t<\vartheta,\ x\in R^n. \tag{18}$$

The set $d\omega^0(t, x)$ is defined as

$$d\omega^0(t, x)=\left\{p^0=\frac{\partial\varphi}{\partial t}(t, x, \alpha^0),\ p=\frac{\partial\varphi}{\partial x}(t, x, \alpha^0):\varphi(t, x, \alpha^0)=\omega^0(t, x)\right\}. \tag{19}$$

The mapping $(p^0, p)\to\langle(\tau, f), (p^0, p)\rangle$ in (18) is linear, hence the directional derivative $D\omega^0(t, x)|(\tau, f)$ can also be given by

$$D\omega^0(t, x)|(\tau, f)=\min_{(p^0, p)\in\operatorname{co} d\omega^0(t,x)}\langle(\tau, f), (p^0, p)\rangle, \qquad t<\vartheta,\ x\in R^n \tag{20}$$

where symbol co S denotes the convex hull of set S.

The following assertions can be proved by using relations (15), (18)–(20) and definition 1.

Lemma 1. The sets $d\omega^0(t, x)$ and co $d\omega^0(t, x)$ are nonempty and compact at any $(t, x) \in (-\infty, \vartheta) \times R^n$. The multivalued mappings $(t, x) \to d\omega^0(t, x)$, $(t, x) \to$ co $d\omega^0(t, x)$ are upper semicontinuous with respect to inclusion.

Lemma 2. The superdifferential $\partial\omega^0(t, x)$ (14) of the value function $\omega^0(\cdot)$ (5) has the form

$$\partial\omega^0(t, x) = \text{co } d\omega^0(t, x), \qquad t < \vartheta, \ x \in R^n. \tag{21}$$

Remark 2. As follows from (15), (18)–(21), and results of [15] the superdifferential of the value function $\partial\omega^0(t, x)$ coincides with the F. Clarke subdifferential $\partial_{\text{Cl}}\omega^0(t, x)$ which is defined in the following way

$$\partial_{\text{Cl}}\omega^0(t, x) = \text{co}\left\{(p^0, p) \in R \times R^n : (p^0, p) = \right.$$

$$\left. = \lim_{\substack{t_k \to t \\ x_k \to x}} \left(\frac{\partial\omega^0}{\partial t}(t_k, x_k), \frac{\partial\omega^0}{\partial x}(t_k, x_k) \right) \right\}. \tag{22}$$

Here points (t_k, x_k) are chosen from an almost everywhere dense set Λ on $(-\infty, \vartheta) \times R^n$. The set Λ is the region of strict differentiability of the quasi-differentiable function $\omega^0(\cdot)$. In the general case of Lipschitz continuous function $\omega(t, x)$ the following inclusion is fulfilled

$$\partial\omega(t, x) \subset \partial\omega_{\text{Cl}}(t, x).$$

5. The dynamic programming

The dynamic programming method interprets the value function $\omega^0(\cdot)$ (5) as a solution of the following Cauchy problem

$$\frac{\partial\omega}{\partial t}(t, x) + \min_{u \in P} \left\langle \frac{\partial\omega}{\partial x}(t, x), f(t, x, u) \right\rangle = 0, \tag{23}$$

$$t < \vartheta, \quad x \in R^n;$$

$$\omega(\vartheta, x) = \sigma(x), \quad x \in R^n,$$

(see [1, 2]). Equation (23) of the Hamilton–Jacobi type is the Bellman equation for the optimal control problem (1)–(4). As a rule, this Cauchy problem has no classical solution. The quasi-differentiable value function $\omega^0(\cdot)$ (5) is a generalized solution [19, 21]. It is the so-called "viscosity" solution of this problem [23]. The following assertion is valid [21].

Theorem 3. A function $\omega(\cdot):(-\infty, \vartheta]\times R^n\to R$ coincides with the value function $\omega^0(\cdot)$ (5) iff

$$\omega(\cdot)\in\Omega, \tag{25}$$

$$\omega(\vartheta, x)=\sigma(x), \qquad \text{for all} \quad x\in R^n, \tag{26}$$

$$\min_{f\in F(t,x)} D\omega(t, x)|(1, f)=0, \qquad \text{for all} \quad t<\vartheta,\ x\in R^n, \tag{27}$$

where Ω is the class of quasi-differentiable functions (5). Sets $F(t, x)$ are defined by (7).

It is easy to establish that equality (27) turns into the Bellman equation (23) at the differentiability points of the value function.

6. Necessary and sufficient optimality conditions

Let us note the following useful properties of the value function $\omega^0(\cdot)$ (5) (see [11, 21, 24]).

Lemma 3. If $(x(\cdot), u(\cdot))$ is an admissible process for the initial position (t_0, x_0) then the function $t\to\rho[t]=\omega^0(t, x(t))$ is absolutely continuous, monotone non-decreasing when $t\to\vartheta$, and the following condition holds

$$\frac{d\rho[t]}{dt}=D\omega^0(t, x(t))|(1, f(t, x(t), u(t)))\geqq 0 \qquad \text{a.e.} \quad t\in[t_0, \vartheta]. \tag{28}$$

Lemma 4. The admissible process $(x^0(\cdot), u^0(\cdot))$ is optimal for the initial position (t_0, x_0) iff either

$$\rho^0[t]=\omega^0(t, x^0(t))=\omega^0(t_0, x_0)\equiv\text{const} \qquad \text{for all} \quad t\in[t_0, \vartheta] \tag{29}$$

or

$$D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u^0(t)))=$$

$$=\min_{u\in P} D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u)) \qquad \text{a.e.} \quad t\in[t_0, \vartheta]. \tag{30}$$

The main result of this paper is given in the following theorem.

Theorem 4. An admissible process $(x^0(\cdot), u^0(\cdot))\in D(t_0, x_0)$ is optimal for the initial position (t_0, x_0) iff conditions (8)–(12) of Theorem 1 hold and the inclusion

$$(\lambda^0(t), \psi^0(t))\in\partial\omega^0(t, x^0(t)) \tag{31}$$

is valid for all $t\in[t_0, \vartheta]$. Here $\omega^0(t, x)$ is the value function (5).

Proof

Necessity. As follows from Assumptions 1)–3) a trajectory $x(\cdot)$ of the considered system under the generalized control is continuously differentiable with respect to the initial time and state [21]. One can easily prove using this fact, differentiability of $\sigma(\cdot)$ and (16), (17) that adjoint variables (8)–(12) satisfy the equalities

$$\frac{\partial\varphi}{\partial t}(t, x, \alpha)=\lambda(t), \qquad \frac{\partial\varphi}{\partial x}(t, x, \alpha)=\psi(t)$$

where $\alpha\in A$ is obtained from an admissible control function $u_t(\cdot)\in U_t$ as the result of the linear transformation $[t, \vartheta]\to[0, 1]: \xi\to\tau$ of the form $\tau=\dfrac{\xi-t}{\vartheta-t}$; the vector-function $(\lambda(\cdot), \psi(\cdot)): [t, \vartheta]\to R^{n+1}$ is the solution of the adjoint system corresponding to the admissible process $(x(\cdot)=x(\cdot, t, x, u_t(\cdot)), u_t(\cdot))\in D(t, x)$. If $(x^0(\cdot)=x^0(\cdot, t_0, x_0, u^0(\cdot)), u^0(\cdot))\in D(t_0, x_0)$ is the optimal process for (t_0, x_0) then the corresponding adjoint variables $(\lambda^0(\cdot), \psi^0(\cdot))$ satisfy the equalitites (see [24])

$$\frac{\partial\varphi}{\partial t}(t, x^0(t), \alpha_t^0)=\lambda^0(t), \quad \frac{\partial\varphi}{\partial x}(t, x^0(t), \alpha_t^0)=\psi^0(t)$$

$$\text{for all } t\in[t_0, \vartheta] \tag{32}$$

Here α_t^0 is obtained from the optimal control function $u_t^0(\cdot)\in U_t$ as the result of the linear transformation $\xi\to\tau: \tau=\dfrac{\xi-t}{\vartheta-t}$ and also

$$u_t^0(\xi)=u^0(\xi) \qquad \text{for all} \quad \xi\in[t, \vartheta], t\in[t_0, \vartheta] \tag{33}$$

$$\varphi(t, x^0(t), \alpha_t^0)=\omega^0(t, x^0(t)) \qquad \text{for all} \quad t\in[t_0, \vartheta]. \tag{34}$$

It follows from (32), (34), (19) and (21) that

$$(\lambda^0(t), \psi^0(t))=(p^0(t), p(t))\in d\omega^0(t, x^0(t))\subset\partial\omega^0(t, x^0(t)). \tag{35}$$

Sufficiency. Let us consider an admissible process $(x^0(\cdot), u^0(\cdot))\in D(t_0, x_0)$ satisfying conditions (8)–(12), and (31). From (27), (20), (31) and (12) we can obtain the following relations

$$\begin{aligned}
0&=\min_{u\in P} D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u))\leqq\\
&\leqq D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u^0(t)))=\\
&=\min_{(p^0, p)\in\partial\omega^0(t, x^0(t))}[p^0\cdot 1+\langle p, f(t, x^0(t), u^0(t))\rangle]\leqq\\
&\leqq\lambda^0(t)\cdot 1+\langle\psi^0(t), f(t, x^0(t), u^0(t))\rangle=0 \qquad \text{a.e.} \quad t\in[t_0, \vartheta].
\end{aligned} \tag{36}$$

In particular, we have

$$D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u^0(t)))=0 \qquad \text{a.e.} \quad t\in[t_0, \vartheta]. \tag{37}$$

According to Lemma 4, condition (37) (see also (30)) is a sufficient optimality condition for the process $(x^0(\cdot), u^0(\cdot))\in D(t_0, x_0)$. Thus, Theorem 4 is proved.

Remark 3. Theorem 4 remains valid if inclusion (31) is replaced by

$$(\lambda^0(t), \psi^0(t))\in d\omega^0(t, x^0(t)) \qquad \text{for all} \quad t\in[t_0, \vartheta]. \tag{38}$$

Here the set $d\omega^0(t, x)$ is defined by (19). This follows from (18), (20) and (35).

Remark 4. To compare conditions (31) and (38) of Theorem 4 with the corresponding results of [25] we recall that the authors of [25] obtained the following necessary optimality conditions

$$\psi^0(t)\in\partial_x(t, x^0(t)) \qquad \text{for all} \quad t\in[t_0, \vartheta], \tag{39}$$

where

$$\partial_x\omega^0(t, x)=\operatorname{co}\left\{p=\lim_{\substack{t_k\to t\\ x_k\to x}}\frac{\partial\omega^0}{\partial x}(t_k, x_k):(t_k, x_k)\in\Lambda\right\}, \tag{40}$$

and Λ is the set of differentiability of the value function $\omega^0(\cdot)$. Inclusion (31) and Remark 2 incorporate the Clarke–Vinter conditions (39)–(40).

Remark 5. We have proved (see [26]) similar results in problem (1)–(4) where the cost functional γ (1) was replaced by γ^*

$$\gamma^*=\gamma^*_{t_0, x_0}(x(\cdot), u(\cdot))=\sigma(x(\vartheta))+\int_{t_0}^{\vartheta} f^0(t, x(t), u(t))dt$$

References

1. *Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., Mishchenko, E. F.*, The mathematical theory of optimal processes. Fizmatgiz, Moscow, 1961.
2. *Bellman, R. E.*, Dynamic programming. Princeton Univ. Press, Princeton, New York, 1957.
3. *Krasovskii, N. N.*, On a problem of optimum control of nonlinear systems. Prikl. mat. i meh., 1959, **23**, *2*, pp. 209–229.
4. *Filippov, A. F.*, On some questions of the theory of optimal control. Vestnik Mosk. Gos. Univ., Ser. mat., meh., astron., fiz., him., 1959, *2*, pp. 25–32.
5. *Rozonoer, L. I.*, The Pontryagin maximum principle in the optimal systems theory, III. Avtomat. i telemeh., 1959, **20**, *12*, pp. 1561–1578.
6. *Krotov, V. F.*, Methods for solving of variational problems on the basis of the sufficient conditions for an absolute minimum, I. Avtomat. i telemeh., 1962, **23**, *12*, pp. 1571–1583.
7. *Gamkrelidze, R. V.*, On sliding optimal states. Dokl. Akad. Nauk SSSR, 1962, **143**, *6*, pp. 1243–1246.
8. *Kružkov, S. N.*, Generalized solutions of nonlinear first order equation with many independent variables, I. Matem. Sbornik, 1966, **70**, *3*, pp. 394–415.
9. *Boltyanskii, V. G.*, Mathematical methods of optimal control. Nauka, Moscow, 1969.
10. *Pshenichniy, B. N.*, Necessary conditions of extremum. Nauka, Moscow, 1969.

11. *Leitman, G.*, An introduction to optimal control. McGraw-Hill Book Company, New York, 1966.
12. *Dem'yanov, V. F., Malozemov, V. N.*, On the theory of nonlinear minmax problems. Uspehi mat. nauk, 1971, **26**, *3*(159), pp. 53–104.
13. *Young, L. C.*, Lectures on the calculus of variations and optimal control theory. W. B. Saunders Company, Philadelphia, London, Toronto, 1969.
14. *Blagodatskh, V. T.*, Sufficient optimality conditions for differential inclusions. Izv. Akad. Nauk SSSR. Ser. matem., 1974, **38**, *3*, pp. 615–624.
15. *Clarke, F. H.*, Generalized gradients and applications. Trans. Amer. Math. Soc., 1975, **205**, *2*, pp. 247–262.
16. *Gabasov, R., Kirillova, F. M.*, Foundations of dynamic programming. Izdat. Belorus. Univ., Minsk, 1975.
17. *Warga, J.*, Optimal control of differential and functional equations. Acad. Press, New York, London. 1972.
18. *Fleming, W. H., Rishel, R. W.*, Deterministic and stochastic optimal control. Springer, New York, 1975.
19. *Hrustalev, M. M.*, Necessary and sufficient optimality conditions in the form of the Bellman equation. Dokl. Akad. Nauk SSSR, 1978, **242**, *5*, pp. 1023–1026.
20. *De Giorgi, E., Marino, A., Tosques, M.*, Problemi di evoluzione in spazi metrici e curve di massima pendenza. Rend. Cl. Sci. Fis. Mat. Accad. Naz., Lincei, 1980, **68**, pp. 180–187.
21. *Subbotin, A. I., Subbotina, N. N.*, On justification of dynamic programming method in an optimal control problem. Izv. Acad. Nauk SSSR, Tehn. kibernet., 1983, no.*2*, pp. 24–32.
22. *Vyazgin, V. A.*, On justification of sufficient conditions in the Weierstrass and Hamilton–Jacobi–Bellman methods. Avtomat. i telemeh., 1984, *4*, pp. 31–37.
23. *Crandall, M. C., Evans, L. C., Lions, P. L.*, Some properties of viscosity solutions of Hamilton–Jacobi equations. Trans. Amer. Math. Soc., 1984, **282**, *2*, pp. 487–502.
24. *Subbotina, N. N.*, Necessary and sufficient optimality conditions for controls and trajectories. A synthesis of an optimal control in game systems. Trans. Inst. Math. Mech., Izd. Ural Centre Acad. Sci. USSR, Sverdlovsk, 1986, pp. 86–96.
25. *Clarke, F. H., Vinter, R. B.*, The relationship between the maximum principle and dynamic programming. SIAM J. Control and Optimiz., 1987, **25**, *5*, pp. 1291–1311.
26. *Subbotina, N. N.*, Necessary and sufficient optimality conditions in terms of the maximum principle and the superdifferential of the value function, Inst. Math. Mech. Ural Branch of Acad. of Sci. USSR, Sverdlovsk, 1988, date dep. VINITI 15.04.88, no.2898-B88.

Принцип максимума и супердифференциал функции цены

Н. Н. СУББОТИНА

(Свердловск)

В работе рассматриваются задачи оптимального управления со свободным правым концом, нелинейной динамикой и терминальным функционалом качества управляемого процесса. Функция, которая фиксированному начальному состоянию процесса ставит в соответствие минимальное значение функционала качества, называется функцией цены. Эта функция является обобщенным, «вязким» решением основного уравнения метода динамического программирования Р. Беллмана. Функция цены в рассматриваемой задаче квазидифференцируема, ее супердифференциал — непустое множество, совпадающее с субдифференциалом Ф. Кларка во всех внутренних точках области определения функции цены.

Показано, что супердифференциал содержит вектор сопряженных переменных из условий принципа максимума Л. С. Понтрягина. Это включение дополняет условия принципа максимума до необходимых и достаточных условий оптимальности управляемых процессов.

Н. Н. Субботина
Институт математики и механики УрО АН СССР
СССР, 620219, Свердловск, ГСП-384,
ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 18 (3), pp. 161—168 (1989)

MEAN-SQUARE STRATEGIES IN STOCHASTIC DIFFERENTIAL GAMES*

S. D. GAIDOV

(*Plovdiv*)

(Received March 1, 1988)

The paper deals with N-person cooperative games in which the dynamics is described by Ito stochastic differential equations. Mean-square strategies are introduced as a close approximation of the absolutely cooperative strategies. Some properties, including the Pareto-optimality, of these strategies are considered. Sufficient conditions for the mean-square strategies are found.

Introduction

A basic notion of an optimality in cooperative differential games is Pareto-optimality. There are many publications on the topic, mainly in the deterministic case [9]. Pareto-optimal strategies in stochastic differential games were considered in [3].

The Pareto-optimality possesses several properties, among which at least two can be interpreted as negative properties [9]. Firstly, the application of Karlin's lemma and the reduction of the game problem to a singlecriterial optimization imply the ambiguity of the strategies. Secondly, the Pareto-optimal strategies can supply some players with values of their cost-functions greater than the guaranteeing (minimax) strategies can do.

These negative properties can be overcome if we narrow down the set of Pareto-optimal strategies. One possibility is to use the so-called mean-square strategies.

Note that in this paper the mean-square strategies are introduced in relation with the absolutely cooperative strategies [4] and then their Pareto-optimality is established. The results presented here were announced without any proof and details in [5].

* This research was supported in part by the Ministry of Culture, Science and Education under Contract 1006.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Formalization of a stochastic differential game

Let us consider the system (game)

$$\Gamma = \langle I, \Sigma, \{\mathcal{U}_i\}_{i\in I}, \{\mathcal{J}_i\}_{i\in I}\rangle .$$

Here $I=\{1, \ldots, N\}$ is the set of players participating in Γ. The evolution of the dynamic system Σ is desribed by a stochastic differential equation of the type

$$(*)dx(t) = f(t, x(t), u_1, \ldots, u_N)dt +$$

$$+ g(t, x(t), u_1, \ldots, u_N)dw(t), \quad t\in[t_0, T]$$

with the initial condition $x(t_0)=x_0\in\mathbf{R}^n$ and 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 $(\Omega, \mathcal{F}, \mathbf{P})$ and adapted to a family $F=\{\mathcal{F}_t, t\in[t_0, T]\}$ of nondecreasing sub-σ-algebras of $\mathcal{F}$. The vector $x(t)\in\mathbf{R}^n$ is the state process and $u_i\in U_i\subset\mathbf{R}^{n_i}$ is the control of the i-th player, $i\in I$.

Let us make the following assumptions about the functions $f(t, x, u_1, \ldots, u_N)$ and $g(t, x, u_1, \ldots, u_N)$. Suppose that

$$f: [t_0, T]\times\mathbf{R}^n\times U_1\times \ldots \times U_N\to\mathbf{R}^n$$

and

$$g: [t_0, T]\times\mathbf{R}^n\times U_1\times \ldots \times U_N\to\mathbf{R}^n\times\mathbf{R}^m$$

have continuous partial derivatives and let $C>0$ be a constant such that

$$|f(t, 0, \ldots, 0)|+|g(t, 0, \ldots, 0)|\leqq C,$$

$$|f_x|+|g_x|+\sum_{i\in I}(|f_{u_i}|+|g_{u_i}|)\leqq C$$

where $|\cdot|$ is a general symbol for the norm in the respective space.

Every player has perfect observations of the state vector $x(t)$ at every moment $t\in[t_0, T]$ and constructs his strategy in the game Γ as an admissible feedback control [2] of the type

$$u_i = u_i(t, x(t)),$$

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

$$|u_i(t, x)|\leqq M_i(1+|x|) \qquad \text{for all} \quad t\in[t_0, T], x\in\mathbf{R}^n;$$

(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 $\mathcal{U}_i$ the set of strategies of the i-th player, $i \in I$ and $\mathcal{U} = \mathcal{U}_1 \times \ldots \times \mathcal{U}_N$. Let, for brevity, the vector of strategies $u = (u_1, \ldots, u_N)$ be called simply a strategy.

The assumptions made above imply the existence and sample path uniqueness [2] of the solution $X = \{x(t), t \in [t_0, T]\}$ of Ito equation $(*)$ corresponding to the control u. Moreover, X is an a.s. continuous Markov process and let $\mathcal{A}(u)$ be its infinitesimal operator (see [1], ch.V, §11):

$$\mathcal{A}(u)W(t, x) = f'(t, x, u)W_x(t, x) +$$

$$+ \frac{1}{2} \operatorname{tr} [g(t, x, u)g'(t, x, u)W_{xx}(t, x)]$$

where prime denotes the vector or matrix transpose. Here $W(t, x)$ is a real-valued function with continuous partial derivatives up to the second order for all $t \in [t_0, T]$, $x \in \mathbf{R}^n$.

Let us consider the continuous functions L_i, Q_i satisfying the following growth conditions:

$$|L_i(t, x, u_1, \ldots, u_N)| \leqq C_i(1 + |x|^K + \sum_{i \in I} |u_i|^K) ,$$

$$|Q_i(t, x)| \leqq C_i(1 + |x|^K) .$$

Here C_i, K are positive constants. Now we introduce the cost-function $\mathcal{J}_i(u)$ of the i-th player:

$$\mathcal{J}_i(u) = \mathbf{E}_{t_0, x_0}\{Q_i(T, x(T)) + \int_{t_0}^{T} L_i(t, x(t), u_1, \ldots, u_N)dt\}, \quad i \in I .$$

Let us recall that every stochastic differential game has its internal mechanism of development. Each player, e.g. the i-th one, chooses his strategy $u_i \in \mathcal{U}_i$ according to some principle of optimality. Then the vector of strategies $u = (u_1, \ldots, u_N)$ is formed. Further, the solution of Ito equation $X = \{x(t), t \in [t_0, T]\}$ is found. Thus, the value of $\mathcal{J}_i(u)$ corresponding to u and X is determined. The goal of each player in the game Γ is to minimize his cost-function.

Definition and properties of mean-square strategies

Recall the following notion of an optimal strategy in a stochastic differential game.

Definition ([4]). The strategy $u^a=(u_1^a, \ldots, u_N^a)$ is called an absolutely cooperative strategy in the game Γ, if for each $u \in \mathscr{U}$ we have

$$\mathscr{J}_i(u) \geqq \min_{u \in \mathscr{U}} \mathscr{J}_i(u) = \mathscr{J}_i(u^a) = \mathscr{J}_i^a, \quad i \in I .$$

Such strategies are rarely met in game theory. They could be classified as nearly utopic (see [9]). Nevertheless, it is quite natural to consider strategies which in some sense are a close approximation of the absolutely cooperative strategy u^a. Thus we come to the following

Definition. The strategy $u^{ms}=(u_1^{ms}, \ldots, u_N^{ms})$ is called a mean-square strategy in the game Γ, if for each $u \in \mathscr{U}$

$$\sum_{i \in I} [\mathscr{J}_i(u) - \mathscr{J}_i^a]^2 \geqq \sum_{i \in I} [\mathscr{J}_i(u^{ms}) - \mathscr{J}_i^a]^2 .$$

Note that in deterministic game theory these strategies were discussed in [8] and [9].

Let us mention two of the properties of the mean-square strategies.

1. The mean-square strategy u^{ms} supplies the cost-functions with the nearest values (in mean-square distance sense) to the "point of utopia" $\mathscr{J}^a=(\mathscr{J}_1^a, \ldots, \mathscr{J}_N^a)$.

2. The strategy u^{ms} is Pareto-optimal.

Let us suppose that u^{ms} is not Pareto-optimal [3]. Then there exists a strategy $\bar{u} \in \mathscr{U}$ such that the system

$$\mathscr{J}_i(\bar{u}) \leqq \mathscr{J}_i(u^{ms}), \quad i \in I$$

holds where at least one inequality is strict. Hence

$$0 \leqq \mathscr{J}_i(\bar{u}) - \mathscr{J}_i^a \leqq \mathscr{J}_i(u^{ms}) - \mathscr{J}_i^a, \quad i \in I$$

and, therefore

$$\sum_{i \in I} [\mathscr{J}_i(\bar{u}) - \mathscr{J}_i^a]^2 < \sum_{i \in I} [\mathscr{J}_i(u^{ms}) - \mathscr{J}_i^a]^2 ,$$

which obviously contradicts the definition of u^{ms}. So, the meansquare strategies are Pareto-optimal.

Sufficient conditions for the mean-square strategies

First we shall consider the following auxiliary proposition.

Lemma. Let $X=\{x(t), t \in [t_0, T]\}$ be the solution of Ito equation $(*)$ with the initial condition $x(t_0)=x_0$. Then there is a positive constant E_i, such that the following estimate holds:

$$\mathrm{Var}_{t_0, x_0}\{Q_i(T, x(T))\} \leqq E_i, \quad i \in I .$$

Proof. Taking into consideration the properties of conditional expectations, the growth condition of the function Q_i and a result from [6] (see Theorem 4, §6, ch.2, part I) we get

$$\mathrm{Var}_{t_0,x_0}\{Q_i(T, x(T))\} \leqq \mathbf{E}_{t_0,x_0}\{Q_i^2(T, x(T))\} \leqq \mathbf{E}_{t_0,x_0}\{C_i^2(1+|x(T)|^K)^2\} \leqq$$
$$\leqq 2C_i^2\mathbf{E}_{t_0,x_0}\{1+|x(T)|^{2K}\} = 2C_i^2(1+\mathbf{E}_{t_0,x_0}\{|x(T)|^{2K}\}) \leqq$$
$$\leqq 2C_i^2[1+K(1+|x_0|^{2K})] = E_i .$$

Now we give sufficient conditions satisfied by the mean-square strategies. For convenience (see [2], ch.VI, §7) suppose $L_i \equiv 0$, $i \in I$, i.e. we consider our game Γ with terminal cost-functions only:

$$\mathscr{I}_i(u) = \mathbf{E}_{t_0,x_0}\{Q_i(T, x(T))\}, \quad i \in I .$$

Theorem. The strategy u^{ms} is a mean-square strategy in the game Γ, if there exist real-valued functions $W^{(i)}(t, x)$ such that for all $t \in [t_0, T]$, $x \in \mathbf{R}^n$ and $i \in I$ the following conditions simultaneously hold:

(a) $W^{(i)}$, $W_t^{(i)}$, $W_x^{(i)}$, $W_{xx}^{(i)}$ are continuous;

(b) $\sum_{i\in I} 2[W_t^{(i)}(t, x)+\mathscr{A}(u)W^{(i)}(t, x)][W^{(i)}(t, x)-\mathscr{I}_i^a]+$
$+[W_x^{(i)}(t, x)]'g(t, x, u)g'(t, x, u)W_x^{(i)}(t, x) \geqq E$ for each $u \in \mathscr{U}$
where $E=(T-t_0)^{-1}\sum_{i\in I} E_i$. E_i are the constants in the above lemma;

(c) $W_t^{(i)}(t, x)+\mathscr{A}(u^{ms})W^{(i)}(t, x)=0$;

(d) $W^{(i)}(T, x)=Q_i(T, x)$.

Proof. Let $x^{ms}(t)$, $t \in [t_0, T]$ be the sample path of the solution of Ito equation $(*)$ corresponding to the strategy u^{ms}. Conditions (c), (d) and Theorem 5, §9, ch. 2, Part II in [6] imply the relation

$$W^{(i)}(t_0, x_0) = \mathbf{E}_{t_0,x_0}\{Q_i(T, x^{ms}(T))\} = \mathscr{I}_i(u^{ms}), \ i \in I .$$

Now let $x(t)$, $t \in [t_0, T]$ be the sample path of the solution of $(*)$ which corresponds to an arbitrary strategy $u \in \mathscr{U}$. Write Ito's formula for $W^{(i)}(t, x)$, $x(t)$ and u (see [2]):

$$dW^{(i)}(t, x(t)) = [W_t^{(i)}(t, x(t))+\mathscr{A}(u)W^{(i)}(t, x(t))]dt +$$
$$+[W_x^{(i)}(t, x(t))]'g(t, x(t), u)dw(t) .$$

Then Ito's formula for $[W^{(i)}(t, x)]^2$, $x(t)$ and u has the form:

$$d[W^{(i)}(t, x(t))]^2 = \{2[W_t^{(i)}(t, x(t))+$$
$$+\mathscr{A}(u)W^{(i)}(t, x(t))]W^{(i)}(t, x(t))+$$

$$+[W_x^{(i)}(t, x(t))]'g(t, x(t), u)g'(t, x(t), u)W_x^{(i)}(t, x(t))\}\,dt +$$

$$+2W^{(i)}(t, x(t))[W_x^{(i)}(t, x(t))]'g(t, x(t), u)dw(t).$$

By integration we get

$$[W^{(i)}(T, x(T))]^2 - [W^{(i)}(t, x(t))]^2 =$$

$$= \int_t^T 2[W_t^{(i)}(\tau, x(\tau)) + \mathscr{A}(u)W^{(i)}(\tau, x(\tau))]W^{(i)}(\tau, x(\tau)) +$$

$$+[W_x^{(i)}(\tau, x(\tau))]'g(\tau, x(\tau), u)g'(\tau, x(\tau), u)W_x^{(i)}(\tau, x(\tau))d\tau +$$

$$+2\int_t^T W^{(i)}(\tau, x(\tau))[W_x^{(i)}(\tau, x(\tau))]'g(\tau, x(\tau), u)dw(\tau).$$

Therefore

$$[W^{(i)}(t, x)]^2 = \mathbf{E}_{t,x}\{[W^{(i)}(T, x(T))]^2 - \int_t^T 2[W_t^{(i)}(\tau, x(\tau)) +$$

$$+\mathscr{A}(u)W^{(i)}(\tau, x(\tau))]W^{(i)}(\tau, x(\tau)) +$$

$$+[W_x^{(i)}(\tau, x(\tau))]'g(\tau, x(\tau), u)g'(\tau, x(\tau), u)W_x^{(i)}(\tau, x(\tau))d\tau\}.$$

Taking into account condition (d), we get

$$[W^{(i)}(t_0, x_0)]^2 = \mathbf{E}_{t_0,x_0}\{Q_i^2(T, x(T)) -$$

$$-\int_{t_0}^T 2[W_t^{(i)}(t, x(t)) + \mathscr{A}(u)W^{(i)}(t, x(t))]W^{(i)}(t, x(t)) +$$

$$+[W_x^{(i)}(t, x(t))]'g(t, x(t), u)g'(t, x(t), u)W_x^{(i)}(t, x(t))dt\}$$

and hence

$$\mathscr{I}_i^2(u^{ms}) = \mathbf{E}_{t_0,x_0}\{Q_i^2(T, x(T)) -$$

$$-\int_{t_0}^T 2[W_t^{(i)}(t, x(t)) + \mathscr{A}(u)W^{(i)}(t, x(t))]W^{(i)}(t, x(t)) +$$

$$+[W_x^{(i)}(t, x(t))]'g(t, x(t), u)g'(t, x(t), u)W_x^{(i)}(t, x(t))dt\}, \quad i \in I.$$

While Ito–Dynkin's formula gives us

$$\mathscr{I}_i(u^{ms}) = W^{(i)}(t_0, x_0) =$$

$$= \mathbf{E}_{t_0,x_0}\{Q_i(T, x(T)) - \int_{t_0}^T W_t^{(i)}(t, x(t)) + \mathscr{A}(u)W^{(i)}(t, x(t))dt\}, \quad i \in I.$$

Now consider the expression

$$\sum_{i\in I}[\mathscr{I}_i^2(u^{ms})-2\mathscr{I}_i^a\mathscr{I}_i(u^{ms})]=$$

$$=\sum_{i\in I}\mathbf{E}_{t_0,x_0}\{Q_i^2(T,x(T))-2\mathscr{I}_i^aQ_i(T,x(T))-$$

$$-\int_{t_0}^{T}2[W_t^{(i)}(t,x(t))+\mathscr{A}(u)W^{(i)}(t,x(t))][W^{(i)}(t,x(t))-\mathscr{I}_i^a]+$$

$$+[W_x^{(i)}(t,x(t))]'g(t,x(t),u)g'(t,x(t),u)W_x^{(i)}(t,x(t))dt\}.$$

Thus

$$\sum_{i\in I}[\mathscr{I}_i^2(u^{ms})-2\mathscr{I}_i^a\mathscr{I}_i(u^{ms})]=\sum_{i\in I}[\mathscr{I}_i^2(u)-2\mathscr{I}_i^a\mathscr{I}_i(u)]+$$

$$+\sum_{i\in I}\{\mathrm{Var}_{t_0,x_0}\{Q_i(T,x(T))\}-\mathbf{E}_{t_0,x_0}\{\int_{t_0}^{T}2[W_t^{(i)}(t,x(t))+$$

$$+\mathscr{A}(u)W^{(i)}(t,x(t))][W^{(i)}(t,x(t))-\mathscr{I}_i^a]+$$

$$+[W_x^{(i)}(t,x(t))]'g(t,x(t),u)g'(t,x(t),u)W_x^{(i)}(t,x(t))dt\}\}.$$

Further, condition (b) and the Lemma imply the inequality

$$\sum_{i\in I}[\mathscr{I}_i^2(u^{ms})-2\mathscr{I}_i^a\mathscr{I}_i(u^{ms})]\leqq\sum_{i\in I}[\mathscr{I}_i^2(u)-2\mathscr{I}_i^a\mathscr{I}_i(u)].$$

Hence

$$\sum_{i\in I}[\mathscr{I}_i(u^{ms})-\mathscr{I}_i^a]^2\leqq\sum_{i\in I}[\mathscr{I}_i(u)-\mathscr{I}_i^a]^2$$

for each $u\in\mathscr{U}$. This completes the proof of the Theorem.

Remark. The problem of the existence of mean-square strategies will be considered in an independent research and published additionally in a forthcoming paper.

Acknowledgement. The contents of this paper was discussed with Dr. J. Stoyanov. I have the pleasure to express my gratitude to him for his useful comments and for our collaboration.

References

1. *Dynkin, E. B.*, Markov Processes. Fizmatgiz, Moscow 1963 (In Russian; Engl. transl.: Springer-Verlag, Berlin 1965).
2. *Fleming, W. H., Rishel, R. W.*, Deterministic and Stochastic Optimal Control. Springer-Verlag, Berlin–Heidelberg–New York 1975.

3. *Gaidov, S. D.*, Pareto-optimality in stochastic differential games. Problems of Control and Information Theory, 1986 **15**(6), pp. 439–450.
4. *Gaidov, S. D.*, Absolutely cooperative strategies in stochastic differential games. Universite de Plovdiv, Travaux scientifiques, Mathematiques, 1986, **24**(2), pp. 133–139.
5. *Gaidov, S. D.*, Two cooperative solutions in stochastic differential games. C. R. Acad. Bulgare Sci., 1988, **41**(1), pp. 19–22.
6. *Gihman, I. I., Skorohod, A. V.*, Stochastic Differential Equations. Naukova Dumka, Kiev 1968 (In Russian; Engl. transl.: Springer-Verlag, Berlin 1972).
7. *Gihman, I. I., Skorohod, A. V.*, Theory of Stochastic Processes III. Nauka, Moscow 1975 (In Russian).
8. *Huang, S. C.*, Note of the mean-square strategy for vector-valued objective functions. JOTA, 1972, **9**(5), pp. 364–366.
9. *Vaisbord, I. M., Zhukovskii, V. I.*, Introduction to Many-Player Differential Games and Their Applications. Sovetskoe Radio, Moscow 1980 (In Russian).

Среднеквадратичные стратегии в стохастических дифференциальных играх

С. Д. ГАЙДОВ

(Пловдив)

В работе рассматриваются кооперативные игры многих лиц, где динамика описывается стохастическими дифференциальными уравнениями Ито. Введены среднеквадратичные стратегии как наиболее близкие к абсолютно кооперативным стратегиям. Рассмотрены некоторые свойства этих стратегий, включая оптимальность по Парето. Установлены достаточные условия для среднеквадратичных стратегий.

S. D. Gaidov
Department of Mathematics
Plovdiv University
BG-4000 Plovdiv
Bulgaria

Problems of Control and Information Theory, Vol. 18 (3), pp. 169—182 (1989)

GENERAL MARKOV MODELS WITH THE INFINITE HORIZON

A. B. PIUNOVSKI

(*Moscow*)

(Received December 24, 1987)

The problem of the optimal control of a uniform Markov chain with the infinite horizon with the presence of the discounting factor and with the average criterion is solved.

The sufficiency of Markov stationary nonrandomized control policies is proved, the Bellman equation is obtained, characteristic properties of canonical systems are shown, and new exactly solvable examples are considered.

1. Introduction

Construction of the optimal control policies of stochastic processes is an urgent applied task [1–13]. It should be noted that discrete-time models [2–11] can be used for the approximation of continuous controlled processes, to the direct study of which [1, 12, 13] are devoted. For discounted models the sufficiency of the Markov stationary selector class was proved, the Bellman equation was obtained and its properties were studied [2–5, 7–9]; for models with the average criterion canonical equations and sufficient conditions of the canonical systems' existence were obtained [9, 11]. The most effective results for Borel models were obtained with the supposition of the boundedness of the reward function [4, 8, 9, 11]. However, in practice the above-mentioned supposition is often violated. For example, while studying queueing systems with controllable intensity it is natural to regard the losses, connected with expectation, as infinitely growing with the queue length growth.

In [3, 5, 7] the method of studying countable discounted models with an unbounded reward function is offered. In the present work the above-mentioned results are extended to the Borel models, including those with the average criterion.

In Section 4 three new examples are considered, whereby only the third one fits into the framework of the known investigations. The proofs of the theorems are given in Section 5.

For continuous time jump models (with continuous parameter) analogous results were obtained in [12, 13].

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. Definitions and notations

Let arbitrary Borel spaces $(X, \mathscr{B}(X))$ the state space, and $(A, \mathscr{B}(A))$ the control space be given. The direct product $H_t = X \times (A \times X)^t$ is called the history space; $\Omega \triangleq H_\infty$, $H \triangleq \bigcup_{t=1}^{\infty} H_t$. The spaces H_t, Ω, H are Borel with σ-algebras of Borel sets $\mathscr{B}(H_t)$, $\mathscr{F} = \mathscr{B}(H_\infty)$, $\mathscr{B}(H)$, respectively. Further, if $(E, \mathscr{B}(E))$ is a Borel space, the symbols $\mathscr{A}(E)$, $\mathscr{U}(E)$ denote the analytical and universally measurable σ-algebras in E. The definitions and properties of Borel spaces and probability measures there on, as well as those of analytical, analytically measurable and universally measurable sets, functions and kernels are stated in [4, 8, 9].

The symbols $\xi_t(\omega)$, $a_t(\omega)$ denote the respective elements of the sequence $\omega = \{\xi_0, a_1, \xi_1, \ldots\} \in \Omega$; $h_t(\omega) = \{\xi_0, a_1, \xi_1, \ldots, a_t, \xi_t\}$. The argument ω, as a rule, is omitted everywhere in the following.

The symbol $\mathscr{F}_t$ denotes σ-algebra in Ω, the preimage of $\mathscr{B}(H_t)$.

Let us assume that the one-step transition probability $P_a(x, \Gamma)$ — a probability measure on $X (\forall a \in A, x \in X)$, $\mathscr{B}(A \times X)$ — measurable at any fixed $\Gamma \in \mathscr{B}(X)$ is given: in other words P is a Borel stochastic kernel on X provided $A \times X$.

Definition. A sequence of universally measurable stochastic kernals $\mu_t(\Gamma | h_{t-1})$ on A provided H_{t-1} is called the control policy $\pi = \{\mu_t\}_{t=1}^{\infty}$.

It is known [4, 8, 9] that every policy $\pi = \{\mu_t\}_{t=1}^{\infty}$ at arbitrary fixed $\xi_0 = x \in X$ has a unique probability measure P_x^π on Ω, which is constructed with the help of the Ionescu–Tulcea theorem.

As the projections $\xi_t \colon \Omega \to X$ and $a_t \colon \Omega \to A$ are measurable, each policy at fixed x specifies matched random processes a_t, ξ_t on the stochastic basis $(\Omega, \mathscr{F}, P_x^\pi, (\mathscr{F}_t)_{t=0,1,2,\ldots})$; σ-algebras $\mathscr{F}$, $\{\mathscr{F}_t\}$ are filled up with sets of zero P_x^π-measure. The P_x^π-measure integral is denoted by M_x^π.

Let the upper semi-analytical [4] function (or, in short, A-function) of reward $R(x, a) \colon X \times A \to \mathbf{R}^1$ be given. The object $Z = \{X, A, P, R\}$, as usual, is called the model.

Definition. The policy $\pi = \{\mu_t\}_{t=1}^{\infty}$ is called a selector if for $\forall t > 0$ the measure μ_t is concentrated at point $a_t = \varphi(h_{t-1})$. The selector is called Markovian, if $\varphi(h_{t-1}) = \varphi'(t, \xi_{t-1})$ and stationary, if $\varphi(h_{t-1}) = \varphi'(\xi_{t-1})$. If the policy (selector) is determined by Borel or analytically measurable kernels (functions), it is called B-policy and A-policy (B-selector, A-selector) respectively.

Alongside the initial model Z with infinite horizon, let us introduce a "derived" model $Z_t^T(S)$, given on the interval $[t, t+1, \ldots, T]$ with the "final" reward $S \colon X \to \mathbf{R}^1$.

The function S is supposed to be $\mathscr{U}(X)$-measurable.

Definition. The symbols $\pi^t = \{\mu_\theta\}_{\theta=1}^{t}$, $\pi_t = \{\mu_\theta\}_{\theta=t+1}^{\infty}$ further denote "parts" of the policy π.

It is obvious, that fixed π^t by a given x determines the unique probability measure $P_H^{\pi^t}$ on H_t, whose preimage on Ω is denoted by $P_x^{\pi^t}$. Let x be fixed and $\psi_t(\omega)$ be a sequence of random variables, for which at all policies π the conditional expectation $M_x^{\pi}[\psi_t|\mathscr{F}_t]$ is determined ($t=0, 1, \ldots$). Let us assume that π^t is fixed and π_t is variable. Then $\mathscr{F}_t$-measurable random variables $M_x^{\pi}[\psi_t|\mathscr{F}_t]$ form a set, for which π_t acts as a parameter. Let us introduce the operation $P_x^{\pi^t}$–ess sup by the set $\{\pi_t\}$: the symbol $P_x^{\pi^t}$–ess $\sup_{\pi_t} M_x^{\pi_t}[\psi_t|\mathscr{F}_t]$ denotes such a $\mathscr{F}_t$-measurable random variable $\gamma(\omega)$ that a) $\forall \pi_t$ $\gamma(\omega) \geqq M_x^{\pi}[\psi_t|\mathscr{F}_t]$; b) if $\tilde{\gamma}(\omega)$ is a $\mathscr{F}_t$-measurable random variable satisfying the demand a), then $\gamma \leqq \tilde{\gamma}$ $P_x^{\pi^t}$-almost surely.

The existence and properties of the essential supremum follow from the results obtained in [16] for the operation P–ess inf whose description can be found also in [17] where the symbol $\wedge$ is used.

Let

$$W_t^T(h_t, \pi^t, \pi_t, S) \triangleq M_x^{\pi}\left[\left\{\sum_{\theta=t+1}^{T} R(\xi_{\theta-1}, a_\theta) + S(\xi_T)\right\}|\mathscr{F}_t\right] \tag{1}$$

$$V_t^T(h_t, \pi^t, S) \triangleq P_x^{\pi^t}\text{–ess}\sup_{\pi_t} W_t^T(h_t, \pi^t, \pi_t, S)\,. \tag{2}$$

It is customary to call the functions W_t^T, V_t^T the policy evaluation and the model $Z_t^T(S)$ evaluation. At $t=0$, $h_t=x$ the argument π^t is omitted. The policy π^* is called ε-optimal in the model $Z_0^T(S)$, if $W_0^T(x, \pi^*, S) \geqq V_0^T(x, S)-\varepsilon$ at all $x \in X$. 0-optimal policy is called optimal.

Remark. In the present paper uniform boundedness of the reward function is not necessary. That is why it is useful to specify the meaning of expressions (1)–(2). Let us denote by $\Phi_t(\omega)$ the formula in the figure brackets (1); $\Phi_t^+(\omega) \triangleq \max\{\Phi_t^-(\omega), 0\}$; $\Phi_t^-(\omega) \triangleq \min\{\Phi_t(\omega), 0\}$. Let us consider policies for which $\forall x \in X$ $\hat{\Phi}_t^+ \triangleq M_x^{\pi}[\Phi_t^+|\mathscr{F}_t] < \infty$, or $\hat{\Phi}_t^- \triangleq M_x^{\pi}[\Phi_t^-|\mathscr{F}_t] > -\infty$ P_x^{π}–almost surely (further: P_x^{π}–a.s.) to be admissible. Then for every admissible policy equation (1) is calculated using formula $W_t^T = \hat{\Phi}_t^+ + \hat{\Phi}_t^-$ and the supremum in (2) is taken for the class of admissible policies and it is assumed to be existing.

The present paper is mainly concerned with the study of models "with the average criterion", in which admissibility and optimality of the policy is defined in the following way.

Definition. Policy π in the initial model Z is called admissible, if it is admissible in the model $Z_0^T(0)$, for all $T>0$ and $\forall x \in X$ there exist finite limits

$$\underline{\lim_{T\to\infty}}\, W_0^T(x, \pi, 0)/T; \qquad \overline{\lim_{T\to\infty}}\, W_0^T(x, \pi, 0)/T.$$

The class of admissible policies is denoted by the symbol Π. Policy $\pi^* \in \Pi$ is called average optimal in the model Z, if

$$\forall x \in X \ \forall \pi \in \Pi \ \underline{\lim}_{T\to\infty} W_0^T(x, \pi^*, 0)/T \geqq \overline{\lim}_{T\to\infty} W_0^T(x, \pi, 0)/T.$$

Definition. Stationary selector φ^* and A-functions f and S on X form a canonical system, if for all $x \in X$, $T=0, 1, \ldots$

$$W_0^T(x, \varphi^*, S) = Tf(x) + S(x) = V_0^T(x, S).$$

The characteristic properties of canonical system are given in Theorem 4 (Section 3). Analogous results for models with bounded function of reward were obtained in [9, 11].

In the models with an infinite horizon the optimality criterion is often determined by the formula:

$$W(x, \pi) \triangleq \overline{\lim}_{T\to\infty} M_x^\pi \left[\sum_{\theta=1}^{T} \beta^\theta \cdot R(\xi_{\theta-1}, a_\theta) \right],$$

where $\beta \in (0, 1)$ is the discounting coefficient.

The value $v(x) = \sup_\pi W(x, \pi)$ is called the evaluation of the discounted model Z^β.

The notions of admissibility and optimality of policies are introduced by analogy with the models $Z_t^T(S)$.

3. Supporting results and main theorems

For the study of models with the average criterion additional theorems concerning models $Z_t^T(S)$ and Z^β will be needed.

Below you can find two statements without detailed proof, which generalize the known results in the case of Borel models with an unbounded reward function.

Let $T \leqq \infty$, $g(x)$ be a certain positive $\mathscr{U}(X)$-measurable function. Let the symbol $L^{g,T}$ denote the metrical space of valid A-functions $F_t(x)$ $(x \in X, t \in \{0, 1, \ldots, T\})$ such that $\exists \sup_{(x,t)} g(x)|F_t(x)| < \infty$. The metric is given by the equality $\rho(F^1, F^2) \triangleq \sup_{(x,t)} g(x)$ $|F_t^1(x) - F_t^2(x)|$. It is easy to prove that $L^{g,T}$ is a complete space. At $T = \infty$ the symbol T is omitted further.

For formulating of theorems the following conditions will be needed.

C1. There exists a positive $\mathscr{U}(X)$-measurable function $g(x)$ such that $|\sup_{a\in A} R(x, a)| \in L^g$; $g(x) \int_X P_a(x, dy)/g(y) \leqq 1$.

C2. a) The set A is a compact.

b) The function $R(x, a)$ at all $x \in X$ is upper semicontinuous with respect to a.

Theorem 1. Let condition C1 be satisfied and $S \in L^g$. Then 1) the Bellman equation

$$\mathscr{V}_{t-1}^T(x) = \sup_{a \in A} \{R(x, a) + \int_X \mathscr{V}_t^T(y) P_a(x, dy)\} \tag{3}$$

with the initial condition

$$\mathscr{V}_T^T(x) = S(x) \tag{4}$$

has the unique solution $\mathscr{V}^T \in L^{g,T}$, whereby $V_t^T(h_t(\omega), \pi^t, S) = \mathscr{V}_t^T(\xi_t(\omega))$ $P_x^{\pi^t}$–a.s.

2) $\forall \varepsilon > 0$ in the model $Z_0^T(S)$ there exists an ε-optimal Markovian A-selector;

3) For the optimality of the policy π^* in the model $Z_0^T(S)$ it is necessary and sufficient, that at all $x \in X$ almost everywhere, according to the measure $P_x^{\pi^*}$, the equality be satisfied:

$$R(\xi_{t-1}, a_t^*) + \int_X \mathscr{V}_t^T(y) P_{a_t^*}(\xi_{t-1}, dy) = \mathscr{V}_{t-1}^T(\xi_{t-1}), \qquad t = 1, 2, \ldots, T. \tag{5}$$

Theorem 2. Let conditions C1, C2 be satisfied and one of the following conditions be true:

1) The transition probability $P_a(x, \Gamma)$ is dominated by a certain σ-finite measure $\mu(\Gamma)$ whereby the density $P_a(x, y) = \dfrac{dP_a(x, \cdot)}{d\mu(\cdot)}(y)$ is continuous according to a and for $\forall x \in X$ the function $P_a(x, y)/g(y)$ is majorated by a certain integrated (according to the measure μ) function: $P_a(x, y)/g(y) \leqq f(x, y)$.

2) The set X is finite or countable, the transition probability $P_a(x, y)$ is continuous according to a and for $\forall a \in A$ $\forall x \in X$ converges the row $\sum_{y \in X} P_a(x, y)/g(y) \triangleq C(x, a)$, the function $C(x, a)$ being continuous according to a.

Then

1) the Bellman equation

$$\sup_{a \in A} \{R(x, a) + \int_X \mathscr{V}(y) P_a(x, dy)\} = \frac{1}{\beta} \cdot \mathscr{V}(x) \tag{6}$$

has the unique solution $\mathscr{V} \in L^g$, coinciding with the evaluation of the model Z^β;

2) in the model Z^β there exists an optimal stationary selector;

3) for the optimality of the policy π^* in the model Z^β it is necessary and sufficient that for all $x \in X$ almost everywhere, according to the measure $P_x^{\pi^*}$ the following equality be satisfied

$$R(\xi_{t-1}, a_t^*) + \int_X \mathscr{V}(y) P_{a_t^*}(\xi_{t-1}, dy) = \frac{1}{\beta} \cdot \mathscr{V}(\xi_{t-1}), \qquad t = 1, 2, \ldots \tag{7}$$

Remark. Conditions C2 and 1), 2) are needed only by the proof of point 2) of the theorem.

The proofs of Theorems 1, 2 are carried out according to the standard pattern [4, 8, 9], if it is noted, that the Bellman operators (3), (6) preserve the spaces $L^{g,T}$, L^g invariant, respectively.

Really, let us consider, for example, equation (6), corresponding to the operator

$$Q: \mathscr{V}(x) \to Q \cdot \mathscr{V}(x) \triangleq$$

$$\triangleq \beta \cdot \sup_{a \in A} \{R(x, a) + \int_X \mathscr{V}(y) P_a(x, dy)\} .$$

If $\mathscr{V} \in L^g$, then $Q \cdot \mathscr{V}(x)$ is A-function according to the Zvonkin lemma [8] and statement 7.47 [4]. Uniform boundedness of the function $g(x)|Q \cdot \mathscr{V}(x)|$ follows from condition C1. The condition $\sup_{(a,x) \in A \times X} g(x) \int_X P_a(x, dy)/g(y) \leqq 1$ is used by the proof of the fact that the Q-operator is a contracting one.

The following three theorems are related to the uniform model Z with the average criterion.

Theorem 3. Let such a positive $\mathscr{U}(X)$-measurable function $g(x)$ exist, that $\sup_{(a,x) \in A \times X} g(x) \int_X P_a(x, dy)/g(y) \leqq 1$. Then, if $\langle \varphi^*, f, S \rangle$ is a canonical system and $S \in L^g$, then the policy φ^* is average optimal.

Theorem 4. Let condition C1 be satisfied. The stationary selector φ^* and A-functions f and $S \in L^g$ form a canonical system if and only if

$$\int_X f(y) P_{\varphi^*(x)}(x, dy) = \sup_{a \in A} \{ \int_X f(y) P_a(x, dy)\} = f(x) ; \tag{8}$$

$$R(x, \varphi^*(x)) + \int_X S(y) P_{\varphi^*(x)}(x, dy) = \sup_{a \in A} \{R(x, a) +$$

$$+ \int_X S(y) P_a(x, dy)\} = S(x) + f(x) . \tag{9}$$

As it is known [9], in the models with a bounded reward function the existence of a minoranta is a sufficient condition of the existence of a canonical system. Below the formulation of this result for the case under consideration is presented.

Definition. The finite measure ν on X is called a minoranta if $0 < \nu(X) < 1$ and for $\forall a \in A$, $\forall x \in X$, $\forall \Gamma \in \mathscr{B}(X)$: $\nu(\Gamma) \leqq P_a(x, \Gamma)$.

Suppose that there exists the minoranta ν and let us consider the discounted uniform model

$$Z^\beta = \{X, A, \tilde{P}, \tilde{R}\} ,$$

where

$$\tilde{P}_a(x, \Gamma) \triangleq \frac{1}{1-v(X)}[P_a(x, \Gamma)-v(\Gamma)],$$

$$\tilde{R}(x, a) \triangleq \frac{R(x, a)}{1-v(X)}; \quad \beta = 1-v(X).$$

Theorem 5. Let all the conditions of Theorem 2 be satisfied for the model Z^β. If, for the initial model Z, all conditions C1 are satisfied, then a canonical system exists.

4. Examples

Examples considered below include as a component of the optimality criterion the speed of entropy production by the controlled process. Such optimization tasks appear in the information theory and also by the construction of random search procedures [18]. In the latter case the controlled process can be interpreted as a random walk about the discrete array knots. Without discussing the aims of this walk, let us remark that all the rest of the conditions being equal, the walk with a greater entropy is considered to be more preferable (from the searching point of view).

Model with an average criterion

Let $X = \{\ \ldots, -2, -1, 0, 1, 2, \ \ldots\ \}$, $A = [-q, q]$ where $q \in \left(0, \frac{1}{2}\right]$ is a derived constant;

$$P_a(x, y) = (1-2q)\delta_{y, x+1} + (q+a)\delta_{y, x-1} + (q-a)\delta_{y, x-2}.$$

Here $\delta_{i,j}$ is the Kronecker symbol. Let us assume

$$R(x, a) = -(1-2q)\ln(1-2q) - (q+a)\ln(q+a) - (q-a)\ln(q-a) + K_x(a+q),$$

where $0 \ln 0 \triangleq 0$, K_x is an arbitrary function.

Remark. In the case under consideration the criterion includes the mean speed $\dot{H}$ of the entropy production by the process ξ_t:

$$\dot{H} = \lim_{T \to \infty} \frac{1}{T} H^T,$$

where

$$H^T \triangleq M_x^\pi \left[-\sum_{\theta=1}^{T} \sum_{y \in X} P_{a_\theta}(\xi_{\theta-1}, y) \ln P_{a_\theta}(\xi_{\theta-1}, y) \right]. \tag{10}$$

The last summand of the function R corresponds to the reward from the jump $x \to x-1$.

The validity of the following statement is verified by means of direct substitution.

Let constant f and function $S(x)$ satisfy the following relation:

$$\exp\left\{F-\left(\frac{1}{2q}-1\right)S(x+1)+\frac{1}{2q}S(x)-S(x-1)\right\}=$$

$$=\exp(K_x)\cdot[1+Y_x], \tag{11}$$

where

$$F \triangleq \frac{1}{2q}[f+(1-2q)\ln(1-2q)+2q\ln(2q)];$$

$$Y_x \triangleq \exp[S(x-2)-S(x-1)-K_x]. \tag{12}$$

Then, for $\langle \varphi^*, f, S\rangle$, where

$$\varphi^*(x) \triangleq q\frac{1-Y_x}{1+Y_x}, \tag{13}$$

equalities (8), (9) are satisfied.

So, if for $\langle f, S\rangle$ (11) is fulfilled and a positive function $g(x)$ exists, for which

a) $\sup\limits_{x\in X} g(x)|K_x| < \infty$;

b) $\sup\limits_{a\in A}\sum\limits_{y\in x}\dfrac{P_a(x, y)}{g(y)} \leqq \dfrac{1}{g(x)}$;

c) $\sup\limits_{x\in X} g(x)|S(x)| < \infty$,

then according to Theorems 3 and 4 the object $\langle \varphi^*, f, S\rangle$ is a canonical system, and policy (13) is average optimal.

Let, for instance $q=2/5$; $K_x = \ln\operatorname{sh}(\tilde{F}+2^x)+\tilde{F}+\ln 2$, where $\tilde{F}>0$ is an arbitrary constant. Then $f=\dfrac{8}{5}\tilde{F}-\dfrac{1}{5}\ln\dfrac{1}{5}-\dfrac{4}{5}\ln\dfrac{4}{5}$; $S(x)=2^{x+2}$, and as $g(x)$ it is possible to take the function $g(x)=\dfrac{1}{1+2^x}$. The verification of properties a), b), c) and equality (11) is easy enough.

The average optimal policy (13) looks like this:

$$\varphi^*(x)=\frac{2}{5}\times[1-2\exp(-2\tilde{F}-2^{x+1})].$$

The discounted model

Let us consider the model with the same elements X, A, P, R and an arbitrary discounting coefficient $\beta \in (0, 1)$. Suppose, that there exists a positive function $g(x)$, for which

$$\sup_{x \in X} g(x) \times |K_x| < \infty, \sup_{a \in A} \sum_{y \in x} \frac{P_a(x, y)}{g(y)} \leqq 1/g(x).$$

Then all the conditions of Theorem 2 (version 2) are fulfilled. It is easy to verify that in this case the Bellman equation is equivalent to the equation:

$$\frac{1}{\beta} \mathscr{V}(x) = \mathscr{V}(x+1) - 2q[\mathscr{V}(x+1) - \mathscr{V}(x-1)] - (1-2q) \times$$

$$\times \ln(1-2q) - 2q \ln(2q) + 2qK_x + 2q \ln(1+K_x), \tag{14}$$

where

$$Y_x \triangleq \exp[\mathscr{V}(x-2) - \mathscr{V}(x-1) - K_x],$$

and the optimal synthesis is determined by the equality

$$\varphi^*(x) = q \frac{1-Y_x}{1+Y_x}. \tag{15}$$

Remark. According to Theorem 2, equation (14) has a unique solution in the class $\mathscr{V} \in L^g$. If arbitrary values $\mathscr{V}(-1)$, $\mathscr{V}(0)$, $\mathscr{V}(1)$ are given and (14) is used for recurrent calculation of the function $\mathscr{V}$, at some step under the logarithm sign one might obtain a negative number, or the constructed function will not belong to the class L^g.

Let us assume, for example, that

$$q = \frac{2}{5}; \quad K_x = \ln \operatorname{sh}(\tilde{F} + 7 \cdot 2^x) + \tilde{F} + \ln 2 + 5 \cdot 2^x, \quad \beta = \frac{1}{2}.$$

One can easily verify that the function $g(x) = \dfrac{1}{1+2^x}$ may be taken and the function $\mathscr{V}(x) = 2^{x+3} + \dfrac{8}{5}\tilde{F} - \dfrac{1}{5}\ln\dfrac{1}{5} - \dfrac{4}{5}\ln\dfrac{4}{5}$ is the unique solution of (14) in the class L^g.

The optimal policy (15) looks like this:

$$\varphi^*(x) = \frac{2}{5}[1 - 2\exp(-2\tilde{F} - 2^{x+1})].$$

Model with an average criterion

Let $X=\{0, 1, \ldots, l\}$;

$$A=\left[-\frac{1}{2}, \frac{1}{2}\right]; \quad P_a(x, y)=\left(\frac{1}{2}+a\right)\delta_{y, x+1}+\left(\frac{1}{2}-a\right)\delta_{y, x},$$

where the expression $x+1$ is understood as a result of modulo $l+1$ addition. Let

$$R(x, a)=L_x\left[-\left(\frac{1}{2}+a\right)\times \ln\left(\frac{1}{2}+a\right)-\left(\frac{1}{2}-a\right)\ln\left(\frac{1}{2}-a\right)\right]+aK_x+M_x$$

where $0 \ln 0 \triangleq 0$, $L_x>0$, K_x, M_x are arbitrary functions.

In the case under consideration condition C1 is satisfied by $g(x)\equiv 1$ and, therefore, Theorems 3, 4 are valid.

Let f be the unique solution of the equation

$$\sum_{x=0}^{l}\left\{K_x-L_x \ln\left[\exp\left(\frac{2f+K_x-2M_x}{2L_x}\right)-1\right]\right\}=0, \tag{16}$$

satisfying the inequality $f>\sup\limits_{x\in X}\{-K_x/2+M_x\}$. Let $Y_x \triangleq \left[\exp\left(\frac{2f+K_x-2M_x}{2L_x}\right)-1\right]^{-1}$. Let us assume, that $S(0)$ is equal to an arbitrary constant, $S(x+1)\triangleq S(x)-K_x-L_x \ln Y_x$. (Obviously, $S(l+1)=S(0)$.)

By means of direct substituting one can verify the validity of relations (8), (9) for the stationary selector

$$\varphi^*(x)=\frac{1}{2}\cdot\frac{1-Y_x}{1+Y_x} \tag{17}$$

and for the functions $f(x)\equiv f$, $S(x)$. Hence, according to Theorem 4, the set $\langle\varphi^*, f, S\rangle$ is a canonical system and according to Theorem 3, the policy (17) is average optimal.

Thus, the optimal policy building is reduced to the solution of equation (16), which in the particular case $l=1$; $L_x\equiv L$ is equivalent to the following (square with regard to $\exp(f/L)$) equation:

$$\left[\exp(f/L)\exp\left(\frac{K_0-2M_0}{2L}-1\right)\right]\times$$

$$\times\left[\exp(f/L)\exp\left(\frac{K_1-2M_1}{2L}\right)-1\right]=\exp\left(\frac{K_0+K_1}{2}\right).$$

5. Conclusions

The results, quoted in Section 3 can be obtained in another way, analysing the model

$$\tilde{Z} = \{X, A, \tilde{P}, \tilde{R}\},$$

where

$$\tilde{P}_a(x, \Gamma) \triangleq g(x) \int_{\Gamma} P_a(x, dy)/g(y); \quad \tilde{R}(x, a) \triangleq g(x) R(x, a).$$

Function $\tilde{R}$ is bounded from above and respective theorems from [4] are valid. In the case where $\tilde{P}$ is a substochastic kernel, it is necessary to introduce an additional absorbing state. Evidently, this idea was proposed for the first time by Van der Wal [14] for countable discounted model. The theory of canonical systems for models with an average criterion and an unbounded reward function had not been applied before. In [6] another technique was used by respective limitations on the functions P, R for the countable model.

All the examples presented in Section 4 are new. Their characteristic property is the entropy type of the criterion. Similar examples can be found in [12, 13, 15] for jump continuous-time models. Thus, one can argue, that models with an entropy criterion form a new class of exactly solvable examples alongside with linear models by the square criterion [1, 9].

6. Proofs of main theorems

Proof of Theorem 3. From the obvious relations

$$M_x^{\pi}\left[\frac{1}{g(\xi_T)}\right] = \frac{1}{g(x)} + M_x^{\pi}\left[\sum_{\theta=1}^{T}\left\{\int_X \frac{P_{a_\theta}(\xi_{\theta-1}, dy)}{g(y)} - \frac{1}{g(\xi_{\theta-1})}\right\}\right] \leqq \frac{1}{g(x)},$$

valid for any policy π, we obtain:

$$M_x^{\pi}[|S(\xi_T)|] \leqq \sup_{\tilde{x} \in X} \{g(\tilde{x})|S(\tilde{x})|\}/g(x).$$

Therefore,

$$\forall \pi \ \forall x \in X \ \lim_{T \to \infty} M_x^{\pi}[S(\xi_T)]/T = 0$$

and

$$\underline{\lim_{T \to \infty}} \ W_0^T(x, \pi, 0)/T = \underline{\lim_{T \to \infty}} \ W_0^T(x, \pi, S)/T;$$

$$\overline{\lim_{T \to \infty}} \ W_0^T(x, \pi, 0)/T = \overline{\lim_{T \to \infty}} \ W_0^T(x, \pi, S)/T.$$

The statement of the theorem follows from the definition of a canonical system. The admissibility of the selector φ^* is obvious.

Proof of Theorem 4. According to Theorem 1 for $\forall T < \infty$ the evaluation of the model $Z_t^T(S)$ has the form: $V_t^T(h_t, \pi^t, S) = \mathcal{V}_t^T(\xi_t)$, where $\mathcal{V}^T \in L^{g \cdot T}$ is the unique solution of equation (3) with initial condition (4).

Necessity. Let $\langle \varphi^*, f, S \rangle$ be a canonical system and $S \in L^g$. Then, from (3) and 3) of Theorem 1, we have:

$$Tf(x) + S(x) = \mathcal{V}_0^T(x) = \sup_{a \in A} \{R(x, a) + \int_X \mathcal{V}_1^T(y) \times$$

$$\times P_a(x, dy)\} = R(x, \varphi^*(x)) + \int_X \mathcal{V}_1^T(y) P_{\varphi^*(x)}(x, dy) \tag{18}$$

Relations (9) follow from (18) by $T = 1$ and from (4). Let $T > 1$. Due to the uniformity of the model $Z_t^T(S)$ we have: $\mathcal{V}_1^T(x) = \mathcal{V}_0^{T-1}(x)$. Hence, from (18) we obtain:

$$Tf(x) + S(x) = R(x, \varphi^*(x)) + \int_X [(T-1)f(y) + S(y)] \cdot P_{\varphi^*(x)}(x, dy) =$$

$$= S(x) + f(x) + (T-1) \int_X f(y) P_{\varphi^*(x)}(x, dy) .$$

(Here the proved equality (9) is used.) Thus,

$$f(x) = \int_X f(y) P_{\varphi^*(x)}(x, dy)$$

On the other hand, (18) gives:

$$\frac{S(x) + f(x)}{T-1} + f(x) = \frac{1}{T-1} \times$$

$$\times \sup_{a \in A} \{R(x, a) + \int_X [(T-1)f(y) + S(y)] P_a(x, dy)\} .$$

In the limit by $T \to \infty$ we have:

$$f(x) = \sup_{a \in A} \{ \int_X f(y) P_a(x, dy)\} .$$

Relation (8) is completely proved.

Sufficiency. Let statements (8), (9) be true. Obviously, under $T = 0$ $W_0^0(x, \varphi^*, S) = S(x) = \mathcal{V}_0^0(x)$. Let us suppose, that for a certain $T \geqq 0$ $W_0^T(x, \varphi^*, S) = Tf(x) + S(x) = \mathcal{V}_0^T(x)$ and let us prove analogous relations for $T+1$.

Let $\pi \in \Pi$. Then

$$W_0^{T+1}(x, \pi, S) = M_x^{\pi}\left[R(\xi_0, a_1) + M_x^{\pi}\left[\sum_{\theta=2}^{T+1} R(\xi_{\theta-1}, a_\theta) + S(\xi_{T+1}) | \mathscr{F}_1 \right]\right] =$$

$$M_x^{\pi}[R(\xi_0, a_1) + \int_X W_1^{T+1}(\xi_0, a_1, y, \pi^1, \pi_1, S) P_{a_1}(\xi_0, dy)] \leqq$$

$$\leqq M_x^{\pi}[R(\xi_0, a_1) + \int_X [Tf(y) + S(y)] \cdot P_{a_1}(\xi_0, dy)] \leqq$$

$$\leqq \sup_{a \in A} \{R(x, a) + \int_X S(y) P_a(x, dy) + T \int_X f(y) P_a(x, dy)\} .$$

Here the uniformity of the model: $\hat{\mathscr{V}}_1^{T+1}(x) = \hat{\mathscr{V}}_0^{T}(x)$, and the induction supposition were used. Thus, considering (8), (9), we obtain:

$$W_0^{T+1}(x, \pi, S) \leqq S(x) + (T+1) f(x) .$$

Similarly, for the policy φ^* we have:

$$W_0^{T+1}(x, \varphi^*, S) = R(x, \varphi^*(x)) + \int_X [Tf(y) + S(y)] P_{\varphi^*(x)}(x, dy) =$$

$$= S(x) + (T+1) f(x) .$$

Hence,

$$\varphi^* \in \Pi, \quad \hat{\mathscr{V}}_0^{T+1}(x) = W_0^{T+1}(x, \varphi^*, S) = (T+1) f(x) + S(x)$$

and the induction statement is proved.

Proof of Theorem 5. According to Theorem 2 there exists an optimal stationary selector φ^* in the model Z^{β} and the equation (6) has a unique solution $\tilde{\mathscr{V}}(\cdot) \in L^g$. With the help of (7) under $t = 1$ and Theorem 4 one can easily verify, that $\langle \varphi^*, \int_X \tilde{\mathscr{V}}(y) \nu(dy), \tilde{\mathscr{V}} \rangle$ is a canonical system.

References

1. *Liptser, R. S., Shiryaev, A. N.*, Statistic of random processes. 696 pp. Moscow: Nauka, 1974; English transl., Springer-Verlag, 1979.
2. *Presman, E. L., Sonin, I. M.*, Successive control by imperfect data. 256 pp. Moscow: Nauka, 1982.
3. *Van Nunen, F.* Markov decision processes — theory and applications. Techn. Hochschule Leipzig. Wissenschaftliche Z., **4**, 1980, 6, 357–371.
4. *Bertsekas, D. P., Shreve, S. E.*, Stochastic optimal contol. N. Y., S. Francisco, London: Academic Press, 1978; Russian transl., Moscow: Nauka, 1985, 280 pp.

5. *Wessels, J.*, Markov programming by successive approximations with respect to weighted supremum norms. J. Mathem. anal. and appl., 1977, **58**, 326–335.
6. *Robinson, D. R.*, Markov decision chains with unbounded costs and applications to the control of queues. Adv. Appl. Prob., 1976, **8**, *1*, 159–176.
7. *Van Nunen, J., Wessels, J.*, Markov decision processes with unbounded rewards. Proc. of the adv. Sem. on Markov decision theory. MCT, Amsterdam, the Netherlands, 1977, pp. 1–24.
8. *Juškevič, A. A., Chitashvili, R. J.*, Controlled random sequences and Markov chains. Uspekhi mathem. nauk, 1982, **37**, *6*, 213–242.
9. *Dynkin, E. B., Juškevič, A. A.*, The controlled Markov processes and their applications. 338 pp. Moscow: Nauka, 1975; English transl., Springer-Verlag, 1979.
10. *Fainberg, E. A.*, Controlled Markov processes with arbitrary numeric criteria. Teor. verojatn. i primen., 1982, **27**, *3*, 456–473.
11. *Juškevič, A. A.*, On one class of strategies in common controlled Markov models. Teor. verojatn. i primen., 1973, **18**, *4*, 815–817.
12. *Khametov, V. M., Piunovski, A. B.*, The optimal control of discounted Markov processes with infinite horizon. — II IFAC Symp. on stoch. Control. Vilnius, USSR, 1986 (preprints). Part 1, pp. 326–329.
13. *Piunovski, A. B., Khametov, V. M.*, New effective solutions of optimality's equations for the controlled Markov chains with continuous parameter (the unbounded price-function). Problems of Control and Information Theory, 1985, **14**, *4*, 303–318.
14. *Van der Wal*, Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games. 251 pp. MCT, Amsterdam, the Netherlands, 1981.
15. *Piunovski, A. B., Khametov, V. M.*, Examples of optimal synthesis for jump Markov models. Investigations on stochastic control problems of economic processes. Moscow: CEMI, 1985, pp. 55–76).
16. *Striebel, Ch.* Optimal control of discrete time stochastic systems. Lect. Notes in Econ. and Math. Systems., 1975, **110**, 208 pp. Springer-Verlag: Berlin, Heidelberg, New York.
17. *Elliott, R. J.*, Stochastic calculus and applications. New York, Heidelberg, Berlin: Springer-Verlag, 1982; Russian transl., Moscow, Mir, 1986, 350 pp.
18. *Rastrigin, L. A.*, Statistical methods of search. Moscow: Nauka, 1968, 376 pp.

Общие марковские модели с бесконечным горизонтом

А. Б. ПИУНОВСКИЙ

(Москва)

В статье рассмотрены задачи оптимального управления цепями Маркова с дисконтированием и при среднем критерии. Доказаны теоремы о достаточности стационарных селекторов и о разрешимости уравнения Беллмана, сформулированы характеристические свойства канонических систем. Основные результаты относятся к случаю, когда функция дохода не ограничена по состоянию. В качестве примеров рассмотрены модели с энтропийным типом критерия, для которых построен точный оптимальный синтез.

А. Б. Пиуновский
Институт физико-технических проблем
СССР, 119034 Москва
Кропоткинская ул., д. 13/7

Problems of Control and Information Theory, Vol. 18 (3), pp. 183—195 (1989)

NEW TIME DOMAIN STABILITY ROBUSTNESS MEASURES FOR LINEAR SYSTEMS*

D. B. PETKOVSKI

(Novi Sad)

(Received April 14, 1988)

This paper presents new bounds for the robust stability of linear time-invariant systems in state space models. Explicit bounds on perturbations are given which do not destabilize the system for structured perturbations. These bounds are superior to those reported in recent literature in two senses: (1) they are less conservative and (2) they can be applied to a more general class of systems and perturbations. The results are readily applicable to practical situations in which a designer has some a priori knowledge of the modelling uncertainties and parameter variations. The robustness results are illustrated by two numerical examples including an aircraft control example.

Notations

R^α = Real vector space of dimension α
$\in$ = Belongs to
$\sigma(\cdot)$ = Singular values of the matrix $(\cdot)$
$\lambda(\cdot)$ = Eigenvalues of the matrix $(\cdot)$
$(\cdot)_s$ = Symmetric part of the matrix $(\cdot)$
$|(\cdot)|$ = Modulus matrix = Matrix with modulus entries
$\pi|\cdot|$ = Perron eigenvalue of a non-negative square matrix

1. Introduction

In the analysis and synthesis of multivariable control systems, a fundamental problem that arises is that the mathematical description of a physical plant (i.e. the nominal model) is always characterized by uncertainty or modeling error. In addition, the parameter variations are usually present in the system dynamics. The robustness of multivariable control systems, i.e. their ability to maintain performance in face of uncertainties and perturbations was extensively studied in the past. In the light of

* This research was supported in part by the U.S.–Yugoslav Joint Fund for Scientific and Technological Cooperation in cooperation with DOE under Grant PP-727.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

this, stability robustness is a primary concern because an unstable control system is of little practical importance.

Although more and more attention has been given on the robustness analysis of multivariable feedback systems, the stability robustness evaluation is still a fundamental problem in control theory that has yet to be completely resolved. The recently published literature on the stability robustness analysis of linear time-invariant systems can be viewed from three perspectives:

(i) frequency domain approach,
(ii) time domain approach,
(iii) frequency domain approach which uses a state space representation of the system.

The frequency domain approach which is based on the transfer function representation was extensively studied in the past (e.g. [1]–[5]). The stability robustness has been traditionally described in terms of gain and phase margin, although the more recent approach focusses on the singular value of the return difference or of the inverse return difference matrix. The methodologies for testing the stability robustness of multivariable control systems using matrix-norm bounds in the frequency domain represent a true advantage in our ability to evaluate the tolerance of multivariable control systems.

Although many of the robustness criteria developed so far are in the frequency domain, it is also useful to analyze stability robustness of multivariable control systems in the time domain, especially when a broader class of parameter perturbations have to be considered arising in the state equations describing the plant. Time domain approach [6]–[12] is primarily based on Lyapunov theory and generally involves checking only a finite number of inequalities, often just one, while the frequency domain methodology requires all criteria over the whole range of frequences to be satisfied.

Recently, new bounds on linear time-invariant perturbations which do not destabilize the system were given, based on the frequency domain approach which using a state space representation of the system [13], [14]. It was shown that these bounds are superior to time-domain stability robustness criteria [6], [11], [12] in the sense that they are less conservative and that they can be applied to a more general class of systems and perturbations.

In this paper, we present a new time-domain stability criterion for linear state space models. A computationally effective algorithm is proposed leading to perturbation bounds superior to those based on frequency-domain approach [13], [14] and time-domain approach [6], [11], [12].

The problem formulation is given in Section 2, where a brief discussion on some common stability robustness tests is also included. In Section 3 a new iterative

algorithm for stability robustness analysis is proposed, and it is shown that the directional information on perturbation matrices can be properly represented and adequately studied in order to overcome the conservatism of robustness tests. In Section 4 the robustness results are illustrated by two numerical examples, including an aircraft control example.

2. Problem formulation

Consider a linear time-invariant model of a physical system with linear time-invariant perturbations

$$\dot{x}(t)=(A+E)x(t)\,, \tag{1}$$

where x is the n-dimensional state vector (R^n), A is an $n\times n$ asymptotically stable matrix, and E is a perturbation matrix. In other words, all the parameter variations and modelling errors are lumped into the matrix E.

It is clear that a satisfactory notation of stability robustness of multivariable dynamic systems must be able to characterize the magnitude of the arbitrary perturbations which may be tolerated without instability. Two types of perturbations have been considered:

(*i*) *Unstructured perturbations*

In this case, only a bound on the norm of the perturbation matrix is given. This approach leads to overly conservative results in many instances, since the robustness criteria do not distinguish the structure of perturbations.

(*ii*) *Structured perturbations*

In this case, the structure of perturbations in E is specified, using the information concerning the nature of perturbations which can physically occur, and the bounds on such structured perturbations are given.

In what follows, a brief discussion on some common approaches to stability robustness analysis of state space models is given. We restrict our attention only to structured perturbations. First, we consider a time-domain stability robustness criterion.

Theorem 1 [12]. Assume that the elements of the perturbation matrix E, Eq. (1), are restricted so that

$$|E_{ij}|\leqq e_{ij} \tag{2}$$

and let $e = \max_{i,j} e_{ij}$. Then system (1) is stable if

$$e < \frac{1}{\sigma_{\max}((|P|U)_s)} = \mu_y \tag{3}$$

where P is the solution of the Lyapunov matrix equation

$$A^T P + PA + 2I = 0, \tag{4}$$

U is a matrix with elements $U_{ij} = \frac{e_{ij}}{e}$, $\sigma_{\max}(\cdot)$ is maximum singular value of $(\cdot)$*, and I is identity matrix.

In [13] an alternative frequency-domain criterion has been given.

Theorem 2 [13]. Assume that the perturbation matrix E, Eq. (1), has the structure

$$E = S_1 E^* S_2 \tag{5}$$

where $S_1 \in R^{n \times p}$, $E^* \in R^{p \times q}$, $S_2 \in R^{q \times n}$, $p \leqq n$, $q \leqq n$, and S_1, S_2 are known matrices. Without any loss of generality, assume that rank $S = p$ or/and rank $S_2 = q$, and let the elements of the perturbation matrix be denoted by E^*_{ij}, and assume that

$$|E^*_{ij}| \leqq e_{ij} e \tag{6}$$

where $e_{ij} \geqq 0$ are given, and $e > 0$ is unknown. Then the perturbed system (1) is stable if

$$e < \frac{1}{\sup_{w \geqq 0} \pi(|S_2(jw - A)^{-1} S_1| U)} = \mu_Q \tag{7}$$

where $U \in R^{q \times p}$ is a matrix with elements given by $u_{ij} = e_{ij}$.

As mentioned, it was shown that the frequency-domain bounds (7) are superior to the time-domain stability criterion (3), i.e. criterion (7) leads to less conservative results.

However, it should be pointed out that the frequency-domain criterion (7), as well as time-domain criterion (3), in many cases are not adequate in describing structured perturbations, since the directional information cannot be properly represented nor adequately studied. Namely, the perturbation characterization given by (3) and (5) or (6) leads to a loss of directional information. Therefore, although the effect of the plant parameter variations and modeling errors may be incorporated into the structured uncertainties, the directional information associated with the

* $\sigma_{\max}(X) = (\lambda_{\max}(XX^T))^{1/2}$.

structured uncertainties is lost and the result is an overly conservative stability robustness bound.

In the next section we circumvent these difficulties by introducing a new iterative algorithm for stability robustness bounds improvement. In addition, directional information (in the state space model) associated with the structured uncertainty is exploited to reduce conservatism.

3. An iterative algorithm

In what follows, we give a new stability robustness test for improved perturbation bounds. These bounds are superior to those reported in [11], [12] and [13] in two senses:

(1) they are less conservative, and
(2) the directional information on structured perturbations can be properly represented and adequately incorporated in the robustness test.

Notice that the scalars e_{ij} in Eqs (2) and (6) are restricted to be positive. In other words, the directional information on structured perturbations cannot be properly represented by Eqs (3) and (7). However, in many practical situations a designer following his intuition and experience has enough information concerning the nature of the perturbations (modelling uncertainties, modelling reductions or parameter variations), which can physically occur, to select the most appropriate directions in the space of all perturbation matrices. For example, in the case of weakly coupled systems [15], [16], the perturbation directions in the open loop matrix A and in the control actuating matrix B are defined as

$$\bar{A} = \begin{bmatrix} 0 & A_{12} & \dots & A_{1k} \\ A_{21} & 0 & \dots & A_{2k} \\ \vdots & & & \\ A_{k1} & A_{k2} & \dots & 0 \end{bmatrix}; \qquad \bar{B} = \begin{bmatrix} 0 & B_{12} & \dots & B_{1k} \\ B_{21} & 0 & \dots & B_{2k} \\ \vdots & & & \\ B_{k1} & B_{k2} & \dots & 0 \end{bmatrix} \tag{8}$$

where the matrices $\bar{A}$ and $\bar{B}$ are known.

In a similar case, in the case of singularly perturbed systems [17]

$$\bar{A} = \begin{bmatrix} 0 & 0 \\ A_{21} & A_{22} \end{bmatrix} \qquad \bar{B} = \begin{bmatrix} 0 \\ B_2 \end{bmatrix}. \tag{9}$$

However, these cases cannot be adequately incorporated in the perturbation characterization given by (2), i.e. (3). To overcome this difficulty, we suppose that the

perturbation matrix E is defined by

$$E = e\bar{E} \tag{10}$$

where $\bar{E}$ is given, and $e > 0$ is unknown.

In the following a time-domain stability robustness methodology (in the lines of [11]) will be used to develop an iterative algorithm for determining computationally the largest positive number e such that the perturbed system:

$$(S_0) \qquad \dot{x}(t) = (A + e\bar{E}) \times (t) \tag{11}$$

where A is a time-invariant asymptotically stable matrix, $\bar{E}$ is given, and $e > 0$ is unknown, remains asymptotically stable.

First, we give the following Theorem:

Theorem 3. System (S_0), Eq. (10), is stable if

$$e < \frac{1}{\sigma_{\max}((|P||\bar{E}|)_s)} \tag{12}$$

where P is the solution of the Lyapunov matrix equation (4) and $\bar{E}$ is a given perturbation matrix.

Proof. The proof follows directly from the proof of Theorem 1 [11] and is omitted.

Now, we review the procedure for stability robustness bounds improvement.

Step 1. Using criterion (12), determine e_0.

Step 2. Consider the perturbed system (S_i) as unperturbed system with

$$A_i = A + \sum_{p=0}^{i-1} e_p \bar{E} \tag{13}$$

and determine e_i, solving the corresponding Lyapunov equation

$$\left(A + \sum_{p=0}^{i-1} e_p \bar{E}\right)^T P + P\left(A + \sum_{p=1}^{i-1} e_p \bar{E}\right) + 2I = 0 \tag{14}$$

and using condition (12), i.e. Theorem 3. In each step the new value for A (i.e. A_i) is determined from Eq. (13).

Step 3. Check that the closed loop system (S_i) is stable. If it is the case, return to Step 2. If not,

$$\bar{e} = \sum_{p=0}^{i-1} e_p . \tag{15}$$

Step 4. The largest $\bar{e}$ is defined as

$$\bar{e} = \sum_{p=0}^{\infty} e_p . \tag{16}$$

Therefore, using this procedure we obtain a sequence $\{(S_p)\}$ of closed loop perturbed systems and a sequence $\{(e_p)\}$ of scalars. For each p the eigenvalues of the corresponding perturbed system (S_p) have negative real parts. If, for any i, there is an eigenvalue with zero real part, we cannot apply Theorem 3, and we shall have

$$\bar{e} = \sum_{p=0}^{i-1} e_p . \tag{17}$$

Otherwise, it may be seen that

$$\bar{e} = \sum_{p=0}^{\infty} e_p \tag{18}$$

which may be equal to $+\infty$. If $\lim_{p\to\infty} \bar{e} \to \infty$, then the Lyapunov equation (14) becomes progressively more ill-conditioned as $p \to \infty$ and the process will have to stop for some finite value of $\bar{e}$, due to numerical difficulties.

Although the proposed algorithm appears to be complicated, it is, in fact, not really difficult to be carried out. The only calculation is the solution of the Lyapunov matrix equation (14), and it is well known that there is sophisticated and widely acceptable software to solve this equation.

4. Numerical examples

In order to illustrate some of the results presented in this paper, we give two numerical examples.

Example 1. First, we consider the same example as the one considered in [11] and [13]. The nominal, time-invariant stable matrix is

$$A = \begin{bmatrix} -3 & -2 \\ 1 & 0 \end{bmatrix}. \tag{19}$$

Table 1 gives bounds of allowable perturbations applying criteria (3), i.e. (12) and (7), when the possible perturbed elements of E have different combinations. As it was pointed out in [13], the frequency-domain criterion gives less conservative results than the time-domain criterion (3), i.e. (12). Table 1 also gives the bound of allowable perturbations which do not disturb system stability when the iterative algorithm, Eqs. (13)–(16), is applied. Notice that μ_y, Eq. (3), is equal to e, Eq. (12), in the first iteration. However, the new bounds $\bar{e}$ are a significant improvement over the bounds based on the frequency-domain criterion (7). As can be seen, the number of iterations varies from 3 to 79, depending on the type of perturbation. This simply means that for this example the time-domain criterion (12) should be checked between

Table 1. Comparison of results obtained for the stability robust bound for structured perturbations for Example 1

Perturbed elements of A	$\begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{matrix}$	a_{11}	a_{12}	a_{21}	a_{22}	$\begin{matrix} a_{11} & a_{12} \\ & \end{matrix}$	$\begin{matrix} a_{11} & \\ & a_{22} \end{matrix}$	$\begin{matrix} a_{11} & \\ a_{21} & \end{matrix}$	$\begin{matrix} & a_{12} \\ & a_{22} \end{matrix}$	$\begin{matrix} & \\ a_{21} & a_{22} \end{matrix}$	$\begin{matrix} & a_{12} \\ a_{21} & \end{matrix}$	$\begin{matrix} a_{11} & a_{12} \\ & a_{22} \end{matrix}$	$\begin{matrix} a_{11} & a_{12} \\ a_{21} & \end{matrix}$	$\begin{matrix} a_{11} & a_{21} \\ & a_{22} \end{matrix}$	$\begin{matrix} & a_{12} \\ a_{21} & a_{22} \end{matrix}$
U	$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 1 & 0 \\ 0 & 0 \end{matrix}$	$\begin{matrix} 0 & 1 \\ 0 & 0 \end{matrix}$	$\begin{matrix} 0 & 0 \\ 1 & 0 \end{matrix}$	$\begin{matrix} 0 & 0 \\ 0 & 1 \end{matrix}$	$\begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}$	$\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$	$\begin{matrix} 1 & 0 \\ 1 & 0 \end{matrix}$	$\begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$	$\begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$	$\begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix}$	$\begin{matrix} 1 & 1 \\ 1 & 0 \end{matrix}$	$\begin{matrix} 1 & 0 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix}$
$\mu_y = e$	0.2361	1.6569	1.6569	0.6558	0.3961	1.0000	0.3820	0.4805	0.3246	0.3028	0.5000	0.3117	0.3972	0.2737	0.2564
μ_Q	0.3295	3.0000	2.0000	1.0000	0.6667	1.5201	0.5612	0.9150	0.5000	0.4000	0.8108	0.4486	0.6848	0.3714	0.3528
$\bar{e}$	0.9913	2.9998	2.0000	1.5583	0.6660	1.9961	0.8981	2.9981	0.4997	1.9758	1.9993	0.5808	1.9966	1.4969	0.7301
Number of iterations	34	8	3	5	8	10	30	20	6	50	12	10	26	27	13
"Exact" bounds	0.3333	3	2	1	0.6667	2	0.5616	1	0.5	0.4	1	0.4495	1	0.3723	0.3542

3 and 79 times. However, it should be also pointed out that the frequency-domain methodology requires that the corresponding criterion should be checked over the range of frequency (over a hundred of discrete frequency points are often necessary). As a consequence, the average time needed to determine $\bar{e}$ was between 14% and 90% of the time needed to determine μ_0, Eq. (7). Table 1 also gives "exact" bounds which provide necessary and sufficient conditions for stability robustness. These "exact" bounds are difficult to be computed generally, but in the 2×2 case they can be obtained by observation [13]. However, it should be pointed out that these "exact" bounds were determined [13] when the directional information was not included in the robustness analysis, and they correspond only to the worst possible perturbation, i.e. they do not distinguish possible different directions of perturbations with the same structure. This explains why some of the allowable perturbation bounds determined by the iterative algorithm are larger then the "exact" bounds given in [13].

Table 2 summarizes the robustness results when the directional information of the perturbation matrices is explicitly incorporated in the stability robustness criterion. Therefore, for different directions of the same perturbation matrix we have different bounds of allowable perturbations which do not disturb system stability.

Table 2. Comparison of stability robustness bounds for Example 1 when the directional information on perturbations is explicitly defined

Perturbed elements of A	$a_{11}a_{21}$		$a_{11}a_{12}$		$\begin{matrix} a_{11}a_{12} \\ a_{21}a_{22} \end{matrix}$		$a_{11}a_{12}a_{21}$	
E	$\begin{matrix} 1 & 0 \\ 1 & 0 \end{matrix}$	$\begin{matrix} -1 & 0 \\ -1 & 0 \end{matrix}$	$\begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}$	$\begin{matrix} -1 & 1 \\ 0 & 0 \end{matrix}$	$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$	$\begin{matrix} -1 & 1 \\ -1 & 1 \end{matrix}$	$\begin{matrix} 1 & 1 \\ 1 & 0 \end{matrix}$	$\begin{matrix} -1 & 1 \\ -1 & 0 \end{matrix}$
μ_y	0.48051	0.48051	1.0000	1.0000	0.23607	0.23607	0.39718	0.39718
$\bar{e}$	2.9981	1.0000	1.9961	2.0000	0.9913	0.3333	1.9966	0.99899

For example, for the structured perturbation

$$U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \tag{20}$$

with two directions

$$\bar{E}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \qquad \bar{E}_2 = \begin{bmatrix} -1 & 0 \\ -1 & 0 \end{bmatrix} \tag{21}$$

for the first case, $\bar{E}_1$, the iterative algorithm gives the bound which is nearly 300% greater than the "exact" bound; while for $\bar{E}_2$ the iterative algorithm gives the bounds which is identical to the "exact" bound.

Example 2. We now consider the same example as the one considered in [11] and [18]. The system chosen is the flare control of the Augmentor Wing Jet STOL Research Aircraft (AWJSRA). The equations for the longitudinal dynamics of the AWJSRA at an airspeed of 110 ft/s and flight path angle of -1° are given by

$$\dot{x}(t) = Ax + Bu \tag{22}$$

where

$$A = \begin{bmatrix} -0.0547 & -0.298 & -0.2639 & -0.0031 & 0.0 \\ 0.16 & -0.4712 & 0.4661 & 0.0437 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.1752 & 0.1236 & -0.1236 & -1.3 & 0.0 \\ -0.0174 & 1.92 & 0.0 & 0.0 & 0.0 \end{bmatrix} \tag{23}$$

$$B = \begin{bmatrix} -0.00315 & -0.0943 \\ 0.0408 & 0.0224 \\ 0.0 & 0.0 \\ -1.1200 & -0.08 \\ 0.0 & 0.0 \end{bmatrix}. \tag{24}$$

The performance index considered is

$$J = \int_0^\infty (x^T Q x + u^T R u) dt \tag{25}$$

with $R = \text{diag}\,[16, 0.5]$ and $Q = qI_5$.

The more detailed derivations and descriptions of the physical process can be found in [18].

In [11] the parameter q was used as a design parameter in order to improve the stability robustness bounds of the nominal closed loop system.

The stability robustness bounds μ_y and $\bar{e}$ and their variations with q are summarized in Table 3. Clearly, it is seen that $\bar{e}$ is much greater than μ_y for the values of q considered. Now, a designer is able to precisely define and determine the allowable perturbations which do not disturb system stability. In this way, he can select an appropriate value for q which leads to satisfactory stability robustness bounds.

It should be pointed out that a greater value of the weighting parameter q leads to a more robust design. However, this conclusion is correct only for the full state feedback based on LQ design methodology. In many practical applications, the whole

Table 3. Variations of μ_y and $\bar{e}$ with q, for LQ Design

q	μ_y	$\bar{e}$	Number of iterations
0.1	0.1218900×10^{-2}	0.1560758×10^{-1}	72
0.25	0.1865912×10^{-2}	0.2197875×10^{-1}	67
0.5	0.2505332×10^{-2}	0.2816889×10^{-1}	64
1.0	0.3274987×10^{-2}	0.3601452×10^{-1}	64
5.0	0.5454320×10^{-2}	0.6323033×10^{-1}	72
10.0	0.6445775×10^{-2}	0.7935480×10^{-1}	76
50.0	0.8408343×10^{-2}	0.1238874	84
10^2	0.9019857×10^{-2}	0.1434407	86
10^4	0.1059858×10^{-1}	0.2189010	90

state vector is rarely available for measurement. A realistic control scheme in such cases involves the use of incomplete state measurement to form a control feedback.

In this case, the output feedback design was based on the output matrix

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{26}$$

Table 4 gives the stability robustness bounds μ_y and $\bar{e}$, and their variations with q. Clearly, it is seen that $\bar{e}$ is much greater than μ_y for the values of q considered. Table 4 also shows that $q = 1$, attains the most robust design. Therefore, in the case of output feedback, the greater value for q does not automatically means a more robust design. The presented results can be used to select an appropriate value of the weighting parameter q to attain a robust design.

Table 4. Variations of μ_y and $\bar{e}$ with q, for output feedback design

q	μ_y	$\bar{e}$	Number of iterations
0.1	0.5642174×10^{-3}	0.1907418×10^{-1}	72
0.25	0.7541758×10^{-3}	0.2281746×10^{-1}	67
0.5	0.8740879×10^{-3}	0.2597963×10^{-1}	64
1.0	0.9119264×10^{-3}	0.2784040×10^{-1}	64
5.0	0.3168635×10^{-3}	0.9833818×10^{-2}	72
7.0	0.4861198×10^{-6}	0.1729670×10^{-4}	199

Conclusions

A new computationally efficient algorithm has been proposed for stability robustness evaluation of linear time-invariant systems in state space models. Unlike many other criteria, this algorithm allows a designer to easily incorporate the directional informations on structural perturbations in stability robustness analysis. In this way, the stability robustness bounds have been increased substantially over the ones recently reported in the literature. Two numerical examples, including an aircraft control example have been given to illustrate the algorithm.

References

1. IEEE Trans. on Autom. Control, Special issue on "Linear Multivariable Control Systems". **AC-26**, *1*, Feb. 1981.
2. *Sandell, N. R.* (ed), Recent Developments in Robustness Theory of Multivariable Systems. LIDS-R-945, Massachusetts Institute of Technology, Cambridge, M.A., 1979.
3. *Lehtomaki, N. A., Sandell, N. R., Athans, M.*, Robustness Results in Linear Quadratic Gaussian Based Multivariable Control Designs. IEEE Trans. on Autom. Control, **AC-26**, pp. 75–92, 1981.
4. *Doyle, J. C.*, Robustness of Multiloop Linear Feedback Systems. Proc. Conference on Decision and Control, San Diego, January 1979.
5. *Safonov, M. G., Athans, M.*, Gain and Phase Margin for Multiloop LOG Regulators. IEEE Trans. on Autom. Control, **AC-22**, pp. 173–179, 1977.
6. *Patel, R. V., Toda, M.*, Quantitative Measures of Robustness for Multivariable System. Proc. Joint Automatic Control Conference, paper TD8-A, 1980.
7. *Petkovski, Dj., Athans, M.*, Robustness of Decentralized Output Feedback Control Designs with Application to Electric Power Systems. Third IMA Conference on Control Theory, Sheffield, Sep. 1980, Academic Press, London, pp. 859–880, 1981.
8. *Petkovski, Dj., Athans, M.*, Robustness Results in Decentralized Multivariable Feedback Design. LIDS-R-1304, Massachusetts Institute of Technology, Cambridge, M.A., 1983.
9. *Petkovski, Dj.*, Robustness of Decentralized Control Systems Subject to Sensor Perturbations. IEE Proc. D., Control Theory and Appl., **132**, *2*, pp. 53–60, 1985.
10. *Petkovski, Dj., Athans, M.*, Robust Decentralized Control of Multiterminal DC/AC Power Systems. Electric Power Systems Research, **9**, pp. 253–262, 1985.
11. *Yedavallie, R. K., Banda, S. S., Ridgely, D. B.*, Time Domain Stability Robustness Measures for Linear Regulators. Journal of Guidance Control and Dynamics, May–June, 1985.
12. *Yedavallie, R. K.*, Perturbation Bounds for Robust Stability in Linear State Space Models. Int. J. Control, **42**, pp. 1507–1517, 1985.
13. *Qin, L., Davison, E. J.*, New Perturbation Bounds for the Robust Stability of Linear State Space Models. Proc. Conference on Decision and Control, Athens, Greece, pp. 751–755, Dec. 1986.
14. *Joung, Y. T., Kuo, T. S., Hsu, S. F.*, Stability Robustness Analysis for State Space Models. Proc. Conference on Decision and Control, Athens, Greece, pp. 745–750, Dec. 1986.
15. *Petkovski, Dj., Rakić, M.*, Series Solution of Feedback Gain for Output Constrained Regulators. Int. J. Control, **30**, *4*, pp. 661–669, 1979.
16. *Petkovski, Dj.*, Application of ε-Method to the Class of Linear Systems Applicable to Chemical Engineering. Computers and Chemical Engineering, **2**, pp. 123–126, 1978.
17. *Kokotović, P. V., O'Malley, Sannuti, P.*, Singular Perturbations and Order Reduction in Control Theory. An Overview, Automatica, **12**, *2*, pp. 123–132, 1976.
18. *Patel, R. V., Toda, M., Sridhar, B.*, Robustness of Linear Quadratic State Feedback Designs in the Presence of System Uncertainty. IEEE Trans. on Autom. Control, **AC-22**, pp. 945–949, Dec. 1977.

Новые меры робастной стабильности во временной области для линейных систем

Д. Б. ПЕТКОВСКИ

(Нови Сад)

Даются новые границы для робастной стабильности линейных инвариантных во времени систем в моделях пространства состояний, которые превосходят известные из литературы границы в двух отношениях: они являются менее консервативными и их можно применять для более общего класса систем и пертурбаций. Результаты иллюстрируются двумя численными примерами.

Djordjija B. Petkovski
Centre for Large Scale Control
and Decision Systems
Faculty of Technical Sciences
Veljka Vlahovića 3
21000 Novi Sad
Yugoslavia

РУССКИЙ ПЕРЕВОД

Проблемы управлевния и теории информации, 18 (3) (1989)

ПРИНЦИП МАКСИМУМА И СУПЕРДИФФЕРЕНЦИАЛ ФУНКЦИИ ЦЕНЫ

Н. Н. Субботина

(*Свердловск*)

В настоящей работе установлена важная связь между двумя центральными результатами теории оптимального управления: принципом максимума Л. С. Понтрягина и методом динамического программирования Р. Беллмана. Установлено, что сопряженные переменные из условий принципа максимума являются обобщенными градиентами функции цены — обобщенного решения уравнения Беллмана. Доказано, что в задаче оптимального управления с терминальным функционалом качества это соотношение дополняет условия принципа максимума до необходимых и достаточных условий оптимальности.

1. Введение

Принцип максимума Л. С. Понтрягина [1] — основной математический инструмент, доставляющий необходимые условия оптимальности управляемых процессов. Получение достаточных условий оптимальности связано либо с выявлением специфики задачи (структуры управляемой системы, функционала качества, свойств семейств экстремалей, найденных с помощью приципа максимума, ...) см., например [3, 5, 9, 14, 21], либо с конструированием глобальных оценок функции цены, которая исходному начальному состоянию управляемой системы ставит в соответствие оптимальный результат (см., например [6, 9, 11, 17, 18, 20–25]). Последнее из отмоченных панравлений исследований связано с методом динамического программирования Р. Беллмана [2], который функцию цены трактует как решение нелинейного дифференциального уравнения в частных производных первого порядка типа Гамильтона–Якоби с нелинейностью типа “min” или “max”. Однако это уравнение, которое принято называть уравнением Беллмана, как правило, не имеет классического решения [1, 8, 11, 17, 22]. Функция цены является обобщенным, «вязким» решением уравнения Беллмана [8, 18, 20, 21, 22]. Она принадлежит классу квазидифференцируемых функций [10, 12, 15, 20], для которых в каждой точке открытой подобласти из области определения непусто множество обобщенных градиентов (суперградиентов), называемое супердифференциалом [19, 22].

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Цель данной работы состоит в том, чтобы установить связь между суперградиентами функции цены и сопряженными переменными из условий принципа максимума.

Теорема 4 настоящей работы показывает, что условия принципа максимума, дополненные соотношением, описывающим эту связь, образуют необходимые и достаточные условия оптимальности. Этот результат является развитием идей работ [20, 23, 24]. В предлагаемых конструкциях использован математический аппарат теорич обобщенных управлений [4, 7, 13, 16].

2. Постановка задачи

Рассмотрим следующую задачу оптимального управления со свободным правым концом. Минимизировать функционал качества γ вида

$$\gamma = \gamma_{t_0, x_0}(x(\cdot), u(\cdot)) = \sigma(x(\vartheta)), \tag{1}$$

полагая, что $(x(\cdot), u(\cdot)) : [t_0, \vartheta] \to R^n \times R^p$ и

$$\dot{x}(t) = f(t, x(t), u(t)) \qquad \text{при п.в.} \quad t \in [t_0, \vartheta] \tag{2}$$

$$u(t) \in P \qquad \text{при п.в.} \quad t \in [t_0, \vartheta] \tag{3}$$

$$x(t_0) = x_0. \tag{4}$$

Здесь $x \in R^n$ — фазовый вектор системы (1), $u \in R^p$ — управляющий параметр, принимающий значения в компакте $P \subset R^p$, ϑ — фиксированный конечный момент времени, $(t_0, x_0) \in (-\infty, \vartheta) \times R^n$ — заданное начальное состояние.

Предполагается, что выполнены следующие требования:

1) функции $f(\cdot) : (-\infty, \vartheta] \times R^n \times P \to R^n$ и $\sigma(\cdot) : R^n \to R$ — непрерывны;

2) определены и непрерывны функции

$$\frac{\partial f}{\partial t}(\cdot) : (-\infty, \vartheta) \times R^n \times P \to R^n,$$

$$\frac{\partial f}{\partial x}(\cdot) : (-\infty, \vartheta) \times R^n \times P \to R^{n \times n},$$

$$\frac{\partial \sigma}{\partial x}(\cdot) : R^n \to R^n,$$

3) $\| f(t, x, u) \| \leqq \varkappa(1 + \|x\|)$, $\varkappa > 0 - \text{const}$, где символом $\|h\|$ обозначается евклидова норма вектора $h \in R^n$.

Обозначим символом U_{t_0} множество измеримых по Борелю управлений $u(\cdot):[t_0, \vartheta]\to R^p$, стесненных ограничением (3). В силу сделанных предположений 1)–3) для всякого $u(\cdot)\in U_{t_0}$ существует единственная траектория $x(\cdot):[t_0,\vartheta]\to R^n$ системы (2)–(4). Назовем пару $(x(\cdot), u(\cdot)$, где $u(\cdot)\in U_{t_0}$, а $x(\cdot)$ — траектория системы (2)–(4), отвечающая этому управлению, — управляемым процессом, допустимым для заданного начального состояния (t_0, x_0), и множество всех таких пар обозначим символом $D(t_0, x_0)$.

Число ω^0

$$\omega^0=\omega^0(t_0, x_0)=\inf_{(x(\cdot), u(\cdot))\in D(t_0, x_0)} \gamma_{t_0, x_0}(x(\cdot), u(\cdot)) \tag{5}$$

называется оптимальным результатом или ценой для начального состояния (t_0, x_0). Функция $(t, x)\to\omega^0(t, x):(-\infty), \vartheta)\times R^n\to R$ называется функцией оптимального результата или функцией цены. Допустимый процесс $(x^0(\cdot), u^0(\cdot))$ называется оптимальным для начального состояния (t_0, x_0), если

$$\gamma_{t_0, x_0}(x^0(\cdot), u^0(\cdot))=\omega^0(t_0, x_0)=\min_{(x(\cdot), u(\cdot))\in D(t_0, x_0)} \gamma_{t_0, x_0}(x(\cdot), u(\cdot)). \tag{6}$$

Для того, чтобы гарантировать существование оптимального процесса, предположим (см. [4, 7]), что выполняется условие;

4) при всех $(t, x)\in(-\infty, \vartheta]\times R^n$ выпуклы множества $F(t, x)$ вида

$$F(t, x)=\{f(t, x, u): u\in P\}. \tag{7}$$

3. Принцип максимума

Необходимые условия оптимальности управляемого процесса $(x^0(\cdot), u^0(\cdot))\in$ $\in D(t_0, x_0)$ в задаче (1)–(4), которые доставляет принцип максимума Понтрягина, можно сформулировать в виде следующего утверждения (см. [1]).

Теорема 1. Если $(x^0(\cdot), u(\cdot))$ — оптимальный процесс для начального состояния (t_0, x_0), то для абсолютно-непрерывной вектор-функции $\lambda^0(\cdot), \psi^0(\cdot)$: $:[t_0, \vartheta]\to R\times R^n$, которая является решением сопряженной системы

$$\frac{d\lambda^0(t)}{dt}=-\left\langle \frac{\partial f}{\partial t}(t, x^0(t), u^0(t)), \psi^0(t)\right\rangle \tag{8}$$

$$\frac{d\psi^0(t)}{dt}=-\left(\frac{\partial f}{\partial x}(t, x^0(t), u^0(t))\right)^T\psi^0(t) \tag{9}$$

с краевыми условиями при $t=\vartheta$

$$\lambda^0(\vartheta)=-\min_{u\in P}\langle\psi^0(\vartheta), f(\vartheta, x^0(\vartheta), u)\rangle \tag{10}$$

$$\psi^0(\vartheta) = \frac{\partial \sigma}{\partial x}(x^0(\vartheta)) \tag{11}$$

при почти всех $t \in [t_0, \vartheta]$ выполняется условие

$$\lambda^0(t) \cdot 1 + \langle \psi^0(t), f(t, x^0(t), u^0(t)) \rangle = 0. \tag{12}$$

Здесь T надстрочное — символ транспонирования, а символом $\langle a, b \rangle$ обозначается скалярное произведение векторов a и b. Заметим, что $\lambda^0(\cdot)$ — решение уравнения (8) с краевым условием (10) описывается следующей формулой

$$\lambda^0(t) = -\min_{u \in P} \langle \psi^0(t), f(t, x^0(t), u) \rangle, \qquad t \in [t_0, \vartheta]. \tag{13}$$

Переменные λ, ψ называются сопряженными переменными. Управляемый процесс $(x^*(\cdot), u^*(\cdot)) \in D(t_0, x_0)$, удовлетворяющий условиям (8)–(12) теоремы 1, называется экстремалью. Геометрический смысл этих условий состоит в том, что при движении вдоль экстремальной траектории $(t, x^*(t))$ в расширенном фазовом пространстве $(-\infty, \vartheta] \times R^n$ вектор скорости движения $v^*(t) = (1, f(t, x^*(t), u^*(t)))$ остается ортогональным вектору сопряженных переменных $s^*(t) = (\lambda^*(t), \psi^*(t))$, построенных для данной экстремали $(x^*(\cdot), u^*(\cdot))$.

Если рассматриваемая экстремаль действительно является оптималью, то при движении вдоль оптимальной траектории функция цены должна оставаться постоянной, т. е. скорость движения по оптимальной траектории должна лежать в плоскости, касательной к поверхности уровня функции цены, по которой скользит траектория. Следовательно, вектор сопряженных переменных, ортогональный к вектору $(1, f(t, x^0(t), u^0(t))) = v^0(t)$ согласно (12), должен совпадать с обобщенным градиентом функции цены. Это рассуждение поясняет необходимость совпадения вектора сопряженных переменных с суперградиентом функции цены на содержательном уровне. Строгое доказательство этого факта содержит теорема 4.

4. Супердифференциал и квазидифференцируемость функции цены

Приведем определение супердифференциала [19, 22] функции цены, который будет использован в формулировке необходимых и достаточных условий оптимальности.

Определение 1. Супердифференциалом функции $\omega(\cdot): (-\infty, \vartheta] \times R^n \to R$ в точке (t, x) называется множество $\partial\omega(t, x)$ вида

$$\partial\omega(t, x) = \{(p^0, p) \in R \times R^n : \limsup_{\substack{\tau \to t \\ y \to x}} [(|\tau - t| + \|y - x\|)^{-1} \times$$

$$\times(\omega(\tau, y)-\omega(t, x)-p^0(\tau-t)-\langle p_0(y-x)\rangle]\leqq 0\}. \tag{14}$$

Для того, чтобы получить формулу для супердифференциала функции цены $\omega^0(\cdot)$ (5), воспользуемся следующим утверждением [20].

Теорема 2. Функция цены $\omega^0(\cdot)$ (5) в задаче (1)–(4) при выполнении условий (1)–(4) принадлежит классу Ω квазидифференцируемых [10, 12] функций $\omega(\cdot)$: $:(-\infty, \vartheta]\times R^n\to R$, представимых в следующем виде

$$\omega^0(t, x)=\min_{\alpha\in A}\rho(t, x, \alpha), \qquad t\leqq\vartheta,\ x\in R^n. \tag{15}$$

Здесь α — параметр, A — метрический компакт $\rho(\cdot),:(-\infty, \vartheta]\times R^n\times$ $\times A\to R$ — непрерывная функция, у которой существуют и непрерывны частные производные $\partial\rho(\cdot)/\partial t$, $\partial\rho(\cdot)/\partial x:(-\infty, \vartheta)\times R^n\times A\to R\times R^n$.

Замечание 1. Роль параметров α в формуле (15) играют, грубо говоря, допустимые управления $u(\cdot)$. Точнее, множество A есть множество обобщенных управлений, т. е. измеримых функций, определенных на стандартном отрезке [0, 1] со значениями во множестве регулярных вероятностных мер на P (см. [7, 13, 16]). В случае, когда значения обобщенного управления α суть меры, сосредоточенные в одной точке, образ этого управления α при линейном преобразовании $[0, 1]\to[t, \vartheta]:\tau\to\xi$ вида $\xi=t+(\vartheta-t)\tau$ совпадает с допустимым управлением $u_t(\cdot)\in U_t$. Таким образом, функция $(t, x)\to\rho(t, x, \alpha)$ при таком фиксированном α есть

$$\rho(t, x, \alpha)=\gamma_{t,x}(x(\cdot), u_t(\cdot))=\sigma(x(\vartheta; t, x, u_t(\cdot))) \tag{16}$$

где $x(\xi)=x(\xi; t, x, u_t(\cdot))$, $t\leqq\xi\leqq\vartheta$ — решение уравнения

$$\frac{dx(\xi)}{dt}=f(\xi, x(\xi), u_t(\xi)) \qquad \text{при п.в.} \quad \xi\in[t, \vartheta], \tag{17}$$

$$x(t)=x, \qquad u_t(\cdot)\in U_t.$$

Из теоремы 2 вытекает, что функция $\omega^0(\cdot)$ (5) локально-липшицева и при всех $(t, x)\in(-\infty, \vartheta)\times R^n$ дифференцируема по любому направлению $(\tau, f)\in$ $\in R\times R^n$. Согласно [10, 12, 15], справедлива следующая формула для производной по направлению:

$$D\omega^0(t, x)|(\tau, f)=\lim_{\lambda\downarrow 0}\lambda^{-1}[\omega^0(t+\lambda\tau, x+\lambda f)-\omega^0(t, x)]=$$

$$=\min_{(p^0, p)\in d\omega^0(t, x)}[\tau, p^0+\langle f, p\rangle], \tag{18}$$

где множество $d\omega^0(t, x)$ определяется следующим образом

$$d\omega^0(t, x) = \left\{ p^0 = \frac{\partial \rho}{\partial t}(t, x, \alpha^0), \quad p = \frac{\partial \rho}{\partial x}(t, x, \alpha^0) : \rho(t, x, \alpha^0) = \omega^0(t, x) \right\}. \tag{19}$$

Функция $(p^0, p) \to \langle (\tau, f), (p^0, p) \rangle$ в формуле (18) линейна по (p^0, p), следовательно, для $D\omega^0(t, x)|(\tau, f)$ имеет место также и равенство

$$D\omega^0(t, x)|(\tau, f) = \min_{(p^0, p) \in \operatorname{co} d\omega^0(t, x)} \langle (\tau, f), (p^0, p) \rangle, \tag{20}$$

где символом co S обозначается выпуклая оболочка множества S. Используя соотношения (15), (18)–(20) и определение 1, можно доказать справедливость следующих утверждений.

Лемма 1. При всех $(t, x) \in (-\infty, \vartheta) \times R^n$ множества $d\omega^0(t, x)$ и co $d\omega^0(t, x)$ — непусты и компактны, а многозначные отображения $(t, x) \to d\omega^0(t, x)$ и $(t, x) \to$ $\to$co $d\omega^0(t, x)$ — полунепрерывны сверху по включению.

Лемма 2. Супердифференциал $\partial\omega^0(t, x)$ (14) функции цены (ш) *(з)* имеет вид

$$\partial\omega^0(t, x) = \operatorname{co} d\omega^0(t, x) \qquad t < \vartheta, \quad x \in R^n, \tag{21}$$

Замечание 2. Из соотношений (15), (18)–(21) и результатов работы [15] следует, что супердифференциал функции цены $\partial\omega^0(t, x)$ совпадает с субдифференциалом Ф. Кларка $\partial_{\mathrm{Cl}}\omega^0(t, x)$, который определяется следующим образом

$$\partial_{\mathrm{Cl}}\omega^0(t, x) = \operatorname{co} \{(p^0, p) \in R \times R^n : (p^0, p) =$$

$$= \lim_{\substack{t_k \to t \\ x_k \to x}} \left(\frac{\partial \omega^0}{\partial t}(t_k, x_k), \frac{\partial \omega^0}{\partial x}(t_k, x_k) \right). \tag{22}$$

В этом определении точки (t_k, x_k) выбираются из всюду плотного в $(-\infty, \vartheta) \times R^n$ множества Λ, на котором квазидифференцируемая функция $\omega^0(\cdot)$ — дифференцируема. В общем случае, для липшицевой функции $\omega(t, x)$ справедливо

$$\partial\omega(t, x) \subset \partial_{\mathrm{Cl}}\omega(t, x).$$

5. Метод динамического программирования

Метод динамического программирования в задаче (1)–(4) интерпретирует функцию цены $\omega^0(\cdot)$ (5) как решение следующей задачи Коши для дифференциального уравнения в частных производных первого порядка типа Гамильтона–

Якоби, которое называется уравнением Беллмана

$$\frac{\partial \omega}{\partial t}(t, x) + \min_{u \in P} \left\langle \frac{\partial \omega}{\partial x}(t, x),\ f(t, x, u) \right\rangle = 0, \quad t < \vartheta, \quad x \in R^n \tag{23}$$

$$\omega(\vartheta, x) = \sigma(x), \qquad x \in R^n \tag{24}$$

(см., например, [1, 2]).

Однако квазидифференцируемая функция цены $\omega^0(\cdot)$ (5) является не классическим решением задачи (23), (24), (т.к. такого может не существовать), а обобщенным, «вязким» решением этой задачи [18, 20, 22]. Справедливо [20] следующее утверждение.

Теорема 3. Функция $\omega(\cdot): (-\infty, \vartheta) \times R^n \to R$ совпадает с функцией цены $\omega^0(\cdot)$ (5) задачи (1)–(4) тогда и только тогда, когда

$$\omega(\cdot) \in \Omega, \tag{25}$$

$$\omega(\vartheta, x) = \sigma(x), \qquad \forall\, x \in R^n \tag{26}$$

$$\min_{f \in F(t, x)} D\omega(t, x) | (1, f) = 0 \qquad \forall\, (t, x) \in (-\infty, \vartheta) \times R^n \tag{27}$$

где Ω — класс функций вида (15), а $F(t, x)$ — множество вида (7).

Нетрудно заметить, что в точках дифференцируемости функции $\omega(t, x)$ равенство (27) обращается в равенство (23).

6. Необходимые и достаточные условия оптимальности

Отметим еще ряд полезных свойств функции цены $\omega^0(\cdot)$ (5) (см. [11, 20, 23]).

Лемма 3. Для любого допустимого процесса $(x(\cdot), u(\cdot))$ из начального состояния (t_0, x_0) функция $t \to \rho[t] = \omega^0(t, x(t))$ — абсолютно-непоеывна, монотонно не убывает при $t \to \vartheta$, и при почти всех $t \in [t_0, \vartheta]$ справедливо равенство

$$\frac{d\rho[t]}{dt} = D\omega^0(t, x(t)) | (1, f(t, x(t), u(t))) \geqq 0. \tag{28}$$

Лемма 4. Для того, чтобы процесс $(x^0(\cdot), u^0(\cdot))$ был оптимален для начального состояния (t_0, x_0) необходимо и достаточно выполнение следующих эквивалентных условий

$$\rho^0[t] = \omega^0(t, x^0(t)) = \omega^0(t_0, x_0) - \text{const} \tag{29}$$

при всех $t \in [t_0, \vartheta]$, или

$$D\omega^0(t, x^0(t)) | (1, f(t, x^0(t), u^0(t))) =$$

$$= \min_{u \in P} D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u)) \tag{30}$$

при почти всех $t \in [t_0, \vartheta]$.

Основным результатом данной работы является следующее утверждение.

Теорема 4. Для того, чтобы процесс $(x^0(\cdot), u^0(\cdot))$ был оптимален для начального состояния (t_0, x_0), необходимо и достаточно выполнение условий (8)–(12) теоремы 1 и условия

$$(\lambda^0(t), \psi^0(t)) \in \partial\omega^0(t, x^0(t)), \qquad \forall t \in [t_0, \vartheta], \tag{31}$$

где $\omega^0(\cdot)$ — функция цены (5).

Доказательство

Необходимость. Полученные в результате соответствующих выкладок соотношения (5.18), (5.22) из работы [23] показывают, что сопряженные переменные $(\lambda^0(t), \psi^0(t))$ (8)–(11) суть величины

$$\lambda^0(t) = \frac{\partial\varphi}{\partial t}(t, x^0(t), \alpha_t^0), \qquad \psi^0(t) = \frac{\partial\varphi}{\partial x}(t, x^0(t), \alpha_t^0) \tag{32}$$

для всех $t \in [t_0, \vartheta)$, где параметр $\alpha_t^0 \in A$ при каждом $t \in [t_0, \vartheta)$ есть образ управления $u_t^0(\cdot) \in U_t$ при отображении $[t, \vartheta] \to [0, 1] : \xi \to \tau$ вида $\tau = \dfrac{\xi - t}{\vartheta - t}$, причем

$$u_t^0(\xi) = u^0(\xi) \qquad \text{при} \quad \xi \in [t, \vartheta], \quad t_0 \leqq t \leqq \vartheta, \tag{33}$$

$$\varphi(t, x^0(t), \alpha_t^0) = \omega^0(t, x^0(t)) \qquad \text{при всех} \quad t \in [t_0, \vartheta]. \tag{34}$$

Из (32), (34), (19), (21) следует, что

$$(\lambda^0(t), \psi^0(t)) = (p^0(t), p(t)) \in d\omega^0(t, x^0(t)) \subset \partial\omega^0(t, x^0(t)). \tag{35}$$

Достаточность. Пусть для некоторого допустимого процесса $(x^0(\cdot), u^0(\cdot))$ из начального состояния (t_0, x_0) выполняются условия (8)–(11) и (31). Используя равенство (27) для функции цены $\omega^0(\cdot)$, а также формулу (20) для $D\omega^0(t, x)|(1, f)$, включение (31) и условие (12), получим следующую цепочку соотношений

$$0 = \min_{u \in P} D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u)) \leqq$$

$$\leqq D\omega^0(t, x^0(t))|(1, f(t, x^0(t), u^0(t))) =$$

$$= \min_{(p^0,\, p)\in \partial\omega^0(t,\, x^0(t))} [p^0 + \langle p,\, f(t, x^0(t), u^0(t))\rangle] \leqq$$

$$\leqq \lambda^0(t)1 + \langle \psi^0(t),\, f(t, x^0(t), u^0(t))\rangle = 0 \tag{36}$$

справедливую при всех $t \in [t_0,\, \vartheta]$. Из (36) следует, что при всех $t \in [t_0,\, \vartheta]$ имеет место равенство

$$D\omega^0(t, x^0(t))|(1,\, f(t, x^0(t), u^0(t))) = 0 \tag{37}$$

а это, согласно лемме 4 (условие (30)), является достаточным условием для оптимальности процесса $(x^0(\cdot),\, u^0(\cdot))$. Теорема 4 доказана.

Замечание 3. Теорема 4 остается верна, если условие (31) заменить условием

$$(\lambda^0(t), \psi^0(t)) \in d\omega^0(t, x^0(t)) \qquad \text{при всех} \quad t \in [t_0,\, \vartheta], \tag{38}$$

где множество $d\omega^0(t, x)$ определено соотношением (19). Справедливость этого утверждения вытекает из (18), (20) и (35).

Замечание 4. (См. [25]). Все результаты данной статьи легко обобщить для случая, когда в рассматриваемой задаче (1)–(4) функционал качества заменен на функционал

$$\gamma^* = \gamma^*_{t_0, x_0}(x(\cdot), u(\cdot)) = \sigma(x(\vartheta)) + \int_t^{\vartheta} f^0(t, x(t), u(t))\, dt.$$

Литература

1. *Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф.* Математическая теория оптимальных процессов. М.: Физматгиз, 1961.
2. *Беллман Р.* Динамическое программирование. М.: ИЛ., 1960.
3. *Красовский Н. Н.* Об одной задаче оптимального регулирования нелинейных систем. Прикл. матем. и механ., 1959, т. **23,** № *2,* с. 209–229.
4. *Филиппов А. Ф.* О некоторых вопросах теории оптимального регулирования. Вестник МГУ, сер. матем. мех., астрон., физ., хим., 1959, № *2,* с. 25–32.
5. *Розоноэр Л. И.* Принцип максимума Л. С. Понтрягина в теории оптимальных систем. III. Автомат. и телемех., 1959, т. **20,** № *12,* с. 1561–1578.
6. *Кротов В. Ф.* Методы решения вариационных задач на основе достаточных условий абсолютного минимума. I. Автомат. и телемехан., 1962, т. **23,** № *12,* с. 1571–1583.
7. *Гамкрелидзе Р. В.* О скользящих оптимальных режимах. Докл. АН СССР, 1962, т. **143,** № *6,* с. 1243–1246.
8. *Кружков С. Н.* Обобщенные решения нелинейных уравнений первого порядка со многими независимыми переменными I. Матем. сборник, 1966, т. **70,** № *3,* с. 394–415.
9. *Болтянский В. Г.* Математические методы оптимального управления. М.: Наука, 1969.
10. *Пшеничный Б. Н.* Необходимые условия экстремума. М.: Наука, 1969.
11. *Лейтман Дж.* Введение в теорию оптимального управления. М., Наука, 1968.
12. *Демьянов В. Ф., Малозёмов В. Н.* К теории нелинейных минимаксных задач. Успехи матем. наук, 1971, т. **26,** № *3*(159), с. 53–104.

13. *Янг Л.* Лекции по вариационному исчислению и теории оптимального управления. М.: Мир, 1974.
14. *Благодатских В. И.* Достаточные условия оптимальности для дифференциальных включений. Изв. АН СССР. Сер. матем., 1974, **38,** No **3,** с. 615–624.
15. *Clarke, F. H.*, Generalized gradients and applications. Trans. Amer. Math. Soc., 1975, Vol. **205,** No. *2*, pp. 247–262.
16. *Варга Дж.* Оптимальное управление дифференциальными и функциональными уравнениями. М.: Наука, 1977.
17. *Флеминг У., Ришел Р.* Оптимальное управление детерминированными и стохастическими системами. М.: Мир, 1978.
18. *Хрусталев М. М.* Необходимые и достаточные условия оптимальности в форме уравнения Беллмана. Докл. АН СССР, 1978, т. **242,** № *5*, с. 1023–1026.
19. *De Giorgi, E., Marino, A., Tosques, M.*, Problemi di evoluzione in spazi metrici e curve di massima pendenza. Rend. Cl. Sci. Fis. Mat. Nat. Accad. Naz. Lincei, 1980, Vol. **68,** pp. 180–187.
20. *Субботин А. И., Субботина Н. Н.* К вопросу обоснования метода динамического программирования в задаче оптимального управления. Изв. АН СССР, Техн. киберн., 1983, № *2*, с. 24–32.
21. *Вязгин В. А.* К обоснованию достаточных условий оптимальности в методах Вейерштрасса и Гамильтона–Якоби–Беллмана. Автоматика и телемехан., 1984, № *4*, с. 31–37.
22. *Crandall, M. C., Evans, L. C., Lions, P. L.*, Some properties of viscosity solutions of Hamilton–Jacobi equations. Trans. Amer. Math. Soc. 1984, Vol. **282,** No. *2*, pp. 487–502.
23. *Субботина Н. Н.* Необходимые и достаточные условия оптимальности управлений и траекторий. Сб.: Синтез оптимального управления в игровых системах. Свердловск: УНЦ АН СССР, 1986, с. 86–96.
24. *Clarke, F. H., Vinter, R. B.*, The relationship between the maximum principle and dynamic programming. SIAM J. Control & Optimization, 1987, Vol. **25,** No. *5*, pp. 1291–1311.
25. *Субботина Н. Н.* Необходимые и достаточные условия оптимальности в терминах принципа максимума и супердифференциала функции цены. АН СССР, УрО, ИММ — Свердловск, 1988. Рукопись деп. в ВИНИТИ 15. 04. 88., № 2898-В88.

PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, 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 V. STREJC
UTIA ČSAV
18208 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

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: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

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. TOM 18 · NUMBER HOMEP 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 СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1989

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 PLC Headington Hill Hall, Oxford OX3 OBW, England

or

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

1989 Subscription Rate DM 535,— 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. 18 (4), pp. 197—221 (1989)

CONTRIBUTIONS TO A THEORY OF ORDERING FOR SEQUENCE SPACES[1]

R. AHLSWEDE, Z. ZHANG

(Bielefeld)

(Received March 4, 1988)

We continue the investigations of [1] and obtain first results concerning optimal orderings of sequences in two simple, but basic, cases:

a non-probabilistic model with active memory and
a probabilistic model with unlimited passive memory.

1. Introduction

In [1] we have presented ideas about a theory of ordering and established some primary but basic results for a non-probabilistic model of ordering sequences under constraints on the operations and the knowledge about the sequences.

The basic model considered in [1] can be formulated as follows. Suppose we have a box that contains at time $t=0$ β balls. We assume that the balls are labelled with numbers from $\mathcal{X}=\{1, \ldots, \alpha\}$. For simplicity, we say a ball "i" instead of a ball labelled by "i". Thus, the content or "state" of the box can be described by a multi-set $s_0=(s_0(1), \ldots, s_0(\alpha))$, where $s_0(i)$ is the number of i's in the box and $\sum_{i=1}^{\alpha} s_0(i)=\beta$. We denote a sequence of balls just by its labelling. An arbitrary n-length sequence of balls, say

$$x^n=(x_1, \ldots, x_n) \in \mathcal{X}^n,$$

enters the box iteratively. At time t, x_t enters just after a person, called the organizer, has pulled out a ball "y_t" from the box. Consequently, the state s_{t-1} should be changed to s_t. We call x^n the input sequence and $y^n=(y_1, \ldots, y_n)$ the output sequence. The organizer's strategy obeys the following rules.

[1] Presented at the International Colloquium on Coding Theory, organized by Osaka University and the Armenian Academy of Sciences, Osaka, June 24–28, 1988.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

(1) The organizer can output only objects which he has in the box. At any time $t(0<t\leq n)$ he can and has to output exactly one ball.
(2) The organizer's strategy may depend on some of his knowledge such as knowledge concerning time, natural order of the balls in the box, the input sequence and the output sequence before the current time.

The subclass of problems, which have been investigated most intensively in [1], is characterized by a triple (π, β, φ). It specifies that β objects from $\mathcal{X}$ fit into the box and that at any time t the organizer $\mathcal{O}$ knows the state of the box s_t, that he can see the incoming letters $x_t, x_{t+1}, \ldots, x_{t+\varphi-1}$ and that he still remembers (or can see) the output letters $y_{t-\pi}, y_{t-\pi+1}, \ldots, y_{t-1}$ when he outputs y_t.

The goal of $\mathcal{O}$ is to minimize the number of output sequences for a specified blocklength n.

We continue here the studies of [1] in two directions.

I. *Active memory*

In addition to the knowledge described by the triple (π, β, φ), which we term "passive memory", the organizer may have storage space, in which he can (and has to) feed a number from $\{0, 1, \ldots, \gamma-1\}$. He uses this storage to remember a certain amount of any information relevant to him. We speak of an active memory. It can be realized by a switching board with γ states. The organizer can turn the state of the board to any one of the states labelled by the numbers $0, 1, \ldots, \gamma-1$ based on his current knowledge and try to "remember" something.

Formally the new model is described by a quadruple $(\pi, \beta, \varphi, \gamma)$. It involves passive and additional active memory. In this notation the case of passive memory only, described by the triple (π, β, φ), can equivalently be described by the quadruple $(\pi, \beta, \varphi, 1)$. Here we study another extremal case, namely that of "pure" active memory that is, the case $(0, \beta, 0, \gamma)$.

We denote the set of the organizer's strategies by $F_\alpha^n(\beta, \gamma)$. Note that here a strategy is a pair of functions (f, g), where for state of the box s and state of the switching board z $f(s, z)$ gives the output and $g(s, z)$ gives the next state of the board. For initial state of the box s_0, initial state of the board z_0 and strategy (f, g) an input sequence x^n determines as output sequence $y^n = y^n(z_0, s_0, x^n, f, g)$. The set of all n-length output sequences under (f, g) is therefore

$$\mathcal{Y}^n(f, g) = \{y^n(z_0, s_0, x^n, f, g): 0 \leqq z_0 \leqq \gamma-1, s_0 \in \mathcal{S}, x^n \in \mathcal{X}^n\}, \tag{1.1}$$

if $\mathcal{S}$ denotes the set of all possible states of the box.

In accordance with the terminology of [1], where $v_\alpha(\pi, \beta, \varphi)$ was defined, we

introduce now

$$N_\alpha^n(0, \beta, 0, \gamma) = \min \{|\mathscr{Y}^n(f, g)| : (f, g) \in F_\alpha^n(\beta, \gamma)\}, \tag{1.2}$$

$$v_\alpha(0, \beta, 0, \gamma) = \lim_{n \to \infty} \frac{1}{n} \log N_\alpha^n(0, \beta, 0, \gamma). \tag{1.3}$$

It is convenient to use the abbreviations

$$\bar{N}_\alpha(\beta, \gamma) = N_\alpha^n(0, \beta, 0, \gamma), \ \bar{v}_\alpha(\beta, \gamma) = v_\alpha(0, \beta, 0, \gamma). \tag{1.4}$$

These quantities are studied in Section 2.

The key new observation is that in case $\gamma = 2$ our strategy of [1] in case $(1, \beta, 0)$ can be simulated by using instead of the one letter knowledge of the past the active memory and that this is optimal.

Section 3, the rest of the paper, is devoted to an analysis of the probabilistic model mentioned under C in Section 2 of [1]. The objects or letters are here produced by a stochastic process, which in the simplest case is a sequence $(X_t)_{t=1}^\infty$ of i.i.d RV's with values in $\mathscr{X} = \{0, 1, \ldots, \alpha - 1\}$ and generic distribution P_X. In Information Theory this is also called a discrete, memoryless source. For a strategy f_n, which depends on the triple (π, β, φ), let $Y^n = Y_1 \ldots Y_n$ be the output sequence corresponding to $X^n = X_1 \ldots X_n$. Let $F_\alpha^n(\pi, \beta, \varphi, P_X)$ be the set of strategies restricted to blocklength n.

We use the "per letter" entropy $\frac{1}{n} H(Y^n)$ as performance criterion and define

$$\eta_\alpha(\pi, \beta, \varphi, P_X) = \lim_{n \to \infty} \min_{f_n \in F_\alpha^n(\pi, \beta, \varphi, P_X)} \frac{1}{n} H(Y^n). \tag{1.5}$$

This is the smallest mean entropy of the output process, which can be achieved by $\mathscr{O}$ with strategies based on his knowledge. It corresponds to the optimal rate $v_\alpha(\pi, \beta, \varphi)$ in our non-probabilistic model. Our new quantity is much harder to analyse.

We consider here the simplest (non-trivial) source, that is the binary symmetric source defined by $P_X(0) = P_X(1) = 1/2$, in the first non-trivial case $\beta = 2$. Further it is assumed that $\mathscr{O}$ knows the ∞-past and has no knowledge about the future.

We give a nice formula for $\eta_2(\infty, 2, 0, P_X)$.

2. Ordering with active memory

This section is entirely devoted to the proof of the following result.

Theorem 1

(a) $\bar{v}_2(\beta, 2) = v_2(1, \beta, 0)$

(b) $v_2(1, \beta, 0) = \log \Psi_\beta$,

where Ψ_β is the positive root of $\lambda^\beta - \lambda^{\beta-1} - 1 = 0$.

Whereas (b) repeats an earlier result, Theorem 6 of [1], (a) establishes a new and interesting connection. Since the organizer $\mathcal{C}$ can use the active memory to remember the last letter send, clearly

$$\bar{\nu}_2(\beta, 2) \leqq \nu_2(1, \beta, 0). \tag{2.1}$$

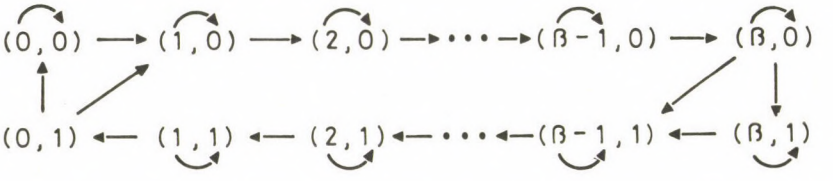
The issue is that in the present situation there is no better way to use the active memory. Actually we show that the strategy which we found in [1] for the case $(1, \beta, 0)$, is also optimal here. Of course now we have to exclude more possibilities and the proof is therefore somewhat more complicated than the one in [1]. However, it is based on similar ideas. We denote by $\bar{s}_t$ the pair (s_t, z_t), where s_t is the state of the box and z_t the state of the board at time t. z_t takes values in $\{0, 1\}$ and s_t can be assumed to count the number of 1's in the box.

We claim that the following strategy (f^*, g^*) is optimal:

$$f^*(s, 0) = 0 \quad \text{for} \quad s < \beta \quad \text{and} \quad f^*(s, 1) = 1 \quad \text{for} \quad s > 0;$$

$$f^*(0, z) = 0, \quad f^*(\beta, z) = 1, \quad \text{and} \quad g^* \equiv f^*. \tag{2.2}$$

Notice that it simply repeats the previous action, if this is possible. We can draw a state transition chart for this strategy. The 2 outgoing arrows correspond to the 2 possible inputs. Loops are included.



Denote by $\mathcal{M}^t(s, z)$ the set of all possible output sequences of length t achievable with $\bar{s}_t = (s, z)$ and some initial state. Since $\mathcal{M}^{t-1}(s, 0) \subset \mathcal{M}^{t-1}(s-1, 0)$ and $\mathcal{M}^{t-1}(s, 1) \subset \subset \mathcal{M}^{t-1}(s+1, 1)$, one readily verifies the following relations for $M^t(s, z) = |\mathcal{M}^t(s, z)|$

(i) $M^t(s, 0) = M^{t-1}(s-1, 0)$ for $2 \leqq s \leqq \beta$
(ii) $M^t(s, 1) = M^{t-1}(s+1, 1)$ for $0 \leqq s \leqq \beta - 2$
(iii) $M^t(0, 0) = M^{t-1}(0, 1) + M^{t-1}(0, 0)$;
$M^t(\beta, 1) = M^{t-1}(\beta, 0) + M^{t-1}(\beta, 1)$
(iv) $M^t(1, 0) = M^t(0, 0)$; $M^t(\beta - 1, 1) = M^t(\beta, 1)$
(v) $M^t(\beta - 1, 1) = M^t(1, 0)$.

Therefore we can conclude that

$$M^t(\beta - 1, 1) = M^{t-1}(\beta - 1, 1) + M^{t-\beta}(\beta - 1, 1) \tag{2.3}$$

and that $M^t(\beta - 1, 1) = \max \{M^t(s, z) : 0 \leqq s \leqq \beta;\ z = 0, 1\}$.

In [1] a similar relationship was used to derive (b).

The equation $\lambda^{\beta}-\lambda^{\beta-1}-1=0$ shows that for smaller β there is less compression of the sequence spaces.

We are going to prove that (f^*, g^*) is an optimal strategy. We begin with a simple observation.

Lemma 1. For every optimal strategy (f, g)

$$f(s, 0)\neq f(s, 1) \quad \text{for} \quad 1\leqq s\leqq\beta-1 .$$

Proof. Assume to the opposite that for some s in the specified range $f(s, 0)=$ $=f(s, 1)$. Then either $\{s'|0\leqq s'\leqq s\}$ or $\{s'|s\leqq s'\leqq\beta\}$ is a closed set of states, that is, starting from a state inside this set, we can never reach a state outside this set. This can be viewed as having a smaller box of size either s or $\beta-s$ and in any case of a size smaller than β. Proceeding inductively in β, we see that (f, g) cannot be optimal.

Lemma 2. For an optimal strategy (f, g) the condition

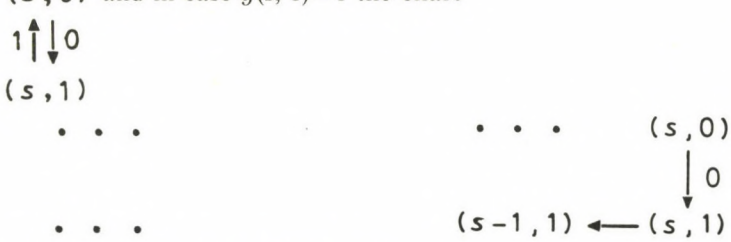
$$(f, g)(s, 0)=(0, x) \quad \text{for any} \quad s\geqq 1 \quad \text{and any} \quad x\in\{0, 1\} \tag{2.4}$$

implies that there exists a directed path in the state transition chart from $(s, 1)$ to $(s, 0)$ of a length not exceeding $2s$.

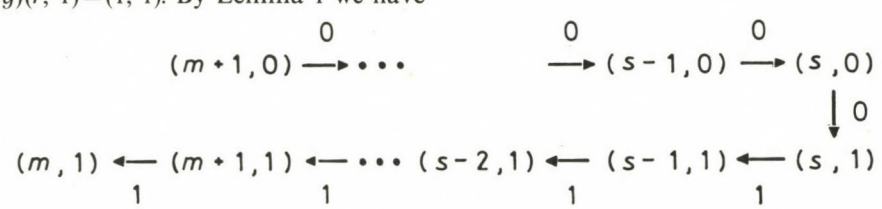
Proof

Case $x=1$: Since $f(s, 0)=0$ we have $(s, 0)$ $\downarrow 0$ $(s, 1)$

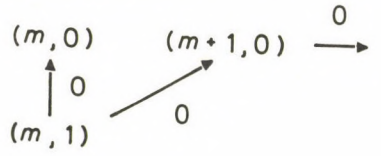
Further, since by Lemma 1 $f(s, 1)=1$ we have in case $g(s, 1)=0$ the desired path $(s, 0)$ and in case $g(s, 1)=1$ the chart



Let now m be the smallest number such that for all r with $m+1\leqq r\leqq s$ $(f, g)(r, 1)=(1, 1)$. By Lemma 1 we have

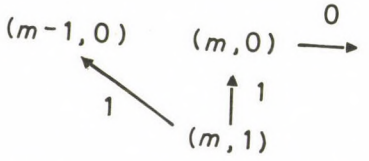


If $g(m, 1)=0$, then we get either



in case $f(m, 1)=0$

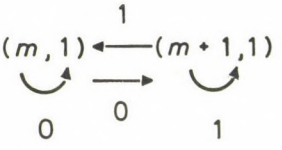
or



in case $f(m, 1)=1$.

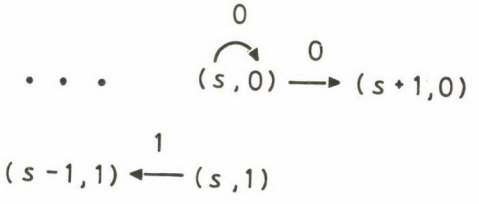
In both cases there is a path from $(s, 1)$ to $(s, 0)$ of a length not exceeding $2s$.

If $g(m, 1)=1$, then by definition of m necessarily $f(m, 1)=0$. This results in a chart



which cannot occur for an optimal strategy.

Case $x=0$: Either $g(s, 1)=0$ and we have a path of length 1 or $g(s, 1)=1$ and the left part of the chart



can be analyzed as in the previous case.

Our next result holds again for all optimal strategies.

Lemma 3. For $\beta \geqq 4$, any optimal strategy (f, g) satisfies in *case* $s=1$

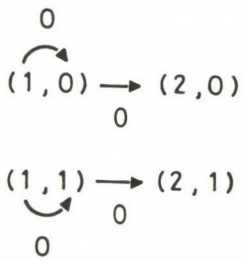
$$\begin{array}{ll} \text{either} & (f, g)(1, 0)=(0, 0) \\ \text{or} & (f, g)(1, 1)=(0, 1) \end{array} \tag{2.5}$$

and in *case* $s=\beta-1$

$$\begin{array}{ll} \text{either} & (f, g)(\beta-1, 0)=(1, 0) \\ \text{or} & (f, g)(\beta-1, 1)=(1, 1). \end{array} \tag{2.6}$$

Proof. By symmetry only one of the two cases has to be established. In case

$s=1$ our statement can be visualized by the following state transition chart:



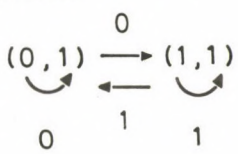
The number at an arrow, in this case the number 0, indicates the output object.

Suppose now that for an optimal strategy our claim for $s=1$ is false.

Case $(f, g)(1, 0)=(0, 1)$: By Lemma 1 thus $f(1, 1)=1$ and we are left with the subcases

(a) $(f, g)(1, 1)=(1, 1)$
(b) $(f, g)(1, 1)=(1, 0)$.

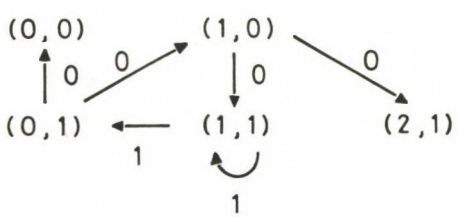
Ad (a): Since always $f(0, 1)=0$, necessarily $(f, g)(0, 1)=(0, x)$. However, if $x=1$, then a part of the chart is of the form



This leads to the relations

$$\begin{aligned} M^t(1, 1) &= M^{t-1}(1, 1)+M^{t-1}(0, 1) \\ &= M^{t-1}(1, 1)+M^{t-2}(1, 1)+M^{t-2}(0, 1) \\ &\geqq M^{t-1}(1, 1)+M^{t-2}(1, 1). \end{aligned}$$

Since the biggest root Ψ_β of $\lambda^\beta-\lambda^{\beta-1}-1=0$ is strictly decreasing in β and since $\beta\geqq 4\geqq 2$, $M^t(1, 1)$ grows to fast. This means that we must have $(f, g)(0, 1)=(0, 0)$ and a chart



From the cycle in this chart we derive the inequality

$$M^t(1, 1)\geqq M^{t-1}(1, 1)+M^{t-3}(1, 1),$$

which results in a rate of growth not smaller than $\log \Psi_3$. Case (a) cannot occur.

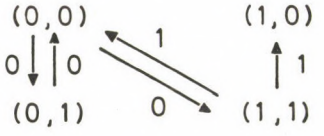
In case (b) we have the following 3 possibilities

(b1) $(f, g)(0, 0)=(0, 0)$
(b2) $(f, g)(0, 0)=(0, 1)$ and $(f, g)(0, 1)=(0, 1)$
(b3) $(f, g)(0, 0)=(0, 1)$ and $(f, g)(0, 1)=(0, 0)$.

One readily checks with the state transition charts that in cases (b1) and (b2) inequalities of the form

$$M^t(.,\ .) \geqq M^{t-1}(.,\ .)+M^{t-3}(.,\ .)$$

hold, whereas in case (b3) we have



Therefore $$M^t(0, 0)=M^{t-1}(0, 1)+M^{t-1}(1, 1),$$

$$M^{t-1}(0, 1)=M^{t-2}(0, 0)$$

and $$M^{t-1}(1, 1) \geqq M^{t-2}(0, 0).$$

In all these cases the rate of growth exceeds $\log \Psi_4$ and (b) cannot occur for an optimal strategy.

We are left with

Case $f(1, 0)=1$: By Lemma 1 $f(1, 1)=0$ and by our supposition necessarily $g(1, 1)=0$. Now just notice that the previous case $(f, g)(1, 0)=(0, 1)$ and the present case $(f, g)(1, 1)=(0, 0)$ differ only in the labelling of states in the active memory.

Lemma 4. For $\beta=2, 3$ there are optimal strategies for which (2.5) and (2.6) hold. (f^*, g^*) is optimal for $\beta=2$.

Proof. Inspection of the previous proof shows that in case $\beta=3$ a strategy violating (2.5) or (2.6) cannot be better than (f^*, g^*), for which (2.5) and (2.6) hold. The case $\beta=2$ requires a more refined analysis. Here by Lemma 1 the optimal f is up to the labelling of the states in the active memory unique, namely, $f=f^*$.

We go again through the cases of the proof of Lemma 3. If (2.5) does not hold, then necessarily

$$(f, g)(1, 0)=(f^*, g)(1, 0)=(0, 1) \tag{2.7}$$

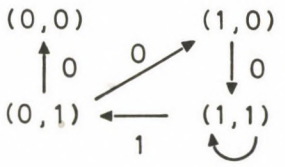
and we are left with the alternatives

(a) $(f, g)(1, 1)=(1, 1)$
(b) $(f, g)(1, 1)=(1, 0)$.

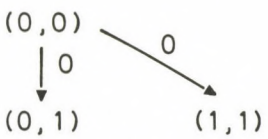
Ad (a): We have seen that in the case $(f, g)(0, 1)=(0, 1)$

$$M^t(1, 1) \geqq M^{t-1}(1, 1)+M^{t-2}(1, 1)$$

and therefore (f^*, g^*) is not superseded. We are left with the case $(f, g)(0, 1)=(0, 0)$ and the chart



Subcase $g(0, 0)=1$:



The output space of n-sequences contains all sequences with 0-strings of an even length. The number $E(n)$ of these sequences satisfies the recursion

$$E(n)=E(n-1)+E(n-2)\,. \tag{2.8}$$

This is the familiar relation for (f^*, g^*), which is therefore again not defeated.

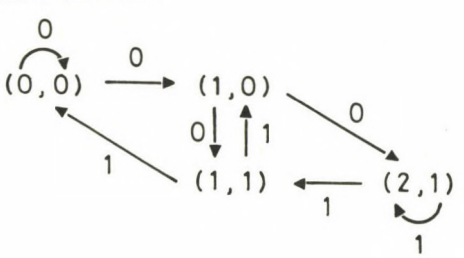
Subcase $g(0, 0)=0$:

Here all output sequences with 0-strings of length $\geqq 2$ and arbitrary 1-strings occur. Their number is bigger than $E(n)$.

Ad (b): We distinguish the cases

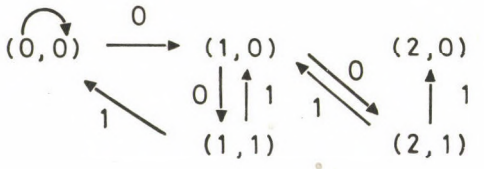
(A) $(f, g)(0, 0)=(0, 0)$ and $(f, g)(2, 1)=(1, 1)$
(B) $(f, g)(0, 0)=(0, 0)$ and $(f, g)(2, 1)=(1, 0)$
(C) $(f, g)(0, 0)=(0, 1)$ and $(f, g)(0, 1)=(0, 1)$
(D) $(f, g)(0, 0)=(0, 1)$ and $(f, g)(0, 1)=(0, 0)$.

Case (A): We have the chart

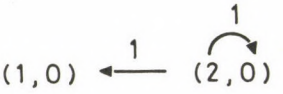


where the set $\mathscr{A}(n)$ of n-sequences with arbitrary 1-strings and 0-strings of length $\geqq 2$ can be produced. Clearly $|\mathscr{A}(n)| > E(n)$ and this case is excluded.

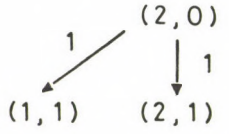
Case (B): We have the chart



and subcases (B_1):



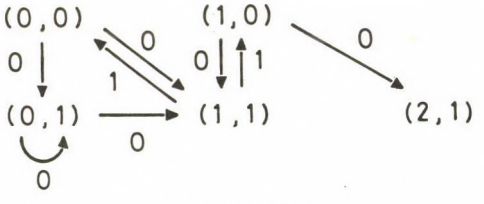
and (B_2):



In the first subcase one can produce $\mathscr{B}_1(n)$, the set of sequences with arbitrary 0-strings and arbitrary 1-strings interrupted by 01-strings of length $\geqq 1$, called "gates".

In the second subcase one can produce $\mathscr{B}_2(n)$, the set of n-sequences with arbitrary 0-strings, 1-strings of even length, 1-strings of even ($\geqq 2$) length followed by a 0, and with 01-strings as gates between these 3 types of strings.

Case (C): We have the chart



Subcase (C_1): $(f, g)(2, 1)=(1, 1)$.

The output space contains $\mathscr{C}_1(n)$, the set of n-sequences with 10-strings between 1-strings and 0-strings of arbitrary length.

Subcase (C_2): $(f, g)(2, 1)=(1, 0)$.

Here we have two more cases.

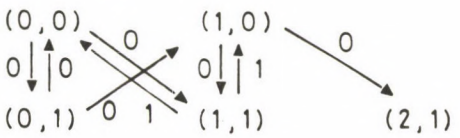
(C_{21}): $(f, g)(2, 0)=(1, 0)$.

$\mathscr{C}_{21}(n)$=set of n-sequences with arbitrary 0-strings, 1-strings of length $\geqq 2$ and 10-strings as gates.

(C_{22}): $(f, g)(2, 0)=(1, 1)$.

$\mathscr{C}_{22}(n)$ = set of n-sequences with arbitrary 0-strings, 1-strings of even length, 1-strings of odd length followed by one 0, and 10-strings as gates.

Case (D): We have the chart



Subcase (D_1): $(f, g)(2, 1)=(1, 1)$.

Subcase (D_2): $(f, g)(2, 1)=(1, 0)$.

Finally we give now the bounds on cardinalities of sets. We observe that

$$|\mathscr{B}_1(n)| = |\mathscr{C}_1(n)| > |\mathscr{C}_{21}(n)|.$$

Further, considering in $\mathscr{C}_{21}(n)$ only 01-strings of length 1 as gates and replacing them by a 11-string of length 1 we get all n-sequences with arbitrary 0-strings and 1-strings of even length. Therefore $|\mathscr{C}_{21}(n)| \geqq E(n)$. Clearly, the same map gives also $|\mathscr{B}_2(n)| \geqq \geqq E(n)$.

Similarly, replacing 10 by 11 we get $|\mathscr{C}_{22}(n)| \geqq E(n)$.

Finally, for subcase (D_1) (and similarly for (D_2)) there are transitions between strings 10, 1x, 1y, 01 and 0z, where x is a 0-string of odd length, y is a 0-string of even length and z is an arbitrary 1-string. It can be shown that $|D_1(n)|, |D_2(n)| \geqq E(n)$.

Our next and main result describes a class of strategies, which includes an optimal strategy. We shall see that (f^*, g^*) is best within this class and therefore an optimal strategy.

Lemma 5. For $\beta \geqq 4$ there is an optimal strategy, say (f, g), satisfying either

$$(f, g)(s, 0)=(0, 0) \quad \text{and} \quad (f, g)(s, 1)=(1, 1) \tag{2.9}$$

for all $2 \leqq s \leqq \beta-2$ or

$$(f, g)(s, 0)=(1, 0) \quad \text{and} \quad (f, g)(s, 1)=(0, 1) \tag{2.10}$$

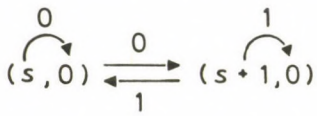
for all $2 \leqq s \leqq \beta-2$.

This result says that the state transition chart of some optimal strategy has the form:

For $2 \leqq s \leqq \beta-2$, either

$$\begin{matrix} (s, 0)\rightarrow \\ \leftarrow(s, 1) \end{matrix} \quad \text{or} \quad \begin{matrix} \leftarrow(s, 0) \\ (s, 1)\rightarrow . \end{matrix}$$

Proof. First notice that validity of (2.9) or (2.10) for $2 \leqq s \leqq \beta - 2$ implies validity of (2.9) for all $2 \leqq s \leqq \beta$ or validity of (2.10) for all $2 \leqq s \leqq \beta$, because we have situations like



and a rate log 2.

Now, by symmetry it suffices to consider the case $f(s, 0) = 0$. This and Lemma 1 leave us with the cases

(a) $(f, g)(s, 0) = (0, 0), \quad (f, g)(s, 1) = (1, 1).$
(b) $(f, g)(s, 0) = (0, 0), \quad (f, g)(s, 1) = (1, 0).$
(c) $(f, g)(s, 0) = (0, 1), \quad (f, g)(s, 1) = (1, 1).$
(d) $(f, g)(s, 0) = (0, 1), \quad (f, g)(s, 1) = (1, 0).$

Clearly (f^*, g^*) has property (2.9).

We want to prove that an optimal strategy satisfies (a). We need to check only that in other cases the strategies cannot be better than (f^*, g^*).

We consider (b) first. There are the following 2 subcases:

(b1) $(f, g)(s-1, 0) = (0, 0)$
(b2) $(f, g)(s-1, 0) = (0, 1).$

For (b1) the following set of states is a closed set:

$$\Lambda = \{(s-1, 0), \quad (u, z) \quad \text{with} \quad s \leqq u \leqq \beta\}.$$

From the induction hypothesis, this leads to the conclusion that for one of the $M^t(\ ,\ ,\ )$ assigned to a state from this set

$$M^t(\ ,\ ,\ ) = M^{t-1}(\ ,\ ,\ ) + M^{t-\beta-s+1}(\ ,\ ,\ ).$$

The strategy cannot be optimal.

In case (b2) by Lemma 2 there exists a path from $(s-1, 0)$ to $(s-1, 1)$ of length not greater than $2(s-1)$. Therefore there exists a cycle from $(s-1, 1)$ to itself of length not greater than $2s \leqq \beta$. It is easy to prove that

$$M^t(s-1, 1) = M^{t-1}(s-1, 1) + M^{t-2s}(s-1, 1),$$

which shows that this strategy cannot be optimal.

For case (c) there are two possibilities:

(c1) $(f, g)(s-1, 0) = (0, 0)$
(c2) $(f, g)(s-1, 1) = (0, 1).$

(c2) is obviously poor. We study only subcase (c1). If $\beta=2s$ we consider the part of the states (s', z) with $s'\geqq s$. Then case (c2) is just case (b). We have already proved that such a strategy cannot be optimal. If $\beta\geqq 2s+1$, then using Lemma 2 we can prove that an inequality

$$M^t(\ .\ ,\ .\ )\geqq M^{t-1}(\ .\ ,\ .\ )+M^{t-\beta}(\ .\ ,\ .\ )$$

will hold. This is just what we need.

Case (d). There are two possibilities:

$$\text{case (d1) } (f, g)(s-1, 0)=(0, 0)$$
$$\text{case (d2) } (f, g)(s-1, 1)=(0, 1).$$

For subcase (d1), the strategy is definitely poor, because of the existence of a cycle of length 3 from $(s-1, 0)$ to itself. Subcase (d2) is similar to subcase (c1). An inequality

$$M^t(\ .\ ,\ .\ )\geqq M^{t-1}(\ .\ ,\ .\ )+M^{t-\beta}(\ .\ ,\ .\ )$$

can be derived, and the strategy cannot be better than (f^*, g^*).

Finally, to prove the claimed optimality of (f^*, g^*) the only thing left is to determine the form of the optimal strategy in Lemmas 3, 4 and 5 for $s=0, 1, \beta-1$ and β.

It is easy to check that in case $(f, g)(1, 1)=(1, 0)$ or $(f, g)(\beta-1, 0)=(0, 1)$, we have inequalities

$$M^t(1, 0)\geqq M^{t-1}(1, 0)+M^{t-\beta+1}(\beta-1, 1)$$

and

$$M^t(\beta-1, 1)\geqq M^{t-1}(\beta-1, 1)+M^{t-\beta}(1, 0).$$

This shows that such a strategy is poorer than (f^*, g^*). Therefore we must have $(f, g)(1, 1)=(1, 1)$ and $(f, g)(\beta-1, 0)=(0, 0)$. It is also readily seen that the form of such an optimal strategy at $s=0$ and $s=\beta$ is the same at that of (f^*, g^*). The proof of the theorem is complete.

3. A first result in a probabilistic model

Here we determine $\eta_\alpha(\pi, \beta, \varphi, P_X)$, which is defined in (1.5), in one of the simplest non-trivial cases, that is,

$$\alpha=2,\quad \pi=\infty,\quad \beta=2,\quad \varphi=0\quad \text{and}\quad P_X(0)=P_X(1)=\frac{1}{2}. \tag{3.1}$$

Since the main results, Theorem 2 and Theorem 3, are stated in terms of concepts which arise during the analysis, we state them in their natural contexts.

We need more notation.

The set of all strategies for blocklength n is simply denoted by F^n.

We will use two kinds of states of the box. The first one is the number of 1's in the box after the last ball X_t has entered the box. We denote this state of the box by S_t. It takes its values in $\{0, 1, 2\}$. Another kind of state of the box is the number of 1's in the box after Y_{t-1} left and before X_t entered the box. This quantity will be denoted by $\hat{S}_t$. It takes values in $\{0, 1\}$. Both, S_t and $\hat{S}_t$, are random variables. Clearly

$$S_t = \hat{S}_t + X_t. \tag{3.1'}$$

After a strategy $f = (f_1, f_2, \ldots)$ has been fixed and $Y^{t-1} = Y_1 \ldots Y_{t-1}$ is the output sequence before time t, then

$$Y_t = f_t(S_t, Y^{t-1}) \tag{3.2}$$

is the output at time t.

After X_{t+1} has entered the box it moves to the state S_{t+1}, where

$$S_{t+1} = S_t - Y_t + X_{t+1}, \quad \hat{S}_{t+1} = S_t - Y_t. \tag{3.3}$$

Conditional on $Y^{t-1} = y^{t-1}$ $\hat{S}_t$ has a distribution

$$P(\hat{S}_t = \varepsilon | Y^{t-1} = y^{t-1}) = P_\varepsilon; \quad \varepsilon = 0, 1. \tag{3.4}$$

Since X_t is independent of $(\hat{S}_t, Y^{t-1})$ by (3.1)

$$P(S_t = 0 | Y^{t-1} = y^{t-1}) = \frac{P_0}{2}$$

$$P(S_t = 1 | Y^{t-1} = y^{t-1}) = \frac{1}{2}$$

$$P(S_t = 2 | Y^{t-1} = y^{t-1}) = \frac{P_1}{2}. \tag{3.5}$$

If $S_t = 0$ (resp. 2), then $\mathcal{O}$ has to send a 0 (resp. 1). Therefore at time t strategies can only differ on the domain $\{(1, y^{t-1}) : y^{t-1} \in \mathcal{Y}^t\}$. How do we find a good strategy? We are guided by the idea to minimize the conditional entropy $H(Y_t | Y^{t-1} = y^{t-1})$ and we define therefore:

$f = (f_1, f_2, \ldots)$ is locally optimal at t, if in the notation of (3.4)

$$f_t(1, y^{t-1}) = \begin{cases} 0 & \text{if} \quad P_0 \geqq P_1 \\ 1 & \text{if} \quad p_0 < P_1 \end{cases}. \tag{3.6}$$

For $t = 1$ we use the letters p and $q = \bar{p}$ instead of p_0 and p_1, that is,

$$P(\hat{S}_1 = 0) = p, \quad P(\hat{S}_1 = 1) = q. \tag{3.7}$$

We allow all initial distributions (p, q), but by symmetry there is no loss of generality, if we always assume

$$p \leqq q. \tag{3.8}$$

We call a strategy $f = (f_1, f_2, \ldots)$ *normal*, if it is locally optimal for every $t = 1, 2, \ldots$.

Our first result can now be stated.

Theorem 2. Under assumptions (3.1) for all p the normal strategy is optimal.

The proof is based on 5 lemmas. They concern estimates on entropies of random variables involving also mixed 1 step strategies $g(r)$, $0 \leqq r \leqq 1$, for which 1 is send with probability r and 0 is send with probability $1-r$. For the initial value p and any r, $0 \leqq r \leqq 1$, let us consider now those strategies which use $g(r)$ at the first step and subsequently, for $t \geqq 2$, follow the normal strategy.

The entropy of the thus produced output process Y^n depends on n, r and p. This justifies the notation

$$H_n^*(p; r) = H(Y^n). \tag{3.9}$$

Notice that by (3.8) $g(1)$ is locally optimal in the first step.

We present and prove now the 5 lemmas, which make statements about $H_n^*(p; r)$. The last lemma says that we shall follow the locally optimal strategy also in the first step, if we use it subsequently. Thus Theorem 2 follows inductively.

In the sequel we frequently use the notation

$$\bar{\mu} = 1 - \mu \quad \text{for} \quad 0 \leqq \mu \leqq 1. \tag{3.10}$$

Lemma 1. $H_n^*(p; r)$ is convex $(\cap)$ in p.

Proof. By our definitions the conditional probabilities $\Pr(Y^n = y^n | \hat{S}_1 = \varepsilon)$ do not depend on p. The unconditional probabilities

$$\Pr(Y^n = y^n; P) = \Pr(Y^n = y^n | \hat{S}_1 = 0)p + \Pr(Y^n = y^n | \hat{S}_1 = 1)\bar{p} \tag{3.11}$$

are therefore linear in p and thus

$$\Pr(Y^n = y^n; \lambda p + \bar{\lambda} p') = \lambda \Pr(Y^n = y^n; p) + \bar{\lambda} \Pr(Y^n = y^n; p'). \tag{3.12}$$

Convexity of $H_n^*(p; r)$ then follows from the convexity of the entropy function.

Lemma 2. $\lambda H_n^*(p; r) + \bar{\lambda} H_n^*(p'; r) \geqq H_n^*(\lambda p + \bar{\lambda} p'; r) + \lambda h(p) + \bar{\lambda} h(p') - h(\lambda p + \bar{\lambda} p')$, where h is the binary entropy function.

Proof. By (3.11) $H_n^*(p; r)$ can be written in the form $H(pP + \bar{p}Q)$ for two distribution P and Q. The inequality can therefore be restated as

$$h(\lambda p + \bar{\lambda} p') - \lambda h(p) - \bar{\lambda} h(p') \geqq H(\lambda(pP + \bar{p}Q) + \bar{\lambda}(p'P + \bar{p}'Q)) -$$
$$- \lambda H(pP + \bar{p}Q) - \bar{\lambda} H(p'P + \bar{p}'Q).$$

The expression to the left can be interpreted as a mutual information $I(J \wedge K)$, where $\Pr(J=1)=\lambda$, $\Pr(J=2)=\bar{\lambda}$ and

$$\Pr(K=1|J=1)=p, \quad \Pr(K=2|J=1)=\bar{p},$$
$$\Pr(K=1|J=2)=p', \quad \Pr(K=2|J=2)=\bar{p}'.$$

Postposing to the "channel" $\begin{pmatrix} p & \bar{p} \\ p' & \bar{p}' \end{pmatrix}$ the "channel" $\begin{pmatrix} P \\ Q \end{pmatrix}$ results in the "channel" $\begin{pmatrix} pP+\bar{p}Q \\ p'P+\bar{p}'Q \end{pmatrix}$, which for the input variable J induces the output variable L. Now $I(J \wedge L)$ is the expression to the right of the inequality, which is thus shown to be a special case of the data-processing inequality.

Lemma 3. $H_n^*(p; 1)$ is monotone increasing in $[0, 1/2]$.

Proof. We proceed by induction in n. Since $H_1^*(p; 1)=h\left(\frac{p}{2}\right)$ and since $\frac{d}{dr}h(r)=\log\frac{1-r}{r}$, we get

$$\frac{dH_1^*(p; 1)}{dp}=\frac{1}{2}\log\frac{2-p}{p}\geqq 0 \quad \text{for} \quad p\in[0, 1/2]$$

and the case $n=1$ is established.

By Lemma 1 we know that $H_n^*(p; 1)$ is convex. It suffices therefore to show that

$$\frac{d}{dp}H_n^*(p; 1)_{|p=0.5}\geqq 0. \tag{3.13}$$

We use now and also later a basic recursion. For the underlying strategy

$$\begin{aligned} H_n^*(p; 1) &= H(Y_1)+H(Y_n, \ldots, Y_2|Y_1) \\ &= h\left(\frac{p}{2}\right)+H(Y_n, \ldots, Y_2|Y_1=0)\Pr(Y_1=0) \\ &\quad +H(Y_n, \ldots, Y_2|Y_1=1)\Pr(Y_1=1) \\ &= h\left(\frac{p}{2}\right)+H_{n-1}^*(0; 1)\frac{p}{2}+H_{n-1}^*\left(\frac{1}{2-p}; 0\right)\left(1-\frac{p}{2}\right) \end{aligned}$$

and therefore

$$H_n^*(p; 1)=h\left(\frac{p}{2}\right)+\frac{p}{2}H_{n-1}^*(0; 1)+\left(1-\frac{p}{2}\right)H_{n-1}^*\left(\frac{1-p}{2-p}; 1\right). \tag{3.14}$$

Hence

$$\frac{d}{dp} H_n^*(p;1) = \frac{1}{2}\log 3 + \frac{1}{2} H_{n-1}^*(0;1) - \frac{1}{2} H_{n-1}^*\left(\frac{1}{3};1\right) + \frac{3}{4}\frac{d}{dp} H_{n-1}^*\left(\frac{1-p}{2-p};1\right)_{|p=0.5} \tag{3.15}$$

Using (3.14) again and the induction hypothesis we obtain

$$\begin{aligned}
&\frac{d}{dp} H_{n-1}^*\left(\frac{1-p}{2-p};1\right)_{|p=0.5} \\
&= -\frac{1}{(1.5)^2}\left[\frac{1}{2} H_{n-2}^*(0;1) - \frac{1}{2} H_{n-2}^*\left(\frac{2}{5};1\right) + \frac{1}{2}\log 5 \right. \\
&\quad \left. - \frac{1}{2}\frac{d}{dp}\left(H_{n-2}^*\left(\frac{1}{3-p};1\right)\right)_{|p=0.5}\right] \geqq \\
&\geqq -\frac{2}{9} H_{n-2}^*(0;1) + \frac{2}{9} H_{n-2}^*\left(\frac{2}{5};1\right) - \frac{2}{9}\log 5.
\end{aligned}$$

Substituting this in (3.15) we get

$$\begin{aligned}
\frac{d}{dp} H_n^*(p;1)_{|p=0.5} \geqq{}& \frac{1}{2}\log 3 + \frac{1}{2} H_{n-1}^*(0;1) - \frac{1}{2} H_{n-1}^*\left(\frac{1}{3};1\right) \\
&- \frac{1}{6} H_{n-2}^*(0,1) + \frac{1}{6} H_{n-2}^*\left(\frac{2}{5};1\right) - \frac{1}{6}\log 5.
\end{aligned}$$

Applying the basic recursion (3.14) to $H_{n-1}^*(0;1)$ and $H_{n-1}^*\left(\frac{1}{3};1\right)$ we continue the derivation with

$$\begin{aligned}
\frac{d}{dp}(H_n^*(p;1)_{|p=0.5} \geqq{}& \frac{1}{2}\log 3 + \frac{1}{2} H_{n-2}^*\left(\frac{1}{2};1\right) - \frac{1}{2} h\left(\frac{1}{6}\right) \\
&- \frac{1}{12} H_{n-2}^*(0;1) - \frac{3}{8} H_{n-2}^*\left(\frac{2}{5};1\right) \\
&- \frac{1}{6} H_{n-2}^*(0;1) + \frac{1}{6} H_{n-2}^*\left(\frac{2}{5};1\right) - \frac{1}{6}\log 5 \\
\geqq{}& \frac{1}{2}\log 3 - \frac{1}{2} h\left(\frac{1}{6}\right) - \frac{1}{6}\log 5
\end{aligned}$$

by induction hypothesis.

Since $\frac{1}{2}\log 3-\frac{1}{2}h\left(\frac{1}{6}\right)-\frac{1}{6}\log 5=\frac{1}{4}\log\frac{5}{4}\geqq 0$, the proof is complete.

Lemma 4. For $p\leqq\frac{1}{2}$ we have

$$H_n^*(p;1)\leqq H_n^*(p;0).$$

Proof. Since $H_n^*\left(\frac{1}{2};0\right)=H_n^*\left(\frac{1}{2};1\right)$, it suffices to show that

$$\frac{d}{dp}(H_n^*(p;0)-H_n^*(p;1))\leqq 0 \quad \text{for} \quad 0\leqq p\leqq 0.5. \tag{3.16}$$

The proof of this inequality repeatedly makes use of (3.14) and needs formidable calculations.

Since $H_n^*(p;0)=H_n^*(1-p;1)$ we get from (3.14)

$$H_n^*(p;0)=h\left(\frac{1-p}{2}\right)+\frac{1-p}{2}H_{n-1}^*(0;1)+\frac{1+p}{2}H_{n-1}^*\left(\frac{p}{1+p};1\right). \tag{3.17}$$

Also from (3.14)

$$H_{n-1}^*\left(\frac{p}{1+p};1\right)=h\left(\frac{1}{2+p}\right)+\frac{p}{2(1+p)}H_{n-2}^*(0;1)+\frac{2+p}{2(1+p)}H_{n-2}^*\left(\frac{1}{2+p};1\right). \tag{3.18}$$

From these two equations we deduce

$$\frac{d}{dp}H_n(p;0)=\frac{1}{2}\log\frac{1-p}{1+p}-\frac{1}{2}H_{n-1}^*(0;1) \tag{3.19}$$
$$+\frac{1}{2}H_{n-1}^*\left(\frac{p}{1+p};1\right)+\frac{1+p}{2}\frac{d}{dp}H_{n-1}^*\left(\frac{p}{1+p};1\right)$$

and

$$\frac{d}{dp}H_{n-1}^*\left(\frac{p}{1+p};1\right)=\frac{2+p}{2(1+p)}\frac{d}{dp}H_{n-2}^*\left(\frac{1}{2+p};1\right) \tag{3.20}$$
$$+\frac{1}{2}\left(\frac{1}{1+p}\right)^2\left[\log\frac{2+p}{p}+H_{n-2}^*(0;1)-H_{n-2}^*\left(\frac{1}{2+p};1\right)\right].$$

Since by Lemma 3 the first term on the right side of (3.20) is negative we conclude that

$$\frac{d}{dp}H_{n-1}^*\left(\frac{p}{1+p};1\right)\leqq\frac{1}{2}\frac{1}{(1+p)^2}\left[\log\frac{2+p}{p}+H_{n-2}^*(0;1)-H_{n-2}^*\left(\frac{1}{2+p};1\right)\right]. \tag{3.21}$$

This and (3.19) imply

$$\frac{d}{dp} H_n^*(p;0) \leqq \frac{1}{2}\left[\log\frac{1-p}{2+p} - H_{n-1}^*(0;1) + H_{n-1}^*\left(\frac{p}{1+p};\ 1\right) + \right.$$
$$\left. + \frac{1}{2(1+p)}\left(\log\frac{2+p}{p} + H_{n-2}^*(0;1) - H_{n-2}^*\left(\frac{1}{2+p};\ 1\right)\right)\right]. \quad (3.22)$$

Similarly, we can prove

$$\frac{d}{dp} H_n^*(p;1) \geqq \frac{1}{2}\left[H_{n-1}^*(0;1) + \log\frac{2-p}{p} - H_{n-1}^*\left(\frac{1}{3-p};\ 1\right) - \right.$$
$$\left. - \frac{1}{2(2-p)}\left(H_{n-2}^*(0;1) - H_{n-2}^*\left(\frac{1}{3-p};\ 1\right) + \log\frac{3-p}{1-p}\right)\right].$$

Using these two inequalities, we deduce

$$\frac{d}{dp}[H_n^*(p;0) - H_n^*(p;1)] \leqq -H_{n-1}^*(0;1) - \frac{1}{2}\log\frac{(2+p)(1+p)}{p(1-p)} +$$
$$+ \frac{1}{2} H_{n-1}^*\left(\frac{1-p}{2-p};\ 1\right) + \frac{1}{2} H_{n-1}^*\left(\frac{p}{1+p};\ 1\right) +$$
$$+ \frac{1}{4(1+p)}\left[H_{n-2}^*(0;1) - H_{n-2}^*\left(\frac{1}{2+p}\right);\ 1\right) + \log\frac{2+p}{p}\Bigg] +$$
$$+ \frac{1}{4(2-p)}\left[H_{n-2}^*(0;1) - H_{n-2}^*\left(\frac{1}{3-p};\ 1\right) + \log\frac{3-p}{1-p}\right].$$

Using (3.14) for all the terms with parameter $n-1$, we obtain

$$\frac{d}{dp}[H_n^*(p;0) - H_n^*(p;1)] \leqq -H_{n-2}^*\left(\frac{1}{2};\ 1\right) + \frac{1-p}{4(2-p)} H_{n-2}^*(0;1)$$
$$+ \frac{3-p}{4(2-p)} H_{n-2}^*\left(\frac{1}{3-p};\ 1\right) + \frac{1}{2} h\left(\frac{1-p}{2(2-p)}\right) + \frac{p}{4(1+p)} H_{n-2}^*(0;1)$$
$$+ \frac{2+p}{4(1+p)} H_{n-2}^*\left(\frac{1}{2+p};\ 1\right) + \frac{1}{2} h\left(\frac{p}{2(1+p)}\right) - \frac{1}{2}\log\frac{(2-p)(1+p)}{p(1-p)}$$
$$+ \frac{1}{4(1+p)}\left[H_{n-2}^*(0;1) - H_{n-2}^*\left(\frac{1}{2+p};\ 1\right) + \log\frac{2+p}{p}\right]$$
$$+ \frac{1}{4(2-p)}\left[H_{n-2}^*(0;1) - H_{n-2}^*\left(\frac{1}{3-p};\ 1\right) + \log\frac{3-p}{1-p}\right].$$

Noting that by Lemma 3 $H^*_{n-2}(p;1)$ is monotone increasing in p and in particular $H^*_{n-2}\left(\frac{1}{2};\ 1\right) \geqq H^*_{n-2}(p;1)$ for all $p \leqq \frac{1}{2}$ we obtain

$$\frac{d}{dp}[H^*_n(p;0)-H^*_n(p;1)] \leqq -\frac{1}{2}\log\frac{(2-p)(1+p)}{p(1-p)} + \frac{1}{2}h\left(\frac{1-p}{2(2-p)}\right) + \frac{1}{2}h\left(\frac{p}{2(1+p)}\right)$$

$$+ \frac{1}{4(1+p)}\log\frac{2+p}{p} + \frac{1}{4(2-p)}\log\frac{3-p}{1-p}$$

$$= \frac{1}{4}\log\frac{16p(1-p)}{(3-p)(2+p)} \leqq \frac{1}{4}\log\frac{4}{(3-p)(2+p)}$$

$$\leqq \frac{1}{4}\log\frac{4}{6+p-p^2} \leqq \frac{1}{4}\log\frac{4}{6} \leqq 0.$$

Lemma 4 is proved.

Lemma 5. For $0 \leqq p \leqq \frac{1}{2}$ and $0 \leqq r \leqq 1$ we have

$$H^*_n(p;1) \leqq H^*_n(p;r).$$

Proof. To save notation we introduce $t = 1-r$. We distinguish between the following 3 cases:

$$(1)\quad t \leqq p; \qquad (2)\quad p < t \leqq 1-p; \qquad (3)\quad t > 1-p.$$

Notice that case (3) can be proved in the same way as case (1) by changing r to t and p to $1-p$.

We use the general form of the basic recursion in case (1).

$$H^*_n(p;r) = \frac{p+t}{2}H^*_{n-1}\left(\frac{t}{p+t};\ 1\right) + \left(1-\frac{p+t}{2}\right)H^*_{n-1}\left(\frac{1-p}{2-p-t};\ 1\right) + h\left(\frac{p+t}{2}\right). \tag{3.23}$$

Lower bounding the first summand with Lemma 1, we get

$$H^*_n(p;r) \geqq \frac{p}{2}H^*_{n-1}(0;1) + \frac{t}{2}H^*_{n-1}(1;1) +$$

$$+ \left(1-\frac{p+t}{2}\right)H^*_{n-1}\left(\frac{1-p}{2-p-t};\ 1\right) + h\left(\frac{p+t}{2}\right). \tag{3.24}$$

Since by Lemma 4 $H^*_{n-1}(1;1)=H^*_{n-1}(0;0)\geqq H^*_{n-1}(0;1)$ we conclude that

$$H^*_n(p;r)\geqq \frac{p}{2}H^*_{n-1}(0;1)+\frac{t}{2}H^*_{n-1}(0;1)+$$

$$+\left(1-\frac{p+1}{2}\right)H^*_{n-1}\left(\frac{1-p}{2-p-t};\ 1\right)+h\left(\frac{p+t}{2}\right). \tag{3.25}$$

Application of Lemma 2 to the 2 central summands gives

$$H^*_n(p;1)\geqq \frac{p}{2}H^*_{n-1}(0;1)+\left(1-\frac{p}{2}\right)H^*_{n-1}\left(\frac{1-p}{2-p};\ 1\right)+\frac{t}{2}h(0)+$$

$$+\left(1-\frac{p+t}{2}\right)h\left(\frac{1-p}{2-p-t}\right)-\left(1-\frac{p}{2}\right)h\left(\frac{1-p}{2-p}\right)+h\left(\frac{t+p}{2}\right). \tag{3.26}$$

The first two summands can be rewritten via (3.14) and by some manipulations

$$H^*_n(p;r)\geqq H^*_n(p;1)-h\left(\frac{p}{2}\right)+\left(1-\frac{p+t}{2}\right)h\left(\frac{1-p}{2-p-t}\right)-$$

$$-\left(1-\frac{p}{2}\right)h\left(\frac{1-p}{2-p}\right)+h\left(\frac{t+p}{2}\right)=$$

$$=H^*_n(p;1)-\frac{p+t}{2}h\left(\frac{p}{p+t}\right)+\frac{1}{2}h(t). \tag{3.27}$$

Since $t\leqq p\leqq\frac{1}{2}$, we have $t\leqq\frac{t}{p+t}$ and $h(t):h\left(\frac{t}{p+t}\right)\geqq t:\frac{t}{p+t}$.

Therefore $h(t)\geqq(p+t)h\left(\frac{p}{p+t}\right)$ and finally $H^*_n(p;r)\geqq H^*_n(p;1)$ in this case.

We now prove the result in case (2).

Here the basic recurrence takes the form

$$H^*_n(p;r)=\frac{p+t}{2}H^*_{n-1}\left(\frac{p}{p+t};\ 1\right)+\left(1-\frac{p+t}{2}\right)H^*_{n-1}\left(\frac{1-t}{2-p-t};\ 1\right)+h\left(\frac{p+t}{2}\right). \tag{3.28}$$

This differs from (3.23) only insofar as at the right side p and $\pm$ are exchanged. Instead of (3.27) we get therefore now

$$H^*_n(p;r)\geqq H^*_n(t;1)-\frac{p+t}{2}h\left(\frac{t}{p+t}\right)+\frac{1}{2}h(p). \tag{3.29}$$

Since $p<t\leqq 1-p$, we have $p\leqq \frac{p}{p+t}$ and $h(p):h\left(\frac{p}{p+t}\right)\geqq p:\frac{p}{p+t}$. Therefore we get in this case

$$H_n^*(p;r)\geqq H_n^*(t;1). \tag{3.30}$$

For $t\leqq \frac{1}{2}$, since $p<t$ by Lemma 3

$$H_n^*(t;r)\geqq H_n^*(p;1)$$

and thus the desired result

$$H_n^*(p;r)\geqq H_n^*(p;1), \tag{3.31}$$

and for $t>\frac{1}{2}$ by Lemma 4

$$H_n^*(t;1)\geqq H_n^*(t;0)=H_n^*(r;1),$$

by Lemma 3 $H_n^*(r;1)\geqq H_n^*(p;1)$, and again the inequality (3.31).

We derive now a formula for the limiting entropy rate $\eta_2(\infty, 2, 0, P_X)$. Suppose that (w.l.o.g.) $\hat{S}_0=0$ and that we follow the optimal strategy. For its analysis we introduce the events $E_k=\{Y^k=01010\ldots\}$ and $D_k=E_k\backslash E_{k+1}$. The D_k's are disjoint and $q(k)\triangleq \text{Prob.}(D_k)$ satisfies

$$\sum_{n=1}^{\infty} q(k)=1. \tag{3.32}$$

Theorem 3. For $P_X(0)=P_X(1)=1/2$

$$\eta_2(\infty, 2, 0, P_X)=\frac{H(q)}{\sum_{k=1}^{\infty} kq(k)}.$$

Proof. By the grouping axiom for entropy

$$H(Y^i|\hat{S}_0=0)=-\sum_{t=1}^{i=1} q(t)\log(t)$$

$$-(1-q(1)-\cdot-q(i-1))\log(1-q(1)-\cdot-q(i-1))$$

$$+\sum_{t=1}^{i-1} q(t)H(Y^{i-t}|\hat{S}_0=0).$$

Therefore we get

$$\sum_{i=1}^{n} H(Y^i|\hat{S}_0=0) = -\sum_{i=1}^{n}\sum_{t=1}^{i-1} q(t)\log q(t)$$

$$-\sum_{i=1}^{n}(1-q(1)-\cdots-q(i-1))\log(1-q(1)-\cdots-q(i-1))$$

$$+\sum_{i=1}^{n}\sum_{t=1}^{i-1} q(t)H(Y^{i-t}|\hat{S}_0=0) = -\sum_{t=1}^{n-1}(n-t)q(t)\log q(t)$$

$$-\sum_{i=1}^{n}\left(1-\sum_{j=1}^{i-1}q(j)\right)\log\left(1-\sum_{j=1}^{i-1}q(j)\right)+\sum_{i=1}^{n-1}\sum_{t=1}^{n-i} q(t)H(Y^i|\hat{S}_0=0)$$

and consequently

$$H(Y^n|\hat{S}_0=0)+\sum_{i=1}^{n-1}\left(1-\sum_{t=1}^{n-i}q(t)\right)H(Y^i|\hat{S}_0=0)$$

$$\geqq -\sum_{t=1}^{n-i}(n-t)q(t)\log q(t). \tag{3.33}$$

Notice that $$\sum_{i=1}^{n-1}\left(1-\sum_{t=1}^{n-i}q(t)\right)=\sum_{i=1}^{n-1}\sum_{t=n-i+1}^{\infty} q(t)$$

and

$$\sum_{t=1}^{\infty} q(t)+\sum_{i=1}^{n-1}\sum_{t=n-i+1}^{\infty} q(1)=\sum_{i=1}^{n}\sum_{t=n-i+1}^{\infty} q(t)$$

$$=\sum_{s=1}^{\infty}\sum_{i=n+1-s}^{n} q(s)=\sum_{s=1}^{\infty} sq(s). \tag{3.34}$$

Since $H(Y^n|\hat{S}_0=0)\geqq H(Y^i|\hat{S}_0=0)$ for $i\leqq n$ from (3.33) and (3.34) we can derive

$$\sum_{s=1}^{\infty} sq(s)H(Y^n|\hat{S}_0=0)\geqq -\sum_{t=1}^{n-1}(n-t)q(t)\log q(t). \tag{3.35}$$

Since for fixed k $-\sum_{t=1}^{n-1}(n-t)q(t)\log q(t)\geqq(n-k)\sum_{t=1}^{k}-q(t)\log q(t)$ we continue with

$$\frac{1}{n}H(Y^n|\hat{S}_0=0)\geq\frac{(n-k)\sum_{t=1}^{k}-q(t)\log q(t)}{n\sum_{s=1}^{\infty}sq(s)}.$$

Let now first n tend to infinity and then k. We obtain

$$\lim_{n\to\infty}\frac{1}{n}H(Y^n|\hat{S}_0=0)\geqq\frac{H(q)}{\sum_{s=1}^{\infty}sq(s)}. \tag{3.36}$$

Instead of (3.33) we derive first the upper bound

$$H(Y^n|\hat{S}_0=0)+\sum_{i=1}^{n-1}\left(1-\sum_{t=1}^{n-i}q(t)\right)H(Y^i|\hat{S}_0=0)\leqq$$

$$\leqq -n\sum_{t=1}^{n-1}q(t)\log q(t)-\sum_{i=1}^{n}\left(1-\sum_{j=1}^{i-1}q(j)\right)\log\left(1-\sum_{j=1}^{i-1}q(j)\right)\leqq nH(q). \tag{3.37}$$

Therefore

$$H(Y^n|\hat{S}_0=0)+\sum_{i=n-k+1}^{n-1}\sum_{t=n-i+1}^{\infty}q(t)H(Y^i|\hat{S}_0=0)\leqq nH(q)$$

and a fortiori

$$H(Y^{n-k+1}|\hat{S}_0=0)\sum_{i=1}^{k}iq(i)\leqq nH(q). \tag{3.38}$$

Since $H(Y^n|\hat{S}_0=0)\leqq H(Y^{n-k+1}|\hat{S}_0=0)+k$ we derive from (3.38)

$$\overline{\lim_{n\to\infty}}\frac{1}{n}H(Y^n|\hat{S}_0=0)\leqq\frac{H(q)}{\sum_{i=1}^{\infty}iq(i)}$$

for all k and therefore also

$$\overline{\lim_{n\to\infty}}\frac{1}{n}H(Y^n|\hat{S}_0=0)\leqq\frac{H(q)}{\sum_{i=1}^{\infty}iq(i)}.$$

This and (3.36) complete the proof.

Remark. The sequence $q(1)$, $q(2)$, obeys a simple rule, so that we can calculate

$$\frac{H(q)}{\sum_{i=1}^{\infty}iq(i)}=0.5989\ldots.$$

The formula in Theorem 3 has a nice structure. It suggests a general principle for arbitrary sources.

Reference

1. *Ahlswede, R., Ye, J. P., Zhang, Z.*, Creating order under constraints on mind and matter. Submitted to Information and Computation.

Вклад в теорию упорядочения пространств последовательностей

Р. АЛСЬЕДЕ, З. ЗАНГ

(Билефелд)

Настоящая работа продолжает исследования, начатые в [1]. Получены первые результаты об оптимальном упорядочении последовательностей в двух простых, но существенных случаях: невероятностная модель с активной памятью и вероятностная модель с неограниченной пассивной памятью.

R. Ahlswede
Universität Bielefeld
Fakultät für Mathematik
Universitätsstraße
4800 Bielefeld 1
Federal Republic of Germany

Problems of Control and Information Theory, Vol. 18 (4), pp. 223—235 (1989)

AIRCRAFT LANDING CONTROL IN THE PRESENCE OF WINDSHEAR

N. D. BOTKIN, V. M. KEIN, V. S. PATSKO, V. L. TUROVA

(Sverdlovsk, Leningrad)

(Received May 31, 1988)

Traditional laws of aircraft control in the landing operate unsatisfactorily when rapid wind changes occur. For this reason, new ways of aircraft landing control are investigated recently [1–6]. This paper deals with the minimax approach based on the differential game theory methods [7, 8]. Applications of differential game theory to the landing problem were considered in [1–3, 9–13].

1. Introduction

There are many papers [3–6, 9, 14, 15] devoted to the investigation of the aircraft motion during take-off and landing with rapid change of wind velocity (windshear). Physical conditions leading to windshear, its mathematical models and aircraft control are analyzed.

This paper deals with the landing problem for middle-sized aircraft in the conditions of wind disturbance. We consider the aircraft motion along the glide path till the moment of passing the runway (RW) threshold. The information we get about the wind is supposed to be minimal. Namely, we suppose that only the deviation boundary of the wind velocity from its nominal value and the nominal value are known. Any other information about the windshear zone and the internal distribution of wind velocity is absent. So, the problem that appears naturally is to find the minimax closed-loop control which can cope with arbitrary variation of the wind velocity in noted boundaries.

Using methods of the antagonistic differential game theory [7, 8] we obtain the minimax solution for auxiliary linear problems. Further, this solution is applied for computer simulation of the motions in a complete nonlinear system. Simulation results deal with the case when the windshear is stipulated by aircraft flight through microburst zone. The microburst is caused by falling mass of air which hits the ground and gives vortex. The mathematical model of the microburst is taken from [14].

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. Nonlinear system of aircraft landing motion

The aircraft motion during landing is described by a differential equations system of 12th order. The state vector includes three coordinates x, y, z of mass center in the coordinate system connected with the RW surface (Fig. 1), angles of pitch ϑ, yaw ψ, bank γ and corresponding linear and angular velocities. These equations are specified, for example, in [16, 17].

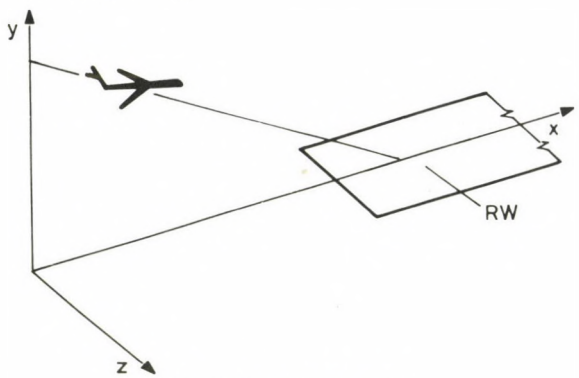


Fig. 1. Coordinate system

The control factors are deviations of the elevator δ_e, the rudder δ_r, the ailerones δ_a and change of thrust force P. Equations of servo and engine dynamics are supplemented to the main system of aircraft motion. So, the noted factors enter in the broadened state vector and new control parameters are the scheduled (commanded) deviations δ_{es}, δ_{rs}, δ_{as}, δ_{ps}. Every parameter has upper and lower restrictions. As a result we get the complete system of differential equations which we write down in vector form as

$$\dot{\xi} = f(\xi, \delta_s, W). \tag{2.1}$$

Here $\delta_s = (\delta_{es}, \delta_{ps}, \delta_{rs}, \delta_{as})^T$ is the control vector, $W = (W_x, W_y, W_z)^T$ is the disturbance vector, consisting of three wind components along the x, y, z axes.

3. Minimax control law

Nominal aircraft motion during landing till the moment of passing RW threshold is a uniform motion (without rotation) along the descending rectilinear glide path.

Control problem is to bring real motion near enough to the nominal motion in the presence of wind disturbance. It is also desirable that the performance of the

control law would not demand any accurate and detailed information about the wind. We assume that for the deviations of wind velocity components W_x, W_y, W_z from the nominal values W_{x0}, W_{y0}, W_{z0} only approximate characteristics of restrictions are known. The nominal values are assumed to be known as well. To solve the problem we apply minimax approach using the methods of differential games.

Effective computer programs have been created [9, 12, 18–20] to find optimal control laws (strategies) in linear differential games with fixed terminal moment and convex payment function, depending on two coordinates of the state vector. System (2.1) is nonlinear. However, we can linearize it, solve auxiliary differential games and use the results to the initial nonlinear system. So, having the nominal values W_{x0}, W_{y0} and W_{z0}, the glide path inclination, the nominal relative velocity, we calculate values of the state variables, corresponding to the nominal motion of system (2.1). The linearization of system (2.1) with respect to the nominal motion gives a linear controllable system, desintegrating to two subsystems of vertical (longitudinal) and lateral motions. The state vector of the vertical motion (VM) subsystem consists of deviations Δx, Δy and some quantities which determine these deviations. The state vector of lateral motion (LM) subsystem consists of deviation Δz and quantities which determine Δz.

For each of the subsystems we consider an auxiliary differential game with fixed terminal time T, geometric restrictions on control variables and wind disturbances and with convex payment function depending on two state vector coordinate at the moment T. In the VM subsystem such coordinates are Δy and $\Delta \dot{y}$, in the LM system these coordinates are Δz and $\Delta \dot{z}$. The first player chooses the control variables to minimize the payment function. The second player, choosing the wind disturbances, maximizes the payment function. It is not necessary to give the moment T any physical meaning in auxiliary problems.

Variables δ_{ps} and δ_{rs} in system (2.1) are intended for the relative velocity and sideslip angle stabilization near the nominal values. So, it is not natural to find closed-loop laws for δ_{ps} and δ_{rs} using the solution of the auxiliary problems mentioned above. Therefore, formulating the auxiliary problems, we assume that the thrust force is a constant and is equal to its nominal value (i.e. $\Delta\delta_{ps} \equiv 0$) and the variation of $\Delta\delta_{rs}$ satisfies a linear differential equation corresponding to traditional rudder control. With that we omit the restriction on $\Delta\delta_{rs}$. As a result we obtain the only control factor $\Delta\delta_{es}$ in the VM subsystem and the control factor $\Delta\delta_{as}$ in the LM subsystem.

To take into account the inertial character of wind velocity variations along the motion, we suppose that variables ΔW_x, ΔW_y, ΔW_z satisfy additional linear differential equations, for example,

$$\Delta \dot{W}_x = k_1(\Delta F_x - \Delta W_x)$$
$$\Delta \dot{F}_x = k_2(w_x - \Delta F_x). \tag{3.1}$$

Here w_x is a new independent variable, constants k_1 and k_2 determine the inertial character of ΔW_x. Similar equations are considered for ΔW_y and ΔW_z. Variables w_x, w_y, w_z are interpreted as new disturbance factors. We add the equations for ΔW_x and ΔW_y to the VM subsystem and the equations for ΔW_z to the LM subsystem.

Solving the auxiliary problems on computer, we find optimal laws for $\Delta\delta_{es}$ and $\Delta\delta_{as}$. These laws are realized by means of sets K_{es} and K_{as} of switch lines [9, 13, 18, 19]. Both sets are defined on a collection of moments τ_i of reverse time counting off the terminal moment T. The sets K_{es} and K_{as} give the desired control laws for components δ_{es} and δ_{as} to initial system (2.1). Using these laws we prognose the time remained till the moment of passing the RW threshold. Depending on the time prognose, certain switch lines are used to choose values δ_{es} and δ_{as}. Simulating motions of system (2.1), we assume that the control factors δ_{ps} and δ_{rs} are obtained by means of control laws accepted nowadays.

So, speaking about the minimax control, we mean the way of finding control factors δ_{es} and δ_{as} from the auxiliary linear differential games. The factors δ_{ps} and δ_{rs} are constructed by traditional methods.

4. The auxiliary linear differential games

The linear VM system is

$$\dot{\mathbf{x}} = A_*\mathbf{x} + B_* u + C_* v, \qquad \mathbf{x} \in R^{11}, \tag{4.1}$$

$$A_* = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.050 & 0 & -0.097 & -2.642 & 0 & 0.063 & 0.050 & 0 & 0.097 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.241 & 0 & -0.639 & 45.278 & 0 & 1.448 & -0.241 & 0 & 0.638 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.007 & -0.501 & -0.526 & -0.383 & 0 & 0 & -0.007 & 0 \\
 & & & & & & -4 & 0 & 0 & 0 & 0 \\
 & & & & & & 0 & -0.5 & 0.50 & 0 & 0 \\
 & & & 0 & & & 0 & 0 & -3 & 0 & 0 \\
 & & & & & & 0 & 0 & 0 & -0.5 & 0.5 \\
 & & & & & & 0 & 0 & 0 & 0 & -3
\end{pmatrix}$$

$$B_* = (0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0)^T, \qquad C_* = \begin{pmatrix} 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0 \end{pmatrix}^T,$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_{11})^T, \qquad u = \Delta\delta_{es}, \qquad v = (w_x, w_y)^T.$$

Here $x_1 = \Delta x$, $x_3 = \Delta y$ are deviations from nominal motion in x and y, respectively; $x_5 = \Delta\vartheta$ is a deviation of pitch angle. The coordinate x_7 is a deviation of the elevator from its trim position. By means of variables x_8, x_9 we describe a variation of ΔW_x. The corresponding equations coincide with (3.1), $x_8 = \Delta W_x$. Similarly, variables x_{10}, x_{11} are used for description of ΔW_y variation ($x_{10} = \Delta W_y$). The control factor of the first player is the scheduled deviation $\Delta\delta_{es}$ of the elevator. The parameters w_x, w_y are used to obtain wind disturbances and belong to the second player.

The restrictions are the following:

$$|\Delta\delta_{es}| \leqq 10 \text{ deg} \frac{\pi}{180}, \qquad |w_x| \leqq 10 \text{ m sec}^{-1}, \qquad |w_y| \leqq 5 \text{ m sec}^{-1}.$$

Introduce a function φ_* which depends on coordinates $x_3 = \Delta y$ and $x_4 = \Delta\dot{y}$. Let M_* be a convex hexagon on the plane x_3, x_4 with apexes $(-3, 0)$, $(-3, 1)$, $(0, 1)$, $(3, 0)$, $(3, -1)$, $(0, -1)$. Suppose

$$\varphi_*(x_3, x_4) = \min \{c \geqq 0 : (x_3, x_4)^T \in cM_*\}.$$

Consider an antagonistic differential game with dynamics (4.1), fixed terminal moment T and payment φ_*. The first player tries to minimize values of the function φ_* at the moment T. The aim of the second player is opposite. The set M_* can be considered as a tolerance for deviations $x_3 = \Delta y$ and $x_4 = \Delta\dot{y}$ at the moment T. The function φ_* indicates a deviation from the tolerance. The optimal strategy of the first player in game (4.1) will be used to define δ_{es} in system (2.1).

The linear LM system is

$$\dot{\mathbf{x}} = A^*\mathbf{x} + B^*u + C^*v, \qquad \mathbf{x} \in R^{11}, \tag{4.2}$$

$$A^* = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.077 & -5.555 & 0 & 9.272 & 0 & -1.485 & 0 & 0.077 & 0 & 0 \\
0 & 0 & 0 & 1.001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.013 & -0.933 & -0.259 & -0.088 & -0.030 & -0.246 & -0.046 & 0.012 & 0 & 0 \\
0 & 0 & 0 & -0.051 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.033 & -2.386 & -0.953 & -0.226 & -1.459 & -0.233 & -0.689 & 0.033 & 0 & 0 \\
 & & & & & & -4 & 0 & 0 & 0 & 4 \\
 & & & & & & 0 & -4 & 0 & 0 & 0 \\
 & & & 0 & & & 0 & 0 & -0.5 & 0.5 & 0 \\
 & & & & & & 0 & 0 & 0 & -3 & 0 \\
0 & -0.058 & -4.202 & -0.365 & -0.397 & -0.136 & -1.105 & -0.207 & 0.058 & 0 & -0.4
\end{pmatrix}$$

$$B^* = (0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0)^T, \qquad C^* = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0)^T,$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_{11})^T, \qquad u = \Delta\delta_{as}, \qquad v = w_z.$$

Here $x_1 = \Delta z$ is a deviation in z from the nominal motion; $x_3 = \Delta\psi$ and $x_5 = \Delta\gamma$ are deviations of yaw and bank angles. The coordinates x_7, x_8 are deviations of the rudder and ailerons ($\Delta\delta_r$ and $\Delta\delta_a$); $x_9 = \Delta\delta_{rs}$. By means of variables x_{10}, x_{11} we describe variation of ΔW_z ($x_{10} = \Delta W_z$). The control factor of the first player is the scheduled deviation $\Delta\delta_{as}$ of the ailerons. The parameter w_z is used to obtain wind disturbance and belongs to the second player.

Restrictions are

$$|\Delta\delta_{as}| \leqq 10 \text{ deg} \frac{\pi}{180}, \qquad |w_z| \leqq 10 \text{ m sec}^{-1}.$$

Introduce a function φ^* depending on coordinates $x_1 = \Delta z$ and $x_2 = \Delta\dot{z}$. Let M^* be a convex hexagon on the plane x_1, x_2 with apexes $(-6, 0)$, $(-6, 1.5)$, $(0, 1.5)$, $(6, 0)$, $(6, -1.5)$, $(0, -1.5)$. Suppose

$$\varphi^*(x_1, x_2) = \min \{c \geqq 0 : (x_1, x_2)^T \in cM^*\}.$$

Consider an antagonistic differential game with dynamics (4.2), fixed terminal moment T and payment φ^*. The first player endeavours to minimize values of the function φ^* at the moment T. The aim of the second player is opposite. The set M^* can be considered as a tolerance for deviations $x_1 = \Delta z$, $x_2 = \Delta\dot{z}$ at the moment T. The function φ^* indicates a deviation from the tolerance. The optimal strategy of the first player in game (4.2) will be used to define δ_{as} in system (2.1).

In systems (4.1) and (4.2) the dimension of linear variables is the meter. Angles are measured in radians and time is in seconds.

For the calculation of coefficients in systems (4.1) and (4.2) we used the following data: the glide path inclination $\Theta = -2.66$ deg, the nominal relative velocity $\tilde{V}_0 = 72.2$ m sec^{-1}, the nominal wind components $W_{x0} = -5$ m sec^{-1}, $W_{y0} = W_{z0} = 0$.

5. Optimal first player strategy in the linear differential game

The main features characteristic to differential games (4.1), (4.2) are the following. Each of the games has a fixed stopping instant and a convex payment function which depends on two coordinates of the phase vector. Besides, the control factor of the first player is scalar. These features simplify the specifying of the optimal first player strategy. The strategy is realized by means of the switch surface in the space t, y_1, y_2 of equivalent [7, 8] second-order game. The relation between vectors $\mathbf{y} = (y_1, y_2)^T$ and $\mathbf{x}$ is described by the formula $\mathbf{y}(t) = X_*(T, t)\mathbf{x}(t)$ ($\mathbf{y}(t) = X^*(T, t)\mathbf{x}(t)$) where $X_*(T, t)$ ($X^*(T, t)$) is a matrix composed of the third and fourth (the first and second) rows of the fundamental Cauchy matrix for the homogeneous part of system (4.1) ((4.2)).

On one side from switch surface the optimal control takes an extremal value of one sign, on the other side the optimal value is opposite. The mathematical proof of the optimality of such a control law (by means of switch surface) and the analysis of its stability are given in [18, 19]. Corresponding computational procedures are stated in [9].

The switch surface is realized on computer as a set of its sections on the given collection of the time moments. These sections are called the switch lines. The switch lines $\Pi_*(\tau)$ for problem (4.1) are shown in Fig. 2. The lines have been built for the reverse time moments $\tau = 7, 11, 15$. Let $\mathbf{x}(t_i)$ be a state of system (4.1) at a moment t_i. If the point $\mathbf{y}(t_i) = X_*(T, t_i)\mathbf{x}(t_i)$ lies in the direction of vector $D_*(t_i) = X_*(T, t_i)B_*$ with respect to the switch line $\Pi_*(\tau_i)$, corresponding to the moment $\tau_i = T - t_i$, then $\Delta\delta_{es} = -10$ on the next step of the discrete scheme of control. We put $\Delta\delta_{es} = +10$ in the opposite situation of the point $\mathbf{y}(t_i)$ with respect to the line $\Pi_*(\tau_i)$. Similarly, the choice of optimal control factor $\Delta\delta_{as}$ in system (4.2) is made with the help of switch lines $\Pi^*(\tau)$ and vector $D^*(t) = X^*(T, t)B^*$.

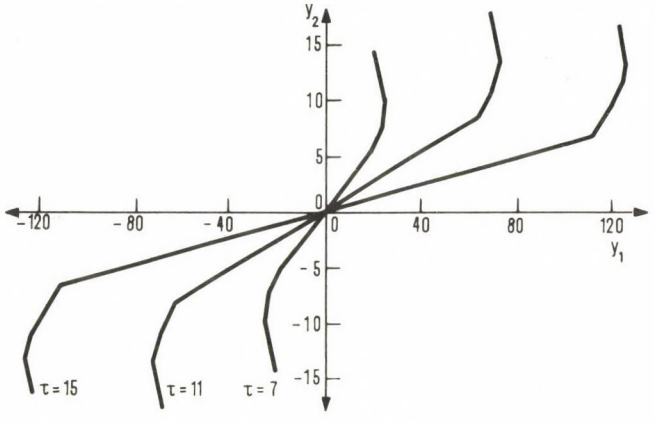


Fig. 2. Switch lines

6. Microburst model

Simulating system (2.1) motions, we suppose that the wind disturbance is caused by the aircraft flight through the microburst zone. The microburst model we used has been taken from [14]. Below we give an outline of this model.

The microburst is idealized as a three-dimensional axially symmetric vortex field. In this field we distinguish the thoroidal region ("core") where the wind velocity, being zero in the center, increases linearly along the radius to the frontier of the core. Outside the core the vortex field is determined by the stream function. Differentiation

of this function gives radial and vertical components of the wind velocity. The radial one is resolved into two components, the first of which is parallel and the second is orthogonal to the RW axis. The microburst is given by three parameters: $\mathscr{V}$ is the modulus of the wind velocity vector in the central microburst part, $\mathscr{H}$ is the altitude of the central part, and $\mathscr{R}$ is the radius of the vortex. The core radius is equal to $0.8\,\mathscr{H}$. The disposition of the microburst with respect to the glide path is determined by two coordinates of its center in the horizontal plane.

The microburst we use for simulation has the following parameters: $\mathscr{V} = 6\,\mathrm{m\,sec}^{-1}$, $\mathscr{H} = 700$ m, $\mathscr{R} = 1200$ m. A computed picture of the wind velocity field in the vertical plane, passing through the microburst center, is shown in Fig. 3.

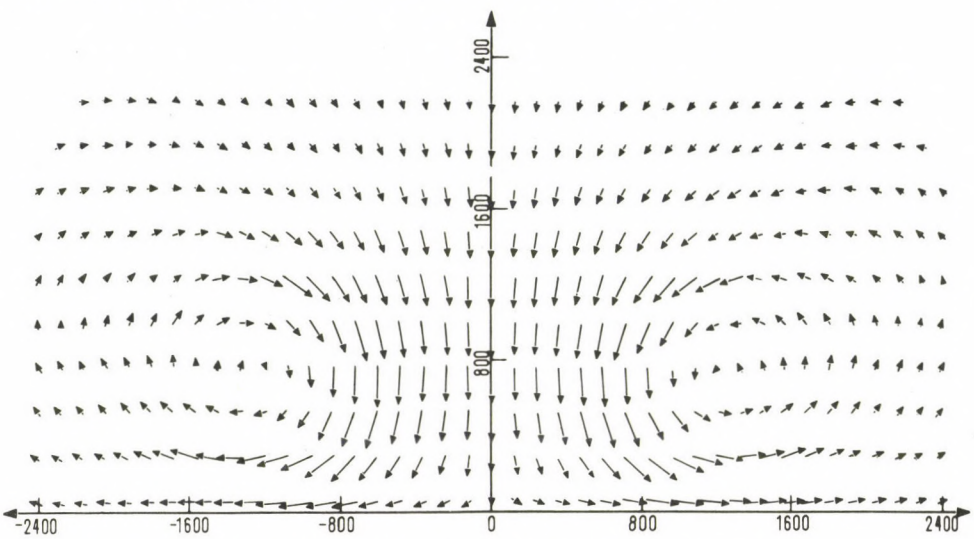


Fig. 3. Wind velocity field in vertical section of microburst

7. Simulation results

Let the initial system (2.1) position in x at the moment t_* be 8000 m from the RW threshold and values of all state coordinates correspond to the nominal motion along the glide path.

Consider two methods of control for system (2.1). The method I_1 uses accepted nowadays autopilot algorithms for constructing δ_{ps}, δ_{es}, δ_{rs} and δ_{as}. The second method I_2 is the minimax law. In this method factors δ_{es}, δ_{as} are constructed by means of switch lines obtained from auxiliary differential games (4.1), (4.2). Factors δ_{ps}, δ_{rs} are constructed with the help of accepted algorithms. Let $T = 15$ sec for problems (4.1), (4.2).

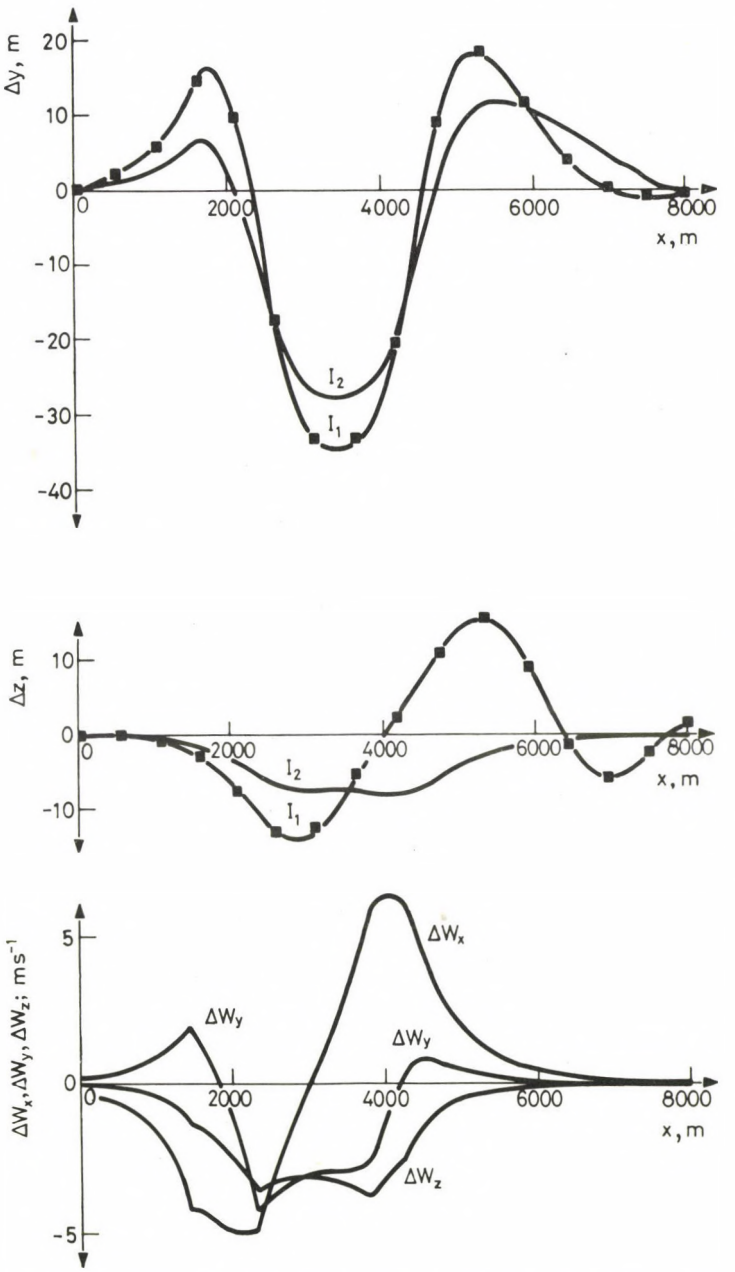


Fig. 4. Landing simulation results. Microburst center coordinates: $DX = 3000$ m, $DZ = 500$ m

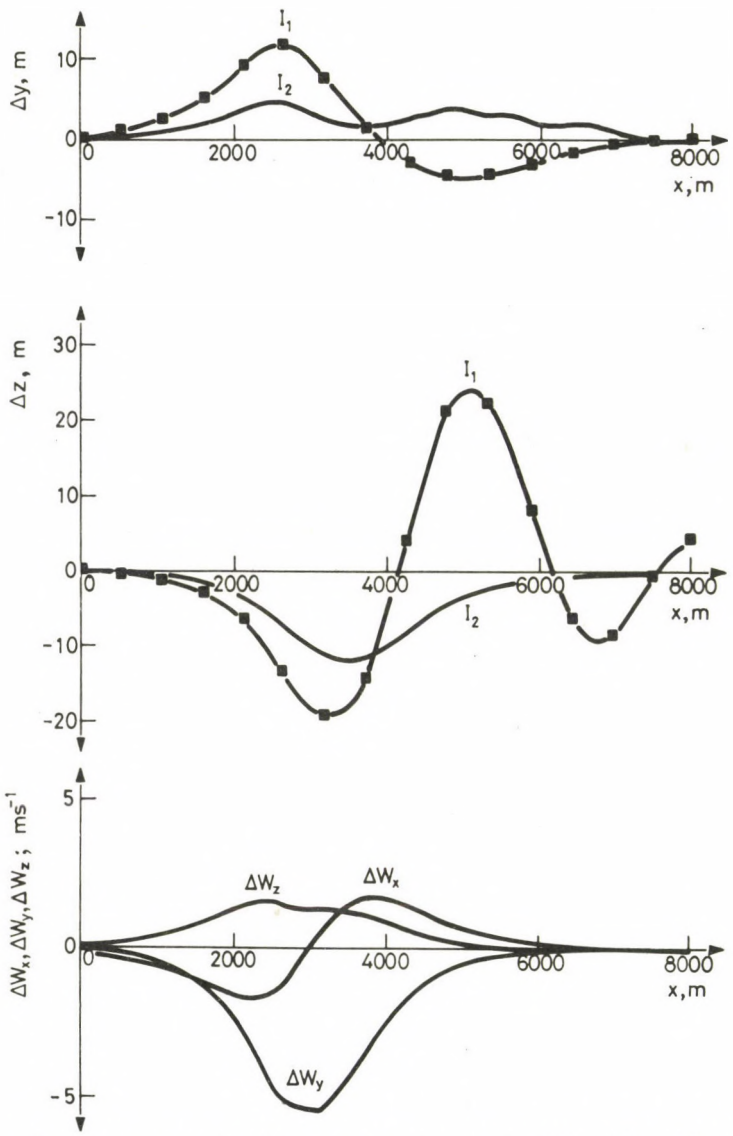


Fig. 5. Landing simulation results. Microburst center coordinates: $DX = 3000$ m, $DZ = 1500$ m

Denote by E a collection of reverse time moments τ_i on the interval $[0, T] = [0, 15]$. We suppose the switch lines have been built for every $\tau_i \in E$. In the method I_2 the switch lines are used in the following way. Let $d(t)$ be the distance in x up to the RW threshold at a moment $t \geqq t_*$ and V_{x0} be the nominal motion velocity in x. Then $s(t) = d(t)/V_{x0}$ is the prognose time remained till passing the RW threshold. For

obtaining $\delta_{es}(\delta_{as})$, if $s(t) \geqq T = 15$ sec, we use the same switch line corresponding to $\tau = T$. If $s(t) < T$, we use the line corresponding to the moment $\tau_i \in E$ nearest to $s(t)$. So, our control is comparatively rough while $d(t) \geqq V_{x0} T \approx 1000$ m. If $d(t) < V_{x0} T$, the control is more qualitative.

In system (2.1) wind velocity components W_x, W_y, W_z are calculated by formulas $W_x = \Delta W_x + W_{x0}$, $W_y = \Delta W_y + W_{y0}$, $W_z = \Delta W_z + W_{z0}$ where ΔW_x, ΔW_y, ΔW_z are taken from the microburst model; $W_{x0} = -5$ m sec^{-1}, $W_{y0} = W_{z0} = 0$. Consider two variants of the microburst center disposition in the horizontal plane: 1) the displacement DX in x from the initial aircraft position is 3000 m (or 5000 m from the RW threshold), the displacement DZ in z orthogonally to the RW axis is 500 m, 2) $DX = 3000$ m, $DZ = 1500$ m.

Simulation results for the control methods I_1, I_2 are shown in Figs 4 and 5. We give graphs of vertical Δy and lateral Δz deviations from the nominal motion and also realizations of deviations ΔW_x, ΔW_y, ΔW_z. The last curves correspond to the control method I_2. Realizations ΔW_x, ΔW_y, ΔW_z for the method I_1 are practically the same. For all the graphs the horizontal axis is the distance passed in x. Figure 4 (5) corresponds to the first (second) variant of microburst center disposition. The time discrete for computing of control factors and wind disturbances was equal to 0.1 sec.

It can be seen that the results for the minimax method I_2 are better than for the traditional method I_1.

8. Concluding remark

In conclusion we emphasize once more that the computation of the minimax control method demands neither accurate information about the disposition of the extremal wind disturbance zone nor any information about the wind velocity field in that zone. It is enough to describe an approximate amplitude of the wind velocity variation. This is the principal difference of the approach, based on the differential game theory, from the methods, given in [4, 6], where such information is essential.

References

1. *Kein, V. M., Parikov, A. N., Smurov, M. Yu.*, On the one method of the optimal control by means of extremal aiming procedure. Prikl. mat. i meh., **44** (1980), No. *3*, pp. 434–440 (in Russian).
2. *Titovskii, I. N.*, Game theoretical approach to the synthesis problem of aircraft control in the landing. Uchenye zapiski Centr. Aerohidrodinam. Inst., **XII** (1981), No. *1*, pp. 85–92 (in Russian).
3. *Kein, V. M.*, Optimization of control systems with minmax criterion. Nauka, Moscow, 1985 (in Russian).
4. *Miele, A., Wang, T., Melvin, W.*, Optimal take-off trajectories in the presence of windshear. J. Opt. Theory and Appl., **49** (1986), No. *1*, pp. 1–45.
5. *Psiaki, M. L., Stengel, R. F.*, Optimal flight paths through microburst wind profiles. J. of Aircraft, **23** (1986), No. *8*, pp. 625–635.

6. *Miele, A., Wang, T., Tzeng, C. Y., Melvin, W. W.*, Optimal abort landing trajectories in the presence of windshear. J. Opt. Theory and Appl., **55** (1987), No. *2*, pp. 165–202.
7. *Krasovskii, N. N., Subbotin, A. I.*, Closed-loop differential games. Nauka, Moscow, 1974 (in Russian).
8. *Subbotin, A. I., Chentsov, A. G.*, Optimization of guaranteed result in control problems. Nauka, Moscow, 1981 (in Russian).
9. *Botkin, N. D., Kein, V. M., Patsko, V. S.*, Model problem of control of lateral motion of aircraft in the landing. Prikl. mat. i meh., **48** (1984), No. *4*, pp. 560–567 (in Russian).
10. *Korneev, V. A., Melikyan, A. A., Titovskii, I. N.*, Stabilization of aircraft path in the presence of wind disturbance in minmax setting. Izv. Akad. Nauk SSSR. Tehn. kibernetika, 1985, **3**, pp. 132–139 (in Russian).
11. *Korneev, V. A.*, Quasi-optimal correction of the aircraft motion in the landing with wind disturbances. Izv. Akad. Nauk SSSR. Tehn. kibernetika, 1987, **3**, pp. 180–184 (in Russian).
12. *Zarkh, M. A., Patsko, V. S.*, Strategy of the second player in a linear differential game. Prikl. mat. i meh., **41** (1987), No. *2*, pp. 193–200 (in Russian).
13. *Botkin, N. D., Kein, V. M., Patsko, V. S.*, Application of differential game theory methods to the problem of aircraft control in the landing. In: Positional control with guaranteed result. Ural Sci. Centr. Akad. Nauk SSSR. Sverdlovsk, 1988, pp. 33–34 (in Russian).
14. *Ivan, M.*, A ring-vortex down-burst model for real-time flight simulation of severe wind shears. AIAA Flight Simulation Technologies Conf., 1985, July 22–24, St. Louis, Miss., 1985, pp. 57–61.
15. *Dietenberger, M. A., Hains, P. A., Luerst, G. K.*, Reconstruction of Pan Am New Orleans accident, J. of Aircraft, **22** (1985), No. *8*, pp. 719–728.
16. Systems of digital aircraft control. Edited by A. D. Aleksandrov, S. M. Fedorov, Machinostroenie, Moscow, 1983 (in Russian).
17. *Ostoslavskii, I. V., Strazheva, I. V.*, Dynamics of flight. Trajectories of aircrafts. Machinostroenie, Moscow, 1969 (in Russian).
18. *Botkin, N. D., Patsko, V. S.*, Universal strategy in a differential game with fixed stopping time. Probl. Control and Inform. Theory, **11** (1982), No. *6*, pp. 419–432.
19. *Botkin, N. D., Patsko, V. S.*, Positional control in a linear differential game. Izv. Akad. Nauk SSSR. Tehn. kibernetika, 1983, *4*, pp. 78–85 (in Russian).
20. Algorithms and programs of solution of linear differential games: (Materials on math. software). Ural Sci. Center. Akad. Nauk SSSR. Sverdlovsk, 1984, p. 295 (in Russian).

Управление самолетом на посадке при сдвиге ветра

Н. Д. БОТКИН, В. М. КЕЙН, В. С. ПАЦКО, В. Л. ТУРОВА

(Свердловск, Ленинград)

Рассматривается задача управления средним транспортным самолетом на посадке в условиях резкого изменения скорости ветра (сдвиг ветра). Процесс посадки исследуется до момента пролета торца взлетно-посадочной полосы. Предполагается, что относительно ветра известны ориентировочно лишь пределы возможных отклонений его скорости от некоторого номинального постоянного значения и само это значение. Какая-либо информация о пространственном расположении зоны сдвига ветра, как и сведения о поле скорости ветра в ней, считаются отсутствующими.

Для решения задачи об управлении привлекаются методы теории позиционных дифференциальных игр [7, 8]. Исходная нелинейная система дифференциальных уравнений линеаризуется относительно номинального движения по прямолинейной глиссаде снижения. Полученнная в результате линейная система распадается на подсистему движения в вертикальной плоскости (продольное движение) и подсистему бокового движения. Для каждой из подсистем ставится вспомогательная линейная дифференциальная игра с фиксированным моментом окончания и выпуклой терминальной функцией платы, зависящей от двух координат фазового вектора. За такие координаты в подсистеме продольного движения берутся отклонения по высоте и его скорость, в подсистеме бокового движения — боковое отклонение и его скорость. Первый игрок выбором

управления минимизирует значения функции платы, интересы второго, распоряжающегося ветровой помехой, противоположны.

Указанные линейные дифференциальные игры поддаются решению на ЭВМ при помощи разработанных в настоящее время численных методов [9, 13, 18–20]. Элементом решения, в частности, является оптимальная (минимаксная) стратегия первого игрока. Оптимальное управление при этом определяется набором линий переключения, имеющих несложную структуру.

Минимаксный способ управления, полученный в рамках вспомогательных задач, используется затем в полной нелинейной системе. Приводимые в статье результаты моделирования процесса посадки относятся к случаю, когда ветровое возмущение обусловлено прохождением самолета через зону микровзрыва. Микровзрыв представляет собой нисходящий поток воздуха, ударяющийся о поверхность земли и растекающийся затем с образованием вихря. Математическая модель микровзрыва заимствована из работы [14].

При моделировании минимаксный способ управления сравнивается с традиционным, принятым в настоящее время. В целом результаты моделирования для минимаксного способа существенно лучше.

Н. Д. Боткин, В. С. Пацко, В. Л. Турова
Институт математики и механики УрО АН СССР, СССР,
620066, Свердловск, ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 18 (4), pp. 237—250 (1989)

SUPERIMPOSED DISTANCE CODES

A. G. DYACHKOV, V. V. RYKOV, A. M. RASHAD

(*Moscow*) (*Kuwait*)

(Received September 9, 1988)

We introduce a new class of superimposed codes [1–4] called superimposed distance codes (SDC). In communication systems with random multiple access (RMA) these codes can be used simultaneously in information transmission for conflict resolution, the multiplicity of which does not exceed a given level [5]. Upper and lower bounds on the rate of SDC are obtained.

1. Definitions, notations and formulations of the results

Let $1 \leqq s < t$, $1 < D < N$ be integers, $[N]$ be the set of integers from 1 to N, $\Omega_j \subset [N]$, $j = \overline{1, t}$, be subsets of $[N]$ such that the number of elements $|\Omega_j| > D$.

Definition 1. A family of subsets $\Omega_1, \Omega_2, \ldots, \Omega_t$ is called an (s, N, D)-*family*, if for any collection of integers $m_1, m_2, \ldots, m_{s+1}$, $m_i \neq m_j$, $m_i \in [t]$ it is true that

$$\left| \Omega_{m_{s+1}} \setminus \bigcup_{k=1}^{s} \Omega_{m_k} \right| \geqq D.$$

The subset Ω_j is identified with the binary (of 0 and 1) column $\mathbf{x}(j) = (x_1(j), x_2(j), \ldots, x_N(j))$, where

$$x_i(j) = \begin{cases} 1, & \text{if} \quad i \in \Omega_j, \\ 0, & \text{if} \quad u \notin \Omega_j. \end{cases}$$

So one can consider the family $\Omega_1, \Omega_2, \ldots, \Omega_t$, where $\Omega_j \subset [N]$, as an $(N \times t)$-matrix (code) $\mathbf{X} = \|x_i(j)\|$, $i = \overline{1, N}$, $j = \overline{1, t}$, which consists of columns (codewords) $\mathbf{x}(1)$, $\mathbf{x}(2), \ldots, \mathbf{x}(t)$.

Definition 2. A matrix $\mathbf{X}$, corresponding to an (s, N, D)-family, is called a code of *strength* s, *length* N, *volume* t, and *superimposed distance* D.

For the particular case $D = 1$ *superimposed distance codes* (SDC) were introduced by Kautz–Singleton [1] and were studied in [2–4]. The similar definition of SDC and its applications to RMA were considered in [5].

Let $t(s, N, D)$ be the maximal possible volume of SDC. Introduce a parameter d, $0 < d < 1$. Denote by the symbol $\lfloor a \rfloor$ the largest integer $< a$. Define for fixed s and

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

d the *rate* of SDC

$$R_s(d)=\overline{\lim_{N\to\infty}}\frac{\log t(s, N, \lfloor Nd\rfloor)}{N}.$$

Here and below we use the logarithm of base 2. The purpose of this paper to obtain upper and lower bounds for $R_s(d)$.

Upper bound. Let

$$h(\alpha)=-\alpha\log\alpha-(1-\alpha)\log(1-\alpha) \tag{1.1}$$

be binary entropy and

$$d_s=\frac{s^s}{(s+1)^{s+1}},\qquad s=1, 2, \ldots. \tag{1.2}$$

Theorem 1. The rate of SDC *is*

$$R_s(d)\leqq U_s(d),$$

where, for $d\geqq d_s$, *the function* $U_s(d)=0$ *and for* $0<d<d_s$ *the sequence* $U_s=U_s(d)>0$ *is defined recurrently*:

a) if $s=1$, *then*

$$U_1=U_1(d)=\begin{cases} h\left(\frac{1}{2}[1-\sqrt{8d(1-2d)}]\right), & \text{if}\quad 0<d<d_1=1/4,\\ 0, & \text{if}\quad d\geqq 1/4;\end{cases}$$

b) *if* $s\geqq 2$, *then*

$$U_s=U_s(d)=\min\{1-d/d_s, U_1(d)/s, \hat{U}_s(d)\},$$

where $\hat{U}_s=\hat{U}_s(d)$ *is the unique solution of the equation*

$$\hat{U}_s=\max_{(1.4)}\left\{h\left(\frac{v}{s}\right)-(v+d)h\left(\frac{v}{(v+d)s}\right)\right\}. \tag{1.3}$$

The maximum in (1.3) *is taken over all* v *for which*

$$0\leqq v\leqq 1-\frac{\hat{U}_s}{U_{s-1}}-d. \tag{1.4}$$

Theorem 1 is a generalization of the upper bound, obtained in [2] for the rate of SDC, when $D=1$. At the first step of our recurrent bound (at $s=1$) we use the bound obtained in [6] for the rate of binary code, which corrects $D=\lfloor Nd\rfloor$ errors. Furthermore, in the proof of Theorem 1, which is given in Section 2, we use the idea of the known Plotkin bound and the arguments similar to [5].

More simple asymptotic formulas for $U_s(d)$ are given by

Corollary 1.1. 1) *For any fixed* $s \geqq 1$ *the rate*

$$R_s(d) \leqq U_s(d) = 1 - d/d_s, \tag{1.5}$$

if $d \to d_s - 0$. 2) *If* $s \to \infty$ *and* $ds \to 0$, *then*

$$R_s(d) \leqq U_s(d) = \hat{U}_s(d) = \frac{(2-ds)\log s}{s^2}(1+o(1)). \tag{1.6}$$

Inequality (1.5) is evident. In the particular case $d=0$ ($D=1$) inequality (1.6) is obtained in [2]. The proof of considered case $ds \to 0$ is similar.

The lower bound. Introduce the Kullback distance

$$K(\alpha, \beta) = \alpha \log \frac{\alpha}{\beta} + (1-\alpha)\log \frac{1-\alpha}{1-\beta} \tag{1.7}$$

which is a non-negative function of parameters α, $0<\alpha<1$, and β, $0<\beta<1$. Fix Q, $0<Q<1$, and d, $0<d<1-Q$. Using (1.7) define

$$\mathscr{A}_s(d, Q) = (1-Q)K\left(\frac{d}{1-Q}, y^s\right) - sK(Q, y), \tag{1.8}$$

where in the right-hand side the parameter $y = y_s(d, Q)$, $0<y<1$, is a root of the equation

$$y = \frac{1}{1-d}[(Q-d) + (1-2Q+Qy)y^s]. \tag{1.9}$$

The lower bound on the rate of SDC is given by

Theorem 2. Let $0<d<d_s$ *and* d_s *be defined by* (1.2). *The rate of* SDC *is*

$$R_s(d) \geqq L_s(d) = s^{-1} \max_{(1.11)} \mathscr{A}_s(d, Q), \tag{1.10}$$

where the maximum in (1.10) *is taken over all* Q *satisfying the condition*

$$(1-Q)Q^s \geqq d. \tag{1.11}$$

This theorem will be proved in Section 2 with the help of the random coding method for the code ensemble X with independent constant weight words. The analytical properties of $\mathscr{A}_s(d, Q)$ are described by

Theorem 3. Let $0<d \leqq (1-Q)Q^s$. *Then the following statements are true.*

1) *In the interval* $0<y<1$ *equation* (1.9) *has a unique solution*

$$y = y_s(d, Q) \geqq Q,$$

where the sign of equality is at $d = (1-Q)Q^s$.

2) *The value* $\mathscr{A}_s(d, Q)$, *defined by* (1.8)–(1.9), *is presented in the form*

$$\mathscr{A}_s(d, Q) = \max_{0<y<1} \left\{ (1-Q)K\left(\frac{d}{1-Q}, y^s\right) - sK(Q, y) \right\}. \tag{1.12}$$

3) *At* $d=(1-Q)Q^s$ *the value* $\mathscr{A}_s(d, Q)=0$. *At* $0<d<(1-Q)Q^s$ *the function* $\mathscr{A}_s(d, Q)>0$, *monotonically decreases with increasing* d *and*

$$\mathscr{A}_s(d, Q) > K(d, Q^s(1-Q)). \tag{1.13}$$

We omit the detailed proof of Theorem 3, only notice that statements 1) and 2) are established by the standard methods of mathematical analysis. Statement 3) follows from 1) and 2) evidently. To prove (1.13) we use (1.12) and the easily obtained inequality

$$\gamma K(\alpha, \beta) > K(\gamma\alpha, \gamma\beta), \qquad 0<\gamma<1.$$

Notice, that

$$\max_{0<Q<1} Q^s(1-Q) = \frac{s^s}{(s+1)^{s+1}} = d_s$$

is reached at $Q=s/(s+1)$. Hence, Theorem 3 means that the lower bound on the rate of SDC has the following properties.

Corollary 2. At $d=d_s$ *the value of* $L_s(d_s)=0$. *At* $0<d<d_s$ *the function* $L_s(d)>0$, *monotonically decreases with increasing* d *and*

$$L_s(d) > s^{-1}K(d, d_s). \tag{1.14}$$

Inequalities (1.5), (1.10) and Corollary 2 yield the important

Corollary 3. At $d \geqq d_s$ *the rate* $R_s(d)=0$, *and at* $0<d<d_s$ *the rate* $R_s(d)>0$.

From inequality (1.14) and Theorem 2 we have

$$R_s(d) > s^{-1}K(d, d_s)$$

if $0<d<d_s$. Notice that this more simple estimation can be easily established with the help of the random coding method for an ensemble $X=\|x_i(j)\|$ with independent identically distributed components $x_i(j)$.

The asymptotics of $L_s(d)$ *at* $s\to\infty$. For fixed $r>0$ define the function

$$E(r) = \lim_{s\to\infty} s^2 \cdot L_s(r/s). \tag{1.15}$$

Theorem 4. At $r \geqq 1/e$ *the value of* $E(r)=0$. *At* $0<r<1/e$ *the function* $E(r)>0$ *is described by*

$$E(r) = \max_{(1.17)} K(r/q, e^{-q}), \tag{1.16}$$

where the maximum in (1.16) *is taken over all* $q>0$, *for which*

$$qe^{-q} \geqq r. \tag{1.17}$$

Furthermore, at $0<r<1/e$ *the value of*

$$E(r) \geqq \frac{\log e}{e} + r \log r. \tag{1.18}$$

Proof. Fix arbitrary $0<r<q$. Let $s \to \infty$ and

$$d = r/s, \qquad Q = 1 - q/s.$$

It is not difficult to verify that the asymptotic expansion of the solution of equation (1.9) has the form

$$y_s(r/s, 1-q/s) = 1 - q/s + O(s^{-2}).$$

So (1.8) means that at $r<q$

$$\mathscr{A}_s(r/s, 1-q/s) = s^{-1} K(r/q, e^{-q}) \cdot (1+o(1)),$$

if $s \to \infty$. Hence, accounting definitions (1.10)–(1.11), we obtain (1.16)–(1.17). Inequality (1.18) is easily established by the calculation of the asymptotics in the right-hand side of (1.14). Theorem 4 is proved.

Some numerical results. Let

$$U_s(0) = \lim_{d \to 0} U_s(d).$$

From numerical values, obtained in [2], we have

$$U_2(0) = 0.3220, \qquad U_6(0) = 0.0830.$$

Similarly defined values of the lower bound are

$$L_2(0) = 0.1824, \qquad L_6(0) = 0.0194.$$

For the function $E(r)$, which defines the main asymptotic member of $L_s(d)$, $s \to \infty$, we have

$$E(0) = 1/\log e = 0.6931, \qquad E(1/2e) = 0.1210.$$

The connection with results of L. A. Bassalygo and M. S. Pinsker. By the analogy with [5] it is easy to understand that SDC can be used in the *synchronous* superimposed RMA-systems simultaneously as for information transmission so for conflict resolution, the multiplicity of which is $\leqq s$. In comparison with the corresponding superimposed codes with "distance" [5], there is the decoding algorithm of SDC, which is essentially more simple. Let t for an RMA-system be the total number of sources and $s \ll t$ be the number of active sources. Then it is evident that the decoding

algorithm of SDC needs only t steps, at that time the full sorting decoding algorithm of the corresponding code from [5] needs t^s steps. Notice also that with the help of Theorem 4 it is possible to obtain the main result of [5] about the maximal summary rate of the binary information transmission for the synchronous RMA-system.

2. The proof of Theorem 1

The upper estimation of Theorem 1 is obtained as an evident combination of the following three auxiliary bounds, which are formulated below as Lemmas 1–3.

Lemma 1. (*The bound of the code distance*). *If* $0<d<1/4$, *then the rate of* SDC

$$R_s(d) \leqq s^{-1} h\left(\frac{1}{2}\left[1-\sqrt{8d(1-2d)}\right]\right).$$

Lemma 2. (*Plotkin bound*). *Let* d_s *be defined by* (1.2). *For any* d, $0<d\leqq d_s$, *the estimation*

$$R_s(d) \leqq 1-d/d_s$$

is true.

Lemma 3. (*Recurrent bound*). *Let* $U_{s-1}=U_{s-1}(d)$ *be some upper bound for the rate* $R_{s-1}(d)$. *Then the value* $\hat{U}_s=\hat{U}_s(d)$, *defined by* (1.3)–(1.4) *is the upper bound for the rate* $R_s(d)$.

Proof of Lemma 1. Consider an $(N \times C_t^s)$-matrix, the columns of which are the Boolean sums of all s-subsets of SDC columns. It is easy to understand that such a (0, 1)-matrix is the binary code correcting $D=\lfloor Nd \rfloor$ errors. Hence, Lemma 1 is the consequence of the known upper bound [6] for the rate of such codes. Lemma 1 is proved.

Proof of Lemma 2. Let $X=\|x_i(j)\|$, $i=\overline{1, N}$, $j=\overline{1, t}$, be SDC corresponding to an (s, N, D)-family $\Omega_1, \Omega_2, \ldots, \Omega_t$. Denote by $\omega_k=(\Omega_{k_1}, \Omega_{k_2}, \ldots, \Omega_{k_{s+1}})$, where $1\leqq k_1<k_2< \ldots <k_{s+1}\leqq t$, an arbitrary $(s+1)$-collection of the family's members. The cardinality number of such collections is equal to C_t^{s+1}. Introduce

$$e(k_m)=\left|\Omega_{k_m} \setminus \bigcup_{\substack{j=1 \\ j\neq m}}^{s+1} \Omega_{k_j}\right|, \qquad b_k=\sum_{m=1}^{s+1} e(k_m).$$

Let

$$r_i=\sum_{j=1}^{t} x_i(j), \qquad i=\overline{1, N},$$

be the number of units in the i-th row of SDC X. Notice that

$$\sum_{k=1}^{C_t^{s+1}} b_k=\sum_{i=1}^{N} r_i C_{t-r_i}^s, \tag{2.1}$$

where, by definition, $C_{t-r}^s=0$, if $s>t-r$.

It is evident that for an (s, N, D)-family $e(k_m) \geqq D$. So, equality (2.1) means that for any (s, N, D)-family

$$C_t^{s+1} \cdot D \cdot (s+1) \leqq \sum_{i=1}^{N} r_i C_{t-r_i}^s. \tag{2.2}$$

At $0 \leqq r \leqq t-s$ the ratio

$$\frac{rC_{t-r}^s}{C_t^{s+1}} \leqq \frac{r(s+1)}{t}\left[\frac{t}{t-s} - \frac{r}{t-s}\right]^s.$$

Let $p_i = r_i/(t-s)$, $u = t/(t-s)$. Then

$$\frac{\sum r_i C_{t-r_i}^s}{C_t^{s+1}} \leqq (s+1) \sum_{i=1}^{N} \frac{1}{u} p_i (u-p_i)^s \leqq \frac{s^s u^s N}{(s+1)^s}, \tag{2.3}$$

where we account that

$$\max_{0 \leqq p \leqq u} p(u-p)^s = \frac{u^{s+1} s^s}{(s+1)^{s+1}}$$

is reached at $p = u/(s+1)$.

From (2.2) and (2.3) follows that for any (s, N, D)-family of volume t the distance

$$D \leqq \left(\frac{t}{t-s}\right)^s \cdot \frac{s^s}{(s+1)^{s+1}} \cdot N. \tag{2.4}$$

Notice that $(1-s/t)^s \geqq 1 - s^2/t$. Hence, inequality (2.4) means that for any (s, N, D)-family

$$N \geqq \frac{D}{d_s}(1 - s^2/t), \tag{2.5}$$

where d_s is defined by (1.2). Note that to obtain (2.5) we used the idea of the similar proof from [5].

Fix an arbitrary integer $n < N$ and $0 < n < \log(t/s)$. In any $(N \times t)$-matrix X there are $t/2^n \geqq s$ columns, which coincide in the first n positions. Let X be an SDC corresponding to an (s, N, D)-family of volume t. Then in the remaining $N-n$ positions the given $t/2^n$ columns (codewords) must form an SDC corresponding to an $(s, N-n, D)$-family of volume $t/2^n$. Now inequality (2.5) shows that for any n, $0 < n < \log(t/s)$ the difference

$$N - n \geqq \frac{D}{d_s}(1 - s^2 2^n/t).$$

Set in the given inequality $n = \log(t/\log t)$. We have

$$\log t \leqq N - \frac{D}{d_s}(1 - s^2/\log t) + \log\log t. \tag{2.6}$$

The definition of the rate $R_s(d)$ and inequality (2.6) yield the statement of Lemma 2. Thus Lemma 2 is proved.

Proof of Lemma 3. Consider the function

$$f_s(d, v)=h\left(\frac{v}{s}\right)-(v+d)h\left(\frac{v}{(v+d)s}\right), \tag{2.7}$$

which was introduced in the definition of (1.3). Notice that (2.7) is the non-negative concave function of the parameter v, $0\leqq v\leqq 1-d$, and $f_s(d, 0)=f_s(d, 1-d)=0$. Define

$$f_s^{(d)}=f_s(d, v_s)=\max_{0\leqq v\leqq 1-d} f_s(d, v), \tag{2.8}$$

where $v_s=v_s(d)$ is the unique extremal value of the parameter v, $0<v<1-d$, in (2.8).

Let $U_{s-1}=U_{s-1}(d)$ be the upper bound of the rate $R_{s-1}(d)$ in the condition of Lemma 3. Introduce $F_s(d, a)$ non-negative continuous function of the parameter a, $0\leqq a\leqq(1-d)U_{s-1}$:

$$F_s(d, a)=\max_{(2.10)} f_s(d, v), \tag{2.9}$$

where the maximum in (2.9) is taken over all v satisfying the condition

$$0\leqq v\leqq 1-d-\frac{a}{U_{s-1}}. \tag{2.10}$$

It is not difficult to understand that for fixed s and d

$$F_s(d, a)=\begin{cases} f_s^{(d)}, & \text{if} \quad 0\leqq a\leqq(1-d-v_s)U_{s-1}, \\ \text{monotonically} \\ \text{decreases}, & \text{if} \quad (1-d-v_s)U_{s-1}<a<(1-d)U_{s-1}, \\ 0, & \text{if} \quad a=(1-d)U_{s-1}. \end{cases}$$

Hence, definitions (1.3)–(1.4) mean that the upper bound of Lemma 3 $\hat{U}_s=\hat{U}_s(d)$ is the unique solution of the equation $a=F_s(d, a)$, and it holds that

$$0<a\leqq F_s(d, a), \quad \text{iff} \quad 0<a\leqq\hat{U}_s(d). \tag{2.11}$$

Consider two arbitrary sequences of integers N_n and t_n, for which there exists an $(s, N_n, \lfloor N_n d\rfloor)$-family of volume t_n, $n=1, 2, \ldots$. For fixed parameters s, t and D denote by $\bar{N}(s, t, D)$ the minimal possible length of SDC. We have

$$\lim_{n\to\infty}\frac{\log t_n}{\bar{N}(s-1, t_n, \lfloor N_n d\rfloor}\leqq R_{s-1}(d)\leqq U_{s-1}(d), \tag{2.12}$$

where the first inequality follows from the definition of the SDC rate, and the second

inequality is the condition of Lemma 3. Define for the considered sequences the value

$$\bar{R}_s = \bar{R}_s(d) = \varlimsup_{n\to\infty} \frac{\log t_n}{N_n}. \tag{2.13}$$

In order to prove Lemma 3 it is sufficient to show that

$$\bar{R}_s(d) \leqq \hat{U}_s(d). \tag{2.14}$$

Let X be an arbitrary SDC with parameters s, N, t and D. Denote by $t(w)$ the cardinal number of codewords of X, having weight w, i.e. containing w units and $N-w$ zeros, $D \leqq w \leqq N-1$. When $t(w)>0$ the upper estimations of w and $t(w)$ were obtained in [2] for the particular case $D=1$. In the case $D>1$ the evident generalizations of the corresponding proofs from [2] lead to the following estimations

$$w \leqq N - \bar{N}(s-1,\ t-1,\ D), \tag{2.15}$$

$$t(w) \leqq K_w \frac{C_N^{k_w+1}}{C_{\lfloor w/k_w \rfloor}^{k_w}}, \tag{2.16}$$

where $k_w = \lfloor (w-D+1)/s \rfloor$, $K_w = \lfloor w/k_w \rfloor^2$.

Remark. With the help of the more accurate method, presented in [4], it is easy to prove that (2.16) is also true when $K_w = 1$.

Now accounting (2.12), (2.15) and the logarithmic asymptotics of the right-hand side of (2.16), we obtain

$$\bar{R}_s \leqq F_s(d,\ \bar{R}_s), \tag{2.17}$$

where $\bar{R}_s$ is defined by (2.13). From (2.11) and (2.17) follows (2.14). Lemma 3 is proved.

3. The proof of Theorem 2

Introduce the $(N,\ t,\ w)$-ensemble of $(N \times t)$-matrices (codes) X, columns (codewords) of which are selected independently from the collection of C_N^w columns, containing the same number $w \geqq D$ units and $N-w$ zeros. Denote by Ω_j, $j = \overline{1,\ t}$, the w-subset, which is identified by the column $x(j)$.

We shall say that the codeword $\mathbf{x}(j)$ is "bad" for code X, if it is not satisfied the definition of SDC. This means, that among the rest $t-1$ codewords there are codewords $\mathbf{x}(l_1)$, $\mathbf{x}(l_2)$, ..., $\mathbf{x}(l_s)$, such that for the corresponding W-subsets the condition

$$\left| \Omega_j \setminus \bigcup_{k=1}^{s} \Omega_{l_k} \right| \leqq D-1 \tag{3.1}$$

is fulfilled.

Denote by $\mathscr{P}_s = \mathscr{P}_s(N, D, W)$ the probability of event (3.1), for the (N, t, W)-ensemble. It is evident that for this ensemble the average number of "bad" words does not exceed $t \cdot C_{t-1}^{s} \cdot \mathscr{P}_s$. From this follows

Lemma 4. If for an (N, t, W)*-ensemble the probability of event* (3.1) *satisfies the condition*

$$C_{t-1}^{s} \cdot \mathscr{P}_s(N, D, W) < 1/2,$$

then there is an (s, N, D)*-family consisting of the* $t/2$ *W-subsets of the set* $[N]$.

The definition of the rate $R_s(d)$ and Lemma 4 yield

Lemma 5. For any fixed Q, $0 < Q < 1 - d$, *the rate of* SDC *satisfies the inequality*

$$R_s(d) \geqq \frac{1}{s} \overline{\lim_{N \to \infty}} \frac{-\log \mathscr{P}_s(N, \lfloor Nd \rfloor, \lfloor N(1-Q) \rfloor)}{N}.$$

Lemma 6. Let $0 < d \leqq Q^s(1-Q)$. *If* $N \to \infty$, *then*

$$\mathscr{P}_s(N, \lfloor Nd \rfloor, \lfloor N(1-Q \rfloor) = \exp\{-N[\mathscr{A}_s(d, Q) + o(1)]\}, \tag{3.2}$$

where the exponent $\mathscr{A}_s(d, Q)$ *is defined by* (1.8)–(1.9).

Theorem 2 is the evident consequence of Lemmas 5 and 6. To complete the inference of Theorem 2 we need to give

Proof of Lemma 6. To describe the probability $\mathscr{P}_s(N, D, W)$ we need some additional notations. Let

$$x_1^s = (x_1, x_2, \ldots, x_s) \in (0, 1)^s, \; z \in (0; 1).$$

We shall write $z = x_1 \vee x_2 \vee \ldots \vee x_s$, if

$$z = \begin{cases} 0, & \text{at} \quad x_1 = x_2 = \ldots = x_s = 0, \\ 1, & \text{otherwise.} \end{cases}$$

For pairs $(x_1^s, z) \in (0, 1)^{s+1}$ we define the function

$$P(z | x_1^s) = \begin{cases} 1, & \text{if} \quad z = x_1 \vee x_2 \vee \ldots \vee x_s, \\ 0, & \text{otherwise.} \end{cases}$$

By the symbol $\|n(x_1^s, z)\|$ we denote an arbitrary collection of non-negative integers, for which

$$\sum_{x_1^s} \sum_{z} h(x_1^s, z) = N$$

and for any $i = \overline{1, s}$ the number

$$n(x_i) = \sum_{x_1^{i-1}} \sum_{x_{i+1}^s} \sum_{z} n(x_1^s, z) = \begin{cases} W, & \text{if} \quad x_i = 1, \\ N - W, & \text{if} \quad x_i = 0. \end{cases}$$

For each collection of this kind we introduce

$$n_1 = n_1(\|n(x_1^s, z)\|) = \sum_{x_1^s} n(x_1^s, 1),$$

$$b(\|n(x_1^s\, z)\|) = \frac{N! \prod_{x_1^s} \prod_z P(z|x_1^s)^{n(x_1^s, z)}}{\prod_{x_1^s} \prod_z n(x_1^s, z)! \, (C_N^w)^S}. \tag{3.3}$$

At $i = \overline{0, D-1}$ define the probabilities

$$P_i = P_i(\|n(x_1^s, z)\|) = \frac{C_{N-n_1}^i C_{n_1}^{W-i}}{C_N^W}. \tag{3.4}$$

Comparing the ratio P_{i+1}/P_i with 1, one can easily verify that $P_0 < P_1 < \ldots < P_{D-1}$, if

$$D - 1 \leqq \frac{W(N+1) - n_1(W+1) - 1}{N+2}. \tag{3.5}$$

Now, we can write that

$$\mathscr{P}_s(N, D, W) = \sum_{\|n(x_1^s, z)\|} b(\|n(x_1^s, z)\|) \sum_{i=0}^{D-1} P_i(\|n(x_1^s, z)\|). \tag{3.6}$$

Consider also the probability

$$\hat{\mathscr{P}}_s = \hat{\mathscr{P}}_s(N, D, W) = \sum_{\|n(x_1^s, z)\|} b(\|n(x_1^s, z)\|) \cdot P_{D-1}(\|n(x_1^s, z)\|). \tag{3.7}$$

Notice that the probabilities $\mathscr{P}_s$ and $\hat{\mathscr{P}}_s$ have the same logarithmic asymptotics ($N \to \infty$), if condition (3.5) is fulfilled.

Let $\mathbf{Q} = (Q(0), Q(1))$ be the fixed probability distribution at (0, 1) and $\tau(\mathbf{Q})$ be the set of probability distribution

$$\tau = \{\tau(x_1^s, z), \quad x_1^s \in (0, 1)^s, \quad z \in (0, 1)\},$$

satisfying the following conditions:

1) if $P(z|x_1^s) = 0$, then $\tau(x_1^s, z) = 0$,

2) the marginal probabilities at x_i, $i = \overline{1, s}$, are fixed and coincide with $\mathbf{Q}$, i.e. $\tau(x_i) = Q(x_i)$.

For $\tau \in \tau(\mathbf{Q})$ define

$$\tau_1 = \sum_{x_1^s} \tau(x_1^s, 1)$$

and introduce non-negative functions

$$\mathscr{H}_s(\mathbf{Q}, \boldsymbol{\tau}) = \sum_{x_1^s} \sum_{z} \tau(x_1^s, z) \log \frac{\tau(x_1^s, z)}{P(z|x_1^s) \prod_{i=1}^{s} Q(x_i)}, \tag{3.8}$$

$$I_s^{(d)}(\mathbf{Q}, \boldsymbol{\tau}) = h(Q(0)) - \tau_1 h\left(\frac{Q(1)-d}{\tau_1}\right) - (1-\tau_1) h\left(\frac{d}{1-\tau_1}\right), \tag{3.9}$$

where $h(\cdot)$ is the binary entropy (1.1). It is not difficult to verify that (3.8)–(3.9) describe the logarithmic asymptotics ($N \to \infty$) of probabilities (3.3)–(3.4), when

$$n(x_1^s, z) = N\tau(x_1^s, z)(1 + o(1)).$$

So, at $N \to \infty$ for probability (3.7) we have

$$\hat{\mathscr{P}}_s(N, \lfloor Nd \rfloor, \lfloor N(1-Q) \rfloor) = \exp\{-N[\hat{\mathscr{A}}_s^{(d)}(\mathbf{Q}) + o(1)]\}, \tag{3.10}$$

$$\hat{\mathscr{A}}_s^{(d)}(\mathbf{Q}) = \min_{\boldsymbol{\tau} \in \tau(\mathbf{Q})} \{\mathscr{H}_s(\mathbf{Q}, \boldsymbol{\tau}) + I_s^{(d)}(\mathbf{Q}, \boldsymbol{\tau})\} =$$

$$= \mathscr{H}_s(\mathbf{Q}, \boldsymbol{\tau}^{(d)}) + I_s^{(d)}(\mathbf{Q}, \boldsymbol{\tau}^{(d)}), \tag{3.11}$$

where $\boldsymbol{\tau}^{(d)} \in \tau(\mathbf{Q})$ is an extremal distribution in (3.11).

To solve the extremal problem (3.11) we can use the standard Lagrange method. In our case functions (3.8)–(3.9) are convex at distribution $\boldsymbol{\tau} \in \tau(\mathbf{Q})$ and the Kuhn–Tucker theorem is applied. Accounting the symmetry of the task for variables x_i, $i = \overline{1, s}$, it is easy to verify that the unique extremal distribution

$$\boldsymbol{\tau}^{(d)} = \{\tau^{(d)}(x_1^s, z), \quad (x_1^s, z) \in (0, 1)^{s+1}\}$$

has the form

$$\tau^{(d)}(x_1^s, z) = \begin{cases} \dfrac{C \prod_{i=1}^{s} \tilde{Q}(x_i) \left[1 - \dfrac{d}{1-\tau_1^{(d)}}\right] P(z|x_1^s)}{1 - \dfrac{Q(1)-d}{\tau_1^{(d)}}}, & \text{if} \quad z = 1, \\ C \cdot \tilde{Q}(0)^s P(z|x_1^s), & \text{if} \quad z = 0, \end{cases} \tag{3.12}$$

where $\tau_1^{(d)}$ is the value of τ_1 for the extremal distribution $\boldsymbol{\tau}^{(d)}$, C is a constant, and the probability distribution $\tilde{\mathbf{Q}} = (\tilde{Q}(0), \tilde{Q}(1))$ will be chosen later, so that $\boldsymbol{\tau}^{(d)} \in \tau(\mathbf{Q})$. The probabilistic sense of components (3.12) means that

$$C = Q(0) + \frac{d}{\tilde{Q}(0)^s}, \qquad \tau_1^{(d)} = 1 - Q(0)\tilde{Q}(0)^s - d. \tag{3.13}$$

Hence,

$$\tau^{(d)}(x_1^s, z) = \begin{cases} \dfrac{\prod_{i=1}^{s} \tilde{Q}(x_i)\,[1 - Q(0)\tilde{Q}(0)^s - d]P(z|x_1^s)}{1 - \tilde{Q}(0)^s}, & \text{if} \quad z = 1, \\ [d + Q(0)\tilde{Q}(0)^s]P(z|x_1^s), & \text{if} \quad z = 0. \end{cases} \tag{3.14}$$

Say, for brevity, $Q(0) = Q$. Probabilities (3.14) satisfy the conditions $\tau(x_i) = Q(x_i)$, $i = \overline{1, s}$. So, the parameter $\tilde{Q}(0) = y$ is the root of the equation

$$1 - Q = \frac{(1-y)(1 - y^s \cdot Q - d)}{1 - y^s} \tag{3.15}$$

which is equivalent to equation (1.9). Now substitute $\tau = \tau^{(d)}$ into (3.8)–(3.9) and evaluate the right-hand side of (3.11), i.e. $\hat{\mathscr{I}}_s^{(d)}(\mathbf{Q})$. After not complicated, but wearing, transformations (omitted here) we obtain that $\hat{\mathscr{I}}_s^{(d)}(\mathbf{Q})$ coincide with $\mathscr{A}_s(d, Q)$, which is defined by (1.8)–(1.9).

To complete the inference of Lemma 6 we need to verify that condition (3.5) is asymptotically fulfilled, when

$$0 < d \leqq Q^s(1 - Q). \tag{3.16}$$

Hence, we must show, that

$$d \leqq (1 - \tau_1^{(d)})(1 - Q) \tag{3.17}$$

if (3.16) is true. From (3.13) we have

$$1 - \tau_1^{(d)} = d + Q y^s,$$

where y is the root of (1.9). So, (3.17) is equivalent to the condition

$$y^s \geqq \frac{d}{1 - Q}$$

which is the consequence of (3.16) and the inequality $y \geqq Q$ from Theorem 3. Lemma 6 is proved.

References

1. *Kauts, W. H., Singleton, R. C.*, Nonrandom binary superimposed codes. IEEE Trans. Inform. Theory, **10,** *4*, 1964, pp. 363–377.
2. *Dyachkov, A. G., Rykov, V. V.*, Bounds on the length of superimposed codes. Problemy Peredachi Informatsii, **18,** *3*, 1982, pp. 7–13 (in Russian).
3. *Dyachkov, A. G., Rykov, V. V.*, A survey of superimposed code theory. Problems of Control and Information Theory, **12,** *4*, 1983, pp. 229–242.

4. *Erdős, P., Frankl, P., Füredi, Z.*, Families of finite sets in which no set is covered by the union of r others. Israel Journal of Math., **51,** *1–2*, 1985, pp. 75–89.
5. *Bassalygo, L. A., Pinsker, M. S.*, Limited multiple-access of a non-synchronous channel. Problemy Peredachi Informatsii, **19,** *8*, 1983, pp. 92–96 (in Russian).
6. *McEliace, P. I., Rodemich, E. R., Rumsey, H., Welch, L. R.*, New upper bounds of the rate of a code via the Delsarte–MacWillams inequalities. IEEE Trans. Inform. Theory, **23,** *2*, 1977, pp. 157–166.

Коды с дизъюнктивным расстоянием

А. Г. ДЬЯЧКОВ, В. В. РЫКОВ, А. М. РАШАД

(Москва) (Кувейт)

Пусть $1 \leqq s < t$, $1 \leqq D < N$ — целые числа, $[N]$ — множество целых чисел от 1 до N, $\Omega_j \subset [N]$, $j = \overline{1, t}$ — подмножества $[N]$ такие что число элементов $|\Omega_j| > D$.

Семейство подмножеств $\Omega_1, \Omega_2, \ldots, \Omega_t$ называется (s, N, D)-семейством, если для любого набора номеров $m_1, m_2, \ldots, m_s, m_{s+1}, m_i \neq m_j$, $m_i \in [t]$:

$$\left| \Omega_{m_{s+1}} \setminus \bigcup_{k=1}^{s} \Omega_{m_k} \right| \geqq D .$$

Пусть $t(s, N, D)$ максимально возможное число членов (s, N, D)-семейства. Введем параметр d, $0 < d < 1$. Символом $\lfloor a \rfloor$ будем обозначать наибольшее целое $\leqq a$. Для фиксированных $s = 1, 2, \ldots$ и d, $0 < d < 1$, определим скорость

$$R(s, d) = \overline{\lim_{N \to \infty}} \frac{\log_2 (s, N, \lfloor Nd \rfloor)}{N} .$$

В работе построены верхние и нижние оценки $R(s, d)$, которые обобщают ранее полученные границы [2, 3] для частного случая $D = 1$.

Пусть $d_s = s^s/(s+1)^{s+1}$. Из результатов работы, в частности, следует, что $R(s, d) = 0$ при $d \geqq d_s$ и $R(s, d) > 0$ при $0 < d < d_s$.

А. Г. Дьячков
Московский Государственный университет им. М. В. Ломоносова,
механико-математический факультет, кафедра теории вероятностей
СССР, 119899, ГСП, Москва, Ленинские Горы

Problems of Control and Information Theory, Vol. 18 (4), pp. 251—260 (1989)

ESTIMATES IN STOCHASTIC PROGRAMMING — CHANCE CONSTRAINED CASE

V. KAŇKOVÁ

(Prague)

(Received August 5, 1988)

Solving a stochastic programming problem without any knowledge of the probability laws we get a statistical problem. It is the problem to estimate the optimal value and the optimal solution. If the probability laws are completely unknown then the empirical distribution function is usually set instead of the theoretical one to obtain some estimates. This idea appeared already in [3], [7], [14]. The statistical behaviour of the corresponding estimates was studied, for example, in [2], [5], [6], [8], [9], [10]. However, the recalled papers concerned the stochastic programming problems with deterministic constraints only.

The chance constrained case can be found in [13], where the convergence rate of the optimal solution estimates is studied.

In this paper we shall deal with the chance constrained case, too. In detail, we shall try to get conditions under which the empirical estimates of the optimal value are consistent. To get this we utilize the results of [7], [9], [11], [12].

The achieved result can be utilized always when the probability laws are unknown and simultaneously it is possible to obtain corresponding random sample of sufficient range. For example, such situation can happen in technical or economic applications (automatic control, distribution systems, planning of production etc.).

Introduction

Let $(\Omega, \mathscr{S}, P)$ be a probability space;

$\xi = \xi(\omega) = [\xi_1(\omega), \ldots, \xi_l(\omega)]$ be an l-dimensional random vector defined on $(\Omega, \mathscr{S}, P)$;

$F(z)$ be the distribution function of the random vector $\xi(\omega)$;

$\xi^k = \xi^k(\omega)$, $k = 1, 2, \ldots$, be a sequence of independent random vectors such, that for every k, $k = 1, 2, \ldots$ the random vector $\xi^k(\omega)$ has the same distribution function as the random vector $\xi(\omega)$;

$F_N(\cdot) = F_N(\cdot, \omega)$ be the empirical distribution function determined by the first N members of the random sequence $\{\xi^k\}_{k=1}^{\infty}$;

$f_i(x)$, $i = 1, 2, \ldots, l$ be real valued, continuous functions defined on E_n;

(E_n, $n \geqq 1$ denotes an n-dimensional Euclidean space).

Let, further, $g(x, z)$ be a real valued, continuous function defined on $E_n \times E_l$.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

If the sets $Z(x)$, $X(\alpha)$, $X_N(\alpha)=X_N(\alpha, \omega)$, $N=1, 2, \ldots$, $\alpha \in \langle 0, 1\rangle$ we defined by

$$Z(x)=\{z \in E_l^+,\ z=(z_1, \ldots, z_l):\ f_i(x) \leqq z_i,\ i=1, 2, \ldots, l\},$$

$$X(\alpha)=\{x \in E_n^+:\ P\{Z(x)\} \geqq \alpha\},$$

$$X_N(\alpha)=X_N(\alpha, \omega)=\{x \in E_n^+:\ P_N\{Z(x)\} \geqq \alpha\},$$

then the aim of the paper is to give the assumptions under which

$$P\left\{\omega:\ \left|\inf_{X_N(\alpha)} E_N g(x, \xi(\omega)) - \inf_{X(\alpha)} Eg(x, \xi(\omega))\right| \underset{(N\to\infty)}{\to} 0\right\}=1. \tag{1}$$

(E and E_N denote the theoretical and the empirical mathematical expectation, respectively, $P\{Z(x)\}=P\{\omega:\ \xi(\omega) \in Z(x)\}$, $P_N=P_N(\cdot, \omega)$ the empirical probability measure corresponding to the distribution function F_N, $E_n^+=\{x \in E_n: x=(x_1, \ldots, x_n):$ $x_i \geqq 0,\ i=1, 2, \ldots, l\}$).

Remark. It can generally happen that some symbols in (1) are not meaningful. However, this situation cannot appear under assumptions which will be respected in this paper.

Convergence results

If we assume

(i) $f_i(x)$, $i=1, 2, \ldots, l$ are real valued, continuous functions on E_n^+ such that
  a) $f_i(\mathbf{0})=0$, $i=1, 2, \ldots, l$, $\mathbf{0} \in E_n$,
  b) $f_i(x)$, $i=1, 2, \ldots, l$ are functions increasing in all components of the vector x,

(ii) $\xi(\omega)$ fulfils the conditions
  a) $P\{\omega:\ \xi(\omega) \in E_l^+\}=1$,
  b) the probability measure of the random vector $\xi(\omega)$ is absolutely continuous with respect to the Lebesgue measure in E_l,

(iii) the probability density corresponding to the probability measure of the random vector ξ is positive on E_l^+,

(iv) $f_i(x) \to +\infty$, $i=1, 2, \ldots, l$ for $\|x\| \to \infty$, $x \geqq 0$ componentwise ($\|\cdot\|$ denotes the Euclidean norm in E_l),

then we can formulate the main result of this paper.

Theorem 1. Let P be a complete probability measure, $\alpha \in (0, 1\rangle$. If assumptions (i), (ii), (iii), (iv) are fulfilled and if $g(x, z)$ is a continuous bounded function on $X(\alpha-\delta) \times E_l^+$ for a $\delta>0$, $\alpha-\delta \in (0, 1\rangle$, then

$$P\left\{\omega:\ \left|\inf_{X(\alpha)} \int g(x, z)dF(z) - \inf_{X_N(\alpha)} \int g(x, z)dF_N(z)\right| \underset{(N\to\infty)}{\to} 0\right\}=1.$$

Since the proof of this theorem is rather complicated and long, we present it, in detail, in the Appendix, separately. Here we give a short survey of the proof containing the essential ideas only.

We obtain immediately from the triangular inequality that

$$\left| \inf_{X(\alpha)} \int g(x,z)dF(z) - \inf_{X_N(\alpha)} \int g(x,z)dF_N(z) \right| \leqq$$

$$\leqq \left| \inf_{X(\alpha)} \int g(x,z)dF(z) - \inf_{X_N(\alpha)} \int g(x,z)dF(z) \right| + \tag{2}$$

$$+ \left| \inf_{X_N(\alpha)} \int g(x,z)dF(z) - \inf_{X_N(\alpha)} \int g(x,z)dF_N(z) \right|.$$

It is easy to see that to prove Theorem 1 it is sufficient to prove the following two assertions:

$$P\left\{\omega: \left| \inf_{X(\alpha)} \int g(x,z)dF(z) - \inf_{X_N(\alpha)} \int g(x,z)dF(z) \right| \underset{(N\to\infty)}{\to} 0 \right\} = 1, \tag{3}$$

$$P\left\{\omega: \left| \inf_{X_N(\alpha)} \int g(x,z)dF(z) - \inf_{X_N(\alpha)} \int g(x,z)dF_N(z) \right| \underset{(N\to\infty)}{\to} 0 \right\} = 1. \tag{4}$$

However, it will be proved in the Appendix that exactly these assertions follow from the next (quite important) Theorems.

Theorem 2. Let P be a complete probability measure, $\alpha \in (0, 1\rangle$. If assumptions (i), (ii), (iii), (iv) are fulfilled and if $f(x)$ is a continuous, real valued function on E_n^+ then

$$P\left\{\omega: \left| \inf_{X_N(\alpha)} f(x) - \inf_{X(\alpha)} f(x) \right| \underset{(N\to\infty)}{\to} 0 \right\} = 1. \tag{5}$$

Theorem 3. Let P be a complete probability measure. If $g(x, z)$ is a continuous, bounded function on $\mathscr{K} \times E_l$, where $\mathscr{K} \subset E_n^+$ is a compact convex set then

$$P\{\omega: |\int g(x,z)dF_N(z) - \int g(x,z)dF(z)| \underset{(N\to\infty)}{\to} 0 \text{ uniformly in } \mathscr{K}\} = 1. \tag{6}$$

The proofs of Theorem 2 and Theorem 3 will be given in the Appendix, too. The main attention will be paid to the case of Theorem 2 while the assertion of Theorem 3 follows practically from the assertion of [7].

In this part of our paper we also present one more rather important result following from Theorem 1.

Corollary. If the assumptions of Theorem 1 are fulfilled then

$$E \inf_{X_N(\alpha)} \int g(x, z) dF_N(z) \underset{(N \to \infty)}{\to} \inf_{X(\alpha)} Eg(x, \xi).$$

Since the proof of the Corollary is very similar to the proof of Theorem 1 in [9], we omit it.

Appendix

The aim of this part of the paper is to give the proof of Theorem 1. For this, we shall prove Theorem 2 and Theorem 3. As the proof of Theorem 2 is rather complicated and long we begin with this one.

First, we recall the definition of the Kuratowski convergence of sets [11].

Definition 1. Let $G \subset E_n$, $G_N \subset E_n$, $N = 1, 2, \ldots$ be closed sets. The sequence G_N converges to G (for $N \to \infty$) in the Kuratowski sense if

$$\lim_N \inf G_N = \lim_N \sup G_N = G,$$

where

$$\lim_N \inf G_N = \{x = \lim_N x_N\colon\ x_N \in G_N \text{ for all but finitely many } N\text{'s}\}, \tag{7}$$

$$\lim_N \sup G_N = \left\{ x = \lim_M x_M\colon\ x_M \in G_{N_M} \text{ for a } \{G_{N_M}\}_{M=1}^{\infty} \text{ subsequence of } \{G_N\}_{N=1}^{\infty} \right\}.$$

Here we prove some auxiliary assertions.

Lemma 1. If assumptions (i), (ii) are fulfilled then sets $X(\alpha)$ for $\alpha \in (0, 1\rangle$ are closed.

Proof. To prove the assertion we have to verify the validity of the implication

$$x_N \in X(\alpha), \qquad N = 1, 2, \ldots, \quad x_N \to x \Rightarrow x \in X(\alpha)$$

for an arbitrary $\alpha \in (0, 1\rangle$.

Therefore, let $\alpha \in (0, 1\rangle$ be arbitrary. It follows, immediately, from assumption (i) that for every $\varepsilon > 0$ there exists $N_0 = N_0(\varepsilon)$ such that

$$\alpha \leqq P\{\omega\colon f_i(x_N) \leqq \xi_i(\omega), \quad i = 1, 2, \ldots, l\} \leqq$$

$$\leqq P\{\omega\colon f_i(x) \leqq \xi_i(\omega), \quad i = 1, 2, \ldots, l\} +$$

$$+ \sum_{i=1}^{l} P\{\omega\colon\ \xi_i(\omega) \in (f_i(x) - \varepsilon, f_i(x) + \varepsilon), \quad \xi_j(\omega) > f_j(x) - \varepsilon, \quad j \neq i, \quad j = 1, 2, \ldots, l\}$$

for $N > N(\varepsilon)$.

However, according to assumption (ii) we get also

$$\alpha \leqq P\{\omega\colon f_i(x) \leqq \xi_i, \quad i = 1, 2, \ldots, l\}.$$

This completes the proof.

Lemma 2. If assumptions (i), (ii) are fulfilled and if $\{\delta_N\}_{N=1}^{\infty}$ is a sequence such that $\alpha + \delta_N \in (0, 1\rangle$, $N = 1, 2, \ldots, \alpha + \delta_N \to \alpha$ then

$$\limsup_N X(\alpha + \delta_N) \subset X(\alpha).$$

Proof. According to the definition, if $x \in \limsup_N X(\alpha + \delta_N)$ and $\delta > 0$ are arbitrary then there exists a subsequence $\{\delta_{N_k}\}_{k=1}^{\infty}$ of the sequence $\{\delta_N\}_{N=1}^{\infty}$, points $x_{N_k} \in X(\alpha + \delta_{N_k})$, $k = 1, 2, \ldots$ and k_0 such that $x_{N_k} \in U(x, \delta)$ for every $k > k_0 = k_0(\delta)$, [$U(x, \delta)$ denotes δ-surroundings of the point x]. Hence

$$\alpha + \delta_{N_k} \leqq P\{\omega\colon f_i(x_{N_k}) \leqq \xi_i(\cdot), \quad i = 1, 2, \ldots, l\} \leqq$$

$$P\{\omega\colon f_i(x) \leqq \xi_i(\omega), \quad i = 1, 2, \ldots, l\} +$$

$$+ \sum_{i=1}^{l} P\left\{\omega\colon \inf_{x' \in U(x,\delta)} f_i(x') \leqq \xi_i(\omega) \leqq \sup_{x' \in U(x,\delta)} f_i(x'),\ \xi_j(\omega) \geqq \right.$$

$$\left. \geqq \inf_{x' \in U(x,\delta)} f_j(x'), \quad j \neq i, \quad j = 1, 2, \ldots, l \right\} \text{ for every } k > k_0.$$

It is easy to see that $\alpha \leqq P\{\omega\colon f_i(x) \leqq \xi_i, i = 1, 2, \ldots, l\}$, too.

As $x \in \limsup X_N(\alpha)$ was an arbitrary point we proved Lemma 2.

Lemma 3. Let assumptions (i), (ii), (iii) be fulfilled. If $\{\delta_N\}_{N=1}^{\infty}$ is a sequence such that $\alpha + \delta_N \in (0, 1\rangle$, $N = 1, 2, \ldots, \alpha + \delta_N \to \alpha$ then

$$X(\alpha) \subset \liminf_N X(\alpha + \delta_N).$$

Proof. Let $x \in X(\alpha)$ be an arbitrary point. To prove the assertion it is necessary to find $x_N \in X(\alpha + \delta_N)$ for every N such that $x_N \to x (N \to \infty)$. For this we divide the set of the natural numbers $\mathcal{N}$ into three disjoint subsets $\mathcal{N}_a$, $\mathcal{N}_b$, $\mathcal{N}_c$

$$\mathcal{N}_a = \{N\colon \delta_N < 0\}$$

$$\mathcal{N}_b = \{N\colon \delta_N > 0, \quad P[\omega\colon f_i(x) \leqq \xi_i(\omega), \quad i = 1, 2, \ldots, l] > \alpha\}$$

$$\mathcal{N}_c = \{N\colon \delta_N > 0, \quad P[\omega\colon f_i(x) \leqq \xi_i(\omega), \quad i = 1, 2, \ldots, l] = \alpha\}.$$

It is easy to see that $\mathcal{N}_a \cup \mathcal{N}_b \cup \mathcal{N}_c = \mathcal{N}$. First we consider the set $\mathcal{N}_a$. Since $\delta_N < 0$, we get $X(\alpha) \subset X(\alpha + \delta_N)$. Consequently, we can define $x_N = x$ (for every $N \in \mathcal{N}_a$).

Further, we consider the set $\mathcal{N}_b$. If there exists only a final number of the elements belonging to the $\mathcal{N}_b$ then we can set $x_N = 0$ (for every $N \in \mathcal{N}_b$). If $\mathcal{N}_b$ is an infinite set then there exists N_0 such that

$$P\{\omega:\ f_i(x) \leqq \xi_i(\omega),\quad i = 1, 2, \ldots, l\} \geqq \alpha + \delta_N \text{ for every } N > N_0.$$

According to this we can set

$$x_N = 0 \qquad \text{for} \quad N < N_0, \quad N \in \mathcal{N}_b,$$

$$x_N = x \qquad \text{for} \quad N > N_0, \quad N \in \mathcal{N}_b.$$

It remains the case $N \in \mathcal{N}_c$. It is easy to see that if $\mathcal{N}_c$ is a finite set then we have already finished the proof of the Lemma. So, let us assume that X is infinite. We denote

$$A(x) = \{y \in E_n^+:\ y_i \leqq x_i,\quad y = (y_1, \ldots, y_n),\quad y_i \geqq 0,\quad x = (x_1, \ldots, x_n)\}.$$

Since $X(\alpha + \delta_N) \cap A(x) \neq \emptyset$ we can set

$$x_N = \operatorname*{arg\,inf}_{y \in X(\alpha + \delta_N) \cap A(x)} \|y - x\| \qquad \text{for} \quad N \in \mathcal{N}_c,$$

where $\|\cdot\|$ denotes the Euclidean norm in E_n. It is easy to see that the proof will be complete if we prove that $x_N \to x$ for $N \to \infty$, $N \in \mathcal{N}_c$. We shall prove it by contradiction. Let us assume that $x_N \nrightarrow X$ for $N \to \infty$, $N \in \mathcal{N}_c$. As $X(\alpha + \delta_N) \cap A(x)$ is a compact set, there exists $a \neq x$, $N_k \subset \mathcal{N}_c$, $x_{N_k} \in X(\alpha + \delta_{N_k})$, $k = 1, 2, \ldots$ such that $x_{N_k} \to a$ for $k \to \infty$. It follows immediately from this that there exists k_0 such that the following inequalities

$$\|x_{N_k} - x\| \geqq \frac{1}{2}\|x - a\|, \qquad \|x_{N_k} - a\| < \frac{1}{2}\|x - a\|$$

hold for every $k > k_0$.

Since $a \leqq x$ (componentwise) the last inequalities are in the contradiction with assumption (iii).

Lemma 4. If assumption (iv) is fulfilled then the sets $X(\alpha)$ for $\alpha \in (0, 1\rangle$ are bounded.

Proof. Let $\alpha \in (0, 1\rangle$ be arbitrary. We shall prove the assertion of Lemma 4 by contradiction. Let us assume that there exists a sequence $\{x_N\}_{N=1}^{\infty}$, $x_N \in X(\alpha)$ such that $\|x_N\| \to \infty$ for $N \to \infty$. Then

$$P\{\omega:\ L \leqq \xi_i(\omega) \text{ for at least one } i \in \{1, 2, \ldots, l\}\} \geqq \alpha$$

for an arbitrary large constant L. However, this is in contradiction with the elementary properties of the probability measure.

If we denote by $\mathcal{C}$ the set of the convex subsets of E_n then the following result is proved in [1].

Lemma 5. If assertion (ii) is fulfilled then

$$P\left\{\omega\colon \sup_{B\in\mathcal{B}} |P_N(B)-P(B)| \underset{(N\to\infty)}{\to} 0\right\}=1,$$

where $P(B)=P\{\omega\colon \xi(\omega)\in B\}$.

Proof. The assertion of Lemma 5 follows immediately from Theorem 1, and Corollary 1 in Appendix 1 of [1].

If, further, we define symbols $\alpha(x)$, $\alpha_N(x)$ in the following way

$$\alpha(x)=P\{Z(x)\} \tag{8}$$

$$\alpha_N(x)=P_N\{Z(x)\}$$

for every $x\in E_n^+$ then we can see that

$$X(\alpha)=\{x\colon \alpha(x)\geqq\alpha\}, \qquad X_N(\alpha)=\{x\colon \alpha_N(x)\geqq\alpha\}.$$

Lemma 6. Let $\delta>0$ be arbitrary, P be a complete probability measure. Under assumptions (i), (ii), the following assertion holds:

$$P\{\omega\colon \text{there exists } N_0 \text{ such that } |P_N(Z(x))-P(z(x))|<\delta$$

$$\text{for every } N>N_0, \quad x\in E_n^+\}=1. \tag{9}$$

Proof. Since it is easy to see that $Z(x)$ is a convex set for every $x\in E_n^+$ we can get the assertion of Lemma 6 from the assertion of Lemma 5, and the properties of the probability measure.

Lemma 7. Let P be a complete probability measure. Let, further, α, $\delta>0$, $\alpha-\delta$, $\alpha+\delta\in(0,1\rangle$ and the assumptions (i), (ii) be fulfilled, then

$$P\{\omega\colon \text{there exists } N_0 \text{ such that } X(\alpha+\delta)\subset X_N(\alpha)\subset X(\alpha-\delta) \text{ for every } N>N_0\}=1.$$

Proof. Let $\delta>0$ and $N\in\mathcal{N}$ be arbitrary. Since it is easy to see that

$$\{\omega\colon |P_N(Z(x))-P(Z(x))|<\delta \text{ for a } x\in X_N(\alpha,\omega)$$

and simultaneously

$$|P_N(Z(x))-P(Z(x))|<\delta \text{ for a } x\in X(\alpha+\delta)\}\supset$$

$$\supset\{\omega\colon |P_N(Z(x))-P(Z(x))|<\delta \text{ for } x\in E_n^+\},$$

and since P is a complete probability measure then by Lemma 6 $P\{\omega\colon$ there exists N_0 such that

$$P[Z(x)]\geqq\alpha-\delta \text{ for a } x\in X_N(\alpha)$$

and simultaneously

$$P_N[Z(x)]\geqq\alpha \text{ for a } x\in X(\alpha+\delta), \text{ for every } N>N_0\}=1.$$

It is evident that the last relation completes the proof.

If we denote $\bar{X}_N(\alpha) = \text{cl } X_N(\alpha)$ for every $\alpha \in (0, 1\rangle$ then we get also

Lemma 8. Let P be a complete probability measure. Let, further, α, $\delta > 0$, $\alpha + \delta \in (0, 1\rangle$ and assumptions (i), (ii) be fulfilled, then

$$P\{\omega\text{: there exists } N_0 \text{ such that } X(\alpha+\delta) \subset \bar{X}_N(\alpha) \subset$$

$$\subset X(\alpha-\delta) \text{ for every } N > N_0\} = 1.$$

Proof. The assertion of Lemma 8 follows immediately from the assertion of Lemma 7 and the assertion of Lemma 1.

Lemma 9. Let P be complete probability measure. If assumptions (i), (ii), (iii), (iv) are fulfilled then

$$P\{\omega\text{: } \bar{X}_N(\alpha) \to X(\alpha) \text{ in the Kuratowski sense}\} = 1$$

for every $\alpha \in (0, 1\rangle$.

Proof. Let $\{\delta_k\}_{k=1}^{\infty}$, $\delta_k > 0$ be a sequence such that $\delta_K \downarrow 0$. It follows from Lemma 8 that

$$P\{\omega\text{: for every } \delta_k,\ k = 1, 2, \ldots \text{ there exists } N_0(k) \in \mathcal{N} \text{ such that}$$

$$X(\alpha+\delta_k) \subset \bar{X}_N(\alpha) \subset X(\alpha-\delta_k) \text{ for every } N > N_0(k)\} = 1.$$

However, since $\bar{X}_N(\alpha)$, $\alpha \in (0, 1\rangle$, $X(\alpha+\delta_k)$, $X(\alpha-\delta_k)$ are closed sets it is easy to see that $\lim\limits_N \inf \bar{X}_N(\alpha)$ and $\lim\limits_N \sup \bar{X}_N(\alpha)$ are closed sets, too, and further

$$P\{\omega\text{: } X(\alpha+\delta_k) \subset \lim_N \inf \bar{X}_N(\alpha) \subset \lim_N \sup \bar{X}_N(\alpha) \subset X(\alpha-\delta_k),$$

$$\text{for every } \delta_k, \quad k = 1, 2, \ldots,\} = 1.$$

Further, according to Lemma 2 and Lemma 3 we get

$$X(\alpha) \subset \lim \inf \bar{X}_N(\alpha) \subset \lim \sup \bar{X}_N(\alpha) \subset X(\alpha)$$

with the probability one. This completes the proof.

A convergence assertion proved in [11] is that is useful for us [Theorem 3.3]. We recall it in a special form only.

Theorem 4. If a) $f(x)$ is a continuous function on E_n^+,
b) $\alpha \in (0, 1\rangle$, $X'_N(\alpha)$, $N = 1, 2, \ldots$ are deterministic, closed sets such that

$$X'_N(\alpha) \to X(\alpha) \text{ in the Kuratowski sense,}$$

c) there exists a closed ball B such that

$$X(\alpha) \subset B, \; X'_N(\alpha) \subset B, \qquad N = 1, 2, \ldots$$

then

$$\min_{X'_N(\alpha)} f(x) \to \min_{X(\alpha)} f(x).$$

Proof of Theorem 2. Let $\delta > 0$, $\alpha - \delta$, $\alpha + \delta \in (0, 1\rangle$ be arbitrary. It follows from Lemma 1 and Lemma 5 that $X(\alpha - \delta)$ is a compact set. According to this and to Lemma 8 it is easy to see that with the probability one there exists $N_0 \in \mathcal{N}$ such that $\bar{X}_N(\alpha)$, $N = N_0, N_0 + 1, \ldots$ are compact sets, too. Since, further

$$\min_{\bar{X}_N(\alpha)} f(x) = \inf_{X_N(\alpha)} f(x) \tag{10}$$

we get from Lemma 9 and Theorem 4 that

$$\inf_{X_N(\alpha)} f(x) \underset{(N \to \infty)}{\to} \inf_{X(\alpha)} f(x) \quad \text{a.s.}$$

Proof of Theorem 3. Since the proof of the corresponding assertion for the case of the maximization is given in [7] (relation (6)), we shall omit it here.

Proof of Theorem 1. It follows from the previous part that to prove Theorem 1 it is sufficient to prove relations (3) and (4). However the validity of relation (3) follows immediately from the assertion of Theorem 2 and from the continuity of the function $\int g(x, z) dF(z)$ on every compact set $\mathcal{K} \subset E_n^+$.

It remains to prove relation (4). For this we can utilize the assertion of Theorem 3. Namely, if we set $\mathcal{K} = X(\alpha - \delta)$ for a $\delta > 0$, $\alpha - \delta \in (0, 1\rangle$ then, according to relation (10) we get the validity of the assertion of Theorem 1.

References

1. *Боровков, А. А.* Математическая статистика, оценка параметров, проверка гипотез. Наука, Москва 1984.
2. *Цибаков, А. В.* Оценки точности метода минимизации эмпирического риска. Проблемы передачи информаяци **17,** *1,* 50–61.
3. *Dupačová, J.*, Experience in Stochastic Programming Models. In: A. Prékopa (ed.): Survey of Math. Programming. Proc. IX. Int. Math. Progr. Symp., Budapest 1976. Akadémiai Kiadó, Budapest, pp. 99–105.
4. *Dupačová, J.*, Stochastic Programming with Incomplete Information. A Survey of Results on Postoptimization and Sensitivity Analysis. Optimization **18** (1987), *4*, pp. 507–532.
5. *Dupačová, J., Wets, R.*, Asymptotic Behaviour of Statistical Estimators and Optimal Solutions for Stochastic Optimization Problems, I and II. IIASA, Laxenburg, Austria WP-86-41 (1986), WP-87-9 (1987).

6. *Dupačová, J., Wets, R.*, Asymptotic Behaviour of Statistical Estimators and the Optimal Solutions of Stochastic Optimization Problems. Annals of Math. Stat. 1988, 4.
7. *Kaňková, V.*, Optimum Solution of a Stochastic Problem with Unknown Parameters. In: Trans. of the Seventh Prague Conference 1974, Academia, Prague 1977, pp. 239–244.
8. *Kaňkova, V.*, An Approximative Solution of a Stochastic Optimization Problem. In: Trans. of the Eighth Prague Conference, Prague 1978, Academia, Prague, pp. 327–332.
9. *Kaňková, V.*, Empirical Estimates in Stochastic Programming. In: Trans. of the Tenth Prague Conference, Prague 1986, Academia, Prague.
10. *Römisch, W., Wakolbinger, A.*, Obtaining Convergence Rates for Approximations in Stochastic Programming. In: Parametric Optimization and Related Topics. Akademie-Verlag, Berlin.
11. *Salinetti, G.*, Approximations for Chance Constrained Programming Problems. Stochastics (1983), **10,** pp. 157–179.
12. *Salinetti, G., Wets, R.*, On the Convergence of Closed-Valued Measurable Multifunctions. Trans. of the Americal Society **266** (1981), *1*, pp. 275–289.
13. *Vogel, S.*, Stability Results for Stochastic Programming Problems. Optimization **19** (1988), *2*, pp. 269–288.
14. *Wets, R.*, A Statistical Approach to the Solution of Stochastic Programs with (Convex) Simple Recourse. Research Report, Univ. Kentucky, USA 1979.

Об оценках в стохастическом программировании — случай с вероятностным ограничением

В. КАНЬКОВА

(Прага)

В задаче стохастического программирования с вполне неизвестными вероятностными законами распределения можно получить оценки оптимального решения и оптимального значения математического ожидания оптимизирующей функции при помощи эмпирической функции распределения случайных элементов. В литературе теория так называемых эмпирических оценок разработана для случая задач с детерминированными ограничениями. Однако теория этих оценок в случае вероятностных ограничений разработана недостаточно.

В настоящей работе даны условия, в которых эмпирические оценки сходятся к теоретическим с вероятностью единица.

V. Kaňková
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. 18 (4), pp. 261—276 (1989)

ON THE PROBLEM OF MINIMAX ESTIMATION OF LINEAR FUNCTIONALS

R. DODUNEKOVA*

(Sofia)

(Received September 20, 1988)

Let S be an unknown function, a signal passing through a channel for information transmission and L be a linear functional. Consider the ratio Δ_0^2/Δ^2, where Δ_0^2 is the linear minimax risk of $L(s)$ obtained by using direct observations (signal + noise) and Δ^2 is the linear minimax risk obtained by using indirect observations (signal + noise → linear filter + noise). We find out a large class of functionals L, for which under some conditions there exists a turning point in the behaviour of this ratio. Up to this point the filter causes great informational losses but after it the losses are practically zero.

1. Introduction

This work is in close relation to [1], so we give the problems considered there and the main results. In radio systems we sometimes meet with the following situation. Let the input signal $s(t)$ be a function from a Hilbert space $L_2[0, 1] = L_2$. This signal passes through a channel with an additive white Gaussian noise $\varepsilon \dot{W}_1$ with intensity ε^2. The result $\dot{X}(t)$ goes through a linear filter described by a linear n-th order differential equation with constant coefficients. The output is the signal $\dot{Z}(t)$, which is the sum of the filter's output and another, independent of the first one, white Gaussian noise $\delta \dot{W}_2$. This information transmission channel can be described by Fig. 1.

The observations $\dot{Z}$ are usually called indirect. When the output of the channel is the process $\dot{X}$, we speak about direct observations.

We consider two problems in this scheme of observations.

(A) We wish to estimate in the minimax sense the value $L(s)$, where L is a linear functional. For this we use $\dot{Z}$ (indirect observations) and also the information that the unknown signal s contains in the convex symmetric set $\sum$ of L_2.

(B) Suppose $\dot{X}$ is observed (direct observations). We want to compare the minimax estimator of $L(s)$ obtained in this case (see [7]) with the estimator in (A) in order to find out if the linear filter causes great informational losses.

* Research partially supported by the Ministry of Science, Culture and Education in Bulgaria, under Contract No. 1035/1988.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

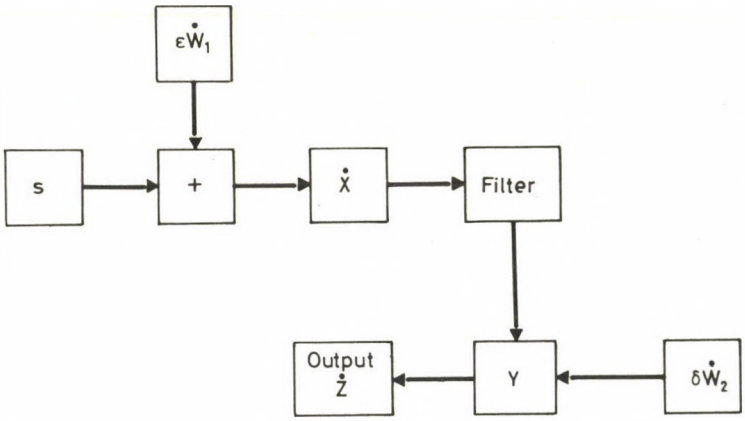


Fig. 1

The processes $X(t)$ and $Z(t)$ have representations

$$X(t)=\int_0^1 s(u)du+\varepsilon W_1(t), \tag{1.1}$$

$$Z(t)=\int_0^1 Y(u)du+\delta W_2(t), \tag{1.2}$$

where $Y(t)$ and $\dot{X}(t)$ are in the following relation

$$Y^{(n)}(t)+a_1 Y^{(n-1)}(t)+\ \ldots\ +a_n Y(t)=\dot{X}(t),$$
$$Y(0)=Y'(0)=\ \ldots\ =Y^{(n-1)}(0)=0 \tag{1.3}$$

and $a_1, a_2, \ldots, a_n$ are constants.

We deal with the class M of all linear estimators of the form

$$\hat{L}=\int_0^1 y(t)dZ(t), \qquad y\in L_2.$$

The linear minimax estimator $\hat{L}_0$ is defined as delivering inf in the relation

$$\Delta^2(\varepsilon,\delta)=\inf_{\hat{L}\in M}\sup_{s\in\Sigma} E|\hat{L}-L(s)|^2,$$

by which the linear minimax risk of $L(s)$ is defined.

Problem (A) is equivalent, as shown in [1], to the problem of estimating $L(s)$ when s is observed with a generalized Gaussian process N, i.e., the "observed" process Z_1 has the representation

$$Z_1=s+N(\varepsilon,\delta). \tag{1.4}$$

This general problem is considered in [2]. Using the results there, [1] gives a constructive theorem for finding $\hat{L}_0$ and $\Delta^2(\varepsilon, \delta)$, which is formulated in terms of the correlation operator $R_{\varepsilon,\delta}$ of the process N in the equivalent problem (1.4).

It is easy to show, that if the functional L is unbounded on $\sum$, then $\Delta^2(\varepsilon, \delta) = \infty$. That is why in problem (B) we assume that L is bounded on $\sum$. Let $\Delta_0^2(\varepsilon)$ be the linear minimax risk in the direct case. In [1] we compare $\Delta_0^2(\varepsilon)$ and $\Delta^2(\varepsilon, \delta)$ when the intensities ε and δ tend to zero commensurable: $\delta = \varepsilon^\gamma$, $\gamma > 0$, and $\sum$ is an ellipsoid:

$$\sum = \left\{ s \in L_2 : \sum_{i=1}^{\infty} s_i^2 \lambda_i^{2k} < 1 \right\}. \tag{1.5}$$

Here $s_i = (s, \varphi_i)$ (inner product) and $\{\varphi_i\}_{i=1}^{\infty}$, $\{\lambda_i\}_{i=1}^{\infty}$ are eigenfunctions and eigenvalues of $R_{1,1}$. This operator is a differential of $2n$-th order, self-adjoined and positive. Functions φ_i are indefinitely differentiable and form an orthogonal normed base in L_2; the eigenvalues are greater than 1 and are equivalent to $(2\pi i)^{2n}$, when $i \to \infty$:

$$\lim_{i \to \infty} \frac{\lambda_i}{(2\pi i)^{2n}} = 1. \tag{1.6}$$

One can show that if $\gamma > \gamma_0 = 1 + \dfrac{1}{2k}$, then the two risks are equivalent when $\varepsilon \to 0$:

$$\lim_{\varepsilon \to 0} \frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon, \varepsilon^\gamma)} = 1, \qquad \gamma > \gamma_0,$$

which means that the informational losses caused by the filter are practically zero. If $\gamma = \gamma_0$, then

$$\frac{1}{2} < \frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon, \varepsilon^\gamma)} < 1.$$

Example 3′ in [8] shows that for some functionals L the ratio $\Delta_0^2(\varepsilon)/\Delta^2(\varepsilon, \varepsilon^\gamma)$ tends to 0, when $\gamma < \gamma_0$ (and $\varepsilon \to 0$), and to c, $\dfrac{1}{2} < c < 1$, when $\gamma = \gamma_0$. It means, then, that γ_0 is a real threshold: when $\varepsilon \to 0$, for $0 < \gamma < \gamma_0$, the informational losses caused by the filter are unmeasurable large; for $\gamma = \gamma_0$ the two risks differ only by a multiplicant practically (but $\Delta_0^2(\varepsilon) < \Delta^2(\varepsilon, \varepsilon^{\gamma_0})$, of course); for $\gamma > \gamma_0$ the two risks are practically the same. One cannot expect that this is valid for all functionals bounded on $\sum$, because, for example, if $\sum \lambda_i l_i^2 < \infty$, then number 1 is a real threshold.

In this paper we give a large class of linear functionals for which the real threshold in the behaviour of the ratio Δ_0^2/Δ^2 is the number γ_0.

2. Basic theorem

The linear functional L, being defined on $\sum$, is not necessarily defined at other points of L_2. But the values $l_i = L(\varphi_i)$, $i = 1, 2, 3, \ldots$, are defined, since the functions $(2\lambda_i)^{-k}\varphi_i$ lie in $\sum$.

Theorem. Let in (1.1)–(1.3) $\delta = \varepsilon^{\gamma}$, $\gamma > 0$ and $\sum$ is the ellipsoid (1.5). Assume that the functional L is bounded on $\sum$ and unbounded on L_2, when it is defined there, and that the sequence $\{l_i^2\}_{i=1}^{\infty}$ is regularly varying with the index of variation $\rho \neq 0$; $\rho \neq 1$; $\rho \neq 4nk - 1$. Then

$$\frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon, \varepsilon^{\gamma})} \underset{\varepsilon \to 0}{\to} 0, \qquad 0 < \gamma < \gamma_0; \tag{2.1}$$

$$\frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon, \varepsilon^{\gamma_0})} \underset{\varepsilon \to 0}{\to} C, \qquad \frac{1}{2} < C < 1; \tag{2.2}$$

$$\frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon, \varepsilon^{\gamma})} \underset{\varepsilon \to 0}{\to} 1, \qquad \gamma > \gamma_0. \tag{2.3}$$

So relations (2.1)–(2.3) show that the number $\gamma_0 = 1 + \frac{1}{2k}$ is a turning point in the problem, when $\varepsilon \to 0$ and L, and $\sum$ are of such kind as the theorem requires. If $0 < \gamma < \gamma_0$, the direct risk is infinitesimal with respect to the indirect risk and then the evaluation in the indirect case is of much worse quality than the one in the direct case. If $\gamma = \gamma_0$, the ratio of the risks is close to a constant, the indirect risk being nevertheless greater than the direct one. If $\gamma > \gamma_0$, the two risks are equivalent and so there are no informational losses caused by the linear filter. Note again more that (2.3) is true for all functionals bounded on $\sum$.

3. Some subsidiary results

In this section we follow [3], [4] and [5] to give some results necessary for the proof and concerning the regularly varying sequences and functions.

We call a sequence $\{\theta(m)\}_{m=1}^{\infty}$ of positive terms regularly varying if there is a sequence of positive terms $\{\alpha(m)\}_{m=1}^{\infty}$ satisfying

$$\theta(m) \sim k\alpha(m), \qquad k > 0, \text{ constant},$$

$$m\left[1 - \frac{\alpha(m-1)}{\alpha(m)}\right] \underset{m \to \infty}{\to} \rho, \qquad \rho \text{ finite}. \tag{3.1}$$

The number ρ is called the index of variation. The case $\rho = 0$ is called slowly varying.

We call a positive function $\mathscr{R}(x)$, $x>0$, regularly varying if there exists a number $\rho\in(-\infty,\infty)$, such that for every $\lambda>0$ it is fulfilled that

$$\lim_{x\to\infty}\frac{\mathscr{R}(\lambda x)}{\mathscr{R}(x)}=\lambda^{\rho}. \tag{3.2}$$

The number ρ is also called the index of variation. The case $\rho=0$ is called slowly varying. The function $\mathscr{R}$ is regularly varying if and only if it has the representation

$$\mathscr{R}(x)=x^{\rho}\mathscr{L}(x), \tag{3.3}$$

where $\mathscr{L}$ is slowly varying. If $\mathscr{L}$ is a slowly varying function and $\sigma>0$, then

$$\lim_{x\to\infty}x^{\sigma}\mathscr{L}(x)=\infty, \qquad \lim_{x\to\infty}x^{-\sigma}\mathscr{L}(x)=0 \tag{3.4}$$

(see [5], Section 1.5).

According to [3], if $\{\theta(m)\}_{m=1}^{\infty}$ is regularly varying sequence with index of variation ρ, then the function $\theta(x)=\theta([x])$, $x>0$ ($[x]$ denotes the integer part of x) is regularly varying with the same index of variation ρ. From (3.1) it follows that

$$\lim_{m\to\infty}\frac{\theta(m-1)}{\theta(m)}=1. \tag{3.5}$$

Note that definition (3.1) is equivalent to a definition which is analogous to (3.2): $\theta(m)$ is regularly varying, if for every $\lambda>0$,

$$\lim_{m\to\infty}\frac{\theta([\lambda m])}{\theta(m)}=\lambda^{\rho}, \quad -\infty<\rho<\infty.$$

We will be interested further in the behaviour of the integrals of the form

$$\int_a^b f(y)\mathscr{L}(xy)dy,$$

where $0\leqq a<b\leqq\infty$, $x>0$, $\mathscr{L}$ is a slowly varying function and f is integrable on $[a,b]$. In [5], Section 2.3, it is proved that

$$\int_a^b f(y)\mathscr{L}(xy)dy \underset{x\to\infty}{\sim} \mathscr{L}(x)\int_a^b f(y)dy, \tag{3.6}$$

if $0<a,\ b<\infty$;

$$\int_a^{\infty} f(y)\mathscr{L}(xy)dy \underset{x\to\infty}{\sim} \mathscr{L}(x)\int_a^{\infty} f(y)dy, \tag{3.7}$$

if $a>0$ and, for some $\alpha>0$ the integral $\int_a^\infty y^\alpha f(y)dy$ converges;

$$\int_0^b f(y)\mathscr{L}(xy)dy \underset{x\to\infty}{\sim} \mathscr{L}(x)\int_0^b f(y)dy, \tag{3.8}$$

if $b<\infty$ and, for some $\beta>0$, the integral $\int_0^b y^{-\beta} f(y)dy$ converges.

4. About the conditions of the theorem

Trying to find linear functionals for which the number $\gamma_0=1+\frac{1}{2k}$ is a threshold we do not need to consider the functionals unbounded on $\sum$ as for them $\Delta^2(\varepsilon,\delta)=\infty$, as it was already noted. Next, we show that the functionals bounded on L_2 can be eliminated, too. When $\sum$ is the ellipsoid (1.5) the two linear minimax risks are found to be

$$\Delta_0^2(\varepsilon)=\sum_{i=1}^\infty \frac{\varepsilon^2 l_i^2}{1+\varepsilon^2\lambda_i^{2k}}, \qquad \Delta^2(\varepsilon,\delta)=\sum_{i=1}^\infty \frac{[\delta^2(\lambda_i-1)+\varepsilon^2]l_i^2}{1+\varepsilon^2\lambda_i^{2k}+\delta^2\lambda_i^{2k}(\lambda_i-1)} \tag{4.1}$$

(see [7], (4.1) and [1], (5.1), correspondingly). When $\delta=\varepsilon^\gamma$ and $\gamma>1$, then, from (4.1) it follows that

$$\Delta^2(\varepsilon,\varepsilon^\gamma) \underset{\varepsilon\to 0}{\sim} \sum_{i=1}^\infty \frac{(\varepsilon^{2\gamma}\lambda_i+\varepsilon^2)l_i^2}{1+\varepsilon^2\lambda_i^{2k}+\varepsilon^{2\gamma}\lambda_i^{2k+1}}. \tag{4.2}$$

This relation allows us to replace $\Delta^2(\varepsilon,\varepsilon^\gamma)$ by the expression on the right-side of (4.2), when considering $\lim_{\varepsilon\to 0}\frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon,\varepsilon^\gamma)}$. Write for it $\mathcal{A}^2(\varepsilon,\gamma)$:

$$\mathcal{A}^2(\varepsilon,\gamma)=\sum_{i=1}^\infty \frac{(\varepsilon^{2\gamma}\lambda_i+\varepsilon^2)l_i^2}{1+\varepsilon^2\lambda_i^{2k}+\varepsilon^{2\gamma}\lambda_i^{2k+1}}. \tag{4.3}$$

Now let the functional L be bounded on L_2. That is equivalent, as it is well known, to the condition $\sum_{i=1}^\infty l_i^2<\infty$. Then from (4.1) we come to $\Delta_0^2(\varepsilon)\underset{\varepsilon\to 0}{\sim}\varepsilon^2\sum_{i=1}^\infty l_i^2$. Taking in (4.3) $\gamma=\gamma_0=1+\frac{1}{2k}$ one can obtain (the detailed calculations are omitted) $\mathcal{A}^2(\varepsilon,\gamma_0)\underset{\varepsilon\to 0}{\sim}\varepsilon^2\sum_{i=1}^\infty l_i^2$ which shows that the number γ_0 does not appear as a threshold in the case of functionals bounded on L_2.

Define $l_0^2 = l_1^2$, $l^2(x) = l^2([x])$, $x > 0$. This is a regularly varying function with the index of variation ρ. Then (see (3.3))

$$l^2(x) = x^\rho \mathscr{L}(x), \qquad l_i^2 = i^\rho \mathscr{L}(i). \tag{4.4}$$

We can verify that in our theorem the conditions for L to be bounded on $\sum$ and $\rho \neq 4nk - 1$ are equivalent to the condition

$$\rho < 4nk - 1. \tag{4.5}$$

For this we need that L is bounded on $\sum$ if and only if

$$\sum_{i=1}^{\infty} \frac{l_i^2}{\lambda_i^{2k}} < \infty \tag{4.6}$$

(see [7], example 3). Let $\rho = 4nk - 1 - 2\varkappa$, $\varkappa > 0$. Using (4.4), (1.6) and (3.4) we can get (the calculations are omitted again)

$$\sum_{i=1}^{\infty} \frac{l_i^2}{\lambda_i^{2k}} \leqq \text{const.} \sum_{i=1}^{\infty} \frac{1}{i^{1+\varkappa}} < \infty$$

i.e., L is bounded on $\sum$. When $\rho = 4nk - 1 + \varkappa$, $\varkappa > 0$, we get

$$\sum_{i=1}^{\infty} \frac{l_i^2}{\lambda_i^{2k}} \geqq \text{const.} \sum_{i=1}^{\infty} \frac{1}{i} = \infty$$

and then

$$\sum_{i=1}^{\infty} l_i^2 = \infty, \tag{4.7}$$

i.e., L is unbounded on $\sum$.

One can show also that the condition

$$\rho > -1 \tag{4.8}$$

is equivalent to the following two conditions of the theorem: L is unbounded on L_2 (when it is defined there) and $\rho \neq -1$.

What happens if $\rho = -1$ or $\rho = 4nk - 1$? In these cases conditions (4.6) and (4.7) may be fulfilled or may not; it depends on the sequence $\mathscr{L}(i)$. For example, let $\rho = -1$. Then $\sum_{i=1}^{\infty} l_i^2 = \sum_{i=1}^{\infty} \frac{\mathscr{L}(i)}{i}$. The function $\mathscr{L}(x) = \ln x$, $x > 0$, is slowly varying, as the function $\mathscr{L}_1(x) = (\ln x)^{-2}$, $x > 1$. It is clear that

$$\sum_{i=1}^{\infty} \frac{\mathscr{L}(i)}{i} = \infty, \qquad \sum_{i=1}^{\infty} \frac{\mathscr{L}_1(i)}{i} < \infty.$$

When $\rho = 4nk - 1$, then

$$\sum_{i=1}^{\infty} \frac{l_i^2}{\lambda_i^{2k}} = \sum_{i=1}^{\infty} \frac{\mathscr{L}(i)}{i}$$

and the situation is the same. Hence, if $\rho = -1$ or $\rho = 4nk - 1$ one has to select the slowly varying functions in (4.4) such that conditions (4.6) and (4.7) would be satisfied, i.e., the functional would be bounded on $\sum$ and unbounded on L_2, and only after that to investigate whether γ_0 is a threshold.

The Theorem does not consider the slowly varying case ($\rho = 0$). It was done only for simplicity. In this case, if, moreover, $\{l_i^2\}_{i=1}^{\infty}$ is monotonic, the Theorem also holds; one can verify this by following the proofs.

5. Preliminary results

Denote

$$\Delta_1^2 = \sum_{i=1}^{\infty} \frac{\varepsilon^{2\gamma} \lambda_i l_i^2}{1 + \varepsilon^2 \lambda_i^{2k} + \varepsilon^{2\gamma} \lambda_i^{2k+1}},$$

$$\Delta_2^2 = \sum_{i=1}^{\infty} \frac{\varepsilon^2 l_i^2}{1 + \varepsilon^2 \lambda_i^{2k} + \varepsilon^{2\gamma} \lambda_i^{2k+1}},$$

and let $\tilde{\Delta}_0^2$, $\tilde{\Delta}_j^2$, $j = 1, 2$ and $\tilde{\Delta}^2(\varepsilon, \gamma)$ be the series obtained from Δ_0^2, Δ_j^2 and $\Delta^2(\varepsilon, \gamma)$ by writting $\tilde{l}_i^2$ instead of l_i^2.

Lemma 5.1. Let $\sum_{i=1}^{\infty} l_i^2 = \infty$ and $\{\tilde{l}_i^2\}_{i=1}^{\infty}$ be a sequence which is equivalent to $\{l_i^2\}_{i=1}^{\infty}$, when $i \to \infty$. If $\gamma > 1$, then

$$\Delta_j^2 \underset{\varepsilon \to 0}{\sim} \tilde{\Delta}_j^2 \qquad \text{and} \qquad \Delta_0^2 \underset{\varepsilon \to 0}{\sim} \tilde{\Delta}_0^2, \quad \Delta^2 \underset{\varepsilon \to 0}{\sim} \tilde{\Delta}^2.$$

Proof. Consider the last relation. Write S_N for the N-th sum of the series $\Delta^2(\varepsilon, \gamma)$ in (4.3):

$$S_N = \varepsilon^2 \sum_{i=1}^{N} \frac{(\varepsilon^{2(\gamma-1)} \lambda_i + 1) l_i^2}{1 + \varepsilon^2 \lambda_i^{2k} + \varepsilon^{2\gamma} \lambda_i^{2k+1}}.$$

Note, that

$$\frac{S_N}{S_Q} \underset{\varepsilon \to 0}{\to} \frac{\sum_{1}^{N} l_i^2}{\sum_{1}^{Q} l_i^2} \underset{Q \to \infty}{\to} 0.$$

Let $\tau>0$ be a small number. Choose N such that $\left|1-\frac{\bar{l}_i^2}{l_i^2}\right|<\frac{\tau}{2}$, if $i>N$, and Q such that

$$\sum_{i=1}^{N} l_i^2\left[\sum_{i=1}^{Q} l_i^2\right]^{-1}<\frac{\tau}{4A}, \qquad A=\max_{i\leqq N}\left|1-\frac{\bar{l}_i^2}{l_i^2}\right|.$$

Finally, let $\varepsilon(\tau)$ be such that if $\varepsilon<\varepsilon(\tau)$, then

$$\frac{S_N}{S_Q}<\frac{\sum_{i=1}^{N} l_i^2}{\sum_{i=1}^{Q} l_i^2}+\frac{\tau}{4A}.$$

Consider

$$|\Delta^2(\varepsilon,\gamma)-\bar{\Delta}^2(\varepsilon,\gamma)|\leqq\sum_{i=1}^{\infty}\frac{(\varepsilon^{2\gamma}\lambda_i+\varepsilon^2)l_i^2}{1+\varepsilon^2\lambda_i^{2k}+\varepsilon^{2\gamma}\lambda_i^{2k+1}}\left|1-\frac{\bar{l}_i^2}{l_i^2}\right|$$

$$\leqq AS_N+\frac{\tau}{2}\Delta^2(\varepsilon,\gamma),$$

$$\left|1-\frac{\bar{\Delta}^2(\varepsilon,\gamma)}{\Delta^2(\varepsilon,\gamma)}\right|\leqq\frac{AS_N}{S_Q}+\frac{\tau}{2}<\tau, \qquad \text{for} \quad \varepsilon<\varepsilon(\tau).$$

Proofs of the remaining relations are analogous and in the case $j=1$ use also the fact that $\sum_{i=1}^{\infty}\lambda_i l_i^2=\infty$, too.

Corollary 5.2. Let $\sum_{i=1}^{\infty} l_i^2=\infty$ and $N>0$ be a fixed integer. If $\gamma>1$, then

$$\Delta_0^2(\varepsilon)\underset{\varepsilon\to 0}{\sim}\varepsilon^2\sum_{i=N+1}^{\infty}\frac{l_i^2}{1+\varepsilon^2\lambda_i^{2k}},$$

$$\Delta^2(\varepsilon,\gamma)\underset{\varepsilon\to 0}{\sim}\sum_{i=N+1}^{\infty}\frac{(\varepsilon^{2\gamma}\lambda_i+\varepsilon^2)l_i^2}{1+\varepsilon^2\lambda_i^{2k}+\varepsilon^{2\gamma}\lambda_i^{2k+1}}.$$

Proof. Write S^N for the series in the right-hand side of the second relation. We have

$$\frac{S^N}{\Delta^2(\varepsilon,\gamma)}=1-\frac{S_N}{S_Q}\to 1, \qquad \text{when} \quad \varepsilon\to 0 \quad \text{and} \quad Q\to\infty.$$

The first relation can be proved similarly.

Remark 5.3. According to [5], Section 1.5, for the regularly varying function $l^2(x)$ in (4.4) there exists a regularly varying function $\bar{l}^2(x)$ with the same index of variation ρ which is monotonic on $[V,+\infty)$, where V is a positive constant, large enough, and $\bar{l}^2(x)\underset{x\to\infty}{\sim}l^2(x)$. If $\gamma>1$ the lemma allows us to write $\bar{l}_i^2$ instead of l_i^2, when

$\varepsilon \to 0$, and according to the corollary, $\overline{l}_i^2$ may be considered monotonic for $i \geqq 1$. Therefore, when $\gamma > 1$ we will assume that $\{l_i^2\}_{i=1}^{\infty}$ is monotonic.

Lemma 5.4. Let the sequence $\{\tilde{\lambda}_i\}_{i=1}^{\infty}$ be equivalent to $\{\lambda_i\}_{i=1}^{\infty}$, when $i \to \infty$. If $\gamma > 1$, then

$$\Delta_0^2(\varepsilon) \underset{\varepsilon \to 0}{\sim} \varepsilon^2 \sum_{i=1}^{\infty} \frac{l_i^2}{1+\varepsilon^2 \tilde{\lambda}_i^{2k}},$$

$$\Delta_1^2(\varepsilon, \gamma) \underset{\varepsilon \to 0}{\sim} \varepsilon^{2\gamma} \sum_{i=1}^{\infty} \frac{\tilde{\lambda}_i l_i^2}{1+\varepsilon^2 \tilde{\lambda}_i^{2k}+\varepsilon^{2\gamma} \tilde{\lambda}_i^{2k+1}},$$

$$\Delta_2^2(\varepsilon, \gamma) \underset{\varepsilon \to 0}{\sim} \varepsilon^2 \sum_{i=1}^{\infty} \frac{l_i^2}{1+\varepsilon^2 \tilde{\lambda}_i^{2k}+\varepsilon^{2\gamma} \tilde{\lambda}_i^{2k+1}},$$

$$\Delta^2(\varepsilon, \gamma) \underset{\varepsilon \to 0}{\sim} \sum_{i=1}^{\infty} \frac{(\varepsilon^{2\gamma} \tilde{\lambda}_i + \varepsilon^2) l_i^2}{1+\varepsilon^2 \tilde{\lambda}_i^{2k}+\varepsilon^{2\gamma} \tilde{\lambda}_i^{2k+1}}.$$

This lemma is proved in [1].

Corollary 5.5. Let $\gamma > 1$. Then

$$\Delta_0^2(\varepsilon) \underset{\varepsilon \to 0}{\sim} \varepsilon^2 \sum_{i=1}^{\infty} \frac{l_i^2}{1+\varepsilon^2 (2\pi i)^{4nk}},$$

$$\Delta_1^2(\varepsilon, \gamma) \underset{\varepsilon \to 0}{\sim} \varepsilon^{2\gamma} \sum_{i=1}^{\infty} \frac{(2\pi i)^{2n} l_i^2}{1+\varepsilon^2 (2\pi i)^{4nk}+\varepsilon^{2\gamma} (2\pi i)^{4nk+2n}},$$

$$\Delta_2^2(\varepsilon, \gamma) \underset{\varepsilon \to 0}{\sim} \varepsilon^2 \sum_{i=1}^{\infty} \frac{l_i^2}{1+\varepsilon^2 (2\pi i)^{4nk}+\varepsilon^{2\gamma} (2\pi i)^{4nk+2n}},$$

$$\Delta^2(\varepsilon, \gamma) \underset{\varepsilon \to 0}{\sim} \sum_{i=1}^{\infty} \frac{[\varepsilon^{2\gamma} (2\pi i)^{2n} + \varepsilon^2] l_i^2}{1+\varepsilon^2 (2\pi i)^{4nk}+\varepsilon^{2\gamma} (2\pi i)^{4nk+2n}}.$$

These statements follow from (1.6) and Lemma 5.4.

Corollary 5.6. Denote

$$I_0 = \varepsilon^2 \int_1^{\infty} \frac{l^2(x)}{1+\varepsilon^2 (2\pi x)^{4nk}} dx, \tag{5.1}$$

$$I_1 = \varepsilon^{2\gamma} \int_1^{\infty} \frac{(2\pi x)^{2n} l^2(x) dx}{1+\varepsilon^2 (2\pi x)^{4nk}+\varepsilon^{2\gamma} (2\pi x)^{4nk+2n}}, \tag{5.2}$$

$$I_2 = \varepsilon^2 \int_1^{\infty} \frac{l^2(x) dx}{1+\varepsilon^2 (2\pi x)^{4nk}+\varepsilon^{2\gamma} (2\pi x)^{4nk+2n}}. \tag{5.3}$$

If $\gamma > 1$, then $\Delta_0^2 \underset{\varepsilon \to 0}{\sim} I_0$, $\Delta_j^2 \underset{\varepsilon \to 0}{\sim} I_j$, $j = 1, 2$; $\Delta^2 \underset{\varepsilon \to 0}{\sim} I_1 + I_2$.

Proof. The function $l^2(x)$ is monotonic (see Remark 5.3). Let it be nondecreasing and take for instance $j=1$. For $i \leqq x < i+1$, $i=1, 2, 3, \ldots$ we have

$$\frac{\varepsilon^{2\gamma}(2\pi i)^{2n} l_i^2}{1+\varepsilon^2[2\pi(i+1)^{4nk}+\varepsilon^{2\gamma}[2\pi(i+1)]^{4nk+2n}} \leqq$$

$$\leqq \frac{\varepsilon^{2\gamma}(2\pi x)^{2n} l^2(x)}{1+\varepsilon^2(2\pi x)^{4nk}+\varepsilon^{2\gamma}(2\pi x)^{4nk+2n}} \leqq$$

$$\leqq \frac{\varepsilon^{2\gamma}[2\pi(i+1)]^{2n} l_{i+1}^2}{1+\varepsilon^2(2\pi i)^{4nk}+\varepsilon^{2\gamma}(2\pi i)^{4nk+2n}},$$

$$\sum_{i=1}^{\infty} \frac{\varepsilon^{2\gamma}(2\pi i)^{2n} l_i^2}{1+\varepsilon^2[2\pi(i+1)]^{4nk}+\varepsilon^{2\gamma}[2\pi(i+1)]^{4nk+2n}} \leqq I_1 \leqq$$

$$\leqq \sum_{i=1}^{\infty} \frac{\varepsilon^{2\gamma}[2\pi(i+1)]^{2n} l_{i+1}^2}{1+\varepsilon^2(2\pi i)^{4nk}+\varepsilon^{2\gamma}(2\pi i)^{4nk+2n}}.$$

As $l_{i+1}^2 \underset{i\to\infty}{\sim} l_i^2$ (see (3.5)) and $[2\pi(i+1)]^{2n} l_{i+1}^2 \underset{i\to\infty}{\sim} (2\pi i)^{2n} l_i^2$, according to Lemma 5.1, the case $j=2$, the series on the right-hand side of I_1 is equivalent, when $\varepsilon\to 0$, to

$$\sum_{i=1}^{\infty} \frac{\varepsilon^{2\gamma}(2\pi i)^{2n} l_i^2}{1+\varepsilon^2(2\pi i)^{4nk}+\varepsilon^{2\gamma}(2\pi i)^{4nk+2n}} \underset{\varepsilon\to 0}{\sim} A_1^2.$$

So does the series on the left-hand side of I_1 (see Lemma 5.4). Hence $A_1^2 \underset{\varepsilon\to 0}{\sim} I_1$. The remaining relations can be proved in a similar way.

Next we transform the integrals (5.1)–(5.3) using (4.4).

$$I_0(\varepsilon) = \varepsilon^2 \int_1^{\infty} \frac{l^2(x)dx}{1+\varepsilon^2(2\pi x)^{4nk}} =$$

$$= \varepsilon^2 \int_1^{\infty} \frac{x^{\rho}\mathscr{L}(x)dx}{1+(2\pi\varepsilon^{1/2nk}x)^{4nk}} = \varepsilon^2 \varepsilon_1^{-(\rho+1)} \int_{\varepsilon_1}^{\infty} f_0(y)\mathscr{L}(\varepsilon_1^{-1}y)dy,$$

where $\varepsilon_1 = 2\pi\varepsilon^{1/2nk}$ and $f_0(y) = y^{\rho}(1+y^{4nk})^{-1}$, $y>0$. Let in (5.2) and (5.3) $\gamma = \gamma_0 = 1+\frac{1}{2k}$. Then

$$I_1(\varepsilon) = \varepsilon^2 \varepsilon_1^{-(\rho+1)} \int_{\varepsilon_1}^{\infty} f_1(y)\mathscr{L}(\varepsilon_1^{-1}y)dy,$$

where $f_1(y)=\dfrac{y^{2n+\rho}}{1+y^{4nk}+y^{4nk+2n}}$;

$$I_2(\varepsilon)=\varepsilon^2\varepsilon_1^{-(\rho+1)}\int_{\varepsilon_1}^{\infty} f_2(y)\mathscr{L}(\varepsilon_1^{-1}y)dy,$$

where $f_2(y)=\dfrac{y^{\rho}}{1+y^{4nk}+y^{4nk+2n}}$. Note that the integrals $\int_0^{\infty} f_j(y)dy, j=0, 1, 2$ converge. In fact, from (4.5) and (4.8) we have that $4nk-(\rho+1)>0$, $(4nk+2n)-(2n+\rho+1)>0$, $4nk+2n-(\rho+1)>0$, which, according to [6] (p. 306), is enough for the convergence of the integrals.

Lemma 5.7. Let $\gamma=\gamma_0$. Then for $j=0, 1, 2$,

$$I_j(\varepsilon)\underset{\varepsilon\to 0}{\sim}\tilde{I}_j(\varepsilon)=\varepsilon^2\varepsilon_1^{-(\rho+1)}\int_0^{\infty} f_j(y)\mathscr{L}(\varepsilon_1^{-1}y)dy.$$

Proof. Consider, for example, the case $j=1$,

$$\frac{1}{I_1(\varepsilon)}\left[\varepsilon^2\varepsilon_1^{-(\rho+1)}\int_0^{\varepsilon_1} f_1(y)\mathscr{L}(\varepsilon_1^{-1}y)dy\right]\leqq$$

$$\leqq\varepsilon_1^{2n+\rho}\int_0^{\varepsilon_1}\frac{(\varepsilon_1^{-1}y)^{2n+\rho}\mathscr{L}(\varepsilon_1^{-1}y)dy}{1+y^{4nk}+y^{4nk+2n}}\left[\int_1^2 f_1(y)\mathscr{L}(\varepsilon_1^{-1}y)dy\right]^{-1}\leqq$$

$$\leqq\varepsilon_1^{2n+\rho}\sup_{0<x<1}[x^{2n+\rho}\mathscr{L}(x)]\int_0^{\varepsilon_1}\frac{dy}{1+y^{4nk}+y^{4nk+2n}}\times$$

$$\times\left[\int_1^2 f_1(y)\mathscr{L}(\varepsilon_1^{-1}y)dy\right]^{-1}\leqq\text{const.}\,\varepsilon_1^{2n+\rho}\left[\int_1^2 f_1(y)\mathscr{L}(\varepsilon_1^{-1}y)dy\right]^{-1}$$

$$\underset{\varepsilon\to 0}{\sim}\text{const.}\frac{\varepsilon_1^{2n+\rho}}{\mathscr{L}(\varepsilon_1^{-1})}=\text{const.}\left[\varepsilon_1^{-(2n+\rho)}\mathscr{L}(\varepsilon_1^{-1})\right]^{-1}\underset{\varepsilon\to 0}{\to}0$$

(for the last relations see (3.6) and (3.4)). The remaining cases can be considered in a similar manner.

Lemma 5.8. Let $\gamma=\gamma_0$. Then for $j=0, 1, 2$, it is fulfilled

$$\int_0^{\infty} f_j(y)\mathscr{L}(\varepsilon_1^{-1}y)dy\underset{\varepsilon\to 0}{\sim}\mathscr{L}(\varepsilon_1^{-1})\int_0^{\infty} f_j(y)dy.$$

Proof. We wish to apply (3.7) and (3.8). Then for every $j=0, 1, 2$, we have to find two positive numbers α and β such that

$$\int_1^\infty y^\alpha f_j(y)dy<\infty, \qquad \int_0^1 y^{-\beta} f_j(y)dy<\infty.$$

Let $j=1$ and take $0<\alpha<4nk-1-\rho$, $0<\beta<2n+\rho$. Then

$$\int_1^\infty y^\alpha f_1(y)dy=\int_1^\infty \frac{y^{2n+\rho+\alpha}}{1+y^{4nk}+y^{4nk+2n}}dy<\infty,$$

for $(4nk+2n)-(2n+\rho+\alpha+1)=4nk-\rho-1-\alpha>0$. We have also

$$\int_0^1 y^{-\beta} f_1(y)dy=\int_0^1 \frac{y^{2n+\rho-\beta}}{1+y^{4nk}+y^{4nk+2n}}dy<\infty.$$

If $j=0, 2$ and $\rho>0$, we take $0<\alpha<4nk-1-\rho$, $0<\beta<\rho$; for $\rho<0$ we take $\alpha=-\rho$, $0<\beta<1+\rho$. So we are allowed to use (3.7) and (3.8):

$$\frac{\int_0^\infty f_j(y)\mathscr{L}(\varepsilon_1^{-1}y)dy}{\mathscr{L}(\varepsilon_1^{-1})\int_0^\infty f_j(y)dy}=\frac{\int_0^1+\int_1^\infty}{\mathscr{L}(\varepsilon_1^{-1})\left[\int_0^1+\int_1^\infty\right]}=$$

$$=\frac{\left[\mathscr{L}(\varepsilon_1^{-1})\int_0^1 f_j(y)dy\right]^{-1}\left[\int_0^1+\int_1^\infty\right]}{1+\int_1^\infty f_j(y)dy\left[\int_0^1 f_j(y)dy\right]^{-1}}\underset{\varepsilon\to 0}{\to}1,$$

which proves the Lemma.

Lemma 5.9. Let $0<r<1$. Then

$$\frac{\Delta_0^2(\varepsilon)}{\Delta_0^2(\varepsilon^r)}\underset{\varepsilon\to 0}{\to}0.$$

Proof. Applying Lemma 5.4, Corollary 5.6 and Lemmas 5.7 and 5.8 we get

$$\Delta_0^2(\varepsilon^r)\underset{\varepsilon\to 0}{\sim}I_0(\varepsilon^r)\underset{\varepsilon\to 0}{\sim}\varepsilon^{2r}\varepsilon_1^{-r(\rho+1)}\mathscr{L}(\varepsilon_1^{-r})\int_0^\infty f_0(y)dy.$$

Then

$$\frac{\Delta_0^2(\varepsilon)}{\Delta_0^2(\varepsilon^r)} \underset{\varepsilon\to 0}{\sim} \varepsilon^{2(1-r)}\varepsilon_1^{-(\rho+1)(1-r)}\frac{\mathscr{L}(\varepsilon_1^{-1})}{\mathscr{L}(\varepsilon_1^{-r})} =$$

$$= \text{const. } \varepsilon_1^{(1-r)(4nk-1-\rho)}\frac{\mathscr{L}(\varepsilon_1^{-1})}{\mathscr{L}(\varepsilon_1^{-r})} =$$

$$= \text{const. } \varepsilon_1^{\varkappa}\mathscr{L}(\varepsilon_1^{-1})\left[\varepsilon_1^{-\varkappa}\mathscr{L}(\varepsilon_1^{-r})\right]^{-1} \underset{\varepsilon\to 0}{\to} 0,$$

where $2\varkappa=(1-r)(4nk-1-\rho)>0$.

6. Proof of the Theorem

Let first $\gamma=\gamma_0$. For $j=1, 2$, Lemma 5.4, Corollary 5.6 and Lemmas 5.7, 5.8 give

$$\frac{\Delta_j^2(\varepsilon,\gamma_0)}{\Delta_0^2(\varepsilon)} \underset{\varepsilon\to 0}{\sim} \frac{I_j(\varepsilon)}{I_0(\varepsilon)} \underset{\varepsilon\to 0}{\sim} \frac{\tilde{I}_j(\varepsilon)}{\tilde{I}_0(\varepsilon)} \to c_j,$$

$$c_j = \int_0^\infty f_j(y)dy\left[\int_0^\infty f_0(y)dy\right]^{-1}. \tag{6.1}$$

But $\Delta^2(\varepsilon,\gamma)=\Delta_1^2(\varepsilon,\gamma)+\Delta_2^2(\varepsilon,\gamma)$, then

$$\frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon,\gamma_0)} \underset{\varepsilon\to 0}{\to} C, \qquad C=\frac{1}{c_1+c_2}.$$

Evidently, $f_2(y)<f_0(y)$ and, by the obvious inequality $1+z^2>z$, $z>0$, we get also

$$f_1(y)=\frac{y^{2n+\rho}}{1+y^{4nk}+y^{4nk+2n}}<\frac{y^{2n+\rho}}{y^{2n}+y^{4nk+2n}}=\frac{y^\rho}{1+y^{4nk}}=f_0(y).$$

Hence, $c_j<1$, $C>\frac{1}{2}$.

From

$$f_1(y)+f_2(y)=\frac{y^\rho(y^{2n}+1)}{1+y^{4nk}(y^{2n}+1)}>\frac{y^\rho}{1+y^{4nk}}=f_0(y)$$

it follows that $C<1$, thus (2.2) is proved. Let us turn to (2.1). From (1.1) it is easy to verify that the risk $\Delta^2(\varepsilon,\varepsilon^\gamma)$ increases when γ decreases, so it is sufficient to show that

$$\frac{\Delta_0^2(\varepsilon)}{\Delta^2(\varepsilon,\gamma)} \underset{\varepsilon\to 0}{\to} 0 \qquad \text{for} \quad 1<\gamma<\gamma_0.$$

Denote $\varepsilon_2 = {}^{\gamma/\gamma_0}$, $1 < \gamma < \gamma_0$. Then, using (6.1), we have

$$\mathcal{A}_1^2(\varepsilon, \gamma) = \varepsilon_2^{2\gamma_0} \sum_{i=1}^{\infty} \frac{\lambda_i l_i^2}{1 + \varepsilon^{2(1-\gamma/\gamma_0)} \varepsilon_2^2 \lambda_i^{2k} + \varepsilon_2^{2\gamma_0} \lambda_i^{2k+1}} >$$

$$> \mathcal{A}_1^2(\varepsilon_2, \gamma_0) \underset{\varepsilon \to 0}{\sim} c_1 \mathcal{A}_1^2(\varepsilon_2),$$

$$\frac{\mathcal{A}_0^2(\varepsilon)}{\mathcal{A}^2(\varepsilon, \gamma)} < \frac{\mathcal{A}_0^2(\varepsilon)}{\mathcal{A}_1^2(\varepsilon, \gamma)} < \frac{\mathcal{A}_0^2(\varepsilon)}{\mathcal{A}_1^2(\varepsilon_2, \gamma_0)} \underset{\varepsilon \to 0}{\sim} \frac{\mathcal{A}_0^2(\varepsilon)}{c_1 \mathcal{A}_0^2 c \varepsilon_2)}.$$

According to Lemma 5.9

$$\frac{\mathcal{A}_0^2(\varepsilon)}{\mathcal{A}_0^2(\varepsilon_2)} = \frac{\mathcal{A}_0^2(\varepsilon)}{\mathcal{A}_0^2(\varepsilon^{\gamma/\gamma_0})} \underset{\varepsilon \to 0}{\to} 0,$$

which proves (2.1). In order to complete the proof, we say again that (2.3) is valid in a more general situation, i.e., when L is bounded on Σ.

References

1. *Dodunekova, R.*, On the estimation of a signal passing through a linear filter. Problemi Peredachi Informatsii (to appear).
2. *Ibragimov, I. A., Has'minski, R. Z.*, On the estimation of linear functionals in Gaussian noise. Teorija verojatnostei i primenenija, **32,** *1*, 1987, pp. 24–34.
3. *Galambos, J., Seneta, E.*, Regularly varying sequences. Proc. Amer. Math. Soc., **41,** *1*, 1973, pp. 110–116.
4. *Bojanić, R., Seneta, E.*, A unified theory of regularly varying sequences. Math. Zeits., **134,** 1973, pp. 91–106.
5. *Seneta, E.*, Regularly Varying Functions. Lecture Notes in Mathematics, **508,** Springer Verlag, 1976.
6. *Gradstein, I. S., Rižik, I. M.*, Tables for Integrals, Sums, Series and Products. Fizmatgiz, Moscow, 1962.
7. *Ibragimov, I. A., Has'minski, R. Z.*, On the nonparametric estimation of a value of a linear functional in the Gaussian noise. Teorija verojatnostei i primenenija, **29,** *1*, 1984, pp. 19–32.
8. *Dodunekova, R. D., Has'minski, R. Z.*, Estimation of a value of a linear functional by indirect observations. Problemi Peredachi Informatsii, **23,** *2*, 1987, pp. 54–60.

О задаче минимаксного оценивания линейних функционалов

Р. ДОДУНЕКОВА

(София)

Пусть $s(t)$ — входящий сигнал, неизвестная функция пространства $L_2[0, 1]$. Этот сигнал проходит через канал с аддитивным белым гауссовским шумом $\varepsilon \dot{W}_1$ интенсивностью ε^2. Результат $\dot{X}(t)$ входит в линейный филтр, описываемый линейным дифференциальным уравнениим порядка n с постоянными коэффициентами. Выходит из канала сигнал $\dot{Z}(t)$, являющийся суммой выхода фильтра и другого белого гауссовского шума $\delta \dot{W}_2$, независимого с первым. Пусть известно, что s принадлежит заданному множеству $\Sigma \subset L_2[0, 1]$ и L есть линейный функционал на Σ. Рассмотрим отношение $\mathcal{A}_0^2/\mathcal{A}^2$, где $\mathcal{A}_0^2$ есть линейный минимаксный риск в задаче оценивания $L(s)$ по прямым

наблюдениям (т.е., по наблюдениям $\dot{X}(t)$), а Δ^2 есть линейный минимаксиный риск в задаче оценивания $L(s)$ по косвенным наблюдениям (т.е., по наблюдениям $\dot{Z}(t)$). В работе указан широкий класс функционалов L, для которых при некоторых условиях существует пороговое число в поведении этого риска. До этого порога оценивание по косвенным наблюдениям несоизмеримо хуже, чем оценивание по прямым наблюдениям. После этого порога Δ_1^2 и Δ^2 практически одни и те же, т.е., линейный фильтр не приводит к практическим информационным потерям.

R. Dodunekova
Department of Probability and Statistics
University of Sofia
P.O. Box 373 1090 Sofia
Bulgaria

Проблемы управления и теории информации, том **18**, *номер 4 (1989)*

УПРАВЛЕНИЕ САМОЛЕТОМ НА ПОСАДКЕ ПРИ СДВИГЕ ВЕТРА

Н. Д. Воткин, В. М. Кейн, В. С. Пацко, В. Л. Турова

(*Свердловск, Ленинград*)

Традиционные способы управления самолетом на посадке плохо работают в условиях резкого изменения скорости ветра. В связи с этим в последнее время исследуются [1–6] новые способы управления на посадке. Статья посвящена минимаксному способу, основанному на методах теории дифференциальных игр [7, 8]. Применение методов теории дифференциальных игр к задаче посадки изучалось ранее в работах [1–3, 9–13].

1. Введение

В настоящее время имеется значительное число работ [3–6, 9, 14, 15], в которых анализируется поведение самолета на взлете и посадке при резком изменении скорости ветра (сдвиг ветра). Изучаются физические условия возникновения сдвига ветра, математические модели этого явления, способы управления самолетом.

В статье исследуется процесс посадки среднего транспортного самолета в условиях ветрового возмущения. Рассматривается движение на предпоследней прямой до момента пролета торца взлетно-посадочной полосы (ВПП). Относительно ветра предполагаются известными ориентировочно лишь пределы возможных отклонений его скорости от некоторого номинального значения и само это значение. Какие-либо сведения о пространственном расположении области сдвига ветра, равно как и информация о распределении скорости ветра в ней, считаются отсутствующими. Таким образом, естественно возникает задача о нахождении минимаксного способа управления самолетом по принципу обратной связи, рассчитанного на любое изменение скорости ветра в оговоренных пределах.

В работе минимаксное решение получено в рамках вспомогательных линейных задач на основе методов теории антагонистических дифференциальных игр [7, 8]. Затем оно используется при моделировании в полной нелинейной

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

системе. Результаты моделирования относятся к случаю, когда сдвиг ветра обусловлен прохождением самолета через зону микровзрыва. Микровзрыв образуется за счет нисходящего потока воздуха, который ударяется о поверхность земли и растекается затем с образованием вихря. Математическая модель микровзрыва взята из работы [14].

2. Нелинейная система движения самолета на посадке

Движение самолета на посадке описывается нелинейной системой дифференциальных уравнений 12-го порядка, где вектор состояния включает три координаты x, y, z центра масс в системе, связанной с поверхностью ВПП (фиг. 1), углы тангажа ϑ, рыскания ψ и крена γ, а также соответствующие линейные и угловые скорости. Конкретный вид уравнений приводится, например, в [16, 17].

Управляющими воздействиями являются отклонения рулей высоты $\delta_в$, направления $\delta_н$, элеронов $\delta_э$ и изменение тяги двигателей P. К уравнениям самолета добавляются уравнения динамики рулевых приводов и двигателей. Указанные выше величины входят теперь в расширенный вектор состояния объекта, а управляющими становятся задающие (командные) $\delta_{вз}$, $\delta_{нз}$, $\delta_{эз}$, $\delta_{рз}$. Каждое из них ограничено снизу и сверху.

Полученную в итоге полную систему дифференциальных уравнений в векторной форме запишем в виде

$$\dot{\xi} = f(\xi, \delta_з, W). \tag{2.1}$$

Здесь $\delta_з = (\delta_{вз}, \delta_{рз}, \delta_{нз}, \delta_{эз})^T$ — векторный параметр управления, $W = (W_x, W_y, W_z)^T$ — векторный параметр помехи, состоящий из компонент скорости ветра по осям x, y, z.

3. Минимаксный способ управления

Номинальное движение самолета на этапе посадки до момента пролета торца ВПП представляет собой равномерное движение (без вращения) по прямолинейной глиссаде снижения.

Задача управления состоит в том, чтобы реальное движение, проходящее в условиях ветрового возмущения, не слишком сильно отличалось от номинального. Желательно также, чтобы закон управления не требовал для своей реализации какой-либо точной и детализированной информации о ветровой помехе.

Мы считаем, что приближенно могут быть известны лишь «грубые» характеристики ветрового возмущения, а именно, пределы отклонений компонент W_x, W_y, W_z скорости ветра от некоторых номинальных значений W_{x0}, W_{y0}, W_{z0}, которые предполагаем заданными. Обратимся к минимаксной постановке задачи управления. Поскольку движение самолета описывается дифференциальным уравнениями, мы приходим к дифференциальной игре.

В настоящее время разработаны эффективные программы [9, 12, 18–20], позволяющие находить на ЭВМ оптимальные законы управления (стратегии) для линейных дифференциальных игр с фиксированным моментом окончания и выпуклой функцией платы, зависящей от двух координат фазового вектора. Система (2.1) не является линейной. Однако мы можем линеаризовать ее относительно номинального движения, поставить вспомогательные линейные задачи и использовать затем их решение в исходной нелинейной системе.

Итак, задав номинальные составляющие W_{x0}, W_{y0}, W_{z0} скорости ветра, угол наклона глиссады и номинал модуля воздушной скорости, вычисляем соответствующие номинальному движению значения фазовых переменных системы (2.1). Линеаризуя систему (2.1) относительно номинального движения, приходим к линейной управляемой системе, распадающейся на две подсистемы вертикального (продольного) и бокового движений [16]. В подсистеме вертикального движения фазовыми переменными являются отклонения Δx, Δy и величины, их определяющие. В подсистему бокового движения входит отклонение Δz и величины, его определяющие.

Для каждой из подсистем поставим вспомогательную дифференциальную игру с фиксированным моментом окончания T, геометрическими ограничениями на управляющие воздействия и воздействия помехи, выпуклой функцией платы, зависящей от двух координат фазового вектора в момент T. В вертикальном движении за такие координаты возьмем Δy и $\Delta \dot{y}$, в боковом — Δz и $\Delta \dot{z}$. Первый игрок, распоряжающийся управляющими воздействиями, минимизирует значения функции платы. Второй, в ведении которого воздействия помехи, максимизирует функцию платы. В рамках вспомогательных задач можно не придавать моменту T какого-либо физического смысла.

Воздействия $\delta_{рз}$ и $\delta_{нз}$ в системе (1.1) на этапе посадки имеют специфическое назначение: стабилизация воздушной скорости и поддержание вблизи нуля величины угла скольжения. Не совсем естественно находить способы формирования $\delta_{рз}$ и $\delta_{нз}$ из решения описанных вспомогательных дифференциальных игр. Поэтому условимся при формулировке вспомогательных задач считать тягу постоянной и совпадающей с номинальной (т.е. $\Delta\delta_{рз} \equiv 0$), а изменение $\Delta\delta_{нз}$ подчиним линейному дифференциальному уравнению, описывающему прин-

ятый в настоящее время закон управления по рулю направления, опустив при этом ограничение на $\Delta\delta_{нз}$. Таким образом, в подсистеме вертикального движения остается одно управляющее воздействие $\Delta\delta_{вз}$, в подсистеме бокового движения — воздействие $\Delta\delta_{эз}$.

Чтобы учесть свойства, характеризующие инерционность изменения скорости ветра вдоль движения, подчиним переменные ΔW_x, ΔW_y, ΔW_z дополнительным линейным дифференциальным уравнениям, например,

$$\Delta\dot{W}_x = k_1(\Delta F_x - \Delta W_x)$$
$$\Delta\dot{F}_x = k_2(w_x - \Delta F_x). \qquad (3.1)$$

Здесь новой независимой переменной является w_x, константы k_1 и k_2 регулируют инерционность изменения ΔW_x. Аналогичные уравнения введем для ΔW_y и ΔW_z. Переменные w_x, w_y, w_z трактуем как воздействия помехи. Уравнения, определяющие ΔW_x и ΔW_y, отнесем к подсистеме вертикального движения, уравнения для ΔW_z — к подсистеме бокового движения.

Решая вспомогательные задачи на ЭВМ, находим оптимальные для них законы управления по $\Delta\delta_{вз}$ и $\Delta\delta_{эз}$. Эти законы реализуются при помощи наборов $K_{вз}$ и $K_{эз}$ линий переключения [9, 13, 18, 19]. Каждый набор задан на сетке моментов τ_i обратного времени, отсчитываемого от момента T.

Наборы $K_{вз}$, $K_{эз}$ линий переключения для $\Delta\delta_{вз}$, $\Delta\delta_{эз}$ и определяют искомый способ управления по $\delta_{вз}$, $\delta_{эз}$ в исходной системе (2.1). При работе с ним делаем текущий прогноз времени, оставшегося до момента пролета торца ВПП. Сооответственно прогнозу используем вполне определенные линии переключения для выбора $\delta_{вз}$ и $\delta_{эз}$. При моделировании движений системы (2.1) воздействия $\delta_{рз}$ и $\delta_{нз}$ условимся вырабатывать на основе принятых в настоящее время законов.

Таким образом, говоря о минимаксном способе управления нелинейной системой (2.1), имеем в виду способ формирования управляющих воздействий $\delta_{вз}$ и $\delta_{эз}$, полученный из решения вспомогательных линейных дифференциальных игр. Воздействия $\delta_{рз}$ и $\delta_{нз}$ формируются традиционными методами.

4. Вспомогательные линейные дифференциальные игры

Линейная система дифференциальных уравнений вертикального движения имеет вид

$$\dot{\mathbf{x}} = A_*\mathbf{x} + B_*u + C_*v, \qquad \mathbf{x} \in R^{11}, \tag{4.1}$$

$$\mathbf{A}_* = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.050 & 0 & -0.097 & -2.642 & 0 & 0.063 & 0.050 & 0 & 0.097 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.241 & 0 & -0.639 & 45.278 & 0 & 1.448 & -0.241 & 0 & 0.638 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.007 & -0.501 & -0.526 & -0.383 & 0 & 0 & -0.007 & 0 \\
 & & & & & & -4 & 0 & 0 & 0 & 0 \\
 & & & & & & 0 & -0.5 & 0.50 & 0 & 0 \\
 & & & 0 & & & 0 & 0 & -3 & 0 & 0 \\
 & & & & & & 0 & 0 & 0 & -0.5 & 0.5 \\
 & & & & & & 0 & 0 & 0 & 0 & -3
\end{pmatrix}$$

$$B_* = (0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0)^T, \qquad C_* = \begin{pmatrix} 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0 \end{pmatrix}^T$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_{11})^T, \qquad u = \Delta\delta_{вз}, \qquad v = (w_x, w_y)^T.$$

Здесь $x_1 = \Delta x$, $x_3 = \Delta y$ — отклонения от номинальных значений по осям x, y; $x_5 = \Delta\vartheta$ — отклонение угла тангажа. Координата x_7 имеет смысл отклонения руля высоты от балансировочного положения. При помощи переменных x_8, x_9 описывается изменение величины ΔW_x. Соответствующие уравнения совпадают с (3.1), при этом $x_8 = \Delta W_x$. Переменные x_{10}, x_{11} служат для описания величины ΔW_y ($x_{10} = \Delta W_y$). Управляющее воздействие первого игрока — заданное отклонение $\Delta\delta_{вз}$ руля высоты. Параметры w_x, w_y служат для формирования ветровой помехи и принадлежат второму игроку. Ограничения:

$$|\Delta\delta_{вз}| \leqq 10^\circ \pi/180, \qquad |w_x| \leqq 10 \text{ м/с}, \qquad |w_y| \leqq 5 \text{ м/с}.$$

Введем функцию φ_*, зависящую от координат $x_3 = \Delta y$ и $x_4 = \Delta\dot{y}$. Для этого рассмотрим на плоскости x_3, x_4 симметричный относительно нуля выпуклый шестиугольник с вершинами $(-3, 0), (-3, 1), (0, 1), (3, 0), (3, -1), (0, -1)$. Положим

$$\varphi_*(x_3, x_4) = \min\{c \geqq 0 : (x_3, x_4)^T \in cM_*\}.$$

Рассмотрим антагонистическую дифференциальную игру с динамикой (4.1), фиксированным моментом окончания T и платой φ_*. Первый игрок минимизирует значения платы φ_* в момент T, второй — максимизирует. Множество M_* можно трактовать как допуск на отклонения $x_3 = \Delta y$, $x_4 = \Delta\dot{y}$ в момент T. Функция φ_* показывает отклонение от допуска. Оптимальная стратегия первого игрока в игре (4.1) будет использована для задания $\delta_{вз}$ в системе (2.1).

Линейная система бокового движения имеет вид

$$\dot{\mathbf{x}} = A^*\mathbf{x} + B^*u + C^*v, \qquad \mathbf{x} \in R^{11}, \tag{4.2}$$

$$\mathbf{A}^* = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.077 & -5.555 & 0 & 9.272 & 0 & -1.485 & 0 & 0.077 & 0 & 0 \\
0 & 0 & 0 & 1.001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.013 & -0.933 & -0.259 & -0.088 & -0.030 & -0.246 & -0.046 & 0.012 & 0 & 0 \\
0 & 0 & 0 & -0.051 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.033 & -2.386 & -0.953 & -0.226 & -1.459 & -0.233 & -0.689 & 0.033 & 0 & 0 \\
 & & & & & & -4 & 0 & 0 & 0 & 4 \\
 & & & & & & 0 & -4 & 0 & 0 & 0 \\
 & & & 0 & & & 0 & 0 & -0.5 & 0.5 & 0 \\
 & & & & & & 0 & 0 & 0 & -3 & 0 \\
0 & -0.058 & -4.202 & -0.365 & -0.397 & -0.136 & -1.105 & -0.207 & 0.058 & 0 & -0.4
\end{pmatrix}$$

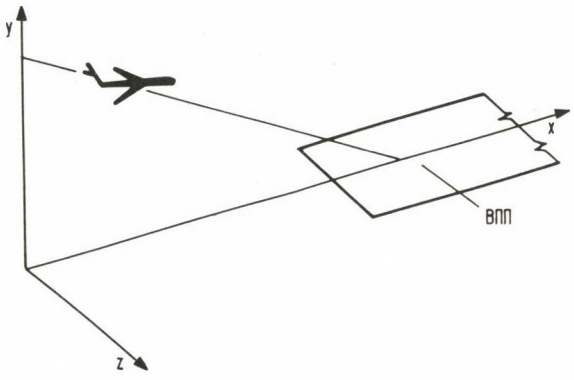
$$B^* = (0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0)^T, \qquad C^* = (0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0)^T,$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_{11})^T, \qquad u = \Delta\delta_{эз}, \qquad v = w_z.$$

Здесь $x_1 = \Delta z$ — отклонение от оси z, $x_3 = \Delta\psi$ и $x_5 = \Delta\gamma$ — отклонения углов рыскания и крена. Координаты x_7, x_8 имеют смысл отклонения руля направления и элеронов ($\Delta\delta_н$, $\Delta\delta_э$), $x_9 = \Delta\delta_{нз}$. При помощи переменных x_{10}, x_{11} описывается изменение величины ΔW_z, при этом $x_{10} = \Delta W_z$. Управляющее воздействие первого игрока — заданное отклонение $\Delta\delta_{эз}$ элеронов. Параметр w_z служит для формирования помехи и принадлежит второму игроку. Ограничения:

$$|\Delta\delta_{эз}| \leqq 10^\circ \frac{\pi}{180}, \qquad |w_z| \leqq 10 \text{ м/с}.$$

Введем функцию φ^*, зависящую от координат $x_1 = \Delta z$ и $x_2 = \Delta\dot{z}$. Для этого рассмотрим на плоскости x_1, x_2 симметричный относительно нуля выпуклый



Фиг. 1. Система координат

шестиугольник M^* с вершинами $(-6, 0)$, $(-6, 1.5)$, $(0, 1.5)$, $(6, 0)$, $(6, -1.5)$, $(0, -1.5)$. Положим

$$\varphi^*(x_1, x_2) = \min \{c \geqq 0 : (x_1, x_2)^T \in cM^*\}.$$

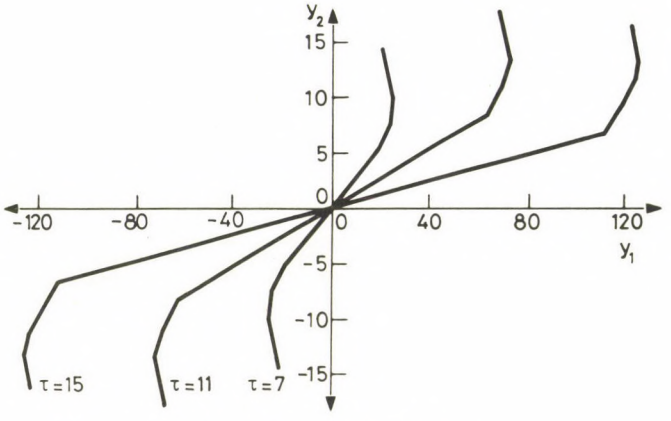
Рассмотрим антагонистическую дифференциальную игру с динамикой (4.2), фиксированным моментом окончания T и платой φ^*. Первый игрок минимизирует значения платы φ^* в момент T, второй максимизирует. Множество M^* — допуск на отклонения $x_1 = \Delta z$, $x_2 = \Delta \dot{z}$ в момент T. Функция φ^* показывает отклонение от допуска. Оптимальная стратегия первого игрока в игре (4.2) будет использована для задания $\delta_{эз}$ в системе (2.1).

В системах (4.1), (4.2) линейные величины измеряются в метрах, углы — в радианах, время — в секундах. Для численного задания систем (4.1), (4.2) были взяты следующие исходные данные: угол наклона глиссады $\Theta = -2.66$, номинал модуля воздушной скорости $\tilde{V}_0 = 72.2$ м/с, номиналы составляющих скорости ветра $W_{x0} = -5$ м/с, $W_{y0} = W_{z0} = 0$.

5. Оптимальная стратегия первого игрока в линейной дифференциальной игре

Специфические особенности дифференциальных игр (4.1), (4.2) состоят в следующем. Каждая из них есть игра с фиксированным моментом окончания и выпуклой функцией платы, зависящей от двух координат фазового вектора. Кроме того, управляющее воздействие первого игрока является скалярным. Эти особенности позволяют сравнительно просто найти оптимальную стратегию первого игрока. Она задается при помощи поверхности переключения в пространстве t, y_1, y_2 эквивалентной [7, 8] игры второго игрока. Связь между

векторами $\mathbf{y}=(y_1, y_2)^T$ и $\mathbf{x}$ описывается формулой $\mathbf{y}(t)=X_*(T,t)\mathbf{x}(t)$ ($\mathbf{y}(t)=X^*(T,t)\mathbf{x}(t)$), где $X_*(T,t)$ ($X^*(T,t)$) — матрица из 3 и 4-ой (1 и 2-ой) строк фундаментальной матрицы Коши однородной части системы (4.1), ((4.2)). По одну сторону от поверхности переключения оптимальное управляющее воздействие принимает экстремальное значение одного знака, по другую сторону — противоположного. Математическое обоснование оптимальности способа управления при помощи поверхности переключения и анализ его устойчивости приведены в работах [18, 19]. Соответствующие численные процедуры разъясняются в [9].



Фиг. 2. Линии переключения

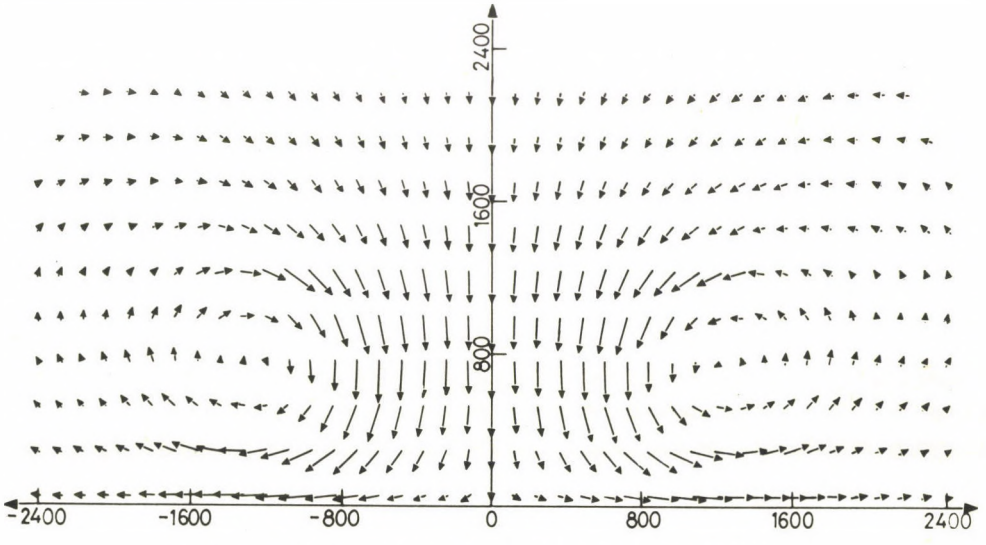
При построении на ЭВМ поверхность переключения реализуется в виде набора сечений на заданной сетке моментов времени. Эти сечения называем линиями переключения. На фиг. 2 показаны линии переключения $\Pi_*(\tau)$ в игре (4.1), просчитанные для моментов обратного времени $\tau=7$, 11, 15. Пусть в момент t_i состояние системы (4.1) есть $\mathbf{x}(t_i)$. Если точка $\mathbf{y}(t_i)=X_*(T, t_i)\mathbf{x}(t_i)$ расположена относительно линии переключения $\Pi_*(\tau_i)$, соответствующей моменту $\tau_i=T-t_i$, в направлении вектора $D_*(t_i)=X_*(T, t_i)B_*$, то на очередном шаге дискретной схемы управления полагаем $\Delta\delta_{вз}=-10$. При противоположном расположении точки $\mathbf{y}(t_i)$ относительно линии $\Pi_*(\tau_i)$ примем $\Delta\delta_{вз}=+10$. Аналогично при помощи линий переключения $\Pi^*(\tau)$ и вектора $D^*(t)=X^*(T,t)B^*$ производится выбор оптимального управляющего воздействия $\Delta\delta_э$, в игре (4.2).

6. Модель микровзрыва

При моделировании движений системы (2.1) будем считать, что ветровое возмущение вызвано попаданием самолета в зону микровзрыва. В настоящее время имеется обширная литература, посвященная микровзрыву. Используемая нами модель взята из [14]. Кратко опишем ее.

Микровзрыв идеализируется в виде трехмерного осесимметричного вихревого поля, в котором выделяется тороидальная область («ядро»), где скорость ветра, начиная от нулевой в центре, линейно увеличивается по радиусу до границы ядра. Вне ядра вихревое поле задается функцией потока. Дифференцирование функции потока дает радиальную и вертикальную составляющие скорости ветра. Первая из них раскладывается затем на две компоненты: параллельно и перпендикулярно оси ВПП. Микровзрыв задается тремя параметрами: $\mathscr{V}$ — скорость ветра в центральной части, $\mathscr{H}$ — высота центральной части, $\mathscr{R}$ — радиус вихря. Радиус ядра вихря полагается равным 0.8 $\mathscr{H}$. Для конкретизации расположения микровзрыва относительно глиссады снижения следует выбрать координаты его центра в горизонтальной плоскости.

Приводимые ниже результаты моделирования относятся к микровзрыву с параметрами $\mathscr{V} = 6$ м/с, $\mathscr{H} = 700$ м, $\mathscr{R} = 1200$ м. Просчитанное на ЭВМ поле скорости ветра в вертикальной плоскости, проходящей через центр микровзрыва, показано на фиг. 3.



Фиг. 3. Поле скорости ветра в вертикалыном сечении микровзрыва

7. Результаты моделирования

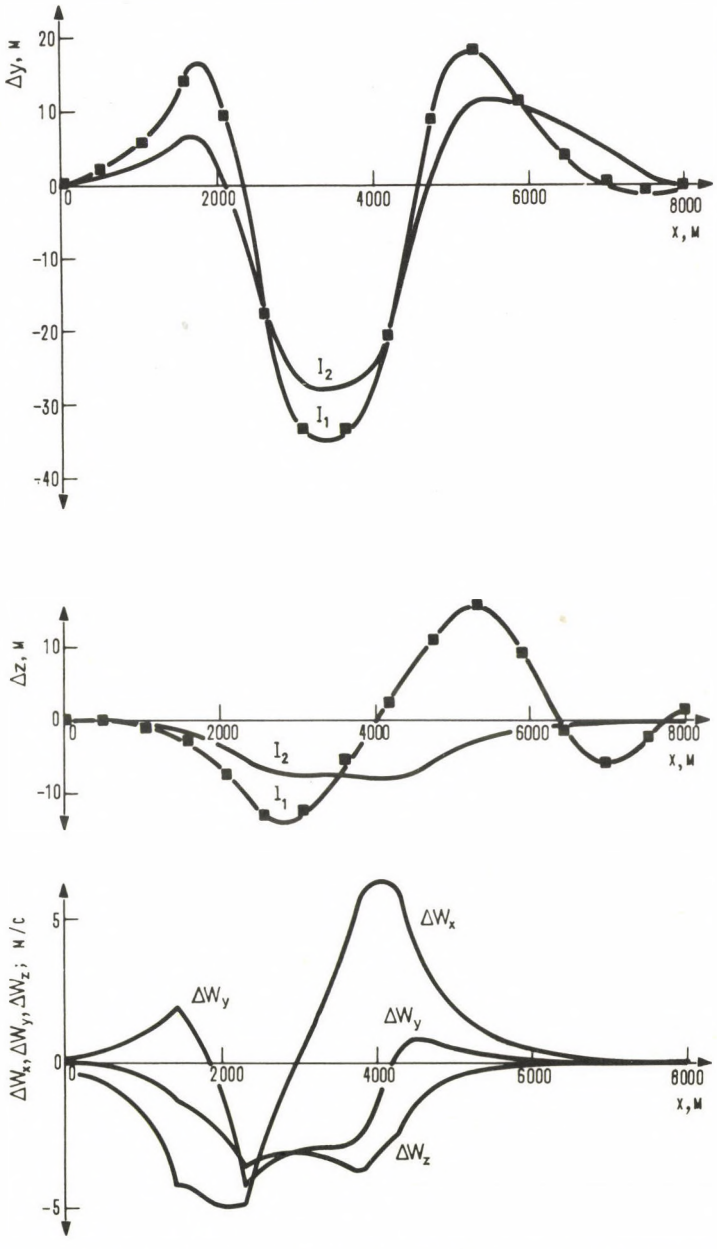
Пусть в момент t_* начальное положение системы (2.1) по оси x находится на расстоянии 8000 м от торца ВПП и значения всех фазовых координат соответствуют номинальному движению по глиссаде.

Рассмотрим два способа управления. Способ I_1 использует алгоритмы формирования $\delta_{вз}$, $\delta_{рз}$, $\delta_{нз}$ и $\delta_{эз}$, принятые в настоящее время. Второй способ — минимаксный. Обозначим его I_2. В способе I_2 воздействия $\delta_{вз}$ и $\delta_{эз}$ формируются при помощи линий переключения, взятых из решения вспомогательных дифференциальных игр (4.1) и (4.2), воздействия $\delta_{рз}$, $\delta_{нз}$ — при помощи принятых в настоящее время алгоритмов. Во вспомогательных задачах положим $T=15$ с.

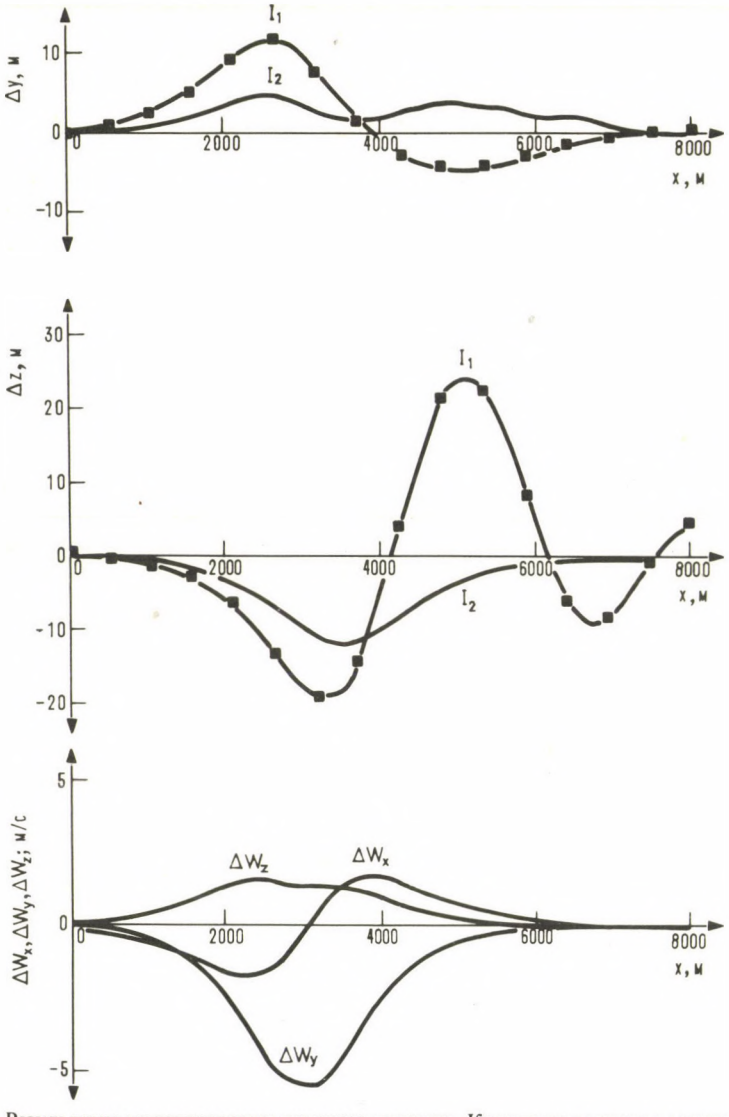
Обозначим через E сетку моментов обратного времени τ_i из промежутка $[0, T]=[0, 15]$, на которой построены линии переключения. Линии переключения применяем в способе I_2 следующим образом. Пусть $d(t)$ — расстояние в момент $t \geqq t_*$ по оси x до торца ВПП, V_{x0} — скорость номинального движения по оси x. Тогда $s(t)=d(t)/V_{x0}$ — прогноз времени до пролета торца ВПП. Пока $s(t) \geqq T=15$ с, для выбора $\delta_{вз}$ (соответственно $\delta_{эз}$) используем одну и ту же линию переключения, отвечающую $\tau = T$. Если $s(t)<T$, то берем линию, соответствующую моменту τ_i из E, ближайшему к $s(t)$. Таким образом, мы управляем сравнительно «грубо» по $\delta_{вз}$ и $\delta_{эз}$, пока расстояние $d(t)$ до торца ВПП больше $V_{x0}T \approx 1000$ м и более качественно при $d(t)<V_{x0}T$. Компоненты W_x, W_y, W_z скорости ветра перед подачей в систему (2.1) просчитываются в виде $W_x=\Delta W_x+W_{x0}$, $W_y=\Delta W_y+W_{y0}$, $W_z=\Delta W_z+W_{z0}$, где ΔW_x, ΔW_y, ΔW_z берутся с модели микровзрыва, $W_{x0}=-5$ м/с, $W_{y0}=W_{z0}=0$. Рассмотрим два варианта расположения центра микровзрыва в горизонтальной плоскости: 1) смещение DX по оси x от начального положения самолета равно 3000 м (5000 м от торца ВПП), смещение DZ по оси z перпендикулярно оси ВПП составляет 500 м, 2) $DX=3000$ м, $DZ=1500$ м.

На фиг. 4, 5 приведены графики изменения вертикального Δy и бокового Δz отклонений системы (2.1) от номинала вдоль движения для способов I_1, I_2, а также реализации отклонений ΔW_x, ΔW_y, ΔW_z скорости ветра. Последние отвечают способу I_2, для способа I_1 они практически те же самые. По горизонтали откладывается расстояние, пройденное по оси x. Фиг. 4 (5) соответствуют первому (второму) варианту расположения центра микровзрыва. Шаг выбора управления и помехи при моделировании был равен 0.1 с.

Видно, что результаты для минимаксного способа I_2 лучше, чем для традиционного I_1.



Фиг. 4. Результаты моделирования процесса посадки. Координаты центра микровзрыва: $DX = 3000$ м, $DZ = 500$ м



Фиг. 5. Результаты моделирования процесса посадки. Координаты центра микровзрыва: $DX = 3000$ м, $DZ = 1500$ м

8. Заключительное примечание

В заключение еще раз подчеркнем, что для расчета минимаксного способа управления не требуется каких-либо точных сведений о пространственном расположении зоны экстремальных ветровых возмущений и тем более сведений

о поле скорости ветра в такой зоне. Достаточно задать примерный размах колебаний скорости ветра относительно номинала и номинальное значение. В этом состоит принципиальное отличие подхода, основанного на теории дифференциальных игр, от методов, где указанная информация является существенной [4, 6].

Литература

1. *Кейн В. М., Париков А. Н., Смуров М. Ю.* Об одном способе оптимального управления по методу экстремального прицеливания. Прикл. мат. и мех., **44,** *3* (1980), с. 434–440.
2. *Титовский И. Н.* Игровой подход к задаче синтеза управления самолетом при заходе на посадку. Ученые записки ЦАГИ, **XII,** *1* (1981), с. 85–92.
3. *Кейн В. М.* Оптимизация систем управления по минимаксному критерию. М.: Наука, 1985, 248 с.
4. *Miele, A., Wang, T., Melvin, W.,* Optimal take-off trajectories in the presence of windshear, J. Opt. Theory and Appl., **49** (1986), No. 1, pp. 1–45.
5. *Psiaki, M. L., Stengel, R. F.,* Optimal flight paths through microburst wind profiles, J. of Aircraft, **23** (1986), No. *8*, pp. 625–635.
6. *Miele, A., Wang, T., Tzeng, C. Y., Melvin, W. W.,* Optimal abort landing trajectories in the presence of windshear, J. Opt. Theory and Appl., **55** (1987), No. *2*, pp. 165–202.
7. *Красовский Н. Н., Субботин А. Н.* Позиционные дифференциальные игры. М.: Наука, 1974, 456 с.
8. *Субботин А. И., Ченцов А. Г.* Оптимизация гарантии в задачах управления. М.: Наука, 1981, 288 с.
9. *Боткин Н. Д., Кейн В. М., Пацко В. С.* Модельная задача об управлении боковым движением самолета на посадке. Прикл. мат. и мех., **48,** *4* (1984), с. 560–567.
10. *Корнеев В. А., Меликян А. А., Титовский И. Н.* Стабилизация глиссады самолета при ветровых возмущениях в минимаксной постановке. Изв. АН СССР. Техн. кибернет., **3** (1985), с. 132–139.
11. *Корнеев В. А.* Квазиоптимальная коррекция движения самолёта на посадке при ветровых возмущениях. Изв. АН СССР. Техн. кибернет., **3** (1987), с. 180–187.
12. *Зарх М. А., Пацко В. С.* Стратегия второго игрока в линейной дифференциальной игре. Прикл. мат. и мех., **51,** *2* (1987), с. 193–200.
13. *Боткин Н. Д., Кейн В. М., Пацко В. С.* Применение методов теории дифференциальных игр к задаче управления самолетом на посадке. В кн.: Позиционное управление с гарантированным результатом. Свердловск: ИММ УрО АН СССР, 1988, с. 33–44.
14. *Ivan, M.,* A ring-vortex downburst model for real time flight simulation of severe wind shears, AIAA Flight simulation technologies conf., 1985, July 22–24. St. Louis, Miss., 1985, pp. 57–61.
15. *Дитенбергер М. А., Хейнс П. А., Луэрс Дж. К.* Реконструкция условий авиакатастрофы в Новом Орлеане. Аэрокосмическая техника, **5** (1986), с. 3–15.
16. Системы цифрового управления самолетом. Под ред. А. Д. Александрова, С. М. Федорова. М.: Машиностроение, 1983, 223 с.
16. *Остославский И. В., Стражева И. В.* Динамика полета. Траектории летательных аппаратов. М.: Машиностроение, 1969, 499 с.
18. *Боткин Н. Д., Пацко В. С.* Универсальная стратегия в дифференциальной игре с фиксированным моментом окончания. Пробл. управления и теории информ., **11,** *6* (1982), с. 419–432.
19. *Боткин Н. Д., Пацко В. С.* Позиционное управление в линейной дифференциальной игре. Изв. АН СССР. Техн. кибернетика, **4** (1983), с. 78–85.
20. Алгоритмы и программы решения линейных дифференциальных игр. Свердловск, 1984, 295 с. (Материалы по мат. обеспечению ЭВМ/АН СССР. УНЦ. ИММ).

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda Vállalat, 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

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: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

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. TOM 18 · NUMBER 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

1989

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 PLC Headington Hill Hall, Oxford OX3 OBW, England

or

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

1989 Subscription Rate DM 535,— 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. 18(5), pp. 277–287 (1989)

STOCHASTIC AIMING IN DETERMINED POSITIONAL CONTROL

D. A. SERKOV

(*Sverdlovsk*)

(Received August 24, 1988)

This paper provides a direct proof of equality of the form of stochastic maximin appropriate for possible further applications to the value function of the differential game when the system is described by an ordinary differential equation, nonlinear with respect to control and in the case when the terminal quality index is a convex function. The proof is based on the idea to involve the conjugate variables for the system and does not lean on neither the results of the theory of partial differential equations nor the results of dynamic programming theory.

Introduction

The paper is concerned with the control problem complicated by uncertain disturbance. We use the conception of the control which was provided and is now under development in the works made in Sverdlovsk [1–5, 9, 11]. The paper is also related to other investigations in the theory of differential games [6–8]. On the other hand, it essentially uses the technique of the maximum principle of Pontryagin [10] in the form adapted for minimax program control by stochastic systems.

In [9] the value function of determined differential games was represented in the form of stochastic program maximin, and the cases when the stochastic maximin contains the mathematical expectation of values of quality indices seems to be most interesting from the point of view of possible further applications.

This paper provides a direct proof of equality of the form of stochastic maximin to the value function of the differential game when the system is described by an ordinary differential equation, nonlinear with respect to control and in the case when the terminal quality index is a convex function. The proof is based on the idea [9] to involve the conjugate variables [10] for the system, but it does not lean on neither the results of the theory of partial differential equations nor the results of dynamic programming theory.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

All the reasonings will be carried out in the case of the linear system and smooth terminal function paying no attention to minor details. Some generalizations will be pointed out in conclusion.

Statement of the problem

Consider the differential game [11] for the control system described by the equation

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu[t] + Cv[t], \quad t \in [t_*, \vartheta], \\ x(t_*) = x_* \in R^n, \end{cases} \tag{1}$$

here, A, B, C — constant matrices; $u[\cdot], v[\cdot]$ — measurable control realizations (of the first and second players) satisfying for almost all $t \in [t_*, \vartheta]$ inclusions $u[t] \in \mathcal{P} \subset R^k$, $v[t] \in \mathcal{Q} \subset R^m$; $\mathcal{P}, \mathcal{Q}$ — convex compacts. The sets of such realizations of controls of the players on the interval $[t_*, \vartheta]$ will be denoted by $\mathcal{U}_{[t_*,\vartheta]}$ and $\mathcal{V}_{[t_*,\vartheta]}$, respectively. Let the quality index γ of the form

$$\gamma = \sigma(x(\vartheta)), \; x(\cdot) = x(\cdot, t_*, x_*, u[\cdot], v[\cdot]) \tag{2}$$

where $\sigma(\cdot) \in C^\infty(R^n)$ is strictly convex and $x(\cdot)$ is the solution of (1). Denote by $\varrho(\cdot)$ the value function [11] of the differential game. Let $\Delta = \{t_* = \tau_1, t^* = \tau_2, \ldots, \tau_{\overline{n}+1} = \vartheta\}$ be some partition of $[t_*, \vartheta]$, $\Delta' = \Delta \setminus \{t_*\}$ — the same for $[t^*, \vartheta]$. Denote by $P_u(\Delta)$, $P_v(\Delta)$ the sets of stochastic non-anticipatory programs of the first and the second players, respectively, produced by the partition Δ [11] i.e., $u(\cdot) \in P_u(\Delta)$ if and only if $u(\cdot) : [t_*, \vartheta] \times \Omega \to \mathcal{P}$ is a measurable function $\Omega = \Omega(\Delta) = \{\omega = (\omega_1, \ldots, \omega_{\overline{n}}) \in [0,1]^{\overline{n}}\}$ that does not depend on $\omega_i, \ldots, \omega_{\overline{n}}$ if $(\tau, \omega) \in [\tau_{i-1}, \tau_i) \times \Omega$ for any $i = 2, \ldots, \overline{n}$). Subsets of the elements from $P_u(\Delta)$ and $P_v(\Delta)$ that does not depend on ω_1 will be denoted by $P'_u(\Delta)$ and $P'_v(\Delta)$ (so, elements from $P'_u(\Delta)$ and $P_u(\Delta')$ differ on the sets of definition: for ones it is $[t_*, \vartheta] \times \Omega'$, for others — $[t^*, \vartheta] \times \Omega'$, $\Omega' = \Omega'(\Delta') = \{\omega' = (\omega_2, \ldots, \omega_{\overline{n}}) \in [0,1)^{\overline{n}-1}\}$). The stochastic maximin $\varrho_*(\cdot)$ [9, 11] for the differential game is defined as follows:

$$\varrho_*(t_*, x_*) = \lim_{d(\Delta)\to 0} \sup_{\Delta} \varrho_{*\Delta}(t_*, x_*), \quad d(\Delta) = \max\{\tau_{i+1} - \tau_i : i = 1, \ldots, \overline{n}\},$$

$$\varrho_{*\Delta}(t_*, x_*) = \sup_{v(\cdot)\in P_v(\Delta)} \inf_{u(\cdot)\in P_u(\Delta)} \int_{\Omega} \sigma(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega)))d\omega. \tag{3}$$

It is known (see, e.g. [11]) that for this differential game, necessary and sufficient conditions for $\varrho_*(\cdot)$ to be equal to $\varrho(\cdot)$ are: $\varrho_*(\cdot)$ is a u-stable and v-stable function and $\varrho_*(\cdot)$ satisfies the boundary condition $\varrho_*(\vartheta, x) = \sigma(x)$ for all $x \in R^n$. The last and the second conditions are fulfilled for rather common classes of systems

and quality indices. We remind the definition of the u-stable property: let $\varepsilon > 0$, $t^* \in [t_*, \vartheta]$, $v[\cdot] \in \mathcal{V}_{[t_*,t^*]}$ then, for these data, there exists $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ such that

$$\varrho_*(t_*, x_*) \geq \varrho(t^*, x(t^*, t_*, x_*, u[\cdot], v[\cdot])) - \varepsilon. \tag{4}$$

Inequality (4) will be satisfied if the relevant inequality

$$\varrho_* = \varrho_{*\Delta}(t_*, x_*) \geq \varrho_{*\Delta'}(t^*, x_{u[\cdot]}) - \varepsilon = \varrho^*(x_{u[\cdot]}) - \varepsilon,$$

$$x_{u[\cdot]} = x(t^*, t_*, u[\cdot], v[\cdot]), \tag{5}$$

$$\varrho_{*\Delta'}(t^*, x_{u[\cdot]}) = \sup_{v(\cdot)\in P_v(\Delta')} \inf_{u(\cdot)\in P_u(\Delta')} \int_{\Omega'} (\sigma(x(\vartheta, t^*, x_{u[\cdot]}, u(\cdot, \omega'), v(\cdot, \omega')))d\omega'$$

will be true for every partition Δ. In the following, we assume the data t_*, t^*, x_*, $v[\cdot] \in \mathcal{V}_{[t_*,t^*]}$ to be unchanging. So, we have to verify whether such $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ satisfying (5) exists or not.

The proof of the u-stable property

Let us construct an extension of system (1) and quality index (2):

$$y = (x, x_{n+1}) \in R^{n+1}$$

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu[t] + Cv[t], \quad t \in [t_*, \vartheta] \\ \dot{x}_{n+1}(t) = u^2[t] = |u[t]|^2_{R^k}, \\ y(t_*) = (x(t_*), x_{n+1}(t_*)) = (x_*, 0) = y_*, \end{cases} \tag{6}$$

$$\sigma_\alpha(y) = \sigma(x) + \alpha \cdot x^2_{n+1}, \quad \alpha > 0. \tag{7}$$

Denote

$$y_{u[\cdot]} = y(t^*, t_*, y_*, u[\cdot], v[\cdot]),$$

where, $y(\cdot, t_*, y_*, u[\cdot], v[\cdot])$ is a solution of (6). Introduce values $\varrho_{*\alpha}$, $\varrho^*_\alpha(\cdot)$ corresponding to the values ϱ_* and $\varrho^*(\cdot)$:

$$\begin{aligned} \varrho_{*\alpha} &= \sup_{v(\cdot)\in P_v(\Delta)} \inf_{u(\cdot)\in P_u(\Delta)} \int_{\Omega} \sigma_\alpha y(\vartheta, t_*, x_*, u(\cdot, \omega), v(\cdot, \omega)))d\omega, \\ \varrho^\alpha_*(y_{u[\cdot]}) &= \sup_{v(\cdot)\in P_v(\Delta')} \inf_{u(\cdot)\in P_u(\Delta')} \int_{\Omega'} \sigma_\alpha(y(\vartheta, t^*, y_{u[\cdot]}, u(\cdot, \omega'), v(\cdot, \omega')))d\omega'. \end{aligned} \tag{8}$$

Lemma 1. For any number $\eta > 0$ there exists $\alpha(\eta) > 0$ such that the inequalities

$$|\varrho_{*\alpha} - \varrho_*| < \eta, \tag{9}$$

$$|\varrho^*_\alpha(y_{u[\cdot]}) - \varrho^*(x_{u[\cdot]})| < \eta \tag{10}$$

are fulfilled for every $\alpha < \alpha(\eta)$ and $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$.

Proof. Taking into account compactness of $\mathcal{P}$ and $\mathcal{Q}$ we obtain the inequality

$$|\sigma(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))) - \sigma_\alpha(y(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))) \le C \cdot \alpha,$$

$$C = \text{ const},$$

for any $u(\cdot) \in P_u(\Delta)$, $v(\cdot) \in P_v(\cdot)$ and almost all $\omega \in \Omega$ (in the sense of Lebesgue). Using the inequality and definitions of ϱ_* and $\varrho_{*\alpha}$ one can obtain (9). Inequality (10) may be verified by a similar reasoning.

Denote by $M(t,\cdot)$ the operator of conditional mathematical expectation with respect to $\omega_1, \dots, \omega_i$ if $t \in [\tau_i, \tau_{i+1})$. Introduce the motion of the conjugate system:

$$\begin{cases} \dot{s}(t,\omega) = -A^T s(t,\omega), \quad (t,\omega) \in [t_*, \vartheta] \times \Omega, \\ \dot{s}_{n+1}(t,\omega) = 0, \\ s(\vartheta,\omega) = \dfrac{\partial \sigma}{\partial x}(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))), \\ s_{n+1}(\vartheta,\omega) = 2\alpha \cdot \displaystyle\int_{[t_*,\vartheta]} u^2(\tau,\omega)\, d\tau. \end{cases} \tag{11}$$

It holds for all $(t,\omega) \in [t_*, \vartheta] \times \Omega$

$$s(t,\omega) = X^T(\vartheta, t)\frac{\partial \sigma}{\partial x}(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))),$$
$$s_{n+1}(t,\omega) = s_{n+1}(\vartheta,\omega),$$

where $X^T(\vartheta, t)$ is the transposed fundamental matrix for (1).

Lemma 2. For any $\overline{v}(\cdot) \in P_v(\Delta)$ there exists unique $u_\alpha(\cdot) \in P_u(\Delta)$ giving minimum to the functional

$$I_\alpha(u(\cdot)) = \int_\Omega \sigma_\alpha(y(\vartheta, t_*, y_*, u(\cdot,\omega), v(\cdot,\omega)))\, d\omega. \tag{12}$$

This non-anticipatory program $u_\alpha(\cdot)$ is the unique element satisfying the following condition: for almost all $(t,\omega) \in [t_*, \vartheta] \times \Omega$

$$\begin{aligned} &\langle B u_\alpha(t,\omega), M(t, s(t,\omega))\rangle_{R^n} + u^2_\alpha(t,\omega) \cdot M(t, s_{n+1}(t,\omega)) = \\ &\quad = \min_{u \in P}\{\langle Bu, M(t, s(t,\omega))\rangle + u^2 \cdot M(t, s_{n+1}(t,\omega))\}, \end{aligned} \tag{13}$$

where $(s(\cdot), s_{n+1}(\cdot))$ is the motion conjugate with respect to the stochastic motion $y(\cdot, t_*, y_*, u_\alpha(\cdot), \overline{v}(\cdot))$.

Scheme of the proof. We show that there are no two different elements in $L_\infty([t_*, \vartheta] \times \Omega)$ satisfying (13). Suppose the contrary: there are $u_1(\cdot)$ and $u_2(\cdot)$ from $P_u(\Delta)$ that make (13) to be true. Using a well-known reasoning [12, chapter II, §2] we can deduce from (13):

$$\langle B(u_1(t,\omega) - u_2(t,\omega)), M(t, X^T(\vartheta, t) \cdot$$

$$\cdot \left(\frac{\partial \sigma}{\partial x}\left(W + \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)d\tau\right) -\right.$$

$$\left. -\left(\frac{\partial \sigma}{\partial x}\left(W + \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)d\tau\right)\right)\right\rangle + \qquad (14)$$

$$+(u_1^2(t,\omega) - u_2^2(t,\omega)) \cdot M\left(t,\ 2\alpha \int\limits_{[t_*,\vartheta]} (u_2^2(\tau,\omega) - u_1^2(\tau,\omega))d\tau\right) \geq 0,$$

$$W = X(\vartheta, t_*)x_* + \int\limits_{[t_*,\vartheta]} X(\vartheta\tau)C\overline{v}(\tau,\omega)d\tau.$$

Because of the non-anticipatory property of $u_1(\cdot)$ and $u_2(\cdot)$ we can introduce everything in the right-hand side of (14) into the common sign $M(t, \cdot)$. Then we integrate the expression with respect to $\omega \in \Omega$ and $t \in [t_*, \vartheta]$ obtaining

$$\int\limits_{\Omega} \left\{\left\langle \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)d\tau - \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)d\tau,\right.\right.$$

$$\frac{\partial \sigma}{\partial x}\left(W + \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)d\tau\right) - \frac{\partial \sigma}{\partial x}\left(W + \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)d\tau\right)\Big\rangle - \qquad (15)$$

$$\left. -2\alpha \cdot \left(\int\limits_{[t_*,\vartheta]} (u_1^2(\tau,\omega) - u_2^2(\tau,\omega))d\tau\right)^2\right\}d\omega \geq 0.$$

For the function $\sigma(\cdot)$, as for a convex and smooth one, the following statements will be fulfilled [13, chapter I, properties 5.4 and 5.5]: for all $x_1, x_2 \in R^n$

a) $$\left\langle \frac{\partial \sigma}{\partial x}(x_1) - \frac{\partial \sigma}{\partial x}(x_2), x_1 - x_2 \right\rangle \geq 0,$$

b) $$\left\langle \frac{\partial \sigma}{\partial x}(x_1) - \frac{\partial \sigma}{\partial x}(x_2),\ x_1 - x_2 \right\rangle = 0,$$

if and only if $x_1 = x_2$.

From (15), a), b) it follows that for almost all $\omega \in \Omega$ the expression under the integral equals zero. Hence, for almost all $\omega \in \Omega$, the following equalities are true:

$$\int\limits_{[t_*,\vartheta]} u_1^2(\tau,\omega)d\tau = \int\limits_{[t_*,\vartheta]} u_2^2(\tau,\omega)d\tau \tag{16}$$

$$\int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)d\tau = \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)d\tau. \tag{17}$$

Equalities (16), (17) give us

$$\begin{aligned} M(t, s^{(1)}(t,\omega)) &= M(t, s^{(2)}(t,\omega)), \\ M(t, s_{n+1}^{(1)}(t,\omega)) &= M(t, s_{n+1}^{(2)}(t,\omega)), \end{aligned} \tag{18}$$

where the upper index shows which of the considered programs $u_1(\cdot)$, $u_2(\cdot)$ the component of the conjugate motion belongs to. Equalities (18) and (13) make the programs $u_1(\cdot)$ and $u_2(\cdot)$ to be equal almost everywhere on $[t_*,\vartheta] \times \Omega$.

The necessity of (13) can be deduced from the necessary conditions of the extremum for the convex and differentiable functionals (see, e.g. [5, chapter II]).

Denote by $B_{n+1}(r)$ a sphere with radius $r \geq 0$ and center $(0,\ldots,0) \in R^{n+1}$. Let us determine a multivalued mapping of the set $B_{n+1} \times \mathcal{U}_{[t_*,t^*]}$ into itself. Let $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ and $\{v_j(\cdot) : j \in N\} \subset P_v(\Delta')$ be any sequence approximating the supremum in the definition of $\varrho_\alpha^*(y_{u[\cdot]})$ in (8). This sequence generates the sequence $\{u_j(\cdot) : j \in N\} \subset P_u(\Delta')$ giving the infimums in (8) for the corresponding $v_j(\cdot)$. Every couple $(u_j(\cdot), v_j(\cdot))$ gives us a motion $(s^{(j)}(\cdot), s_{n+1}^{(j)}(\cdot))$ of conjugate system (11). The set of limit points of the set $\{M(t^*-0, (s^{(j)}(t^*,\cdot), s_{n+1}^{(j)}(t^*,\cdot))) : j \in N\}$ for all possible sequences $\{v_j(\cdot) : j \in N\}$ mentioned above, we denote by $S(u[\cdot])$. For a sufficiently large r the inclusion $S(u[\cdot]) \subset B_{n+1}(r)$ will be fulfilled for every $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$.

Any element $(s, s_{n+1}) \in B_{n+1}(r)$ provides a solution $(s(\cdot), s_{n+1}(\cdot))$ of the conjugate system satisfying boundary condition $s(t^*) = s$, $s_{n+1}(t^*) = s_{n+1}$ and the solution generates a subset $\mathcal{U}(s, s_{n+1}) \subset \mathcal{U}_{[t_*,t^*]}$ of elements $u[\cdot]$ satisfying equalities

$$\begin{aligned} &\langle Bu[\tau], s(\tau)\rangle + u^2[\tau] \cdot s_{n+1}(\tau) = \\ &= \min_{u \in P}\{\langle Bu, s(\tau)\rangle + u^2 \cdot s_{n+1}(\tau)\}, \quad \tau \in [t_*, t^*]. \end{aligned}$$

One may check that for the mapping

$$\{u[\cdot], (s, s_{n+1})\} \longrightarrow \{\mathcal{U}(s, s_{n+1}), s(u[\cdot])\}$$

Kakutani's theorem [14] is true. Let $\{u_\alpha[\cdot], (s^\alpha, s^\alpha_{n+1})\}$ be the fixed point of this mapping and the sequences $\{v_j(\cdot) : j \in N\}$ $\{u_j(\cdot) : j \in N\}$ generate (in the above sense) the point $(s^\alpha, s^\alpha_{n+1})$. Denote by

$$u_{\alpha j}(\tau, \omega') = \begin{cases} u_\alpha[\tau], & (\tau, \omega') \in [t_*, t^*) \times \Omega', \\ u_j(\tau, \omega), & (\tau, \omega') \in [t^*, \vartheta] \times \Omega', \end{cases} \quad j \in N,$$

$$v_{\alpha j}(\tau, \omega') = \begin{cases} v_\alpha[\tau], & (\tau, \omega') \in [t_*, t^*) \times \Omega', \\ v_j(\tau, \omega), & (\tau, \omega') \in [t^*, \vartheta] \times \Omega', \end{cases} \quad j \in N,$$

$$u_{vj}(\cdot) = \operatorname{argmin}\{I_{\alpha j}(u(\cdot)) : u(\cdot) \in P_u(\Delta)\}, \quad j \in N;$$

where

$$I_{\alpha j}(u(\cdot)) = \int_\Omega \sigma_\alpha(y(\vartheta, t_*, y_*, u(\cdot, \omega), v_{\alpha j}(\cdot, \omega')))d\omega.$$

Programs $u_{vj}(\cdot)$ will belong to the set $P'_u(\Delta)$ because $v_{\alpha j}(\cdot) \in P'_v(\Delta)$, $j \in N$. For these sequences (in accordance with their definitions) the following expressions will be true:

$$\begin{aligned} I_{\alpha j}(u_{vj}(\cdot)) &\le \varrho_{*\alpha}, \quad j \in N, \\ \lim_{j \to \infty} I_{\alpha j}(u_{\alpha j}(\cdot)) &= \varrho^*_\alpha(y_{u_\alpha[\cdot]}). \end{aligned} \tag{20}$$

Suppose for a moment that for a subsequence of indices $j(i)$ the equality

$$\lim_{i \to \infty} |u_{vj(i)}(\cdot) - u_{\alpha j(i)}(\cdot)|_{L_2([t_*, \vartheta] \times \Omega)} = 0 \tag{21}$$

is fulfilled. Then using the local Lipschitz property of the functionals $I_{\alpha j}(\cdot)$ [13, chapter I, §2] we should obtain

$$\lim_{i \to \infty} I_{\alpha j(i)}(u_{\alpha j(i)}(\cdot)) - I_{\alpha j(i)}(u_{vj(i)}(\cdot)) = 0. \tag{22}$$

From (20) and (22) it follows that

$$\varrho_{*\alpha} \ge \varrho^*_\alpha(y_{u_\alpha[\cdot]}). \tag{23}$$

And, finally, using Lemma 1 we should choose such a small $\alpha(\varepsilon)$ that the desired inequality (5) would follow from (23) for $u_{\alpha(\varepsilon)}[\cdot] \in \mathcal{U}_{[t_*, t^*]}$. We shall briefly outline the proof of (21). Lemma 2 gives for $u_{\alpha j}(\cdot)$ and $u_{vj}(\cdot)$ the inequalities

$$\int_{\Omega'} \left\{ \left\langle \int_{[t_*,\vartheta]} X(\vartheta,\tau)B(u_{\alpha j}(\tau,\omega') - u_{vj}(\tau,\omega'))d\tau, \right.\right.$$

$$\left.\left. \frac{\partial\sigma}{\partial x}\left(W_j + \int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_{vj}(\tau,\omega')d\tau \right) \right\rangle + \right. \tag{24}$$

$$\left. +2\alpha \cdot \int_{[t_*,\vartheta]} (u^2_{\alpha j}(\tau,\omega') - u^2_{vj}(\tau,\omega'))d\tau \cdot \int_{[t_*,\vartheta]} u^2_{vj}(\tau,\omega')d\tau \right\} d\omega' \geq 0,$$

$$\int_{\Omega'} \left\{ \left\langle \int_{[t_*,\vartheta]} X(\vartheta,\tau)B(u_{vj}(\tau,\omega') - u_{\alpha j}(\tau,\omega'))d\tau, \right.\right.$$

$$\left.\left. \frac{\partial\sigma}{\partial x}(W_j + \int_{[t^*,\vartheta]} X(\vartheta,\tau)Bu_{\alpha j}(\tau,\omega')d\tau) \right\rangle + \right. \tag{25}$$

$$\left. +2\alpha \cdot \int_{[t^*,\vartheta]} (u^2_{vj}(\tau,\omega') - u^2_{\alpha j}(\tau,\omega'))d\tau \cdot \int_{[t^*,\vartheta]} u^2_{\alpha j}(\tau,\omega')d\tau \right\} d\omega' \geq 0,$$

$$W_j = X(\vartheta,t_*)x_* + \int_{[t_*,\vartheta]} X(\vartheta,\tau)Cv_{\alpha j}(\tau,\omega')d\tau.$$

Using the properties of the fixed point $\{u_\alpha[\cdot],\ (s^\alpha, s^\alpha_{n+1})\}$ and again Lemma 2, it is possible to prove the estimate

$$\int_{\Omega'} \left\{ \left\langle \int_{[t_*,t^*]} X(\vartheta,\tau)B(u_{vj}(\tau,\omega') - u_{\alpha j}(\tau,\omega'))\,d\tau, \right.\right.$$

$$\left.\left. \frac{\partial\sigma}{\partial x}\left(W_j + \int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_{\alpha j}(\tau,\omega')d\tau \right) \right\rangle + \right. \tag{26}$$

$$+2\alpha \cdot \int\limits_{[t_*,t^*]} (u_{vj}^2(\tau,\omega') - u_{\alpha j}^2(\tau,\omega'))d\tau \cdot \int\limits_{[t_*,\vartheta]} u_{\alpha j}^2(\tau,\omega')d\tau \Bigg\} d\omega' \geq -\delta_j,$$

$$\delta_j \geq 0, \quad \lim_{j\to\infty} \delta_j = 0. \tag{26}$$

The sum of (24)–(26) gives us inequalities similar to (15) with $-\delta_j$ in the right-hand side. These inequalities and properties a), b) of the function $\sigma(\cdot)$ give us

$$\lim_{i\to\infty} \left| \int\limits_{[t_*,\vartheta]} u_{\alpha j(i)}^2(\tau,\omega')d\tau - \int\limits_{[t_*,\vartheta]} u_{vj(i)}^2(\tau,\omega')d\tau \right| = 0,$$

$$\lim_{i\to\infty} \left| \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_{\alpha j(i)}(\tau,\omega')d\tau - \int\limits_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_{vj(i)}(\tau,\omega')d\tau \right|_{R^n} = 0 \tag{27}$$

for almost all $\omega' \in \Omega'$ and for some subsequences of indices $j(i)$. Equalities (27) lead to the convergence of the conjugate motions determining the programs $u_{\alpha j(i)}(\cdot)$, $u_{vj(i)}(\cdot)$ (see (13) of Lemma 2). Finally, the convergence causes equality (21).

Conclusion

The provided scheme of the proof remains valid for the systems of the form

$$\dot{x}(t) = A(t,v[t])x(t) + B(t,v[t])u[t] + C(t)v[t], \tag{28}$$

where the dependence of $A(\cdot), B(\cdot), C(\cdot)$ on v and t is only restricted by the conditions of the existence and uniqueness theorem for equation (28).

The conditions of the strict convexity and of the smoothness on the function $\sigma(\cdot)$ may be omitted. For our needs it is sufficient that $\sigma(\cdot)$ be a convex function, because when introducing the function $\sigma_\alpha(\cdot)$, we may slightly deform $\sigma(\cdot)$ in such a way that these properties will be inherent for the corresponding part of $\sigma_\alpha(\cdot)$ and because of this Lemma 1 remains true.

Acknowledgement

The author is grateful to N.N. Krasovskii for the statement of the problem and discussion of the article.

References

1. *Krasovskii, N. N., Subbotin, A. I.*, Positional differential games. Moscow, Nauka, 1974; English transl.: New York, Springer, 1987.
2. *Osipov, Yu. S.*, Differential games in systems with aftereffects, Dokl. AN SSSR, 1971, vol. **196**, No. *4* (in Russian).
3. *Osipov, Yu. S.*, On the theory of differential games in distributed parameter systems, Dokl. AN SSSR, 1975, vol. **223**, No. *6* (in Russian).
4. *Kurzhanskii, A. B.*, Control and observation under uncertainty. Moscow, Nauka, 1977 (in Russian).
5. *Subbotin, A. I., Chentsov, A. G.*, Optimization of guarantee in control problems. Moscow, Nauka, 1981 (in Russian).
6. *Isaacs, R.*, Differential games. New York, Wiley, 1965.
7. *Friedman, A.*, Differential games. New York, Academic Press, 1975.
8. *Chernous'ko, F. L., Melikyan, A. A.*, Game problems of control and search. Moscow, Nauka, 1978 (in Russian).
9. *Krasovskii, N. N., Tret'yakov, V. E.*, Stochastic program synthesis for positional differential game, Dokl. AN SSSR, 1981, vol. **259**, No. *1* (in Russian).
10. *Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., Mischenko, E. F.*, The mathematical theory of optimal processes. Moscow, Nauka, 1961 (in Russian).
11. *Krasovskii, N. N.*, Control of the dynamical system. Moscow, Nauka, 1985 (in Russian).
12. *Kinderlehrer, D., Stampacchia, G.*, An introduction to variational inequalities and their applications. New York, Academic Press, 1980.
13. *Ekeland, I., Temam, R.*, Convex analysis and variational problems. Amsterdam, North-Holland, 1976.
14. *Kantorovich, L. V., Akilov, G. P.*, Functional analysis. Moscow, 1977.

Стохастическое прицеливание в детерминированном позиционном управлении

Д. А. СЕРКОВ

(Свердловск)

Известно представление цены дифференциальной игры в виде стохастического максимина (с.м.). Наибольший интерес с точки зрения возможных численных реализаций представляет с.м. в форме, содержащей математическое ожидание значений показателя качества на случайных движениях. В работе приводится непосредственное доказательство равенства такой формы с.м. цене дифференциальной игры для одного класса нелинейных по управлению систем и выпуклого терминального показателя качества.

Доказательство основано на идее использования сопряженных переменных принципа максимума Л. С. Понтрягина и не использует результаты теории уравнений в частных производных и динамического программирования.

Д. А. Серков
Институт математики и механики
УрО АН СССР,
СССР, 620219, Свердловск, ГСП–384,
ул. С. Ковалевскои, 16.

Problems of Control and Information Theory, Vol. 18(5), pp. 289–309 (1989)

DETECTION OF CHANGES IN A SIMPLE REGRESSION MODEL OF A RANDOM PROCESS

J. MICHÁLEK

(*Prague*)

(Received October 23, 1988)

The author studies the construction of maximal likelihood estimates of change moment in the behaviour of expected value under unchanging covariance functions. A test of the hypothesis "a change occurred" against the alternative "no change occurred" is suggested, too.

The detection of changes in the behaviour of random processes is a very important statistical decision problem from the practical point of view. The first who investigated instants of changes in the case of random processes in continuous time was Shiryaev [4], [5] and [6]. He studied the detection of instants of change for the standard Wiener process. The Bayes method for estimation is used. There exists the survey paper [7] summarizing all the important results made in this field up to 1980. The main goal of this paper is to suggest MLE of instants of change in the behaviour of expected value and to test the hypothesis "a change occurred" against the alternative "no change occurred" for some classes of random processes.

Let a random process $\{x(\cdot)\}$ with finite second moments be observed in the interval $\langle 0, T\rangle$. Let

$$m(t) = E\{x(t)\}, \quad R(s,t) = E\{(x(s) - m(s))\overline{(x(t) - m(t))}\}.$$

A change in the behaviour $\{x(\cdot)\}$ consists in occurring an instant $\tau \in (0,T)$ at which the expected value $m(\cdot)$ of $\{x(\cdot)\}$ changes. Thus, we assume that in the interval $\langle 0, \tau\rangle$

$$m(t) = \varphi_1(t)$$

and for the interval $(\tau, T\rangle$

$$m(t) = \varphi_2(t)$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

where $\varphi_1(\cdot) \neq \varphi_2(\cdot)$ in $(\tau, T\rangle$. We will assume, further, that a change in the behaviour of $m(\cdot)$ at a moment τ does not effect the covariance function $R(\cdot, \cdot)$. That means, for every $\tau_1, \tau_2 \in (0, T)$

$$\operatorname{cov}_{\tau_1}(x(s), x(t)) = \operatorname{cov}_{\tau_2}(x(s), x(t)) = R(s, t),$$

for every $s, t \in \langle 0, T\rangle$. If no change occurs, i.e. $m(t) = \varphi_1(t)$ for every $t \in \langle 0, T\rangle$, we denote the corresponding covariance function by $\operatorname{cov}_T(\cdot, \cdot)$. The opposite situation, i.e. $m(t) = \varphi_2(t)$ for every $t \in \langle 0, T\rangle$, will be denoted by $\operatorname{cov}_0(\cdot, \cdot)$. We will assume that the covariance function does not change in these marginal cases, too. We can put, for simplicity, $\varphi_1(\cdot) \equiv 0$. This simple regression model can be considered as a special generalization of that studied by Hájek [1]. Now, the following problem arises: how to estimate the unknown time instant τ of a possible change of $m(\cdot)$ if such a change happened. A suitable estimate of τ will be sought among random variables derived from linear combinations

$$\xi = \sum_{i=1}^{n} \lambda_i x(t_i), \quad t_i \in \langle 0, T\rangle.$$

Let $\mathcal{L}$ be the linear set of all these combinations, let

$$\langle \xi, \eta \rangle_\tau = E_\tau\{\xi\overline{\eta}\} = \operatorname{cov}_\tau(\xi, \eta) + E_\tau\{\xi\} E_\tau\{\overline{\eta}\},$$

$\xi, \eta \in \mathcal{L}$. Evidently, under our assumptions,

$$\langle \xi, \eta \rangle_\tau = \langle \xi, \eta \rangle_T + E_\tau\{\xi\} E_\tau\{\overline{\eta}\}$$

for every $\tau \in \langle 0, T\rangle$ because $E_T \xi = E_T \eta = 0$. Hence, of course,

$$\|\xi\|_\tau^2 = \|\xi\|_T^2 + |E_\tau\{\xi\}|^2$$

for every $\xi \in \mathcal{L}$ and we see: if $\|\xi\|_\tau \to 0$ for arbitrary $\tau \in \langle 0, T)$ then $\|\xi\|_T \to 0$. On the contrary, in general, the opposite implication does not hold. As long as the expected value $E_\tau(\cdot)$ will be a bounded linear functional with respect to the norm $\|\cdot\|_T$ on $\mathcal{L}$, then both norms $\|\cdot\|_T$, $\|\cdot\|_\tau$ will be equivalent. If $\|\xi\|_\tau = 0$, then $E_\tau\{\xi\} = 0$, and hence $\|\xi\|_T = 0$. On the other hand, if $\|\xi\|_\tau = 0$, then

$$\|\xi\|_\tau^2 = |E_\tau\{\xi\}|^2$$

what implies $\xi = \text{const}$ a. e. $[P_\tau]$. At the same moment we have $\xi = 0$ a. e. $[P_T]$ and the measures P_τ, P_T must be mutually orthogonal in case of const $\neq 0$. But, if $E_\tau\{\cdot\}$ is bounded on $\mathcal{L}$, i.e. there exists $K_\tau < \infty$ such that

$$|E_\tau\{\xi\}| \leq K_\tau \|\xi\|_T$$

for every $\xi \in \mathcal{L}$, then the fact $\|\xi\|_T = 0$ implies $E_\tau\{\xi\} = 0$ and $\xi = 0$ a. e. $[P_\tau]$, too. Under this assumption a singular situation cannot occur. In general, the closure $\overline{\mathcal{L}}_\tau$ of $\mathcal{L}$ with respect to the norm $\|\cdot\|_\tau$ can differ from the closure $\overline{\mathcal{L}}_T$ made by means of the norm $\|\cdot\|_T$ if both the norms are not equivalent. Such a situation, as we have seen, can lead to the orthogonality of the corresponding measures. This possibility is typical for Gaussian measures as we will see later.

Lemma 1. In our regression model the norms $\{\|\cdot\|_\tau, \tau \in \langle 0, T\rangle\}$ are mutually equivalent (i.e. $\|\xi\|_{\tau_1} = 0 \Leftrightarrow \|\xi\|_{\tau_2} = 0$; $\|\xi\|_{\tau_1} \to 0 \Leftrightarrow \|\xi\|_{\tau_2} \to 0$) if and only if $E_\tau\{\cdot\}$ is bounded with respect to the norm $\|\cdot\|_T$ on $\mathcal{L}$ for every $\tau \in \langle 0, T)$.

Proof. If all the norms $\{\|\cdot\|_\tau, \tau \in \langle 0, T\rangle\}$ are mutually equivalent in the considered sense then they generate a unique topology on the set $\mathcal{L}$, let $\overline{\mathcal{L}}$ be the closure of $\mathcal{L}$ with respect to this topology. As $\mathcal{L} \subset \mathcal{L}_2(\Omega, \sigma, P_T)$ and all random variables $x(t)$, $t \in \langle 0, T\rangle$ are defined on a measurable space, (Ω, σ), $(\overline{\mathcal{L}}, \|\cdot\|_T)$ is a Hilbert space with the inner product $\langle\cdot,\cdot\rangle_T$. Let $\xi \to 0$ with respect to the norm $\|\cdot\|_T$. Then, thanks to our model,

$$\|\xi\|_\tau^2 = \|\xi\|_T^2 + |E_\tau\{\xi\}|^2$$

under the equivalence of $\{\|\cdot\|_\tau, \tau \in \langle 0, T\rangle\}$ $\|\xi\|_\tau \to 0$, too. Thus

$$E_\tau\{\xi\} \to 0.$$

Since $E_\tau\{\cdot\}$ is a linear functional on $\overline{\mathcal{L}}$ its continuity at the null element in $\overline{\mathcal{L}}$ is equivalent to its boundedness. On the contrary, when $|E_\tau\xi| \le K_\tau\|\xi\|_T$ for every $\tau \in \langle 0, T)$ then

$$\|\xi\|_\tau^2 \le \|\xi\|_T^2 + K_\tau^2\|\xi\|_T^2 = (1 + K_\tau^2)\|\xi\|_T^2$$

and the norm $\|\cdot\|_T$ dominate all the norms $\{\|\cdot\|_\tau, \tau \in \langle 0, T)\}$. In this case, as follows from our regression model, every norm $\|\cdot\|_\tau$ plays the dominating role for the norm $\|\cdot\|_T$ because $|E_\tau\{\xi\}|^2 \ge 0$. QED

At this moment, we mention the Gaussian case. One can prove for Gaussian measures with the same covariance function that the mutual equivalence of the norms $\{\|\cdot\|_\tau, \tau \in \langle 0, T\rangle\}$ implies the mutual absolute continuity of the corresponding measures that does not hold in a general case. A detailed information on this can be found e.g. in [2]. The case, where all norms are mutually equivalent, will be called regular. Next, we will consider a regular case only.

Let us consider the Hilbert space $(\overline{\mathcal{L}}, \langle\cdot,\cdot\rangle_0)$. Since $E_\tau\{\cdot\}$ is, in the regular case, bounded with respect to $\|\cdot\|_0$, there exists such an element $u(\tau) \in \overline{\mathcal{L}}$, that

$$E_\tau\{\xi\} = E_0\{\xi\overline{u(\tau)}\}$$

for every $\xi \in \overline{\mathcal{L}}$. Then

$$E_\tau\{x(s)\} = E_0\{x(s)\overline{u(\tau)}\} = 0$$

for every $s \leq \tau$ and simultaneously

$$E_\tau\{x(s)\} = E_0\{x(s)\overline{u(\tau)}\} = \varphi_2(s)$$

for every $\tau < s \leq T$.

We see that $u(\tau) \overset{\perp}{_0} \overline{\mathcal{L}}\{(x(s), s \leq \tau\} = \overline{\mathcal{L}_\tau}$ which is the subspace of $\overline{\mathcal{L}}$ generated by all the random variables $x(s)$, $s \leq \tau$. We see, further, that

$$E_0\{x(s)(\overline{1 - u(\tau)})\} = E_0\{x(s)\} - E_0\{x(s)\overline{u(\tau)}\} = \\ = \varphi_2(s) = E_\tau\{x(s)\} = 0$$

for every $s > \tau$ because $E_0\{x(s)\} = \varphi_2(s)$ in this case. In other words,

$$1 - u(\tau) \overset{\perp}{_0} \overline{\mathcal{L}}_{(\tau,T\rangle},$$

where $\overline{\mathcal{L}}_{(\tau,T\rangle} = \overline{\mathcal{L}}\{x(s),\ \tau < s \leq T\}$.

Lemma 2. Let $\xi \in \overline{\mathcal{L}}_\tau \cap \overline{\mathcal{L}}_{(\tau,T\rangle}$. Then $E_0\{\xi\} = 0$.

Proof. If $\xi \in \overline{\mathcal{L}}_\tau$ then $E_\tau\{\xi\} = E_0\{\xi\overline{u(t)}\} = 0$. At the same moment, as assumed, $\xi \in \overline{\mathcal{L}}_{(\tau,T\rangle}$ and hence

$$E_0\{\xi(\overline{(1 - u(t))}\} = 0.$$

This means, of course, $E_0\{\xi\} = 0$. QED

This elementary lemma yields very important conclusions. Let the considered process $\{x(t),\ t \in \langle 0, T\rangle\}$ be continuous in the quadratic mean sense, i.e. for every $t \in \langle 0, T\rangle$

$$\|x(t+h) - x(t)\|_0 \underset{h\to 0}{\longrightarrow} 0.$$

Surely, then the function $\varphi_2(\cdot)$ must be continuous in $\langle 0, T\rangle$ because

$$\varphi_2(t+h) - \varphi_2(t) = E_0\{x(s+h)\} - E_0\{x(s)\} \underset{h\to 0}{\longrightarrow} 0.$$

Such a situation occurs when the covariance function $R(\cdot,\cdot) = E_T\{x(\cdot)\overline{x(\cdot)}\}$ and the function $\varphi_2(\cdot)$ are continuous. Then $x(\tau) \in \overline{\mathcal{L}}_\tau$ and simultaneously $x(\tau) \in \overline{\mathcal{L}}_{(\tau,T\rangle}$, and by means of Lemma 2 we obtain $E_0\{x(t)\} = \varphi_2(\tau) = 0$. This fact holds in every regular case. If $\lim\limits_{t\searrow\tau} \varphi_2(t)$ were different of 0 then the norms $\|\cdot\|_0$, $\|\cdot\|_\tau$ could not be equivalent in our regression model and a singular case would set in. When for every $\tau \in (0, T\rangle$ the norm $\|\cdot\|_\tau$ is equivalent to $\|\cdot\|_0$ then, as follows from our considerations, $\varphi_2(\cdot) \equiv 0$. In this case, of course, nothing is to be distinguished.

Theorem 1. Let a covariance function $R(\cdot,\cdot)$ be continuous in $\langle 0, T\rangle^2$ and let $\varphi_2(\cdot)$ also be continuous. Then in our regression model the norm $\|\cdot\|_\tau$ is equivalent to $\|\cdot\|_T$ only at those points where $\varphi_2(\tau) = 0$.

Proof. Since in our model

$$E_0\{|x(\tau+h)-x(\tau)|^2\} = E_T\{|x(\tau+h)-x(\tau)|^2\} + |\varphi_2(\tau+h)-\varphi_2(\tau)|^2,$$

thus $x(\tau) \in \overline{\mathcal{L}}_{\langle\tau,T\rangle}$, and hence $E_0\{x(\tau)\} = \varphi_2(\tau) = 0$ by Lemma 2. If

$$\lim_{t \searrow \tau} \varphi_2(t) = \varphi_2(\tau) \neq 0$$

then $E_0\{|x(\tau+h)-x(\tau)|^2\} \xrightarrow[h \searrow 0]{} \varphi_2(\tau)$ and evidently the norm $\|\cdot\|_\tau$ would not be equivalent to the norm $\|\cdot\|_T$. In the Gaussian case, thanks to the dichotomy of Gaussian measures, we would obtain immediately that $P_\tau \perp P_T$. But, this means that the appropriate jump at $\tau(\varphi_2(\tau) \neq 0)$ can be distinguished with probability 1. QED

A similar conclusion can be, of course, made in case of random processes possessing derivatives in the quadratic mean sense. Let there exist the derivative $\{\dot{x}(t), t \in \langle 0,T\rangle\}$ of $\{x(t), t \in \langle 0,T\rangle\}$ continuous in the quadratic mean. $\dot{x}(\cdot)$ exists if and only if the generalized second partial derivative of $R(\cdot,\cdot)$ exists together with the first derivative of the function $\varphi_2(\cdot)$ in $\langle 0,T\rangle$. Then

$$E_T\{\dot{x}(s)\overline{\dot{x}(t)}\} = \frac{\partial^2 R(s,t)}{\partial s \partial t}$$

and

$$E_0\{\dot{x}(s)\} = \dot{\varphi}_2(s),$$

$s,\ t \in \langle 0,T\rangle$. It is clear that $\dot{x}(s) \in \overline{\mathcal{L}}$ for every $s \in \langle 0,T\rangle$. When the underlying process $\{x(t),\ t \in \langle 0,T\rangle\}$ satisfies our regression model then its derivative $\{\dot{x}(t),\ t \in \langle 0,T\rangle\}$ does, too. Surely

$$E_\tau\{\dot{x}(s)\overline{\dot{x}(t)}\} = \frac{\partial R(s,t)}{\partial s \partial t} + E_\tau\{\dot{x}(s)\}E_\tau\{\overline{\dot{x}(t)}\}$$

for every $s,\ t \in \langle 0,T\rangle$. Applying Theorem 1 we obtain that in case

$$\dot{\varphi}(\tau) = \lim_{h \to 0} \frac{\varphi(\tau+h)-\varphi(\tau)}{h} \neq 0$$

our model is singular. It means, in order to ensure the equivalence of the norms $\|\cdot\|_\tau$, $\|\cdot\|_T$ on $\overline{\mathcal{L}}$ we must require both $\varphi(t) = 0$ and $\dot{\varphi}(\tau) = 0$. If the process $\{x(t),\ t \in \langle 0,T\rangle\}$ has the n-th derivative in the quadratic mean sense we must demand, for the regularity of our model,

$$\varphi(\tau) = 0,\ \dot{\varphi}(\tau) = 0,\ \ddot{\varphi}(\tau) = 0, \ldots, \varphi^{(n)}(\tau) = 0.$$

Under these circumstances we must specify our model more precisely. As we put $\varphi_1(\cdot) \equiv 0$ those instants of changes in the behaviour of expected value are interesting for us at which the function $\varphi_2(\cdot)$ is vanishing, too. These demands are fulfilled by the following modification. Let a function $\varphi(\cdot)$, complex in general, with $\varphi(0) = 0$ be given on $\langle 0, T\rangle$. Let us set

$$E_0\{x(t)\} = \varphi(t)$$

for every $t \in \langle 0, T\rangle$,

$$E_\tau\{x(t)\} = \varphi(t - \tau)$$

for every $t \in (\tau, T)$ and

$$E_T\{x(t)\} = 0$$

for every $t \in \langle 0, T\rangle$. The covariance function of $\{x(t)\}$, $t \in \langle 0, T\rangle\}$ is without any changes. In a regular case, which is mainly interesting for us, there exists such a random variable $u(\tau) \in \overline{\mathcal{L}}$ that

$$E_\tau\{\xi\} = E_0\{\xi\overline{u(\tau)}\}.$$

As $E_T\{\xi\} = 0$ for every $\xi \in \overline{\mathcal{L}}$ $u(t) = 0$ a. e. Further, as it was shown before,

$$u(\tau) \overset{\perp}{_0} \overline{\mathcal{L}}_\tau$$

and

$$1 - u(\tau) \overset{\perp}{_0} \overline{\mathcal{L}}_{\langle \tau, T\rangle}.$$

Let us investigate, for illustration, the following example with a standard Wiener process. Let a process $\{x(t), t \in \langle 0, T\rangle\}$ be observed, where for $0 \leq t \leq \tau$, $\tau \in \langle 0, T\rangle$

$$x(t) = w(t),$$

for $\tau < t \leq T$

$$x(t) = w(t) + \varphi(t - \tau)$$

($\varphi(0) = 0, \varphi(\cdot)$ is a continuous function defined on $\langle 0, T\rangle$); $\{w(t), t \in \langle 0, T\rangle\}$ is a standard Wiener process. The random variable $u(\tau)$ will be sought among all variables of the form

$$\int_0^T g(\tau, u)dx(u)$$

where $g(\cdot, \cdot)$ is a non-random function and the integral is understood in the sense of Ito. The variable $u(\tau)$ must satisfy

1. $$0 \equiv E_0\left\{x(s)\int_0^T g(\tau, u)dx(u)\right\} \quad \text{for every } s \leq \tau$$

2. $$\varphi(t-\tau) = E_0\left\{x(t)\int_0^T g(\tau,u)dx(u)\right\} \quad \text{for every } t > \tau.$$

Under the hypothesis P_0 we can write

$$0 = E_0\left\{(w(s)+\varphi(s))\left(\int_0^T g(\tau,u)dw(u) + \int_0^T g(\tau,u)d\varphi(u)\right)\right\}.$$

Thus,

$$\int_0^s g(\tau,u)du + \varphi(s)\int_0^T g(\tau,u)d\varphi(u) = 0.$$

If $\int_0^T g(\tau,u)d\varphi(u) \neq 0$, $\dot{\varphi}(s)$ exists and

$$\dot{\varphi}(s) = \frac{g(\tau,u)}{Ig}, \quad Ig = \int_0^T g(\tau,u)d\varphi(u).$$

In case $Ig = 0$,

$$0 = \int_0^s g(\tau,u)du$$

for every $s \leq \tau$, and hence $g(\tau,u) = 0$ a. e. for $0 \leq u \leq \tau$. Then, of course,

$$u(\tau) = \int_\tau^T g(\tau,u)dx(u).$$

The second condition gives

$$\varphi(t-\tau) = \int_0^t g(\tau,u)du + \varphi(t)Ig$$

as follows easily from the properties of the stochastic integral. Under $Ig \neq 0$, derivation with respect t yields

$$g(\tau,t) = \dot{\varphi}(t-\tau) - \dot{\varphi}(t)Ig \quad \text{for } t < \tau.$$

As fas as $Ig = 0$, then

$$\varphi(t-\tau) = \int_0^t g(\tau,u)du = \int_\tau^t g(\tau,u)du$$

because the first condition, in this case, demands $g(\tau,u) = 0$ a. e. on $\langle 0,\tau\rangle$. At this moment, surely,

$$\dot{\varphi}(t-\tau) = g(\tau,t)$$

for $t > \tau$ and $0 = Ig = \int_0^T g(\tau,u)d\varphi(u) = \int_\tau^T \dot{\varphi}(u-\tau)\dot{\varphi}(u)du$ as we supposed. Further, we have

$$Ig = \int_0^\tau g(\tau,u)d\varphi(u) + \int_\tau^T g(\tau,u)d\varphi(u) =$$

$$= -\int_0^\tau Ig\dot{\varphi}(u)d\varphi(u) + \int_\tau^T (\dot{\varphi}(u-\tau) - \dot{\varphi}(u)Ig)\,d\varphi(u) =$$

$$= -Ig\int_0^\tau (\dot{\varphi}(u))^2du + \int_\tau^T \dot{\varphi}(u-\tau)du - Ig\int_\tau^T (\dot{\varphi}(u))^2du =$$

$$= -Ig\int_0^T (\dot{\varphi}(u))^2du + \int_\tau^T \dot{\varphi}(u-\tau)\dot{\varphi}(u)du.$$

We have shown that

$$Ig = \frac{\int_\tau^T \dot{\varphi}(u-\tau)\dot{\varphi}(u)du}{1+\int_0^T (\dot{\varphi}(u))^2du}.$$

When $Ig = 0$, $u(\tau) = \int_\tau^T \dot{\varphi}(u-\tau)dx(u)$; in the other case, $Ig \neq 0$

$$u(\tau) = \int_\tau^T \dot{\varphi}(u-\tau)dx(u) - Ig\int_0^T \dot{\varphi}(u)dx(u).$$

For $\tau = T$ we obtain $u(T) = 0$, and for $\tau = 0$

$$u(0) = \frac{\int_0^T \dot{\varphi}(u)dx(u)}{1+\int_0^T (\dot{\varphi}(u))^2du}.$$

The random variable $u(0)$ is the projection of 1 into the space $\overline{\mathcal{L}}_{\langle 0,T\rangle}$ with respect to $\langle\cdot,\cdot\rangle_0$. When we consider the measure P_T as a dominating measure instead of P_0 the situation is getting somewhat simpler. Under the norm $\|\cdot\|_T$

$$E_T\{x(s)\} = 0$$

for every $s \in \langle 0, T\rangle$, and hence $x(s) = w(s)$. Otherwise,

$$E_\tau\{x(s)\} = E_T\{x(s)v(\tau)\} = \begin{cases} 0 & \text{for } s \leq \tau \\ \varphi(s-\tau) & \text{for } s > \tau. \end{cases}$$

Then, let again

$$v(\tau) = \int_0^T g(\tau, u)dx(u) = \int_0^T g(\tau, u)dw(u)$$

under the measure P_T. This implies

$$E_T\left\{w(s)\int_0^T g(\tau, u)dw(u)\right\} = 0 \text{ for } s \leq \tau$$

which gives $g(\tau, s) = 0$ a. e. in $\langle 0, \tau\rangle$. On the other hand,

$$\varphi(t-\tau) = E_T\left\{w(t)\int_0^T g(\tau, u)dw(u)\right\} = \int_0^t g(\tau, u)du.$$

But,

$$\int_0^t g(\tau, u)du = \int_0^\tau g(\tau, u)du + \int_\tau^t g(\tau, u)du = \int_\tau^t g(\tau, u)du,$$

hence

$$\varphi(t-\tau) = \int_\tau^t g(\tau, u)du,$$

i.e.

$$\dot{\varphi}(t-\tau) = g(\tau, t)$$

for every $\tau < t \leq T$. We see that

$$v(\tau) = \int_\tau^T \dot{\varphi}(u-\tau)dx(u).$$

The properties of the stochastic integral yield easily

$$E_T\{v^2(\tau)\} = \int_\tau^T (\dot{\varphi}(u-\tau))^2 du = \int_0^{T-\tau} (\dot{\varphi}(u))^2 du.$$

Then we can express the corresponding Radon–Nikodym derivative

$$\frac{dP_\tau}{dP_T} = \exp\left\{\int_\tau^T \dot{\varphi}(u-\tau)dx(u)\right\} \exp\left\{-\frac{1}{2}\int_0^{T-\tau} [\dot{\varphi}(u)]^2 du\right\}.$$

In the simplest case, when $\varphi(t) = at$, one can immediately write

$$\frac{dP_\tau}{dP_T} = \exp\left\{a(x(T) - x(\tau)) - \frac{1}{2}a^2(T-\tau)\right\}.$$

This formula gives a possibility to determine MLE of τ. We will look for such a value $\tilde{\tau}$ what is a solution of the following optimization problem

$$\max_{\langle 0,T\rangle}\left\{-ax(\tau) + \frac{1}{2}a\tau^2\right\}.$$

We must realize, of course, that a solution of the previous equation is meaningful only under rejection of the hypothesis P_T, i.e. a change has occurred. The test of the simple hypothesis P_T against the composite hypothesis $\{P_\tau,\ \tau \in \langle 0,T\rangle\}$ can be based on the behaviour of the statistic $\frac{1}{T}\int_0^T x(t)dt$ which is Gaussian under P_T with $E_T\left\{\frac{1}{T}\int_0^T x(t)dt\right\} = 0$ and $D\left\{\frac{1}{T}\int_0^T x(t)dt\right\} = \frac{T}{3}$.

On the other hand, under validity of P_τ, the random variable $\frac{1}{T}\int_0^t x(s)ds =$ $= \frac{1}{T}\int_0^T w(s)ds + \frac{1}{T}\int_\tau^T \varphi(t-\tau)dt$ is also Gaussian, but $E_\tau\left\{\frac{1}{T}\int_0^t x(s)ds\right\} =$ $= \frac{1}{T}\int_\tau^T \varphi(t-\tau)dt = \frac{1}{T}\int_0^{T-\tau} \varphi(u)du$. In case $\varphi(u) = au$ we obtain, of course,

$$E_\tau\left\{\frac{1}{T}\int_0^T x(s)ds\right\} = \frac{a}{T}\frac{(T-\tau)}{2} > 0.$$

Let us fix $\tau \in \langle 0, T)$ and let us construct the best test (based on the Neyman–Pearson lemma) of the simple hypothesis P_T against the simple alternative P_τ. As it is well known, the best test is given by the maximal likelihood ratio

$$\frac{dP_\tau}{dP_T}(x(\cdot)) \gtrless L(\alpha)$$

where $L(\alpha)$ is determined by the relation

$$P_T\left\{x(\cdot) : \frac{dP_\tau}{dP_T}(x(\cdot)) > L(\alpha)\right\} = \alpha.$$

As, in a general case for a standard Wiener process,

$$\frac{dP_\tau}{dP_T}(x(\cdot)) = \exp\left\{\int_\tau^T \dot{\varphi}(u-\tau)dx(u) - \frac{1}{2}\int_\tau^T (\dot{\varphi}(u-\tau))^2 du\right\}$$

then

$$\frac{dP_\tau}{dP_T}(x(\cdot)) > L(\alpha),$$

if and only if $\int_\tau^T \dot{\varphi}(u-\tau)dx(u) - \frac{1}{2}\int_\tau^T (\dot{\varphi}(u-\tau))^2 du > \ln L(\alpha)$. The condition

$P_T\left\{x(\cdot) : \int_\tau^T \dot{\varphi}(u-\tau)dx(u) \geq \frac{1}{2}\int_\tau^T (\dot{\varphi}(u-\tau))^2 du + \ln L(\alpha)\right\} = \alpha$ gives

$$L(\alpha) = \exp\left\{K(\alpha)\sqrt{D_\tau} - \frac{1}{2}D_\tau\right\}$$

where

$$\int_{K(\alpha)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \alpha,$$

$$D_\tau^T = \int_\tau^T (\dot{\varphi}(u-\tau))^2 du.$$

Then the hypothesis P_T is rejected when

$$\frac{\int_\tau^T \dot{\varphi}(u-\tau)dx(u)}{\sqrt{D_\tau}} > K(\alpha).$$

The second-kind error equals

$$1 - P_\tau \left\{ \frac{\int_\tau^T \dot{\varphi}(u-\tau)dx(u)}{\sqrt{D_\tau}} > K(\alpha) \right\} =$$

$$= P_\tau \left\{ \frac{\int_\tau^T \dot{\varphi}(u-\tau)dx(u)}{\sqrt{D_\tau}} \leq K(\alpha) \right\} =$$

$$= P_T \left\{ \frac{\int_\tau^T \dot{\varphi}(u-\tau)dw(u)}{\sqrt{D_\tau}} \leq K(\alpha) - \sqrt{D_\tau} \right\} =$$

$$= \int_{-\infty}^{K(\alpha)-\sqrt{D_\tau}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}u^2\right\} du \leq \int_{-\infty}^{K(\alpha)-\sqrt{D_\tau}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}u^2\right\} du = 1 - \alpha.$$

Thus, we have proved that our test is unbiased because for every

$$\tau \in \langle 0, T)$$

$$P_\tau \left\{ x(\cdot) : \frac{dP_\tau}{dP_T}(x(\cdot)) \geq L(\alpha) \right\} \geq \alpha.$$

In order not to reject the hypothesis P_T we have to obtain

$$\sup_{\tau \in \langle 0,T)} \left\{ \frac{\int_\tau^T \dot{\varphi}(u-\tau)dx(u)}{\sqrt{D_\tau}} \right\} < K(\alpha)$$

when the alternative is composite. It means, if at least one $\tau_0 \in \langle 0, T)$ occurs satisfying the inequality

$$\frac{\int_{\tau_0}^T \dot{\varphi}(u-\tau_0)dx(u)}{\sqrt{D_{\tau_0}}} > K(\alpha)$$

the hypothesis P_T must be rejected.

Remark on Gaussian measures

In our model defined by the same covariance function we can assert more in the case of Gaussian measures than in the general case. Thanks to the dichotomy of Gaussian measures, the case of non-equivalent norms $\|\cdot\|_{P_1}$, $\|\cdot\|_{P_2}$ on the linear set $\mathcal{L}$ generated by the random variables $x(t)$, $t \in \langle 0, T\rangle$, leads to the orthogonality of the corresponding measures P_1, P_2. In a regular case we can determine an explicit form of the Radon–Nikodym derivative

$$\frac{dP_2}{dP_1} = \exp\left\{\eta - \frac{1}{2}\|\eta\|^2_{P_1}\right\}$$

where $\|\eta\|^2_{P_1} = E_{P_1}\{\eta^2\}$. Two Gaussian measures P_1, P_2 with the same covariance function are equivalent if and only if there exists a random variable $\eta \in \overline{\mathcal{L}}$ such that

$$E_{P_2}\{x(t)\} = E_{P_1}\{x(t)\eta\}.$$

In our case, the random variable $v(\tau)$ plays the role of η.

Process with independent increments

Let us consider on $\langle 0, T\rangle$ a random process $\{x(t)\}$ with $x(0) = 0$ having independent increments. Let us denote

$$F(\lambda) = E_T\{x^2(\lambda)\}, \quad F(0) = 0.$$

We will look for a random variable $v(\tau)$ in the form of the stochastic integral

$$v(\tau) = \int_0^T g(\tau, u)dx(u).$$

Then $E_T\left\{x(s)\int_0^T g(\tau, u)dx(u)\right\} = 0$ for every $s \leq \tau$ and simultaneously $E_T\left\{x(t)\int_0^T g(\tau, u)dx(u)\right\} = \varphi(t - \tau)$ for every $t > \tau$.

By means of stochastic integrals the preceding conditions can be expressed as

$$\int_0^s g(\tau, u)dF(u) = 0$$

for every $s \in \langle 0, \tau \rangle$, and

$$\int_0^t g(\tau, u) dF(u) = \varphi(t - \tau)$$

for every $t \in (\tau, T\rangle$. These demands yield that

$$g(\tau, s) = 0 \text{ a. e. } [F] \text{ on } \langle 0, \tau \rangle,$$

and, at the same time,

$$\frac{d\varphi(t-\tau)}{dF(t)} = g(\tau, t)$$

for $t > \tau$. Hence,

$$v(\tau) = \int_\tau^T \frac{d\varphi(t-\tau)}{dF(t)} dx(t)$$

with

$$D_\tau = E_T\{v^2(\tau)\} = \int_\tau^T \left(\frac{d\varphi(u-\tau)}{dF(t)} \right)^2 dF(t).$$

In case of Gaussian measures

$$\frac{dP_\tau}{dP_T}(x(\cdot)) = \exp\left\{ \int_\tau^T \frac{d\varphi(t-\tau)}{dF(t)} dx(t) - \frac{1}{2} \int_\tau^T \left(\frac{d\varphi(t-\tau)}{dF(t)} \right)^2 dF(t) \right\}$$

and this formula gives a possibility to consider the MLE of τ.

Gauss–Markov process

The previous result for processes with independent increments can be used in the Gauss–Markov case. Let $\{x(t),\ t \in \langle 0, T\rangle\}$ be a Gauss–Markov process with a covariance function $R(\cdot, \cdot)$. Then

$$E_T\{x(s)x(t)\} = R(s,t) = \sigma_s \sigma_t \frac{h_s}{h_t}$$

in case $s \leq t$ where $\sigma_s = R(s,s)^{\frac{1}{2}}$ and $h_s = (\sigma_0 \sigma_s)/(R(0,s))$. Let us assume that $R(0,s) \neq 0$ for every $s \in \langle 0, T\rangle$. Following [1] the process $\{y(t)\}$, $t \in \langle 0, T\rangle$ defined by

$$y(t) = x(t) \cdot \frac{\sigma_0}{R(0,t)}$$

is a process with independent increments with

$$E_T\{y(s)y(t)\} = h_s^2$$

for $s \leq t$. Let $E_\tau\{x(s)\} = 0$ for $s \in \langle 0, \tau\rangle$ and $E_\tau\{x(t)\} = \varphi(t-\tau)$ for $t \in (\tau, T\rangle$. Then $E_\tau\{y(s)\} = 0$ and $E_\tau(t)\} = \varphi(t-\tau)\cdot(\sigma_0/R(0,t))$. One can write

$$v_\tau(y(\cdot)) = \int_\tau^T \frac{\sigma_0\left(\frac{\varphi(t-\tau)}{R(0,t)}\right)}{d(h_t^2)} dy(t) =$$

$$= \int_\tau^T \frac{d\left(\frac{\varphi(t-\tau)}{R(0,t)}\right)}{d\left(\frac{\sigma_t^2}{R^2(0,t)}\right)} d\left(\frac{x(t)}{R(0,t)}\right).$$

The corresponding Radon–Nikodym derivative has the form

$$\frac{dP_\tau}{dP_T} = \exp\left\{v_\tau(y(\cdot)) - \frac{1}{2}\|v_\tau(y(\cdot))\|_T^2\right\} =$$

$$= \exp\left\{\int_\tau^T \frac{d\left(\frac{\varphi(t-\tau)}{R(0,t)}\right)}{d\left(\frac{\sigma_t^2}{R^2(0,t)}\right)}\right\} \exp\left\{-\frac{1}{2}\int_\tau^T \left(\frac{d\left(\frac{(\varphi(t-\tau)^2)}{R(0,t)}\right)}{d\left(\frac{\sigma_t^2}{R^2(0,t)}\right)}\right)^2 d\left(\frac{\sigma_t^2}{R^2(0,t)}\right)\right\}.$$

For a stationary Markov process with $\sigma_t = \sigma_0$ and $h_t = e^{at}(a > 0)$ and under assumption that $\varphi(\cdot)$ has the first derivative $\dot{\varphi}(\cdot)$

$$\frac{dP_\tau}{dP_t} = \exp\left\{\int_\tau^T \left(\frac{d(\varphi(t-\tau)e^{at})}{de^{2at}}\right) d\left(x(t)e^{at}\right)\right\} \times$$

$$\times \exp\left\{\frac{1}{2}\int_\tau^T \left(\frac{d\varphi(t-\tau)e^{at}}{de^{2at}}\right)^2 de^{2at}\right\} =$$

$$= \exp\left\{\int_\tau^T \frac{1}{2}(\dot{\varphi}(t-\tau) + a\varphi(t-\tau))x(t)dt + \int_\tau^T \frac{1}{2a}(\dot{\varphi}(t-\tau) + a\varphi(t-\tau))dx(t)\right\} \times$$

$$\times \exp\left\{-\frac{1}{4a}\int_\tau^T [\dot{\varphi}(t-\tau) + a\varphi(t-\tau)]^2 dt\right\}.$$

Weakly stationary processes

From a practical point of view, the most important class of random processes is that of stationary ones. Let $\{x(t), t \in \langle 0, T\rangle\}$ be a weakly stationary process with continuous covariance function. Such a process can be expressed in the form of stochastic integral

$$x(t) = \int_{-\infty}^{+\infty} e^{it\lambda} d\Phi(\lambda)$$

where $\Phi(\cdot)$ is an orthogonally scattered random measure. Let us seek for the random variable $v(\tau)$ in the form

$$v(\tau) = \int_{-\infty}^{+\infty} g(\tau, \lambda) d\Phi(\lambda).$$

Let $F(\cdot)$ be the spectral measure of $x(\cdot)$. Then we need

$$E_T \left\{ x(s) \int_{-\infty}^{+\infty} \overline{g(\tau, \lambda) d\Phi(\lambda)} \right\} = \int_{-\infty}^{+\infty} e^{is\lambda} \overline{g(\tau, \lambda)} dF(\lambda) = 0$$

for every $s \in \langle 0, \tau\rangle$, and

$$E_T \left\{ x(t) \int_{-\infty}^{+\infty} \overline{g(\tau, \lambda) d\Phi(\lambda)} \right\} = \int_{-\infty}^{+\infty} e^{it\lambda} \overline{g(\tau, \lambda)} dF(\lambda) = \varphi(t - \tau)$$

for every $t \in (\tau, T\rangle$. We have obtained a system of integral equations. If there exists a solution of this system then such a solution must be unique. A very important case is given by a spectral density function of the form

$$\frac{dF(\lambda)}{d\lambda} = f(\lambda) = \frac{1}{2\pi} \frac{|P(i\lambda)|^2}{|Q(i\lambda)|^2}$$

where $P(z) = \sum_{k=0}^{m} b_{m-k} z^k$, $Q(z) = \sum_{k=0}^{n} a_{n-k} z^k$, $n \geq m$, and all roots of $P(z) = 0$, $Q(z) = 0$ lie in the left complex half-plane. An expression of $v(\tau)$ will be looked for by means of the theory of reproducing kernel Hilbert space (RKHS) presented e.g. in [3]. First, we will investigate the case with $m = 0$. Now, as proved in [3], the corresponding RKHS is formed by all complex functions defined on $\langle 0, T\rangle$ having

square integrable derivatives up to the order n. The inner product in RKHS is defined by

$$\langle \psi, \varphi \rangle = a_0^2 \int_0^T \psi^{(n)}(s)\overline{\varphi}^{(n)} ds + \sum_{k=0}^{n-1} A_{n-k} \int_0^T \psi^{(k)}(s)\overline{\varphi}^{(k)} ds +$$

$$+ \sum_{0 \le j} \sum_{k \le n-1} \left(\psi^{(j)}(0)\overline{\varphi}^{(k)}(0) + \psi^{(j)}(T)\overline{\varphi}^{(k)}(T) \right)$$

$$\sum_{i=\max(0,j+k-1-n)}^{\min(j,k)} (-1)^{j-1} a_{n-i} a_{n+i-j-k-1}$$

where

$$A_{n-k} = \sum_{l=0}^{2k} a_{n-l} a_{n-2k+l}.$$

We look for a solution of the equation $\varphi(t-\tau) = \langle v(\tau),\, x(t) \rangle$ by means of the substitution

$$\psi^{(k)}(s)ds \leftrightarrow dX^{(k)}(s)$$

where $X(s) = \int_0^s x(u)du$. The random variable $v(\tau)$ solving the equations

$$\langle v(\tau), x(s) \rangle = 0 \text{ for } s \in \langle 0, \tau \rangle \text{ and } \langle v(\tau), x(t) \rangle = \varphi(t-\tau)$$

for $t \in (\tau, T\rangle$ has the following form in this case

$$v(\tau) = a_0^2 \int_0^\tau \varphi_\tau^{(n)}(s) d\overline{X^{(n)}}(s) + \sum_{k=0}^{n-1} A_{n-k} \int_0^T \varphi_\tau^{(k)}(s) d\overline{X}^{(k)}(s) +$$

$$+ \sum_{0 \le j} \sum_{k \le n-1} \alpha_{jk} \left(\varphi_\tau^{(j)}(0)\overline{X}^{(k+1)}(0) + \varphi_\tau^{(j)}(T)\overline{X}^{(k+1)}(T) \right),$$

where

$$\alpha_{jk} = \sum_{i=\max(0,j+k-1-n)}^{\min(j,k)} (-1)^{j-i} a_{n-i} a_{n+i-j-k-1},$$

$$\varphi_\tau(s) = 0 \text{ for } s \in \langle 0, \tau \rangle,\ \varphi_\tau(t) = \varphi(t-\tau) \text{ for } t \in (\tau, T\rangle.$$

Again, for illustration, let us consider the simplest case with the spectral density function

$$f(\lambda) = \frac{1}{2\pi} \frac{1}{|a_1 + a_0 i\lambda|^2}.$$

Let $\varphi(u) = \alpha u$, then $\dot{\varphi}(u-\tau) = \alpha$ for $u > \tau$, $\dot{\varphi}(u-\tau) = 0$ for $u \in \langle 0, \tau \rangle$,

$$v(\tau) = a_0^2 \int_\tau^T \alpha dx(s) + A_1 \int_\tau^T \alpha(s-t)x(s)ds + \alpha_{00}\alpha(T-\tau)x(T) =$$

$$= \alpha a_0^2(x(T) - x(\tau)) + \alpha A_1 \int_\tau^T (s-\tau)x(s)ds + \alpha_{00}\alpha(T-\tau)x(T),$$

$$A_1 = a_1^2,\ \alpha_{00} = a_0 a_1.$$

We can, similarly as it was made sooner, consider the maximal likelihood ratio in the Gaussian case given, in a general form, by

$$\frac{dP_\tau}{dP_T} = \exp\left\{v(\tau) - \frac{1}{2}D_\tau\right\}.$$

Testing the simple hypothesis P_T (no change) against the simple alternative P_τ (at τ change occurred) in our model we will see that the Neyman–Pearson fundamental lemma gives the unbiased tests because

$$P_\tau\left\{x(\cdot) : \frac{dP_\tau}{dP_T}(x(\cdot)) > L(\alpha)\right\} =$$

$$= P_\tau\left\{x(\cdot) : \frac{v(\tau)}{\sqrt{D_\tau}} > K(\alpha)\right\} =$$

$$= P_T\left\{x(\cdot) : \frac{v(\tau) + D_\tau}{\sqrt{D_\tau}} > K(\alpha)\right\} =$$

$$= P_T\left\{x(\cdot) : \frac{v(\tau)}{\sqrt{D_\tau}} > K(\alpha) - \sqrt{D_\tau}\right\} =$$

$$= \int_{K(\alpha)-\sqrt{D_\tau}}^{\infty} N(0,1) \geq \alpha.$$

Under this fact we suggest the testing statistic in the form

$$S(T) = \sup_{\tau \in \langle 0,T)} \left\{\frac{v(\tau)}{\sqrt{D_\tau}}\right\}.$$

When $S(T) > K(\alpha)$ we reject the hypothesis P_T, on the other hand, if $S(T) \leq K(\alpha)$ we can admit the hypothesis P_T.

Now, we return to a general rational spectral density function corresponding to the ARMA-model

$$(*) \qquad g(\lambda) = \frac{1}{2\pi} \left| \frac{\sum_{k=0}^{m} b_{m-k}(i\lambda)}{\sum_{k=0}^{m} a_{n-k}(i\lambda)} \right|^2$$

where $m < n$, the coefficients a_k, b_k are real and all roots of the equations

$$\sum_{k=0}^{n} a_{n-k}\lambda^k = 0, \quad \sum_{k=0}^{m} b_{m-k}\lambda^k = 0$$

possess negative real parts. When a process $\{x(t),\ t \in \langle 0, T\rangle\}$ has a spectral density function of the form $(*)$ with $m = 0$ then the process $\{z(t)\}$, $t \in \langle 0, T\rangle$

$$z(t) = \sum_{k=0}^{m} b_{m-k} x^{(k)}(t)$$

has a spectral density function $g(\cdot)$. In this case, the corresponding RKHS is created by all, complex in general, functions defined on $\langle 0, T\rangle$ having quadratically integrable derivatives up to the order $n - m$. A solution of a general filtration problem can be reduced to that with $m = 0$. A more detailed explanation is given in [3]. When the function $\varphi_\tau(\cdot)$ belongs to the space $\mathcal{L}^2_{2n-2m}\langle 0, T\rangle$ then the corresponding random variables $v(\tau)$ can be expressed by means of trajectories $z(t)$, $t \in \langle 0, T\rangle$

$$v(\tau) = \int_0^T z(t) L^* L\psi_\tau(t)\,dt +$$

$$+ \sum_{j=0}^{n-m-1} z^{(j)}(T) \sum_{k=0}^{n-1-j} (-1)^k (L\psi_\tau(T))^{(k)} a_{n-j-k-1} +$$

$$+ \sum_{j=0}^{n-m-1} z^{(j)}(0) \sum_{k=0}^{n-1-j} (-1)^j (L^* L\psi_\tau(0))^{(k)} a_{n-j-k-1},$$

where

$$\varphi_\tau(t) = \sum_{k=0}^{m}\sum_{j=0}^{m} b_{m-k} b_{m-j} (-1)^k \frac{d^{k+1}}{dt^{k+1}} \psi_\tau(t)$$

and

$$\sum_{k=0}^{n-1-j} (-1)^k (L\psi_\tau(T))^{(k)} a_{n-j-k-1} = 0$$

for $n - m \leq j \leq n - 1$. Similarly,

$$\sum_{k=0}^{n-1-j} (-1)^k (L^* \psi_t(0))^{(k)} a_{n-j-k-1} = 0$$

for $n - m \leq j \leq n - 1$. The operator

$$L\psi_\tau(t) = \sum_{k=0}^{n} a_{n-k} (\psi_\tau(t))^{(k)}$$

and its adjoint operator

$$L^*\psi_\tau(t) = \sum_{k=0}^{n} (-1)^k a_{n-k} (\psi_\tau(t))^{(k)}$$

are defined on the space $\mathcal{L}^2\langle 0, T\rangle$.

References

1. *Hájek, J.*, On a simple regression model in Gaussian processes. Trans. of the Second Prague Conference on Inf. Theory, Dec. Functions and Random Proc., Academia, Prague 1960, 185–197.
2. *Rozanov, Y. A.*, Gaussian Infinite-dimensional Probability Distribution Functions, Nauka, Moscow 1968 (in Russian).
3. *Hájek, J.*, On linear statistical problems in stochastic processes. Czechosl. Math. Journal **12** (87), (1962), No. *3*, 404–444.
4. *Shiryaev, A. N.*, Problem of fastest detection of breakdown of stationary mode. Dokl. Akad. Nauk SSSR, **138** (1961), No. *5*, 1039–1042.
5. *Shiryaev, A. N.*, On optimal methods in fastest detection problem. Teor. Veroyatn. Ee Primen. **8** (1963), No. *1*, 26–51.
6. *Shiryaev, A. N.*, Some exact expressions in the change problem. Teor. Veroyatn. Ee Primen. **10** (1965), No. *2*, 380–385.
7. *Kligiene, N., Telksnins, L.*, Methods of detecting instants of change of random process properties. Automation and Remote Control. March 1984, New York (Translated from Russian).

Обнаружение изменений в простой регрессионной модели случайного процесса

Й. МИХАЛЕК

(Прага)

Автор исследует возможность нахождения максимально вероятной оценки точки изменения среднего значения случайного процесса при неизменной ковариационной

функции. Также предложен тест проверки гипотезы «появление изменения» против альтернативной гипотезы «изменение не появилось».

J. Michálek
Institute of Information Theory and Automation
Pod vodárenskou věží 4
182 08 Prague 8, Czechoslovakia

Problems of Control and Information Theory, Vol. 18(5), pp. 311-322 (1989)

A NONPARAMETRIC ANALYSIS OF PROPORTIONAL HAZARD REGRESSION MODEL

P. VOLF

(*Prague*)

(Received October 15, 1988)

The proportional hazard model of regression is widely used for the analysis of lifetime data. The model is suitable even in the presence of censoring. The complete solution for the parametrized Cox's regression model has been examined frequently. We present a method for the identification of the proportion of hazards as a nonparametrized continuous function. The estimate of the "average" cumulative hazard function is obtained, too. The results are strongly consistent and asymptotically normal, it follows from the well-known large sample properties of the cumulative hazard function estimator.

Keywords: lifetime, nonparametric estimation, proportional hazard, regression.

Introduction

In the field of the survival analysis, one of the more widely used regression models is the Cox's one, which assumes the parametrized exponential proportion of hazard functions. Let us imagine that the proportion of the hazards is given by a nonspecified continuous function $B(x)$. In terms of the cumulative hazard function, the model acquires the form

$$L(t, x) = A(t) \cdot B(x),$$

with x as the regressor variable. When analysing survival data, we may often consider either the linear regression model for the logarithm of lifetime, or the proportional hazard model for lifetime. In both cases, the first step of the analysis may consist in nonparametric estimation. Procedures for nonparametric inference in linear regression model are well developed. But, when the proportional hazard model is considered, there is still need for a method of nonparametric estimation of the hazard proportion. Our contribution suggests such a method of identification of the model. Evidently, the functions A, B cannot be identified unambiguously. We offer their nonparametric estimates, provided that the function $B(x)$ is normed by some condition.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Let N independent random variables $Y(x_i)$ characterize the survival times of N cases, x_i's can acquire their values from an interval $\mathcal{X} \subset R_1$. But the realizations are censored from the right side by means of random variables $V(x_i)$, which are independent mutually, as well as of survival times. The observed variables are $T(x_i) = \min\{Y(x_i), V(x_i)\}$ and $\delta(x_i) = I[Y(x_i) \leq V(x_i)]$. Denote the distribution function of $Y(x)$ by $F(t,x)$, the distribution function of $V(x)$ as $G(t,x)$, their supplements by $P(t,x) = 1 - F(t,x) = \exp -L(t,x)$, $Q(t,x) = 1 - G(t,x)$. From practical point of view there must exist a finite time T as a possible upper value of Y and V. Theoretically, we shall examine the behaviour of cumulative hazards up to some $\mathcal{T} < T$, where $\mathcal{T}$ is such that $P(\mathcal{T}, x) \cdot Q(\mathcal{T} \cdot x) > 0$ for every $x \in \mathcal{X}$ and, naturally, $P(\mathcal{T}, x) < 1$.

Method of estimation

For an arbitrary point $z \in \mathcal{X}$ let us define its neighbourhood as the interval $O_{d_N}(z) = \{x \in \mathcal{X} : |x - z| \leq d_N\}$.

(i) Let us fix K representative points in $\mathcal{X} : z_1, \ldots, z_K$. For every level of regressor values (the level is represented by z_j) we shall estimate the cumulative hazard function. We shall use the standard estimator, cf. [1], but constructed from the realizations in $O_{d_N}(z_j)$, for given $d_N > 0$:

$$L_N(t, z_j) = \sum_{i=1}^{N} \frac{\delta(x_i) \cdot I[T(x_i) \leq t] \cdot I[x_i \in O_{d_N}(z_j)]}{R_N(i, z_j)},$$

where $R_N(i, z_j) = \sum_{l=1}^{N} I[T(x_l) \geq T(x_i)] \cdot I[x_l \in O_{d_N}(z_j)]$ is the number of observations in $O_{d_N}(z_j)$ with results no less that $T(x_i)$.

(ii) Suppose that function $B(x)$ is normed by such way, that

$$\frac{1}{K} \sum_{j=1}^{K} B(z_j) = 1. \tag{1}$$

Then $A(t) = \frac{1}{K} \sum_{j=1}^{K} L(t, z_j)$ and the estimator is directly obtained,

$$A_N(t) = \frac{1}{K} \sum_{j=1}^{K} L_N(t, z_j).$$

(iii) The value of function B at point $z \in \mathcal{X}$ can be estimated by the least squares method:

$$B_N(z) = \arg\min_B \sum^* \{L_N(T(x_i), z) - A_N(T(x_i)) \cdot B\}^2,$$

where $L_N(t, z)$ is the estimate of C.H.F. at z, by method (i), the sum $\sum^*$ is over $\{i : x_i \in \bigcup_{j=1}^{K} O_{d_N}(z_j) \bigcup O_{d_N}(z),\ T(x_i) \leq \mathcal{T}\}$.

Evidently, the properties of the cumulative hazard estimation are crucial for the quality of identification of our model. For the case without regression, good asymptotic properties of the estimator of the cumulative hazard function (or of the distribution function, respectively) were proved in [1] and [2]. In order to use the arguments of these proofs, we state several assumptions.

Assumption 1. Function $B(x)$ is positive and continuous on $\mathcal{X}$. Function $A(t)$ is nonnegative, nondecreasing and continuous on $[0, \mathcal{T}]$.

Assumption 2. Distribution function $G(t, x)$ is continuous in both arguments.

Now, as the procedure of step (i) resembles the kernel-type nonparametric estimation, some suppositions are inevitable about the design of variable x. Let us denote by $M_N(z)$ the number of x_i's in $O_{d_N}(z)$. Supposing $N \to \infty$ we choose $d_N \to 0$ such that $N \cdot d_N \to \infty$. When the design of x_i's is uniform sufficiently, $M_N(z)$ increases to infinity, too. Therefore our requirement is the following:

Assumption 3

$$0 < \liminf_{N\to\infty} M_N(z)/(N \cdot d_N) \leq \limsup_{N\to\infty} M_N(z)/(N \cdot d_N) < \infty$$

for every $z \in \mathcal{X}$.

Strong consistency of the cumulative hazard function estimation

Let z be a point from $\mathcal{X}$. When the number of realizations N tends to infinity, the width of neighbourhood $d_N \to 0$, but $d_N \cdot N \to \infty$. According to Assumption 3, the number of points in $O_{d_N}(z)$, $M = M_N(z)$ tends to infinity. In this section we try to follow the arguments of the proof of strong uniform consistency from [2]. There the product-limit estimator (PLE) of the survival function is considered, and the same properties shall be derived easily for our estimator of cumulative hazard function.

In our model, the random variables corresponding to x_i's from $O_{d_N}(z)$ are not distributed identically, but thanks to Assumptions 1, 2, their distributions differ only slightly for small d_N. At the same time, our distribution functions are supposed to be continuous in t, we need not care about the problems arising from discontinuities and ties, although such problems are not principal. In [2] they are

solved by means of Lemma A. Therefore Lemma A or its equivalent may be omitted here.

(I) Let us re-index the observations by such a way that $x_1, \ldots, x_M$ are all x_i's from $O_{d_N}(z)$ and $T(x_1) \leq T(x_2) \leq \ldots \leq T(x_M)$. Now, we can define the PL estimator of survival function $P(t, z)$ as

$$P_N(t,z) = \prod_{i=1}^{M} \left(\frac{M-i}{M-i+1} \right)^{\delta(x_i) I[T(x_i) \leq t]} \qquad \text{if } T(x_1) \leq t \leq T(x_M),$$

with $P_N(t,z) = 1$ for $t < T(x_1)$ and $P_N(t,z) = 0$ for $t \geq T(x_M)$.

(II) Let us divide the time interval $(0, \mathcal{T}]$ by means of the points $0 = \xi_0 < \xi_1 < \ldots < \xi_k = \mathcal{T}$ to subintervals $I_j = (\xi_{j-1}, \xi_j]$, $j = 1, \ldots, k$. Denote $p_j(x) = P\{Y(x) > \xi_j \mid Y(x) > \xi_{j-1}\}$. Then the estimate of "PL-type"

$$p_{N,j}(z) = \prod_{i=1}^{M} \left(\frac{M-i}{M-i+1} \right)^{\delta(x_i) I[T(x_i) \in I_j]}$$

is a logical estimate for $p_j(z)$ and

$$P_N(\xi_l, z) = \prod_{j=1}^{l} p_{N,j}(z).$$

Let us introduce several other variables-frequencies and theoretical probabilities of connected events. By "# $\{A\}$" we mean the number of elements in set A.

$$N_i = \#\{T(x_i).\xi_{j-1}\} \qquad \Pi_j^N(x) = P\{T(x) > \xi_{j-1}\}$$
$$D_j = \#\{T(x_i) \in I_j, \delta(x_i) = 1\} \qquad \Pi_j^D(x) = P\{T(x) \in I_j, \Pi_j^D \delta(x) = 1\}$$
$$W_j = \#\{T(x_i) \in I_j, \delta(x_i) = 0\} \qquad \Pi_j^W(x) = P\{T(x) \in I_j, \delta(x) = 0\}$$
$$A_j = 1 - D_j/(N_j - W_j) \qquad a_j(x) = 1 - \Pi_j^D(x) / \left(\Pi_j^N(x) - \Pi_j^W(x) \right)$$
$$B_j = 1 - D_j/N_j \qquad b_j(x) = 1 - \Pi_j^D(x) / \Pi_j^N(x)$$

$$i = 1, \ldots, M; \quad j = 1, \ldots, k; \quad x_i, x \in O_{d_N}(z).$$

(III) Let us remind again the second part of [2] —"Preliminary results" and several lemmas stated there.

Evidently, the following holds:

*Lemma B**

$$A_j \leq p_{N,j}(z) \leqq B_j.$$

Instead of Lemma C we can formulate a similar proposition:

*Lemma C**

Under Assumptions 1–3, almost surely

$$\lim_{n\to\infty} A_j = a_j(z) \text{ and } \lim_{N\to\infty} B_j = b_j(z), \text{ for all } j = 1, \ldots, k.$$

Proof. Probabilities $\Pi_j^N(x), \Pi_j^D(x), \Pi_j^W(x)$, as well as $a_j(x)$ and $b_j(x)$ are continuous functions of x. Variables D_j, W_j, N_j may be presented as the sums of independent binomial random variables, for instance

$$D_j = \sum_{i=1}^{M} e_j(x_i), \text{ where } e_j(x_i) = \begin{cases} 1, & \text{with probability } \Pi_j^D(x_i) \\ 0, & \text{with probability } 1 - \Pi_j^D(x_i). \end{cases}$$

Than SLLN guarantees that

$$\lim_{N\to\infty} \frac{1}{M} \left(D_j - \sum_{i=1}^{M} \Pi_j^D(x_1) \right) = 0 \text{ a.s.} \tag{2}$$

and similarly, for N_j and W_j, $j = 1, \ldots, k$.

At the same time $|x_i - z| \leq d_N \to 0$, therefore

$$\begin{aligned} a_j(z) &= \Pi_{j+1}^N(z) / \left(\Pi_j^N(z) - \Pi_j^W(z)\right) = \\ &= \lim \left[\Pi_{j+1}^N(x_i) / \left(\Pi_j^N(x_i) - \Pi_j^W(x_i)\right)\right]. \end{aligned} \tag{3}$$

Then

$$\lim(A_j - a_j(z)) = \lim \left(\frac{N_{j+1}}{N_j - W_j} - \frac{\sum \Pi_{j+1}^N(x_i)}{\sum \Pi_j^N(x_i) - \sum \Pi_j^W(x_i)} \right) = 0 \text{ a.s.}$$

The convergence of B_j to $b_j(z)$ could be proved similarly. □

(IV) Lemmas D and E of [2] deal with theoretical probabilities of certain events. From Lemma D we shall use the fact that

$$b_j(x) - a_j(x) \leq \frac{(F(\xi_j, x) - F(\xi_{j-1}, x)) \cdot (G(\xi_j, x) - G(\xi_{j-1}, x))}{P(\xi_j, x) \cdot Q(\xi_j, x)}, \tag{4}$$

from Lemma E the relation

$$a_j(x) \leq p_j(x) \leq b_j(x) \tag{5}$$

is important.

Now, our main result can be proposed and proved:

Theorem 1. Let Assumptions 1–3 hold and z be a point from $\mathcal{X}$. Then

$$\lim_{N\to\infty} \sup_{t\le T} |P_N(t,z) - P(t,z)| = 0 \text{ a.s.}$$

Proof. Let us choose $\varepsilon > 0$, denote $\delta = P(T,z) \cdot Q(T,z) > 0$. Points $\xi_1, \dots, \xi_{k-1}$ can be fixed by such a way that in every interval I_j $(j = 1, \dots, k)$ the variation of $F(t,z)$ is less than $\delta\varepsilon/2$.

From Lemmas B* and C* it follows, with probability one, that for N sufficiently large

$$a_j(z) - \frac{\varepsilon}{4k} \le p_{N,j}(z) \le b_j(z) + \frac{\varepsilon}{4k}, \quad j = 1, 2, \dots, k.$$

Together with (5) it leads to the relation

$$|p_{N,j}(z) - p_j(z)| \le b_j(z) - a_j(z) + \frac{\varepsilon}{2k}$$

and, with the help of (4), we get $b_j(z) - a_j(z) \le [G(\xi_j, z) - G(\xi_{j-1}, z)] \cdot \frac{\delta\varepsilon}{2} \cdot \frac{1}{\delta}$.

Summarizing all previous relations, for every ξ_j, $j = 1, \dots, k$, we obtain that

$$|P_N(\xi_j, z) - P(\xi_j, z)| = \left| \prod_{l\le j} p_{N.l}(z) - \prod_{l\le j} p_l(z) \right| \le$$

$$\le \sum_{l\le j} |p_{N,l}(z) - p_l(z)| \le \varepsilon.$$

Now, let us suppose that t lies between ξ_{j-1} and ξ_j. With probability one the following holds:

For fixed $e > 0$ there exists N_e such that as soon as $N > N_e$,

$$P(\xi_j, z) - e \le P_N(\xi_j, z) \le P_N(\xi_{j-1}, z) \le P(\xi_{j-1}, z) + e.$$

Therefore

$$|P_N(t,z) - P(t,z)| \le 2e + \frac{\varepsilon\delta}{2}. \qquad \square$$

We wish to prove the same result for our estimate of the cumulative hazard function. After re-indexing according to (I)

$$L_N(t,z) = \begin{cases} \displaystyle\sum_{i=1}^{M} \frac{\delta(x_i)}{M-i+1} \cdot I[T(x_i) \le t] & \text{for } t \ge T(x_1) \\ 0 & \text{for } t < T(x_1). \end{cases} \tag{6}$$

The connection between the estimates of survival function and of cumulative hazard function was shown also by Breslow and Crowley in [1]. The result is well known, we shall re-formulate it in order to obtain the statement suitable for us.

Lemma 1. Under Assumptions 1–3

$$\lim_{N\to\infty} \sup_{t\le\tau} \sqrt{M}|L_N(t,z) + \ln P_N(t,z)| = 0 \text{ a.s.}$$

Proof

$$-\ln P_N(t,z) = \sum_{i=1}^{M} \delta(x_i) I[T(x_i) \le t] \cdot \{\ln(M-i) - \ln(M-i+1)\} =$$

$$= \sum \delta(x_i) I[T(x_i) \le t] \cdot \int_{M-i}^{M-i+1} \frac{ds}{s} \cdot \sqrt{M}|L_N(t,z) + \ln P_N(t,z)| \le$$

$$\le \sqrt{M} \sum \delta(x_i) I[T(x_i) \le t] \left(\frac{1}{M-i} - \frac{1}{M-i+1} \right) \le$$

$$\le \sqrt{M} \sum_{i=1}^{M} \frac{I[T(x_i) \le \tau]}{(M-i+1)^2} \le \frac{\sqrt{M} \cdot M}{R_N(\tau)^2},$$

where

$$R_N(\tau) = \#\{T(x_i) > \tau,\ x_i \in O_{d_N}(z)\}.$$

Evidently, $\frac{R_N(\tau)}{M} \xrightarrow{\text{a.s.}} P(\tau,z) \cdot Q(\tau,z) = \delta > 0$, it follows from the SLLN and from the continuity of $P(t,x)$ and $Q(t,x)$. That is why

$$\frac{\sqrt{M} \cdot M}{R_N(\tau)^2} = O\left(M^{-\frac{1}{2}}\right) \text{ a.s.} \qquad \square$$

Corollary 1. Under Assumptions 1–3, for arbitrary $z \in \mathcal{X}$,

$$\lim_{N\to\infty} \sup_{t\le\tau} |L_N(t,z) - L(t,z) = 0 \text{ a.s.}$$

Remark. Assumptions 1, 2 about the continuity in variable x are utilized maximally. It is possible that a more thorough analysis of the problem enables us to weaken this assumption. But later, when proving the asymptotical normality, among others, we demand the existence of

$$\lim_{N\to\infty} \frac{1}{M} \sum_{i=1}^{M} Q(t, x_i)$$

and this assumption is not significantly weaker.

Asymptotic normality of the estimates

Let us consider the random sample, described in (I) of the previous part, with $M = M_N(z) \to \infty$ and $d_N \to 0$. Then the estimate of C.H.F. is given by (6). In Chapter 7 of [1] the same problem was solved, but for the case of identically distributed variables (i.e. without regression). In order to follow the arguments of [1] we have to extend our assumptions:

Assumption 4. Let functions $B(x)$ and $G(t,x)$ be Lipschitz-continuous as to variable $x \in \mathcal{X}$.

Assumption 5. Let d_N be such that $\sqrt{M_N(z)} \cdot d_N \to 0$.

The C.H.F. at point z can be rewritten as

$$L(t,z) = \int_0^t \frac{d\tilde{F}(s,z)}{1 - H(s,z)},$$

where

$$\tilde{F}(s,z) = P\{T(z) < s,\ \delta(z) = 1\} = \int_0^s Q(u,z)dF(u,z),$$

$$H(s,z) = P\{T(z) < s\} = 1 - P(s,z)Q(s,z).$$

Both these characteristics can be estimated empirically by means of relative frequencies in the sample $\{T(x_i),\ \delta(x_i) :\ i = 1,\dots,M\}$

$$\tilde{F}_N(s,z) = \frac{1}{M}\#\{T(x_i) < s,\ \delta(x_i) = 1\},$$

$$H_N(s,z) = \frac{1}{M}\#\{T(x_i) < s\}.$$

Now, define random process $(\nu_N(t),\ \varrho_N(t))$ on $(0, \mathcal{T}]$ by

$$\nu_N(t) = \sqrt{M}(H_N(t,z) - H(t,z)), \quad \varrho_N(t) = \sqrt{M}(\tilde{F}_N(t,z) - \tilde{F}(t,z)).$$

Theorem 2. Let Assumptions 1–5 be fulfilled. Then bivariate random process $(\nu_N(t),\ \varrho_N(t))$ on $(0, \mathcal{T}]$ converges weakly to a bivariate Gaussian process $(\nu(t),\ \varrho(t))$ having a mean 0 and a covariance structure given for $0 \le s \le t \le \mathcal{T}$ by

$$\begin{aligned}
&\operatorname{cov}(\nu(s),\nu(t)) = H(s,z)(1 - H(t,z)),\\
&\operatorname{cov}(\varrho(s),\varrho(t)) = \tilde{F}(s,z)(1 - \tilde{F}(t,z)),\\
&\operatorname{cov}(\nu(s),\varrho(t)) = \tilde{F}(s,z) - H(s,z)\tilde{F}(t,z),\\
&\operatorname{cov}(\nu(t),\varrho(s)) = \tilde{F}(s,z)(1 - H(t,z)).
\end{aligned} \tag{7}$$

Proof. Again, let us suppose that the empirical variables are expressed by normed sums of independent binomial random variables:

$$H_N(t,z) = \frac{1}{M}\sum_{i=1}^{M} \varepsilon(t,x_i), \quad \tilde{F}_N(t,z) = \frac{1}{M}\sum_{i=1}^{M} \tilde{\varepsilon}(t,x_i),$$

where $P\{\varepsilon(t,x_i)=1\} = H(t,x_i), \quad P\{\tilde{\varepsilon}(t,x_i)=1\} = \tilde{F}(t,x_i)$.

Then the bivariate random process given by $v_N(t) = \sqrt{M}(H_N(t,z) - \frac{1}{M}\sum H(t,x_i))$, $r_N(t) = \sqrt{M}(\tilde{F}_N(t,z) - \frac{1}{M}\sum \tilde{F}(t,x_i))$ has a mean of zero and a covariance structure given by the covariances of random variables ε and $\tilde{\varepsilon}$. For instance, for $0 \le s \le t \le \tau$

$$\operatorname{cov}(v_N(s), r_N(t)) = \frac{1}{M}\sum \operatorname{cov}(\varepsilon(s,x_i), \tilde{\varepsilon}(t,x_i)) =$$

$$= \frac{1}{M}\sum (\tilde{F}(s,x_i) - H(s,x_i)\tilde{F}(t,x_i)).$$

Obviously, under Assumptions 1–3 the covariances tend to covariances (7) of $(\nu(t), \varrho(t))$ and the C.L. Theorem ensures that the process $(\nu_N(t), r_N(t))$ is asymptotically Gaussian on $(0, \tau]$.

To finish the proof, we only have to show that $\sqrt{M}(\frac{1}{M}\sum H(t,x_i) - H(t,z)) \to 0$ (and the same with $\tilde{F}$). But when Assumption 4 holds, even functions $H(t,x)$ and $\tilde{F}(t,x)$ are Lipschitz-continuous in x and, therefore, a positive constant c_t exists that

$$\sqrt{M}\left|\frac{1}{M}\sum H(t,x_i) - H(t,z)\right| \le \sqrt{M}\cdot c_t \cdot d_N.$$

The last term tends to zero, it follows from Assumption 5. □

The estimate of C.H.F. can be rewritten as $L_N(t,z) = \int_0^t (1 - H_N(s,z))^{-1} \cdot d\tilde{F}_N(s,z)$. Similarly as in [1], we then have the expansion

$$\sqrt{M}(L_N(t,z) - L(t,z)) = A_N + B_N + R_{1N} + R_{2N},$$

where

$$A_N(t) = \int_0^t \nu_N \cdot (1-H)^{-2} d\tilde{F}$$

$$B_N(t) = \varrho_N(t)\cdot(1 - H(t,z))^{-1} - \int_0^t \varrho_N \cdot (1-H)^{-2} dH$$

$$R_{1N}(t) = M^{-\frac{1}{2}} \int_0^t \nu_N^2 (1-H)^{-2}(1-H_N)^{-1} d\tilde{F}$$

$$R_{2N}(t) = \int_0^t \nu_N (1-H)^{-1}(1-H_N)^{-1} d(\tilde{F}_N - \tilde{F}).$$

Theorem 3. Let Assumptions 1–5 hold. Then the random function $\sqrt{M}(L_N(t,z) - L(t,z))$ on $(0, \tau)$ converges weakly to the Gaussian process Z defined by

$$Z(t) = \int_0^t \nu(1-H)^{-2} d\tilde{F} + \varrho(t)(1-H(t,z))^{-1} - \int_0^t \varrho(1-H)^{-2} dH.$$

This process has zero mean and its covariance function is, for $0 \leq s \leq t \leq \tau$, given by

$$\operatorname{cov}(Z(s), Z(t)) = C(s,z) = \int_0^s (1-H)^{-2} d\tilde{F} = \int_0^s (1-H)^{-1} P^{-1} dF.$$

Proof. We could follow step by step the proof of Theorem 4 in [1], part 7.

Our next aim is to estimate the function $A(t)$ and the values $B(z_j)$, provided they are tied by condition (1).

Corollary 2

(a) Let Assumptions 1–3 hold, then $A_N(t) = \dfrac{1}{K} \sum_{j=1}^{K} L_N(t, z_j)$ is a strongly consistent estimator of $A(t)$, uniformly in $t \in [0, \tau]$.

(b) Let Assumptions 1–5 hold, suppose that $I_j = \lim_{N \to \infty} M_N(z_j)/M_N^*$ exist, with $M_N^* = \sum_{j=1}^{K} M_N(z_j)$. Then $W_N(t) = \sqrt{M_N^*}\{A_N(t) - A(t)\}$ on $(0, \tau]$ converges weakly to the Gaussian random function $W(t)$, which has zero mean and a covariance structure for $0 \leq s \leq t \leq \tau$ given by

$$\operatorname{cov}(W(s), W(t)) = \frac{1}{K^2} \sum_{j=1}^{K} C(s, z_j)/I_j.$$

Remark. From Assumption 3 we may deduce that the limits I_j's are greater than zero.

Finally, the estimation of $B(z)$ by the least squares method yields

$$B_N(z) = \sum^* L_N(T_i, z) \cdot A_N(T_i) / \sum^* A_N^2(T_i),$$

where T_i stands for $T(x_i)$ and the sums $\sum^*$ have the same meaning as in (iii).

Corollary 3. The estimates $B_N(z)$ are strongly consistent as soon as Assumptions 1–3 are fulfilled.

Numerous papers were dealing with the large sample properties of PLE as well as the smoothing techniques of curve estimation. It could represent an alternative source for the proofs of the propositions presented by our contribution. [3], [4] should be mentioned as examples of the papers solving the above problems.

The author is indebted to the referees for their contructive suggestions concerning especially with the alternative approaches to the problem.

References

1. *Breslow, N., Crowley, J.*, A large sample study of the life table and product limit estimates under random censorship. Annals of Statist. **2** (1974), 437–453.
2. *Winter, B. B., Földes, A., Rejtő, L.*, Glivenko–Cantelli theorems for the product limit estimate. Problems of Control and Inform. Theory **7** (1978), 213–225.
3. *Bickel, P. J., Breiman, L.*, Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit tests. Annals of Probab. **11** (1983), 185–214.
4. *Gill, R.D.*, Large sample behaviour of the product-limit estimator on the whole line. Annals of Statist. **11** (1983), 49–58.

Непараметрический анализ регрессии в модели с пропорциональной интенсивностью отказов

П. ВОЛЬФ

(Прага)

Предполагается модель с пропорциональной интенсивностью отказов для зависимости времени жизни на предсказывающей переменной x. Тогда кумулятивная интенсивность отказов выражена произведением $L(t, x) = A(t) \cdot B(x)$. Время жизни исследовано

с цензурированием. Предлагаются оценки нормированной функции $B(x)$ и функции $A(t)$. Доказана их сильная состоятельность и асимптотическая нормальность.

P. Volf
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Prague 8, Czechoslovakia

Problems of Control and Information Theory, Vol. 18(5), pp. 323–338 (1989)

ROBUST CONTROLLER DESIGN FOR SYSTEMS WITH INTERVAL PARAMETER DESIGN

C.J. GOH, C.C. LIM, K.L. TEO, D.J. CLEMENTS

(*Nedlands, Adelaide, Nedlands, Kensington*)

(Received July 20, 1988)

A computational technique to design a robust controller for systems with interval variable plant parameters is proposed. The well-known result of Kharitonov is used to transform the design problem into a nonlinearly constrained mathematical programming problem. Two examples of moderate size are computed to illustrate the feasibility of the technique.

Introduction

Since the control engineering community has become aware of the work of Kharitonov [1], there has been considerable interest in applying this result to various problems in analysis and design. For example, see [2, 3, 4, 5, 6, 7] and the references cited therein. Kharitonov's result shows that the stability, with respect to the open left half s-plane, of a class of polynomials with coefficients independently variable over a finite interval is equivalent to the stability, with respect to the open left half s-plane, of just four carefully chosen polynomials from within the class. Further analysis [6] has shown that for low order systems, not all the Kharitonov polynomials are required to be considered for stability. To be more precise, if $\overline{n}$ is the degree of the monic closed loop polynomial, then only $\overline{n} - 2$ (for $\overline{n} = 3, 4, 5$) specific Kharitonov polynomials are required to be considered. If $\overline{n} \geq 6$, it is shown that all four polynomials are necessary.

Kharitonov's result as it stands is not completely satisfactory. First, it requires that the coefficients of the polynomials are independently variable. If coefficients depend (often linearly) on another smaller set of independently variable parameters, then, in these situations, the Kharitonov condition is only sufficient for stability with respect to the open left s-plane. Second, in practice, one is often interested in stability with respect to a region to satisfy certain design specifications other than the open left half plane. In this case, Kharitonov's condition is only necessary for stability.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

A recent result of Bartlett et al. [8] reconciles both of the above problems by showing that the stability, with respect to a convex region, of a class of polynomials with coefficients linear in a set of independent variables is equivalent to the stability, with respect to the same convex region, of a certain subclass of those polynomials. This is a nice theoretical result but reduces computationally to 2^n root locus problems where n is the number of independent variables. For small n, this is acceptable for purposes of analysis, but it is difficult to see how this result might be incorporated directly into a design procedure.

Thus, for the purpose of realistic control system design, it appears that Kharitonov's condition is still the best option available.

In this paper we present a numerical technique for satisfying Kharitonov's condition as a part of a computational procedure to design a stabilizing controller. Essentially, the Kharitonov polynomials are being expanded in their eigenvalue structure forms. By treating these eigenvalues as design variables, and desiring them to be as close as possible to some nominal design values, a nonlinearly constrained mathematical programming problem can be formulated. The optimization is subject to constraints on the stability criteria on the relevant eigenvalue to ensure overall robustness. Numerical experiences have indicated that the technique is both computationally feasible and adaptable to a variety of related problems. We shall present two numerical examples to verify the validity of the design. The first example was previously computed by Soh et al. [2] and the second is of larger dimension.

Problem formulation

Consider a linear SISO plant characterised by

$$A(s)y - B(s)u \tag{1a}$$

where

$$A(s) = \sum_{i=0}^{n} a_i s^i, \; a_n = 1 \tag{1b}$$

and

$$B(s) = \sum_{i=0}^{n-1} b_i s^i \tag{1c}$$

are polynomials of order n and $n-1$, respectively. The plant is assumed causal and controllable.

The aim of this paper is to present a technique for finding a robust controller for a plant whose dynamics are subject to variation such that the corresponding closed-loop system is not only stable but also having its resultant poles made close

to its assigned nominal locations. As the plant dynamics are known only within certain degree of certainty and they may vary due to changes in operating condition, the variation can be modelled by expressing the plant parameters to be within certain specified intervals. The robust controller design can now be performed by finding a set of controller coefficients that will satisfy the stability requirements for these plant parameteres varying within their respective intervals.

To state the design concept formally, let us specify the structure of the controller in a general form as

$$P(s)u - T(s)y = 0 \tag{2a}$$

where

$$P(s) = \sum_{i=0}^{m} p_i s^i, \; p_m = 1 \tag{2b}$$

$$T(s) = \sum_{i=0}^{m} \tau_i s^i \tag{2c}$$

are polynomials of order $m(< n)$.

The closed-loop characteristic polynomial is

$$D(s) = A(s)P(s) + B(s)T(s). \tag{3}$$

From (1b)–(1c) and (2b)–(2c), it follows that

$$D(s) = s^{m+n} + \sum_{k=0}^{m+n-1} d_k s^k \tag{4a}$$

where

$$d_k = \sum_{i+j=k} (p_i a_j + \tau_i b_j). \tag{4b}$$

Let

$$\mathbf{a}' = [a_0, a_1, \ldots, a_n], \quad \mathbf{b}' = [b_0, b_1, \ldots, b_{n-1}],$$

$$\boldsymbol{\tau}' = [\tau_0, \tau_1, \ldots, \tau_m], \quad \mathbf{p}' = [p_0, p_1, \ldots p_m],$$

and

$$\mathbf{d}' = [d_0, d_1, \ldots, d_{m+n-1}, d_{m+n}].$$

Clearly, $d_{m+n} = 1$.

Denote $\mathbf{v}' = [\mathbf{a}', \mathbf{b}']$ as the plant parameter vector and

$\mathbf{x}' = [\boldsymbol{\tau}', \mathbf{p}'] \in R^{2m+1}$ as the controller parameter vector.

To stabilize the closed-loop system characterized by (4), we need to determine a controller parameter vector $\mathbf{x}$ so that the poles of the corresponding closed-loop

polynomial are stable. There already exist methods in the literature which can be used to solve this well-knovn pole-placement problem [9, 10].

When the plant parameter vector $\mathbf{v}$ is uncertain, the controller vector $\mathbf{x}$ obtained from equation (4) may not be able to stabilize the closed-loop system for a small perturbation of $\mathbf{v}$ from its nominal value. Here a method which can be used to design a controller is developed for the system when v is not exactly known, but known to lie within an interval $[\mathbf{v}^-, \mathbf{v}^+]$, i.e.,

$$a_i^- \leq a_i \leq a_i^+, \quad i = 0, 1, \ldots, n,$$

and

$$b_i^- \leq b_i \leq b_i^+, \quad i = 0, 1, \ldots, n-1$$

where a_i^-, a_i^+, $i = 0, 1, \ldots, n$ and b_i^-, b_i^+, $i = 0, 1, \ldots, n-1$, are given fixed constants representing the upper and lower bounds to the uncertain parameters. From (4) it is easy to show that the corresponding set of closed-loop polynomials $D(s)$ is characterised by

$$D(s) = s^{m+n} + \sum_{k=0}^{m+n-1} d_k s^k \tag{5a}$$

with

$$d_k \in [d_k^-, \ d_k^+] \tag{5b}$$

where

$$d_k^- = \sum_{i+j=k} \{\min(p_i a_j^-, \ p_i a_j^+) + \min(\tau_i b_j^-, \ \tau_i b_j^+)\} \tag{5c}$$

and

$$d_k^+ = \sum_{i+j=k} \{\max(p_i a_j^-, \ p_i a_j^+) + \max(\tau_i b_j^-, \ \tau_i b_j^+)\}. \tag{5d}$$

To continue, the nominal plant parameter vector $\mathbf{v}^0$ is taken as the mid-point of its interval, i.e.

$$\mathbf{v}^0 = \frac{1}{2}[\mathbf{v}^- + \mathbf{v}^+].$$

Furthermore, we also define a nominal controller parameter vector $\mathbf{x}^0$ which may or may not result in an overall stable closed-loop system. More likely it will not by virtue of the uncertain plant parameters. However, the final controller is desired to be close to this nominal controller while preserving closed-loop stability.

It should be pointed out, however, that this design criteria is only a nominal one so as to illustrate a general approach to a difficult problem. In practice, other more relevant criteria may be considered in the same spirit.

The robust pole-placement problem can thus be formally stated as follows:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}^0\|^2 \tag{6}$$

subject to $\mathbf{x} \in \vartheta$, where

$$\|\mathbf{x} - \mathbf{x}^0\|^2 = \sum_{i=0}^{2m} (x_i - x_i^0)^2$$

and the set ϑ consists of all those controller vectors $\mathbf{x} \in R^{2m+1}$ such that the corresponding zeroes of the closed-loop polynomial $D(s)$ are in the left hand side of the complex s-plane for all prescribed uncertainties in the plant parameter vector $\mathbf{v} \in [\mathbf{v}^-, \mathbf{v}^+]$ or equivalently for all $d_k \in [d_k^-, d_k^+]$, $k = 0, 1, \ldots, n + m - 1$.

Robust controller design

The aim of this section is to formulate the optimization problem (6) into a nonlinearly constrained optimization problem. To begin, let us introduce some notation.

Let

$$\mathbf{K}' = [1,\ 1,\ -1,\ -1,\ 1,\ 1, \ldots] \in N^{m+n} \tag{7}$$

where

$$K_i = (-1)^{\text{Int}\left(\frac{i+1}{2}\right)-1}. \tag{8}$$

Then, consider the following four polynomials (Kharitonov polynomials):

$$D_j(s) = \sum_{i=0}^{m+n} d_{ji} s^i, \quad j = 1, 2, 3, 4, \tag{9a}$$

where, for each $i = 0, 1, \ldots, m + n - 1$ and $j = 1, 2, 3, 4$.

$$d_{ij} = \begin{cases} 1 & \text{if } i = m + n \\ d_i^- & \text{if } K_{i+j} = 1 \\ d_i^+ & \text{if } K_{i+j} = -1. \end{cases} \tag{9b}$$

Theorem 1. Let Ξ be the set consisting of all those controller vectors $\mathbf{x} \in R^{2m+1}$ such that the corresponding zeroes of the above four Kharitonov polynomials $D_j(s)$, $j = 1, 2, 3, 4$ are on the left side of the complex s-plane. Then $\Xi = \vartheta$, where ϑ is defined in the optimization problem (6).

Proof. The proof of this theorem follows obviously from the Kharitonov theorem (cf. Ref. [1]).

Remark. We shall, for generality, assume that $m + n \geq 6$ for all subsequent analysis. For $m + n \leq 5$, the problem may be simplified accordingly using the result of [6].

For computatioal purposes, we shall refer to the Kharitonov polynomials as appeared in (9) as in the "natural" form. One can also express the same polynomial in a "structural" form which naturally illuminates the eigen-structure of the polynomials [11].

For each $j = 1, \ldots, 4$, the structural form of the Kharitonov polynomial can be appropriately defined as [11]

$$Q_j(s;\beta) = \prod_{k=1}^{\frac{m+n}{2}} \{s^2 + s(\varrho_{k1}^j + \varrho_{k2}^j) + [\varrho_{k1}^j \varrho_{k2}^j + (\sigma_k^j)^2]\} \tag{10}$$

when $m+n$ is even

$$Q_j(s;\beta) = (s+\lambda_j) \prod_{k=1}^{\frac{m+n-1}{2}} \{s^2 + s(\varrho_{k1}^j + \varrho_{k2}^j) + [\varrho_{k1}^j \varrho_{k2}^j + (\sigma_k^j)^2]\} \tag{11}$$

when $m+n$ is odd.

Let

$$\lambda' = [\lambda_1, \ldots, \lambda_4] \tag{12}$$

$$(\sigma^j)' = [\sigma_1^j, \sigma_2^j, \ldots, \sigma_{\hat{n}}^j] \tag{13}$$

where

$$\hat{n} = \begin{cases} \dfrac{m+n}{2} & \text{if } m+n \text{ is even} \\ \dfrac{m+n-1}{2} & \text{if } m+n \text{ is odd} \end{cases}$$

$$(\varrho^j)' = [\varrho_{11}^j, \varrho_{12}^j, \varrho_{21}^j, \varrho_{22}^j, \cdots \varrho_{\hat{n}1}^j, \varrho_{\hat{n}2}^j] \tag{14}$$

and

$$\beta' = [\mathbf{p}', \tau', (\varrho^1)', \ldots, (\varrho^4)', (\sigma^1)', \ldots, (\sigma^4)', \lambda']. \tag{15}$$

Theorem 2. For $m+n$ even, if

$$h_{jk} = \sigma_k^j(\varrho_{k2}^j - \varrho_{k1}^j) = 0, \quad \begin{matrix} j = 1, \ldots, 4 \\ k = 1, \ldots, \hat{n} = \dfrac{m+n}{2} \end{matrix} \tag{16}$$

and

$$g_{ij} = D_j(s) - Q_j(s_i;\beta) = 0, \quad \begin{matrix} j = 1, \ldots, 4 \\ i = 1, \ldots, m+n+1, \end{matrix} \tag{17}$$

where s_i, $i = 1, \ldots, m+n+1$, are distinct points in R and $D_j(s)$ is defined by (9), then it is necessary and sufficient that

$$D_j(s) = Q_j(s;\beta) \quad \forall s \in R. \tag{18}$$

Proof. The sufficiency part of the theorem is obvious. Thus, we need only to prove the necessary part. Since for each $j = 1, \ldots, 4$, $D_j(s)$ is a polynomial of even degree $m+n$, there must be $m+n$ zeroes. For these $m+n$ zeroes, some may appear as complex conjugate pairs, and others as real zeroes. In either case, a pair of complex conjugate pair or a pair of real zeroes can both be expressed as

$$\begin{aligned}[s + (\varrho_{k1}^j + i\sigma_k^j)][s + (\varrho_{k2}^j + i\sigma_k^j)] = \\ = s^2 + s[\varrho_{k1}^j + \varrho_{k2}^j] + [\varrho_{k1}^j \varrho_{k2}^j + (\sigma_k^j)^2] + \\ + i\sigma_k^j[\varrho_{k1}^j - \varrho_{k2}^j].\end{aligned} \tag{19}$$

Here $i = \sqrt{-1}$.

Since $D_j(s)$ must be a real-valued function of s, (16) must hold for (19) to be real. Consequently, a product of all these pairs of factors will constitute the structural form of (10). Furthermore, it is well known that two polynomials of degree $m+n$ are identical for all s in R if they are equal at $m+n+1$ distinct points. It thus follows that (18) is equivalent to (17), and the proof is complete.

For the case of odd $m+n$, at least one of the zeroes must be real which we singled out as the factor $(s+\lambda_j)$ in (15). By using a similar argument as that given for Theorem 2, the following theorem must also hold.

Theorem 3. For $m+n$ odd, if

$$\begin{aligned}\bar{h}_{jk} = \sigma_k^j(\varrho_{k2}^j - \varrho_{k1}^j) = 0, \quad & j = 1, \ldots, 4 \\ & k = 1, \ldots, \hat{n} = \frac{m+n-1}{2}\end{aligned} \tag{20}$$

and

$$\begin{aligned}\bar{g}_{ij} = D_j(s;\beta) - \bar{Q}(s_i;\beta) = 0, \quad & j = 1, \ldots, 4 \\ & i = 1, \ldots, m+n+1\end{aligned} \tag{21}$$

where s_i, $i = 1, \ldots, m+n+1$ are distinct points in R and $D_j(s_i)$ is defined by (9). Then

$$D_j(s) = \bar{Q}_j(s;\beta) \quad \forall s \in R. \tag{22}$$

By Theorem 1 and Theorem 2 (respectively Theorem 3), we can easily transform the optimization problem (6) with $m+n$ being even (respectively being odd) into a linearly constrained optimization. Since both cases can be handled similarly, we shall only consider the case when $m+n$ is odd. The result is given in the following theorem.

Theorem 4. Let $m+n$ be odd, then, the optimization problem (6) is equivalent to:

$$\min J = \|\mathbf{x} - \mathbf{x}^0\|^2 \tag{23a}$$

subject to

$$\varrho_{k\nu}^{j} \geq \varepsilon,\ k = 1, \ldots, \hat{n} = \frac{m+n-1}{2};\ \nu = 1, 2;\ j = 1, \ldots, 4 \tag{23b}$$

where ε is some arbitrarily small positive real to ensure stability

$$\overline{h}_{jk}(\boldsymbol{\beta}) = 0,\ j = 1, 2, 3, 4;\ k = 1, \ldots, \hat{n} = \frac{m+n-1}{2} \tag{23c}$$

$$\overline{g}_{ij}(\boldsymbol{\beta}) = D_j(s_i) - Q_j(s_i; \boldsymbol{\beta}) = 0,\ j = 1, \ldots, 4; \tag{23d}$$

$$i = 1, \ldots, m+n+1.$$

Solution procedure

From the previous section, we recall that the optimization problem (6) is equivalent to the nonlinearly constrained optimization problem (23) which, in turn, can be solved by the software package NLPQL (cf. Ref. [12]). Just as any other mathematical programming algorithm, we need to calculate the gradients of the objective function (23a) and the constraints (23b)–(23d). For brevity, we shall only present the gradient for the constraint (23d), as the gradients for the objective function and the other constraints follow in a straightforward manner.

We have, for $\nu = 1, \ldots, m+n+1;\ j = 1, \ldots, 4;\ k = 1, \ldots, \hat{n}$

$$\begin{aligned} \overline{g}_{\nu j}(\boldsymbol{\beta}) = D_j(s_\nu) - \overline{Q}_j(s_\nu; \boldsymbol{\beta}) = \sum_{i=0}^{m+n} d_{ji} s_\nu^i - \\ -(s_\nu + \lambda_j) \prod_{k=1}^{\hat{n}} \{s_\nu^2 + s_\nu(\varrho_{k1}^{j} + \varrho_{k2}^{j}) + [\varrho_{k1}^{j}\varrho_{k2}^{j} + (\sigma_k^j)^2]\}. \end{aligned} \tag{24a}$$

The gradients with respect to λ_j, ϱ_{k1}^{j}, ϱ_{k2}^{j} and σ_k^j can be readily shown to be:

$$\frac{\partial \overline{g}_{\nu j}(\boldsymbol{\beta})}{\partial \lambda_j} = -\prod_{k=1}^{\hat{n}} \{s_\nu^2 + s_\nu(\varrho_{k1}^{j} + \varrho_{k2}^{j}) + (\varrho_{k1}^{j}\varrho_{k2}^{j} + (\sigma_k^j)^2)\} \tag{24b}$$

$$\begin{aligned} \frac{\partial \overline{g}_{\nu j}(\boldsymbol{\beta})}{\partial \varrho_{k1}^{j}} = -(s_\nu + \lambda_j)(s_\nu + \varrho_{k2}^{j}) \cdot \\ \cdot \prod_{k' \neq k} \left[s_\nu^2 + s_\nu(\varrho_{k'1}^{j} + \varrho_{k'2}^{j}) + (\varrho_{k'1}^{j}, \varrho_{k'2}^{j} + (\sigma_{k'}^{j})^2)\right] \end{aligned} \tag{24c}$$

$$\frac{\partial \bar{g}_{\nu j}(\boldsymbol{\beta})}{\partial \varrho^j_{k2}} = -(s_\nu + \lambda_j)(s_\nu + \varrho^j_{k1}) \cdot \cdot \prod_{k' \neq k} \left[s_\nu^2 + s_\nu(\varrho^j_{k'1} + \varrho^j_{k'2}) + (\varrho^j_{k'1}\varrho^j_{k'2} + (\sigma^j_{k'})^2) \right] \tag{24d}$$

$$\frac{\partial \bar{g}_{\nu j}(\boldsymbol{\beta})}{\partial \sigma^j_k} = -2\sigma^j_k(s_\nu + \lambda_j) \cdot \cdot \prod_{k' \neq k} \left[s_\nu^2 + s_\nu(\varrho^j_{k'1} + \varrho^j_{k'2}) + (\varrho^j_{k'1}, \varrho^j_{k'2} + (\sigma^j_{k'})^2) \right]. \tag{24e}$$

The gradients with respect to the controller parameters are somewhat more tricky. As the coefficients to the Kharitonov polynomials (9) are non-differentiable functions of the controller parameters ((5c) and (5d)), the problem is, strictly speaking, a non-smooth optimization problem, the point of singularity being at the origin where the derivative is discontinuous. To allow solution in the present framework, we have to introduce a concept of enforced smoothing which will be taken up in the next section. Elsewhere, apart from the singular point, the gradient exists and can be shown to be

$$\frac{\partial \bar{g}_{\nu j}(\boldsymbol{\beta})}{\partial p_k} = \sum_{i=0}^{m+n-1} \frac{\partial d_{ji}}{\partial p_k} s^i_\nu \qquad \begin{matrix} \nu = 1, \ldots, m+n+1 \\ j = 1, \ldots, 4 \\ k = 0, \ldots, m-1 \end{matrix} \tag{24f}$$

$$\frac{\partial \bar{g}_{\nu j}(\boldsymbol{\beta})}{\partial \tau_k} = \sum_{i=0}^{m+n-1} \frac{\partial d_{ji}}{\partial \tau_k} s^i_\nu \qquad \begin{matrix} \nu = 1, \ldots, m+n+1 \\ j = 1, \ldots, 4 \\ k = 0, \ldots, m-1 \end{matrix} \tag{24g}$$

$$\frac{\partial d_{j\nu}}{\partial p_i} = \begin{cases} a^-_{\nu-i} & \text{if } K_{\nu+j} p_i > 0 \\ a^+_{\nu-i} & \text{if } K_{\nu+j} p_i < 0 \end{cases} \tag{24h}$$

$$\frac{\partial d_{j\nu}}{\partial \tau_i} = \begin{cases} b^-_{\nu-i} & \text{if } K_{\nu+j} \tau_i > 0 \\ b^+_{\nu-i} & \text{if } K_{\nu+j} \tau_i < 0. \end{cases} \tag{24i}$$

Enforced smoothing

Recall that the coefficients of the polynomials D_j, $j = 1, \ldots, 4$, are given by (9b) where

$$d^-_\nu = \sum_{i+j=\nu} \{\min(p_i a^-_j, p_i a^+_j) + \min(\tau_i b^-_j, \tau_i b^+_j)\} \tag{25a}$$

$$d_\nu^+ = \sum_{i+j=\nu} \{\max(p_i a_j^-, p_i a_j^+) + \max(\tau_i b_j^-, \tau_i b_j^+)\} \tag{25b}$$

Thus, strictly speaking, the optimization problem (23) is a nonsmooth one. In spite of this, the NLPQL software package is found to work well in solving this problem. For a more rigorous justification, we shall nonetheless propose a method based on the concept of enforced smoothing. To begin, we define

$$H_j(p_i) = \min(p_i a_j^-, p_i a_j^+) \tag{26a}$$

$$K_j(\tau_i) = \min(\tau_i b_j^-, \tau_i b_j^+) \tag{26b}$$

$$M_j(p_i) = \max(p_i a_j^-, p_i a_j^+) \tag{26c}$$

$$N_j(\tau_i) = \max(\tau_i b_j^-, \tau_i b_j^+) \tag{26d}$$

In view of these nonsmooth functions defined above, we see that the discontinuities only occur when $p_i = 0$ or $\tau_i = 0$.

For any $\varepsilon > 0$, define

$$\hat{H}_j^\varepsilon(p_i) = \frac{1}{2\varepsilon}\{a_j^-(p_i - \frac{\varepsilon}{2})^2 + a_j^+(p_i + \frac{\varepsilon}{2})^2\} \tag{27a}$$

$$\hat{K}_j^\varepsilon(\tau_i) = \frac{1}{2\varepsilon}\{b_j^(\tau_i - \frac{\varepsilon}{2})^2 + b_j^+(\tau_i + \frac{\varepsilon}{2})^2\} \tag{27b}$$

$$\hat{M}_j^\varepsilon(p_i) = \frac{1}{2\varepsilon}\{a_j^+(p_i - \frac{\varepsilon}{2})^2 + a_j^-(p_i + \frac{\varepsilon}{2})^2\} \tag{27c}$$

$$\hat{N}_j^\varepsilon(\tau_i) = \frac{1}{2\varepsilon}\{b_j^+(\tau_i - \frac{\varepsilon}{2}) + b_j^-(\tau_i + \frac{\varepsilon}{2})^2\}. \tag{27d}$$

Then for each i, j and $\varepsilon > 0$, we consider the function

$$H_j^\varepsilon(p_i) = \begin{cases} H_j(p_i) & \text{if } p_i \le -\dfrac{\varepsilon}{2} \\ \hat{H}_j^\varepsilon(p_i) & \text{if } -\dfrac{\varepsilon}{2} \le p_i \le \dfrac{\varepsilon}{2} \\ H_j(p_i) & \text{if } p_i \ge \dfrac{\varepsilon}{2}. \end{cases} \tag{28}$$

The functions $K_j^\varepsilon(\tau_i)$, $M_j^\varepsilon(p_i)$ and $N_j^\varepsilon(\tau_i)$ are defined similarly. Clearly, for each i, j

$$|H_j^\varepsilon(p_i) - H_j(p_i)| \le \frac{\varepsilon}{2}(a_j^+ + a_j^-) \tag{29a}$$

$$|M_j^\varepsilon(p_i) - M_j(p_i)| \le \frac{\varepsilon}{2}(a_j^+ + a_j^-) \tag{29b}$$

$$|K_j^\varepsilon(\tau_i) - K_j(\tau_i)| \leq \frac{\varepsilon}{2}(b_j^+ + b_j^-) \tag{29c}$$

$$|N_j^\varepsilon(\tau_i) - N_j(\tau_i)| \leq \frac{\varepsilon}{2}(b_j^+ + b_j^-). \tag{29d}$$

Next, let

$$d_{\nu,\varepsilon}^- = \sum_{i+j=\nu} \{H_j^\varepsilon(p_i) + K_j^\varepsilon(\tau_i)\} \tag{30a}$$

$$d_{\nu,\varepsilon}^+ = \sum_{i+j=\nu} \{M_j^\varepsilon(p_i) + N_j^\varepsilon(\tau_i)\} \tag{30b}$$

$$D_j^\varepsilon(s_i) = \sum_{\nu=0}^{m+n} d_{j\nu}^\varepsilon s_i^\nu, \tag{31}$$

where, for each $\nu = 0, 1, \ldots, m+n-1$ and $j = 1, \ldots, 4$, $d_{j\nu}^\varepsilon$ is as defined by (9b) but with d_ν^- and d_ν^+ replaced, respectively, by d_ν^- and d_ν^+. Note that D_j^ε, $j = 1, \ldots, 4$, are now continuously differentiable. From (29), it follows that

$$|D_j^\varepsilon(s_i) - D_j(s_i; \boldsymbol{\beta}| \leq c\varepsilon \tag{32}$$

where the constant c is independent of ε. At this stage, we can write down an approximate problem to the optimization problem (23), as follows:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}^0\|^2 \tag{33a}$$

subject to

$$\varrho_{k\nu}^j \geq \varepsilon, \ldots, k = 1, \ldots, \frac{m+n}{2};\ \nu = 1, 2;\ j = 1, \ldots, 4 \tag{33b}$$

$$\bar{h}_{jk}(\boldsymbol{\beta}) = 0, k = 1, \ldots \frac{m+n}{2};\ j = 1, \ldots, 4 \tag{33c}$$

$$\bar{g}_{\nu j}^\varepsilon(\boldsymbol{\beta}) = D_j^\varepsilon(s_\nu) - \bar{Q}_j(s_\nu; \boldsymbol{\beta}) = 0,\ j = 1, \ldots, 4;\ \nu = 1, \ldots, m+n+1. \tag{33d}$$

For each $\varepsilon > 0$, the approximate problem (33) is a standard nonlinearly constrained optimization problem. Furthermore, by (32), we see that it converges to the original optimization problem (23) as $\varepsilon \downarrow 0$. Thus, in practice, we may solve the approximate optimization problem (33) with a sufficiently small value of ε, say 10^{-5}. We believe that the solution so obtained should be closed enough to the solution of the original problem. If the solution of (33) with $\varepsilon = 10^{-5}$ is appreciably different from that of (23), the model of the problem must be re-examined.

Numerical examples

In order to illustrate the robust controller design procedure developed in the previous sections, two examples are presented.

Example 1. The proposed method is applied to the same design problem as in [2] for controlling a second order process. The transfer function of the process is

$$G(s) = \frac{s + [c, d]}{s^2 - 2.2s - [a, b]}$$

where

$$a = 2.2$$

$$b = 2.6$$

$$c = 0.5$$

$$d = 1.5.$$

The controller to be designed is such that the closed-loop system response is stable for all the parameter variations within the given intervals.

Let the structure of the controller be in the general form as (2), i.e.

$$\frac{u}{y} = \frac{\tau_1 s + \tau_0}{s + p_0} \tag{35}$$

with the controller assumed stable; i.e., $p_0 \geq 0$. The closed-loop characteristic polynomial is:

$$s^2 + \alpha s^2 + [w_1, v_1]s + [w_0, v_0] = 0 \tag{36}$$

where

$$\alpha = \tau_1 + p_0 - 2.2 \tag{37}$$

$$[w_1, v_1] = \tau_1[c, d] + \tau_0 - [a, b] - 2.2p_0 \tag{38}$$

$$[w_0, v_0] = \tau_0[c, d] - p_0[a, b]. \tag{39}$$

From (38), we have

$$w_1 = \tau_0 - 2.2p_0 + \min[\tau_1 c, \tau_1 d] - b \tag{40a}$$

$$v_1 = \tau_0 - 2.2p_0 + \max[\tau_1 c, \tau_1 d] - a \tag{40b}$$

and from (39):

$$w_0 = \min[\tau_0 c, \tau_0 d] - p_0 d \tag{40c}$$

$$v_0 = \max[\tau_0 c, \tau_0 d] - p_0 a. \tag{40d}$$

The only Kharitonov polynomials that need to be considered explicitly is in this case [6]:

$$d_2(s) = s^3 + \alpha s^2 + w_1 s + v_0 = 0. \tag{41}$$

From the previous analysis, the characteristic polynomial (36) is stable iff (41) is stable.

With the nominal controller parameters taken as

$$(\mathbf{x}^*)' = (\tau_0^*, \tau_1^*, p_1^*), \tag{42}$$

and using the proposed algorithm, the problem can be readily solved and the desired controller parameters after a few iterations are found to be

$$\tau_1 = -2.745 \times 10^2$$

$$\tau_0 = \quad 1.742 \times 10^2$$

and

$$p_1 = \quad 1.659 \times 10^2.$$

Example 2. The transfer function of a helicopter which is inherently unstable [13] is written as

$$G(s) = \frac{25(s + 0.03)}{[s + (0.4 \pm \delta)][s^2 + 0.36 + 0.16]} = \frac{B(s)}{A(s)}, \quad n = 3 \tag{43}$$

where δ is a constant varying within certain intervals.

The automatic stabilization controller for $thp + 25X$ helicopter is assumed to take the general structure

$$\frac{u}{y} = \frac{\tau_0 + \tau_1 s + \tau_2 s}{p_0 + p_1 s + s^2} = \frac{T(s)}{P(s)}, \quad m = 2 \tag{44}$$

and that the resultant closed-loop characteristic polynomial

$$A(s)P(s) + T(s)B(s) = 0 \tag{45}$$

is stable.

If $\delta = 0$, i.e. no parameter variation, we can use the pole-placement technique to place the roots as desired. Let the closed-loop polynomial be

$$(s + 0.4)(s + 0.2 \pm \imath 0.5)(s + 1)^2 = 0 \tag{46}$$

where $\imath = \sqrt{-1}$. Then the nominal parameters are

$$p_0 = 1.563, \; p_1 = 2.760, \; p_2 = 1.0,$$

and

$$\tau_0 = 0.0213, \; \tau_1 = 0.0642, \; \tau_2 = 0.0544;$$

with

$$a_0 = 0.064,\ a_1 = 0.016,\ a_2 = 0.04,\ a_3 = 1$$

and

$$b_0 = 0.75,\ b_1 = 25.$$

When $\delta = 0.2$, the parameters of $A(s)$ polynomial become:

$$a_0 = [0.032,\ 0.096]$$
$$a_1 = [-0.056,\ 0.088]$$
$$a_2 = [-0.16,\ 0.34]$$

while $B(s)$ remains unchanged.

With these intervals, the nominal controller will give rise to unstable roots in the relevant Kharitonov polynomials (note than in this case there are only three of such). However, if we derive the final controller parameters to be as close as possible to these initial nominal values, and by using the resulting roots of initial Kharitonov polynomial parameters as an initial guess, the algorithm converges after 5 iterations to

$$p_0 = 1.5626,\ p_1 = 2.7593,$$

$$\tau_0 = 0.02722,\ \tau_2 = 0.08649,\ \text{and}\ \tau_2 = 0.06916$$

with roots of all the relevant Kharitonov polynomials being stable.

To verify the obtained results, digital simulation of the helicopter with the parameter δ varying within $[-0.2, 0.2]$ stabilized by the designed controller is conducted. It is observed that its closed-loop responses are stable for $\delta < 0.2$. When $\delta = 0.2$, the system is marginally stable as expected. Thus, if a stable response is needed when $\delta = 0.2$, we remark that the controller should be designed with a slightly larger interval variation of say $[-0.3, 0.3]$.

Concluding remarks

We have presented an alternative formulation and solution procedure to the robust design of systems with interval parameter variation. The method has been illustrated by the computed examples to be feasible for problems of moderate size. It, however, has several limitations. Firstly, the present approach may not work for large scale problems. Furthermore, a recent paper [14] has pointed out that the present method fails to satisfy certain constraint qualifications and hence the success of the numerical computation is sensitive to the choice of initial starting point. The method, nevertheless, provides a starting point in this direction of research. Work in overcoming these shortcomings is underway.

References

1. *Kharitonov, V. L.* (1978), Asymptotic stability of an equilibrium position of a family system of linear differentiable equations, Differentsial'nye Uravneniya, **14**, *11*, pp. 2086–2088.
2. *Soh, Y. C., Evans, R. J., Petersen, I. R., Betz, R. E.* (1987), Robust pole assignment. Automatica, **23**(*5*), pp. 601–610.
3. *Soh, C. B., Berger, C. S., Dabke, K. P.* (1985), On the stability properties of polynomials with perturbed coefficients. IEEE Tr. Aut. Control, **AC–30**, *10*, pp. 1033–1036.
4. *Barmish, B. R.* (1984), Invariance of the strict Hurwitz property for polynomials with perturbed coefficients. IEEE Trans. Aut. Control, **AC–29**, *10*, pp. 935–936.
5. *Cieslik, J.* (1987), On possibilities of the extension of Kharitonov's stability test for interval polynomials to the discrete-time case. IEEE Trans. Aut. Control, **AC–32**, *9*, pp. 237–238.
6. *Anderson, B. D. O., Jury, E. I., Mansour, M.* (1987), On Robust Hurwitz Polynomials. IEEE Trans. Aut. Control, **AC–32**, *10*, pp. 909–912.
7. *Wei, K. H., Yedavalli, T.* (1987), Invariance of Strict Hurwitz Property for Uncertain Polynomials with Dependent Coefficients. IEEE Trans. Aut. Control, **AC–32**, *10*, pp. 907–908.
8. *Bartlett, A. C., Hollot, C. V., Lin, H.* (1987), Root locations on an entire polytope of polynomials: it suffices to check the edges. IEEE Trans. Aut. Control, submitted.
9. *Kucera, V.* (1979), Discrete linear control. John Wiley, New York.
10. *Astrom, K. J., Wittenmark, B.* (1984), Computer controlled systems — theory and design. Prentice-Hall, Englewood Cliffs, N.J.
11. *Goh, G. J., Teo, K. L.* (1988), On minimax eigenvalue problems via constrained optimization. Journal of Optimization Theory and Applications, **57**, *1*, pp. 59–68.
12. *Schittkowski, K.* (1985), NLPQL: A Fortran subroutine for solving constrained nonlinear programming problems. Operations Research Annuals, **5**, pp. 485–500.
13. *Dorf, R. C.* (1986), Modern control system. Addison-Wesley, Reading, Massachusetts.
14. *Panier, E. R.* (1989), On the need for a special purpose algorithm for minimax eigenvalue problems. Journal of Optimization Theory and Application, to appear.

Проектирование робастного контроллера для систем с интервальными параметрами

Ц. Й. ГОХ, Ц. Ц. ЛИМ, К. Л. ТЕО, Д. Й. КЛЕМЕНТС

(Недландс, Аделаида, Недландс, Кенсингтон)

Предлагается метод вычисления для проектирования робастного контроллера. Применяются хорошо известные результаты Харитонова для преобразования проблем проектирования в задаче нелинейного программирования. Вычисляются два примера для иллюстрации применяемости данного подхода.

C.J. Goh
Department of Mathematics
University of Western Australia
Nedlands
WA 6009, Australia

C. C. Lim
Department of Electrical Engineering
University of Adelaide
SA 5001, Australia

K. L. Teo
Department of Mathematics
University of Western Australia
Nedlands
WA 6009, Australia

D. J. Clements
Department of Electrical Engineering
University of New South Wales
NSW 2033, Australia

Problems of Control and Information Theory, Vol. 18(5), pp. 339-358 (1989)

AN ADAPTIVE PRECISION ALGORITHM FOR NUMERICAL SOLUTION OF OPTIMAL CONTROL PROBLEMS BY SUCCESSIVE APPROXIMATION METHOD

E. SCHEIBER

(*Braşov*)

(Received July 20, 1988)

An implementable version of the successive approximation method to solve optimal control is derived. The numerical integration of the differential systems is considered. The algorithm presented here is adaptive in the sense that we use low accuracy numerical integration while we are far from the solution and we improve the accuracy as the solution is approached.

1. Introduction

The successive approximation method has been found to be computationally quite efficient to solve continuous optimal control problems. Several varieties of the successive approximation method were elaborated, but in all cases the exact precision of the intermediate integration is assumed.

In this paper we establish an algorithm based on approximate solution of differential systems. The convergence theorem of the algorithm is an application of Theorem A.1.26 in Polak (1971). Our approach is similar to Kiessig and Polak (1973). Here we use the variety of the successive approximation method studied by Mayne and Polak (1975), Liubushin (1979, 1982). The developed algorithm and the corresponding convergence theorem consider, instead of the exact solutions (as in the referred papers), the numerical solutions based on a discretization method of the differential systems. Thus the algorithm is implementable and we are closely to what happens when we solve an optimal control on a digital computer.

Essentially the new control is obtained disturbing the old control in a subset in order to minimize the cost functional. Other varieties of the method can be found in Krylov, Tchernousko (1962, 1972), Vasiliev (1980), Vasiliev, Tjatiushkin (1981, 1983). The adaptive precision method supposes to use low accuracy numerical integration while we are far from the solution and improve the accuracy as the solution is approached.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

2. The optimal control problem

Consider the following optimal control problem for a fixed $T > 0$:

$$\text{minimize} \quad I(u) = G(x(T)) \tag{2.1}$$

subject to

$$\dot{x}(t) = f(x(t),\, u(t),\, t) \qquad t \in [0, T] \tag{2.2}$$

$$x(0) = x_0 \tag{2.3}$$

$$u(t) \in \Omega \qquad t \in [0, T] \tag{2.4}$$

where $x(t) \in R^n$, $u(t) \in R^q$ are the state of the system and the control at time t, respectively. Ω is a given convex and compact subset of R^q with

$$\Omega \subseteq \{x \in R^q : \|x\| \leq r\}. \,^{\dagger}$$

Let U be the set of admissible controls, i.e.

$$U = \{u : [0, T] \longrightarrow \Omega \text{ is continuous except at a countable number of points}\}.$$

We assume the following hypotheses be satisfied

(H1) The functions $f : R^n \times R^q \times R \longrightarrow R^n$ and $G : R^n \longrightarrow R$ are continuous together with their partial derivatives f_x, f_u, f_{xx}, f_{xu}, f_{uu}, G_x, G_{xx}.

(H2) There exists $M > 0$ such that

$$\langle x, f(x, u, t)\rangle \leq M(1 + \|x\|^2), \; \forall (x, u, t) \in R^n \times \Omega \times [0, T].$$

Given $u \in U$, denote by $p^u : [0, T] \longrightarrow R^n$ the solution of the Cauchy problem

$$\dot{p}(t) = -H_x(x(t),\, u(t),\, p(t),\, t) \tag{2.5}$$

$$p(T) = -G_x(x(T)) \tag{2.6}$$

where $H : R^n \times R^q \times R^n \times R \longrightarrow R$ is the Hamiltonian defined by

$$H(x, u, p, t) = \langle p, f(x, u, t)\rangle.$$

† Throughout this paper $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product, respectively.

From (H1) and (H2) it results that for any $u \in U$ the solutions x^u and p^u of systems (2.2)–(2.3) and (2.5)–(2.6), respectively, exists for any $t \in [0, T]$, and that there exists a positive constant M_1 such that

$$\|x^u(t)\| \leq M_1 \text{ and } \|p^u(t)\| \leq M_1 \quad \forall t \in [0, T], \tag{2.7}$$

for any $u \in U$. Consequently the cost functional I is bounded below. Denote $B = \{x \in R^n : \|x\| \leq M_1\}$.

(H3) For any $u \in U$, there exists a $\overline{u} \in U$ such that

$$H(x^u(t), \overline{u}(t), p^u(t), t) = \max\{H(x^u(t), w, p^u(t), t) : w \in \Omega\}.$$

We denote by $W(u, t)$ and $\mu(u)$ the functions defined by

$$W(u, t) = H(x^u(t), \overline{u}(t), p^u(t), t) - H(x^u(t), u(t), p^u(t), t)$$

and respectively

$$\mu(u) = \int_0^T W(u, t)dt.$$

We have

$$W(u, t) \geq 0 \qquad \forall t \in [0, T] \qquad \text{and } \mu(u) \geq 0$$

for any $u \in U$. If u_* is the optimal control of problem (2.1)–(2.4) then

$$W(u_*, t) = 0 \qquad \forall t \in [0, T] \qquad \text{and } \mu(u_*) = 0.$$

Let the metric $d : U \times U \longrightarrow R$ be defined by

$$d(u_1, u_2) = \int_0^T \|u_1(t) - u_2(t)\|dt.$$

The following properties are known

Theorem 2.1 (Mayne and Polak, 1975). The function $\mu(u)$ is uniformly continuous with respect to the metric d.

Theorem 2.2 (Rozonoer, 1959). There exists a constant $C > 0$ such that for all $u, v \in U$

$$I(v) - I(u) = -\int_0^T [H(x^u(t), v(t), p^u(t), t) - H(x^u(t), u(t), p^u(t), t)]dt + \mathcal{R}$$

with $|\mathcal{R}| \leq Cd^2(u, v)$.

3. The adaptive precision algorithm

For any integer $i \in N$, let $\eta_i = T/2^i$ and let $t_{i,s} = s\eta_i$, $s \in \{0, 1, \ldots, 2^i\}$. Then we define U_i as the set of all functions $u : [0, T] \longrightarrow \Omega$ such that the components of u are real valued step functions having mesh points at $t_{i,s}$, which are continuous from right and which satisfy $u(T) = u(T - 0)$. If $i_1 < i_2$ then $U_{i_1} \subset U_{i_2}$.

Now, let consider the admissible set of controls as

$$U = \overline{\bigcup_{i \in N} U_i},$$

where the closure is in the sense of the of the topology of the Banach space $\mathcal{B}$ of the bounded functions $u : [0, T] \longrightarrow R^q$ with the uniform norm.

For any $u \in U_i$ we define the sequences $(x^u_{i,s})_{0 \leq s \leq 2^i}$, $(x'^u_{i,s})_{0 \leq s \leq 2^i}$ and $(p^u_{i,s})_{0 \leq s \leq 2^i}$ as follows

$$x^u_{i,s+1} = x^u_{i,s} + \eta_i F_1(x^u_{i,s}, u(t_{i,s}), t_{i,s}, \eta_i)$$

$$s = 0, 1, \ldots, 2^i - 1$$

$$x^u_{i,0} = x_0$$

and

$$\begin{pmatrix} x'^u_{i,s} \\ p^u_{i,s} \end{pmatrix} = \begin{pmatrix} x'^u_{i,s+1} \\ p^u_{i,s+1} \end{pmatrix} + \eta_i F_2(x'^u_{i,s+1}, p^u_{i,s+1}, u(t_{i,s}), \eta_i)$$

$$s = 0, 1, \ldots, 2^i - 1$$

$$\begin{pmatrix} x'^u_{i,2^i} \\ p^u_{i,2^i} \end{pmatrix} = \begin{pmatrix} x^u_{i,2^i} \\ -G_x(x^u_{i,2^i}) \end{pmatrix}$$

where F_1 and F_2 are one step integration functions for the differential equations

$$\dot{x}(t) = f(x(t), u(t), t) \qquad t \in [0, T]$$

and

$$\begin{pmatrix} \dot{x}(t) \\ \dot{p}(t) \end{pmatrix} = \begin{pmatrix} f(x(t), u(t), t) \\ -H_x(x(t), u(t), p(t), t) \end{pmatrix} \quad t \in [0, T],$$

respectively. We define the sequence $(\overline{u}_{i,s})_{0 \leq s \leq 2^i - 1}$ by

$$H(x^u_{i,s}, \bar{u}_{i,s}, p^u_{i,s+1}, t_{i,s}) =$$
$$= \max\{H(x^u_{i,s}, w, p^u_{i,s+1}, t_{i,s}) : w \in \Omega\} \tag{3.1}$$
$$s = 0, 1, \ldots, 2^i - 1.$$

Using these sequences we construct the step functions $\mathbf{x}^u_i : [0,T] \longrightarrow R^n$, $\mathbf{x}'^u_i : [0,T] \longrightarrow R^n$, $\mathbf{p}^u_i : [0,T] \longrightarrow R^n$ and $\bar{u}_i : [0,T] \longrightarrow R^q$ as follows

$$\mathbf{x}^u_i(t) = \begin{cases} x^u_{i,s} & \text{if } t \in [t_{i,s}, t_{i,s+1}) \\ x^u_{i,2^i-1} & \text{if } t = T \end{cases}$$

$$\mathbf{x}'^u_i(t) = \begin{cases} x'^u_{i,s+1} & \text{if } t \in [t_{i,s}, t_{i,s+1}) \\ x'^u_{i,2^i} & \text{if } t = T \end{cases}$$

$$\mathbf{p}^u_i(t) = \begin{cases} p^u_{i,s+1} & \text{if } t \in [t_{i,s}, t_{i,s+1}) \\ p^u_{i,2^i} & \text{if } t = T \end{cases}$$

$$\bar{u}_i(t) = \begin{cases} \bar{u}_{i,s} & \text{if } t \in [t_{i,s}, t_{i,s+1}) \\ \bar{u}_{i,2^i-1} & \text{if } t = T. \end{cases}$$

We introduce the following assumption

(H4) The one-step integration formulas used previously are such that there exist an integer $i_0 \geq 0$ and a positive number K satisfying

$$\|x^u(t) - x^u_i(t)\| \leq K\eta_i$$

and

$$\left\| \begin{pmatrix} x^u(t) \\ p^u(t) \end{pmatrix} - \begin{pmatrix} \mathbf{x}'^u_i(t) \\ \mathbf{p}^u_i(t) \end{pmatrix} \right\| \leq K\eta_i$$

for all $u \in U_i$ and for all $i \geq i_0$.

Hypothesis (H4) is satisfied by the Euler–Cauchy method under a Lipschitz assumption on f and $(\partial/\partial x)f$ (cf. Klessig and Polak, 1973).

There exists $M_1 > 0$ such that

$$\|\mathbf{x}^u_i(t)\| \leq M_1 \text{ and } |\mathbf{p}^u_i(t)\| \leq M_1 \quad \forall t \in [0,T] \tag{3.2}$$

for any $u \in U_i$ and any $i \geq i_0$. Without loss of generality we may assume that M_1 in (2.7) and (3.2) is the same.

Finally, for any $i \in N$, we define the functions $W_i : U_i \times \{t_{i,s} : s \in \{0, 1, \ldots, 2^i\}\} \to R$, $I_i : U_i \to R$ as $\mu_i : U_i \to R$ as follows

$$W_i(u, t_{i,s}) = H(x^u_{i,s}, \bar{u}_{i,s}, p^u_{i,s+1}, t_{i,s}) - H(x^u_{i,s}, u_{i,s}, p^u_{i,s+1}, t_{i,s})$$
$$I_i(u) = G(x^u_{i,2^i})$$
$$\mu_i(u) = \eta_i \sum_{s=0}^{2^i-1} W_i(u, t_{i,s})$$

where $u(t) = u_{i,s}$ for $t \in [t_{i,s}, t_{i,s+1})$, $s \in \{0, 1, \ldots, 2^i - 1\}$.

In order to minimize the cost functional we disturb the old control in a subset obtaining the new control. The connection between the analytical elements and the corresponding discretized elements and the change of the cost functional are evaluated in some preliminary results.

Proposition 3.1 For any $\varepsilon > 0$ there exists $i_1 \in N$ such that for any $i > i_1$ and any $u \in U_i$

$$|I(u) - I_i(u)| < \varepsilon.$$

Proof. Let $\varepsilon > 0$. There exists $C > 0$ such that $\|G_x(x)\| \leq C$ for any $x \in B$. Let i_1 be an integer such that $i_1 > i_0$ and

$$\frac{T}{2^{i_1}} < \frac{\varepsilon}{KC}.$$

If $i \geq i_1$ and $u \in U_i$ then, by (H4), we have

$$|I(u) - I_i(u)| = |G(x^u(T)) - G(x^u_{i,2^i})| \leq C\|x^u(T) - x^u_{i,2^i}\| \leq CK\eta_i < \varepsilon.$$

The relation between μ and μ_i is stated in the next proposition.

Proposition 3.2. For any $\varepsilon > 0$ there exists $i_1 \in N$ such that for any $i \geq i_1$ and any $u \in U_i$

$$|\mu(u) - \mu_i(u)| < \varepsilon.$$

Proof. Let $\varepsilon > 0$. If $u \in U_i$ $(i \in N)$ then $u(t) = u_{i,s}$, for $t \in [t_{i,s}, t_{i,s+1})$, $s \in \{0, 1, \ldots, 2^i - 1\}$. We have

$$\mu(u) - \mu_i(u) = \int_0^T W(u, t)dt - \eta_i \sum_{s=0}^{2^i-1} W_i(u, t_{i,s}) =$$

$$= \sum_{s=0}^{2^i-1} \int_{t_{i,s}}^{t_{i,s+1}} [W(u, t) - W_i(u, t_{i,s})]dt =$$

$$= \sum_{s=0}^{2^i-1} \int_{t_{i,s}}^{t_{i,s+1}} [H(x^u(t), \overline{u}(t), p^u(t), t) - H(x^u(t), u(t), p^u(t), t) - \tag{3.3}$$

$$-H(x^u_{i,s}, \overline{u}_{i,s}, p^u_{i,s+1}, t_{i,s}) + H(x^u_{i,s}, u_{i,s}, p^u_{i,s+1}, t_{i,s})]\, dt =$$

$$= \sum_{s=0}^{2^i-1} \int_{t_{i,s}}^{t_{i,s+1}} [H(x^u(t), \overline{u}(t), p^u(t), t) - H(x^u(t), u_{i,s}, p^u(t), t) -$$

$$-H(x^u_{i,s}, \overline{u}_{i,s}, p^u_{i,s+1}, t_{i,s}) + H(x^u_{i,s}, u_{i,s}, p^u_{i,s+1}, t_{i,s})]\, dt.$$

The Hamiltonian $H(x,u,p,t)$ is uniformly continuous on the compact set $B \times \Omega \times B \times [0,T]$. Since the function $(x,p,t) \longrightarrow \max\{H(x,w,p,t) : w \in \Omega\}$ is continuous, it is uniformly continuous in the compact set $B \times B \times [0,T]$. It results that there exists $\delta > 0$ such that $\|x_1 - x_2\| < \delta$, $\|u_1 - u_2\| < \delta$, $\|p_1 - p_2\| < \delta$, $|t_1 - t_2| < \delta$ implies

$$|H(x_1,u_1,p_1,t_1) - H(x_2,u_2,p_2,t_2)| < \delta/2T$$

$$|\max\{H(x_1,w,p_1,t_1) : w \in \Omega\} - \max\{H(x_2,w,p_2,t_2) : w \in \Omega\}| < \varepsilon/2T.$$

By (H4) there exists $i_1 \in N$, $i_1 \geq i_0$ such that for any $i \geq i_1$ and $u \in U_i$

$$\|x^u(t) - x^u_{i,s}\| = \|x^u(t) - \mathbf{x}^u_i(t)\| < \delta$$
$$\|p^u(t) - p^u_{i,s+1}\| = \|p^u(t) - \mathbf{p}^u_i(t)\| < \delta$$
$$|t - t_{i,s}| \leq \eta_i < \delta$$

for any $t \in [t_{i,s}, t_{i,s+1})$. Consequently, we have

$$|H(x^u(t), u_{i,s}, p^u(t), t) - H(x^u_{i,s}, u_{i,s}, p^u_{i,s+1}, t_{i,s})| < \varepsilon/2T$$

and

$$|H(x^u(t), \overline{u}(t), p^u(t), t) - H(x^u_{i,s}, \overline{u}_{i,s}, p^u_{i,s+1}, t_{i,s})| =$$
$$= |\max\{H(x^u(t), w, p^u(t), t) : w \in \Omega\} -$$
$$- \max\{H(x^u_{i,s}, w, p^u_{i,s+1}, t_{i,s}) : w \in \Omega\}| < \varepsilon/2T,$$

$\forall t \in [t_{i,s}, t_{i,s+1})$, $\forall s \in \{0,1,\ldots,2^i-1\}$. Then, from (3.3), we find $|\mu(u) - \mu_i(u)| < \varepsilon$.

Proposition 3.3. For any $m \leq i$ and $u \in U_i$ there exists $j \in \{1,2,\ldots,2^{m-1}\}$ such that

$$\frac{1}{2\eta_m}\, \eta_i \sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} W_i(u, t_{i,s}) < \frac{\mu_i(u)}{T}. \tag{3.4}$$

Proof. We suppose that there exists an $u \in U_i$ and $m \leq i$ such that

$$\frac{1}{2\eta_m}\eta_i \sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} W_i(u, t_{i,s}) \geq \frac{\mu_i(u)}{T}$$

$$\forall j \in \{1, 2, \ldots, 2^{m-1}\}.$$

Then we obtain

$$\mu_i(u) = \eta_i \sum_{s=0}^{2^i-1} W_i(u, t_{i,s}) = \eta_i \sum_{j=1}^{2^{m-1}} \sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} W_i(u, t_{i,s}) <$$

$$< \frac{2\eta_m}{T} 2^{m-1} \mu_i(u) = \mu_i(u)$$

relations which are contradictory.

Let $i, m \in N$, $m \leq i$, $u \in U_i$ and $j \in \{1, 2, \ldots, 2^{m-1}\}$ satisfying (3.4). We define the sequence $(\tilde{u}_{i,s})_{0 \leq s \leq 2^i-1}$ as

$$\tilde{u}_{i,s} = \begin{cases} \tilde{u}_{i,s} & \text{for } s \in \{(2j-2)2^{i-m}, \ldots, 2j \cdot 2^{i-m} - 1\} \\ u_{i,s} & \text{for } s \in \{0, 1, \ldots, 2^i - 1\} \setminus \\ & \qquad \setminus \{2j-2)2^{i-m}, \ldots, 2j \cdot 2^{i-m} - 1\}, \end{cases} \tag{3.5}$$

and the function $\tilde{u} : [0, T] \longrightarrow \Omega$ by

$$\tilde{u}(t) = \begin{cases} \tilde{u}_{i,s} & \text{if } t \in [t_{i,s}, t_{i,s+1}) \\ \tilde{u}_{i,2^i-1} & \text{if } t = T. \end{cases} \tag{3.6}$$

Clearly, $\tilde{u} \in U_i$.

Proposition 3.4. Let $v \in U$ be such $\mu(v) \neq 0$. For any $m \in N$ and $\varepsilon > 0$ there exist $i_2 \in N$ with $i_2 \geq m$, $K_1 > 0$ and a neighborhood V of v, in $\mathcal{B}$, such that for any $i \geq i_2$ and any $u \in V \cap U_i$ we have

$$I(\tilde{u}) - I(u) < 2\varepsilon\eta_m - \frac{\eta_m}{2T}\mu(v) + K_1\eta_m^2 :$$

where $\tilde{u}$ is defined by (3.6).

Proof. Let $j \in \{1, 2, \ldots, 2^{m-1}\}$ and $i \in N$, $i \geq m$. We denote by $J^i_{m,j}$ the interval $[(2j-2)2^{i-m}\eta_i,\ 2j \cdot 2^{i-m}\eta_i] = [(2j-2)\eta_m, 2j\eta_m]$. The length of the interval $J^i_{m,j}$ is $2\eta_m$. Using Theorem 2.2 we have

$$I(\tilde{u}) - I(u) =$$

$$= -\int_0^T [H(x^u(t), \tilde{u}(t), p^u(t), t) - H(x^u(t), u(t), p^u(t), t)]dt + \mathcal{R} =$$

$$= -\int_{J^i_{m,j}} [H(x^u(t), \overline{u}(t), p^u(t), t) - H(x^u(t), u(t), p^u(t), t)]dt + \mathcal{R} = \tag{3.7}$$

$$= -\sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} \int_{t_{i,s}}^{t_{i,s+1}} [H(x^u(t), \overline{u}_{i,s}, p^u(t), t) - H(x^u(t), u_{i,s}, p^u(t), t)]dt +$$

$$+\eta_i \sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} [H(x^u_{i,s}, \overline{u}_{i,s}, p^u_{i,s+1}, t_{i,s}) -$$

$$-H(x^u_{i,s}, u_{i,s}, p^u_{i,s+1}, t_{i,s})] - \eta_i \sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} W_i(u, t_{i,s}) + \mathcal{R}$$

and

$$|\mathcal{R}| \leq Cd^2(\tilde{u}, u) = C\left(\int_{J^i_{m,j}} \|\tilde{u}(t) - u(t)\|dt\right)^2 \leq 16Cr^2\eta_m^2 = K_1\eta_m^2$$

where $K_1 = 16Cr^2$. From Theorem 2.1, with $\varepsilon = \mu(v)/2$ there exists a neighborhood V (in $\mathcal{B}$) of v such that for any $u \in V$

$$|\mu(u) - \mu(v)| < \frac{\mu(v)}{2}.$$

From Proposition 3.2, with $\varepsilon = \mu(v)/4$ there exists $i_1 \in N$, $i_1 \geq m$ such that

$$|\mu(u) - \mu_i(u)| < \frac{\mu(v)}{4}$$

for any $u \in U_i$ and for any $i \geq i_1$. For $i \geq i_1$ and $u \in V \cap U_i$ it results that

$$|\mu_i(u) - \mu(v)| \leq |\mu_i(u) - \mu(u)| + |\mu(u) - \mu(v)| < \frac{3}{4}\mu(v)$$

and hence

$$\mu_i(u) > \frac{\mu(v)}{4}. \tag{3.8}$$

Using the uniform continuity of the Hamiltonian $H(x,u,p,t)$ in the compact set $B \times \Omega \times B \times [0,T]$, there exists $\delta > 0$ such that

$$|H(x_1,u_1,p_1,t_1) - H(x_2,u_2,p_2,t_2)| < \varepsilon/2$$

for any $(x_1,u_1,p_1,t_1),(x_2,u_2,p_2,t_2)$ satisfying $\|x_1 - x_2\| < \delta$, $\|u_1 - u_2\| < \delta$, $\|p_1 - p_2\| < \delta$, $|t_1 - t_2|\delta$. From hypothesis (H4) there exists $i_2 \in N$, $i_2 \geq i_1$ such that for any $i \geq i_2$

$$\|x^u(t) - x^u_{i,s}\| = \|x^u(t) - \mathbf{x}^u_i(t)\| < \delta$$
$$\|p^u(t) - p^u_{i,s+1}\| = \|p^u(t) - \mathbf{p}^u_i(t)\| < \delta, \qquad \eta_i < \delta$$

and consequently

$$|H(x^u(t),u_{i,s},p^u(t),t) - H(x^u_{i,s},u_{i,s},p^u_{i,s+1},t_{i,s})| < \varepsilon/2$$
$$|H(x^u(t),\overline{u}_{i,s},p^u(t),t) - H(x^u_{i,s},u_{i,s},p^u_{i,s+1},t_{i,s})| < \varepsilon/2$$

for all $t \in [t_{i,s},t_{i,s+1})$. For $i \geq i_2$, relations (3.7), (3.4) and (3.8) imply

$$I(\tilde{u}) - I(u) \leq \eta_i \sum_{s=(2j-2)2^{i-m}}^{2j\cdot 2^{i-m}-1} \varepsilon - \frac{2\eta_m}{T}\mu_i(u) + K_1\eta_m^2 < 2\varepsilon\eta_m - \frac{\eta_m}{2T}\nu(v) + K_1\eta_m^2.$$

Proposition 3.5. Let $v \in U$ be such that $\mu(v) \neq 0$. Then there exist i_2, $m \in N$ with $i_2 \geq m$ and a neighborhood V, in $\mathcal{B}$, of v such for any $i \geq i_2$ and for any $u \in V \cap U_i$ we have

$$I_i(\tilde{u}) < I_i(u)\frac{1}{2^{m+2}}\mu_i(u),$$

where $\tilde{u}$ is defined by (3.6).

Proof. Choose $\varepsilon > 0$ and $m \in N$ such that

$$0 < \varepsilon < \frac{\mu(v)}{32T}$$

and

$$\eta_m = \frac{T}{2^m} < \frac{1}{K_1}\left(\frac{\mu(v)}{16T} - 2\varepsilon\right).$$

The constant K_1 is defined in Proposition 3.4. From the same proposition there exist an $i_1 \in N$, $i_1 \geq m$ and a neighborhood V of v, in $\mathcal{B}$, such that for any $i \geq i_1$ and any $u \in V \cap U_i$

$$I(\tilde{u}) - I(u) < 2\varepsilon\eta_m - \frac{\eta_m}{2T}\mu(v) + K_1\eta_m^2.$$

From the proof of Proposition 3.4 we have

$$|\mu(u) - \mu(v)| < \frac{\mu(v)}{2} \qquad \forall u \in V.$$

Using Propositions 3.1 and 3.2, with $\varepsilon = \frac{\eta_m}{16T}\mu(v)$ there exists $i_2 \geq i_1$, $i_2 \in N$ such that for any $i \geq i_2$ and any $u \in U_i$

$$\left|[I_i(\tilde{u}) - I_i(u) + \frac{\eta_m}{4T}\mu_i(u) - [I(\tilde{u}) - I(u) + \frac{\eta_m}{4T}\mu(u)]\right| < \frac{\eta_m}{16T}\mu(v).$$

For $i \geq i_2$ and $u \in V \cap U_i$ it results that

$$I_i(\tilde{u}) - I_i(u) + \frac{\eta_m}{4T}\mu_i(u) < I(\tilde{u}) - I(u) + \frac{\eta_m}{4T}\mu(u) + \frac{\eta_m}{16T}\mu(v) <$$

$$< I(\tilde{u}) - I(u) + \frac{7\eta_m}{16T}\mu(v) < 2\varepsilon\eta_m - \frac{\eta_m\mu(v)}{16T} + K_1\eta_m^2 < 0,$$

that is the desired result, because $\eta_m/4T = 1/2^{m+2}$.

Now we are able to state the adaptive precision algorithm for solving the continuous optimal control problem.

Step 1. Select $i_0 \in N$, $\varepsilon_0 > 0$ and $u^0 \in U_{i_0}$. Set $k = 0$, $i = i_0$, $m = 1$, $\varepsilon = \varepsilon_0$.

Step 2. Compute $(x^k_{i,s})_{0,\leq s\leq 2^i}$, $(p^k_{i,s})_{0,\leq s\leq 2^i}$.

Step 3. Compute $I_i(u^k) = G(x^k_{i,2^i})$.

Step 4. Compute $\bar{u}_{i,s}$ such that

$$H(x^k_{i,s}, \bar{u}_{i,s}, p^k_{i,s+1}, t_{i,s}) =$$

$$= \max\{H(x^k_{i,s}, w, p^k_{i,s+1}, t_{i,s}) : {}_{w\in\Omega\}, s\in\{0,1,\ldots,2^i-1\}} .$$

Step 5. Compute $\mu_i(u^k)$.

Step 6. Find $j \in \{1, 2, \ldots, 2^{m-1}\}$ satisfying (3.4).

Remark. According to Proposition 3.3 there exists $j \in \{1, 2, \ldots, 2^{m-1}\}$ satisfying (3.4).

Step 7. Compute the sequence $\tilde{u} = (\tilde{u}_{i,s})_{0,\leq s\leq 2^i-1}$, according to (3.5) and compute $I_i(\tilde{u})$.

Step 8. If $I_i(\tilde{u}) < I_i(u) - \frac{1}{2^{m+2}}\mu_i(u)$ then go to Step 10, else, go to Step 9.

Remark. According to Proposition 3.5 there exists i and m satisfying the desired inequality.

Step 9. If $m < i$ then set $m = m + 1$ and go to Step 6, else,set $i = i + 1$ and go to Step 2.

Step 10. If $I_i(\tilde{u}) - I_i(u) \leq -\varepsilon$ then go to Step 11, else set $i = i + 1$, $\varepsilon = \varepsilon/2$ and go to Step 2.

Step 11. Set $u^{k+1} = \tilde{u}$, $k = k + 1$ and go to Step 2.

The algorithm has the same form as the prototype algorithm defined in Polak (1971, p.283) or Kiessig and Polak (1973). The search function $A_i : U_i \longrightarrow 2^{U_i}$ is defined by

$$A_i(u) - \{\tilde{u} : \text{ computed from } u \text{ according to (3.6)}\}$$

and the desirable set is

$$\Delta = \{u \in U : \mu(u) = 0\}.$$

In applying Theorem A.1.26 in Polak (1971), the convergence properties of our algorithm will be established with respect to the topology of the Banach space $\mathcal{B}$. The algorithm has the following property

Theorem 3.6. Suppose that assumptions (H1)–(H4) are satisfied. Consider any sequence constructed by our algorithm. If the sequence is finite because the algorithm jammed up between Step 2 and Step 10 with the last element u^k then $\mu(u^k) = 0$. If the sequence is infinite and if u_* is an accumulation point of the sequence then $\mu(u_*) = 0$.

Proof. We must verify the conditions of Theorem A.1.26 in Polak (1971), which are, in our case

(i) $\overline{\bigcup_{i \in N} U_i} = U$.

(ii) For any $v \in U$, $\mu(v) \neq 0$ there exist a neighborhood V of v, a $\delta(v) < 0$ and an integer $N(v) \geq 0$ satisfying

$$I_i(u'') - I_i(u') \leq \delta(v) < 0$$

for all $u' \in V \cup U_i$, for all $u'' \in A_i(u')$ and for all $i \geq N(v)$.

(iii) There is a sequence $(q_s)_{s=0}^{\infty}$ such that

$$\sum_{s=0}^{\infty} q_s < +\infty$$

and $|I_i(u) - I(u)| \leq q_s$ for all $u \in U_i$ and for all $i \geq s$.

Let us do the checking.

1° It is obvious that (i) is satisfied.

2° Suppose that $v \in U$ is such that $\mu(v) \neq 0$. Using Proposition 3.5 there exist m, $i_2 = N(v) \in N$, $m \leq i_2$ and a neighborhood V of v, in $\mathcal{B}$, such that for any $i \geq N(v)$, for any $u' \in V \cap U_i$ and for any $u'' \in A(u')$

$$I_i(u'') - I_i(u') < -\frac{1}{2^{m+2}}\mu_i(u).$$

Since inequality (3.8) is valid, it results that

$$I_i(u'') - I_i(u') < -\frac{1}{2^{m+4}}\mu(v) = \delta(v) < 0.$$

3° There exists $C > 0$ such that $\|G_x(x)\| \leq C$ for any $x \in B$. Then we have

$$|I(u) - I_i(u)| = |G(x^u(T)) - G(x^u_{i,2}i)| \leq$$

$$\leq C\|x^u(T) - x^u_{i,2}i\| \leq CKT/2^i,$$

for any $u \in U_i$ and any $i \geq i_0$. Hence (iii) is satisfied too.

4. Numerical examples

We have solved the following optimal control problems.

Example 1 ° (Sakawa, Shindo, 1980)

minimize $I(u) = \int_0^{2.5} (x_1^2 + u^2)dt$

$\dot{x}_1 = x_2$ $\qquad x_1(0) = -5.0$

$\dot{x}_2 = -x_1 + 1.4\,x_2 - 0.14\,x_2^3 + 4\,u$ $\qquad x_2(0) = -5.0$

$|u(t)| \leq 1 \qquad \forall t \in [0,\ 2.5]$

This optimal control problem may be rewritten as

minimize $I(u) = y_1(2.5)$

$\dot{y}_1 = y_2^2 + u^2$ $\qquad y_1(0) = 0.0$

$\dot{y}_2 = y_3$ $\qquad y_2(0) = -5.0$

$\dot{y}_3 = -y_2 + 1.4\,y_3 - 0.14\,y_3^3 + 4\,u$ $\qquad y_3(0) = -5.0$

$|u(t)| \leq 1 \qquad \forall t \in [0,\ 2.5]$

The costate system of the optimal control problem is given by

$\dot{p}_1 = 0$ $\qquad p_1(2.5) = -1.0$

$\dot{p}_2 = -2p_1y_2 + p_3$ $\qquad p_2(2.5) = 0.0$

$\dot{p}_3 = -p_2 - 1.4\,p_3 + 0.42\,p_3y_3^2$ $\qquad p_3(2.5) = 0.0$

and the expressions for $\overline{u}$ and $W(u,t)$ are

$\overline{u} = 2$ sgn p_3

$W(u,t) = p_1(\overline{u}^2 - u^2) + 4\,p_3(\overline{u} - u).$

The numerical results are presented in Table 1.

Example 2 ° (Vasiliev, Tjatiushkin, 1983)

minimize $I(u) = x_1^2(5) + x_2^2(5)$

$$\dot{x}_1 = x_2 \qquad x_1(0) = 5.0$$
$$\dot{x}_2 = u - \sin x_1 \qquad x_2(0) = 0.0$$
$$|u(t)| \leq 1 \qquad \forall t \in [0, 5].$$

The numerical results are presented in Table 2.

Example 3 ° (Kiessig, Polak, 1973)

minimize $I(u) = x_1(1)$

$$\dot{x}_1 = x_2^2 + x_3^2 \qquad x_1(0) = 0.0$$
$$\dot{x}_2 = x_3 \qquad x_2(0) = 1.0$$
$$\dot{x}_3 = -(0.01 + 4\pi^2)x_2 - 0.2x_3 + u \qquad x_3(0) = 0.0$$
$$|u(t)| \leq 1 \qquad \forall t \in [0, 1].$$

The numerical results are presented in Table 3.

Table 1

Iter. Number	Discret. Param.	Cost Functional	R
1	3	64.794	42.159
2	3	41.717	.39566
3	3	41.176	.12379E-01
3	4	42.641	.24390E-01
4	4	42.631	.24390E-03

The ended mode indicator : 0
The minimal cost-function : 42.631195

The Optimal Control

U(	.00000000	)=	1.0000000
U(	.15625000	)=	1.0000000
U(	.31250000	)=	1.0000000
U(	.46875000	)=	1.0000000
U(	.62500000	)=	1.0000000
U(	.78125000	)=	1.0000000
U(	.93750000	)=	1.0000000
U(	1.0935700	)=	.72686633
U(	1.2500000	)=	.49603235
U(	1.4062500	)=	.29519614
U(	1.5625000	)=	.12663747
U(	1.7187500	)=	.44250043E-02
U(	1.8750000	)=	-.58008658E-01
U(	2.0312500	)=	-.60135456E-01
U(	2.1875000	)=	-.24453221E-01
U(	2.3437500	)=	.00000000

Table 2

Iter. Number	Discret. Param.	Cost Functional	R
1	3	44.631	55.137
2	3	16.250	4.3421
3	3	13.615	11.271
3	4	13.571	8.7012
3	5	13.568	7.7635
4	5	12.421	.86176
5	5	12.095	.51259
5	6	12.095	.38609
6	6	11.966	.10247
6	7	11.966	.15361
7	7	11.963	.93746E-01
8	7	11.930	.84338E-01
8	8	11.930	.72336E-01
9	8	11.914	.10670E-01
10	8	11.912	.10955E-01
10	9	11.912	.87888E-02

The ended mode indicator : 0
The minimal cost-function : 11.912089

The Optimal Control			
U(	.00000000	)=	1.0000000
U(	.97656250E-02	)=	1.0000000
U(	.19531250E-01	)=	1.0000000
U(	.29296875E-01	)=	1.0000000
U(	.39062500E-01	)=	1.0000000
U(	.48828125E-01	)=	1.0000000
U(	.58593750E-01	)=	1.0000000
U(	.68359375E-01	)=	1.0000000
U(	.78125000E-01	)=	1.0000000
U(	.87890625E-01	)=	1.0000000
U(	.97656250E-01	)=	1.0000000
U(	.10742187	)=	1.0000000
U(	.11718750	)=	1.0000000
U(	.12695312	)=	1.0000000
U(	.13671875	)=	1.0000000
U(	.14648437	)=	1.0000000

```
U(   .86914062   )=    1.0000000
U(   .87890625   )=    1.0000000
U(   .88867187   )=    1.0000000
U(   .89843750   )=    1.0000000
U(   .90820312   )=    1.0000000
U(   .91796875   )=    1.0000000
U(   .92773437   )=    1.0000000
U(   .93750000   )=    1.0000000
U(   .94726562   )=    1.0000000
U(   .95703125   )=    1.0000000
U(   .96679687   )=    1.0000000
U(   .97656250   )=   -1.0000000
U(   .98632812   )=   -1.0000000
U(   .99609375   )=   -1.0000000
U(   1.0058594   )=   -1.0000000
U(   1.0156250   )=   -1.0000000
U(   1.0253906   )=   -1.0000000
```

```
U(   4.4628906   )=   -1.0000000
U(   4.4726562   )=   -1.0000000
U(   4.4824219   )=   -1.0000000
U(   4.4921875   )=   -1.0000000
U(   4.5019531   )=   -1.0000000
U(   4.5117187   )=   -1.0000000
U(   4.5214844   )=   -1.0000000
U(   4.5312500   )=   -1.0000000
U(   4.5410156   )=   -1.0000000
U(   4.5507812   )=    1.0000000
U(   4.5605469   )=    1.0000000
U(   4.5703125   )=    1.0000000
U(   4.5800781   )=    1.0000000
U(   4.5898437   )=    1.0000000
U(   4.5996094   )=    1.0000000
```

```
.                .
.                .
.                .
U(  4.8632812  )=   1.0000000
U(  4.8730469  )=   1.0000000
U(  4.8828125  )=   1.0000000
U(  4.8925781  )=   1.0000000
U(  4.9023437  )=   1.0000000
U(  4.9121094  )=   1.0000000
U(  4.9218750  )=   1.0000000
U(  4.9316406  )=   1.0000000
U(  4.9414062  )=   1.0000000
U(  4.9511719  )=   1.0000000
U(  4.9609375  )=   1.0000000
U(  4.9707031  )=   1.0000000
U(  4.9804687  )=   1.0000000
U(  4.9902344  )=   1.0000000
STOP
```

Table 3

Iter. Number	Discret. Param.	Cost Functional	R
1	3	18.117	1.8289
2	3	16.392	.00000

The ended mode indicator : 0
The minimal cost-function : 16.391563

The Optimal Control

```
U(  .00000000  )=   1.0000000
U(  .12500000  )=   1.0000000
U(  .25000000  )=   1.0000000
U(  .37500000  )=   1.0000000
U(  .50000000  )=  -1.0000000
U(  .62500000  )=  -1.0000000
U(  .75000000  )=  -1.0000000
U(  .87500000  )=   1.0000000
STOP
```

References

1. *Kiessig, R., Polak, E.*, 1973, An Adaptive Precision Gradient Method for Optimal Control. SIAM J. Control, **11**, No.1, 80–93.
2. *Mayne, D.Q., Polak, E.*, 1975, First Order Strong Variation Algorithms for Optimal Control. JOTA, **16**, No.3/4, 277–301.
3. *Polak, E.*, 1971, Computational Methods in Optimization. Academic Press, New York.
4. *Sakawa, Y., Shindo, Y.*, 1980, Global Convergence of an Algorithm for Optimal Control. IEEE Trans. on Autom. Control, **25**, No.6, 1149–1153.
5. Васильев, Ф. П., 1980, Численные методы решения экстремальных залач. Наука, Москва.
6. Васильев, О. В., Тятюшкин, А. Н., 1981, Об одном методе решения задач опитмального управления, основанном на принципе максимума. Ж. вычисл. матем. и матем. физ., **21,** No. *6*, 1367–1384.
7. Васильев, О. В., Тятюшкин, А. Н., 1983, Опыт решения задач оптимального управления на основе необходимых условии оптимальности типа принципа максимума. Вопросы устойчивости и оптимизации динамических систем. Изд. Иркутского Унив., Иркутск, 43–64.
8. Крылов, И. А., Черноусько, Ф. Л., 1962, О методе последовательных приближений для решения задач оптимального управления. Ж. вычисл. матем. и матем. физ., **2,** No. *6*, 1132–1139.
9. Крылов, И. А., Черноусько, Ф. Л., 1972, Алгоритм метода последовательных приближений для задач оптимального управления. Ж. вычисл. матем. и матем. физ., **12,** No. *1*, 14–34.
10. Любушин, А. А., 1979, Модификации и исследование сходимости метода последовательных приближений для задач оптимального управления. Ж. вычисл. матем. и матем. физ., **19,** No. *6*, 1414–1421.
11. Любушин, А. А., 1982, О применений модификации метода последовательных приближений для решения задач оптимального управления. Ж. вычисл. матем. и матем. физ., **22,** No. *1*, 30–35.
12. Розоноэр, Л. И., 1959, Принцип максимума Л. С. Понтрягина в теории оптимальных систем. Автоматика и телемехан., **20,** No. *10*, 1320–1334; No. *11*, 1441–1458.

Адаптивный алгорифм для численного решения задач оптимального управления с методом последовательных приближений

Е. ШАЙБЕР

(Брашов)

Выводится реализационный вариант метода последовательных приближений для решения задач оптимального управления. Изучается численное интегрирование для систем дифференциальных уравнений.

Алгоритм называется адаптивным в том смысле, что точность численных интегрированных схем увеличится с приближением к решению задач.

E. Scheiber
Department of Mathematics
University of Brasov
str. Karl Marx 50
2200 Braşov, Romania

РУССКИЙ ПЕРЕВОД

Проблемы управления и теории информации, том **18**, номер *5 (1989)*

СТОХАСТИЧЕСКОЕ ПРИЦЕЛИВАНИЕ В ДЕТЕРМИНИРОВАННОМ ПОЗИЦИОННОМ УПРАВЛЕНИИ

Д. А. Серков

(Свердловск)

В данной статье приводится непосредственное доказательство равенства удобной для возможных численных реализаций формы стохастического максимина цене дифференциальной игры для одного класса нелинейных по управлению систем и выпуклого терминального показателя качества. Доказательство основывается на идее использования сопряженных переменных принципа максимума Понтрягина.

Введение

В предлагаемой статье рассматривается задача об управлении в условиях неопределенной информации о помехе. Принята математическая концепция такого управления, предложенная и развиваемая в исследованиях, которые проводятся в Свердловске [1–5, 9, 11]. Статья примыкает и к другим работам по теории дифференциальных игр [6–8].

В ней существенно используется аппарат принципа максимума Л. С. Понтрягина [10] в форме, модернизированной для минимаксного программного управления стохастической системой.

В работе [9] дано представление цены детерминированной позиционной дифференциальной игры в виде стохастического максимина. При этом наибольший интерес, с точки зрения возможных численных реализаций, представляет стохастический максимин, содержащий математическое ожидание значений, принимаемых показателем качества на случайных движениях.

В данной статье приводится непосредственное доказательство равенства такой формы стохастического максимина цене дифференциальной игры для одного класса нелинейных по управлению систем и выпуклого терминального показателя качества. Доказательство основывается на предложенной в [9] идее привлечения сопряженных переменных принципа максимума

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Понтрягина [10], но не использует результаты теории уравнений в частных производных и динамического программирования.

Все рассуждения будут проведены на примере наиболее простой системы и удобного показателя качества, чтобы не отвлекать внимание на второстепенные детали. На возможные обобщения будет указано в последнем пункте.

Постановка задачи

Рассмотрим позиционную дифференциальную игру [11] для системы линейных уравнений

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu[t] + Cv[t], \quad t \in [t_*, \vartheta], \\ x(t_*) = x_* \in R^n, \end{cases} \tag{1}$$

здесь A, B, C — постоянные матрицы; $u[\cdot]$, $v[\cdot]$ — измеримые реализации управления первого и второго игроков, удовлетворяющие при почти всех $t \in [t_*, \vartheta]$ включениям $u[t] \in \mathcal{P} \subset R^k$, $v(t) \in \mathcal{Q} \subset R^m$, $\mathcal{P}$, $\mathcal{Q}$ — выпуклые компакты. Множества допустимых реализаций управлений игроков на интервале $[t_*, \vartheta]$ будем обозначать $\mathcal{U}_{[t_*,\vartheta]}$ и $\mathcal{V}_{[t_*,\vartheta]}$. В качестве терминального показателя качества возьмем строго выпуклую функцию $\sigma(\cdot)$ из $C^\infty(R^n)$:

$$\gamma = \sigma(x(\vartheta, t_*, x_*, u[\cdot], v[\cdot])), \tag{2}$$

$x(\cdot, t_*, x_*, u[\cdot], v[\cdot])$ — решение системы (1). Обозначим $\varrho(\cdot)$ цену дифференциальной игры.

Пусть $\Delta = \{t_* = \tau_1, \ t^* = \tau_2, \ldots, \tau_{\overline{n}+1} = \vartheta\}$ — некоторое разбиение отрезка $[t_*, \vartheta]$, $\Delta' = \Delta \setminus \{t_*\}$ — разбиение отрезка $[t^*, \vartheta]$. Обозначим $P_u(\Delta)$, $P_v(\Delta)$ множества стохастических неупреждающих программ первого и второго игроков, соответственно, порожденных разбиением Δ [11] (можно считать, что $u(\cdot) \in P_u(\Delta)$ есть измеримая функция, определенная на $[t_*, \vartheta] \times \Omega$, $\Omega = \Omega(\Delta) = \{\omega = (\omega_1, \ldots, \omega_{\overline{n}}) \in [0,1)^{\overline{n}}\}$ почти всюду удовлетворяющая включению $u(\tau, \omega) \in \mathcal{P}$ и не зависящая от $\omega_i, \ldots, \omega_n$, если $\tau \in [\tau_{i-1}, \tau_i)$ для любого $i = 2, \ldots, \overline{n}$). Подмножества из $P_u(\Delta)$ и $P_v(\Delta)$ элементов, не зависящих от ω_1, обозначим, соответственно, $P'_u(\Delta)$ и $P'_v(\Delta)$ (таким образом, элементы множеств $P'_u(\Delta)$ и $P_u(\Delta')$ отличаются областью определения: у первых это $[t_*, \vartheta] \times \Omega'$, у вторых — $[t^*, \vartheta] \times \Omega'$, $\Omega' = \Omega'(\Delta) = \{\omega' = (\omega_2, \ldots, \omega_{\overline{n}}) \in [0,1)^{\overline{n}-1}\}$). Стохастический максимин $\varrho_*(\cdot)$ [9, 11] для данной игры определим соотношением

$$\varrho_*(t_*, x_*) = \lim_{d(\Delta)\to 0} \sup_{\Delta} \varrho_{*\Delta}(t_*, x_*),$$

$$\varrho_{*\Delta}(t_*, x_*) = \sup_{v(\cdot)\in P_v(\Delta)} \inf_{u(\cdot)\in P_u(\Delta)} \int_{\Omega} \sigma(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega)))\, d\omega, \qquad (3)$$

$$d(\Delta) = \max\{\tau_{i+1} - \tau_i: \quad i = 1, \ldots, \overline{n}\}.$$

Известно, что для выполнения равенства $\varrho_*(\cdot) = \varrho(\cdot)$ необходимы и достаточны следующие свойства [11] функции $\varrho_*(\cdot)$: u- и v-стабильность и выполнение краевого условия $\varrho_*(\vartheta, x) = \sigma(x), \quad x \in R^n$. Последние два свойства установлены при достаточно общих предположениях [11]. Приведем определение свойства u-стабильности функции $\varrho_*(\cdot)$: пусть $\varepsilon > 0$, $t^* \in [t_*, \vartheta]$ и $v[\cdot] \in \mathcal{V}_{[t_*,\vartheta]}$ — произвольная реализация управления второго игрока на промежутке $[t_*, t^*]$. Для этих данных найдется реализация управления первого игрока $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ такая, что

$$\varrho_*(t_*, x_*) \geq \varrho_*(t^*, x(t^*, t_*, x_*, u[\cdot], v[\cdot])) - \varepsilon. \qquad (4)$$

Следовательно, для доказательства (4) достаточно для любого разбиения установить аналогичное неравенство

$$\varrho_* = \varrho_{*\Delta}(t_*, x_*) \geq \varrho_{*\Delta'}(t^*, x_{u[\cdot]}) - \varepsilon = \varrho^*(x_{u[\cdot]}) - \varepsilon, \qquad (5)$$

где

$$x_{u[\cdot]} = x(t^*, t_*, x_*, u[\cdot], v[\cdot]),$$

$$\varrho_{*\Delta'}(t^*, x_{u[\cdot]}) = \sup_{v(\cdot)\in P_v(\Delta')} \inf_{u(\cdot)\in P_u(\Delta')} \int_{\Omega'} \sigma(x(\vartheta, t^*, x_{u[\cdot]}, u(\cdot,\omega'), v(\cdot,\omega')))\, d\omega'.$$

Параметры t_*, t^*, x_*, $v[\cdot] \in \mathcal{V}_{[t_*,t^*]}$ в дальнейшем будем считать зафиксированными. Итак, следует доказать существование функции $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$, удовлетворяющей неравенству (5).

Доказательство u-стабильности функции

Построим «расширение» системы (1) и показателя качества (2):

$$y = (x, x_{n+1}) \in R^{n+1}$$

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu[t] + Cv[t], \quad t \in [t_*, \vartheta] \\ \dot{x}_{n+1}(t) = u^2[t] = |u[t]|^2_{R^k}, \\ y(t_*) = (x(t_*), x_{n+1}(t_*)) = (x_*, 0) = y_*; \end{cases} \qquad (6)$$

$$\sigma_\alpha(y) = \sigma(x) + \alpha \cdot x_{n+1}^2, \quad (\alpha > 0). \tag{7}$$

Обозначим

$$y_{u[\cdot]} = y(t^*, t_*, y_*, u[\cdot], v[\cdot]),$$

где $y(\cdot, t_*, y_*, u[\cdot], v[\cdot])$ есть решение системы (6). Введем величины, аналогичные величинам ϱ_* и $\varrho^*(x_{u[\cdot]})$:

$$\begin{aligned} \varrho_{*\alpha} &= \sup_{v(\cdot)\in P_v(\Delta)} \inf_{u(\cdot)\in P_u(\Delta)} \int_\Omega \sigma_\alpha(y(\vartheta, t_*, y_*, u(\cdot,\omega), v(\cdot,\omega)))\, d\omega, \\ \varrho_\alpha^*(y_{u[\cdot]}) &= \sup_{v(\cdot)\in P_v(\Delta')} \inf_{u(\cdot)\in P_u(\Delta')} \int_{\Omega'} \sigma_\alpha(y(\vartheta, t^*, y_{u[\cdot]}, u(\cdot,\omega'), v(\cdot,\omega')))\, d\omega'. \end{aligned} \tag{8}$$

Лемма 1. Для любого $\eta > 0$ найдется $\alpha(\eta) > 0$, такое, что для любых $\alpha < \alpha(\eta)$, $\overline{u}[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ верны неравенства

$$|\varrho_{*\alpha} - \varrho_*| < \eta, \tag{9}$$

$$|\varrho_\alpha^*(y_{\overline{u}[\cdot]}) - \varrho^*(x_{u[\cdot]})| < \eta. \tag{10}$$

Доказательство. В силу ограниченности множества значений управлений первого и второго игроков, для любых $u(\cdot) \in P_u(\Delta)$, $v(\cdot) \in P_v(\Delta)$ и почти всех $\omega \in \Omega$ имеем неравенства

$$|\sigma(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))) - \sigma_\alpha(y(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega)))| \le C\cdot\alpha, \quad C = \text{const}.$$

Отсюда, используя определения $\varrho_{*\alpha}$ и ϱ_*, получаем (9). Аналогично доказывается (10).

Обозначим $M(t,\cdot)$ оператор условного математического ожидания относительно первых i компонент случайной величины ω при $t \in [\tau_i, \tau_{i+1})$.

Введем управления для сопряженных переменных:

$$\begin{cases} \dot{s}(t,\omega) = -A^T s(t,\omega), \quad (t,\omega) \in [t_*, \vartheta] \times \Omega, \\ \dot{s}_{n+1}(t,\omega) = 0, \\ s(\vartheta,\omega) = \dfrac{\partial\sigma}{\partial x}(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))), \\ s_{n+1}(\vartheta,\omega) = 2\alpha \cdot \displaystyle\int_{[t_*,\vartheta]} u^2(\tau,\omega)\, d\tau. \end{cases} \tag{11}$$

Нетрудно проверить, что при всех (t,ω)

$$\begin{aligned} s(t,\omega) &= X^T(\vartheta,t)\frac{\partial\sigma}{\partial x}(x(\vartheta, t_*, x_*, u(\cdot,\omega), v(\cdot,\omega))), \\ s_{n+1}(t,\omega) &= s_{n+1}(\vartheta,\omega); \end{aligned}$$

здесь $X^T(\vartheta, t)$ — транспонированная фундаментальная матрица системы (1).

Лемма 2. Для любой $\overline{v}(\cdot) \in P_v(\Delta)$ существует единственная $u_\alpha(\cdot) \in \in P_u(\Delta)$, дающая минимум функционалу

$$I_\alpha(u(\cdot)) = \int_\Omega \sigma_\alpha(y(\vartheta, t_*, y_*, u(\cdot, \omega), \overline{v}(\cdot, \omega)))\, d\omega. \tag{12}$$

Неупреждающая программа $u_\alpha(\cdot)$ однозначно определяется условием: для всех $(t, \omega) \in [t_*, \vartheta] \times \Omega$

$$\begin{aligned} \langle Bu_\alpha(t,\omega), M(t, s(t,\omega))\rangle_{R^n} + u_\alpha^2(t,\omega) \cdot M(t, s_{n+1}(t,\omega)) = \\ = \min_{u \in P}\{\langle Bu, M(t, s(t,\omega))\rangle_{R^n} + u^2 \cdot M(t, s_{n+1}(t,\omega))\}, \end{aligned} \tag{13}$$

где $(s(\cdot), s_{n+1}(\cdot))$ — движение, сопряженное к движению $y(\cdot, t_*, y_*, u_\alpha(\cdot), v(\cdot))$.

Схема доказательства. Покажем, что условию (13) не могут одновременно удовлетворять два различных в $L_\infty([t_*, \vartheta] \times \Omega)$ элемента. Предположим противное: существуют $u_1(\cdot)$ и $u_2(\cdot)$ из $P_u(\Delta)$, для которых верно (13). Применяя известный прием [12, гл. II, п. 2], получим из (13) неравенство

$$\begin{aligned} \Bigg\langle B(u_1(t,\omega) - u_2(t,\omega)), \\ M\left(t, X^T(\vartheta, t) \cdot \left(\frac{\partial \sigma}{\partial x}\left(W + \int_{[t_*,\vartheta]} X(\vartheta,\tau) B u_2(\tau,\omega)\, d\tau\right) - \right.\right. \\ \left.\left. - \frac{\partial \sigma}{\partial x}\left(W + \int_{[t_*,\vartheta]} X(\vartheta,\tau) B u_1(\tau,\omega)\, d\tau\right)\right)\right)\Bigg\rangle + \\ + (u_1^2(t,\omega) - u_2^2(t,\omega)) \cdot M\left(t,\ 2\alpha \int_{[t_*,\vartheta]} (u_2^2(\tau,\omega) - u_1^2(\tau,\omega))\, d\tau\right) \geq 0, \end{aligned} \tag{14}$$

где

$$W = X(\vartheta, t_*)x_* + \int_{[t_*,\vartheta]} X(\vartheta,\tau) C \overline{v}(\tau,\omega)\, d\tau.$$

Воспользовавшись неупреждаемостью функций $u_1(\cdot)$ и $u_2(\cdot)$, внесем в правой части (14) всё под единый знак $M(t, \cdot)$. Затем, проинтегрировав по $\omega \in \Omega$ и $t \in [t_*, \vartheta]$, получим:

$$\int_{\Omega}\left\{\left\langle\int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)\,d\tau-\int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)\,d\tau,\right.\right.$$

$$\left.\frac{\partial\sigma}{\partial x}\left(W+\int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)\,d\tau\right)-\frac{\partial\sigma}{\partial x}\left(W+\int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)\,d\tau\right)\right\rangle - \tag{15}$$

$$\left.-2\alpha\cdot\left(\int_{[t_*,\vartheta]}(u_1^2(\tau,\omega)-u_2^2(\tau,\omega))\,d\tau\right)^2\right\}d\omega\geq 0.$$

Из свойств строго выпуклых гладких функций следуют соотношения [13, гл. 1, предл. 5.4 и 5.5]: для всех $x_1,\ x_2\in R^n$

а)
$$\left\langle\frac{\partial\sigma}{\partial x}(x_1)-\frac{\partial\sigma}{\partial x}(x_2),x_1-x_2\right\rangle\geq 0,$$

и

б)
$$\left\langle\frac{\partial\sigma}{\partial x}(x_1)-\frac{\partial\sigma}{\partial x}(x_2),x_1-x_2\right\rangle=0,$$

тогда и только тогда, когда $x_1=x_2$.

Сопоставляя (15), а), б), заключаем: при почти всех $\omega\in\Omega$ подинтегральное выражение в (15) равно нулю. Следовательно, при почти всех $\omega\in\Omega$ верны равенства

$$\int_{[t_*,\vartheta]} u_1^2(\tau,\omega)\,d\tau=\int_{[t_*,\vartheta]} u_2^2(\tau,\omega)\,d\tau \tag{16}$$

$$\int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_1(\tau,\omega)\,d\tau=\int_{[t_*,\vartheta]} X(\vartheta,\tau)Bu_2(\tau,\omega)\,d\tau. \tag{17}$$

Отсюда при почти всех $(t,\omega)\in[t_*,\vartheta]\times\Omega$ получаем соотношения:

$$\begin{aligned}M(t,s^{(1)}(t,\omega))&=M(t,s^{(2)}(t,\omega)),\\ M(t,s_{n+1}^{(1)}(t,\omega))&=M(t,s_{n+1}^{(2)}(t,\omega)),\end{aligned} \tag{18}$$

где верхний индекс указывает, к какой из рассматриваемых программ $u_1(\cdot)$, $u_2(\cdot)$ относится компонента сопряженного движения. Из (18) и (13) следует искомое равенство.

Необходимость условия (13) нетрудно получить из необходимого условия минимума для выпуклых дифференцируемых по Гато функционалов, (см., например, [13, гл. II, предл. 2.1]).

Обозначим $B_{n+1}(r)$ шар радиуса $r \geq 0$ в пространстве R^{n+1} с центром в нуле. Построим многозначное отображение множества $B_{n+1}(r) \times \mathcal{U}_{[t_*,t^*]}$ в себя. Пусть $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ и $\{v_j(\cdot): j \in N\} \subset P_v(\Delta')$ — любая последовательность, аппроксимирующая верхнюю грань в определении величины $\varrho^*_\alpha(y_{u[\cdot]})$ (см. (8)). Этой последовательности отвечает последовательность $\{u_j(\cdot): j \in N\} \subset P_u(\Delta')$ программ, дающих нижние грани при подходящих $v_j(\cdot)$ в выражении (8) (существование таких программ обусловлено выпуклостью минимизируемого выражения и слабой компактностью множества $P_u(\Delta')$ в $L_2([t^*, \vartheta] \times \Omega')$). Данной паре последовательностей соответствует последовательность сопряженных движений, определяемых уравнениями (11). Множество всех предельных элементов множества $\{M(t^* - 0, (s^{(j)}(t^*, \cdot), s^j_{n+1}(t^*, \cdot))): j \in N\}$, когда просматриваются все возможные максимизирующие последовательности $\{v_j(\cdot): j \in N\} \subset P_v(\Delta)$, обозначим $S(u[\cdot])$. При достаточно большом r для всех $u[\cdot] \in \mathcal{U}_{[t_*,t^*]}$ будет выполнено включение $S(u[\cdot]) \subset B_{n+1}(r)$.

Каждому $(s, s_{n+1}) \in B_{n+1}(r)$, как краевому условию, поставим в соответствие решение сопряженной системы $s(t^*) = s$, $s_{n+1}(t) = s_{n+1}$, а на основе последнего построим подмножество управлений $\mathcal{U}(s, s_{n+1}) \subset \mathcal{U}_{[t_*,t^*]}$, описываемых условиями

$$\langle Bu[\tau], s(\tau)\rangle + u^2[\tau] \cdot s_{n+1}(\tau) = \\ = \min_{u \in P} \{\langle Bu, s(\tau)\rangle + u^2 \cdot s_{n+1}(\tau)\}, \quad \tau \in [t_*, t^*].$$

Можно проверить, что отображение

$$\{(s, s_{n+1}), u[\cdot]\} \longrightarrow \{s(u[\cdot]), \mathcal{U}(s, s_{n+1})\}$$

удовлетворяет условиям теоремы Какутани о неподвижной точке [14]. Пусть $\{(s^\alpha, s^\alpha_{n+1}, u_\alpha[\cdot]\}$ есть неподвижная точка этого отображения, а последовательности $\{v_j(\cdot): j \in N\}$ и $\{u_j(\cdot): j \in N\}$ указанным способом порождают $(s^\alpha, s^\alpha_{n+1})$.

Обозначим

$$u_{\alpha j}(\tau, \omega') = \begin{cases} u_\alpha[\tau], & (\tau, \omega') \in [t_*, t^*) \times \Omega', \\ u_j(\tau, \omega), & (\tau, \omega') \in [t^*, \vartheta] \times \Omega', \end{cases} \quad j \in N,$$

$$v_{\alpha j}(\tau, \omega') = \begin{cases} v_\alpha[\tau], & (\tau, \omega') \in [t_*, t^*) \times \Omega', \\ v_j(\tau, \omega), & (\tau, \omega') \in [t^*, \vartheta] \times \Omega', \end{cases} \quad j \in N, \tag{19}$$

$$u_{vj}(\cdot) = \operatorname{argmin} \{I_{\alpha j}(u(\cdot)): \quad u(\cdot) \in P_u(\Delta)\}, \quad j \in N;$$

здесь

$$I_{\alpha j}(u(\cdot)) = \int_{\Omega} \sigma_\alpha(y(\vartheta, t_*, y_*, u(\cdot,\omega), v_{\alpha j}(\cdot,\omega'))) \, d\omega.$$

В силу того, что $v_{\alpha j}(\cdot) \in P'_v(\Delta)$ программы $u_{vj}(\cdot)$ будут принадлежать $P_u(\Delta)$ (т.е. они не будут зависеть от ω_1). Для этих последовательностей (по их определению) справедливы соотношения:

$$\begin{aligned} I_{\alpha j}(u_{vj}(\cdot)) &\le \varrho_{*\alpha}, \quad j \in N, \\ \lim_{j\to\infty} I_{\alpha j}(u_{\alpha j}(\cdot)) &= \varrho^*_\alpha(y_{u_\alpha[\cdot]}). \end{aligned} \tag{20}$$

Предположим, что для некоторой подпоследовательности индексов $j(i)$ выполнено

$$\lim_{i\to\infty} \left\| u_{vj(i)}(\cdot) - u_{\alpha j(i)}(\cdot) \right\|_{L_2([t_*,\vartheta]\times\Omega')} = 0. \tag{21}$$

Тогда из равностепенной локальной липшивости функционалов $I_{\alpha j}(\cdot)$ [13, гл. 1, §2, п. 3] будет следовать

$$\lim_{i\to\infty} I_{\alpha j(i)}(u_{\alpha j(i)}(\cdot)) - I_{\alpha j(i)}(u_{vj(i)}(\cdot)) = 0. \tag{22}$$

А из (22) и (20) получим

$$\varrho_{*\alpha} \ge \varrho^*_\alpha(y_{u_\alpha[\cdot]}). \tag{23}$$

И, наконец, пользуясь леммой 1, выбрав достаточно малое $\alpha(\varepsilon)$, получим из неравенства (23) искомое неравенство (5) для $u_{\alpha(\varepsilon)}[\cdot] \in \mathcal{U}_{[t_*,t^*]}$.

Опишем кратко доказательство равенства (21). Из леммы 2 для $u_{\alpha j}(\cdot)$ и $u_{vj}(\cdot)$ следует

$$\begin{aligned} \int_{\Omega'} \Bigg\{ \Bigg\langle \int_{[t_*,\vartheta]} X(\vartheta,\tau) B(u_{\alpha j}(\tau,\omega') - u_{vj}(\tau,\omega') \, d\tau, \\ \frac{\partial\sigma}{\partial x}\left(W_j + \int_{[t_*,\vartheta]} X(\vartheta,\tau) B u_{vj}(\tau,\omega') \, d\tau \right) \Bigg\rangle + \\ +2\alpha \cdot \int_{[t_*,\vartheta]} (u^2_{\alpha j}(\tau,\omega') - u^2_{vj}(\tau,\omega')) \, d\tau \cdot \int_{[t_*,\vartheta]} u^2_{vj}(\tau,\omega') \, d\tau \Bigg\} \, d\omega' \ge 0, \end{aligned} \tag{24}$$

$$\int_{\Omega'} \left\{ \left\langle \int_{[t^*,\vartheta]} X(\vartheta,\tau) B(u_{vj}(\tau,\omega') - u_{\alpha j}(\tau,\omega')\, d\tau, \right.\right.$$

$$\left.\left. \frac{\partial \sigma}{\partial x} \left(W_j + \int_{[t^*,\vartheta]} X(\vartheta,\tau) B u_{\alpha j}(\tau,\omega')\, d\tau \right) \right\rangle + \right. \tag{25}$$

$$\left. + 2\alpha \cdot \int_{[t^*,\vartheta]} (u_{vj}^2(\tau,\omega') - u_{\alpha j}^2(\tau,\omega'))\, d\tau \cdot \int_{[t^*,\vartheta]} u_{\alpha j}^2(\tau,\omega')\, d\tau \right\} d\omega' \geq 0,$$

$$W_j = X(\vartheta,t_*)x_* + \int_{[t_*,\vartheta]} X(\vartheta,\tau) C v_{\alpha j}(\tau,\omega')\, d\tau.$$

Используя лемму 2 и свойства неподвижной точки $\{(s^\alpha, s^\alpha_{n+1}), u_\alpha[\cdot]\}$, можно установить оценку

$$\int_{\Omega'} \left\{ \left\langle \int_{[t_*,t^*]} X(\vartheta,\tau) B(u_{vj}(\tau,\omega') - u_{\alpha j}(\tau,\omega')\, d\tau, \right.\right.$$

$$\left.\left. \frac{\partial \sigma}{\partial x} \left(W_j + \int_{[t_*,\vartheta]} X(\vartheta,\tau) B u_{\alpha j}(\tau,\omega')\, d\tau \right) \right\rangle + \right. \tag{26}$$

$$\left. + 2\alpha \cdot \int_{[t_*,t^*]} (u_{vj}^2(\tau,\omega') - u_{\alpha j}^2(\tau,\omega'))\, d\tau \cdot \int_{[t_*,\vartheta]} u_{\alpha j}^2(\tau,\omega')\, d\tau \right\} d\omega' \geq -\delta_j,$$

$$\delta_j \geq 0, \quad \lim_{j\to\infty} \delta_j = 0.$$

Суммируя (24)–(26), используя а) и б), нетрудно показать, что для некоторой подпоследовательности $j(i)$ при почти всех $\omega' \in \Omega'$ имеют место соотношения:

$$\lim_{i\to\infty} \left\| \int_{[t_*,\vartheta]} u_{\alpha j(i)}^2(\tau,\omega')\, d\tau - \int_{[t_*,\vartheta]} u_{vj(i)}^2(\tau,\omega')\, d\tau \right\| = 0,$$

$$\lim_{i\to\infty} \left\| \int_{[t_*,\vartheta]} X(\vartheta,\tau) B u_{\alpha j(i)}(\tau,\omega')\, d\tau - \int_{[t_*,\vartheta]} X(\vartheta,\tau) B u_{vj(i)}(\tau,\omega')\, d\tau \right\| = 0. \tag{27}$$

Пределы (27) обусловливают сходимость сопряженных переменных, фигурирующих в соотношении (13) леммы 2, которое определяет программы $u_{\alpha_{j(i)}}(\cdot)$, $u_{v_{j(i)}}(\cdot)$. Из этой сходимости вытекает равенство (21).

Замечания

Приведенная схема доказательства остается верной для системы вида:

$$\dot{x}(t) = A(t, v[t])x(t) + B(t, v[t])u[t] + C(t)v[t], \tag{28}$$

в которой зависимость матриц $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ от t и v ограничивается лишь условиями теоремы существования и единственности решения системы (28) при любых $u[\cdot] \in \mathcal{U}_{[t_*,\vartheta]}$, $v[\cdot] \in \mathcal{V} - [t_*, \vartheta]$.

Строгая выпуклость функции $\sigma(\cdot)$ и ее гладкость не играют существенной роли. Достаточно, чтобы $\sigma(\cdot)$ была просто выпуклой функцией, т.к. при переходе к $\sigma_\alpha(\cdot)$ можно слегка деформировать $\sigma(\cdot)$, придав ей эти свойства. Лемма 1 при этом остается верна.

Автор выражает благодарность Н. Н. Красовскому за постановку задачи и обсуждение статьи.

Литература

1. Красовский Н. Н., Субботин А. И. Позиционные дифференциальные игры. М.: Наука, 1974.
2. Осипов Ю. С. Дифференциальные игры систем с последействием. Докл. АН СССР, 1971, т. **196**, № 4.
3. Осипов Ю. С. К теории дифференциальных игр в системах с распределенными параметрами. Докл. АН СССР, 1975, т. **223**, № 6.
4. Куржанский А. Б. Управление и наблюдение в условиях неопределенности. М.: Наука, 1977.
5. Субботин А. И., Ченцов А. Г. Оптимизация гарантии в задачах управления. М.: Наука, 1981.
6. Айзекс Р. Дифференциальные игры. М.: Мир, 1967.
7. *Friedman A.*, Differential games, New York, Acad. Press, 1975.
8. Черноусько Ф. Л., Меликян А. А. Игровые задачи управления и поиска. М.: Наука, 1978.
9. Красовский Н. Н., Третьяков В. Е. Стохастический программный синтез для позиционной дифференциальной игры. Докл. АН СССР, 1981, т. **259**, № 1.
10. Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. Математическая теория оптимальных процессов. М.: Физматгиз, 1961.

11. **Красовский Н. Н.** Управление динамической системой. М.: Наука, 1985.
12. **Киндерлерер Д., Стампаккья Г.** Введение в вариационные неравенства и их приложения. М.: Мир, 1983.
13. **Экланд И., Темам Р.** Выпуклый анализ и вариационные проблемы. М.: Мир, 1979.
14. **Канторович Л. В., Акилов Г. П.** Функциональный анализ. М.: Наука, 1977.

Typesetted by TYPOTEX GT, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

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: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

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. 18 · NUMBER 6
TOM HOMEP

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1989

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 PLC Headington Hill Hall, Oxford OX3 OBW, England

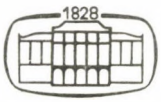
or

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

1989 Subscription Rate DM 535,— 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. 18(6), pp. 359–380 (1989)

IMPROPER MATHEMATICAL PROGRAMMING PROBLEMS

I. I. EREMIN, A. A. VATOLIN

(*Sverdlovsk*)

(Received October 10, 1988)

A mathematical programming problem is said to be proper if it is solvable together with its dual and both problems share the same optimal value, otherwise it is improper. This survey deals with finite and infinite dimensional improper problems. The duality theory for such problems is developed. Limiting processes approximating primal and dual optimal values are studied. The structure of solvability sets of minimax problems depending on parameters is investigated and correction methods for such problems are constructed.

1. Introduction

The importance of mathematical programming (MP) models in solving economic planning problems is well known. The constraint system of such problems (in mathematical terms — inequality and equality system), which describes resource and technological constraints, environmental constraints, directive requirements, etc. may turn out to be inconsistent (contradictory). Such models appear as a result of resource shortage, unrealistic planning, lack of capacities, inaccuracy of economic information, accounting of contradictory requirements, accounting of normals of the production negative influence on the environment, etc. The construction of contradictory models is rather typical in planning practice. Avoiding the contradictory nature of the simplest models by means of their correction (relaxing the constraints or disregarding some of them, information correction, etc.) does not present any difficulty. Examining more complex models, describing more complicated situations leads to the necessity of accounting appearance of improper (contradictory) models and, hence, working out their theory (above all — duality), methods of numerical analysis and the corresponding software [1].

Let

$$\sup\{f(x) | f_j(x) \le 0, \quad j = 1, \dots, m, \ x \ge 0\} \tag{1.1}$$

be an MP problem over R^n, and

$$\sup\{\langle c, x\rangle | Ax \le b, \ x \ge 0\} \tag{1.2}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

its linear case, i.e. a linear programming (LP) problem. Let

$$\Phi(x,u) = f(x) - \sum_{j=1}^{m} u_j f_j(x) \text{ and } L(x,u) = \langle c,x\rangle - \langle Ax-b,u\rangle$$

be the Lagrangean functions of problems (1.1) and (1.2), respectively. Note that (1.1) may be equivalently presented as $\sup\limits_{x\geq 0}\ \inf\limits_{u\geq 0} \Phi(x,u) (= \overline{\Phi})$. The problem

$$\inf_{u\geq 0}\sup_{x\geq 0}\Phi(x,u) = \underline{\Phi} \tag{1.1}*$$

is said to be dual to (1.1). If $\Phi(x,u) \equiv L(x,u)$, then the dual of (1.2) can be reduced to the form

$$\inf\{\langle b,u\rangle | A^T u \geq c,\ u\geq 0\}. \tag{1.2}*$$

Problem (1.1) is said to be proper, if

$$\overline{\Phi} = \underline{\Phi} = \Phi(\overline{x},\overline{u}) = \sup_{x\geq 0}\Phi(x,\overline{u}) = \inf_{u\geq 0}\Phi(\overline{x},u), \quad \overline{x}\geq 0, \quad \overline{u}\geq 0;$$

otherwise it is improper. The emptiness of the set $\{x\geq 0 | f_j(x)\leq 0,\ j=1,\ldots,m\}$ is a special case of impropriety. Problem (1.2) is proper iff it is solvable. Improper LP (and also MP) problems can be classified according to whether each of the feasible sets M and M^* of problems (1.2) and (1.2)*, respectively, is empty or not: 1. $M=\emptyset$, $M^*\neq\emptyset$; 2. $M\neq\emptyset$, $M^*=\emptyset$; 3. $M=\emptyset$, $M^*=\emptyset$ (improper problems of the 1st, 2nd and 3rd kind).

By approximation (correction) of an improper MP problem we shall mean a formal but substantially interpretable way of reducing (mapping) it to a proper problem. Consider two examples.

Example 1. With problems (1.1) and (1.2) we associate

$$\sup\{f(x) - \langle \Delta c, x\rangle | f_j(x)\leq b_j + \Delta b_j, \quad j=1,\ldots,m, x\geq 0\}, \tag{1.3}$$

$$\sup\{\langle c-\Delta c, x\rangle | Ax\leq b+\Delta b, \quad x\geq 0\}; \tag{1.4}$$

here $(\Delta c, \Delta b) = \Delta$ are correcting parameters. Let K_0 be a feasible set of Δ's, $K = \{\Delta\in K_0 |$; problem (1.4) is proper$\}$, $\varphi(\Delta)$ be a function estimating the quality of approximation (correction). Consider an approximation problem of the form

$$\min\{\varphi(\Delta) | \Delta\in K\}. \tag{1.5}$$

If $\tilde{\Delta} = (\Delta\tilde{c}, \Delta\tilde{b})$ is an optimal solution of problem (1.5), then $\max\{\langle c-\Delta\tilde{c}, x\rangle | Ax \leq \leq b+\Delta\tilde{b},\ x\geq 0\}$ may be considered as a compromise model.

Example 2. Let the restriction system of problem (1.1) be splitted into two subsystems: $f_j(x) \le 0$, $j \in J_0$, $x \ge 0$, and $f_j(x) \le 0$, $j \in J_1$, the first of which being consistent. Using a residual function $d(x)$ (e.g. $d(x) = \sum\limits_{j \in J_1} \overline{u}_j f_j^+(x)$, $\overline{u}_j > 0$, $j \in J_1$) we may introduce a set $\tilde{M} = \operatorname{Arg\,min}\{d(x) | f_j(x) \le 0,\ j \in J_0,\ x \ge 0\}$ and an approximation problem

$$\max\{f(x) | x \in \tilde{M}\}.$$

Here and henceforth Arg P denotes the optimal set of problem P; if $\alpha \in R$, then $\alpha^+ = \max\{0, \alpha\}$; if $x = (x_1, \ldots, x_k)^T \in R^k$, then $x^+ = (x_1^+, \ldots, x_k^+)^T$; $R_+^k = \{x \in R^k | x \ge 0\}$.

2. Duality in improper LP problems [1, 2]

Duality plays the central role in the theory of mathematical programming and is the source of methods for solving problems of this class. This is also true for improper MP problems. The duality scheme for improper MP problems P and P^* can be described as follows: problems C and $C^\#$ are associated with the dual MP problems P and P^* according to the same transformation scheme π; problems C and $C^\#$ are connected by duality relations; they play the role of approximating problems for P and P^*. Let us show this in a scheme:

$$\begin{array}{ccc} & \overset{\pi}{\longrightarrow} & \\ P & & C \\ (*)\ \downarrow & & \downarrow\ (\#) \\ & \overset{\pi}{\longrightarrow} & \\ P^* & & C^\# \end{array} \qquad (\#)$$

If P and P^* are problems (1.2) and (1.2)*, then C and $C^\#$ may take the form:

$$\begin{aligned} C:\quad & \sup\{\langle c, x\rangle - \sum_{j=1}^{m_0} R_j \|(A_j x - b^j)^+\|_{p(j)} | \\ & A_0 x \le b^0,\ x \ge 0,\ \|x^i\|_{q(i)} \le r_i,\ i = 1, \ldots, n_0\}, \\ C^\#:\quad & \inf\{\langle b, u\rangle + \sum_{i=1}^{n_0} r_i \|(c^i - B_i^T u)^+\|_{q(i)}^* | \\ & B_0^T u \ge c^0,\ u \ge 0,\ \|u^j\|_{p(j)}^* \le R_j,\ j = 1, \ldots, m_0\}. \end{aligned}$$

Here $R_j \ge 0$, $r_i \ge 0$, $j = 1, \ldots, m_0$, $i = 1, \ldots, n_0$ are parameters; $\{\|\cdot\|_{p(j)}\}$, $\{\|\cdot\|_{q(i)}\}$ are arbitrary monotone (on the non-negative orthant of the corresponding space) norms; $\{\|\cdot\|_{p(j)}^*\}$, $\{\|\cdot\|_{q(i)}^*\}$ are norms dual to them (these norms are also assumed to be monotone);

$$A = \begin{pmatrix} A_0 \\ \vdots \\ A_{m_0} \end{pmatrix} = (B_0, \ldots, B_{n_0});$$

the partitions of A into $\{A_j\}$ and $\{B_i\}$ determine the partitions $c = (c^0, \ldots, c^{n_0})^T$, $x = (x^0, \ldots, x^{n_0})^T$ and $b = (b^0, \ldots, b^{m_0})^T$, $u = (u^0, \ldots, u^{m_0})^T$, respectively.

Consider, e.g. the following norms, which are monotone (i.e. $x \geq y \geq 0$ implies $\|x\| \geq \|y\|$) together with their duals:

$$\|w\|_0 = \max_{1 \leq j \leq k} \alpha_j |w_j|, \qquad \|w\|_0^* = \sum_{j=1}^{k} \alpha_j^{-1} |w_j|,$$

$$\|w\|_1 = \sum_{j=1}^{k} \alpha_j |w_j|, \qquad \|w\|_1^* = \max_{1 \leq j \leq k} \alpha_j^{-1} |w_j|,$$

$$\|w\|_2 = \left(\sum_{j=1}^{k} \alpha_j w_j^2 \right)^{1/2}, \qquad \|w\|_2^* = \left(\sum_{j=1}^{k} \alpha_j^{-1} w_j^2 \right)^{1/2},$$

where $w = (w_1, \ldots, w_k) \in R^k$, $\alpha_j > 0$, $j = 1, \ldots, k$.

We shall assume that $M_0 = \{x \geq 0 | A_0 x \leq b^0\} \neq \emptyset$, $M_0^{\#}\{u \geq 0 | B_0^T u \geq c^0\} \neq \neq \emptyset$. If necessary, we can assign the value $\emptyset$ to some of the submatrices $\{B_i\}$, $\{A_j\}$. Let $f_R(x)$ and $f_r^{\#}(u)$ be objective functions and $M(r)$ and $M^{\#}(R)$ be feasible sets of the problems C and $C^{\#}$, where $R = (R_1, \ldots, R_{m_0})$, $r = (r_1, \ldots, r_{n_0})$.

Theorem 2.1. 1) If $\bar{x} \in M(r)$, $\bar{u} \in M^{\#}(R)$, then $f_R(\bar{x}) \leq f_r^{\#}(\bar{u})$ so that $f_R(\bar{x}) = f_r^{\#}(\bar{u})$ implies $\bar{x} \in \operatorname{Arg} C$, $\bar{u} \in \operatorname{Arg} C^{\#}$.

2) Let $\|\bar{x}^i\|_i < r_i$, $i = 1, \ldots, n_0$ for some $\bar{x} \in M_0$ and let problem C be solvable; then problem $C^{\#}$ is solvable, and the optimal values of the two problems are equal.

The inverse of 2) also holds.

We represent some particular realizations of problems C and $C^{\#}$:

$$C_0: \quad \max\{\langle c, x\rangle - \langle R, (Ax - b)^+\rangle | \, 0 \leq x \leq r\},$$

$$C_0^{\#}: \quad \min\{\langle b, u\rangle + \langle r, (c - A^T u)^+\rangle | \, 0 \leq u \leq R\},$$

$$C_1: \quad \max\{\langle c, x\rangle - R_0 \|(Ax - b)^+ \| x \geq 0\},$$

$$C_1^{\#}: \quad \min\{\langle b, u\rangle | A^T u \geq c, u \geq 0, \|u\|^* \leq R_0\},$$

$$C_2: \quad \max\{\langle c, x\rangle | Ax \leq b, x \geq 0, \|x\| \leq r_0\},$$

$$C_2^{\#}: \quad \min\{\langle b, u\rangle + r_0 \|(c - A^T u)^+\|^* | \, u \geq 0\},$$

where all norms are monotone.

Theorem 2.2. 1) Problems C_0 and $C_0^{\#}$ are solvable and their optimal values are equal (for arbitrary realization of (1.2)).

2) If $\|\overline{u}\| < R_0$ for some $\overline{u} \in M^* = \{u \geq 0 | A^T u \geq c\}$, then problems C_1 and $C_1^{\#}$ are solvable and their optimal values are equal.

3) If $\|\overline{x}\| < r_0$ for some $\overline{x} \in M = \{x \geq 0 | Ax \leq b\}$, then problems C_2 and $C_2^{\#}$ are solvable and their optimal values are equal.

We shall now present a more general realization of the scheme (#).

Let R, ϵ_i, R_j, r_i, δ_j, $i = 1, \ldots, n_0$, $j = 1, \ldots, m_0$ be positive parameters; let Φ and Ψ be functions over R^ν, where $\nu = n + m - \nu_1 - \nu_2$, $x^0 \in R^{\nu_1}$, $u^0 \in R^{\nu_2}$. With problems (1.2) and (1.2)* we associate

$$D: \quad \sup\{\langle c, x\rangle - R\Phi(\varepsilon_1 x^1, \ldots, \varepsilon_{n_0} x^{n_0}; R_1(A_1 x - b^1)^+, \ldots \\ \ldots, R_{m_0}(A_{m_0} x - b^{m_0})^+) | A_0 x \leq b^0,\ x \geq 0\}, \tag{2.1}$$

$$D^{\#}: \quad \inf\{\langle b, u\rangle + R\Psi(r_1(c^1 - B_1^T u)^+, \ldots, r_{n_0}(c^{n_0} - B_{n_0}^T u)^+; \\ \delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}) | B_0^T u \geq c^0,\ u \geq 0\}. \tag{2.2}$$

We introduce the notation for the arguments of Φ and Ψ:

$$\Gamma(x) = (\varepsilon_1 x^1, \ldots, \varepsilon_{n_0} x^{n_0}; R_1(A_1 x - b^1)^+, \ldots, R_{m_0}(A_{m_0} x - b^{m_0})^+),$$
$$\Gamma^{\#}(u) = (r_1(c^1 - B_1^T u)^+, \ldots, r_{n_0}(c^{n_0} - B_{n_0}^T u)^+, \delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}).$$

We shall now show how to choose function Φ and Ψ and the parameters in D and $D^{\#}$ to ensure the duality between (2.1) and (2.2). A function Ω mapping R^k into $R^1 U\{+\infty\}$ is said to be admissible if (a) Ω is convex and lower semi-continuous, dom $\Omega \subset R_+^k$, and (b) Ω is a monotone non-decreasing function on R_+^k and $\Omega(0) = 0$. We assume that Φ and Ψ are admissible functions.

Let Ω^* be a conjugate function of Ω. Put

$$\Omega^{\#}(z) = \Omega^*(z) + i(z | R_+^k),$$

where $i(\cdot | W)$ is the indicator function of W, and it is zero if z is in W and $+\infty$, otherwise. The following property holds: if Ω is admissible, then $\Omega^{\#}$ is also admissible and $(\Omega^{\#})^{\#} = \Omega$. In what follows it will be assumed that $\Psi = \Phi^{\#}$ and

$$\varepsilon_i r_i = \delta_j R_j = R^{-1},\ i = 1, \ldots, n_0,\ j = 1, \ldots, m_0. \tag{2.3}$$

Remark 2.1. The systems $A_0 x \leq b^0$, $x \geq 0$ and $B_0^T u \geq c^0$, $u \geq 0$ are assumed to be consistent. This may correspond to the selection of directive restrictions, which are justified (i.e. their fulfilment is possible).

Remark 2.2. Using the property $(\Phi^{\#})^{\#} = \Phi$ it is easily seen that the scheme (#) : $D \xrightarrow{\#} D^{\#}$ has the reciprocity property, i.e. $(D^{\#})^{\#} = D$.

Remark 2.3. D and $D^{\#}$ are convex programs.

General form (2.1)–(2.2) of the problems D and $D^{\#}$ implies a great number of particular realizations corresponding to a variety of approaches to the approximation of improper problems (1.2) and (1.2)*. First of all we note that if $n_0 = 0$, $m_0 = 0$, then D and $D^{\#}$ convert into (1.2) and (1.2)*. If, e.g., (1.2) is an improper problem of the 1st kind, then we can put $n_0 = 0$ in (2.1) and (2.2), and problems D and $D^{\#}$ become

$$D_1: \quad \sup\{\langle c, x\rangle - R\Phi(R_1(A_1x - b^1)^+, \ldots, R_{m_0}(A_{m_0}x - b^{m_0})^+)| \\ A_0x \le b^0, \quad x \ge 0\}, \tag{2.4}$$

$$D_1^{\#}: \quad \inf\{\langle b, u\rangle + R\Phi^{\#}(\delta_1 u^1, \ldots, \delta_{m_0}u^{m_0})| \\ A^T u \ge c, \quad u \ge 0\}. \tag{2.5}$$

If Φ in D is a piecewise linear convex function, then D is called an l-problem. The objective functions in D and $D^{\#}$ are

$$f(x) = \langle c, x\rangle - R\Phi(\Gamma(x)), \quad f^{\#}(u) = \langle b, u\rangle + \Phi^{\#}(\Gamma^{\#}(u)).$$

Let f, $f^{\#}$ and $\tilde{M}$, $\tilde{M}^{\#}$ be optimal values and optimal sets of problems (2.1) and (2.2), respectively. Put

$$M = \{x \ge 0 | A_0x \le b^0, \ \Gamma(x) \in \text{dom } \Phi\},$$
$$M^{\#} = \{u \ge 0 | B_0^T u \ge c^0, \ \Gamma^{\#}(u) \in \text{dom } \Phi^{\#}\}.$$

We can now formulate a statement concerning the duality relating problems (2.1) and (2.2).

Theorem 2.3. The following statements hold:

1) $\tilde{f} \le \tilde{f}^{\#}$.

2) If $\tilde{f} < +\infty$ and D satisfies a constraint qualification (we do not concretize it here), then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$ and $\tilde{M}^{\#} \ne \emptyset$.

3) If $\tilde{f}^{\#} > -\infty$ and $D^{\#}$ satisfies a constraint qualification, then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$ and $\tilde{M} \ne \emptyset$.

4) If $\{x \in M | f(x) \ge \alpha\}$ is non-empty and bounded for some α, then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$ and $\tilde{M} \ne \emptyset$.

5) If $\{u \in M^{\#} | f^{\#}(u) \le \alpha\}$ is non-empty and bounded for some α, then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$ and $\tilde{M}^{\#} \ne \emptyset$.

6) Let D be an l-problem and either M or $M^{\#}$ be non-empty. Then $\tilde{f} = \tilde{f}^{\#}$, moreover, in this case $\tilde{f} \ne \pm\infty$ implies $\tilde{M} \ne \emptyset$ and $\tilde{M}^{\#} \ne \emptyset$.

We shall now consider the case when Φ depends on the norms of the residuals $(A_jx - b^j)^+$:

$$\Phi(\Gamma(x)) = \varphi(\varepsilon_1\|x^1\|_{q(1)}, \ldots, \varepsilon_{n_0}\|x^{n_0}\|_{q(n_0)}; \\ R_1\|(A_1x - b^1)^+\|_{p(1)}, \ldots, R_{m_0}\|(A_{m_0}x - b^{m_0})^+\|_{p(m_0)}) + \\ + i(\Gamma(x)|R_+^{\nu}), \tag{2.6}$$

here $\{\|\cdot\|_{q(i)}\}$, $\{\|\cdot\|_{p(j)}\}$ are monotone (together with their duals), and φ is a function on $R^{n_0+m_0}$.

Proposition. Let φ be a lower semi-continuous proper convex function on $R^{n_0+m_0}$ and a monotone non-decreasing function on $R_+^{n_0+m_0}$; let $\varphi(0) = 0$ and either $\varphi(v) = \varphi(|v_1|, \ldots, |v_{n_0+m_0}|)$ for all $v \in R^{n_0+m_0}$ or $\operatorname{dom}\varphi \subset R_+^{n_0+m_0}$. Then Φ defined in (2.6) is admissible and

$$\Phi^{\#}(\Gamma^{\#}(u)) = \varphi^*(r_1\|(c^1 - B_1^T u)^+\|_{q(1)}^*, \ldots, r_{n_0}\|(c^{n_0} - B_{n_0}^T u)^+\|_{q(n_0)}^*;$$

$$\delta_1\|u^1\|_{p(1)}^*, \ldots, \delta_{m_0}\|u^{m_0}\|_{p(m_0)}^*) + i(\Gamma^{\#}(u)|R_+^{\nu}).$$

Thus, for Φ defined in (2.6), D and $D^{\#}$ become

$$\sup\{\langle c, x\rangle - R\varphi(\varepsilon_1\|x^1\|_{q(1)}, \ldots, \varepsilon_{n_0}\|x^{n_0}\|_{q(n_0)};$$

$$R_1\|(A_1 x - b^1)^+\|_{p(1)}, \ldots, R_{m_0}\|(A_{m_0} x - b^{m_0})^+\|_{p(m_0)})|A_0 x \le b^0,\ x \ge 0\}, \quad (2.7)$$

$$\inf\{\langle b, u\rangle + R\varphi^*(r_1\|(c^1 - B_1^T u)^+\|_{q(1)}^*, \ldots, r_{n_0}\|(c^{n_0} - B_{n_0}^T u)^+\|_{q(n_0)}^*;$$

$$\delta_1\|u^1\|_{p(1)}^*, \ldots, \delta_{m_0}\|u^{m_0}\|_{p(m_0)}^*|B_0^T u \ge c^0,\ u \ge 0\}. \quad (2.8)$$

Put, e.g.

$$\varphi(v) = \sum_{i=1}^{n_0} \alpha_i^{-1}|v_i|^{\alpha_i} + \sum_{j=1}^{m_0} \beta_j^{-1}|v_{n_0+j}|^{\beta_j}. \quad (2.9)$$

Here $1 \le \alpha_i \le +\infty$, $1 \le \beta_j \le +\infty$, $i = 1, \ldots, n_0$, $j = 1, \ldots, m_0$. The following laws are used to operate with $\pm\infty$: $-(+\infty) = -\infty$, $\gamma \pm \infty = \pm\infty$,

$$(+\infty)^{-1}\gamma^{+\infty} = \begin{cases} 0, & 0 \le \gamma \le 1, \\ +\infty, & \gamma > 1 \end{cases}$$

for each real γ. Then direct calculation shows that $\omega^*(\gamma) = \sigma^{-1}|\gamma|^{\sigma}$ for $\omega(\gamma) = \alpha^{-1}|\gamma|^{\alpha}$, where $1 \le \alpha \le +\infty$, $\alpha^{-1} + \sigma^{-1} = 1$ (in the latter equality we put $(+\infty)^{-1} = 0$ if $\alpha = +\infty$ or $\sigma = +\infty$). Thus, we have

$$\varphi^*(v) = \sum_{i=1}^{n_0} \sigma_i^{-1}|v_i|^{\sigma_i} + \sum_{j=1}^{m_0} \tau_j^{-1}|v_{n_0+j}|^{\tau_j},$$

where σ_i, τ_j satisfy

$$\alpha_i^{-1} + \sigma_i^{-1} = 1, \quad \beta_j^{-1} + \sigma_j^{-1} = 1, \quad 1 \le \sigma_i \le +\infty,$$

$$1 \le \tau_j \le +\infty, \quad i = 1, \ldots, n_0, \quad j = 1, \ldots, m_0. \quad (2.10)$$

Note that φ defined in (2.9) satisfies the assumptions of the Proposition; so we put $R = 1$ and obtain the following particular realization of problems (2.7) and (2.8):

$$\sup\{\langle c, x\rangle - \sum_{i=1}^{n_0} \alpha_i^{-1} \|\varepsilon_i x^i\|_{q(i)}^{\alpha_i} - \sum_{j=1}^{m_0} \beta_j^{-1} \|R_j (A_j x - b^j)^+\|_{p(j)}^{\beta_j} | \quad A_0 x \le b^0,\ x \ge 0\} \tag{2.11}$$

$$\inf\{\langle b, u\rangle + \sum_{i=1}^{n_0} \sigma_i^{-1} \|r_i (c^i - B_i^T u)^+\|_{q(i)}^{*\sigma_i} + \sum_{j=1}^{m_0} \tau_j^{-1} \|\delta_j u^j\|_{p(j)}^{*\tau_j} | B_0^T u \ge c^0,\ u \ge 0\}, \tag{2.12}$$

where parameters ε_i, r_i, R_j, δ_j, α_i, β_j, σ_i, τ_j satisfy (2.3) and (2.10) with $R = 1$.

We present some examples of the problems obtained from (2.11), (2.12) by the variation of the parameters:

1) If $B_0 = A$, $n_0 = 1$, $\alpha_1 = \sigma_1 = 2$, $\beta_j = 1$, $\tau_j = +\infty$, $j = 1, \dots, m_0$, we obtain the problems:

$$\sup\{\langle c, x\rangle - \sum_{j=1}^{m_0} R_j \|(A_j x - b^j)^+\|_{p(j)} - \varepsilon \|x\|_q^2 | A_0 x \le b^0,\ x \ge 0\},$$

$$\inf\{\langle b, u\rangle + (4\varepsilon)^{-1} \|(c - A^T u)^+\|_q^{*2} | u \ge 0,\ \|u^j\|_{p(j)}^* \le R_j,\ j = 1, \dots, m_0\},$$

where R_j and ε are arbitrary positive parameters. In particular, if $\|x\|_q$ is the Euclidean norm, then the first of the above problems can be considered as the Tikhonov [3] regularization of the problem

$$\sup\{\langle c, x\rangle - \sum_{j=1}^{m_0} R_j \|(A_j x - b^j)^+\|_{p(j)} | A_0 x \le b^0,\ x \ge 0\}$$

approximating improper (of the 1st kind) problem (1.2).

2) If $A_0 \ne \emptyset$, $B_0 = A$, $m_0 = 1$, $\beta_1 = \tau_1 = 2$, we obtain the problems

$$\sup\{\langle c, x\rangle - R\|(Ax - b)^+\|_p^2 | x \ge 0\},$$

$$\inf\{\langle b, u\rangle + \delta \|u\|_p^{*2} | A^T u \ge c,\ u \ge 0\}$$

(where $R > 0$, $\delta > 0$, $R\delta = 1/4$), which correspond to the case when problem (1.2) is improper of the 1st kind.

3) If $A_0 \ne \emptyset$, $B_0 \ne \emptyset$, $m_0 = n_0 = 1$, $\alpha_1 = \beta_1 = \sigma_1 = \tau_1 = 2$, we obtain the following symmetric form of problems (2.11), (2.12):

$$\sup\{\langle c, x\rangle - R\|(Ax - b)^+\|_p^2 - \varepsilon \|x\|_p^2 | x \ge 0\},$$

$$\inf\{\langle b, u\rangle + r\|(c - A^T u)^+\|_q^{*2} + \delta \|u\|_p^{*2} | u \ge 0\}$$

corresponding to the case when problems (1.2) and (1.2)* are improper of the 3rd kind. Here R, ε, r, δ are positive parameters satisfying $\varepsilon r = R\delta = 1/4$.

4) If $\alpha_j = +\infty$, $\beta_j = 1$, $\sigma_i = 1$, $\tau_j = +\infty$, $j = 1, \dots, m_0$, $i = 1, \dots, n_0$, we obtain the problems C and $C^{\#}$ proposed by one of the authors in [4].

3. Duality in improper convex programming problems

Consider the convex programming problem

$$\sup\{f_0(x) \mid f_j(x) \leq 0,\ j = 1, \ldots, m,\ x \in M\}, \tag{3.1}$$

where $\{-f_0, f_1, \ldots, f_m\}$ are proper convex functions, M is a convex set. We assume that

$$M = \operatorname{dom}(-f_0),\ M \subset \operatorname{dom} f_j,\ \operatorname{ri} M \subset \operatorname{ri} \operatorname{dom} f_j,\ j = 1, \ldots, m,$$

where 'ri' stands for the relative interior. Set

$$F(x,u) = f_0(x) - \sum_{j=1}^{m} u_j f_j(x),\ u = (u_1, \ldots, u_m)^T,$$

$$M^* = \{u \mid \sup_{x \in M} F(x,u) < +\infty\},\ g_0(u) = \sup_{x \in M} F(x,u).$$

Then the problem dual to (3.1) becomes

$$\inf\{g_0(u) \mid u \in M^*,\ u \geq 0\}. \tag{3.2}$$

Let (3.1) be an improper problem of the 1st kind, i.e. its feasible set is empty and the feasible set of (3.2) is non-empty. As in Section 2, we fix a partition $(f_1, \ldots, f_m) = (F_0, \ldots, F_m)$, so that the $F_j(x)$ are vectors of appropriate dimensions. With problems (3.1), (3.2) we associate

$$\sup\{f_0(x) - R\Phi(R_1 F_1^+(x), \ldots, R_{m_0} F_{m_0}^+(x)) \mid F_0(x) \leq 0,\ x \in M\}, \tag{3.3}$$

$$\inf\{g_0(u) + R\Phi^{\#}(\delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}) \mid u \in M^*,\ u \geq 0\} \tag{3.4}$$

where positive parameters R, R_j, δ_j satisfy $R_j \delta_j = R^{-1}$, $j = 1, \ldots, m_0$; Φ is an admissible function.

Let $\tilde{f}$, $\tilde{f}^{\#}$ and $\tilde{M}$, $\tilde{M}^{\#}$ be optimal values and optimal sets of problems (3.3), (3.4). Let $f(x)$ be the objective function in (3.3). Put $\Phi'(z) \equiv \Phi(z^+)$, $\Gamma_0(x) = (R_1 F_1(x), \ldots, R_{m_0} F_{m_0}(x))$.

Theorem 3.1 [2]. The following statements hold:

1) $\tilde{f} \le \tilde{f}^{\#}$.

2) Let $\tilde{f} < +\infty$ and $\overline{x} \in \operatorname{ri} M$, $\Gamma_0(\overline{x}) \in \operatorname{ri} \operatorname{dom} \Phi'$ for some feasible $\overline{x}$ such that $f_j(\overline{x}) < 0$ for all non-affine f_j entering into F_0. Then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$, $\tilde{M}^{\#} \neq \emptyset$.

3) Let the functions $\{-f_0, f_1, \ldots, f_m\}$ be lower semicontinuous, and let the set $\{x \in M | f(x) \ge \alpha,\ F_0(x) \le 0\}$ be non-empty and bounded for some α. Then $\tilde{f} = \tilde{f}^{\#}$, $\tilde{M} \neq \emptyset$.

Now, let (3.1) be an arbitrary convex programming problem with $M = R^n_+$, i.e.

$$C: \quad \sup\{f_0(x) | f_j(x) \le 0,\ j = 1, \ldots, m,\ x \ge 0\}. \tag{3.5}$$

For ease of presentation, we assume that the functions f_j, $j = 0, 1, \ldots, m$, are differentiable.

The transfer $g(x) \xrightarrow{(l)} l_p(x)$ from a differentiable function $g(x)$ to the linear function

$$l_p(x) = \langle \nabla g(p), x - p \rangle + g(p),$$

where $p \in R^n$, is called a linearization of g at the point p. The transfer $C \xrightarrow{(l)} L_p$ from the MP problem C to the LP problem L_p occuring at the linearization of all f_j in C is called a linearization of C (at the point p). We obtain the following duality scheme:

$$C \xrightarrow{(l)} L_p \xrightarrow{(*)} L_p^* \xrightarrow{(x=p)} L_x^* \equiv C^*. \tag{3.6}$$

By (3.6), we have

$$C^*: \quad \inf\{F(x,u) - \langle v, x \rangle | \nabla_x F(u,x) \ge 0,\ (x,v) \ge 0\},$$

where $F(x,u) = f_0(x) - \sum\limits_{j=1}^{m} u_j f_j(x)$. If we remove the constraint $x \ge 0$ from (3.5), then C^* becomes

$$\inf\{F(x,u) | \nabla_x F(x,u) = 0,\ u \ge 0\}.$$

If (3.5) is an improper problem, then the following transformation of (3.6) suggests itself:

$$\begin{array}{ccccccc}
C & \xrightarrow{(l)} & L_p & \xrightarrow{\pi} & P_p & \xrightarrow{(p=x)} & P_x \equiv P \\
 & & {\scriptstyle (*)}\downarrow\uparrow & & \downarrow\uparrow & & \downarrow {\scriptstyle (\#)} \\
 & & L_p^* & \xrightarrow[\Pi']{\pi} & P_p^{\#} & \xrightarrow{(p=x)} & P_x^{\#} \equiv P^{\#} \\
\downarrow & & & & & & \uparrow \\
C^* & & & & & & \longrightarrow
\end{array} \tag{3.7}$$

Denoting by Π the transition $C \longrightarrow P$ in (3.7), we have

$$\begin{array}{ccc} C & \xrightarrow{\Pi} & P \\ \downarrow & & \downarrow (\#) \\ C^* & \xrightarrow{\Pi'} & P^{\#} \end{array} \tag{3.8}$$

By (3.8) (where the transition $C \xrightarrow{\#} C^{\#}$ of Section 2 is applied), we get

$$P: \quad \sup\{f_0(x) - \sum_{j=1}^{m_0} R_j \|F_j(x)\|_{p(j)} |$$
$$F_0(x) \le 0, x \ge 0, \|x^i\|_{q(i)} \le r_i, i = 1, \ldots, n_0\} \ (= \tilde{f}); \tag{3.9}$$

where $(f_1, \ldots, f_m) = (F_0, \ldots, F_{m_0}) = F(x)$. System $F_0(x) \le 0$, $x \ge 0$ is assumed to be consistent.

To formulate problem $P^{\#}$ in (3.8), we introduce the notation

$$A^x = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1}, & \ldots, & \frac{\partial f_1(x)}{\partial x_n} \\ \ldots, & \ldots, & \ldots \\ \frac{\partial f_m(x)}{\partial x_1}, & \ldots, & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}, \quad b_x = A^x x - F(x),$$

$$c_x = \nabla f_0(x) = \left(\frac{\partial f_0(x)}{\partial x_1}, \ldots, \frac{\partial f_0(x)}{\partial x_n} \right)^T.$$

Let the partition A_j^x, $B_i(x)$, b_x^j, c_x^i, $j = 0, 1, \ldots, m_0$, $i = 0, 1, \ldots, n_0$, of A^x, b_x, c_x correspond to that of A, b, c in Section 2 such that the horizontal partition corresponds to the partition $(f_1, \ldots, f_m) = (F_0, \ldots, F_{m_0})$ in (3.9). Now we have

$$P^{\#}: \quad \inf\{F(x,u) - \langle \nabla_x F(x,u), x \rangle + \sum_{i=1}^{n_0} r_i \|(c_x^i - B_i^T(x)u)^+\|_{q(i)}^* |$$

$$B_0^T(x)u \ge c_x^0, \quad u \ge 0,$$
$$\|u^j\|_{p(j)}^* \le R_j, \ j = 1, \ldots, m_0\} \ (= \tilde{f}^{\#}). \tag{3.10}$$

Note that a more unwieldy presentation of P could be obtained using the notation A_j^x, b_x^j, but this would be equivalent to (3.9). Let $\Phi(x)$ and M denote the objective function and the feasible set in (3.9).

Theorem 3.2. The following statements hold for problems (3.9) and (3.10):

1) $\tilde{f} \leq \tilde{f}^{\#}$.

2) Let P be solvable and satisfy a constraint qualification. Then $P^{\#}$ is solvable and $\tilde{f} = \tilde{f}^{\#}$.

3) If, for some α, the set $M \cap \{x|\Phi(x) \geq \alpha\}$ is non-empty and bounded, then P is solvable and $\tilde{f} = \tilde{f}^{\#}$.

4) Let (3.5) be an LP problem and all the norms in (3.9) be piecewise linear. If the feasible sets of P and $P^{\#}$ are non-empty, then these problems are solvable and $\tilde{f} = \tilde{f}^{\#}$.

Remark 3.1. As before, all the norms in (3.9) and (3.10) are assumed to be monotone.

4. Duality for improper problems in infinite dimensional spaces [2,5]

The most obvious difficulty hindering the transference of the duality results to the case of infinite dimensional spaces is the impossibility of using operation "+" (projection on R^n_+) appearing in P and $P^{\#}$. Various approaches to the generalization of this operation (such as, in Hilbert space, interpreting it as a projection on the cone dual to the cone of "non-negative vectors" or, in Banach space, generalizing it using even more complex construction involving two pairs of dual cones) lead to the necessity of imposing some additional restrictions on the cones, norms and functions employed. As it seemed important for us, on the one hand, to retain the generality of the assumptions and, on the other hand, to preserve the simplicity and the naturalness of the whole construction, a different concept of generalization was taken. Namely, a general pair of problems D and $D^{\#}$ is constructed below, which, in particular, includes all formulations obtained by the generalization of operation "+".

Consider the following dual pair of LP problems:

$$L: \quad \sup\{\langle c, x\rangle | b - Ax \in K,\ x \in C\}, \tag{4.1}$$

$$L^*: \quad \inf\{\langle b, u\rangle | A^*u - c \in C^*,\ u \in K^*\}. \tag{4.2}$$

Here $x \in X$, $u \in U$, $c \in X^*$, $b \in U^*$; X, U are real, locally convex spaces with continuous duals X^*, U^*; $A: X \to U^*$ is a continuous linear operator and A^* is its conjugate; $C \subset X$ and $K \subset U^*$ are closed convex cones; $C^* = \{x^* \in X^* | \langle x^*, x\rangle \geq 0\ \forall x \in C\}$, $K^* = \{u \in U | \langle u, u^*\rangle \geq 0\ \forall u^* \in K\}$.

In order to introduce the partition of A we need a more concrete definition of problems L and L^*. Let

$$X = \prod_{i=1}^{n} X_i,\ U = \prod_{j=1}^{m} U_j,\ X^* = \prod_{i=1}^{n} X_i^*,\ U^* = \prod_{j=1}^{m} U_j^*,$$

where X_i, X_i^*, U_j, U_j^* are reflexive Banach spaces and their (strong) conjugates. If we put $\langle x^*, x\rangle = \sum_{i=1}^{n} \langle x_i^*, x_i\rangle$, where $x = (x_1, \ldots, x_n) \in X_1 \times \ldots \times X_n = X$, $x^* = (x_1^*, \ldots, x_n^*) \in X^*$, then X and X^* (and, similarly, U and U^*) are reflexive Banach spaces, e.g. with the norms

$$\|x\| = \sum_{i=1}^{n} \|x_i\|, \quad \|x^*\| = \max_{i \le i \le n} \|x_i^*\|.$$

Set

$$C = C_1 \times \ldots \times C_n, \quad K = K_1 \times \ldots \times K_m,$$
$$C^* = C_1^* \times \ldots \times C_n^*, \quad K^* = K_1^* \times \ldots \times K_m^*,$$

where $C_i \subset X_i$, $C_i^* \subset X_i^*$, $K_j \subset U_j^*$, $K_j^* \subset U_j$ are closed convex cones and their duals. Let $A_{ji} : X_i \to U_j^*$, $j = 1, \ldots, m$, $i = 1, \ldots, n$ be linear continuous operators and A_{ji}^* be their conjugates. Thus, problems (4.1) and (4.2) take the form

$$L: \quad \sup\{\langle c, x\rangle = \sum_{i=1}^{n} \langle c_i, x_i\rangle \mid b_j - A_j x \in K_j,\ j = 1, \ldots, m,\ x \in C\},$$

$$L^*: \quad \inf\{\langle b, u\rangle = \sum_{j=1}^{m} \langle b_j, u_j\rangle \mid A_i' u - c_i \in C_i^*,\ i = 1, \ldots, n,\ u \in K^*\}$$

where, by definition, $A_j x = \sum_{i=1}^{n} A_{ji} x_i$, $A_i' u = \sum_{j=1}^{m} A_{ji}^* u_j$, $c = (c_1, \ldots, c_n) \in X^*$, $b = (b_1, \ldots, b_m) \in U^*$.

As in the case of finite dimensions, we fix n_0, m_0 $(0 \le n_0 \le n, 0 \le m_0 \le m)$, thus choosing the subsystems $b_j - A_j x \in K_j$, $j = 1, \ldots, m_0$, $A_i' u - c_i \in C_i^*$, $i = 1, \ldots, n_0$ of the restriction systems.

We introduce the functions Φ and Ψ. Function $\Phi(y)$ mapping $Y = X_{n_0+1} \times \ldots \times X_n \times U_{m_0+1}^* \times \ldots \times U_m^*$ into $R \cup \{+\infty\}$ is said to be admissible if (a) Φ is proper, convex and lower semi-continuous,

$$\operatorname{dom} \Phi \subset C_{n_0+1} \times \ldots \times C_n \times U_{m_0+1}^* \times \ldots \times U_m^*;$$

and (b) for each $x \in C_{n_0+1} \times \ldots \times C_n$, $u^1 - u^2 \in K_{m_0+1} \times \ldots \times K_m$ implies $\Phi(x, u^1) \ge \Phi(x, u^2)$.

Function $\Psi(z)$ mapping $Z = X_{n_0+1}^* \times \ldots \times X_n^* \times U_{m_0+1} \times \ldots \times U_m$ into $R \cup \{+\infty\}$ is said to be admissible if (a) Ψ is proper, convex and lower semi-continuous,

$$\operatorname{dom} \Psi \subset X_{n_0+1}^* \times \ldots \times X_n^* \times K_{m_0+1}^* \times \ldots \times K_m^*;$$

and (b) for each $u \in K_{m_0+1}^* \times \ldots \times K_m^*$, $x^1 - x^2 \in C_{n_0+1}^* \times \ldots \times C_n^*$ implies $\Psi(x^1, u) \ge \Psi(x^2, u)$.

We fix positive parameters R, ε_i, R_j, r_i, δ_j such that

$$\varepsilon_i r_i = R_j \delta_j = R^{-1},\ i = n_0 + 1, \ldots, n,\ j = m_0 + 1, \ldots, m,$$

and consider the problems

$$D: \quad \sup\{\langle c, x\rangle - R\Phi(\varepsilon_{n_0+1} x_{n_0+1}, \ldots, \varepsilon_n x_n; R_{m_0+1}(A_{m_0+1}x - b^{m_0+1}), \ldots \\ \ldots, R_m(A_m x - b^m)) | b_j - A_j x \in K_j,\ j = 1, \ldots, m_0,\ x \in C\},$$

$$D^{\#}: \inf\{\langle b, u\rangle + R\Psi(r_{n_0+1}(c_{n_0+1} - A'_{n_0+1}u), \ldots, r_n(c_n - A'_n u); \\ \delta_{m_0+1}u_{m_0+1}, \ldots, \delta_m u_m) | A'_i u - c_i \in C_i^*,\ i = 1, \ldots, n_0,\ u \in K^*\},$$

where Φ and Ψ are admissible functions over corresponding spaces.

Let $\tilde{f}$, $\tilde{f}^{\#}$, M, $M^{\#}$, $\tilde{M}$, $\tilde{M}^{\#}$ be optimal values and feasible and optimal sets of the problems D and $D^{\#}$; let $\Gamma(x)$ and $\Gamma^{\#}(u)$ be expressions appearing in the arguments of Φ and Ψ in D and $D^{\#}$.

Theorem 4.1. Let in D and $D^{\#}$ one of the functions Φ and Ψ be admissible and the other be its conjugate. Then

1) both function Φ and Ψ are admissible and mutually conjugated, and $\tilde{f} \le \tilde{f}^{\#}$;

2) if $\tilde{f} < +\infty$ and

$$\Gamma(x) \in \text{ int dom } \Phi,\ b_j - A_j x \in \text{ int } K_j,\ j = 1, \ldots, m_0,$$

for some $x \in M$, then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$ and $\tilde{M}^{\#} \neq \emptyset$

3) if $\tilde{f}^{\#} > -\infty$ and

$$\Gamma^{\#}(u) \in \text{ int dom } \Psi,\ A'_i u - c_i \in \text{ int } C_i^*,\ i = 1, \ldots, n_0,$$

for some $u \in M^{\#}$, then $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$ and $\tilde{M} \neq \emptyset$.

5. Infinite linear programs with a duality gap

Approximation by cones [6]. Consider a pair of dual infinite LP problems

$$\inf\{\langle c, x\rangle | Ax \in b + Q,\ x \in P\} = v, \tag{5.1}$$

$$\sup\{\langle b, u\rangle | -A^* u + c \in P^*,\ u \in Q^*\} = v^* \tag{5.2}$$

where X, Y (as well as Z, U) are real linear spaces paired under the bilinear functional $\langle \cdot, \cdot \rangle$; $A : X \to Z$ is a linear transformation, $A^* : U \to Y$ is the adjoint transformation; $Q \subset Z$ and $P \subset X$ are closed convex cones; $Q^* \subset U$ and $P^* \subset Y$

are their duals. We denote the weak topology on Z by $\sigma(Z,U)$ and the Mackey topology on Z by $\tau(Z,U)$.

A sequence of closed convex cones $\{Q_n\}$ is said to be strictly decreasing to the cone Q (notation: $\{Q_n\} \to Q$) if

1) $Q_{n+1} \subset Q_n$, $Q \setminus \{0\} \subset \tau - \operatorname{int} Q_n \forall n$, $Q = \bigcap_{n=1}^{\infty} Q_n$;

2) for some n the cone Q_n is locally $\sigma(Z,U)$-compact and $Q_n \setminus \{0\} \subset \subset \tau - \operatorname{int} Q_{n-1}$.

In what follows v_n will denote the optimal value of problem (5.1) where P (or P and Q) is replaced by P_n (or by P_n and Q_n).

Theorem 5.1. Let $\{P_n\} \to P$ and $v_n = v_n^* \in R$ for all n. Then either $\lim_{n\to\infty} v_n = v^* = v$, or

$$\lim_{n\to\infty} v_n = \sup\{\langle b,u\rangle | A^*u = c,\ u \in Q^*\}.$$

Theorem 5.2. Let $\{P_n\} \to P$, $\{Q_n\} \to Q$ and $v_n \in R$ for some n. Then $v \in R$ implies $\lim_{n\to\infty} v_n = v^*$.

Infinite LP problem over R_∞ [7]. Let R_∞ be a space of sequences $x = (x_1, x_2, \ldots)$, $x_i \in R$, $i = 1,2,\ldots$ Let R'_∞ be a finite sequence space, i.e. a subspace of R_∞ consisting of all $x = (x_1, x_2, \ldots) \in R_\infty$ with finitely many non-zero components. For $a \in R_\infty$ and $x \in R'_\infty$ we put $\langle a, x\rangle = \sum_i a_i x_i$. Set

$$a_{0\cdot} = (a_{01}, a_{02}, \ldots) \in R_\infty,\ b = (b_1, b_2, \ldots) \in R_\infty,$$
$$a_{i\cdot} = (a_{i1},\ a_{i2}, \ldots) \in R_\infty,\ a_{\cdot j} = (a_{1j}, a_{2j}, \ldots) \in R_\infty,$$

i.e. $a_{i\cdot}$, $i = 1,2,\ldots$, is a row and $a_{\cdot j}$, $j = 1,2,\ldots$ is a column of the matrix $A_{\cdot\cdot} = (a_{ij})$. Consider a pair of dual infinite programs over R'_∞

$$L_{\cdot\cdot} : \inf\{\langle a_{0\cdot}, x\rangle | \langle a_{i\cdot}, x\rangle \le b_i,\ i = 1,2,\ldots\} = v_{\cdot\cdot},$$
$$L^*_{\cdot\cdot} : \sup\{\langle -b, y\rangle | \langle a_{\cdot j}, y\rangle = a_{0j},\ y \ge 0,\ j = 1,2,\ldots\} = v^*_{\cdot\cdot}$$

and a family of approximating finite LP problems

$$L_{mn} :\ \min\left\{\sum_{j=1}^{m} a_{0j}x_j \Big| \sum_{j=1}^{m} a_{ij}x_j \le b_i,\ i = 1,\ldots,n\right\} = v(m,n).$$

Put

$$\overline{\gamma} = \sup_{\{m_k,n_k\}} \overline{\lim_k}\, v(m_k,n_k),\ \underline{\gamma} = \inf_{\{m_k,n_k\}} \underline{\lim_k}\, v(m_k,n_k),$$

where sup and inf are taken among all sequences $\{m_k\} \to \infty$, $\{n_k\} \to \infty$.

We say that problem $L_{\cdot\cdot}$ is finite-definite if there exists m_0 such that the system $\sum_{j=1}^{m} a_{ij}x_j \leq b_i$, $i = 1, 2, \ldots$, is consistent and finite-definite in the sense of [8] for all $m \geq m_0$.

Theorem 5.3. Let $L_{\cdot\cdot}$ be finite-definite and $v_{\cdot\cdot}$ be finite. Then $v_{\cdot\cdot} = \overline{\gamma}$, $v^*_{\cdot\cdot} = \underline{\gamma}$ and for each sequence $\{m_k\} \to \infty$ ($\{n_k\} \to \infty$) there exists $\{n_k\}$ ($\{m_k\}$) such that

$$\lim_{k\to\infty} v(m_k, \tilde{n}_k) = \overline{\gamma} \; (\lim_{k\to\infty} v(\tilde{m}_k, n_k) = \underline{\gamma})$$

as soon as $\tilde{n}_k > n_k$ ($\tilde{m}_k > m_k$).

This theorem shows that $v_{\cdot\cdot} > v^*_{\cdot\cdot}$ iff $\overline{\gamma} > \underline{\gamma}$. An example can be presented, where $v_{\cdot\cdot} > v^*_{\cdot\cdot}$ and each approximating sequence $\{v(m_k, n_k)\}$ tends either to $v_{\cdot\cdot}$, or to $v^*_{\cdot\cdot}$.

To remove the finite-definiteness hypothesis we introduce a regularized approximating family

$$L^{(t)}_{mn}: \; \min\{\sum_{i=1}^{m+1} \lambda_i \langle a_{0\cdot}, te^m_i\rangle \Big| \sum_{i=1}^{m+1} \lambda_i = 1, \lambda_i \geq 0,$$

$$\sum_{i=1}^{m+1} \lambda_i (\langle a_{k\cdot}, te^m_i\rangle - b_k) \leq 0, \; k = 1, \ldots, n\} = v(m, n, t),$$

where $e_i \in R^m$, cone $\{e_1, \ldots, e_{m+1}\} = R^m$.

Theorem 5.4. Let $v_{\cdot\cdot}$ be finite. Then, for each sequence $\{m_k\} \to \infty$ there exist $\tilde{t}(m_k)$ and $\tilde{n}(t_k)$ (for all $t_k > \tilde{t}(m_k)$) such that $\lim_k v(m_k, n_k, t_k) = v_{\cdot\cdot}$ as soon as $t_k \geq \tilde{t}(m_k)$, $n_k \geq \tilde{n}(t_k)$.

For the works concerning the approximation of the optimal values of problems having the duality gaps, known also as construction of "perfect duality" and "limiting Lagrangeans", see e.g. [9] and references there.

6. Correction of improper problems

The notion of correction (approximation) of improper problems was discussed in Section 1. Taking into account that the extremum problems can usually be re-formulated in the form of minimax or maximin problems we shall consider the correction of the latter problems. If e.g. $F(u, x, \Delta b, \Delta c)$ is the Lagrangean function of problem (1.3) (and is convex in $(u, \Delta c)$ and concave in $(x, \Delta b)$), then $K = \{\Delta |$ problem (1.3) is proper $\}$ is equal to the set of all Δ's ensuring the existence of a saddle point for $F(.,., \Delta b, \Delta c)$ on R^{m+n}_+. Thus (1.5) is a particular case of problem (6.2) below. To deal with the correction problem (6.2) (or (6.3)) we

need some results concerning the solvability set K. The convex analytic tools are taken mainly from [10].

Solvability sets [11]. Consider the maximin and minimax problems

$$\sup_u \inf_x F(u,x,p,q), \quad \inf_x \sup_u F(u,x,p,q) \tag{6.1}$$

where F is a proper closed saddle function (concave in $(u,q) \in R^{m_1+n_2}$ and convex in $(x,p) \in R^{m_2+n_1}$; $m_1+m_2=m$, $n_1+n_2=n$) mapping R^{m+n} into $R \cup \{-\infty\} \cup \cup\{+\infty\}$, and p, q are parameters. The solvability sets of problems (6.1) are

$$K = \{(p,q) \in R^n | F(\cdot,\cdot,p,q) \text{ has a saddle point}\},$$

$$K^1 = \{(p,q) | \sup_u \inf_x F \in R\}, K^2 = \{(p,q) | \inf_x \sup_u F \in R\}.$$

With F we associate a proper saddle function

$$F^1(u^*,x^*,p,q) = \sup_x \inf_u \{\langle u^*,u\rangle + \langle x^*,x\rangle - F(u,x,p,q)\}$$

and the sets (where $P_q(\cdot)$ denotes the projection on the subspace of vector q):

$$D = \{(p,q) | (0,0,p,q) \in \text{cl dom } F^1\},$$

$$Q_1 = \{p | (0,p) \in \text{cl dom}_1 F^1 \setminus \text{dom}_1 F^1\} \times (P_q(\text{dom } F) \setminus \text{ri } P_q(\text{dom } F)),$$

$$Q_2 = (P_p(\text{dom } F) \setminus \text{ri } P_p(\text{dom } F)) \times \{q | (0,q) \in \text{cl dom}_2 F^1 \setminus \text{dom}_2 F^1\}.$$

Function $\text{cl}_u F$ is obtained from F by closing it in u (for each (x,p,q)). Denote real $F = \{(u,x,p,q) | F(u,x,p,q) \in R\}$. Consider the following conditions on F:

$$A_q : P_q (\text{real cl}_u F) = P_q (\text{dom } F); \quad A_p : P_p (\text{real cl}_x F) = P_p (\text{dom } F);$$

$$A_1 : P_{(u,q)} (\text{real cl}_x F) = \text{dom}_1 F; \quad A_2 : P_{(x,p)} (\text{real cl}_u F) = \text{dom}_2 F.$$

Theorem 6.1. A_q implies $K^1 \subset D \cup Q_1$. A_p implies $K^2 \subset D \cup Q_2$. A_q and A_p imply $K \subset K^1 \cap K^2 \subset D$.

Theorem 6.2. A_1 and A_p imply $K^2 \subset D$. A_2 and A_q imply $K^1 \subset D$.

Theorem 6.3. Let $K \subset D$, $0 \in P_{(u^*,x^*)}$ (ri dom F). Then $F_0^1(p,q) \equiv \equiv F^1(0,0,p,q)$ is proper and closed, and ri $D = \{(p,q) | (0,0,p,q) \in \text{ri dom } F^1\} = = \text{ri dom } F_0^1 \subset K \subset \text{cl dom } F_0^1 = D$.

Correction using convex criterion [11]. With (6.1) we associate the correction problem of the form

$$\text{minimize} \quad \Phi(p,q)$$

$$\text{subject to} \quad (p,q) \in K \cap S, \tag{6.2}$$

where $\Phi(p, q)$ is a proper closed convex function, S is a convex set (in what follows we put $S = \operatorname{dom} \Phi$). Problem (6.2) may also be used to investigate the structure of the solvability sets (by solving (6.2) with various Φ and S).

Let $(p_2, q_1) \in \operatorname{ri} K$. Let p_1^* and q_2^* satisfy either the conditions

$$A': \quad p_1^* \in \operatorname{ri}\{p^* | (0, 0, -p^*) \in \partial_{u,x,p} F(u, x, p, q_1),\ (u, x, p) \in R^{m+n_1}\},$$
$$q_2^* \in \operatorname{ri}\{q^* | (0, 0, -q^*) \in \partial_{u,x,q} F(u, x, p_2, q),\ (u, x, q) \in R^{m+n_2}\},$$

or the conditions A'' obtained from A' by omitting symbols 'ri'. Thus,

$$(0, 0, -p_1^*) \in \partial_{u,x,p} F(u_1, x_1, p_1, q_1),\ (0, 0, -q_2^*) \in \partial_{u,x,q} F(u_2, x_2, p_2, q_2).$$

Put

$$c_1 = \langle p_1^*, p_1 \rangle + F(u_1, x_1, p_1, q_1),\ c_2 = \langle q_2^*, q_2 \rangle + F(u_2, x_2, p_2, q_2).$$

We introduce the set

$$\begin{aligned} \operatorname{dom} H_{\alpha,\beta} = &\{(u, y, v, x, p, q) | (p, q) \in \operatorname{dom} \Phi,\ (u, x, p, q_1) \in \\ &\in \operatorname{dom} F,\ (v, y, p_2, q) \in \operatorname{dom} F\} = \{(u, y, v, x, p, q) | \\ &(u, y) \in \operatorname{dom}_1 H_{\alpha,\beta},\ (v, x, p, q) \in \operatorname{dom}_2 H_{\alpha,\beta}\} \end{aligned}$$

as the domain of the saddle function

$$H_{\alpha,\beta}(u, y, v, x, p, q) = \Phi(p, q) + \alpha F(u, x, p, q_1) + \alpha \langle p_1^*, p \rangle -$$
$$-\beta F(v, y, p_2, q) - \beta \langle q_2^*, q \rangle - \alpha c_1 + \beta c_2,\ (u, y, v, x, p, q) \in \operatorname{dom} H_{\alpha,\beta},$$
$$H_{\alpha,\beta} = +\infty,\ (u, y) \in \operatorname{dom}_1 H_{\alpha,\beta},\ (v, x, p, q) \notin \operatorname{dom}_2 H_{\alpha,\beta},$$
$$H_{\alpha,\beta} = -\infty,\ (u, y) \notin \operatorname{dom}_1 H_{\alpha,\beta}.$$

Here α, β are positive parameters. Let $S_{\alpha,\beta}$ be the set of saddle points of $H_{\alpha,\beta}$ and σ be the optimal value of (6.2). Denote

$$\Phi_0(p, q) = \Phi(p, q) + i((p, q) | \operatorname{cl} K),\ S'_{\alpha,\beta} = P_{(p,q)}(S_{\alpha,\beta}),$$
$$\sigma(\alpha, \beta) = \sup_{u,y} \inf_{x,v,p,q} H_{\alpha,\beta},\ d(A, B) = \sup_{a \in A} \inf_{b \in B} \|a - b\|.$$

The following theorem reduces the solving of problem (6.2) to finding saddle points of the auxiliary function $H_{\alpha,\beta}$.

Theorem 6.4. Let $K \subset D$, $-\infty < \sigma < +\infty$,

$$0 \in P_{(u^*,x^*)}(\operatorname{ri} \operatorname{dom} F^1),\ \operatorname{ri} \operatorname{dom} \Phi \cap \operatorname{ri} K \neq \emptyset,$$

and let A'' be satisfied. Then, for each $\alpha > 0$, $\beta > 0$

1) $H_{\alpha,\beta}$ is a proper closed function;

2) $\sigma(\alpha, \beta)$ is a monotone non-decreasing function of $(\alpha, \beta) > 0$, $\sigma(\alpha, \beta) = \inf_{x,v,p,q} \sup_{u,y} H_{\alpha,\beta} \in R$, $\sigma = \lim_{\alpha \to +0, \beta \to +0} \sigma(\alpha, \beta) = \inf_{\alpha > 0, \beta > 0} \sigma(\alpha, \beta)$;

3) $S'_{\alpha,\beta} \subset \operatorname{dom} F_0^1 \cap \operatorname{dom} \Phi \subset \operatorname{cl} K \cap \operatorname{dom} \Phi$;

4) $\sigma \leq \Phi(p,q) \leq \sigma(\alpha,\beta)\ \forall (p,q) \in S'_{\alpha,\beta}$.

If, in addition, $0 \in \operatorname{ri} \operatorname{dom} \Phi_0^*$ and A' is satisfied, then

5) $S_{\alpha,\beta} \neq \emptyset$ and what is more $0 \in \operatorname{ri} \operatorname{dom} H^*_{\alpha,\beta}$;

6) $S_0 = \operatorname{Arg} \min \Phi_0 \neq \emptyset$, $d(S'_{\alpha,\beta}, S'_0) \to 0$ $(\alpha \to +0, \beta \to +0)$.

Correction using concave-convex criterion. Under the hypothesis of Theorem 6.3 $\operatorname{cl} K = \{(p,q) | p \in P, q \in Q\}$ where P and Q are closed convex sets. Consider the correction problem of the form

$$\sup_{p \in P} \inf_{q \in Q} \Psi(p,q) \quad (= \sigma), \tag{6.3}$$

where $\Psi(p,q)$ is a proper closed concave–convex function. For (6.3) a proposition similar in form to Theorem 6.4 holds (see [11]). If F is of the form

$$F(u,x,p,q) = F_0(u,x) - \langle p,u\rangle - \langle q,x\rangle,$$

where F_0 is a proper closed concave–convex function, then another method for solving problem (6.3) (and also problem (6.2) if Φ is of the form $\Phi(p,q) = \Phi_1(p) + \Phi_2(q)$) can be proposed. Note that in this case equality $\operatorname{cl} K = P \times Q \neq \emptyset$ always holds.

Consider the concave–convex function $G_\alpha = \Psi(p,q) - \alpha F(u,x,p,q)$ of variables $z = (p,x)$, $w = (q,u)$, and the conditions

A_3 : $\operatorname{ri} \operatorname{dom}_1 \Psi \cap \operatorname{ri} \operatorname{dom}_1 F_0^* = P_0 \neq \emptyset, Q_0 = \operatorname{ri} \operatorname{dom}_2 \Psi \cap \operatorname{ri} \operatorname{dom}_2 F_0^* \neq \emptyset$;

A_4 : for some $p_0 \in P_0$, $q_0 \in Q_0$ all level sets of the functions $\Psi(p_0,\cdot) + i(\cdot | \operatorname{dom}_2 F_0^*)$, $\Psi(\cdot, q_0) - i(\cdot | \operatorname{dom}_1 F_0^*)$ are bounded (here F_0^* is the conjugate of F_0).

If A_3 holds then G_α is a proper closed function. If A_4 also holds, then, for sufficiently small $\alpha > 0$, G_α has the finite saddle value σ_α and the non-empty set of saddle points S_α [10]. The following theorem reduces the solving of problem (6.3) to finding saddle points of G_α.

Theorem 6.5 [12]. Let A_3 and A_4 be satisfied. Then

1) if $(\overline{p}, \overline{x}, \overline{q}, \overline{u}) \in S_\alpha$ then $(\overline{u}, \overline{x})$ is a saddle point of $F(\cdot,\cdot,\overline{p},\overline{q})$ and $S'_\alpha = P_{(p,q)}(S_\alpha) \subset K$;

2) $\sigma_\alpha \to \sigma$, $d(S'_\alpha, S_0) \to 0$ $(\alpha \to +0)$ (here S_0 is a non-empty set of saddle points of Ψ with respect to $P \times Q$). If, in addition, $S_0 \subset K$, then

3) $d(S'_\alpha, S_1) \to 0$ $(\alpha \to +0)$ where S_1 is a non-empty set of saddle points of F_0^* with respect to S_0.

If the values of Ψ^* are easily found, then the described method may be simplified. Set $g_\alpha(u,x) = \alpha F_0(u,x) - \Psi^*(-\alpha u, -\alpha x)$. We denote by κ_α the saddle value and by Z_α the set of saddle points of g_α.

Corollary [12]. Let A_3 and A_4 be satisfied and g_α be closed. Then

1) $Z_\alpha \neq \emptyset$ for sufficiently small $\alpha > 0$;

2) if $(\overline{u}, \overline{x}) \in Z_\alpha$, then $(\overline{u}, \overline{x})$ is a saddle point of $F(\cdot, \cdot, \overline{p}, \overline{q})$ for all $(\overline{p}, \overline{x}) \in$ $\in \partial F_0(\overline{u}, \overline{x}) \cap (-\alpha^{-1} \partial \Psi^*(-\alpha\overline{u}, -\alpha\overline{x}))$;

3) $\kappa_\alpha \to -\sigma$, $d(Z'_\alpha, S_0) \to 0$ $(\alpha \to +0)$, where $Z'_\alpha = \bigcup\limits_{(u,x)\in Z_\alpha} \partial F_0(u, x) \cap$ $\cap(-\alpha^{-1} \partial \Psi^*(-\alpha u, -\alpha x))$.

If, in addition, $S_0 \subset K$, then

4) $d(Z'_\alpha, S_1) \to 0$ $(\alpha \to +0)$.

Note that if Ψ is strongly concave–convex, then Ψ^* is differentiable and has finite values so that

$$Z'_\alpha = \bigcup_{(u,x)\in Z_\alpha} (-\alpha^{-1} \bigtriangledown \Psi^*(-\alpha u, -\alpha x)).$$

E.g. for the quadratic function $\Psi(p, q) = (\|q\|^2 - \|p\|^2)\frac{1}{2}$ we have

$$g_\alpha(u, x) = \alpha F_0(u, x) - \frac{\alpha^2}{2}(\|u\|^2 - \|x\|^2),$$

$$Z'_\alpha = \alpha(\overline{u}_\alpha, \overline{x}_\alpha), \quad \text{where} \quad \{(\overline{u}_\alpha, \overline{x}_\alpha)\} = Z_\alpha.$$

Some other approaches to the correction of improper problems and inconsistent equation and inequality systems (construction of the generalized solutions) as well as applications can be found, e.g. in [1, 11, 13–20].

7. Concluding remarks

In this survey the theory of improper linear and convex programming problems was mainly presented and especially the duality theory. Less attention was paid to the correction of such problems, i.e. to the methods for construction of compromise models according to some criterion. However, it is undoubtedly an important question, in particular, from the point of view of applications. The applied program packages are developed for construction of compromise models. We can refer, e.g., to the DELTA-PLAN-ES package developed in the Institute of Mathematics and Mechanics of Ural Department of the Academy of Sciences of the USSR.

A fairly ample bibliography (about 200 entries) on the improper mathematical programming problems is given in [13].

References

1. *Eremin, I.I., Mazurov, V.D., Astaf'ev, N.N.*, Improper Problems of Linear and Convex Programming. Moscow, Nauka, 1983.
2. *Eremin, I.I., Vatolin, A.A.*, Duality in improper mathematical programming problems. Preprint. Sverdlovsk, Ural Sci. Centre, 1985, 52 pp.
3. *Tikhonov, A.N., Arsenin, V.Y.*, Methods for Solving Incorrect Problems. Moscow, Nauka, 1979.
4. *Eremin, I.I.*. Duality in improper problems of linear and convex programming. Dokl. AN SSSR, 1981, **256**, *2*, 272–276.
5. *Eremin, I.I., Vatolin, A.A.*, Duality in improper infinite-dimensional linear and convex programs. In: Metody approksimatsii nesobstvennyh zadach matematicheskogo programmirovaniya. Sverdlovsk, Ural Sci. Centre, 1984, 3–20.
6. *Trofimov, S.P.*, Analysis of duality relations in semi-infinite and infinite linear programs. Cand. thesis. Sverdlovsk, Instit. of Math. and Mech., Ural. Depart., Acad. Sci. USSR, 1988, 18 pp.
7. *Astaf'ev, N.N.*, Linear Inequalities and Convexity. Moscow, Nauka, 1982.
8. *Chernikov, S.N.*, Linear Inequalities. Moscow, Nauka, 1968.
9. *Karney, D.F., Morley, T.D.*, Limiting Lagrangeans: A primal approach. J. Opt. Theory and Appl., 1986, **48**, *1*, 163–174.
10. *Rockafellar, R.T.*, Convex Analysis. Princeton, N.J., 1970.
11. *Vatolin, A.A.*, Solvability sets and correction of saddle functions and inequality systems. Preprint. Sverdlovsk, Ural Depart. Acad. Sci. USSR, 1989, 90 pp.
12. *Popov, L.D.*, Linear correction of improper convex–concave minimax problems using maximin criterion. Zh. vychisl. matem. i matem. Phis., 1986, **26**, *9*, 1325–1338.
13. *Eremin, I.I.*, Contradictory Models of Optimal Planning. Moscow, Nauka, 1988.
14. *Vatolin, A.A.*, On the linear programming problems with interval coefficients. Zh. vychisl. matem. i matem. Phis., 1984, **24**, *11*, 1629–1637.
15. *Skarin, V.D.*, On an approach to the analysis of improper problems of linear programming. Zh. vychisl. matem. i matem. Phis., 1986, **26**, *3*, 439–448.
16. *Plotnikov, S.V.*, On the cyclic projection on the system of convex sets with empty intersection. In: Nesobstvennye zadachi optimizatsii. Sverdlovsk: Ural Sci. Centre, 1982, 60–66.
17. *Frolov, V.N.*, Optimization of Planning Programs under Poorly Coordinated Restrictions. Moscow, Nauka, 1986.
18. *Vereskov, A.I., Tret'jakov, N.V.*, On the application of augmented Lagrangeans to the correction of inconsistent convex programs. Izvestiya AN SSSR, Tekh. Kib., 1988, *1*, 3–12.
19. *Bulavskij, V.A.*, Generalized solutions and regularization of inequality systems. In: Vychislitelnye metody lineinoj algebry. Novosibirsk, Nauka, 1985, 161–174.
20. *Mangasarian, O.L.*, Normal solutions of linear programs. Math. Progr. Study, 1984, **22**, 206–216.

Несобственные задачи математического программирования

И. И. ЕРЕМИН, А. А. ВАТОЛИН

(Свердловск)

Задача математического программирования называется собственной, если она и двойственная к ней задача разрешимы и их оптимальные значения совпадают. В противном случае задача называется несобственной. В статье, носящей обзорный характер, рассматриваются несобственные задачи линейного и выпуклого программирования (в том числе бесконечномерные). Для них развивается теория двойственности. Для задач с разрывом в двойственности изучаются предельные процессы, приближающие оптимальные значения прямой и двойственной задач. Исследуются различные подходы к коррекции минимаксных задач, зависящих от параметров.

И. И. Еремин
А. А. Ватолин
Институт математики и механики УрО АН СССР
СССР, 620066, Свердловск, ул. С. Ковалевской, 16.

Problems of Control and Information Theory, Vol. 18(6), pp. 381–395 (1989)

HIERARCHIES OF PARALLEL PROBABILISTIC SEARCHING ALGORITHMS WITH POSSIBLE DATA ACCESS CONFLICTS

I. KRAMOSIL

(*Prague*)

(Received May 9, 1988)

In order to find whether there is an element possessing a tested property in a large basic space, a number of processors sample, in parallel and at random, elements from this space and test those of them which were not sampled twice or more times at the same step, in order to avoid data access conflicts. Higher-level processors sample at random the lower-level ones and ask them for elements with the tested property, supposing some of them were found. The output of the unique highest-level processor is taken as a statistical decision function answering the introduced question. The probability of error is investigated as well as the time computational complexity of this algorithm in cases when this probability is below a given threshold value uniformly for all subsets of the basic space.

1. Introduction

Many important and interesting problems in applied mathematics and artificial intelligence can be solved in a way the substantial part of which consists of searching for an at least one element with a given property within a finite but, as a rule, very large set of possible candidates. Remember, e.g. the searching for a divisor of a natural number or for another witness of its composedness when testing the primality property, the searching for a counter-example when proving or rather disproving an assertion, or the searching for a unifying substitution or for matching pairs of clauses in the resolution-based methods of automated deduction. As a rule, it is rather simple to verify whether an already obtained element possesses the tested property or not, but the searching itself may be highly time-consuming, as the set of potential candidates is very large and poorly or unappropriately if at all, structured, and the set of acceptable solutions is very small or even unique, so that the blind exhaustive search is the only solution at hand. Very often, the time computational complexity of the searching procedure forms the substantial part of the time computational complexity of the algorithm in question as a whole and ultimately decides about its practical applicability.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

When looking for a solution of a searching problem outside the scope of classical, i.e. sequential and deterministic algorithms, randomization and parallelization are the first general ideas to come into one's mind. The most simple randomization, taken alone, does not solve the problem, as can be almost immediately seen. Or, consider a set A with N elements and only one of them possessing the tested property and consider a sequence of statistically independent random samples from A each of them ascribing the same probability $1/N$ of sampling to each $a \in A$. An easy calculation yields that we have to take at least $(\ln(1/\varepsilon))N$ samples to be sure that the only element with the tested property will be sampled at least once with a probability exceeding $1 - \varepsilon$, $0 < \varepsilon < 1$ being given a priori.

Supposing parallelization were applied in its idealized form commonly taken in the theory of non-deterministic algorithms, each element of the basic set A could be tested by a particular processor (testing device) and if the time computational complexity of this testing procedure does not depend on the cardinality of A, the problem is solved within a constant time. However, we have perfectly neglected the time (and other) demands involved by the necessity to synchronize the work of particular processors, i.e. to assure that each element of A is tested by just one processor, as well as the demands following from the inspection of outputs of these processors in order to check, whether at least one of them has found an element with the tested property. In case this checking is performed by a sequential and deterministic algorithm, and if each element is tested by a particular processor, the checking problem is nothing else than a copy (duplicate) of the original searching problem in question.

An almost immediate idea is to take $\sqrt{N}$ ($N = \text{card } A$) processors each of them inspecting different $\sqrt{N}$-tuple of elements of A, with trivial modifications if N is not a square. Not taking into consideration the time necessary to separate the elements of A into $\sqrt{N}$-tuples, this searching strategy reduces the time complexity of the searching problem into the $O(\sqrt{N})$-class and we shall see, in what follows, that this is, in a sense, the optimal reduction, reachable by two-level hierarchies of processors, also under much weaker and in probabilistic terms described conditions concerning the testing and inspection procedures and with no co-operation or synchronization among the processors assumed.

2. Informal preliminaries on hierarchical structures and their elements

In this chapter we introduce, in an informal way, the demands imposed into the elementary unit-processors, constituting the hierarchical structures under consideration, as well as the assumptions concerning the possibilities of mutual communication and interactions among these processors. Our aim was to settle as weak conditions as necessary for the processors to be allowed to serve as brickstones in reasonable algorithmical structures. Processors are separated into different levels

enumerated by natural numbers so that we speak about processors of first level, second level, third level, etc. Given a basic set A and its subset V, each processor of the first level is able to sample at random, and with respect to a given probability distribution, an element x from A and to test, within a constant time considered often as a time unit, whether x belongs to V or not. If $x \in V$, the processor outputs a fixed vlaue, say 1, if $x \in A - V$, it outputs a fixed value, say 0 (to be consistent with the usual definition of the characteristic function, or identifier, $\chi_V(\cdot)$ of the subset V of A). In case A is finite, most often the uniform probability distribution over A will be taken into consideration. Each first-level processor works on the black-box principle, so that we have no insight into the way in which the predicate $x \in V$ is tested so that there is no way to modify appropriately the sampling mechanism, neither we have any supplementary knowledge about the set V which could be used to improve the work of particular processors or their structure as a whole. The number of simultaneously working first-level processors is not supposed to be limited a priori (unlimited parallelization), but we shall see that under certain realistic assumptions a too rapid increase of this number of the simultaneously working processors is beyond any sense as it makes worse the total quality of the hierarchical algorithm in question as a whole. In spite of [1], [2] or [3], where random samples were taken as non-conflict in the sense that if two or more processors sample the same element from A in the same time instant (step), all of them may test it, let us accept, in what follows, a more realistic assumption which is, in a sense, something like a dual extremum — if two or more processors sample the same element from A in the same step, the element is blocked and it is not accessible to any processor at this step. However, the processors are not supposed to be able to distinguish this case from that one when $x \in A - V$ has been sampled, so that they output zero value in this situation. Each first-level processor is able to repeat its activity in the next time instant or step, to take, independently and from the same probability distribution over A, another random sample from A, to test it and to cumulate, on its output, the information whether at least one element from A has been found in a finite sequence of samples. If it is the case, the processor outputs a unit value; otherwise it outputs zero value.

The supposed abilities of higher-level processors can be described in an analogous way. Namely, each n-th level processor behaves like a first-level one just with the basic set A replaced by the set of (outputs of) the $(n-1)$-th level processors and with V replaced by the subset of those $(n-1)$-th level processors which yield the unit output value. Hence, each n-th level processor makes, possibly independently repeated, random samples from the set of $(n-1)$-th level processors and test their outputs; the output value of the n-th level processor in question is 1, if there is at least one $(n-1)$-th level processor with unit output value among the sampled one. If it is not the case, the n-th level processor outputs zero. Again, (the outputs of) processors sampled twice or more times at the same step (by different higher-level processors) are inaccessible in this step. The testing oracles at all levels are supposed to be reliable; i.e. they work without any danger of failure. As a rule, there

is a single processor of the highest level and its output is taken as the final result of all the testing procedure in question. If this output value is 1, it is interpreted as the decision that V is not empty and this decision is evidently true by virtue of that we have told (at least one element from V must have been sampled by some of the first-level processors). The zero value of this highest-level output is understood as the decision that V is empty, but this decision is evidently charged by a positive probability of error (except for trivial cases), which we have to investigate in more detail and explicitly quantify in what follows.

3. Two-level hierarchies

Consider a finite set $A = \{a_1, a_2, \ldots, a_N\}$ and $V \subset A$, the pair $\langle A, V \rangle$ is called a *searching problem*. Let $\{X_{i,j}\}_{i=1}^{m}{}_{j=1}^{n}$ be a system of statistically independent and identically distributed random variables defined on an abstract probability space $\langle \Omega, \mathcal{S}, P \rangle$, taking their values in A and such that, for each $i \leq m$, $j \leq n$, $K \leq N$,

$$P(\{\omega : \omega \in \Omega, \quad X_{i,j}(\omega) = a_K\}) = 1/N. \tag{1}$$

Let $\{Z_l\}_{l=1}^{k}$ be a sequence of statistically independent and identically distributed random variables defined on $\langle \Omega, \mathcal{S}, P \rangle$, taking their values from the set $\{1, 2, \ldots, m\}$ of integers and such that, for each $l \leq k$ and $j \leq m$,

$$P(\{\omega : \omega \in \Omega, \quad Z_l(\omega) = j\}) = 1/m. \tag{2}$$

Random variables $\{X_{ij}\}$ and $\{Z_l\}$ are also supposed to be statistically independent. According to the terminology used in the references mentioned above, the structure

$$\langle \{X_{ij}\}_{i=1}^{m}{}_{j=1}^{n}, \{Z_l\}_{l=1}^{k} \rangle \tag{3}$$

is called *two-level hierarchical parallel probabilistic searching algorithm* (HPPSA) *for searching problem* $\langle A, V \rangle$ *with possible data access conflicts* (to distinguish it from the conflict-free models investigated earlier) and is denoted by $\mathcal{X}$ or $\mathcal{X}(A, V, m, n, k)$ to introduce explicitly its parameters m, n, k. Set, for $i \leq m$, $j \leq n$ and $\omega \in \Omega$,

$$\gamma(i, j, \omega) = \chi_V(X_{ij}(\omega)) \prod_{k=1, k \neq i}^{m} \left[1 - \chi_{\{X_{ij}(\omega)\}}(X_{kj}(\omega))\right], \tag{4}$$

and define, on $\langle \Omega, \mathcal{S}, P \rangle$, a random variable $\mathcal{X}_0$ taking its values from $\{0, 1\}$ in this way:

$$\{\omega : \quad \omega \in \Omega, \ \mathcal{X}_0(\omega) = 1\} = \tag{5}$$

$$= \{\omega : \quad \omega \in \Omega, \sum_{l=1}^{k} \sum_{j=1}^{n} \gamma(Z_l(\omega), j, \omega) > 0\},$$

$\mathcal{X}_0 = 0$, otherwise.

As can be easily seen, $\gamma(i,j,\omega) = 1$ iff $X_{ij}(\omega)$ is in V and differs from all $X_{kj}(\omega)$ with $k \neq i$. Supposing that $X_{i1}(\omega), X_{i2}(\omega), \ldots, X_{in}(\omega)$ are samples from A taken by the i-th processor, then $\gamma(i,j,\omega) = 1$ iff the i-th processor samples an element from V in the j-th step and, at the same step, no other processor samples the same element. Hence, the i-th processor tests this element and outputs the unit value. Moreover, $\sum_{j=1}^{n} \gamma(i,j,\omega) > 0$ holds iff the event just described occurs at least once in the sequence of n steps. But only the processors the indices of which are sampled at random by variables $Z_1, Z_2, \ldots, Z_k$ are asked for their output values, so that $\chi_0(\omega) = 1$ iff there is at least one among the sampled first-level processors which outputs unit value, so that $\chi_0(\omega)$ is nothing else than the output value of the unique second- (i.e. highest-)level processor in the two-level hierarchy in question, as defined informally in the previous chapter. The more simple and early investigated case of conflict-free samples can be obtained from χ, by simply setting $\gamma(i,j,\omega) = \chi_V(X_{ij}(\omega))$. Evidently,

$$\sum_{l=1}^{k}\sum_{j=1}^{n}\gamma(Z_l(\omega),j,\omega) \leq \sum_{l=1}^{k}\sum_{j=1}^{n}\chi_V(X^{(\omega)}_{Z_l(\omega),j}). \tag{6}$$

Considering a searching problem $\langle A, V\rangle$ with $V \neq \emptyset$ and a two-level HPPSA χ, the random event $\chi_0(\omega) = 0$ can be intepreted in such a way that χ has made an error. The following assertion offers certain estimations for the corresponding probability of error.

The assumption that the basic set A is finite is not taken in order to assure the theoretical computability, as it is the case in many definitions of deterministic algorithms, and it could be easily omitted supposing (1) were replaced by another probability distribution over A. However, at least in the most natural case of infinite countable A, each distribution over A depends on the ordering of A, i.e. prefers some elements and neglects other ones, and introduces some more parameters, not justifiable within the framework of our theoretical model, but ultimately deciding about the qualities of the HPPSA in question. In other words, the quality of the HPPSA would be determined by its behaviour over a finite subset of A the selection of which will be a matter of pure arbitrariness, from the point of view of our model. On the other hand, the uniform probability distribution over a finite A, defined by (1), obeys the well-known Laplace principle, depends only on the cardinality N of A, and enables to propose and prove assertions describing the qualities of a HPPSA as functions of a single argument N, i.e., in the form common for deterministic algorithms. It is why we suppose the set A to be finite with its cardinality playing the role of the only free parameter in our reasonings and computations, postponing a more detailed investigation of parallel probabilistic searching algorithms with infinite basic sets for another occasion.

Theorem 1. Let $\langle A, V\rangle$ be a searching problem with card $A = N$, card $V = v > 0$, let $\mathcal{X} = \mathcal{X}(A, V, m, n, k)$ be a two-level HPPSA for $\langle A, V\rangle$, then

$$\left(1 - \frac{v}{N}\right)^{nk} < P(\{\omega : \omega \in \Omega,\ \mathcal{X}_0(\omega) = 0\}) < \\ < \left(1 - \frac{m}{N}\left(1 - \frac{1}{N}\right)^{m-1}\right)^n + \left(1 - \frac{1}{m}\right)^k. \tag{7}$$

Proof. Consider, first, fixed $a_0 \in V$, $i_0 \leq m$, $j_0 \leq n$. The output value of the i_0-th processor in the j_0-th step is 1, iff this processor samples a_0 in this step (this event occurs with the probability $1/N$ due to the supposed uniform probability distribution of all X_{ij}) and if none of the others $m-1$ processors samples a_0 at the same time. Due to the supposed statistical independence the combined probability reads as

$$\left(\frac{1}{N}\right)\left(1 - \frac{1}{N}\right)^{m-1}. \tag{8}$$

The same element of A cannot be tested by two or more processors at the same time, so that the probability of being tested adds for different processors and with the probability

$$\left(\frac{m}{N}\right)\left(1 - \frac{1}{N}\right)^{m-1} \tag{9}$$

in the total a_0 is tested in the j_0-th step. For different steps, these events are statistically independent so that the probability of being tested at least once in n steps for a_0 reads as

$$1 - \left(1 - \left(\frac{m}{N}\right)\left(1 - \frac{1}{N}\right)^{m-1}\right)^n. \tag{10}$$

If this is the case, the set of first-level processors with unit output value after n steps is non-empty, hence, for each $l \leq k$, the probability that Z_l samples (the index of) such a processor is at least $1/m$. So, with probability at least

$$1 - \left(1 - \frac{1}{m}\right)^k \tag{11}$$

at least once such an index is sampled and, because of (5), $\mathcal{X}_0(\omega) = 1$. So, supposing that $V = \{a_0\}$ and multiplying the conditional probability (11) by the probability (10) of the conditioning random event, we obtain that

$$P(\{\omega : \omega \in \Omega,\ \mathcal{X}_0(\omega) = 1\}) \geq \\ \geq \left[1 - \left(1 - \left(\frac{m}{N}\right)\left(1 - \frac{1}{N}\right)^{m-1}\right)^n\right]\left[1 - \left(1 - \frac{1}{m}\right)^k\right], \tag{12}$$

so that

$$P(\{\omega : \omega \in \Omega,\ \chi_0(\omega) = 0\}) < \\ < \left(1 - \left(\frac{m}{N}\right)\left(1 - \frac{1}{N}\right)^{m-1}\right)^n + \left(1 - \frac{1}{m}\right)^k. \tag{13}$$

However, the probability of the correct decision $\chi_0(\omega) = 1$ is an increasing function of the cardinality v of V, so that (13) holds for all $V \neq \emptyset$ and the right-hand side inequality in (7) is proved.

The left-hand side inequality in (7) is more easy to prove. Consider the conflict-free modification of our algorithm given by the random variable χ_0. The probability of discovering an element of V in the case of χ_0 is evidently at most the same as for the conflict-free case when this probability reads as

$$p = 1 - \left(1 - \frac{v}{N}\right)^n. \tag{14}$$

Moreover, in this conflict-free case the occurrences of unit outputs for different processors are statistically independent, so that the simple binomial scheme with independent successes with probability p at each step can be used to obtain an upper bound for the probability that $\chi_1(\omega) = 1$. For $0 \leq s \leq m$, set

$$r_s = P(\{\text{card } W(\omega) = s\}). \tag{15}$$

This probability can be easily estimated using the binomial rule, so that

$$r_s \leq \binom{m}{s} p^s (1-p)^{m-s}. \tag{16}$$

If there are just s unit-valued outputs among the m ones, then with the probability

$$1 - \left(1 - \frac{s}{m}\right)^k \tag{17}$$

at least one of them is sampled by some $Z_1, Z_2, \ldots, Z_k$ so that, combining (5), (16) and (17),

$$P(\{\omega : \omega \in \Omega, \chi_0(\omega) = 1\}) = \sum_{s=0}^{m} r_s \left(1 - \left(1 - \frac{s}{m}\right)^k\right) \leq \\ \leq \sum_{s=0}^{m} \binom{m}{s} p^s (1-p)^{m-s} \left(1 - \left(1 - \frac{s}{m}\right)^k\right) = \tag{18} \\ = E_{p,m}\left(1 - \left(1 - \frac{s(\cdot)}{m}\right)^k\right),$$

where $E_{p,m}$ denotes the operator of the expected value with respect to the binomial distribution with m samples and probability p of success in one sample. As, for $f = 1 - (1 - (x/m))^k$ and for $0 \leq x \leq m$,

$$\frac{d^2 f}{dx^2} = \frac{d^2}{dx^2}\left(1 - \left(1 - \frac{x}{m}\right)^k\right) = -\frac{k(k-1)}{m^2}\left(1 - \frac{x}{m}\right)^{k-2} \leq 0, \tag{19}$$

evidently

$$E_{p,m} f(s(\cdot)) \leq f(E_{p,m} s(\cdot)) = 1 - (1-p)^k, \tag{20}$$

due to the well-known fact that $E_{p,m} s = mp$ for a random variable s with binomial probability distribution and parameters m, p. Hence, (14) yields

$$\begin{aligned} P(\{\omega : \omega \in \Omega,\ \chi_0(\omega) = 1\}) &< 1 - (1-p)^k = \\ &= 1 - \left(1 - \frac{v}{N}\right)^{nk}, \end{aligned} \tag{21}$$

so that

$$P(\{\omega : \omega \in \Omega,\ \chi_0(\omega) = 0\}) > \left(1 - \frac{v}{N}\right)^{nk}, \tag{22}$$

and Theorem 1 is proved. □

4. Computational complexity for two-level hierarchies

Theorem 2. Let the notations and conditions of Theorem 1 hold. For each $\varepsilon > 0$ there exist reals $c_1(\varepsilon)$, $c_2(\varepsilon)$ and $c_3(\varepsilon)$ independent of N such that if $N \geq 4c_1^2$, $m \geq \lceil c_1\sqrt{N} \rceil$, $n \geq \lceil c_2\sqrt{N} \rceil$ and $k \geq \lceil c_3\sqrt{N} \rceil$, then

$$P(\{\omega : \omega \in \Omega,\ \chi_0(\omega) = 0\}) < \varepsilon. \tag{23}$$ □

Proof. Due to the well-known analytical inequality

$$\left(1 - \frac{1}{m}\right)^{\ln(2/\varepsilon)m} < e^{-\ln(2/\varepsilon)} = \varepsilon/2, \tag{24}$$

we have to prove that

$$\left(1 - \frac{\lceil c_1\sqrt{N} \rceil}{N}\left(1 - \frac{1}{N}\right)^{\lceil c_1\sqrt{N}\rceil}\right)^{\lceil c_2\sqrt{N}\rceil} > \frac{\varepsilon}{2} \tag{25}$$

for appropriate c_1 and c_2, and to set $c_3 = c_2 \ln(2/\varepsilon)$. Set $\varepsilon_0 = \varepsilon/2$, set

$$c_1 = c_1(\varepsilon) = c_2 = c_2(\varepsilon) = \sqrt{2\ln(1/\varepsilon_0)}, \tag{26}$$

so that $c_1 c_2/2 > \ln(1/\varepsilon_0)$ and

$$\varepsilon_0 > e^{-c_1 c_2/2} = \left(e^{-c_1/2}\right)^{c_2} > \left(\left(1 - \frac{c_1/2}{\sqrt{N}}\right)^{\sqrt{N}}\right)^{c_2}, \tag{27}$$

as

$$a_N = \left(1 - \frac{x}{\sqrt{N}}\right)^{\sqrt{N}} \tag{28}$$

forms a monotonously increasing bounded sequence the subsequence $\{a_{N^2}\}_{N=1}^{\infty}$ of which tends to e^{-x}, hence $a_N \nearrow e^{-x}$ as well. Moreover,

$$\left(\left(1 - \frac{c_1/2}{\sqrt{N}}\right)^{\sqrt{N}}\right)^{c_2} = \left(1 - \left(\frac{c_1}{\sqrt{N}} - \frac{c_1/2}{\sqrt{N}}\right)\right)^{c_2\sqrt{N}}. \tag{29}$$

If $N \geq 4c_1^2$, then

$$\frac{c_1/2}{\sqrt{N}} \geq \frac{c_1^2}{N}, \tag{30}$$

so that

$$\left(1 - \left(\frac{c_1}{\sqrt{N}} - \frac{c_1/2}{\sqrt{N}}\right)\right)^{c_2\sqrt{N}} \leq \left(1 - \left(\frac{c_1}{\sqrt{N}} - \frac{c_1^2}{\sqrt{N}}\right)\right)^{c_2\sqrt{N}} = \tag{31}$$

$$= \left(1 - \frac{c_1}{\sqrt{N}}\left(1 - \frac{c_1}{\sqrt{N}}\right)\right)^{c_2\sqrt{N}} < \left(1 - \frac{c_1}{\sqrt{N}}\left(1 - \frac{1}{N}\right)^{c_1\sqrt{N}}\right)^{c_2\sqrt{N}},$$

as $(1-x)^n > 1 - nx$ for each $x \in (0,1)$, hence,

$$\left(1 - \frac{1}{N}\right)^{c_1\sqrt{N}} > 1 - \frac{c_1\sqrt{N}}{N} = 1 - \frac{c_1}{\sqrt{N}}. \tag{32}$$

Evidently

$$\left(1 - \frac{c_1}{\sqrt{N}}\left(1 - \frac{1}{N}\right)^{c_1\sqrt{N}}\right)^{c_2\sqrt{N}} > \left(1 - \frac{c_1}{\sqrt{N}}\left(1 - \frac{1}{N}\right)^{(c_1\sqrt{N})-1}\right)^{c_2\sqrt{N}}, \tag{33}$$

so that (27) and (33) yield

$$\varepsilon_0 > \left(1 - \frac{m}{N}\left(1 - \frac{1}{N}\right)^{m-1}\right)^n, \tag{34}$$

supposing that $m \geq \lceil c_1\sqrt{N} \rceil$, $n \geq \lceil c_2\sqrt{N} \rceil$ and $N \geq 4c_1$. The assertion is proved. □

Theorem 3. Let $\mathcal{X}$ be a HPPSA for searching problem $\langle A, V\rangle$ with card $A = N$, card $V = 1$. If $n = n(N)$ and $k = k(N)$ are such that $nk \in o(N)$, then there exists, for each $\varepsilon < 1$, $N_0 = N_0(\varepsilon)$ such that, for $N \geq N_0$,

$$P(\{\mathcal{X}_1(\omega) = 0\}) > \varepsilon. \qquad \square \quad (35)$$

Proof. An easy calculation yields

$$\begin{aligned} \lim_{N\to\infty} P(\{\mathcal{X}_1(\omega) = 0\}) &\geq \lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^{nk} = \\ &= \lim_{N\to\infty} \left(\left(1 - \frac{1}{N}\right)^{N}\right)^{\frac{nk}{N}} = \lim_{N\to\infty} e^{-\frac{nk}{N}} = 1, \end{aligned} \quad (36)$$

which proves the assertion. $\square$

The results proved in Theorems 2 and 3 deserve a more detailed comment because of their interpretation. Let us recall that the two-level HPPSA formally described by the structure $\langle \{X_{ij}\}_{i=1}^{m}{}_{j=1}^{n}, \{Z_l\}_{l=1}^{k}\rangle$ works in the following way. First, each first-level processor sequentially takes n random samples from A and outputs 1 or 0 according to whether an element from V has been discovered (i.e. sampled and tested) or not. Then, the second-level processor (supervisor) takes, again sequentially, k random samples from the set of (indices of) the first-level processors and outputs the final result according to the rules described above. Hence, accepting the simplifying assumptions that each random sample from A needs α time units and each random sample from $\{1, 2, \ldots, m\}$ needs β units (independently of N and m), the expression $\alpha n + \beta k +$ const may serve as a very rough estimation of the total time complexity of the HPPSA $\mathcal{X}(m, n, k)$. Now, Theorem 2 claims that it is possible to keep the probability of error, connected with the algorithm $\mathcal{X}(m, n, k)$, below a given $\varepsilon > 0$ with $\alpha n(N) + \beta k(N)$ in $O(\sqrt{N})$-class. Moreover, Theorem 3 claims that this time complexity cannot be qualitatively reduced (i.e. reduced to $o(\sqrt{N})$-class), as an $O(\sqrt{N})$-time complexity is necessary to keep below ε a lower bound of the actual probability of error. In both cases, what is actually needed is that the product of n and k must be in $O(N)$, but when taking $m(N) = d_1 N^q$, $k(N) = d_2 N^{1-q}$ for some q, $0 < q \neq 1/2$, $q < 1$, then $\alpha n + \beta k$ is in $O(N^{\max(1-q,q)})$ and this result would be qualitatively worse than that with $q = 1/2$, as $O(N^{1/2}) \subset \sigma(N^q)$ for each $q > 1/2$.

The following four closing remarks seem to be worth introducing.

(1) The results of Theorems 2 and 3 can be seen as a specification of or a support for more general results of Reif [4] about the quadratic speed-up as the optimum one reachable by two-level algorithmic hierarchies based on statistically independent parallel random samples.

(2) Being in $O(\sqrt{N})$, the optimal time complexity for HPPSA $\mathcal{X}$ is in the same class as the time complexity of a deterministic fully synchronized and co-operating

parallel algorithm with $\sqrt{N}$ processors testing separate $\sqrt{N}$-tuples of elements, when neglecting the time demands for co-operation and inspection of outputs.

(3) A more detailed optimization (i.e. minimization) of the expression $\alpha n+\beta k$ by choosing optimal multiplicative constants when defining $m(N)$, $n(N)$ and $k(N)$ would be possible, following the way of optimization used in [3] for the case of conflict-free algorithms. However, such a computation seems to be rather a matter of technical routine of optimization the extent of which would make this paper undesirably long, so this computation could be left to the reader.

(4) The claim that $\alpha n+\beta k$ in $O(\sqrt{N})$ is sufficient in the sense and for the reasons introduced above should be re-considered and modified supposing that α and β depend on N and m. In [3] we investigate the natural case when $\alpha = \alpha_0 \log_2 N$ and $\beta = \beta_0 \log_2 m$ for conflict-free algorithms and the optimum parameters m, n and k are proved to differ from those for fixed α and β only by certain $\sigma(\sqrt{N})$-functions; it seems to be a matter of rather technical routine to deduce analogous results also in our case.

5. Many-level hierarchies

When looking for a way how to improve the qualitites of two-level HPPSA's investigated above in order to reduce the time computational complexity for the searching problem $\langle A, V\rangle$ into the $\sigma(\sqrt{N})$-class, the straightforward idea is to use appropriate many-level hierarchies. Let us begin with a formal definition which is an immediate analogy of the definition of two-level HPPSA's; we shall turn back to a more intuitive description later.

A many-level hierarchical parallel probabilistic searching algorithm (MLHPPSA) *for searching problem* $\langle A, V\rangle$ *with possible data access conflicts* is a system

$$\mathcal{X} = \langle \{X_{ij}^k\}_{k=1,\ i=1,\ j=1}^{K,\ N_k,\ n_k} \rangle \tag{37}$$

of mutually statistically independent random variables defined on the fixed probability space $\langle \Omega, \mathcal{S}, P\rangle$, such that $N_K = 1$, $N_k > 1$ for $k < K$, with each X_{ij}^k taking its values in the set $A_{k-1} = \{1, 2, \ldots, N_{k-1}\}$ of integers (where $A_0 = A$) and with the uniform probability distribution, hence, for each $k \leq K$, $i \leq N_k$, $j \leq n_k$ and $r \leq N_{k-1}$ (where $N_0 = N$),

$$P(\{\omega :\ \omega \in \Omega,\ X_{ij}^k(\omega) = r\}) = N_{k-1}^{-1}. \tag{38}$$

As in the more simple case of two-level hierarchies, the system $\mathcal{X}$ defines a two-valued random variable $\mathcal{X}_0$ which can be interpreted as a statistical decision

function for the searching problem $\langle A, V\rangle$. The inductive definition reads as follows. Set

$$v_0 = V,$$

$$V_r = V_r(\omega) = \left\{ i : i \leq N_r, \sum_{j=1}^{n_r} \gamma(i,j,r,\omega) > 0 \right\}, \tag{39}$$

where

$$\gamma(i,j,r,\omega) = \chi_{V_{r-1}(\omega)}\left(X_{ij}^r(\omega)\right) \cdot \prod_{k=1,k\neq i}^{m} \left[1 - \chi_{\{X_{ij}^r(\omega)\}}\left(X_{kj}^r(\omega)\right)\right]. \tag{40}$$

This agrees, for $\gamma(i,j,\omega) = \gamma(i,j,1,\omega)$, with the definition of HPPSA above, when

$$V_1(\omega) = \left\{ i : i \leq N_1, \sum_{j=1}^{n_1} \gamma(i,j,\omega) > 0 \right\}, \tag{41}$$

for $r = K$ we obtain

$$\begin{aligned} V_K(\omega) = \{1\} = A_K \text{ iff } \sum_{j=1}^{n_K} \gamma(1,j,K,\omega) > 0, \\ V_K(\omega) = \emptyset \text{ iff } \sum_{j=1}^{n_K} \gamma(1,j,K,\omega) = 0. \end{aligned} \tag{42}$$

Evidently, $V_k(\omega) = \emptyset$ implies $V_l(\omega) = \emptyset$ for each $k \leq l \leq K$. Finally, set

$$\begin{aligned} \{\omega : \omega \in \Omega,\ \chi_0(\omega) = 1\} = \bigcap_{k=0}^{K} \{\omega : \omega \in \Omega, \quad V_k(\omega) \neq \emptyset\}, \\ \chi_0(\omega) = \chi_0(\langle A, V\rangle)(\omega) = 0 \text{ otherwise.} \end{aligned} \tag{43}$$

As before, the random event $\chi_0(\omega) = 1$ is taken as the decision that $V \neq \emptyset$ and this decision is always correct, being based on the positive testing of at least one element from V by a first-level processor. The random event $\chi_0(\omega) = 0$ is taken as the decision that $V = \emptyset$ and it may be charged by an error, or, elements from V can be either disregarded by first-level processors or the report about their finding can be disregarded by higher-level processors. Hence, the value $P(\{\omega : \omega \in \Omega, \chi_0(\omega) = 0\})$ may be taken, if $V \neq \emptyset$, as the probability of error, connected with the MLHPPSA in question, with the aim to keep this value below a threshold value $\varepsilon > 0$, uniformly for all nonempty $V \subset A$. For our further reasonings dealing with the time computational complexity of a MLHPPSA meeting this demand the following lemma may be of use.

Lemma 1. For each $\varepsilon, \delta > 0$ there exists a natural number $n_0 = n_0(\varepsilon, \delta)$ independent of N such that for all $n \geq n_0$ and for $m = \lfloor \delta N \rfloor - 1$

$$\left(1 - \frac{m}{N}\left(1 - \frac{1}{N}\right)^{m-1}\right)^n < \varepsilon. \qquad \square \quad (44)$$

Proof. The inequality

$$\left(1 - \frac{1}{N}\right)^N < e^{-1} < \left(1 - \frac{1}{N}\right)^{N-1} \quad (45)$$

together with $m = \lfloor \delta N \rfloor - 1$ yields that

$$\left(1 - \frac{m}{N}\left(1 - \frac{1}{N}\right)^{m-1}\right)^n = \left(1 - \delta\left(1 - \frac{1}{N}\right)^{\lceil \delta N \rceil - 2}\right)^n < (1 - \delta e^{-\delta})^n, \quad (46)$$

so that

$$n > \frac{\ln \varepsilon}{\ln(1 - \delta e^{-\delta})} = \frac{\ln(1/\varepsilon)}{\delta e^{-\delta}} = e^{\delta} \delta^{-1} \ln(1/\varepsilon) \quad (47)$$

implies (44). The lemma is proved. $\square$

Recalling the interpretation of (44) offered by (10), we can easily see that an analogous threshold value for conflict-free random samples reads $n_0 = \delta^{-1} \ln(1/\varepsilon)$. Hence, the corresponding time computational complexity increases (as $e^{\delta} > 1$ for $\delta > 0$) when admitting data access conflicts, but the increase is only a multiplicative one (e^{δ} does not depend on N).

So, let us have a δ, $0 < \delta < 1$, and suppose, just for the sake of simplicity of the following considerations, that $N(= \text{card } A)$ is of the form $(1/\delta)^K$. Set $N_0 = N$, $N_i = \delta N_{i-1}$, $i = 1, 2, \ldots, K$, hence, $N_k = \delta^k N_0$, and consider N_i processors of the i-th level. Given $\varepsilon > 0$, set

$$\varepsilon_1 = \varepsilon / K, \quad n_0(\delta, \varepsilon) = (e^{\delta}/\delta)(\ln 1/\varepsilon_1). \quad (48)$$

Each of the first-level processors (there are N_1 in total) takes n_0 independent and sequential random samples from the uniform probability distribution over the basic set A, and those among the sampled elements which were not sampled, simultaneously, by another processor are tested as far as their membership in the set V is concerned. If

$$\alpha(m, n, N) = \left(1 - \frac{m}{N}\left(1 - \frac{1}{N}\right)^{m-1}\right)^n, \quad (49)$$

then with probability at least $1 - \alpha(N_1, n_0, N)$ at least one first-level processor takes the unit output value (i.e. reports an element from V). Moreover, n_0 and N_1 are chosen in such a way that

$$1 - \alpha(N_1, n_0, N) \geq 1 - \varepsilon_1 = 1 - (\varepsilon/K). \quad (50)$$

There are N_2 second-level processors and each of them takes n_0 independent and sequential random samples from the uniform probability distribution over the set A_1 of (indices of) outputs of the first-level processors (hence, card $A_1 = N_1 = \delta N$). For those (indices of the) first-level processors among the sampled ones which were not sampled, simultaneously, by another processor, the sampling second-level processor tests, whether their output values were 1 or not, i.e. test the membership of the sampled first-order processor in question in the subset V_1 of A_1 defined by (41). If the result of this test is positive, then the report about the sampled element from V occurs on the output of the corresponding second-level processor and this output takes the unit value (otherwise, the zero value). If at least one first-level processor discovered an element from V, so that $V_1 \neq \emptyset$, then with probability at least $1 - \alpha(N_2, n_0, N_1)$, an element from V is reported also at the second level. But, as can be easily seen,

$$\alpha(N_2, n_0, N_1) = \alpha(N_1, n_0, N), \tag{51}$$

so that

$$1 - \alpha(N_2, n_0, N_1) > 1 - (\varepsilon/K). \tag{52}$$

Now, the induction step is evident: there are N_3 third-level processors and each of them looks for a report about an element from V among n_0 non-colliding random samples from the (outputs of the) second-level processors. Supposing such a report is found, the corresponding third-level processor takes it by outputting the unit output value and so on. Combining the corresponding conditional probabilities and using the supposed statistical independence of all random variables in question we obtain that if $V \neq \emptyset$, then with probability at least

$$(1 - \varepsilon_1)^K > 1 - K\varepsilon_1 = 1 - \varepsilon, \tag{53}$$

the report about an element from V reaches the output of the unit K-th (i.e. the highest) level, hence the probability of error is majorized by ε.

At each level we have taken n_0 sequential samples and the operations on different levels are also executed in sequential time order, so that we have made, in total,

$$\begin{aligned} &Kn_0(\log_{1/\delta} N)\ (e^\delta/\delta)\ \ln(\varepsilon/\log_{1/\delta} N)^{-1} = \\ &=(\log_{1/\delta} N)\ (e^\delta/\delta)\ \ln(\varepsilon/\log_{1/\delta} N - \ln\varepsilon) \end{aligned} \tag{54}$$

sequential samples. Taking the unit time complexity for each sample (independent of the cardinality of the corresponding sample space) and neglecting the other operations, expression (54) approximates the time complexity of the suggested special MLHPPSA answering, within the probability of error uniformly majorized by the given $\varepsilon > 0$, the question whether $V = \emptyset$ or not. As in the conflict-free case, this complexity is, again, in the $O(\log N \log\log N)$-class, the only difference being represented by the multiplicative constant e^δ.

A more detailed optimization of the suggested MLHPPSA including the optimization of the corresponding multiplicative constant, as well as a more detailed investigation of MLHPPSA's in general, i.e. with different n_i's and with N_i's not in the form of a geometric sequence, would deserve further research.

References

1. *Kramosil, I.*, Extremum-searching hierarchical parallel probabilistic algorithms. Kybernetika **24** (1988), *2*, pp. 110–121.
2. *Kramosil, I.*, Prallel algorithms in knowledge systems (in Czech). Res. rep. No. 1507, Inst. of Inform. Theory and Automat., 1987.
3. *Kramosil, I.*, Parallel probabilistic searching and sorting algorithms. Submitted for publication.
4. *Reif, J.H.*, On synchronous parallel computations with independent probabilistic choice. SIAM Journal on Computing **13** (1984), *1*, pp. 46–55.

Иерархия параллельных вероятностных алгоритмов с возможностью конфликтов при подходе к данным

И. КРАМОСИЛ

(Прага)

Чтобы найти, какой-нибудь элемент владеющий заданным качеством находится в данном основном пространстве, параллельные процессоры случайно избирают элементы из этого пространства и проберяют те, которые не избранны другим процессором в том же шаге, чтобы исключить конфликты при подходе к данным. Процессоры высшего уровня случайно избирают процессоры низшего уровня и требуют элементов с проверяемым качеством, если каких-нибудь найдено. Выход единственного процессора самого высшего уровня определяет статистическую решающую функцию отвечающую на данный вопрос. Исследуется вероятность ошибки и временная вычислительная сложность алгорифма в случае, когда эта вероятность сверху оценивана данным порогом равномерно для всех подмножеств основного пространства.

I. Kramosil
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Prague 8
Czechoslovakia

Problems of Control and Information Theory, Vol. 18(6), pp. 397–407 (1989)

M-ESTIMATORS AND GNOSTICAL ESTIMATORS OF LOCATION

J. NOVOVIČOVÁ

(*Prague*)

(Received September 20, 1988)

This paper establishes that six from eight estimators of Kovanic [7], derived from non-statistical (so-called gnostical) considerations are particular cases of M-estimators (maximum likelihood type estimators) proposed by Huber [5]. Consistency and asymptotical normality of these estimators are proved and an influence function is derived.

1. Introduction

Several classes of estimators for a location parameter have been proposed which are less sensitive to the outlying observations and incorrect assumptions concerning the form of the basic distribution. The most extensively studied ones are generalizations of the maximum likelihood estimators or M-estimators. Various criterion functions lead to various M-estimators. The question is then that of the proper choice of function generating the respective M-estimators. Kovanic's [7] non-statistical (gnostical) considerations lead to eight gnostical criterion functions generating the gnostical estimators of a location. In the present paper there is established that six from eight gnostical estimators of a location are equivalent with M-estimators. Using known statistical properties of the M-estimators, the corresponding properties of the gnostical estimators are derived. Fabian [3] has shown that three from the eight mentioned estimators of Kovanic [7] are particular cases of α-estimators (proposed by Vajda [11]), which are minimum distance-type estimators. Using statistical properties of the general α-estimators Vajda [12] has established the corresponding statistical properties of the three Kovanic's estimators. Novovičová [10] has shown that two of the remaining five estimators suggested by Kovanic [7] (i.e. the estimators defined by the equations in the lines 2 and 6 of Table 2 on p. 311) are particular cases of a more general class of estimators, the so-called M_r-estimators for $r = 2$. Statistical properties of these gnostical estimators have been analysed.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

In this paper we concentrate on the remaining three estimators (lines 1, 3, 4 of Table 2 on p. 311 of Kovanic [7]) and on the three estimators studied previously by Fabian [3] and Vajda [12]. Since the estimators in lines 3, 4 happen to be identical, new results are established in fact for two of the Kovanic's estimators (lines 1 and 3, 4).

Let us also note that Novovičová [9] established, that the class of Kovanic's gnostical estimators for the parameters of the linear regression model [8] are equivalent to the class of M-estimators of the regression model with random carriers.

2. M-estimators of location

Let $x_1, x_2, \ldots, x_n$, $x_i \in R^1$, $n = 1, 2, \ldots$ be a sequence of independent observations from the population with distribution function $F(x - \Theta_0)$, where $\Theta_0 \in R^1$ is an unknown location parameter to be estimated after observing a vector of data $\mathbf{x}=(x_1, \ldots, x_n) \in R^n$. Given a non-constant real function $\rho : R^1 \to R^1$, the corresponding M-estimator $M_n(\mathbf{x})$ of Θ_0 is defined as the value that minimizes

$$S_n(\mathbf{x}, \Theta) = \sum_{i=1}^{n} \rho(x_i - \Theta) \tag{1}$$

i.e.

$$M_n(\mathbf{x}) = \arg\min_{\Theta} S_n(\mathbf{x}, \Theta) \tag{2}$$

or, if we take the derivative $\psi(x_i - \Theta) = -d\rho(x_i - \Theta)/d\Theta$, M-estimator of Θ_0 is defined implicitly as a solution of the equation

$$\sum_{i=1}^{n} \psi(x_i - \Theta) = 0 \tag{3}$$

with respect to Θ.

The class of M-estimators of location has been introduced by Huber [5], who then studied their properties in a series of papers; the results may be also found in Huber's monograph [6]. This class of estimators contains in particular all maximum likelihood estimators ($\psi(x) = -f'(x)/f(x)$, $x \in R^1$, where f is the assumed density of the underlying distribution), the sample mean ($\psi(x) = x$), the sample median ($\psi(x) = \operatorname{sgn}(x)$). Various choices of ψ are described, e.g. in Adrews et al. [1]. If M_n is to be resistant to the outliers and to long-tailed distributions, we should select a bounded ψ-function. An important example is the "Huber psi-function" defined as $\psi_c(x) = \operatorname{sgn}(x) \min(|x|, c)$, where c is a properly chosen parameter. Other possible choices for ψ are the so-called "redescending" functions, i.e. functions satisfying

$\psi(x) \to 0$ when $x \to \mp\infty$, which give a better performance of M_n at very long-tailed distributions. The corresponding M-estimators, called redescending, are studied e.g. by Collins [2]; see also Huber [6].

3. Gnostical estimators of location

Kovanic [7] proposed to estimate an unknown value $z_0 \in (0, \infty)$ after observing the vector of data $\mathbf{z} = (z_1, \ldots, z_n)$ by means of mappings G_n^k: $(0, \infty)^n \to (0, \infty)$ for $k = -2, -1, 0, 1, 2$, where

$$G_n^k(\mathbf{z}) = \arg\min_{z_0} J_k(\mathbf{z}, z_0), \tag{4}$$

for

$$J_k(\mathbf{z}, z_0) = \begin{cases} \sum_{i=1}^{n} f_i^k, & \text{if } k = -2, -1, 1, 2 \\ -\sum_{i=1}^{n} \ln f_i, & \text{if } k = 0, \end{cases} \tag{5}$$

with

$$f_i = \frac{2}{\left(\frac{z_i}{z_0}\right)^2 + \left(\frac{z_i}{z_0}\right)^{-2}}, \quad i = 1, 2, \ldots, n. \tag{6}$$

In order to verify this statement let us take into account that if (4) holds, then $z_0 = G_n^k(\mathbf{z})$ is a solution of the equation

$$\frac{d}{dz_0} J_k(\mathbf{z}, z_0) = 0. \tag{7}$$

Equation (7) can be rewritten in the following form:

$$\frac{d}{dz_0} \sum_{i=1}^{n} \left[\frac{2}{\left(\frac{z_i}{z_0}\right)^2 + \left(\frac{z_i}{z_0}\right)^{-2}} \right]^k = 0, \quad \text{if } k = -2, -1, 1, 2$$

$$\sum_{i=1}^{n} \frac{\left(\frac{z_i}{z_0}\right)^2 - \left(\frac{z_i}{z_0}\right)^{-2}}{\left(\frac{z_i}{z_0}\right)^2 + \left(\frac{z_i}{z_0}\right)^{-2}} = 0, \quad \text{if } k = 0,$$

what are the six equations in lines 1, 3, 4, 5, 7, 8 of Table 2 on p. 311 of Kovanic [7]. Let us note that the equations in lines 3 and 4 of the mentioned Table 2 are identical.

4. Relation between M-estimators and gnostical estimators of location

Using the hyperbolic cosine function ch x, $x \in R^1$, we see from definition (6) of f_i, that

$$f_i = \frac{1}{\operatorname{ch}\,(2\ln(z_i/z_0))}.$$

Therefore, (4) is equivalent to

$$G_n^k(\mathbf{z}) = \arg\min_{z_0} \sum_{i=1}^{n} \frac{1}{k}\left(1 - \frac{1}{\operatorname{ch}^k(2\ln(z_i/z_0))}\right), \text{ if } k = -2,-1,0,1,2, \tag{8}$$

$$G_n^k(\mathbf{z}) = \arg\min_{z_0} \sum_{i=1}^{n} \left(-\ln \frac{1}{\operatorname{ch}(2\ln(z_i/z_0))}\right), \text{ if } k = 0.$$

If (8) holds, then $z_0 = G_n^k(\mathbf{z})$ is a solution of the equation

$$\sum_{i=1}^{n} \frac{\operatorname{sh}(2\ln(z_i/z_0))}{\operatorname{ch}^{k+1}(2\ln(z_i/z_0))} = 0 \text{ for } k = -2,-1,0,1,2. \tag{9}$$

Transforming the data z_i, $i = 1,2,\ldots,n$ and the parameter by

$$2\ln z_i = x_i,\ i = 1,2,\ldots,n,\ 2\ln z_0 = \Theta_0 \tag{10}$$

one obtaines that (4) is equivalent to

$$2\ln G_n^k(\mathbf{z}) = \arg\min_{\Theta} S_n^k(\mathbf{x},\Theta),$$

where $S_n^k(\mathbf{x}.\Theta)$ is defined by (1) for $\rho(x_i - \Theta) = \rho_k(x_i - \Theta)$ and $\rho_k(x_i - \Theta)$ given by

$$\rho_k(x_i - \Theta) = \begin{cases} \dfrac{1}{k}\left(1 - \dfrac{1}{\operatorname{ch}^k(x_i - \Theta)}\right), & \text{if } k = -2,-1,1,2 \\ -\ln \dfrac{1}{\operatorname{ch}(x_i - \Theta)}, & \text{if } k = 0. \end{cases} \tag{11}$$

If (11) holds, then $\Theta = M_n^k(\mathbf{x})$ is a solution of the equation

$$\sum_{i=1}^{n} \psi_k(x_i - \Theta) = 0, \quad k = -2,-1,0,1,2$$

where $\psi_k(x) = d\rho_k(x)/dx$, $x \in R^1$ is given by

$$\psi_k(x) = \frac{\operatorname{sh} x}{\operatorname{ch}^{k+1} x}, \text{ if } k = -2,-1,0,1,2. \tag{12}$$

Moreover, it follows from here that if $M_n^k(\mathbf{x})$ for $k = -2, -1, 0, 1, 2$ is an M-estimator of Θ_0 corresponding to the ρ_k-function defined by (11), i.e.

$$M_n^k(\mathbf{x}) = \arg\min_{\Theta} \sum_{i=1}^{n} \rho_k(x_i - \Theta), \tag{13}$$

then

$$G_n^k(\mathbf{z}) = \exp\left\{M_n^k(\mathbf{x})/2\right\} \tag{14}$$

for all $\mathbf{z} \in (0, \infty)^n$ and $\mathbf{x} \in R^n$ related by the one-to-one relation (10) and for all $n = 1, 2, \ldots$.

Let us note that

$$\frac{1}{\operatorname{ch} x} = g(x) \cdot \pi, \quad x \in R^1, \tag{15}$$

where $g(x)$ is the probability denstiy from the family of hyperbolic secant densities

$$q(x) = c(a) \cdot \operatorname{sech}^a x = c(a) \operatorname{ch}^{-a} x, \quad a > 0$$

with norming constant $c(a)$ (in particular, $c(1) = \frac{1}{\pi}$, $c(2) = \frac{1}{2}$). Density $g(x)$ is from this family for $a = 1$.

Using relation (15), M-estimators M_n^k for $k = -2, -1, 1, 2$ have been defined by the property that minimize $S_n^k(\mathbf{x}, \Theta)$ for $\rho_k(x)$ of the form

$$\rho_k(x) = \begin{cases} \frac{1}{k}\left[1 - (\pi g(x))^k\right], & \text{if } k = -2, -1, 1, 2 \\ -\ln g(x), & \text{if } k = 0. \end{cases} \tag{16}$$

The corresponding $\psi_k(x)$ has then the form

$$\psi_k(x) = \frac{g'(x)}{g(x)} \cdot (g(x))^k. \tag{17}$$

Function ψ_k for $k = 1, 2$ is redescending. Function ψ_k for $k = -2, -1, 0$ is increasing. For $k = 0, 1, 2$ function ψ_k is bounded, while for $k = -1, -2$ is unbounded. Function ψ_k for $k = -2, -1, 0, 1, 2$ is odd.

5. Properties of gnostical estimators

Using known statistical properties of the general α-estimators, Vajda [12] established the corresponding statistical properties of the three gnostical estimators, i.e. in our notation, the estimators G_n^k for $k = 0, 1, 2$. He derived that these

estimators are consistent and asymptotically normally distributed under mild conditions on distribution function F.

In this section we shall concentrate our attention to the statistical properties of M-estimators M_n^k for $k = -2, -1$, i.e. to M-estimators which are equivalent to three gnostical estimators in lines 1, 3, 4 of Table 2 of Kovanic [7].

Let us note that function $\psi_k(x)$, $x \in R^1$ generating the M-estimator M_n^k for $k = -2, -1$ and defined by (12) is continuous, increasing and strictly negative (positive) for negative (positive) values of x, such that $\psi_k(x) = -\psi_k(-x)$. Moreover, function ψ_k is unbounded.

Let us assume that F is an unknown member of a family $\mathcal{F}$ of distribution functions with densities $f(x - \Theta_0)$, where f is even, bounded, unimodal, and almost everywhere differentiable function. Let $f'(x)$ be the derivative of $f(x)$ if f is differentiable at x and zero, otherwise, and $f'(x) \neq 0$ almost everywhere. We shall assume for $k = -2, -1$

$$\int_{R^1} \psi_k'(x) dF(x) < \infty \tag{18}$$

and

$$\int_{R^1} \psi_k^2(x - \Theta) dF(x) < \infty, \quad \Theta \in R^1. \tag{19}$$

Let us define

$$\lambda(\Theta) = \int_{R^1} \psi_k(x - \Theta) dF(x - \Theta_0). \tag{20}$$

Theorem 1 (*"Consistency of* M_n^k*"*). For $k = -2, -1$ estimator M_n^k is consistent in the sense that

$$\lim_{n \to \infty} M_n^k = \Theta_0 \tag{21}$$

almost surely.

Proof. If we prove that there is a Θ_0 such that $\lambda(\Theta) > 0$ for $\Theta < \Theta_0$ and $\lambda(\Theta) < 0$ for $\Theta > \Theta_0$, then (21) follows from Lemma 3 of Huber [5].

It follows from (20) that

$$\lambda(\Theta) = \int_{R^1} \psi_k(x - (\Theta - \Theta_0)) dF(x). \tag{22}$$

Function $\psi_k(x) f(x)$ is odd, so that it follows from (22)

$$\lambda(\Theta) = \int_0^\infty [\psi_k(x - (\Theta - \Theta_0) - \psi_k(x + (\Theta - \Theta_0))] \, dF(x).$$

From definition (12) of $\psi_k(x)$ and the basic relation for hyperbolic functions we have

$$\frac{\mathrm{sh}(x-(\Theta-\Theta_0))}{\mathrm{ch}^{k+1}(x-(\Theta-\Theta_0))}-\frac{\mathrm{sh}(x+(\Theta-\Theta_0))}{\mathrm{ch}^{k+1}(x+(\Theta-\Theta_0))}=$$

$$=(k+3)\mathrm{sh}(k(\Theta-\Theta_0))\cdot\mathrm{ch}(kx).$$

Therefore, it holds that

$$\lambda(\Theta)=(k+3)\mathrm{sh}(k(\Theta-\Theta_0))\int_0^\infty \mathrm{ch}(kx)f(x)dx.$$

Since $\mathrm{ch}(kx)$ is positive and $(k+3)\mathrm{sh}(k(\Theta-\Theta_0))$ is negative for $k=-1,-2$ and $\Theta>\Theta_0$, it holds that

$$\lambda(\Theta)<0.$$

Similarly, one proves that

$$\lambda(\Theta)>0 \quad \text{for } \Theta<\Theta_0.$$

The desired assertion now follows from Lemma 3 of Huber [5].

Theorem 2 (*"Asymptotic normality"*). For $k=-2,-1$ the M-estimator M_n^k is asymptotically normally distributed in the sense that

$$n^{1/2}(M_n^k-\Theta_0)\xrightarrow[d]{}N(0,V(\psi_k,F)) \quad \text{as } n\to\infty,$$

where $\xrightarrow[d]{}$ denotes convergence in distribution and

$$V(\psi_k,F)= \tag{23}$$

$$=\int_{R^1}\frac{\mathrm{sh}(x-\Theta_0)}{\mathrm{ch}^{k+1}(x-\Theta_0)}dF(x)\left(\int_{R^1}\frac{1-k\,\mathrm{sh}^2(x-\Theta_0)}{\mathrm{ch}^{k+2}(x-\Theta_0)}f(x)dx\right)^{-2}.$$

Proof. From the fact that function $\psi_k(x)f(x)$ is even and from (22) it follows that

$$\lambda(\Theta_0)=0. \tag{24}$$

Let us assume for simplicity and without loss of generality that $\Theta_0=0$. By interchanging the order of integration and differentiation one obtains

$$\lambda'(0)=\left[d/d\Theta\int_{R^1}\psi_k(x-\Theta)dF(x)\right]_{\Theta=0}= \tag{25}$$

$$= -\int_{R^1} \psi_k'(x)dF(x) = -\int_{R^1} \psi_k'(x)f(x)dx,$$

where ψ_k is defined by (12) and

$$\psi_k'(x-\Theta) = d\psi(x-\Theta)/d(x-\Theta) = -d\psi(x-\Theta)/d\Theta =$$

$$= \left(\frac{1}{\mathrm{ch}(x-\Theta)}\right)^k \frac{1-k\,\mathrm{sh}^2(x-\Theta)}{\mathrm{ch}^2(x-\Theta)} = \frac{1-k\,\mathrm{sh}^2(x-\Theta)}{\mathrm{ch}^{k+2}(x-\Theta)},\ x,\Theta \in R^1.$$

The finiteness of the integral in (25) follows from assumption (18). From the fact that function $\psi_k'(x)f(x)$ is even and almost everywhere positive it follows that

$$\lambda'(0) = -\int_{R^1} \frac{1-k\,\mathrm{sh}^2 x}{\mathrm{ch}^{k+2} x} f(x)dx < 0. \tag{26}$$

From the assumptions on F, properties of ψ_k for $k = -1, -2$ and from (24), (26) and (19) it follows that conditions (i) to (iii) of Lemma 4 of Huber [5] are satisfied and, therefore, we may immediately apply this lemma to conclude that the desired assertion holds.

We see that $V(\psi_k, F)$ is finite for a narrower class of distributions that in the case of bounded ψ-functions.

Theorem 3. The variance $V(\psi_k, F)$ in Theorem 2 satisfies

$$V(\psi_k, F) \geq [I(F)]^{-1},$$

where

$$I(F) = \int_{R^1} (f'(x)/f(x))^2 f(x)dx$$

is the Fisher information for F.

Proof. We find by partial integration that

$$\lambda'(0) = \int_{R^1} \psi_k(x)f'(x)dx. \tag{27}$$

The Schwartz inequality yields the inequality

$$V(\psi_k, F) \geq [I(F)]^{-1}.$$

We have strict inequality unless

$$\frac{\mathrm{sh}\ x}{\mathrm{ch}^{k+1} x} = -p\frac{f'(x)}{f(x)},$$

where p is some constant, that is, unless

$$f(x) = \text{const} \cdot \exp\{-\rho_k(x)/p\}$$

and then the M-estimator is the maximum likelihood estimator. Therefore, the estimator M_n^k is asymptotically efficient only if

$$f(x) = \text{const} \cdot \exp\left\{-\frac{1}{pk}\left(1 - \left(\frac{1}{\operatorname{ch} x}\right)^k\right)\right\}$$

for $k = -2, -1$.

The influence curve of an estimator T_n was introduced by Hampel [4] as a measure of the local sensitivity of an estimator, when the underlying distribution is subject to an infinitesimal contamination. It is a measure of the sensitivity of the functional counterpart $T(F)$ of T_n and is defined by

$$IC(x;\ F, T) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left[T\left((1-\varepsilon)F + \varepsilon\delta_x\right) - T(F)\right],$$

where δ_x is the distribution concentrated at the point x, $x \in R^1$, ε is the size of contamination.

Theorem 4. For $k = -2, -1$ and $\Theta_0 \in R^1$, the influence curve of M_n^k satisfies the relation

$$IC(x;\ M^k, F) = \frac{\operatorname{sh}(x - \Theta_0)}{\operatorname{ch}^{k+1}(x - \Theta_0)} \cdot \left(\int_{R^1} \frac{\operatorname{sh} x}{\operatorname{ch}^{k+1} x} f'(x)dx\right)^{-1}. \tag{28}$$

Proof. In accordance with Huber [6] the influence curve of the M-estimator generated by the ψ-function is given by

$$IC(x; M, F) = \psi\left(x - M(F)\right)\left(\int \psi'(x - M(F))dF(x)\right)^{-1}. \tag{29}$$

Substituting expressing (12) for ψ in (29) and using (27) gives (28).

6. Conclusion

Due to the fact that the function ψ_k for $k = -2, -1$ is unbounded, the influence function of the estimator M_n^k is unbounded. It means that these estimators are not resistant to the outliers and to the long-tailed distributions. On the contrary, these estimators (and thus the three corresponding Kovanic's gnostical estimators)

are more sensitive to the outliers than the estimators obtained by the least squares method. Kovanic's estimators correspondig to the M-estimators generated by ψ_k-function for $k = 0, 1, 2$ are resistant to the outliers, due to boundedness of ψ_k-function for $k = 0, 1, 2$.

References

1. *Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W.* (1972), Robust estimates of location. Survey and Advances, Princeton University Press.
2. *Collins, J.R.* (1977), Upper bounds on asymptotic variances of M-estimators of location. Ann. Statist, **5**, 646–657.
3. *Fabian, Z.* (1987), Point estimation in case of small data sets. Trans. 10th Prague Conf. on Information Theory, ..., Academia, Praha.
4. *Hampel, P.J.* (1974), Influence curve and its role in robust estimation. J. Amer. Statist. Assoc. **62**, 1179–1186.
5. *Huber, P.J.* (1964), Robust estimation of a location parameter. Ann. Math. Statist. **35**, 73–101.
6. *Huber, P.J.* (1981), Robust Statistics. J. Wiley, New York.
7. *Kovanic, P.* (1984), Gnostical theory of small sample real data. Probl. Control Inf. Theory, **13**, 303–319.
8. *Kovanic, P.* (1986), A new theoretical and algorithmical basis for estimation, identification and control. Automatica, **22**, *6*, 657–674.
9. *Novovičová, J.*, M-estimators and gnostical estimators for identification of linear regression model (submitted to Automatica).
10. *Novovičová, J.* (1990), M_r-estimators and gnostical estimators of a location. Probl. Control Inf. Theory, **19**, *2* (to appear).
11. *Vajda, I.* (1983), A new general approach to minimum distance estimation. Trans. 9th Prague Conf. on Information Theory, ..., 103–112. Academia, Praha.
12. *Vajda, I.* (1988), Minimum-distance and gnostical estimators. Probl. Control Inf. Theory, **17**, *5*, 253–266.

M-оценки и гностические оценки параметра сдвига

Й. НОВОВИЧОВА

(Прага)

В этой статье показано, что шесть оценок Кованица [7], которые выведены из нестатистических соображений, могут быть представлены как M-оценки (оценки типа

максимального правдоподобия) введенные Хубером [5]. Состоятельность и асимптотическая нормальность этих оценок доказана и выведена кривая влияния для этих оценок.

Jana Novovičová
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Prague 8
Czechoslovakia

Problems of Control and Information Theory, Vol. 18(6), pp. 409–420 (1989)

ON THE CONDITIONS OF CONTROL MODEL CLOSED FORM AND PROPERTIES OF THE REACHABLE SET FOR A DEFINITE CLASS OF PROBLEMS

S. L. GOLDSHTEIN, E. B. SOLONIN

(*Sverdlovsk*)

(Received December 9, 1988)

A problem of closed form solving of the linear task of programmed control involving a priori limited resources with the quadratic quality criterion is considered. Here the new concept of the reduced reachable set is introduced and the theorem describing its main properties is proved as well. In this case the fact that the finite state of an object belongs to the reduced reachable set (RS) provides the existence of the closed control model.

1. Introduction

Solving a control task when the object state is known at the not too many time instants can be realized as a sequence solution of programmed tasks. If the controllable action proceeds quickly and the controlling device possesses a low computing capacity it would be preferable to use closed mathematical models for the program synthesis. By program synthesis in a closed model it is meant control estimation procedure $U(t)$ according to the formula available. This formula includes time t and parameters, which are known from the problem statement. In the important case when the control resources are constrained by a priori restrictions, due to the nature of the controlling device or an object, construction of such models is more often impossible. However, this problem can be formulated and solved for linear objects and quadratic quality criterion. Suppose we have an object, whose motion may be described as a vector linear differential equation

$$\dot{x} = A(t)x + B(t)U, \tag{1}$$

where $x \in R^n$, $U \in R^r$, $t \in [t_0, T]$. Here T is a sufficiently remote time instant beyond which the motion of an object is of no interest. Matrices $A(t)$ and $B(t)$ are supposed to be uniformly continuous in $[t_0, T]$. Thus control $U(t)$ is realized on the length $[t_0, \Theta]$, $\Theta \in (t_0, T]$. Assume that the object described by equation (1) can be controllable and stable in Lyapunov sense.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Suppose we have the initial $x^0 = x(t_0)$ and final $x^k = x(\Theta)$ positions of an object. Then it is necessary to construct a closed mathematical model of programmed control $U(t)$ shifting the object from position $\{t_0, x^0\}$ to $\{\Theta, x^k\}$ position and giving minimum to the criterion

$$I = \left(\int_{t_0}^{\Theta} \|u(\tau)\|^2 d\tau \right)^{1/2}, \tag{2}$$

where $\|u(\tau)\|$ is the Euclidean norm of $U(\tau)$. When there are no a priori control restrictions, the solution of the set problem for motion (1) and criterion (2) is known [1]:

$$u(t) = B'(t)X'(\Theta, t)Q^{-1}(t_0, \Theta)\{x^k - X(\Theta, t_0)x^0\}, \tag{3}$$

where $X(\Theta, t)$ is the fundamental matrix of equation (1) and $Q^{-1}(t_0, \Theta)$ is the matrix, found by $X(\Theta, t)$ and $B(t)$, the prime ($'$) means transposition. Hence, this solution is true for any finite X^0, X^k, t_0, Θ. Let us constrain the control $U(t)$ by a priori restrictions

$$\mu_i^{(1)} \le u_i(t) \le \mu_i^{(2)}, \tag{4}$$

where $t \in [t_0, \Theta]$, $i = 1, \ldots, r$; $\mu_i^{(1)} \ne \mu_i^{(2)}$ are predetermined numbers. Assume relationship (3) is a closed control model for the problem with restriction (4). In this case the program control synthesis $U(t)$ according to model (3) under the fixed values x^0, t_0, Θ is not possible for all x^k.

The present paper shows that the existence of the closed mathematical model can be provided only by x^k involvement into the specifically found reduced reachable set $\tilde{G}(x^0, t_0, \Theta)$. The construction of such sets is considered and their main properties are investigated.

2. The reduced reachable set

This set has to satisfy the following conditions. The finite points $x^k = x(\Theta)$ of trajectories formed from position $\{t_0, x^0\}$ by the type of control (3), satisfying conditions (4), belong to $\tilde{G}(\cdot)$. At the same time for any point $X^k \in \tilde{G}(\cdot)$, there exists a control $U(t)$ of type (3), satisfying conditions (4) and transferring an object from position $\{t_0, x^0\}$ to the position $\{\Theta, x^k\}$.

Let some preliminary remarks be made. Some properties of the continuity of reduced reachable sets will be defined later. Let $d\{\tilde{G}(x'), \tilde{G}(x'')\}$ be the distance between reduced reachable sets $\tilde{G}(x')$, and $\tilde{G}(x'')$, where x is a variable according to which the continuity is determined. According to definition (2)

$$d\{\tilde{G}(x'), \tilde{G}(x'')\} = \inf[\alpha > 0 : \tilde{G}(x') \subset S(\tilde{G}(x'', \alpha), \tilde{G}(x'') \subset S(\tilde{G}(x'), \alpha)],$$

where $S(\tilde{G}(\cdot), \alpha)$ is the α-neighborhood of the set $\tilde{G}(\cdot)$. Here continuity $\tilde{G}(x)$ for x occurs when for any $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that

$$d\{\tilde{G}(x'), \tilde{G}(x'')\} < \varepsilon, \quad |x' - x''| < \delta(\varepsilon). \tag{5}$$

In order to prove the continuity properties of reduced RS we turn to the control $\tilde{u}(t)$, whose components are constrained by the inequalities

$$|\tilde{u}_i(t)| \leq \mu_i, \tag{6}$$

where $i = 1, \ldots, r$, $t \in [t_0, \Theta]$, $\mu_i > 0$. Vector $\mu = \{\mu_1, \ldots, \mu_r\}$ is now determined as a limiting vector. The bound between vectors $u(t)$ and $\tilde{u}(t)$ is given by the relationship

$$\tilde{u}_i(t) = u_i(t) - (\mu_i^{(1)} + \mu_i^{(2)})/2, \tag{7}$$

where $\mu_i = (\mu_i^{(2)} - \mu_i^{(1)})/2$. Equation (1) after transformation (7) is reduced to

$$\dot{x} = A(t)x + B(t)\tilde{u} + C(t), \tag{8}$$

where $C(t) = \{C_i(t), i = 1, \ldots, n\}$ is a uniformly continuous vector with respect to $[t_0, T]$,

$$C_i(t) = \frac{1}{2}\sum_{l=1}^{r} B_{il}(t)(\mu_l^{(1)} + \mu_l^{(2)}).$$

According to (1) the control $\tilde{u}(t)$ also has form (3), where $\tilde{x}^k = x^k - \int_{t_0}^{\Theta} X(\Theta, \tau)C(\tau)d\tau$ can be substituted for x^k. Further, the equation of motion (9) will also be considered

$$\dot{x} = A(t)x + B(t)\tilde{u}. \tag{9}$$

Let the reduced RS, constructed for motion (9), be expressed in terms of $\tilde{G}(x^0, t_0, \Theta, \mu)$. Thus, the property of the continuity is found to be true both for reduced RS $\tilde{G}(x^0, t_0, \Theta, \mu)$, and $\tilde{G}(x^0, t_0, \Theta)$ as well.

In fact, applying the equivalence (1) and (8) for any $x^k = x(\Theta) \in \tilde{G}(x^0, t_0, \Theta)$, the following Cauchy formula can be written:

$$x(\Theta) = X(\Theta, t_0)x^0 + \int_{t_0}^{\Theta} X(\Theta, \tau)B(\tau)\tilde{u}(\tau)d\tau + \int_{t_0}^{\Theta} X(\Theta, \tau)C(\tau)d\tau, \tag{10}$$

where the first two members of the right part correspond to equation (9) and, therefore, they correspond to the points $x(\Theta) \in \tilde{G}(x^0, t_0, \Theta, \mu)$. Note that the third addend in (10) does not depend on $\tilde{u}(t)$ and x^k and, therefore, the transition from

$\tilde{G}(x^0, t_0, \Theta, \mu)$ to $\tilde{G}(x^0, t_0, \Theta)$ can be done only by the vector shift which is the same for all x^k.

Let $\tilde{G}_1(x^0, t_0, \Theta, \mu)$ and $\tilde{G}_2(x^0, t_0, \Theta', \mu)$ be constructed for motion (9) and restrictions (6), and $\tilde{G}_a(x^0, t_0, \Theta)$ and $\tilde{G}_b(x^0, t_0, \Theta')$ — for motion (1) and restriction (4), moreover, $u(t)$ and $\tilde{u}(t)$ being connected with relationship (7). Let $\tilde{G}(x^0, t_0, \Theta, \mu)$ be continuously dependent on Θ. Then $|\Theta - \Theta'| < \delta^*(\varepsilon)$ can be chosen for any $\varepsilon > 0$ so that $d\{\tilde{G}_1(\cdot), \tilde{G}_2(\cdot)\} < \varepsilon/2$. As the third summand on the right part of (10) has the uniform continuity, it satisfies the Lipshitz condition

$$\left| \int_{t_0}^{\Theta} X(\Theta, \tau) C(\tau) d\tau - \int_{t_0}^{\Theta'} X(\Theta', \tau) C(\tau) d\tau \right| \leq K |\Theta - \Theta'|.$$

Having chosen $|\Theta - \Theta'| < \delta(\varepsilon) = \min\left\{\delta^*(\varepsilon), \frac{\varepsilon}{2K}\right\}$, we have an inequality $d\left\{\tilde{G}_a(\cdot), \tilde{G}_b(\cdot)\right\} < \varepsilon$ which means the continuity $\tilde{G}(x^0, t_0, \Theta)$ according to Θ.

Let us now prove an auxiliary statement.

Lemma 1. Let $u^{(1)}(t)$ be the control program of type (3) for some x^0, t_0, x^k, Θ. Then the program $u^{(2)}(t) = Cu^{(1)}(t)$, where C is an arbitrary number, can be expressed by formula (3) for the same x^0, t_0, Θ and some $\tilde{x}^k$.

As $u^{(1)}(t)$ is the known control program, the right side of the corresponding Cauchy formula can be substituted for x^k in (3):

$$u^{(1)}(t) = D \cdot \int_{t_0}^{\Theta} X(\Theta, \tau) B(\tau) u^{(1)}(\tau) d\tau,$$

where $D = B'(t) X'(\Theta, t) Q^{-1}(t_0, \Theta)$. Turn to the program $u^{(2)}(t)$:

$$u^{(2)}(t) = Cu^{(1)}(t) = D \cdot \int_{t_0}^{\Theta} X(\Theta, \tau) B(\tau) Cu^{(1)}(\tau) d\tau =$$

$$= D \cdot \int_{t_0}^{\Theta} X(\Theta, \tau) B(\tau) u^{(2)}(\tau) d\tau = D \cdot \{X(\Theta, t_0) x^0 +$$

$$+ \int_{t_0}^{\Theta} X(\Theta, \tau) B(\tau) u^{(2)}(\tau) d\tau - X(\Theta, t_0) x^0\} =$$

$$= B'(t) X'(\Theta, t) Q^{-1}(t_0, \Theta) \{\tilde{x}^k - X(\Theta, t_0) x^0\},$$

where $\tilde{x}^k = X(\Theta, t_0) x^0 + \int_{t_0}^{\Theta} X(\Theta, \tau) B(\tau) u^{(2)}(\tau) d\tau$, as it was shown.

3. Properties of the reduced reachable set

Let us investigate the nature of the relationship of reduced *RS* $\tilde{G}(x^0, t_0, \Theta, \mu)$ to restricting vector μ. The following statement holds:

Lemma 2. The reduced reachable set $\tilde{G}(\cdot, \mu)$ continuously depends on μ.

Fix the moment $\Theta \in (t_0, T]$. The next step is to consider two reduced *RS*: $\tilde{G}(x^0, t_0, \Theta, \mu_*)$ and $\tilde{G}(x^0, t_0, \Theta, \mu^*)$ where $\mu^* = (1+\xi)\mu_*$, ξ is defined to be a small number. In accordance with (5) it is necessary to show that with any $\varepsilon > 0$ there is $\xi(\varepsilon) > 0$ which is such a function that, if $|\xi| < \xi(\varepsilon)$, then:

1. for any $x^{(1)} \in \tilde{G}(\cdot, \mu^*)$ we can find $x^{(2)} \in \tilde{G}(\cdot, \mu^*)$ such that $|x^{(2)} - x^{(1)}| < \varepsilon$;

2. for any $x^{(2)} \in \tilde{G}(\cdot, \mu^*)$ we can find $x^{(1)} \in \tilde{G}(\cdot, \mu_*)$ and $|x^{(1)} - x^{(2)}| < \varepsilon$.

Take an arbitrary point $x^{(1)} \in \tilde{G}(\cdot, \mu_*)$. The trajectory formed from the position $\{t_0, x^*\}$ by the control program $\tilde{u}_*(t)$ corresponds to this point. In this case the control program is definned to be of type (3) and it operates under the limiting vector μ_*.

Let us construct the program $\tilde{u}^*(t)$ according to the formula

$$\tilde{u}^*(t) = (1 + \xi)\tilde{u}_*(t).$$

Thus according to Lemma 1, such a program can also be expressed by formula (3) with some $x^k = x^{(2)}$. Obviously, the limiting vector for $\tilde{u}^*(t)$ is found to be $\mu^* = (1 + \xi)\mu_*$, therefore $x^{(2)} \in \tilde{G}(\cdot, \mu^*)$. Using the Cauchy formula we shall have neighbourhood estimation $x^{(1)}$ and $x^{(2)}$:

$$|x^{(2)} - x^{(1)}| = \Big| \int_{t_0}^{\Theta} X(\Theta, \tau)B(\tau)[\tilde{u}^*(\tau) - \tilde{u}_*(\tau)]d\tau \Big| \leq$$

$$\leq |\xi| \cdot \Big| \int_{t_0}^{\Theta} X(\Theta, \tau)B(\tau)\tilde{u}_*(\tau)d\tau \leq |\xi| \cdot \int_{t_0}^{\Theta} \|X(\Theta, \tau)\| \cdot \|B(\tau)\| \cdot |\tilde{u}_*(\tau)|d\tau,$$

where $\|Z\| = \max\limits_{|x| \leq 1} |Zx|$ is the matrix Z norm. Due to the uniform continuity of matrices $X(\Theta, \tau)$ and $B(\tau)$ according to τ, their normal values are uniformly constrained for $[t_0, \Theta]$, $\Theta \in (t_0, T]$:

$$\|X(\Theta, \tau)\| \leq \eta_1, \quad \|B(\tau)\| \leq \eta_2,$$

where η_1, η_2 are constants. The value $|\tilde{u}_*(\tau)|$ is constrained for $[t_0, \Theta]$ by the number $m = \left(\sum\limits_i \mu_{*i}^2\right)^{1/2}$ and, moreover, by the number

$$M = \max\left\{m, \left(\sum_i (\mu_i^*)^2\right)^{1/2}\right\}. \tag{11}$$

Thus we have

$$|x^{(2)} - x^{(1)}| \leq \eta_1 \eta_2 M(\Theta - t_0)|\xi| \leq \eta_1 \eta_2 M(T - t_0)|\xi|.$$

Having chosen $|\xi| < \xi(\varepsilon)$, where

$$\xi(\varepsilon) = \frac{\varepsilon}{\eta_1 \eta_2 M \cdot (T - t_0)} \tag{12}$$

we shall have $|x^{(2)} - x^{(1)}| < \varepsilon$, whatever the number $\varepsilon > 0$ is, as it was shown. The second point of the lemma is similarly proved involving some changes. Due to the choice of number M (11) we shall have here the same expression (12) for $\xi(\varepsilon)$.

Let us continue the study of the properties of reduced RS. Next to be considered are the domains of attainability $\tilde{G}(x^0, t_0, \Theta)$, $\Theta \in (t_0, T]$, constructed for motion (1) and restrictions (4). Here the following theorem holds.

Theorem. The reduced reachable set $\tilde{G}(x^0, t_0, \Theta)$ is a convex compact set in R^n and it continuously depends on Θ.

In order to be able to deal with it, we shall prove the convexity and compactness constructively and give a practical algorithm for constructing the reduced RS. Using formula (3) the inequalities (4) with fixed t can be written as follows:

$$\begin{cases} \varphi_i(x^k) + \beta_i \geq \mu_i^{(1)} \\ \varphi_i(x^k) + \beta_i \leq \mu_i^{(2)}, \end{cases} \tag{13}$$

where $\varphi_i(x^k)$ are linear forms of x^k, and β_i are constants. It is known [3] that any system of linear inequalities, (13) in particular, determines the convex set. In this case it is closed. This set includes all x^k belonging to hyperplanes $\varphi_i(x^k) = \mu_i^{(1)} - \beta_i$, $\varphi_i(x^k) = \mu_i^{(2)} - \beta_i$ or lying between them.

Construct the convex closed set:

$$\tilde{G}(x^0, t_0, \Theta) = \bigcap_{i=1}^{r} \tilde{G}^{(i)}(\cdot), \tag{14}$$

$$\tilde{G}^{(i)}(\cdot) = \bigcap_{t=t_0}^{\Theta} \tilde{G}_t^{(i)}(\cdot). \tag{15}$$

It is the reduced RS needed. Indeed, in construction $\tilde{G}(\cdot)$ it includes all the trajectories formed from position $\{t_0, x^0\}$ by controls of type (3), all r of the components satisfy inequalities (4) with all t on the length $[t_0, \Theta]$. Besides the convexity and closed form the reduced RS domain of attainability $\tilde{G}(x^0, t_0, \Theta)$ possesses the restriction properties. Let $G(x^0, t_0, \Theta)$ be the reachable set constructed in a usual way in the class of admissible controls with the same restrictions (2). It is known

[4] that $G(\cdot)$ is the constrained set. At the same time as the control of type (3) is one of the cases of admissible controls, the inclusion $\tilde{G}(\cdot) \subset G(\cdot)$ which proves that the restriction $\tilde{G}(\cdot)$ holds true. This inclusion served as a basis in choosing the name for the sets $\tilde{G}(\cdot)$.

Thus, the reduced *RS* $\tilde{G}(x^0, t_0, \Theta)$ is convex, closed and constrained in R^n, therefore, it is a convex compact.

Let us prove the continuity $\tilde{G}(\cdot)$ according to Θ. Turn first to the reduced *RS* $\tilde{G}(x^0, t_0, \Theta, \mu)$. Fix $\Theta_1 \in (t_0, T)$ (for $\Theta_1 = T$ the left-side continuity occurs, proving like lower) and consider reduced *RS* $\tilde{G}_a(x^0, t_0, \Theta_1, \mu)$ and $G_b(x^0, t_0, \Theta_2, \mu)$. It has to be shown that for any $\varepsilon > 0$ there exists such $\delta(\varepsilon) > 0$ that if $|\Theta_2 - \Theta_1| = |\delta| < \delta(\varepsilon)$, then

1. for any $X_* \in \tilde{G}_a(\cdot)$ we can find $X^* \in \tilde{G}_b(\cdot)$ that $|X^* - X_*| < \varepsilon$;
2. for any $X^* \in \tilde{G}_b(\cdot)$ we can find $X_* \in \tilde{G}_a(\cdot)$ that $|X^* - X_*| < \varepsilon$.

Take an arbitrary point $X_* \in \tilde{G}_a(\cdot)$. The control program $u^{(1)}(t)$ of type (3) transferring an object from position $\{t_0, X^0\}$ to position $\{\Theta_1, X_*\}$ corresponds to the point. Construct the program using the same point $X^k = X_*$ and the moment $\Theta = \Theta_2$ in formula (3). It follows that $X_* \in \tilde{G}_c(\cdot) = \tilde{G}_c(X^0, t_0, \Theta_2, \mu^*)$. Because of control restrictions of type (3) at any final time instant, vector μ^* can be expressed as follows:

$$\mu^* = (1 + \xi)\mu, \tag{16}$$

where ξ is a certain positive number (the case $\xi \leq 0$ is out of interest, because $\tilde{G}_c(\cdot) \in \tilde{G}_b(\cdot)$ and we can take x_* instead of x^*). We will show that with the right choice δ, $|\delta| = |\Theta_2 - \Theta_1|$, the number ξ may be as small as desired.

Control $u(t)$ of type (3) continuously depends on Θ in $(t_0, T]$ and satisfies the Lipshitz condition for t in $[t_0, \Theta]$ [1]. Take i-components of $\tilde{u}^{(1)}(t)$ and $\tilde{u}^{(2)}(t)$ vectors. Consider first the case $\Theta_2 < \Theta_1$. Because of continuity for any $\xi > 0$, we can find such $\delta_1(\xi, \Theta_1, \mu_i)$ than on condition $|\delta| < \delta_1(\cdot)$ inequality $|\tilde{u}_i^{(2)}(t) - \tilde{u}_i^{(1)}(t)| < \mu_i \xi$ is true for any $t \in [t_0, \Theta_2]$ (we have to remind of $\mu_i > 0$). Therefore, we obtain estimation

$$|\tilde{u}_i^{(2)}(t)| < |\tilde{u}_i^{(1)}(t)| + \mu_i \xi \leq \mu_i(1 + \xi). \tag{17}$$

Now let $\Theta_2 > \Theta_1$. The estimation analogous to (17) holds true for the length $[t_0, \Theta_1]$, where both controls are determined:

$$|\tilde{u}_i^{(2)}(t)|_{\{t \in [t_0, \Theta_1]\}} < \mu_i + \frac{\mu_i \xi}{2}, \tag{18}$$

if $|\delta| < \delta_2(\xi, \Theta_1, \mu_i)$. The control $\tilde{u}_i^{(2)}(t)$ satisfies the Lipshitz condition for the length $[\Theta_1, \Theta_2]$. It follows that $\tilde{u}_i^{(2)}(t)$ has a bounded variation for the same length [5] and

$$\bigvee_{\Theta_1}^{\Theta_2} \tilde{u}_i^{(2)}(t) \leq \lambda |\Theta_2 - \Theta_1| = \lambda |\delta|.$$

Choosing $|\delta| < \delta_3(\xi, \mu_i) = \frac{\mu_i \xi}{2\lambda}$ we obtain an inequality

$$\bigvee_{\Theta_1}^{\Theta_2} \tilde{u}_i^{(2)}(t) < \frac{\mu_i \xi}{2}. \tag{19}$$

If we take $|\delta| < \delta_*(\xi, \Theta_1, \mu_i) = \min\{\delta_2(\cdot), \delta_3(\cdot)\}$, then inequalities (18), (19) will be realized at the same time. Here the result is

$$|\tilde{u}_i^{(2)}(t)|_{\{t \in [t_0, \Theta_2]\}} < \mu_i(1 + \xi). \tag{20}$$

Now we take $|\delta| < \delta^*(\cdot) = \min\{\delta_1(\cdot), \delta_*(\cdot)\}$. Estimations (17) and (20) do not depend on the sign of difference $\Theta_2 - \Theta_1$. Finally, choosing $|\delta| < \delta(\xi, \Theta_1) = \min_i\{\delta^*(\cdot)\}$ provides inequalities (17), (20) hold true for all r components of vector $\tilde{u}^{(2)}(t)$ and for any $\xi > 0$. Consequently, the number ξ in (16) can be made as small as desired.

In accordance with Lemma 2 for $x_* \in \tilde{G}_c(x^0, t_0, \Theta_2, \mu^*)$, there exists such a point $x^* \in \tilde{G}_b(x^0, t_0, \Theta_2, \mu^*)$ when the number ξ in (16) satisfies the inequality $|\xi| < \xi(\varepsilon)$ where $\xi(\varepsilon)$ is expressed by formula (12), then $|x^* - x_*| < \varepsilon$ is true for any $\varepsilon > 0$. So, for any $\varepsilon > 0$, we can choose such $\delta(\xi, \Theta) = \delta(\xi(\varepsilon), \Theta)$, that on condition $|\delta| < \delta(\cdot)$ the inequality $|x^* - x_*| < \varepsilon$ holds true. All that proves the validity of the first item. The proving of the second item is similar with clear changes.

Finally, we can use the fact (see Section 2) that the property of continuity $\tilde{G}(x^0, t_0, \Theta, \mu)$ extends to the reduced RS $\tilde{G}(x^0, t_0, \Theta)$ and we may conclude that $\tilde{G}(x^0, t_0, \Theta)$ continuously depends on Θ. So the Theorem is proved.

It is interesting to note that in spite of a sufficient difference between the reduced RS $\tilde{G}(\cdot)$ and usual RS $\tilde{G}(\cdot)$, their properties shown in the above theorem coincide.

4. Example

Consider a simple example of constructing the reduced RS and a closed model of programmed control optimizing criterion (2). Suppose we have a biphase object with a controllable interphase transfer $u_1(t)$ and outside material flow $u_2(t)$ directed to one of the phases. In both phases uncontrollable processes leading to the leakage of active substance occur. Those leakages are linear for coordinates. Masses and concentrations of active substance in phases can act as coordinates x_1, x_2. Such an object can be described by an autonomous system containing two differential equations:

$$\begin{cases} \dot{x}_1 = -a_1 x_1 - b_1 u_1 + b_3 u_2 \\ \dot{x}_2 = -a_2 x_2 + b_2 u_1, \end{cases} \tag{21}$$

where $a_i > 0$, $t_i \geq$, $0 \leq u_i(t) \leq \mu_i$ $(i = 1, 2)$. Assume that the moment t_0 is equal to zero. The aim of the control is to transfer an object from position $\{0, x^0\}$ to position $\{\Theta, x^k\}$. The object described in system (21) is controllable if one of the conditions: 1) $b_1 b_2 b_3 \neq 0$; 2) $(a_1 - a_2) b_1 b_2 \neq 0$ is fulfilled. After Lyapunov, the object (21) is asymptotically stable.

Controls $u_1(t)$, $u_2(t)$ developed from formula (3) take the form

$$\begin{cases} u_1(t) = b_2 q(\alpha_2 C_1 + \alpha_3 C_2) e^{a_2 t} - b_1 p(\alpha_1 C_1 + \alpha_2 C_2) e^{a_1 t} \\ u_2(t) = b_3 p(\alpha_1 C_1 + \alpha_2 C_2) e^{a_1 t}, \end{cases} \tag{22}$$

where $p = e^{-a_1 \Theta}$, $q = e^{-a_2 \Theta}$, $C_1 = x_1^k - x_1^0 p$, $C_2 = x_2^k - x_2^0 q$. Coefficients α_i are elements of matrix Q^{-1}:

$$\alpha_1 = Q_{11}^{-1} = \frac{b_2^2}{2a_2 y}(1 - q^2), \quad \alpha_2 = Q_{12}^{-1} = Q_{21}^{-1} = \frac{b_1 b_2}{(a_1 + a_2) y}(1 - pq),$$

$$\alpha_3 = Q_{22}^{-1} = \frac{b_1^2 + b_3^2}{2a_1 y}(1 - p^2),$$

$$y = \det Q = b_2^2 \cdot \left[\frac{b_1^2 + b_3^2}{4a_1 a_2}(1 - p^2)(1 - q^2) - \frac{b_1^2}{(a_1 + a_2)^2}(1 - pq)^2 \right].$$

Construct the reduced RS $\tilde{G}(\cdot) = \tilde{G}^{(1)}(\cdot) \cap \tilde{G}^{(2)}(\cdot)$, using formulas (13)–(15). The set $\tilde{G}^{(1)}(\cdot)$ can be found from the condition $0 \leq u_1(t) \leq \mu_1$, $t \in [0, \Theta]$, the set $G^{(2)}(\cdot)$ from the condition $0 \leq u_2(t) \leq \mu_2$, $t \in [0, \Theta]$.

Sets $\tilde{G}_t^{(1)}(\cdot)$, intersection of which (15) results in $\tilde{G}^{(1)}(\cdot)$, are constrained by straight lines

$$x_2^k = x_2^0 q + \frac{\mu_1 - (b_2 \alpha_2 q e^{a_2 t} - b_1 \alpha_1 p e^{a_1 t})(x_1^k - x_1^0 p)}{b_2 \alpha_3 q e^{a_2 t} - b_1 \alpha_2 p e^{a_1 t}} \tag{23}$$

$$x_2^k = x_2^0 q - \frac{(b_2 \alpha_2 q e^{a_2 t} - b_1 \alpha_1 p e^{a_1 t})(x_1^k - x_1^0 p)}{b_2 \alpha_3 q e^{a_2 t} - b_1 \alpha_2 p e^{a_1 t}}. \tag{24}$$

Accordingly relationships (23), (24) are obtained from the conditions: $u_1(t) = \mu_1$, $u_1(t) = 0$. Segments of straight lines (23), (24) are formed the round of the set $\tilde{G}^{(1)}(\cdot)$ with $t = 0$ and $t = \Theta$. The set of lines (23) have the round for the interval $(0, \Theta)$, which is also a part of the bound of $\tilde{G}^{(1)}(\cdot)$ (the round for the set (24) is degenerated into a point). The round equations can be determined from the conditions $u_1(t) = \mu_1$, $\dot{u}_1(t) = 0$, $0 < t < \Theta$ and take the form:

$$x_1^k = x_1^0 p + \frac{\mu_1 \dot{D}_2}{D_1 \dot{D}_2 - D_2 \dot{D}_1}, \quad x_2^k = x_2^0 q + \frac{\mu_1 \dot{D}_1}{D_1 \dot{D}_2 - D_2 \dot{D}_1}, \tag{25}$$

where $D_1 = D_1(t) = b_2\ \alpha_2 q e^{a_2 t} - b_1 \alpha_1 p e^{a_1 t}$, $D_2 = D_2(t) = b_1 \alpha_2 p e^{a_1 t} - b_2 \alpha_3 q e^{a_2 t}$. The point $x^k = x(\Theta)$ of the round (25) being the final trajectory point formed from position $\{0, x^0\}$ by the control $u_1(t)$ (22) corresponds to each t, $t \in (0, \Theta)$. At the moment t the given control is at maximum, equal to μ_1.

It follows from type (22) of the control $u_2(t)$ that the set $\tilde{G}^{(2)}(\cdot)$ is enclosed between straight lines

$$\begin{cases} x_2^k = x_2^0 q - \frac{\alpha_1}{\alpha_2}(x_1^k - x_1^0 p) + \frac{\mu_2}{\alpha_2 b_3} \\ x_2^k = x_2^0 q - \frac{\alpha_1}{\alpha_2}(x_1^k - x_1^0 p), \end{cases} \tag{26}$$

where the first relationship is obtained from the condition $u_2(\Theta) = \mu_2$ and the second from $u_2(t) = 0$.

Besides equations (23)–(26) we used the means of computer graphics to choose segments of the bounds of sets $\tilde{G}_t^{(1)}(\cdot)$, $\tilde{G}^{(2)}(\cdot)$, which are at the same time the bound of the reduced *RS* (14). Note, that the computation of infinite intersection (15) for $n > 2$ in practice should be replaced by computing the finite intersection at discrete instants t_i, $t_i \in [t_0, \Theta]$.

The following coefficients of system (21): $a_1 = 3 \cdot 10^{-5}$, $a_2 = 15 \cdot 10^{-5}$, $b_1 = b_2 = 10^{-3}$, $b_3 = 1$ and the restrictions: $\mu_1 = 1$, $\mu_2 = 6 \cdot 10^{-3}$ were chosen for numerical computation. Here $x^0 = \{50, 50\}$ was chosen as an initial point. The reduced *RS* $\tilde{G}(\cdot)$, corresponding to the moments $\Theta_1 = 10500$ (a) and $\Theta_2 = 7000$ (b) are shown in Fig. 1. For comparison the reachable sets $G(\cdot)$ are given here. The latter are constructed with the same initial data in the class of admissible controls according to generally accepted methods (dot-dash line).

The closed control model obtained (22) was tested by experiment involving an approximate model constructed by the method of convex programming in the class of admissible controls. The experiment was realized using the computer. The comparison shows that the approximate model has the speed which is 50–100 times less than of model (22). The estimate was made according to the expenditure of computing resources for control program synthesis. It is seen that computing accuracy with respect to model (22) is limited only by errors due to the computer used and accounts for 6–7 significant decimal figures, whereas the approximate model is not able to provide the accuracy exceeding 3–4 significant figures for an admissible time. Thus, the results of the computing experiment confirm the efficiency of the method proposed in the given article. One of the negative moments of reducing possible controls to the controls of type (3) is the decrease of the reachable set $\tilde{G}(\cdot)$ size compared with $G(\cdot)$. However, in many cases it is of no principal significane. The combined use of models is quite possible when the model of type (3) is used in the reduced *RS* $\tilde{G}(\cdot)$ and the corresponding approximate model is used in the set $G(\cdot) \setminus \tilde{G}(\cdot)$.

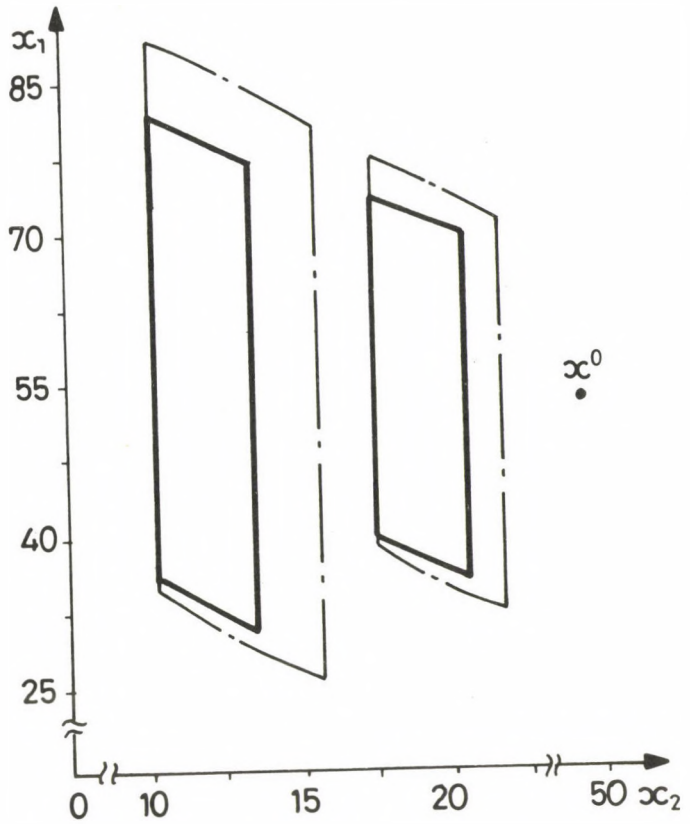


Fig. 1

5. Conclusion

The article presents the condition of control model close with a priori restrictions (4) for linear objects (1) and control quality criterion (2). This condition means that the object final state x^k belongs to the reduced reachable set. In realizing the condition the model of type (3) to be effective from the point of accuracy and the minimum expenditures of computing resources can be used for the synthesis of control programs. The properties of the reduced reachable set and the method of its construction are considered in the process of proving the formulated theorem.

References

1. Красовский Н. Н., Теория управления движением. Москва, Наука, 1968.
2. Варга Дж., Оптимальное управление дифференциальными и функциональными уравнениями. Москва, Наука, 1977.
3. Рокафеллар Р., Выпуклый анализ. Москва, Мир, 1973.
4. Ли Э., Маркус Л., Основы теории оптимального управления. Москва, Наука, 1972.
5. Фихтенгольц Г. М., Курс дифференциального и интегрального исчисления, **3**, Москва, Наука, 1969.

Об условиях замкнутости модели управления и свойствах области достижимости для одного класса задач

С. Л. ГОЛЬДШТЕЙН, Е. Б. СОЛОНИН

(Свердловск)

В статье рассмотрена проблема решения в замкнутой форме линейной задачи программного управления с априорно ограниченными ресурсами при квадратичном критерии качества. Использование замкнутых моделей управления, для которых характерны высокая точность и минимум расхода вычислительных ресурсов, наиболее эффективно в ситуации, когда состояние объекта становится известным лишь в небольшое число моментов времени и управление осуществляется на основе решения последовательности программных задач.

Введено понятие редуцированной области достижимости и доказана теорема, устанавливающая её основные свойства: выпуклость, компактность и непрерывную зависимость от момента окончания движения. Условием существования замкнутой модели управления служит в рассматриваемом случае принадлежность требуемого конечного состояния объекта к редуцированной области достижимости.

Рассмотрен также способ построения указанных областей и приведен конкретный пример решения поставленной задачи управления для двухфазного объекта. Поставлен вычислительный эксперимент и дано обсуждение его результатов.

С. Л. Гольдштейн, Е. Б. Солонин
Уральский политехнический институт,
кафедра вычислительной техники.
СССР, 620002 Свердловск К-2, Втузгородок,
главный учебный корпус

Problems of Control and Information Theory, Vol. 18(6), pp. 421–434 (1989)

LINEAR FILTERING OF SOLUTIONS OF STOCHASTIC INTEGRAL EQUATIONS IN NON-GAUSSIAN CASE

L. E. SHAIKHET, M. L. SHAFIR

(Donetsk)

(Received February 24, 1988)

The problem of constructing a linear optimal in quadratic mean sense estimation of non-gaussian process given by stochastic Volterra's equation is treated.

Introduction

The problem of constructing optimal in quadratic mean sense estimation for solution of Volterra's equation was considered in [1–3]. When the solution of equation is a Gaussian process the optimal estimation is presented [2,3] by an integral over the observed process. In this case the integrand is a determinate function and defined by Fredholm's integral equation. If the solution of the equation is not a Gaussian process then the estimation is not a linear function from the observed process [4]. In the present work, analogously to [2,3], the estimation in the class of linear estimations, optimal in quadratic mean sense for the non-gaussian process is constructed. Analogously to [5], the process of observation keeps a delay. The principle, different from [2,3] in the non-gaussian case is that it is necessary to use the expanding stochastic integral [10] and the multiple stochastic Ito's integrals [7] for constructing the resolvent formula for the solution of Volterra's equation.

1. Problem statement

Consider the problem of constructing optimal in quadratic mean sense estimation $\hat{x}(T)$ of the value of unobserved stochastic process $x(T)$, given by the stochastic integral equation

$$x(t) = \eta(t) + \int_0^t K_1(t,s)x(s)\,dw(s) + \int_0^t K_2(t,s),x(s)ds \qquad (1.1)$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

over the observation

$$dy(t) = A(t)x(t-h)dt + d\xi(t), \tag{1.2}$$

$$x(s) = 0, \quad s < 0.$$

Here $A(t)$, $K_1(t,s)$, $K_2(t,s)$, $0 \le s \le t \le T$ are piecewise continuous determinate functions, $h \ge 0$, $w(t)$ is a standard Wiener process, $\xi(t)$ is a process with independent increases,

$$Md\xi(t)dw(t) = N(t)dt, \quad M(d\xi(t))^2 = N_1(t)dt, \tag{1.3}$$

$\eta(t)$ is independent of $w(t)$ and $\xi(t)$ is a stochastic process for which $\sup\limits_{0\le t\le T} M\eta^2(t) < \infty$. All processes are defined on the same probability space $\{\Omega, f, P\}$ and they are measurable with respect to the family of σ-fields $f_t \subset f$.

It is known [4] that the solution of problem (1.1), (1.2) has a form

$$\hat{x}(T) = M(x(T)/f_T^y), \tag{1.4}$$

where $f_T^y = \sigma\{y(s), 0 \le s \le T\}$. If the joint distribution of processes $x(t)$ and $y(t)$ is Gaussian then estimate (1.4) concurres with the linear optimal estimate $m_0(T)$ having the form

$$m_0(T) = \int_0^T u_0(t)dy(t). \tag{1.5}$$

However, the solution of equation (1.1) is not a Gaussian process. Therefore, estimate (1.4) is a nonlinear functional from the trajectory $y(t)$, $0 \le t \le T$, and is not represented by (1.5).

Analogously with the Gaussian case [2,3], estimate (1.5) is optimal in quadratic mean sense in the class of linear estimates

$$m_u(T) = \int_0^T u(t)dy(t) \tag{1.6}$$

where u is an arbitrary function from the Hilbert space of square-integrable on the $[0,T]$ functions $L_2[0,T]$ will be constructed. There is a subspace $L_2\{m_u(T)\}$, $u \in L_2[0,T]$ formed by the class of estimates (1.6) in Hilbert space $L_2(\Omega)$. Therefore, the optimal estimate $m_0(T)$ in the class $L_2\{m_u(T)\}$ is [11] a projection $x(T)$ on $L_2\{m_u(T)\}$, which is defined simply from the correlation

$$M[(x(T) - m_0(T))m_u(T)] = 0. \tag{1.7}$$

Substituting (1.5), (1.6) in (1.7) we can obtain the equation for u_0 from the optimal linear estimate (1.5).

2. Stochastic Volterra's integral equations

Stochastic integral equations are treated in [8–10]. The solution of (1.1) is obtained by the method of consecutive approximations in [10] and is connected with the introduction of stochastic integral from the anticipatory functionals.

Lemma 1. Let $K_1(t,\tau)$, $K_2(t,\tau)$, $\eta(\tau)$, $0 \le \tau \le t \le T$. From equation (1.1) the following conditions are satisfied:

$$\sup_{0\le\tau\le t\le T} |K_1(t,\tau)| < \infty, \tag{2.1}$$

$$\sup_{0\le\tau\le t\le T} |K_2(t,\tau)| < \infty, \tag{2.2}$$

$$\sup_{0\le\tau\le T} |\eta(\tau)| < \infty, \quad P-\text{ a.s.}, \tag{2.3}$$

$$\sup_{0\le t\le T} M\eta^2(t) < \infty. \tag{2.4}$$

Define the iterations [10], $i = 1, 2$, $0 \le \tau \le t \le T$,

$$r_{0,i}(t,\tau) = K_i(t,\tau), \tag{2.5}$$

$$r_{n+1,i}(t,\tau) = \int_\tau^t K_1(t,s) r_{n,i}(s,\tau) dw(s) + \int_\tau^t K_2(t,s) r_{n,i}(s,\tau) ds. \tag{2.6}$$

Then the solution of (1.1) exists, uniquely with precision to stochastic equivalence and satisfies the condition

$$\sup_{0\le\tau\le T} Mx^2(t) < \infty \tag{2.7}$$

and is defined by

$$x(t) = \eta(t) + \int_0^t R_1(t,\tau)\eta(\tau) dw(\tau) + \int_0^t R_2(t,\tau)\eta(\tau) d\tau - \int_0^t R_1(t,\tau) K_1(\tau,\tau)\eta(\tau) d\tau, \tag{2.8}$$

$$R_i(t,\tau) = \sum_{n=0}^{\infty} r_{n,i}(t,\tau), \quad i = 1,2, \quad 0 \le \tau \le t \le T. \tag{2.9}$$

Series (2.9) converges in $L_2(\Omega)$ uniformly on $0 \le \tau \le t \le T$.

Proof. Existence and uniqueness of solution (1.1) with stochastic heterogeneous item under conditions (2.1)–(2.4) where shown just as in [10] for the equation with determinate heterogeneous item. Convergence of series (2.9) was shown in [10]. The representation (2.8) will be proved. It is worth noting that $R_1(t,\tau)$ and $R_2(t,\tau)$ is measureable with respect to $\sigma\{w(u), \tau \le u \le t\}$ and integral $\int_0^t R_1(t,\tau)\eta(\tau)dw(\tau)$ is understood as an expanding stochastic integral in accordance with the definition in [10]. Functions $R_i(t,\tau)$, $i = 1,2$, $0 \le \tau \le t \le T$, have an orthogonal expansion [7] as a functional from the Wiener process

$$R_i(t,\tau) = MR_i(t,\tau) + \sum_{k=1}^{\infty} \int\limits_{T_k(t,\tau)} \varphi_{k,i}(t,\tau;\tau_1,\dots,\tau_k)dw(\tau_1)\dots dw(\tau_k), \qquad (2.10)$$

where

$$T_k(t,\tau) = \{(\tau_1,\dots,\tau_k) : \tau \le \tau_1 \le \dots \le \tau_k \le t\},$$

$$\int\limits_{T_k(t,\tau)} \varphi_{k,i}(t,\tau;\tau_1,\dots,\tau_k)dw(\tau_1)\dots dw(\tau_k) =$$

$$= \int_0^t \Big(\int_0^{\tau_k} \Big(\dots \Big(\int_0^{\tau_2} \varphi_{k,i}(t,\tau;\tau_1,\dots,\tau_k)dw(\tau_1)\Big)\dots\Big)dw(\tau_{k-1})\Big)dw(\tau_k),$$

functions $\varphi_{k,i}(t,\tau;\cdot)$, $k = 1,2,\dots$, $i = 1,2$, is defined simply in $L_2(T_k(t,\tau))$. In [10] the resolvent formula (2.8) is obtained under the conditions of smoothness on the kernels and the free member of equation (1.1). Under this restrictions the integral $\int_0^t R_1(t,\tau)\ \eta(\tau)dw(\tau)$ exists and is defined by

$$\int_0^t R_1(t,\tau)\eta(\tau)dw(\tau) = \int_0^t \varphi_{1.1}(t,\tau;\tau)\eta(\tau)d\tau + \qquad (2.11)$$

$$+\sum_{k=0}^{\infty} \int\limits_{T_{k+1}(t,0)} \varphi_{k,1}(t,\tau_1;\tau_2,\dots,\tau_{k+1})\eta(\tau_1)dw(\tau_1)\dots dw(\tau_{k+1}) +$$

$$+\sum_{k=2}^{\infty} \int_0^t \int\limits_{T_{k-1}(t,\tau)} \varphi_{k,1}(t,\tau;\tau,\tau_1,\dots,\tau_{k-1})dw(\tau_1)\dots dw(\tau_{k-1})\eta(\tau)d\tau.$$

However, as contended in [10] this integral exists also for the process for which the following inequality is carried out

$$\sum_{k=0}^{\infty} \int\limits_{T_{k+1}(t,0)} \varphi_{k,1}^2(t,\tau_1;\tau_2,\ldots,\tau_{k+1})d\tau_1\ldots d\tau_{k+1} < \infty, \tag{2.12}$$

where $\varphi_{0,1}(t,\tau) = MR_1(t,\tau)$, and it converges in $L_2(\Omega)$

$$R_1(t,\tau) = \varphi_{1,1}(t,\tau;\tau) + \sum_{k=2}^{\infty} \int\limits_{T_{k-1}(t,\tau)} \varphi_{k,1}(t,\tau;\tau,\tau_1,\ldots,\tau_{k-1})dw(\tau_1)\ldots dw(\tau_{k-1}). \tag{2.13}$$

Then the conditions of smoothness on the kernels and the free member of equation (1.1) are not necessary. It will be shown that (2.12) is carried out and the process $R_1(t,\tau)$ exists under conditions of Lemma 1. For $R_1(t,\tau)$ from (2.9) analogously to (2.7), we obtain

$$\sup_{0\le\tau\le t\le T} MR_1^2(t,\tau) < \infty.$$

Then, allowing for continuity of dot product in $L_2(\Omega)$

$$\sup_{0\le\tau\le t\le T} M\left(\sum_{k=0}^{\infty} \int\limits_{T_k(t,\tau)} \varphi_{k,1}(t,\tau;\tau_1,\ldots,\tau_k)dw(\tau_1)\ldots dw(\tau_k)\right)^2 =$$

$$= \sup_{0\le\tau\le t\le T} \sum_{k=0}^{\infty} \int\limits_{T_k(t,\tau)} \varphi_{k,1}^2(t,\tau;\tau_1,\ldots,\tau_k)d\tau_1\ldots d\tau_k < \infty.$$

We can integrate the series on the right part of the equality over τ on $[0,t]$, and allowing for Lebesgue's theorem [12] to go to termwise integrating of the series and obtaining (2.12). Applying Fubini's theorem for iterated integrals of the type $dwdt$ in iterations (2.6), we obtain the representation for $r_{n,1}(t,\tau)$ by means of iterated stochastic integrals

$$r_{n,1}(t,\tau) = \varphi_{0,1}^{(n)}(t,\tau) + \sum_{k=1}^{n} \int\limits_{T_k(t,\tau)} \varphi_{k,1}^{(n)}(t,\tau;\tau_1,\ldots,\tau_k)dw(\tau_1)\ldots dw(\tau_k),$$

where $\varphi_{k,1}^{(n)}$, $k = 0,1,\ldots,n$, $n = 1,2,\ldots$, are defined simply by functions $K_1(t,\tau)$ and $K_2(t,\tau)$. Construct the iterations

$$\tilde{r}_{n,1}(t,\tau) = \sum_{k=1}^{n} \int\limits_{T_{k-1}(t,\tau)} \varphi_{k,1}^{(n)}(t,\tau;\tau,\tau_1,\ldots,\tau_{k-1})dw(\tau_1)\ldots dw(\tau_{k-1}),$$

$n = 1, 2, \ldots$, $0 \le \tau \le t \le T$. Analogously to (2.9), it is shown that the series $\sum_{n=1}^{\infty} r_{n,1}^{\sim}(t,\tau)$ converges in $L_2(\Omega)$ uniformly on $0 \le \tau \le t \le T$. From the uniqueness of the orthogonal expansion of Wiener's functions it follows that

$$R_1^{\sim}(t,\tau) = \sum_{n=1}^{\infty} r_{n,1}^{\sim}(t,\tau), \quad 0 \le \tau \le t \le T \tag{2.14}$$

and $R_1^{\sim}(t,\tau)$ exists. In accordance with (5.6) from [10], the consecutive approximations of solution (1.1) takes the form

$$x_n(t) = \eta(t) + \int_0^t \sum_{k=1}^{n} r_{k,1}(t,\tau)\eta(\tau)dw(\tau) +$$

$$+ \int_0^t \left(\sum_{k=1}^{n} r_{k,2}(t,\tau) - K_1(\tau,\tau) \sum_{k=1}^{n-1} r_{k,1}(t,\tau) \right) \eta(\tau) \Bigg) d\tau, \quad 0 \le t \le T.$$

Estimate $M(x(t) - x_n(t))^2$, $0 \le t \le T$, where $x(t)$ is defined by (2.8)

$$M(x(t) - x_n(t))^2 \le 3 \left\{ M \left(\int_0^t \left(R_1(t,\tau) - \sum_{k=1}^{n} r_{k,1}(t,\tau) \right) \eta(\tau)dw(\tau) \right)^2 + \right.$$

$$+ M \left(\int_0^t \left(R_2(t,\tau) - \sum_{k=1}^{n} r_{k,2}(t,\tau) \right) \eta(\tau)d\tau \right)^2 +$$

$$\left. + M \left(\int_0^t K_1(\tau,\tau) \left(R_1(t,\tau) - \sum_{k=1}^{n-1} r_{k,1}(t,\tau) \right) \eta(\tau)d\tau \right)^2 \right\}.$$

From the definition of the expanding stochastic integral and correlations (2.4), (2.9) it follows that the items on the right part converge to zero if $n \to \infty$. Thus, sequence $x_n(t)$ converges in $L_2(\Omega)$ to the process $x(t)$, $0 \le t \le T$, defined by (2.8). From the unequality of solution (1.1) it follows that (2.8) defines the solution of (1.1). Thus, the Lemma is proved.

It is necessary to find integrands $\varphi_{1,1}(t,\tau;\tau_1)$, $\varphi_{2.1}(t,\tau;\tau_1,\tau_2)$, $\varphi_{1.2}(t,\tau;\tau_1)$ and $MR_1(t,\tau)$, $MR_2(t,\tau)$ from expansions (2.10) for the solution of problem (1.1)–(1.2).

Lemma 2. Let $R_1(t,\tau)$, $R_2(t,\tau)$, $0 \le \tau \le t \le T$, the resolvent kernels of equation (1.1). Then $MR_i(t,\tau)$, $i = 1,2$, satisfy Volterra's integral equation of second kind

$$MR_i(t,\tau) = K_i(t,\tau) + \int_\tau^t K_2(t,s) MR_i(s,\tau) ds, \quad 0 \le \tau \le t \le T, \tag{2.15}$$

the integrands $\varphi_{n,i}(t,\tau;\tau_1,\dots,\tau_n)$, $i = 1,2$, $n = 1,2,\dots$, from the expansions (2.10) satisfy Volterra's equations of second kind (for fixed $0 \le \tau \le \tau_1 \le \dots \le \tau_n$), $\tau_n \le t \le T$,

$$\varphi_{1,i}(t,\tau;\tau_1) = K_1(t,\tau_1) MR_i(\tau_1,\tau) + \int_{\tau_1}^t K_2(t,s)\varphi_{1,i}(s,\tau;\tau_1) ds, \tag{2.16}$$

$$\varphi_{n,i}(t,\tau;\tau_1,\dots,\tau_n) = K_1(t,\tau_n)\varphi_{n-1,i}(\tau_n,\tau;\tau_1,\dots,\tau_{n-1}) + \tag{2.17}$$

$$+ \int_{\tau_n}^t K_2(t,s)\varphi_{n,i}(s,\tau;\tau_1,\dots,\tau_n) ds.$$

Proof. In accordance with (2.5), (2.6), (2.9) the resolvent kernels as limits of consequtive approximations satisfy the integral equations (it is proved analogously to Theorem 5.A of [10])

$$R_i(t,\tau) = K_i(t,\tau) + \int_\tau^t K_1(t,s) R_i(s,\tau) dw(s) + \tag{2.18}$$

$$+ \int_\tau^t K_2(t,s) R_i(s,\tau) ds, \quad i = 1,2, \quad 0 \le \tau \le t \le T.$$

Since functions $R_i(t,\tau)$, $i = 1,2$, have unique presentation (2.10) and series (2.6) converges uniformly for $0 \le \tau \le t \le T$ then

$$\lim_{h\to\infty} \sup_{0\le\tau\le t\le T} M\Big(R_i(t,\tau) - \sum_{k=0}^{n} \int_{T_k(t,\tau)} \varphi_{k,i}(t,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k)\Big)^2 = 0, \tag{2.19}$$

and using the properties of Ito's integral we obtain

$$\int_\tau^t K_1(t,s) R_i(s,\tau) dw(s) =$$

$$= \sum_{k=0}^{\infty} \int_{\tau}^{t} K_1(t,s) \int_{T_k(s,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) dw(s),$$

series converges in $L_2(\Omega)$ on the right part. From (2.19), for $i = 1, 2$

$$\lim_{n\to\infty} M \left(\int_{\tau}^{t} K_2(t,s) \left[R_i(s,\tau) - \right.\right.$$

$$\left.\left. - \sum_{k=0}^{n} \int_{T_k(s,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) \right] ds \right)^2 \le$$

$$\le (t-\tau) \lim_{n\to\infty} \int_{\tau}^{t} K_2^2(t,s) M \left[R_i(s,\tau) - \right. \tag{2.20}$$

$$\left. - \sum_{k=0}^{n} \int_{T_k(s,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) \right]^2 ds = 0.$$

Then (2.18) takes the form, $i = 1, 2$, $0 \le \tau \le t \le T$,

$$\sum_{k=0}^{\infty} \int_{T_k(t,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) = K_i(t,\tau) + \tag{2.21}$$

$$+ \sum_{k=0}^{\infty} \int_{\tau}^{t} K_1(t,s) \int_{T_k(s,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) dw(s) +$$

$$+ \sum_{k=0}^{\infty} \int_{\tau}^{t} K_2(t,s) \int_{T_k(s,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) ds.$$

Applying Fubini's theorem for iterated integrals of the type $dwdt$ we obtain

$$\int_{\tau}^{t} K_2(t,s) \int_{T_k(s,\tau)} \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) dw(\tau_1)\dots dw(\tau_k) ds =$$

$$= \int_{T_k(t,\tau)} \int_{\tau_k}^{t} K_2(t,s) \varphi_{k,i}(s,\tau;\tau_1,\dots,\tau_k) ds dw(\tau_1)\dots dw(\tau_k).$$

From (2.21) and the uniqueness of functions $\varphi_{k,i}$, $k = 0, 1, 2, \ldots$, $i = 1, 2$, in (2.10), we obtain equations (2.15) and

$$\varphi_{k,i}(t, \tau; \tau_1, \ldots, \tau_k) = K_1(t, \tau_k)\varphi_{k-1,i}(\tau_k, \tau; \tau_1, \ldots, \tau_{k-1}) +$$

$$\int_{\tau_k}^{t} K_2(t, s)\varphi_{k,i}(s, \tau; \tau_1, \ldots, \tau_k)ds, \quad 0 \le \tau \le \tau_1 \le \ldots \le \tau_k \le t,$$

where $\varphi_{0,i}(s, \tau) = MR_i(s, \tau)$. Thus, Lemma 2 is proved.

3. The constructing of linear optimal estimation

Theorem. Under conditions (2.1)–(2.4) the optimal estimation in the form $m_u(T) = \int_0^T u(\tau)dy(\tau)$, $u \in L_2[0, T]$, of the value of process $x(T)$ given by equation (1.1) over the observation of process (1.2) is defined by the function $u_0 \in L_2[0, T]$ satisfying Fredholm's integral equation

$$u_0(s)N_1(s) + A(s)\int_h^T R(s - h, \tau - h)u_0(\tau)A(\tau)d\tau + \tag{3.1}$$

$$+N(s)\int_{s+h}^{T} \Phi(\tau - h, s)A(\tau)u_0(\tau)d\tau + A(s)\int_0^{s-h} \Phi(s - h, \tau)N(\tau)u_0(\tau)d\tau =$$

$$= N(s)\Phi(T, s) + A(s)R(T, s - h), \quad s \in [h, T],$$

where $R(t, \tau) = Mx(t)x(\tau)$, $(t, \tau) \in [0, T] \times [0, T]$, is a correlation function of process x, $R(t, \tau) = 0$, $\tau < 0$, function $\Phi(t, \tau)$ is defined by the equality

$$\Phi(t, \tau) = MR_1(t, \tau)M\eta(\tau) + \int_0^{\tau} \varphi_{2,1}(t, s; s, \tau)M\eta(s)ds +$$

$$+\int_0^{\tau} \varphi_{1,2}(t, s; \tau)M\eta(s)ds - \int_0^{\tau} K_1(s, s)\varphi_{1,1}(t, s; \tau)M\eta(s)ds,$$

$$0 \le \tau \le t \le T,$$

functions $MR_1(t, \tau)$, $\varphi_{1,1}(t, \tau; \tau_1)$, $\varphi_{2,1}(t, \tau; \tau_1, \tau_2), \varphi_{1,2}(t, \tau; \tau_1)$ are defined in Lemma 2.

Proof. Consider condition (1.7)

$$Mx(T)m_u(T) = \int_0^T A(\tau)u(\tau)R(T,\tau-h)d\tau + Mx(T)\int_0^T u(\tau)d\xi(\tau).$$

Allowing for (2.8), (2.11) and continuity of dot scalar in $L_2(\Omega)$ we obtain

$$M\left(\int_0^T R_1(T,s)\eta(s)dw(s)\cdot\int_0^T u(\tau)d\xi(\tau)\right) =$$

$$= \sum_{k=0}^{\infty} M\left(\int_0^T \int_{T_k(T,s)} \varphi_{k,1}(T,s;\tau_1,\ldots,\tau_k)dw(\tau_1)\ldots dw(\tau_k)\eta(s)dw(s)\cdot\int_0^T u(s)d\xi(s)\right).$$

Using definition (2.11) and properties of Ito's integral we obtain

$$M\left(\int_0^T MR_1(T,s)\eta(s)dw(s)\cdot\int_0^T u(\tau)d\xi(\tau)\right) = \int_0^T MR_1(T,s)M\eta(s)N(s)u(s)ds,$$

$$M\left(\int_0^T\int_s^T \varphi_{1,1}(T,s;\tau_1)dw(\tau_1)\eta(s)dw(s)\cdot\int_0^T u(\tau)d\xi(\tau)\right) =$$

$$= M\left[\left(\int_0^T\int_0^{\tau_1}\varphi_{1,1}(T,s;\tau_1)\eta(s)dw(s)dw(\tau_1)+\int_0^T\varphi_{1,1}(T,\tau;\tau)\eta(\tau)d\tau\right)\times\right.$$

$$\left.\times\int_0^T u(\tau)d\xi(\tau)\right] = 0,$$

$$M\left(\int_0^T\int_{T_2(T,s)}\varphi_{2,1}(T,s;\tau_1,\tau_2)dw(\tau_1)dw(\tau_2)\eta(s)dw(s)\cdot\int_0^T u(\tau)d\xi(\tau)\right) =$$

$$= \int_0^T u(\tau)N(\tau)\int_0^\tau \varphi_{2,1}(T,s;s,\tau)M\eta(s)dsd\tau,$$

$$M\left(\int_0^T \int_{T_k(T,s)} \varphi_{k,1}(T,s;\tau_1,\dots,\tau_k)dw(\tau_1)\dots dw(\tau_k)\eta(s)dw(s)\cdot\int_0^T u(\tau)\ d\xi(\tau)\right)=0,$$

$$k=3,4,\dots$$

Analogously to (2.20) we can go to termwise integrating of series (2.10) and obtain, for $i=1,2$,

$$M(\int_0^T R_i(T,s)\eta(s)ds\cdot\int_0^T u(\tau)d\xi(\tau))=$$

$$=\sum_{k=0}^{\infty} M\left(\int_0^T \int_{T_k(T,s)} \varphi_{k,i}(T,s;\tau_1,\dots,\tau_k)dw(\tau_1)\dots dw(\tau_k)\eta(s)ds\cdot\int_0^T u(\tau)d\xi(\tau)\right).$$

Using Fubini's theorem for iterated integrals of type $dwdt$ we obtain, $i=1,2$,

$$M\left(\int_0^T MR_i(T,s)\eta(s)ds\cdot\int_0^T u(\tau)d\xi(\tau)\right)=0,$$

$$M\left(\int_0^T\int_s^T \varphi_{1,i}(T,s;\tau_1)dw(\tau_1)\eta(s)ds\cdot\int_0^T u(\tau)d\xi(\tau)\right)=$$

$$=\int_0^T u(\tau)N(\tau)\int_0^\tau \varphi_{1,i}(T,s;\tau)M\eta(s)ds d\tau,$$

$$M\left(\int_0^T \int_{T_k(T,s)} \varphi_{k,i}(T,s;\tau_1,\dots,\tau_k)dw(\tau_1)\dots dw(\tau_k)\eta(s)ds\cdot\int_0^T u(\tau)d\xi\ (\tau)\right)=0,$$

$$k=2,3,\dots$$

Thus

$$Mx(T)\int_0^T u(\tau)d\xi(\tau)=\int_0^T \Phi(T,\tau)N(\tau)u(\tau)d\tau.$$

From the independent increases of process ξ it follows that

$$Mx(t)\int_0^T u(s)d\xi(s)=\int_0^t \Phi(t,\tau)N(\tau)u(\tau)d\tau. \tag{3.2}$$

Allowing (3.2) and that $\Phi(t,\tau)=0$, $\tau>t$,

$$M m_0(T)m_u(T)=\int_0^T u(s)A(s)\int_0^T R(s-h,\tau-h)u_0(\tau)A(\tau)d\tau ds+$$

$$+\int_0^T u(s)A(s)\int_0^{s-h}\Phi(s-h,\tau)u_0(\tau)N(\tau)d\tau ds+$$

$$+\int_0^T u(s)N(s)\int_{s+h}^T \Phi(\tau-h,s)u_0(\tau)A(\tau)d\tau ds+\int_0^T u(s)u_0(s)N_1(s)ds.$$

Finally,

$$M\left[(x(T)-m_0(T))\,m_u(T)\right]=\int_0^T u(s)\left[A(s)R(T,s-h)+\Phi(T,s)N(s)-\right.$$

$$-A(s)\int_h^T R(s-h,\tau-h)u_0(\tau)A(\tau)d\tau-N(s)\int_{s+h}^T \Phi(\tau-h,s)A(\tau)u_0(\tau)d\tau-$$

$$\left.-A(s)\int_0^{s-h}\Phi(s-h,\tau)N(\tau)u_0(\tau)d\tau-u_0(s)N_1(s)\right]ds.$$

From here, since function u is arbitrary, we obtain equation (3.1) for function u_0. Thus, the theorem is proved.

Remark. If $N_1(t)\neq 0$, $t\in[0,T]$, then equation (3.1) is a Fredholm's equation of second kind and it has a unique solution. If $N_1(t)=0$ for some t from $[0,T]$ then equation (3.1) is Fredholm's equation of third kind. If $N_1(t)\equiv 0$ then (3.1) is a Fredholm's equation of first kind. Methods of numerical solution of Fredholm's equations of all kinds is treated in [13].

4. Conclusion

Consider the problem which has a concrete physical sense, the solution of which reduces to the necessity of solving stochastic Volterra's equations and to estimate the solutions of such equations.

The small free swings of the pendulum is described by the equation

$$\ddot{x}(t)+a^2x(t)=0,\quad x(0)=x_0,\quad \dot{x}(0)=0,$$

where $x(t)$ is the corner of deflection of the pendulum from the vertical line in moment t. Suppose that the coefficient a^2 is subjected to additive indignation by the white noise. Then the equation of the motion takes the form

$$\ddot{x}(t) + (a^2 + b_1\dot{w}_1(t))x(t) = 0, \quad x(0) = x_0, \quad \dot{x}(0) = 0. \tag{4.1}$$

Observations $y(t)$ over the corner of deflections of the pendulum $x(t)$ also are subjected to an additive indignation by white noise and realized by the law

$$\dot{y}(t) = x(t) + b_2\dot{w}_2(t). \tag{4.2}$$

It is required to construct the estimation of the value of the corner of deflection $x(t)$ in a moment T over the observations $y(t)$, $t \leq T$. Integrating (4.1), (4.2) we receive

$$x(t) = x_0 - a^2 \int_0^t (t-s)x(s)ds - b_1 \int_0^t (t-s)x(s)dw_1(s),$$

$$dy(t) = x(t)dt + b_2 dw_2(t),$$

i.e. the problem of optimal filtering in the form (1.1), (1.2).

References

1. *Kleptsina, M.L., Veretennikov, A.Yu.*, On filtering and properties of conditional laws of Ito–Volterra processes. Statistics and control of stochastic processes. Steklov Seminar, 1984. Optimization Software, Inc., Publication Division, N.Y., 1985, pp. 179–196.
2. *Kolmanovsky, V.B., Shaikhet, L.E.*, The filtering of integral equations. In: Mathematical theory of system (Bobilev, N.A., Boltyansky, V.G. et al.). Moscow, Nauka, 1986, pp. 55–56 (in Russian).
3. *Kolmanovsky, V.B., Shaikhet, L.E.*, About estimation of solutions of linear integral equations. PMM, 1987, **51**, *5*, pp. 775–781 (in Russian).
4. *Liptser, R.S., Shiryayev, A.N.*, Statistics of random processes. Moscow, Nauka, 1974, 696 pp. (in Russian).
5. *Chernousko, F.L., Kolmanovsky, V.B.*, Optimal control for stochastic perturbations. Moscow, Nauka, 1978, 352 pp. (in Russian).
6. *Skorokhod, A.V.*, About the generalization of stochastic integral. Theory of probability and its applications, 1975, **20**, *2*, pp. 223–237 (in Russian).
7. *Ito, K.*, Multiple Wiener integral. Journ. Math. Soc. Japan, 1951, **3**, pp. 157–169.
8. *Shevlyakov, A.Yu.*, About one class of stochastic integral equations. Theory of stochastic processes. Kiev, Naukova dumka, 1979, pp. 118–127 (in Russian).
9. *Ito, K.*, On the existence and uniqueness of solutions of stochastic integral equations of the Volterra Type. Kodai Math. J., 1979, **2**, pp. 158–170.
10. *Birger, M.C., Mizel, V.J.*, Theorems of Fubini type for iterated stochastic integrals. Transactions of the American Mathematical Society, 1979, **252**, pp. 249–274.

11. *Gikhman, I.I., Skorokhod, A.V.*, Theory of random processes. Vol.**1**, Moscow, Nauka, 1971, 664 pp. (in Russian).
12. *Kolmogorov, A.N., Fomin, S.V.*, Elements of theory of functions and functional analysis, 4th ed. Moscow, Nauka, 1976, 543 pp. (in Russian).
13. *Verlan, A.F., Sizikov, V.S.*, Integral equations: methods, algorithms, programs. Reference appliance. Kiev, Naukova dumka, 1986, 544 pp. (in Russian).

Линейная фильтрация решений стохастических интегральных уравнений в негауссовском случае

Л. Е. ШАЙХЕТ, М. Л. ШАФИР

(Донецк)

Построена линейная оптимальная в среднеквадратическом смысле оценка

$$m(T) = \int_0^T u(s)dy(s)$$

значения $x(T)$ негауссовского процесса, заданного стохастическим интегральным уравнением Вольтерра

$$x(t) = \eta(t) + \int_0^t K_1(t,s)x(s)dw(s) + \int_0^t K_2(t,s)x(s)ds,$$

по наблюдениям

$$dy(t) = A(t)x(t-h)dt + d\xi(t).$$

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

Problems of Control and Information Theory, Vol. 18(6), pp. 435-444 (1989)

AN ALGORITHM TO FIND THE GLOBAL OPTIMUM OF LEFT-TO-RIGHT HIDDEN MARKOV MODEL PARAMETERS

A. FARAGÓ, G. LUGOSI

(*Budapest*)

(Received January 5, 1989)

A central problem concerning the so-called Hidden Markov Models (HMM) is the following: given a finite observation sequence, find a HMM (with a restricted number of states), which chould produce the observation sequence with the highest possible probability. However, the existing algorithms, such as the Baum–Welch (forward–backward) re-estimation procedure, converge only to a local optimum at an uncertain speed of convergence. We show that with a slight and practically justifiable modification of the objective function the situation becomes much better: there exists a fast non-iterative algorithm to find the global optimum for the class of left-to-right models which play central role in some applications, such as speech recognition.

1. Introduction

Hidden Markov Models (HMM) play an increasing role in a number of applications, such as automatic speech recognition, since they proved to be a useful tool for modelling the time varying properties of certain signals, e.g. human speech.

Concerning the application of Markov models it is worth citing Rabiner and Juang [1]: "The basic theory of Markov chains has been known to mathematicians and engineers for close to 80 years, but it is only in the past decade that is has been applied explicitly to problems in speech processing. One of the major reasons why speech models, based on Markov chains, have not been developed until recently was the lack of a method for optimizing the parameters of the Markov model to match observed signal parameters."

Though such methods are available now, the optimization problem still does not have a satisfactory solution. The reason is that the known algorithms do not guarantee a global optimum, and their application is computationally expensive, being iterative methods withoud good practical bounds for the number of iterations.

In section 3 we present a fast non-iterative algorithm, which finds the global optimum of left-to-right HMM's with a slightly modified objective function. Left-to-

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

right HMM's form the most important subclass of all HMM's for speech processing applications.

The reader is assumed to be familiar with the basic concepts and problems of HMM's, but for the sake of self-containment we give all the necessary definitions. For a tutorial introduction to HMM's see [1]; for a more subtle treatment see e.g. [2], [3], [4], [6], [7].

2. Preliminaries

To describe (discrete observation) HMM's we use the following notations:

T = length of observation sequence,
N = number of states,
M = number of possible observation symbols,
$Q = \{q_1, \dots, q_N\}$ set of states,
$V = \{v_1, \dots, v_M\}$ set of observation symbols,
x_t = observation at time t, a realization of the random variable X_t taking values from V,
$x = (x_1, x_2, \dots, x_T)$ observation sequence, a realization of $X = (X_1, \dots, X_T)$,
s_t = state at time t, a realization of the r.v. S_t taking values from Q,
$s = (s_1, s_2, \dots, s_T)$ state sequence,
$A = \{a_{ij}\}$ $(i, j \in \{1, \dots, N\})$ state transition probabilities, that is, $a_{ij} = P(S_{t+1} = q_j \mid S_t = q_i)$,
$B = \{b_j(k)\}$ $(j \in \{1, \dots, N\},\ k \in \{1, \dots, M\})$ observation symbol probabilities, that is, $b_j(k) = P(X_t = v_k \mid S_t = q_j)$,
$\Pi = \{\pi_i\}$ $(i \in \{1, \dots, N\})$ initial state distribution, that is, $\pi_i = P(S_1 = q_i)$.

A HMM is given by the triplet $M = (A, B, \Pi)$. It generates an observation sequence as follows:

1. Initialization: choose an initial state s_1, according to the initial state distribution Π.

2. General step at time t: choose an observation x_t, according to the distribution $b_j(k)$, if $s_t = q_j$. Choose the next state s_{t+1}, according to the state transition probabilities.

The probability of the observation sequence x in the model M is denoted by $P_M(x)$.

The state structure of an HMM can be visualized by a state transition diagram (graph) in the following way: assign a node to each state and draw a directed arc from the ith node to the jth node iff $a_{ij} > 0$. The state transition graph is useful to represent structural restrictions imposed on the HMM, which means that some state transitions are not allowed, that is, they must have zero probability.

Two distinguished types of structural restrictions are the left-to-right and the linear models. An HMM is called left-to-right iff its state transition graph con-

tains no directed cycle (except loops, which are assumed to exist at each node). Equivalently, if a state has been left once by the system, it cannot be visited again. Another equivalent definition is, that the state transition matrix can be made triangular (with an appropriate permutation of state labeling).

Linear models are special cases of left-to-right HMM's. In a linear model all the states can be strung on a path and no other arc is allowed except loops. Equivalently, in a linear HMM, with an appropriate numbering of states, $a_{i,i+1} > 0$ holds for $i = 1, \ldots, N-1$, but $a_{ij} = 0$ holds if $j \neq i$ or $j \neq i+1$. The initial state in a linear model is always meant to be the "leftmost" state, that is, the initial state distribution Π is degenerated here, being concentrated on just a single state.

It is worth mentioning that left-to-right models play a central role in speech recognition applications, which served as a basic motivation for our research.

3. Global optimization of left-to-right models

3.1. Change of the objective function

The estimation problem is the following: Given a finite observation sequence: $x = (x_1, x_2, \ldots, x_T)$, estimate the state transition and the emission distributions of an HMM of a given structure, namely a left-to-right model. The usual way is to try to find the maximum likelihood estimator, that is, to maximize the objective function $P_M(x)$. However, the existing algorithms, such as the Baum–Welch (forward–backward) reestimation procedure, converge only to a local optimum at an uncertain speed of convergence.

Typically, in speech recognition each word (or another unit of speech) is represented by one (or more) optimized model and in the recognition phase we choose the model M_i, for which $P_{M_i}(x)$ is maximum for the input observation sequence to be recognized. The tool of this is the so-called forward–backward algorithm [3]. However, literature reports that the Viterbi algorithm works equally well for recognition (see [1]). That is, while $P_M(x) = \sum_s P_M(x, s)$, where the summation is taken over all possible state sequences, one can use instead only the most significant term $P_M(x, s^*)$, which is provided by the Viterbi algorithm. Here s^* is a state sequence maximizing $P_M(x, s)$. Such a sequence is called a *characteristic state sequence* (with respect to x and M). In other words,

$$s^* = \arg\max_s P_M(x, s).$$

(if the maximizing sequence is not unique, we can take any of the characteristic state sequences).

Now one can raise the following idea: if the objective function $P_M(x, s^*)$ works properly in the recognition phase, why not use it in the model optimization phase

as well? Jelinek [2] ("Viterbi extraction") and Rabiner, Wilpon and Juang [8] ("Segmental k-Means") presented similar methods, which provide convergence to a local optimum of this objective function, called "state-optimized joint likelihood".

We succeeded in finding a fast algorithm globally maximizing the new objective function as we shall see in section 3.2.

3.2. Algorithm for linear models with N states

We consider first the simplest case when the HMM is restricted to be linear with exactly N states. This is the core of our method, since the general left-to-right problem can easily be reduced to this case, as we shall see in section 3.3.

Suppose the states are numbered in the natural order corresponding to the linear model (from left to right). The initial state is always assumed to be q_1, while the terminal state is q_N.

Here we do not need two indices fro state transition probabilities: the a_{ij}'s are replaced by p_i's, where $p_i = P(S_{t+1} = q_{i+1} \mid S_t = q_i)$ $(i = 1, \ldots, N-1)$ and $p_N = 0$. Thus, the model is determined by the p_i's and the $b_i(k)$'s, the observation symbol probabilities. These quantities are to be chosen optimally, to maximize $P_M(x, s^*)$ for a given observation sequence $x = (x_1, x_2, \ldots, x_T)$. It is assumed that $T > N$, otherwise the problem becomes trivial. Further, let us mention for the sake of clarity, that s^* is not a fixed state sequence, it depends on x and M in the expression $P_M(x, s^*)$.

To describe the algorithm, we introduce some notations. For $i < j$ $x^{(i,j)}$ denotes the subsequence $(x_i, x_{i+1}, \ldots, x_j)$ of x. We write $v_k \in x^{(i,j)}$ if symbol v_k occurs in $x^{(i,j)}$. Denote by $f(k, i, j)$ the relative frequency of occurrences of v_k in $x^{(ij)}$, that is, the number of occurrences of v_k in $x^{(ij)}$ divided by the length of $x^{(ij)}$, which is $j - i + 1$. Finally, let $F(i, j)$ be defined by

$$F(i,j) = \prod_{v_k \in x^{(i,j)}} f(k,i,j)$$

where the product is taken only over those v_k's, which occur in $x^{(ij)}$, so the factors of zero-value are omitted.

Now we are prepared to describe the algorithm:

Algorithm 1:

Step 1. Construct a trellis with $N-1$ columns and with T nodes in each column. Denote by v_{ij} the jth node in the ith column $(i = 1, \ldots, N-1;$ $j = 1, \ldots, T)$. Add two additional nodes v_{00} and v_{NT} to the trellis, so $i = j = 0$ and $i = N$, $j = T$ will be allowed in the algorithm, as well. Draw an arc from v_{ij} to $v_{i+1,k}$ iff $i \le j \le T - N + i$, $j + 1 \le k \le T - N + i$ and $0 \le i \le N - 1$ hold.

To the arcs assign the following weights:

$$w(v_{ij}, v_{i+1,k}) = c(j,k) + \log F(j+1,k)$$

where

$$c(j,k) = \begin{cases} \log\left[\left(1 - \frac{1}{k-j+1}\right)^{k-j+1} \cdot \frac{1}{k-j+1}\right] & \text{if } i \le N-2 \\ 0 & \text{if } i = N-1. \end{cases}$$

Step 2. Run the Viterbi algorithm on the trellis constructed in Step 1 to find a maximum-weight path from v_{00} to v_{NT}. The Viterbi algorithm is given here by the recurrence

$$\phi(v_{00}) = 0$$
$$\phi(v_{i,k}) = \max_j(\phi(v_{i-1,j}) + w(v_{i-1,j}, v_{ik})), \ (i = 1, \ldots, N).$$

The maximum is taken over all j's, for which the arc $(v_{i-1,j}, v_{i,k})$ exists. We also have to keep record the nodes of the maximizing path, denote it by

$$v_{0,l_0}, v_{1,l_1}, v_{2,l_2}, \ldots, v_{i,l_i}, \ldots, v_{N,l_N} \quad (l_0,\ l_n = T).$$

Step 3. Define the model parameters by

$$p_i = \frac{1}{l_i - l_{i-1} + 1} \quad (i = 1, \ldots, N-1)$$
$$p_n = 0$$
$$b_i(k) = f(k, l_{i-1} + 1, l_i) \quad (i = 1, \ldots, N)$$

(where the l_i's have been obtained in Step 2).

The correctness of the algorithm is guaranteed by the following theorem:

Theorem 1: For any given observation sequence x, the model parameters obtained by Algorithm 1 globally maximize the objective function $P_M(x, s^*)$.

To prove Theorem 1, we need a lemma, which is widely used in information theory:

Lemma 1: Let $c_i > 0$, $y_i > 0$ be real numbers $(i = 1, \ldots, n)$ with $\sum_{i=1}^{n} y_i = 1$. Then the function $f(y_1, \ldots, y_n) = \sum_{i=1}^{n} c_i \log y_i$ attains its unique global maximum iff

$$y_i = \frac{c_i}{\sum_{j=1}^{n} c_j} \qquad (i = 1, \ldots, N).$$

The proof of Lemma 1 is based on Jensen's inequality and can be found in any comprehensive textbook on information theory (see e.g. [5]).

Proof of Theorem 1: Let s be any state sequence, starting at q_1 and terminating at q_N. (It is enough to consider only the state sequences terminating at the final state q_N, because decreasing the number of states the optimal value of the objective function cannot increase. We shall see the proof of this statement later, in Lemma 2.) By the linearity of the model, s is uniquely determined by the natural numbers $k_1, \ldots, k_i, \ldots, k_N$, where k_i means the multiplicity of q_i in s, that is, the number of time-units spent in state q_i.

The probability of spending k_i time-units in q_i and then proceeding to q_{i+1} is

$$(1-p)^{k_i} p_i \quad (i = 1, \ldots, N-1). \tag{1}$$

Denote by l_i the total time spent in the first i states. Then

$$l_i = \sum_{j=1}^{i} k_i \quad \text{and} \quad k_i = l_i - l_{i-1} \quad (i = 1, \ldots, N;\ l_0 = 0).$$

The system emits the observation subsequence $x^{(l_{i-1}+1, l_i)} = x_{l_{i-1}+1}, \ldots, x_{l_i}$ while being in state q_i. The probability of this event is

$$b_i(x_{l_{i-1}+1}) \cdot \ldots \cdot b_i(x_{l_i}) \tag{2}$$

where $b_i(x_t)$ stands for $b_i(k)$ if $x_t = v_k$.

Thus, $P_M(x, s)$ can be expressed by composing $N-1$ factors of type (1) and N factors of type (2) (taking into consideration that one needs no factor of type (1) for q_N):

$$P_M(x, s) = \left(\prod_{i=1}^{N-1} (1-p_i)^{k_i} p_i \right) \cdot \left(\prod_{i=1}^{N} b_i(x_{l_{i-1}+1}) \cdot \ldots \cdot b_i(x_{l_i}) \right). \tag{3}$$

Taking the logarithm of (3) we get

$$\begin{aligned} \log P_M(x, s) = & \sum_{i=1}^{N-1} (k_i \log(1-p_i) + \log p_i) + \\ & + \sum_{i=1}^{N} (\log b_i(x_{l_{i-1}+1}) + \ldots + \log b_i(x_{l_i})). \end{aligned} \tag{4}$$

Expression (4) contains $2N-1$ main summands (written in brackets), and any model parameter affects just one of them. So, for fixed $k_1, \ldots, k_n$, the summands can be independently maximized by Lemma 1, substituting

$$\begin{aligned} p_i &= \frac{1}{k_i + 1} \quad (i = 1, \ldots, N-1) \\ b_i(k) &= f(k, l_{i-1}+1, l_i). \end{aligned} \tag{5}$$

Resubstituting (5) into (4) (with $k_i = l_i - l_{i-1}$) we obtain

$$\log P_M(x,s) = \sum_{i=1}^{N-1} \log \left(\left(1 - \frac{1}{l_i - l_{i-1}+1}\right)^{\frac{1}{l_i - l_{i-1}}} \cdot \frac{1}{l_i - l_{i-1}+1} \right) + \\ + \sum_{i=1}^{N} \log F(l_{i-1}+1, l_i). \tag{6}$$

So we have to maximize (6) with respect to the l_i's, to obtain a global optimum. But, as the reader can immediately check, exactly this objective function is maximized by the Viterbi algorithm in Algorithm 1. Thus, the proof is completed.

3.3. The general left-to-right case

The optimization of general left-to-right models with arbitrary restriction on the state transition graph can easily be reduced to the linear case, solved in Section 3.2.

We introduce some notations. If M is an HMM, denote by $G(M)$ the state transition graph of M (defined in Section 2). Let $\mathcal{B}$ be the set of the state transition graphs of all left-to-right models. A restriction is expressed by a given subset $\mathcal{B}^* \subseteq \mathcal{B}$ and we intend to maximize $P_M(x, s^*)$ over all models M for which $G(M) \in \mathcal{B}^*$. For example $\mathcal{B}^*$ can contain all left-to-right state diagrams with at most N states. But $\mathcal{B}^*$ can express more sophisticated restrictions, as well. Such a $\mathcal{B}^*$ could be, for example, a set of state diagrams with at most N states, which contain a directed Hamiltonian path (a path traversing all nodes), etc.

In what follows, we assume that $\mathcal{B}^*$ is closed under deletion of arcs, that is, if an arc is not present in the state diagram, then the corresponding model parameter must be zero, but the presence of the arc does not force the parameter to be strictly positive. On the other hand, loops are always assumed to be present at each node (which also allows the corresponding parameter to be zero).

For any $\mathcal{B}^* \subseteq \mathcal{B}$, let $L(\mathcal{B}^*)$ be the longest linear state transition graph, which is a subgraph of some $G \in \mathcal{B}^*$. Clearly, $L(\mathcal{B}^*)$ is unique upto isomorphism. (Isomorphic graphs are considered to be identical.)

The following lemma shows that in general left-to-right HMM optimization it is enough to restrict ourselves to longest linear models:

Lemma 2: For any given observation sequence x and state transition graph restriction $\mathcal{B}^* \subseteq \mathcal{B}$

$$\max_{G(M) \in \mathcal{B}^*} P_M(x, s^*) = \max_{G(M) = L(\mathcal{B}^*)} P_M(x, s^*) \tag{7}$$

holds.

Proof. Let M be a model, for which $G(M) \in \mathcal{B}^*$ holds and $P_M(x, s^*)$ is maximum. (By continuity and closedness arguments one can always assume the existence of an optimum model, on both sides of (7).)

Let s_1 be the first state of a corresponding characteristic state sequence s^*. Suppose, the initial state distribution is not concentrated on s_1, that is, $\pi_i > 0$ for some $q_i \neq s_1$. Then, by decreasing π at q_i and increasing at s_1, $P_M(x, s^*)$ will clearly not decrease. So it is enough to restrict ourselves to the case when the initial distribution is concentrated on s_1.

Now let s_t and s_{t+1} be two different states of s^*, which follow each other when traversing s^* (we can assume that s^* contains at least two distinct states, otherwise the situation becomes trivial). Suppose, M allows a transition from s_i to $q_i \neq s_{t+1}$, with positive probability. Then we can increase the transition probability from s_t to s_{t+1} at the price of decreasing it from s_t to q_i. This clearly not decreases the value of $P_M(x, s^*)$. So we can concentrate the state transition probabilities on a single path without decreasing the objective function.

Thus, it is enough to restrict ourselves to linear models. It remains to show that it is sufficient to consider the possible longest linear model.

Suppose we have a linear model M with N states. Let s^* be a characteristic state sequence, which spends k_i time units in state q_i. Since $T > N$ is assumed to hold, there is a state q_i, for which $k_i > 1$ in s^*. Split up q_i into two consecutive states q_i, q_i^*, making the model longer by one.

Set

$$p_i^* = \begin{cases} 1, & \text{if } i < N \\ 0, & \text{if } i = N \end{cases}$$

and

$$b_i^*(k) \equiv b_i(k)$$

where p_i^* and $b_i^*(k)$ are the parameters of the new state. Call the new augmented model M^* and let s' be a state sequence in M^*, which contains q_j $(j \neq i)$ k_j times (the same as in s^*), but q_i and q_i^* occurs with multiplicity $k_i - 1$ and 1 in s', respectively.

Then we have

$$P_{M^*}(x, s') = \frac{1}{1 - p_i} \cdot P_M(x, s^*) \quad (p_i < 1)$$

so

$$P_M(x, s^*) < P_{M^*}(x, s') \leq P_M(x, s^{*\prime})$$

where $s^{*\prime}$ is a characteristic state sequence in M^*. Thus, increasing the number of states improves the linear model. The proof is completed. □

Now, we can summarize the algorithm for the general left-to-right case.

Algorithm 2:

Step 1. Determine N, the number of states of the possible longest linear model in question.

(In most practical cases it is trivially given by the restrictions. Otherwise, we have to find the longest paths in cycle-free directed graphs, which can be done by a Viterbi-type algorithm.)

Step 2. Run Algorithm 1 for an N-state linear model.

Remark. If we want, e.g. to find an optimum model in the set of all left-to-right HMM's with at most N_0 states, then the result of Step 1 in algorithm 2 is simply N_0.

4. Concluding remark

Our main result is the fast algorithm, which finds the global optimum of left-to-right HMM parameters, described in Section 3.

In many practical cases the task is the estimation from several different observation sequences, not only from a single one. There are two ways to generalize the new objective function when estimating from multiple observation: $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. Perhaps the most reasonable generalization of $P_M(x, s^*)$ is:

$$\prod_{i=1}^{n} \left\{ \max_{s} P_M(x^{(i)}, s^{(i)}) \right\}.$$

Unfortunately, we cannot solve this problem, however, the maximization of the related function:

$$\max_{s} \prod_{i=1}^{n} \left\{ P_M(x^{(i)}, s^{(i)}) \right\}$$

can be done based on Algorithm 2 (the details are omitted here).

5. Acknowledgement

This research has been done as a part of the speech processing research, directed by associate professor G. Gordos, at the Technical University of Budapest. The authors express their thanks for his support and encouragement.

References

1. *Rabiner, L.R., Juang, B.H.*, An Introduction to Hidden Markov Models, IEEE ASSP Magazine, January 1986.

2. *Jelinek, F.*, Continuous Speech Recognition by Statistical Methods, Proc. IEEE, **64**, *4*, pp. 532–556, April 1976.
3. *Levinson, S.E., Rabiner, L.R., Sondhi, M.M.*, An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, B.S.T.J., **62**, *4*, pp. 1035–1074, April 1983.
4. *Bahl, L.R., Jelinek, F., Mercer, L.R.*, A Maximum Likelihood Approach to Speech Recognition, IEEE Trans. Pattern Anal. and Machine Intelligence, **PAMI-5**, *2*, March 1983.
5. *Ash, R.*, Information Theory, J. Wiley, 1965, p. 16.
6. *Baum, L.E., Eagon, J.A.*, An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology, Bull. AMS, **73** (1967), pp. 360–363.
7. *Baum, L.E., Petrie, T., Soules, G., Weiss,N.*, A Maximization Technique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Ann. Math. Stat., **41** (1970), pp. 164–171.
8. *Rabiner, L.R., Wilpon, J.G., Juang, B.H.*, A Segmental *k*-means Training Procedure for Connected Word Recognition, AT&T Technical Journal, **65**, May/June 1986.

Алгоритм нахождения глобального оптимума параметров «слева-направо» модели спрятанной цепи Маркова

А. ФАРАГО, Г. ЛУГОШИ

(Будапешт)

В статье разработан алгоритм идентификации модели, описывающей поведение конечной однородной цепи Маркова, по наблюдениям случайной последовательности с конечным множеством состояний, стохастически связанной с марковской цепью. Предполагается, что переходная матрица цепи Маркова является верхней треугольной матрицей (в частности, и ленточной матрицей). Оцениванию подлежат переходные вероятности цепи Маркова и условные вероятности наблюдаемого процесса. Полученный в работе алгоритм идентификации отличается от известных тем, что он отыскивает глобальный, а не локальный максимум целевой функции. Кроме того, алгоритм неитеративен, что значительно сокращает затраты машинного времени, необходимого на реализацию алгоритма.

A. Faragó
G. Lugosi
Institute of Communication Electronics
Technical University of Budapest
H–1111 Budapest, Stoczek u. 2.
Budapest
Hungary

РУССКИЙ ПЕРЕВОД

Проблемы управления и теории информации, том 18, номер 6 (1989)

НЕСОБСТВЕННЫЕ ЗАДАЧИ МАТЕМАТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

И. И. ЕРЕМИН, А. А. ВАТОЛИН

(Свердловск)

1. Введение

Роль моделей математического программирования (МП) в решении планово-экономических задач общеизвестна. Система ограничений в таких задачах (формально — система неравенств и уравнений), отражающих ресурсные и технологические ограничения, ограничения среды, директивные требования и т.д., может оказаться несовместной (противоречивой). Истоками таких моделей могут служить: ресурсный дефицит, завышенные директивные задания, отсутствие резервов производственных мощностей, неточность экономической информации, учет противоречивых требований, учет нормативов на отрицательное воздействие производства на среду и др. Возникновение противоречивых моделей — довольно обычная для практики ситуация. В простейших моделях преодоление противоречивости модели за счет ее коррекции (ослабления ограничений или выбрасывания их части, коррекции информации и т.д.) не представляет особого труда. Изучение более сложных моделей, отражающих существенно более сложные ситуации, приводит к принципиальной необходимости учета появления несобственных (противоречивых) моделей, а следовательно, необходимости разработки их теории (в первую очередь — двойственности), методов численного анализа и его программного обеспечения [1].

Запишем задачу МП над E^n:

$$\sup\{f(x) \mid f_j(x) \le 0,\ j = 1, \dots, m,\ x \ge 0\} \tag{1.1}$$

и ее частный случай — задачу линейного программирования (ЛП):

$$\sup\{\langle c, x\rangle \mid Ax \le b,\ x \ge 0\}. \tag{1.2}$$

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

Пусть $\Phi(x,u) = f(x) - \sum_{j=1}^{m} u_j f_j(x)$ — функция Лагранжа для (1.1); для (1.2) такой функцией будет $L(x,u) = \langle c,x\rangle - \langle Ax - b, u\rangle$. Постановка (1.1) эквивалентна (по значению) задаче $\sup\limits_{x\geq 0}\inf\limits_{u\geq 0}\Phi(x,u) = \overline{\Phi}$. Задача

$$\inf_{u\geq 0}\sup_{x\geq 0}\Phi(x,u) = \underline{\Phi} \tag{1.1}^*$$

называется двойственной к (1.1). Последняя для случая ЛП эквивалентна задаче

$$\inf\{\langle b,u\rangle \mid A^T u \geq c,\ u\geq 0\}. \tag{1.2}^*$$

Задача (1.1) называется собственной, если

$$\overline{\Phi} = \underline{\Phi} = \Phi(\overline{x},\overline{u}) = \sup_{x\geq 0}\Phi(x,\overline{u}) = \inf_{u\geq 0}\Phi(\overline{x},u),\ \ \overline{x}\geq 0,\ \ \overline{u}\geq 0;$$

в противном случае она называется несобственной. Условие $M = \{x \geq 0 \mid f_j(x)\leq 0,\ \ j = 1,\dots,m\} = \emptyset$ реализует частный случай несобственности. В случае задачи (1.2) свойство собственности эквивалентно ее разрешимости. В несобственном случае возможны три альтернативы: 1) $M = \emptyset$, $M^* \neq \emptyset$; 2) $M \neq \emptyset$, $M^* = \emptyset$; 3) $M = \emptyset$, $M^* = \emptyset$. Здесь M и M^* — допустимые множества задач (1.2) и (1.2)*, соответственно. В зависимости от выполнимости одного из условий 1)–3) говорят о несобственной задаче 1-го, 2-го или 3-го рода.

С несобственной задачей МП (или ЛП) свяжем понятие ее аппроксимации (или коррекции). Под этим мы будем понимать тот или иной способ формального, но содержательно интерпретируемого сведения (отображения) исходной задачи к задаче собственной. Поясним это на примерах.

Пример 1. Задаче (1.1) поставим в соответствие задачу с приращениями Δb_j и Δc_i:

$$\sup\{f(x) - \langle \Delta c, x\rangle \mid f_j(x)\leq b_j + \Delta b_j,\ \ j = 1,\dots,m,\ \ x\geq 0\}. \tag{1.3}$$

Для (1.2) задача (1.3) примет вид:

$$\sup\{\langle c - \Delta c, x\rangle \mid Ax \leq b + \Delta b,\ \ x\geq 0\}. \tag{1.4}$$

Введя функцию $\varphi(\Delta)$ качества коррекции, саму задачу оптимальной коррекции можно записать в виде

$$\min\{\varphi(\Delta) \mid \Delta \in K\}, \tag{1.5}$$

где $K = \{\Delta \in K_0 /\ (1.4) \text{ — собственная}\}$, K_0 — допустимое множество приращений $\Delta = [\Delta c, \Delta b]$. Если $\tilde{\Delta} \in \operatorname{Arg}(1.5)$, то $\max\{\langle c - \tilde{\Delta}c, x\rangle \mid Ax \leq b + \Delta\tilde{b},\ x\geq 0\}$

— компромиссная модель; здесь $\tilde{\Delta} = [\Delta\tilde{c}, \Delta\tilde{b}]$, Arg (1.5) — оптимальное множество задачи (1.5).

Пример 2. Пусть ограничения задачи (1.1) разбиты на две подсистемы: $f_j(x) \le 0$, $j \in J_0$, $x \ge 0$ и $f_j(x) \le 0$, $j \in J_1$, первая из которых совместна. Введя функцию невязки $d(x)$ для второй подсистемы (например, $d(x) = \sum\limits_{j \in J_1} \overline{u}_j f_j^+(x)$, $\overline{u}_j > 0$, $\forall j \in J_1$), можно определить множество

$$\tilde{M} = \operatorname{Arg}\min\{d(x) \mid f_j(x) \le 0,\ j \in J_0,\ x \ge 0\}$$

и задачу

$$\max\{f(x) \mid x \in \tilde{M}\}. \tag{1.6}$$

Последняя может рассматриваться в качестве аппроксимирующей.

В дальнейшем будут использованы следующие обозначения:

Arg ... — оптимальное множество задачи, обозначение (или номер) которой стоит на месте многоточия.

Если $\alpha \in E^1$, т.е. α — число, то $\alpha^+ = \max\{0, \alpha\}$; если $x = [x_1, \dots, x_k]^T \in E^k$, то $x^+ = [x_1^+, \dots, x_k^+]^T$, $E_+^k = \{x \in E^k \mid x \ge 0\}$.

2. Двойственность для несобственных задач ЛП [1, 2]

Теория двойственности в математическом программировании играет большую роль. Она позволяет более глубоко проникнуть в математическую суть задач и является генератором эффективных методов их решения. Двойственность для несобственных задач играет такую же роль. Схема двойственности состоит в следующем. Паре двойственных задач МП: P и P^* по единой схеме π ставится в соответствие пара задач C и $C^\#$, аппроксимирующих P и P^* и связанных между собой классической теоремой двойственности (в стандартной формулировке). Схематически это можно изобразить так:

$$\begin{array}{ccc} P & \xrightarrow{\pi} & C \\ (*)\ \downarrow & & \downarrow\ (\#) \\ P^* & \xrightarrow{\pi} & C^\# \end{array} \qquad (\#)$$

Применительно к задачам (1.2) и $(1.2)^*$ задачи C и $C^\#$ могут быть реализованы в следующем виде:

$$C: \quad \sup\Big\{\langle c, x\rangle - \sum_{j=1}^{m_0} R_j \|(A_j x - b^j)\|_{p(j)} \mid A_0 x \le b^0,\ x \ge 0,$$

$$\|x^i\|_{q(i)} \le r_i,\ i = 1, \dots, n_0\Big\},$$

$$C^*: \quad \inf\Big\{\langle b,u\rangle + \sum_{i=1}^{n_0} r_i \|(c^i - B_i^T u)^+\|^*_{q(i)} \mid B_0^T u \geq c^0, \quad u \geq 0,$$

$$\|u^j\|^*_{p(j)} \leq R_j, \quad j = 1,\ldots,m_0\Big\}.$$

Здесь $R_j \geq 0$, $r_i \geq 0$, $j = 1,\ldots,m_0$, $i = 1,\ldots,n_0$ — числовые параметры; $\{\|\cdot\|_{p(j)}\}, \{\|\cdot\|_{q(i)}\}$ — произвольный набор монотонных (на неотрицательных ортантах соответствующих пространств) норм; $\{\|\cdot\|^*_{p(j)}\}$, $\{\|\cdot\|^*_{q(i)}\}$ — им сопряженные нормы, которые также предполагаются монотонными;

$$A = \begin{bmatrix} A_0 \\ \vdots \\ A_{m_0} \end{bmatrix} = [B_0,\ldots,B_{n_0}];$$

$c^T = [c^0,\ldots,c^{n_0}]$, $x^T = [x^0,\ldots,x^{n_0}]$ — разбиения векторов c и x, соответствующие разбиению матрицы A на вертикальные подматрицы $\{B_i\}$; $b^T = [b^0,\ldots,b^{m_0}]$, $u^T = [u^0,\ldots,u^{m_0}]$ — то же самое, но для разбиения A на горизонтальные подматрицы $\{A_j\}$.

Норма $\|x\|$, $x \in E^k$ называется монотонной на E^k_+, если из $x \geq y \geq 0$ следует $\|x\| \geq \|y\|$.

Приведем примеры норм, монотонных вместе со своими сопряженными:

$$\|w\|_0 = \max_{1\leq j\leq k} \alpha_j |w_j|, \quad \|w\|_1 = \sum_{j=1}^k \alpha_j |w_j|, \quad \|w\|_2 = \left(\sum_{j=1}^k \alpha_j w_j^2\right)^{\frac{1}{2}},$$

то

$$\|w\|^*_0 = \sum_{j=1}^k \alpha_j^{-1} |w_j|, \quad \|w\|^*_1 = \max_{1\leq j\leq k} \alpha_j^{-1} |w_j|, \quad \|w\|^*_2 = \left(\sum_{j=1}^k \alpha_j^{-1} w_j^2\right)^{\frac{1}{2}},$$

где $\alpha_j > 0$, $j = 1,\ldots,k$.

Будем считать, что $M_0 = \{x \geq 0 \mid A_0 x \leq b_0\} \neq 0$ и $M_0^{\#} = \{u \geq 0 \mid B_0^T u \geq\geq c^0\} \neq \emptyset$. Подматрицам из числа $\{B_i\}$ и $\{A_j\}$ можно придавать пустые ($= \emptyset$) значения. Целевые функции задач C и $C^{\#}$ обозначим через $f_R(x)$ и $f_r^{\#}(u)$, а через $M(r)$ и $M^{\#}(R)$ — их допустимые множества.

Теорема 2.1. 1) Для любых $\overline{x} \in M(r)$ и $\overline{u} \in M^{\#}(R)$ справедливо неравенство $f_R(\overline{x}) \leq f_r^{\#}(\overline{u})$. Отсюда: из $f_R(\overline{x}) = f_r^{\#}(\overline{u})$ следует $\overline{x} \in \operatorname{Arg} C$, $\overline{u} \in \operatorname{Arg} C^{\#}$.

2) Пусть задача C разрешима и $\exists \overline{x} \in M_0$: $\|\overline{x}^i\|_i < r_i$, $i = 1,\ldots,n_0$. Тогда задача $C^{\#}$ также разрешима и их оптимальные значения совпадают.

Справедлив и обратный вариант теоремы.

Выше $R = [R_1, \ldots, R_{m_0}]$, $r = [r_1, \ldots, r_{n_0}]$.

Рассмотрим некоторые частные реализации задач C и $C^{\#}$.

$$
\begin{aligned}
C_0 &: \quad \max\{\langle c, x\rangle - \langle R, (Ax-b)^+\rangle \mid 0 \le x \le r\},\\
C_0^{\#} &: \quad \min\{\langle b, u\rangle + \langle r, (c - A^T u)^+\rangle \mid 0 \le u \le R\},\\
C_1 &: \quad \max\{\langle c, x\rangle - R_0\|(Ax-b)^+\| \mid x \ge 0\},\\
C_1^{\#} &: \quad \min\{\langle b, u\rangle \mid A^T u \ge c,\ u \ge 0,\ \|u\|^* \le R_0\},\\
C_2 &: \quad \max\{\langle c, x\rangle \mid Ax \le b,\ x \ge 0,\ \|x\| \le r_0\},\\
C_2^{\#} &: \quad \min\{\langle b, u\rangle + r_0\|(c - A^T u)^+\|^* \mid u \ge 0\}.
\end{aligned}
$$

В записи задач C_1 и $C_1^{\#}$, C_2 и $C_2^{\#}$ нормы произвольные, но монотонные.

Теорема 2.2. 1) Задачи C_0 и $C_0^{\#}$ разрешимы при произвольной реализации задачи L и их оптимальные значения совпадают.

2) Из $\overline{u} \in M^* = \{u \ge 0 \mid A^T u \ge c\} \ne \emptyset$, $\|\overline{u}\|^* < R_0$ вытекает разрешимость C_1 и $C_1^{\#}$ и совпадение их оптимальных значений.

3) Из $\overline{x} \in M = \{x \ge 0 \mid Ax \le b\} \ne \emptyset$, $\|\overline{x}\| < r_0$ вытекает разрешимость C_2 и $C_2^{\#}$ и совпадение их оптимальных значений.

Схема (#) допускает более общую реализацию по сравнению с C и $C^{\#}$. Опишем ее.

Пусть R, ε_i, R_j, r_i, δ_j, $i = 1, \ldots, n_0$, $j = 1, \ldots, m_0$ — система положительных параметров; Φ, Ψ — функции ν-мерного аргумента, где $\nu = n + m - \nu_1 - \nu_2$, а ν_1 и ν_2 — размерности векторов x^0 и u^0.

Сформулируем задачи

$$
\begin{aligned}
D &: \sup\{\langle c, x\rangle - R\Phi(\varepsilon_1 x^1, \ldots, \varepsilon_{n_0} x^{n_0};\ R_1(A_1 x - b^1)^+, \ldots\\
&\quad \ldots, R_{m_0}(A_{m_0} x - b^{m_0})^+) \mid A_0 x \le b^0,\ x \ge 0\}; \qquad (2.1)\\
D^{\#} &: \inf\{\langle b, u\rangle + R\Psi(r_1(c^1 - B_1^T u)^+, \ldots\\
&\quad \ldots, r_{n_0}(c^{n_0} - B_{n_0}^T u)^+; \delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}) \mid B_0^T u \ge c^0,\ u \ge 0\}. \qquad (2.2)
\end{aligned}
$$

Аргументами функций Φ, Ψ являются соответственно векторы из R^{ν}:

$$
\begin{aligned}
\Gamma(x) &= [\varepsilon_1 x^1, \ldots, \varepsilon_{n_0} x^{n_0}; R_1(A_1 x - b^1)^+, \ldots, R_{m_0}(A_{m_0} x - b^{m_0})^+],\\
\Gamma^{\#}(u) &= [r_1(c^1 - B_1^T u)^+, \ldots, r_{n_0}(c^{n_0} - B_{n_0}^T u)^+; \delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}].
\end{aligned}
$$

Ниже будет указана связь между функциями Φ, Ψ и системами констант задач D и $D^{\#}$, которая обеспечит для задач (2.1) и (2.2) выполнение соотношений двойственности. Прежде всего, определим класс функций, из которого могут выбираться Φ и Ψ. Пусть $E_+^k = \{z \in R^k \mid z \ge 0\}$. Функцию $\Omega(z)$, отображающую E^k в $E^1 \cup \{+\infty\}$, назовем допустимой, если она удовлетворяет условиям: 1) Ω — выпуклая, полунепрерывная снизу функция, $\operatorname{dom}\Omega \subset E_+^k$; 2) Ω является монотонно не убывающей на E_+^k и $\Omega(0) = 0$.

В качестве Φ, Ψ будут браться допустимые функции.

Далее, функции Ω над пространством R^k поставим в соответствие функцию $\Omega^\#$ над тем же пространством, определяемую равенством

$$\Omega^\#(z^*) = \Omega^*(z^*) + \delta(z^* \mid E_+^k),$$

где Ω^* — функция, сопряженная к Ω; δ — индикаторная функция, т.е.

$$\delta(z \mid W) = \begin{cases} 0, & z \in W, \\ +\infty, & z \notin W. \end{cases}$$

Имеет место свойство: если функция Ω допустима, то допустима и $\Omega^\#$, причем выполняется равенство $(\Omega^\#)^\# = \Omega$. Подчиним константы задач D и $D^\#$ равенствам

$$\varepsilon_i r_i = \delta_j R_j = R^{-1}, \quad i = 1, \ldots, n_0, \quad j = 1, \ldots, m_0 \tag{2.3}$$

и положим в $D^\#$: $\Psi = \Phi^\#$. Далее под $D^\#$ будем понимать задачу (2.2), в которой $\Psi = \Phi^\#$.

Замечание 1. Подсистемы $A_0 x \le b^0$, $x \ge 0$ и $B_0^T u \ge c^0$, $u \ge 0$ систем ограничений исходных задач L и L^* будут предполагаться, как и прежде, совместными. Содержательно это может соответствовать выделению директивных ограничений в предположении их обоснованности.

Замечание 2. Равенство $\Phi = (\Phi^\#)^\#$ обеспечивает выполнение свойства $D \equiv (D^\#)^\#$.

Замечание 3. D и $D^\#$ представляют собой задачи выпуклого программирования.

Выписанная общая форма задач D и $D^\#$ допускает большое число конкретных реализаций, соответствующих различным подходам к аппроксимации исходных несобственных задач L и L^*.

Прежде всего, отметим, что в случае $n_0 = 0$, $m_0 = 0$ задачи D и $D^\#$ совпадают с L и L^*. Если, например, L — несобственная 1-го рода, то частными реализациями задач D и $D^\#$ будут

$$D_1 : \ \sup\{\langle c, x\rangle - R\Phi(R_1(A_1 x - b^1)^+, \ldots$$
$$\ldots, R_{m_0}(A_{m_0} x - b^{m_0})^+) \mid A_0 x \le b^0, \ \ x \ge 0\}, \tag{2.4}$$

$$D_1^\# : \ \inf\{\langle b, u\rangle + R\Phi^\#(\delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}) \mid A^T u \ge c, u \ge 0\}. \tag{2.5}$$

Эти задачи получены из D и $D^\#$ при $n_0 = 0$.

Задачу D, в которой допустимая функция Φ является выпуклой и кусочно-линейной, назовем l-задачей.

Введем обозначения для целевых функций задач D и $D^{\#}$:

$$f(x) = \langle c, x\rangle - R\Phi(\Gamma(x)), \quad f^{\#}(u) = \langle b, u\rangle + \Phi^{\#}(\Gamma^{\#}(u)).$$

Оптимальные значения и оптимальные множества задач D и $D^{\#}$ обозначим соответственно через f, $\tilde{f}^{\#}$ и $\tilde{M}$, $\tilde{M}^{\#}$.

Сформулируем теорему двойственности, связывающую задачи D и $D^{\#}$.

Теорема 2.3. Для задач D, $D^{\#}$ справедливы утверждения:

1. $\tilde{f} \le \tilde{f}^{\#}$.

2. Если $\tilde{f} < +\infty$ и задача D удовлетворяет (в той или иной форме) условию регулярности, то $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$, причем $\tilde{M}^{\#} \neq \emptyset$.

3. Если $\tilde{f}^{\#} > -\infty$ и задача $D^{\#}$ удовлетворяет условию регулярности, то $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$, причем $\tilde{M} \neq \emptyset$.

4. Если для некоторого $\alpha \in E^1$ множество $\{x \in M \mid f(x) \ge \alpha\}$ непусто и ограничено, то $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$, причем $\tilde{M} \neq \emptyset$.

5. Если для некоторого $\alpha \in E^1$ множество $\{u \in M^{\#} \mid f^{\#}(u) \le \alpha\}$ непусто и ограничено, то $-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty$, причем $\tilde{M}^{\#} \neq \emptyset$.

6. Если $M \neq \emptyset$ или $M^{\#} \neq \emptyset$, причем D является ℓ-задачей, то $\tilde{f} = \tilde{f}^{\#}$. При этом, если $\tilde{f} \neq \pm\infty$, то $\tilde{M} \neq \emptyset$, $\tilde{M}^{\#} \neq \emptyset$.

Ниже будет специально рассмотрен случай, когда функция зависит от норм невязок. А именно, пусть функция Φ имеет следующий конкретный вид:

$$\Phi(\Gamma(x)) = \varphi(\varepsilon_1\|x^1\|_{q(1)}, \dots, \varepsilon_{n_0}\|x^{n_0}\|_{q(n_0)}; R_1\|(A_1x - b^1)^+\|_{p(1)}, \dots$$
$$\dots, R_{m_0}\|(A_{m_0}x - b^{m_0})^+\|_{p(m_0)}) + \delta(\Gamma(x) \mid R_+^{\nu}). \tag{2.6}$$

Здесь $\{\|\cdot\|_{q(i)}\}$, $\{\|\cdot\|_{p(j)}\}$ — наборы монотонных норм (вместе со своими сопряженными) в соответствующих пространствах, φ — функция над $E^{n_0+m_0}$.

Утверждение. Пусть нормы $\{\|\cdot\|_{q(i)}\}$, $\{\|\cdot\|_{p(j)}$ — монотонны, функция φ — выпуклая монотонно не убывающая на множестве $E_+^{n_0+m_0}$, удовлетворяющая условию $\varphi(0) = 0$, $\varphi(z) > 0$, $0 \neq z \in E_+^{n_0+m_0}$. Тогда функция Φ, определенная в (2.6), допустима, причем

$$\Phi^{\#}(\Gamma^{\#}(u)) = \varphi^*(r_1\|(c^1 - B_1^Tu)^+\|_{q(1)}^*, \dots, r_{n_0}\|(c^{n_0} - B_{n_0}^Tu)^+\|_{q(n_0)}^*;$$

$$\delta_1\|u^1\|_{p(1)}^*, \dots, \delta_{m_0}\|u^{m_0}\|_{p(m_0)}^*) + \delta(\Gamma^{\#}(u) \mid E_+^{\nu}).$$

При выполнении предположений сформулированного утверждения задачи (2.1), (2.2), соответствующие конкретному виду (2.6) функции Ф, запишутся в виде

$$D: \sup\{\langle c,x\rangle - R\varphi(\varepsilon_1\|x^1\|_{q(1)},\dots,\varepsilon_{n_0}\|x^{n_0}\|_{q(n_0)}; R_1\|(A_1x-b^1)^+\|_{p(1)},\dots \\ \dots, R_{m_0}\|(A_{m_0}x-b^{m_0})^+\|_{p(m_0)}) \mid A_0x\le b^0,\ x\ge 0\}, \tag{2.7}$$

$$D^{\#}: \inf\{\langle b,u\rangle + R\varphi^*(r_1\|(c^1-B_1^Tu)^+\|^*_{q(1)},\dots,r_{n_0}\|(c^{n_0}-B_{n_0}^Tu)^+\|^*_{q(n_0)}; \\ \delta_1\|u^1\|^*_{p(1)},\dots,\delta_{m_0}\|u^{m_0}\|^*_{p(m_0)} \mid B_0^T\ge c^0,\ u\ge 0\}. \tag{2.8}$$

Пусть, например, в задаче (2.7) функция φ имеет вид:

$$\varphi(v)=\sum_{i=1}^{n_0}\alpha_i^{-1}|v_i|^{\alpha_i}+\sum_{j=1}^{m_0}\beta_j^{-1}|v_{n_0+j}|^{\beta_j} \tag{2.9}$$

где $1\le\alpha_i\le+\infty$, $1\le\beta_j\le+\infty$, $i=1,\dots,n_0$, $j=1,\dots,m_0$. При этом оперирование с бесконечными величинами производится по правилам: $-(+\infty)=$ $=-\infty$, $\gamma\pm\infty=\pm\infty$,

$$(+\infty)^{-1}\gamma^{+\infty}=\begin{cases}0, & 0\le\gamma\le 1\\ +\infty, & \gamma>1\end{cases}$$

(γ — произвольное действительное число). Тогда непосредственная проверка показывает, что сопряженной к функции $\omega(\gamma)=\alpha^{-1}|\gamma|^{\alpha}$, где $1\le\alpha\le\le+\infty$, является $\omega^*(\gamma)=\sigma^{-1}|\gamma|^{\sigma}$, в которой параметр σ выбирается из условия $\alpha^{-1}+\sigma^{-1}=1$, причем в данном условии по определению считается $(+\infty)^{-1}=0$. Отсюда получаем

$$\varphi^*(v)=\sum_{i=1}^{n_0}\sigma_i^{-1}|v_i|^{\sigma_i}+\sum_{j=1}^{m_0}\tau_j^{-1}|v_{n_0+j}|^{\tau_j},$$

где параметры σ_i, τ_j находятся из условий

$$\alpha_i^{-1}+\sigma_i^{-1}=1,\quad \beta_j^{-1}+\tau_j^{-1}=1,\quad 1\le\sigma_i\le+\infty, \\ 1\le\tau_j\le+\infty,\quad i=1,\dots,n_0,\quad j=1,\dots,m_0. \tag{2.10}$$

Учитывая, что функция φ вида (2.9), очевидно, удовлетворяет условиям утверждения, и, полагая $R=1$, получаем соответствующую ей частную реализацию задач (2.7), (2.8):

$$D: \sup\Big\{\langle c,x\rangle-\sum_{i=1}^{n_0}\alpha_j^{-1}\|\varepsilon_ix^i\|^{\alpha_i}_{q(i)}- \\ -\sum_{j=1}^{m_0}\beta_j^{-1}\|R_j(A_jx-b^j)^+\|^{\beta_j}_{p(j)}\mid A_0x\le b^0,\ x\ge 0\Big\}, \tag{2.11}$$

$$D^{\#}: \ \inf\Big\{\langle b,u\rangle + \sum_{i=1}^{n_0}\sigma_i^{-1}\|r_i(c^i - B_i^T u)^+\|_{q(i)}^{*\sigma_i} + \\ + \sum_{j=1}^{m_0}\tau_j^{-1}\|\delta_j u^j\|_{p(j)}^{*\tau_j} \mid B_0^T u \geq c^0, \ \ u \geq 0\Big\}, \tag{2.12}$$

где параметры ε_i, r_i, R_j, δ_j, α_i, β_j, σ_i, τ_j выбираются из условий (2.3) (при $R = 1$), (2.10).

Приведем несколько примеров задач, получающихся из (2.11), (2.12) при различных вариантах выбора параметров.

1) При $B_0 = A$, $n_0 = 1$, $\alpha_1 = \sigma_1 = 2$, $\beta_j = 1$, $\tau_j = +\infty$, $j = 1,\dots,m_0$ получаем задачи:

$$\sup\Big\{\langle c,x\rangle - \sum_{j=1}^{m_0} R_j\|(A_j x - b^j)^+\|_{p(j)} - \varepsilon\|x\|_q^2 \mid A_0 x \leq b^0, \ \ x \geq 0\Big\},$$

$$\inf\{\langle b,u\rangle + (4\varepsilon)^{-1}\|(c - A^T u)^+\|_q^{*2} \mid u \geq 0, \ \ \|u^j\|_{p(j)}^* \leq R_j, \ \ j = 1,\dots,m_0\},$$

где R_j, ε — произвольные положительные параметры. В частности, если $\|x\|_q$ — евклидова норма, то первую из выписанных задач можно рассматривать как регуляризацию (по Тихонову [3]) задачи

$$\sup\left\{\langle c,x\rangle - \sum_{j=1}^{m_0} R_j\|(A_j x - b^j)^+\|_{p(j)} \mid A_0 x \leq b^0, \ \ x \geq 0\right\},$$

аппроксимирующей несобственную задачу 1-го рода L.

2) При $A_0 = \emptyset$, $B_0 = A$, $m_0 = 1$, $\beta_1 = \tau_1 = 2$ получаем задачи:

$$\sup\{\langle c,x\rangle - R\|(Ax - b)^+\|_p^2 \mid x \geq 0\},$$

$$\inf\{\langle b,u\rangle + \delta\|u\|_p^{*2} \mid A^T u \geq c, \ \ u \geq 0\},$$

где $R > 0$, $\delta > 0$, $R\delta = 1/4$, отвечающие случаю несобственности 1-го рода задачи L.

3) При $A_0 = \emptyset$, $B_0 = \emptyset$, $m_0 = n_0 = 1$, $\alpha_1 = \beta_1 = \sigma_1 = \tau_1 = 2$ задачи (2.11), (2.12) запишутся в следующем симметричном виде:

$$\sup\{\langle c,x\rangle - R\|(Ax - b)^+\|_p^2 - \varepsilon\|x\|_q^2 \mid x \geq 0\},$$

$$\inf\{\langle b,u\rangle + r\|(c - A^T u)^+\|_q^{*2} + \delta\|u\|_p^{*2} \mid u \geq 0\},$$

соответствующем случаю несобственности 3-го рода задач L и L^* (положительные параметры R, ε, r, δ выбираются из условий $\varepsilon r = R\delta = 1/4$).

4) При $\alpha_j = +\infty$, $\beta_i = 1$, $\sigma_i = 1$, $\tau_j = +\infty$, $j = 1, \ldots, m_0$, $i = 1, \ldots, n_0$ получаем задачи C и $C^{\#}$, предложенные одним из авторов в работе [4].

3. Двойственность для несобственных задач выпуклого программирования 1-го рода [2]

Задачу выпуклого программирования рассмотрим в форме

$$\sup\{f_0(x) \mid f_j(x) \le 0, \quad j = 1, \ldots, m, \quad x \in M\} \tag{3.1}$$

где $\{-f_0, f_1, \ldots, f_m\}$ — выпуклые собственные функции, M — выпуклое множество. В дальнейшем будем считать, что

$$M = \operatorname{dom}(-f_0), \quad M \subset \operatorname{dom} f_j, \quad \operatorname{ri} M \subset \operatorname{ri} \operatorname{dom} f_j, \quad j = 1, \ldots, m.$$

Обозначив $F(x,u) = f_0(x) - \sum\limits_{j=1}^{m} u_j f_j(x)$, $u^T = (u_1, \ldots, u_m)$,

$$M^* = \left\{u \mid \sup_{x \in M} F(x,u) < +\infty\right\}, \quad g_0(u) = \sup_{x \in M} F(x,u),$$

запишем двойственную к (3.1) задачу в виде

$$\inf\{g_0(u) \mid u \in M^*, u \ge 0\}. \tag{3.2}$$

Рассмотрим случай, когда (3.1) — несобственная задача 1-го рода, т. е. ее допустимое множество пусто, а допустимое множество задачи (3.2) — не пусто. Как и в 2, зафиксируем разбиение $(f_1, \ldots, f_m) = (F_0, \ldots, F_m)$ и рассмотрим задачи

$$\sup\{f_0(x) - R\Phi(R_1 F_1^+(x), \ldots, R_{m_0} F_{m_0}^+(x)) \mid F_0(x) \le 0, \quad x \in M\}, \tag{3.3}$$

$$\inf\{g_0(u) + R\Phi^{\#}(\delta_1 u^1, \ldots, \delta_{m_0} u^{m_0}) \mid u \in M^*, \quad u \ge 0\}, \tag{3.4}$$

где положительные параметры R, R_j, δ_j выбраны из условия $R_j \delta_j = R^{-1}$, $j = 1, \ldots, m_0$.

Пусть $\tilde{f}$, $\tilde{f}^{\#}$ — оптимальные значения, а $\tilde{M}$, $\tilde{M}^{\#}$ — оптимальные множества задач (3.3), (3.4). Целевую функцию задачи (3.3) обозначим через $f(x)$.

Теорема 3.1. Справедливы следующие утверждения:

1. $\tilde{f} \le \tilde{f}^{\#}$.

2. Пусть $\tilde{f} < +\infty$ и существует такой допустимый вектор $\overline{x}$ задачи (3.3), что $\overline{x} \in \operatorname{ri} M$, причем $f_j(\overline{x}) < 0$ для всех не аффинных функций из числа тех, которые составляют вектор F_0. Тогда

$$-\infty < \tilde{f} = \tilde{f}^{\#} < +\infty, \quad \tilde{M}^{\#} \ne \emptyset.$$

3. Пусть функции $\{-f_0, f_1, \ldots, f_m\}$ полунепрерывны снизу и для некоторого $\alpha \in E^1$ множество $\{x \mid f(x) \geq \alpha, \quad F_0(x) \leq 0, \quad x \in M\}$ — не пусто и ограничено. Тогда $\tilde{f} = \tilde{f}^{\#}$, $\tilde{M} \neq \emptyset$.

Остановимся на общей схеме формирования двойственности для несобственных задач выпуклого программирования в их произвольной реализации, т.е. без предположения того, что имеет место несобственность 1-го рода. Рассмотрим задачу (3.1), в которой $M = E^n_+$, т. е.

$$C: \quad \sup\{f_0(x) \mid f_j(x) \leq 0, \quad j = 1, \ldots, m, \quad x \geq 0\}. \tag{3.5}$$

Для упрощения описания общей схемы двойственности для задачи (3.5) будем предполагать ее гладкой, т. е. функции $\{f_j(x)\}_0^m$ дифференцируемыми.

Если $g(x)$ — некоторая дифференцируемая функция, то ее линеаризация в точке $p \in E^n$ состоит в переходе от $g(x)$ к линейной функции

$$l_p(x) = \langle \nabla g(p), x - p \rangle + g(p): \quad g(x) \xrightarrow{(l)} l_p(x).$$

Под линеаризацией в точке p гладкой задачи математического программирования, пусть в форме (3.5), понимается переход к задаче линейного программирования L_p, которая получается из C путем линеаризации всех входящих в нее функций:

$$C \xrightarrow{(l)} L_p.$$

Тогда формальной схемой формирования двойственности в нелинейном программировании будет следующая схема:

$$C \xrightarrow{(l)} L_p \xrightarrow{(*)} L_p^* \xrightarrow{(p=x)} L_x^* \equiv C^*. \tag{3.6}$$

Действуя согласно этой схеме, двойственная к C примет вид

$$C^*: \quad \inf\{F(x,u) - \langle v, x \rangle \mid \nabla_x F(x,u) \geq 0, \quad (x,v) \geq 0\},$$

где $F(x,u) = f_0(x) - \sum\limits_{j=1}^{m} u_j f_j(x)$. Если в (3.5) не выделять требование $x \geq 0$, то двойственная задача будет иметь вид

$$\inf\{F(x,u) \mid \nabla_x F(x,u) = 0, \quad u \geq 0\}.$$

Схема (3.6) является общей для задач математического программирования при стандартном условии их разрешимости, поскольку для несобственных задач МП обычная двойственность теряет свою содержательность,

и в качестве естественного приема формирования двойственности в общем случае напрашивается схема:

$$\begin{array}{ccccccc} C & \xrightarrow{(l)} & L_p & \xrightarrow{\pi} & P_p & \xrightarrow{(p=x)} & P_x \equiv P \\ & & {\scriptstyle (*)}\downarrow\uparrow & & \downarrow\uparrow & & \downarrow {\scriptstyle (\#)} \\ & & L_p^* & \xrightarrow[\Pi']{\pi} & P_p^\# & \xrightarrow{(p=x)} & P_x^\# \equiv P^\# \\ \downarrow & & & & & & \uparrow \\ C^* & & & & & & \end{array} \tag{3.7}$$

Кратко:

$$\begin{array}{ccc} C & \xrightarrow{\Pi} & P \\ \downarrow & & \downarrow {\scriptstyle (\#)} \\ C^* & \xrightarrow{\Pi'} & P^\# \end{array} \tag{3.8}$$

где Π — отображение, являющееся композицией верхних переходов, а Π' — по определению.

Если проделать все преобразования, соответствующие отображению Π, то получим следующий вид задачи

$$P: \ \sup\left\{ f_0(x) - \sum_{j=1}^{m_0} R_j \|F_j(x)\|_{p(j)} \mid F_0(x) \le 0, \ \ x \ge 0, \right.$$
$$\left. \|x^i\|_{q(i)} \le r_i, \ \ i = 1, \ldots, n_0 \right\} \ (= \tilde{f}), \tag{3.9}$$

где $[f_1, \ldots, f_m] = [F_0, \ldots, F_{m_0}] = F(x)$; смысл остальных символов остается прежним. Выделенная подсистема $F_0(x) \le 0$, $x \ge 0$ предполагается совместной.

Для того, чтобы записать задачу $P^\#$, получаемую в силу схемы (3.8), введем обозначения:

$$A^x = \begin{bmatrix} \dfrac{\partial f_1(x)}{\partial x_1}, & \ldots, & \dfrac{\partial f_1(x)}{\partial x_n} \\ \ldots & \ldots & \ldots \\ \dfrac{\partial f_m(x)}{\partial x_1}, & \ldots, & \dfrac{\partial f_m(x)}{\partial x_n} \end{bmatrix}, \ b_x = A^x x - F(x),$$

$$c_x = \nabla f_0(x) = \left[\frac{\partial f_0(x)}{\partial x_1}, \ldots, \frac{\partial f_0(x)}{\partial x_n} \right]^T.$$

Пусть A_j^x, $B_i(x)$, b_x^j, c_x^i, $j = 0, 1, \ldots, m_0$, $i = 0, 1, \ldots, n_0$ — разбиение матрицы A^x и векторов b_x, c_x, соответствующее разбиению матрицы A и векторов b и c в 2, при этом горизонтальное разбиение согласовано с разбиением

системы ограничений задачи (3.5) на подсистемы $F_j(x) \le 0$, $j = 0, 1, \ldots, m_0$ в (3.9).

После введенных обозначений задача $P^\#$ примет вид

$$P^\# : \ \inf\{F(x,u) - \langle \nabla_x F(x,u), x\rangle + \sum_{i=1}^{n} r_i \|(c_x^i - B_i^T(x))^+\|_{q(i)}^* \mid B_0^T(x)u \ge c_x^0,$$

$$u \ge 0, \ \ \|u^j\|_{p(j)}^* \le R_j, \ \ j = 1, \ldots, m_0\} \ \ (= \tilde{f}^\#). \tag{3.10}$$

Заметим, что задачу P можно было бы выписать через символы A_j^x и b_x^j, но эта запись была бы эквивалентной (3.9), но только более громоздкой по форме.

Теорема 3.2. Пусть (3.5), т. е. P — задача выпуклого программирования с дифференцируемыми функциями $f_j(x)$, $j = 0, 1, \ldots, m$.

Для задач P и $P^\#$ справедливы утверждения:

1) $\tilde{f} \le \tilde{f}^\#$.

2) Если P разрешима и удовлетворяет тому или иному условию регулярности, то $-\infty < \tilde{f} = \tilde{f}^\# < +\infty$, причем значение $\tilde{f}^\#$ в задаче $P^\#$ достижимо.

3) Если при некотором α множество $M \cap \{x : \Phi(x) \ge \alpha\}$ не пусто и ограничено, то $-\infty < \tilde{f} = \tilde{f}^\# < +\infty$, причем значение $\tilde{f}$ в задаче P достижимо; здесь M — допустимое множество задачи P, $\Phi(x)$ — ее целевая функция.

4) Если C — задача линейного программирования и все нормы, фигурирующие в задаче P, кусочно-линейны, то из непустоты допустимых множеств задач P и $P^\#$ вытекает их разрешимость и совпадение оптимальных значений, т.е. $\tilde{f} = \tilde{f}^\#$.

Замечание. Все нормы, фигурирующие в формулировках задач P и $P^\#$, предполагаются, как и ранее, монотонными.

4. Двойственность для несобственных задач в бесконечномерных пространствах [2, 5]

При перенесении результатов по двойственности на бесконечномерный случай первая бросающаяся в глаза трудность состоит в невозможности применения операции "+", проектирования на E_+^k, фигурирующей в определении задач P и $P^\#$. Различные попытки обобщения этой операции (например, рассмотрение ее как проектирование на конус, сопряженный конусу "неотрицательных векторов" — в гильбертовом пространстве, или обобщение ее с помощью еще более сложной конструкции, использующей уже две пары сопряженных конусов — в банаховом пространстве) приводят к необходимости введения дополнительных ограничений на виды используемых конусов. Так как нам представлялось важным сохранение, с одной

стороны, общности предположений, а с другой — достаточной простоты и естественности всей конструкции, то для обобщения был избран иной путь. А именно, ниже строится пара задач D, $D^{\#}$ общего вида (содержащая в себе, в частности, и все постановки, возникающие в результате обобщения операции "+").

Выпишем общий вид исходной пары двойственных задач линейного программирования в бесконечномерных пространствах:

$$L: \quad \sup\{\langle c, x\rangle \mid b - Ax \in K, \quad x \in C\}, \tag{4.1}$$

$$L^*: \quad \inf\{\langle b, u\rangle \mid A^*u - c \in C^*, \quad u \in K^*\}; \tag{4.2}$$

здесь $x \in X$, $u \in U$, $c \in X^*$, $b \in U^*$; X, U — некоторые вещественные линейные топологические пространства, X^*, U^* — сопряженные им пространства (непрерывных линейных функционалов), A — непрерывный линейный оператор, действующий из X в U^*; A^* — сопряженный ему оператор; $C \subset X$, $K \subset U^*$ — выпуклые замкнутые конусы,

$$C^* = \{x^* \in X^* \mid \langle x^*, x\rangle \geq 0, \quad \forall x \in C\},$$
$$K^* = \{u \in U \mid \langle u, u^*\rangle \geq 0, \quad \forall u^* \in K\}$$

— сопряженные им конусы. Нам будет удобнее считать все упомянутые пространства рефлексивными банаховыми.

Для дальнейшего потребуется конкретизировать задачи L и L^* путем введения разбиения оператора A, аналогично "разрезу" матрицы A задачи (1.2). Это повлечет за собой соответствующее разбиение введенных пространств и конусов. В целях краткости изложения предположим вместо этого сразу, что

$$X = \prod_{i=1}^{n} X_i, \quad U = \prod_{j=1}^{m} U_j, \quad X^* = \prod_{i=1}^{n} X_i^*, \quad U^* = \prod_{j=1}^{m} U_j^*,$$

где X_i, X_i^*, U_j, U_j^*, $i = 1, \ldots, n$, $j = 1, \ldots, m$ — рефлексивные банаховы пространства и (сильно) сопряженные им. Тогда, если положить $\langle x^*, x\rangle = \sum_{i=1}^{n} \langle x_i^*, x_i\rangle$, где $x = [x_1, \ldots, x_n] \in X_1 \times \ldots \times X_n = X$, $x^* = [x_1^*, \ldots, x_n^*] \in X^*$, то X, X^* (и аналогично U, U^*) будут рефлексивными банаховыми пространствами, например, с нормами

$$\|x\| = \sum_{i=1}^{n} \|x_i\|, \quad \|x^*\| = \max_{1 \leq i \leq n} \|x_i^*\|.$$

Здесь и дальше все нормы обозначаются знаком $\|\cdot\|$. Пусть, далее

$$C = C_1 \times \ldots \times C_n, \quad K = K_1 \times \ldots \times K_m,$$
$$C^* = C_1^* \times \ldots \times C_n^*, \quad K^* = K_1^* \times \ldots \times K_m^*,$$

где $C_i \subset X_i$, $C_i^* \subset X_i^*$, $K_j \subset U_j^*$, $K_j^* \subset U_j$ выпуклые замкнутые конусы и сопряженные к ним. Наконец, пусть A_{ji}, $j = 1, \ldots, m$, $i = 1, \ldots, n$ — линейные непрерывные операторы, действующие из X_i в U_j^*; A_{ji}^* — сопряженные им операторы. Таким образом, задачи (4.1), (4.2) могут быть переписаны в виде

$$L: \quad \sup\left\{\langle c, x\rangle = \sum_{i=1}^{n}\langle c_i, x_i\rangle \mid b_j - A_j x \in K_j, \quad j = 1, \ldots, m, \quad x \in C\right\},$$

$$L^*: \quad \inf\left\{\langle b, u\rangle = \sum_{j=1}^{m}\langle b_j, u_j\rangle \mid A_i' u - c_i \in C_i^*, \quad i = 1, \ldots, n, \quad u \in K^*\right\},$$

где по определению $A_j x = \sum\limits_{i=1}^{n} A_{ji} x_i$, $A_i' u = \sum\limits_{j=1}^{m} A_{ji}^* u_j$, $c = [c_1, \ldots, c_n] \in X^*$, $b = [b_1, \ldots, b_m] \in U^*$.

Аналогично конечномерному случаю, зафиксировав номера $0 \le n_0 \le n$, $0 \le m_0 \le m$, выделим подсистемы ограничений задач L и L^*:

$$b_j - A_j x \in K_j, \quad j = 1, \ldots, m_0, \quad A_i' u - c_i \in C_i^*, \quad i = 1, \ldots, n_0.$$

Введем функции Φ и Ψ. Функцию $\Phi(y)$ над пространством $Y = X_{n_0+1} \times \ldots \times X_n \times U_{m_0+1}^* \times \ldots \times U_m^*$ назовем допустимой, если она удовлетворяет условиям:

1) Φ — выпуклая, собственная, полунепрерывная снизу,

$$\operatorname{dom}\Phi \subset C_{n_0+1} \times \ldots \times C_n \times U_{m_0+1}^* \times \ldots \times U_m^*;$$

2) для любого фиксированного $[x_{n_0+1}, \ldots, x_n] \subset C_{n_0+1} \times \ldots \times C_n$ функция $\Phi(x_{n_0+1}, \ldots, x_n, \cdot)$ монотонно не убывает по конусу $K_{m_0+1} \times \ldots \times K_m$.

Функцию $\Psi(z)$ над пространством $Z = X_{n_0+1}^* \times \ldots \times X_n^* \times U_{m_0+1} \times \ldots \times U_m$ назовем допустимой, если она удовлетворяет условиям:

1) Ψ — выпуклая, собственная, полунепрерывная снизу,

$$\operatorname{dom}\Psi \subset X_{n_0+1}^* \times \ldots \times X_n^* \times K_{m_0+1}^* \times \ldots \times K_m^*;$$

2) для любого фиксированного $[u_{m_0+1}, \ldots, u_m] \in K_{m_0+1}^* \times \ldots \times K_m^*$ функция $\Psi(\cdot, u_{m_0+1}, \ldots, u_m)$ монотонно не убывает по конусу $C_{n_0+1}^* \times \ldots \times C_n^*$.

Зафиксировав систему положительных параметров R, ε_i, R_j, r_i, δ_j, удовлетворяющих условиям

$$\varepsilon_i r_i = R_j \delta_j = R_{-1}, \quad i = n_0 + 1, \ldots, n, \quad j = m_0 + 1, \ldots, m,$$

выпишем задачи

$$D: \quad \sup\{\langle c, x\rangle - R\Phi(\varepsilon_{n_0+1} x_{n_0+1}, \ldots, \varepsilon_n x_n; \; R_{m_0+1}(A_{m_0+1}x - b^{m_0+1}), \ldots$$
$$\ldots, R_m(A_m x - b^m)) \mid b_j - A_j x \in K_j, \quad j = 1, \ldots, m_0, \quad x \in C\},$$

$$D^{\#}: \quad \inf\{\langle b,u\rangle + R\Psi(r_{n_0+1}(c_{n_0+1}-A'_{n_0+1}u),\dots r_n(c_n-A'_nu);$$
$$\delta_{m_0+1}u_{m_0+1},\dots,\delta_m u_m) \mid A'_iu-c_i\in C_i^*,\ \ i=1,\dots,n_0,\ \ u\in K^*\},$$

в которых Φ, Ψ — допустимые функции над соответствующими пространствами.

По аналогии с конечномерным случаем, через $\tilde f$, $\tilde f^{\#}$ и M, $M^{\#}$, $\tilde M$, $\tilde M^{\#}$ обозначим оптимальные значения и допустимые и оптимальные множества задач D и $D^{\#}$, через $\Gamma(x)$, $\Gamma^{\#}(u)$ — аргументы функций Φ, Ψ.

Теорема 4.1. Пусть в рассматриваемых задачах D, $D^{\#}$ одна из функций Φ, Ψ является допустимой, а вторая — ей сопряженной. Тогда

1) функция Φ, Ψ допустимы и взаимно сопряжены, $\tilde f\le\tilde f^{\#}$;

2) если $\tilde f<+\infty$ и выполняется условие

$$\exists\overline{x}\in M,\ \Gamma(\overline{x})\in\operatorname{int}\operatorname{dom}\Phi,\ b_j-A_j\overline{x}\in\operatorname{int}K_j,\ j=1,\dots,m_0,$$

то $-\infty<\tilde f=\tilde f^{\#}<+\infty$, причем $\tilde M^{\#}\ne\emptyset$;

3) если $\tilde f^{\#}>-\infty$ и выполняется условие:

$$\exists\overline{u}\in M^*,\ \Gamma^{\#}(\overline{u})\in\operatorname{int}\operatorname{dom}\Psi,\ A_i\overline{u}-c_i\in\operatorname{int}C_i^*,\ i=1,\dots,n_0,$$

то $-\infty<\tilde f=\tilde f^{\#}<+\infty$, причем $\tilde M\ne\emptyset$.

5. Бесконечномерные задачи линейного программирования с разрывом двойственности

Аппроксимация по конусам [6]. Рассмотрим пару двойственных задач бесконечного линейного программирования (БЛП):

$$\inf\{\langle c,x\rangle \mid Ax\in b+Q,\ \ x\in P\}=v, \tag{5.1}$$

$$\sup\{\langle b,u\rangle \mid -A^*u+c\in P^*,\ \ u\in Q^*\}=v^*, \tag{5.2}$$

где (X,Y), (Z,U) — дуальные пары локально выпуклых ТВП с заданными на них билинейными формами; $A: X\to Z$ — линейный непрерывный оператор, $A^*: U\to Y$ — сопряженный к нему; $Q\subset Z$, $P\subset X$ — выпуклые замкнутые конусы, $Q^*\subset U$, $P^*\subset Y$ — сопряженные к ним конусы, определяемые по формуле

$$Q^*=\{u\mid\langle q,u\rangle\ge 0\ \ \forall q\in Q\};\ \ c\in Y,\ \ b\in Z.$$

Определение. Будем говорить, что последовательность выпуклых замкнутых конусов $\{Q_n\}$ строго убывает к конусу Q, и обозначать это $\{Q_n\} \to Q$, если

1. $Q_{n-1} \subset Q_n$, $Q \backslash \{0\} \subset \tau - \operatorname{int} Q_n \ \forall n$, $Q = \bigcap\limits_{n=1}^{\infty} Q_n$;

2. Для некоторого $n \in N$ конус Q_n локально $\sigma(Z,U)$-компактен и $Q_n \backslash \{0\} \subset \tau - \operatorname{int} Q_{n-1}$.

В дальнейшем будем обозначать одной и той же буквой v_n оптимальное значение задачи (5.1) с той или иной аппроксимацией конусов P и Q.

Теорема 5.1. Пусть $\{P_n\} \to P$ и для всех n выполняется соотношение двойственности $v_n = v_n^* \in R$. Тогда либо $\lim\limits_{n\to\infty} v_n = v^* = v$, либо

$$\lim_{n\to\infty} v_n = \sup\{\langle b, u\rangle \mid A^* u = c, \quad u \in Q^*\}. \tag{5.3}$$

Теорема 5.2. Пусть $\{P_n\} \to P$ и $\{Q_n\} \to Q$ и для некоторого n v_n конечно. Если v конечно, то $\lim\limits_{n\to\infty} v_n = v^*$.

Задача бесконечного линейного программирования над R_∞ [7]. Предварительно введем некоторые обозначения: R_∞ — пространство числовых последовательностей $x = (x_1, x_2, \ldots) \in R_\infty$; R'_∞ — подпространство из R_∞, элементы которого $x = (x_1, x_2, \ldots)$ содержат не более конечного числа отличных от нуля компонент x_i; для $a \in R_\infty$ и $x \in R'_\infty$ положим $\langle a, x\rangle = \sum\limits_i a_i x_i$. Пусть

$$\begin{aligned} a_{0\cdot} &= (a_{01}, a_{02}, \ldots) \in R_\infty, \quad b = (b_1, b_2, \ldots) \in R_\infty, \\ a_{i\cdot} &= (a_{i1}, a_{i2}, \ldots) \in R_\infty, \quad i = 1, 2, \ldots, \\ a_{\cdot j} &= (a_{1j}, a_{2j}, \ldots) \in R_\infty, \quad j = 1, 2, \ldots \end{aligned}$$

(т. е. $a_{i\cdot}$ — строка, $a_{\cdot j}$ — колонка матрицы $A_{\cdot\cdot} = (a_{ji})$). Сформулируем пару взаимнодвойственных задач бесконечного линейного программирования над R'_∞:

$$\begin{aligned} L_{\cdot\cdot}: &\quad \inf\{\langle a_{0\cdot}, x\rangle \mid \langle a_{i\cdot}, x\rangle \le b_i \ \ i = 1, 2, \ldots\} = v_{\cdot\cdot}, \\ L^*_{\cdot\cdot}: &\quad \sup\{\langle -b, y\rangle \mid \langle a_{\cdot j}, y\rangle = -a_{0j}, \ \ y \ge 0, \ \ j = 1, 2, \ldots\} = v^*_{\cdot\cdot}. \end{aligned}$$

Выпишем аппроксимирующее семейство конечных задач ЛП:

$$L_{mn}: \quad \min\left\{\sum_{j=1}^{m} a_{0j} x_j \mid \sum_{j=1}^{m} a_{ij} x_j \le b_i, \ \ i = 1, \ldots, n\right\} = v(m,n).$$

Положим

$$\overline{\gamma} = \sup_{\{m_k, n_k\}} \overline{\lim_k}\, v(m_k, n_k), \quad \underline{\gamma} = \inf_{\{m_k, n_k\}} \underline{\lim_k}\, v(m_k, n_k),$$

где операции sup и inf берутся по всем последовательностям $\{m_k\} \to \infty$, $\{n_k\} \to \infty$.

Скажем, что задача $L_{\cdot\cdot}$ финитно-определена, если, начиная с некоторого m_0, система $\left\{ \sum_{j=1}^{m} a_{ij}x_j \le b_i,\ i = 1, 2, \dots \right\}$ совместна и финитно-определена $(m \ge m_0)$ в смысле работы [8].

Теорема 5.3. Пусть задача $L_{\cdot\cdot}$ финитно-определена и значение $v_{\cdot\cdot}$ конечно. Тогда $v_{\cdot\cdot} = \overline{\gamma}$, $v^*_{\cdot\cdot} = \underline{\gamma}$ и для любой последовательности $\{m_k\} \to \infty$ $(\{n_k\} \to \infty)$ существует $\{n_k\}$ $(\{m_k\})$ такая, что $\lim\limits_k v(m_k, \tilde{n}_k) = \overline{\gamma}$ $(\lim\limits_k v(\tilde{m}_k, n_k) = \underline{\gamma})$, как только $\tilde{n}_k > n_k$ $(\tilde{m}_k > m_k)$.

Отсюда следует, что $v_{\cdot\cdot} > v^*_{\cdot\cdot}$ тогда и только тогда, когда $\overline{\gamma} > \underline{\gamma}$. Можно привести пример, иллюстрирующий ситуацию теоремы, когда $v_{\cdot\cdot} > v^*_{\cdot\cdot}$ и произвольная аппроксимирующая последовательность $\{v(m_k, n_k)\}$ сходится либо к $v_{\cdot\cdot}$, либо к $v^*_{\cdot\cdot}$.

Для снятия условия финитной определенности задачи $L_{\cdot\cdot}$ выпишем регуляризованное аппроксимирующее семейство

$$L^{(t)}_{m,n}: \ \min\left\{ \sum_{i=1}^{m+1} \lambda_i \langle a_{0\cdot}, te^m_i \rangle \mid \sum_{i=1}^{m+1} \lambda_i = 1, \ \ \lambda_i \ge 0, \right.$$
$$\left. \sum_{i=1}^{m+1} \lambda_i (\langle a_{k\cdot}, te^m_i \rangle - b_k) \le 0, \ \ k = 1, \dots, n \right\} = v(m, n, t),$$

здесь $e_i \in R^m$, $\operatorname{cone}\{e_1, \dots, e_{m+1}\} = R^m$.

Теорема 5.4. Пусть значение $v_{\cdot\cdot}$ конечно. Тогда для любой последовательности $\{m_k\} \to \infty$ существуют $\tilde{t}(m_k)$ и $\tilde{n}(t_k)$, $\forall t_k > \tilde{t}(m_k)$ такие, что $\lim v(m_k, n_k, t_k) = v_{\cdot\cdot}$, как только $t_k \ge \tilde{t}(m_k)$, $n_k \ge \tilde{n}(t_k)$.

Относительно работ, в которых исследуется аппроксимация оптимальных значений задач с разрывом в двойственности ("предельные лагранжианы", построение "совершенной двойственности") см., например, [9] и имеющуюся там библиографию.

6. Коррекция несобственных задач

Смысл коррекции несобственных задач был разъяснен в пар. 1. Ввиду эквивалентности задачи на экстремум постановке смысла sup inf или inf sup, вопросы коррекции мы проиллюстрируем именно на них. Ниже терминология будет соотнесена с [10].

Множества разрешимости [11]. Вогнутой по $(u, q) \in R^{m_1+n_2}$ и выпуклой по $(x, p) \in R^{m_2+n_1}$ собственной замкнутой функции $F(u, x, p, q)$ $(m_1 +$

$m_2 = m$, $n_1 + n_2 = n$), отображающей R^{m+n} в $R \cup \{-\infty\} \cup \{+\infty\}$ соответствуют задачи на максимин и минимакс

$$\sup_u \inf_x F(u, x, p, q), \quad \inf_x \sup_u F(u, x, p, q) \tag{6.1}$$

(p, q — параметры) и их множества разрешимости:

$$K = \{(p, q) \in R^n \mid F(\cdot, \cdot, p, q) \text{ имеет седловую точку}\}$$
$$K^1 = \{(p, q) \mid -\infty < \sup_u \inf_x F < +\infty\},$$
$$K^2 = \{(p, q) \mid -\infty < \inf_x \sup_u F < +\infty\}.$$

С собственной седловой функцией

$$F^1(u^*, x^*, p, q) = \sup_x \inf_u \{\langle u^*, u\rangle + \langle x^*, x\rangle - F(u, x, p, q)\}$$

свяжем множества ($P_q(\cdot)$ обозначает проекцию на подпространство координат вектора q; cl — замыкание; ri — относительную внутренность):

$$D = \{(p, q) \mid (0, 0, p, q) \in \operatorname{cl} \operatorname{dom} F^1\},$$
$$Q_1 = \{p \mid (0, p) \in \operatorname{cl} \operatorname{dom}_1 F^1 \backslash \operatorname{dom}_1 F^1\} \times (P_q(\operatorname{dom} F) \backslash P_q(\operatorname{ri} \operatorname{dom} F)),$$
$$Q_2 = (P_p(\operatorname{dom} F) \backslash P_p(\operatorname{ri} \operatorname{dom} F)) \times \{q \mid (0, q) \in \operatorname{cl} \operatorname{dom}_2 F^1 \backslash \operatorname{dom}_2 F^1\}.$$

Пусть $\operatorname{cl}_u F$ — замыкание F по u, $\operatorname{real} F = \{(u, x, p, q) \mid F(u, x, p, q) \in R\}$. Введем четыре типа условий на F:

$$A_q: \ P_q(\operatorname{real} \operatorname{cl}_u F) = P_q(\operatorname{dom} F); \quad A_p: \ P_p(\operatorname{real} \operatorname{cl}_x F) = P_p(\operatorname{dom} F);$$

$$A_1: \ P_{(u,q)}(\operatorname{real} \operatorname{cl}_x F) = \operatorname{dom}_1 F; \quad A_2: \ P_{(x,p)}(\operatorname{real} \operatorname{cl}_u F) = \operatorname{dom}_2 F.$$

Теорема 6.1. 1) Если выполнено A_q, то $K^1 \subset D \cup Q_1$. Если выполнено A_p, то $K^2 \subset D \cup Q_2$. При выполнении обоих условий $K \subset K^1 \cap K^2 \subset D$.

2) Если выполнено A_1, A_p, то $K^2 \subset D$. Если выполнено A_2, A_q, то $K^1 \subset D$.

3) Пусть $K \subset D$, $0 \in P_{(u^*, x^*)}(\operatorname{ri} \operatorname{dom} F^1)$. Тогда функция $F_0^1(p, q) \equiv F^1(0, 0, p, q)$ собственна и замкнута, $\operatorname{ri} D = \{(p, q) \mid (0, 0, p, q) \in \operatorname{ri} \operatorname{dom} F^1\} = \operatorname{ri} \operatorname{dom} F_0^1 \subset K \subset \operatorname{cl} \operatorname{dom} F_0^1 = D$.

Коррекция по выпуклому критерию [11]. Постановкам (6.1) сопоставим задачу коррекции вида

$$\inf\{\Phi(p, q) \mid (p, q) \in K \cap S\}, \tag{6.2}$$

где $\Phi(p, q)$ — произвольная выпуклая, собственная, замкнутая функция, S выпукло (в дальнейшем считаем $S = \operatorname{dom} \Phi$). Задача (6.2) может использоваться (ввиду свободы в выборе Φ, S) и как инструмент исследования множеств разрешимости.

Пусть $(p_2, q_1) \in \operatorname{ri} K$. Точки p_1^*, q_2^* выберем либо из условий

$$p_1^* \in \operatorname{ri}\{p^* \mid (0,0,-p^*) \in \partial_{u,x,p} F(u,x,p,q_1),\ (u,x,p) \in R^{m+n_1}\},$$
$$q_2^* \in \operatorname{ri}\{q^* \mid (0,0,-q^*) \in \partial_{u,x,q} F(u,x,p_2,q),\ (u,x,q) \in R^{m+n_2}\}, \qquad (6.3)$$

либо из условий (6.3)′, получающихся из (6.3) исключением знаков ri. Конкретно p_1^*, q_2^* соответствуют субградиенты в точках:

$$(0,0,-p_1^*) \in \partial_{u,x,p} F(u_1,x_1,p_1,q_1),\ (0,0,-q_2^*) \in \partial_{u,x,q} F(u_2,x_2,p_2,q_2).$$

Пусть $c_1 = \langle p_1^*, p_1\rangle + F(u_1,x_1,p_1,q_1)$, $c_2 = \langle q_2^*, q_2\rangle + F(u_2,x_2,p_2,q_2)$. Множество

$$\begin{aligned}\operatorname{dom} H_{\alpha,\beta} &= \{(u,y,v,x,p,q) \mid (p,q) \in \operatorname{dom}\Phi,\ (u,x,p,q_1) \in \operatorname{dom} F,\ (v,y,p_2,q) \in \operatorname{dom} F\\ &= \{(u,y,v,x,p,q) \mid (u,y) \in \operatorname{dom}_1 H_{\alpha,\beta},\ (v,x,p,q) \in \operatorname{dom}_2 H_{\alpha,\beta}\}\end{aligned}$$

является эффективной областью вогнутой по (u,y) и выпуклой по (v,x,p,q) функции $H_{\alpha,\beta}(u,y,v,x,p,q)$, вводимой следующим образом:

$$\begin{aligned}H_{\alpha,\beta} =& \Phi(p,q) + \alpha F(u,x,p,q_1) + \alpha\langle p_1^*, p\rangle - \beta F(v,y,p_2,q) - \beta\langle q_2^*, q\rangle - \alpha c_1 + \beta c_2,\\ &(u,y,v,x,p,q) \in \operatorname{dom} H_{\alpha,\beta},\\ H_{\alpha,\beta} =& +\infty,\ (u,y) \in \operatorname{dom}_1 H_{\alpha,beta},\ (v,x,p,q) \notin \operatorname{dom}_2 H_{\alpha,beta},\\ H_{\alpha,\beta} =& -\infty,\ (u,y) \notin \operatorname{dom}_1 H_{\alpha,\beta}.\end{aligned}$$

Здесь α, β — произвольные положительные параметры. Пусть $S_{\alpha,\beta}$ — множество седловых точек $H_{\alpha,\beta}$, σ — оптимальное значение задачи (6.2), δ — индикаторная функция,

$$\Phi_0(p,q) = \Phi(p,q) + \delta((p,q) \mid \operatorname{cl} K),\ \ S'_{\alpha,\beta} = P_{(p,q)}(S_{\alpha,\beta}),$$

$$\sigma(\alpha,\beta) = \sup_{u,y}\inf_{x,v,p,q} H_{\alpha,\beta},\ \ d(A,B) = \sup_{a\in A}\inf_{b\in B}|a-b|.$$

В следующей теореме решение задачи (6.2) сведено к нахождению седловых точек вспомогательной функции $H_{\alpha,\beta}$.

Теорема 6.2. Пусть $K \subset D$, $-\infty < \sigma < +\infty$, $0 \in P_{(u^s,x^s)}(\operatorname{ri}\operatorname{dom} F^1)$, $\operatorname{ri}\operatorname{dom}\Phi \cap \operatorname{ri} K \neq \emptyset$ и выполняется (6.3)′. Тогда при любых $\alpha > 0$, $\beta > 0$:

1) Функция $H_{\alpha,\beta}$ собственна и замкнута;

2) $\sigma(\alpha,\beta)$ конечно, не зависит от порядка применения операций sup, inf в его определении, монотонно возрастает по $\alpha > 0$, $\beta > 0$,

$$\sigma = \lim_{\alpha\to+0,\beta\to+0}\sigma(\alpha,\beta) = \inf_{\alpha>0,\beta>0}\sigma(\alpha,\beta);$$

3) $S'_{\alpha,\beta} \subset \operatorname{dom} F^1(0,0,\cdot,\cdot) \cap \operatorname{dom} \Phi \subset \operatorname{cl} K \cap \operatorname{dom} \Phi$;

4) $\sigma \le \Phi(p,q) \le \sigma(\alpha,\beta) \quad \forall (p,q) \in S'_{\alpha,\beta}$.

Если, кроме того, $0 \in \operatorname{ri} \operatorname{dom} \Phi_0^*$ и выполняется (6.3), то

5) $S_{\alpha,\beta} \ne \emptyset$, более того, $0 \in \operatorname{ri} \operatorname{dom} H^*_{\alpha,\beta}$ $(\forall \alpha > 0, \beta > 0)$;

6) $S_0 = \operatorname{Arg\,min} \Phi_0 \ne \emptyset$, $d(S'_{\alpha,\beta}, S_0) \to 0 (\alpha \to +0,\ \beta \to +0)$.

Коррекция по вогнуто-выпуклому критерию. При выполнении условий пункта 3) теоремы 6.1 $\operatorname{cl} K = \{(p,q) \mid p \in P,\ q \in Q\}$, где P, Q выпуклы и замкнуты. Рассмотрим задачу коррекции

$$\sup_{p \in P} \inf_{q \in Q} \Psi(p,q), \tag{6.4}$$

где $\Psi(p,q)$ вогнуто-выпукла, собственна и замкнута. Для (6.4) справедлив результат, аналогичный по форме теореме 6.2 (см. [11]). Для случая, когда F имеет вид

$$F(u,x,p,q) = F_0(u,x) - \langle p,u\rangle - \langle q,x\rangle$$

(F_0 выпукло-вогнута, собственна, замкнута), можно предложить более простой метод решения задачи (6.4) (и задачи (6.2) с сепарабельной функцией цели $\Phi(p,q) = \Phi_1(p) + \Phi_2(q)$), основанный на несколько ином подходе. Заметим, что для данного вида F всегда $\operatorname{cl} K = P \times Q \ne \emptyset$.

Введем $G_\alpha = \Psi(p,q) - \alpha F(u,x,p,q)$ — вогнуто-выпуклую при $\alpha > 0$ функцию аргументов $z = (p,x)$, $w = (q,u)$ и условия

$$A_3: \quad \operatorname{ri} \operatorname{dom}_1 \Psi \cap \operatorname{ri} \operatorname{dom}_1 F_0^* = P_0 \ne \emptyset, \quad Q_0 = \operatorname{ri} \operatorname{dom}_2 \Psi \cap \operatorname{ri} \operatorname{dom}_2 F_0^* \ne \emptyset,$$

A_4: существуют $p_0 \in P_0$, $q_0 \in Q_0$ такие, что все множества уровня функций $\Psi(p_0,\cdot) + \delta(\cdot \mid \operatorname{dom}_2 F_0^*)$ и $\Psi(\cdot,q_0) - \delta(\cdot \mid \operatorname{dom}_1 F_0^*)$ ограничены (здесь и выше F_0^* — функция, сопряженная к F_0).

При выполнении A_3 функция $G_\alpha(z,w)$ является собственной и замкнутой, а при выполнении A_4 для достаточно малых $\alpha > 0$ имеет конечное седловое значение σ_α и не пустое множество седловых точек S_α [10]. В следующей теореме решение задачи (6.4) сведено к отысканию точек из S_α.

Теорема 6.3 [12]. Пусть выполнены условия A_3, A_4. Тогда

1) $S_\alpha \ne \emptyset$ при достаточно малых $\alpha > 0$;

2) $(\overline{p},\overline{x},\overline{q},\overline{u}) \in S_\alpha \Rightarrow (\overline{u},\overline{x})$ — седловая точка функции $F(u,x,\overline{p},\overline{q})$, так что $S'_\alpha = P_{(p,q)}(S_\alpha) \subset K$;

3) $\sigma_\alpha \to \sigma$, $d(S'_\alpha, S_0) \to 0$ при $\alpha \to +0$; здесь σ — оптимальное значение в (6.4), S_0 — непустое множество седловых точек функции Ψ относительно $P \times Q$.

Если же в задаче (6.4) $S_0 \subset K$, то

4) $d(S'_\alpha, S_1) \to 0$ при $\alpha \to +0$, где S_1 — непустое множество седловых точек функции F_0^* относительно S_0.

В случае, когда значения Ψ^* вычисляются сравнительно легко, метод может быть значительно упрощен. Пусть $g_\alpha(u,x) = \alpha F_0(u,x) - \Psi^*(-\alpha u, -\alpha x)$. Обозначим через κ_α седловые значения, а через z_α — множество седловых точек этой вогнуто-выпуклой функции. Имеет место

Следствие [12]. Если выполнены все предположения теоремы 6.3 и g_α замкнута, то

1) $Z_\alpha \neq \emptyset$ при достаточно малых $\alpha > 0$;

2) $(\overline{u},\overline{x}) \in Z_\alpha \Rightarrow (\overline{u},\overline{x})$ — седловая точка функции $F(u,x,\overline{p},\overline{q})$ при всех $(\overline{p},\overline{q}) \in \partial F_0(\overline{u},\overline{x}) \cap (-\alpha^{-1}\partial\Psi^*(-\alpha\overline{u},-\alpha\overline{x}))$;

3) $\kappa_\alpha \to -\sigma$, $d(Z'_\alpha, S_0) \to 0$ при $\alpha \to +0$; здесь

$$Z'_\alpha = \bigcup_{(u,x)\in Z_\alpha} \partial F_0(u,x) \cap (-\alpha^{-1}\partial\Psi^*(-\alpha u, -\alpha x)).$$

Если же для задачи (5.4) $S_0 \subset K$, то

4) $d(Z'_\alpha, S_1) \to 0$ при $\alpha \to +0$.

Заметим, что если Ψ — сильно вогнуто-выпукла, то Ψ^* — всюду конечна и дифференцируема, так что

$$Z'_\alpha = \bigcup_{(u,x)\in Z_\alpha} -\alpha^{-1}\nabla\Psi^*(-\alpha u, -\alpha x).$$

Например, для квадратичной функции $\Psi(p,q) = (\|q\|^2 - \|p\|^2)\cdot\frac{1}{2}$ имеем равенства:

$$g_\alpha(u,x) = \alpha F_0(u,x) - \frac{\alpha^2}{2}(\|u\|^2 - \|x\|^2), \quad Z'_\alpha = \alpha(\overline{u}_\alpha, \overline{x}_\alpha),$$

где $\{(\overline{u}_\alpha, \overline{x}_\alpha)\} = Z_\alpha$.

Некоторые другие подходы к коррекции несобственных задач и несовместных систем уравнений и неравенств, а также приложения можно найти, например, в [1, 11, 13–20].

7. Заключение

В данном обзоре были изложены, в основном, вопросы теории несобственных задач линейного и выпуклого программирования, а именно — теории двойственности. Меньшее внимание уделено вопросам коррекции таких задач, т. е. методам построения компромиссных моделей по тому или иному критерию компромисса. Однако этот вопрос несомненно важный, в частности, с прикладной точки зрения. Для методов поиска компромиссных моделей создаются пакеты прикладных программ. Можно указать, например, на пакет ДЕЛЬТА-ПЛАН-ЕС-ЭВМ, разработанный в Институте математики и механики Уральского отделения АН СССР.

Достаточно большая библиография (около 200 работ) по несобственным задачам математического программирования приведена в [13].

Литература

1. Еремин И.И., Мазуров В.Д., Астафьев Н.Н. Несобственные задачи линейного и выпуклого программирования. М. Наука, 1983.
2. Еремин И.И., Ватолин А.А. Двойственность для несобственных задач математического программирования. Препринт. Свердловск, УНЦ АН СССР, 1985, 52 с.
3. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М. Наука, 1979.
4. Еремин И.И. Двойственность для несобственных задач линейного и выпуклого программирования. ДАН СССР, 1981, **256**, *2*, 272–276.
5. Еремин И.И., Ватолин А.А. Двойственность для несобственных бесконечномерных задач линейного и выпуклого программирования. Методы аппроксимации несобственных задач математического программирования. Сб. статей. Свердловск, УНЦ АН СССР, 1984, 3–20.
6. Трофимов С. П. Анализ соотношений двойственности для задач полубесконечного и бесконечного линейного программирования. Автореф. диссерт., Свердловск, ИММ УрО АН СССР, 1988, 18 с.
7. Астафьев Н. Н. Линейные неравенства и выпуклость. М. Наука, 1982.
8. Черников С. Н. Линейные неравенства. М. Наука, 1968.
9. *Karney D. F., Morley T. D.* Limiting Lagrangeans: A primal approach. J. Opt. Theory and Appl., 1986, **48**, *1*, 163–174.
10. Рокафеллар Р. Т. Выпуклый анализ. М. Мир, 1973.
11. Ватолин А. А. Множества разрешимости и коррекция седловых функций и систем неравенств. Препринт. Свердловск, УрО АН СССР, 1989, 90 с.
12. Попов Л. Д. Линейная коррекция несобственных выпукло-вогнутых минимаксных задач по максиминному критерию. ЖВМ и МФ, 1986, **26**, *9*, 1325–1338.
13. Еремин И. И. Противоречивые модели оптимального планирования. М. Наука, 1988.
14. Ватолин А. А. О задачах линейного программирования с интервальными коэффициентами. ЖВМ и МФ, 1984, **24**, *11*, 1629–1637.
15. Скарин В. Д. Об одном подходе к анализу несобственных задач линейного программирования. ЖВМ и МФ, 1986, **26**, *3*, 439–448.
16. Плотников С. В. О циклическом проектировании на систему выпуклых множеств с пустым пересечением. Несобственные задачи оптимизации. Свердловск, УНЦ АН СССР, 1982, 60–66.
17. Фролов В. Н. Оптимизация плановых программ при слабо согласованных ограничениях. М. Наука, 1986.
18. Вересков А. И., Третьяков Н. В. Об использовании модифицированных функций Лагранжа для корректировки несовместных задач выпуклого программирования. Известия АН СССР Техн. киберн., 1988, *1*, 3–12.
19. Булавский В. А. Обобщение решения и регуляризация систем неравенств. Вычислительные методы линейной алгебры. Новосибирск, Наука, 1985, 161–174.
20. *Mangasarian O. L.* Normal solutions of linear programs. Math. Progr. Study, 1984 **22**, 206–216.

Typesetted by TYPOTEX GT, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

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: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

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